GENOMIC PERSPECTIVES ON EVOLUTION IN BRACKEN FERN

by

Joshua P. Der

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biology

Approved:

Paul G. Wolf
Major Professor

Michael E. Pfrender
Committee Member

Carol D. VonDohlen
Committee Member

Karen Mock
Committee Member

Paul Cliften
Committee Member

Byron Burnham
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2010
ABSTRACT

Genomic Perspectives on Evolution in Bracken Fern

by

Joshua P. Der, Doctor of Philosophy
Utah State University, 2010

Major Professor: Paul G. Wolf
Department: Biology

The fern genus *Pteridium* comprises a number of closely related species distributed throughout the world. Collectively they are called bracken ferns and have historically been treated as a single species, *Pteridium aquilinum*. Bracken is notorious as a toxic weed that colonizes open fields and poisons livestock. Bracken is also easily cultured and has become one of the most intensively studied ferns. Bracken has been used as a model system for the study of the fern life cycle, fern gametophyte development, the pheromonal mechanism of sex determination, toxicology, invasion ecology, and climate change. This dissertation places bracken within a global evolutionary perspective and establishes bracken as an emerging system for evolutionary genomics in ferns. Bracken samples from around the world were examined for chloroplast DNA variation to infer historical phylogenetic and biogeographic evolutionary events. New high-throughput DNA sequencing technologies and bioinformatic approaches were used to determine the complete chloroplast genome sequence in bracken, to identify novel RNA editing sites in chloroplast transcripts, and to identify gene sequences that are expressed in the gametophyte stage of the fern life
cycle. These data represent an important genomic resource in ferns and were examined within a functional and evolutionary perspective. Several novel approaches and analyses were developed in the course of this research. Results presented in this dissertation provide novel insights into fern biology and land plant evolution.

(105 pages)
I dedicate this dissertation to my family.

To Kristal, whose love and support has made this work possible.

To Cora and Tyler who have shown me the love and joy of fatherhood.

To my parents who fostered in me a sense of curiosity and a love of nature.

To Jessica, whose friendship helped shape the man that I have become;

your memory lives on in my heart.
ACKNOWLEDGMENTS

I wish to thank my advisor, Paul Wolf, and my committee members, Michael Pfrender, Carol vonDohlen, Paul Cliften, Eugene Schupp, and Karen Mock. I also wish to thank my coauthors on published and future manuscripts resulting from this work for significant contributions to the research presented in this dissertation. In addition to Paul Wolf, these people include John Thomson and Jeran Stratford (Chapter 2); Aaron Duffy, Matt Kusner, Chen Gu, and Paul Overvoorde (Chapter 3); and Michael Barker, Norman Wickett, and Claude dePamphilis (Chapter 4).

Numerous people contributed plant material used in Chapter 2 and are credited in Appendix B. Jessie Roper, Mark Ellis, Aaron Duffy, Jacob Davidson, Jeran Stratford, Amanda Grusz, Eric Wafula, Paul Cliften and Michael Pfrender assisted with lab or bioinformatics work. Alan Smith, Laura Forrest, Mark Ellis, Aaron Duffy, and Kristal Watrous provided useful comments on portions of text or complete manuscripts included in this dissertation.

This research was funded in part by National Science Foundation grant DEB-0228432 to Paul Wolf. The Genomics Education Partnership with grant 52005780 from the Howard Hughes Medical Institute to Professor Sarah Elgin (Washington University), with additional support from the National Human Genome Research Institute, and the The Genome Center at Washington University School of Medicine contributed a portion of the genome sequence data used in Chapter 3. Jeffrey Boore and Brent Mishler assisted with the collection of another portion of the genome sequence data. Kenneth White and Ninglin Yin donated resources from Center for Integrated Biosystems (CIB) at Utah State University (USU) to collect additional genome sequence data.
The CIB provided initial funding for my proposal to sequence expressed genes in bracken gametophytes with a student research grant. Additional funds from the Office of the Vice President for Research at USU, in a research catalyst grant to Paul Wolf, enabled me to utilize the Roche 454 sequencing technology to analyze the complete gametophyte transcriptome in bracken. Keithanne Mockaitis supervised the cDNA library preparation and transcriptome sequencing at the Center for Genomics and Bioinformatics at Indiana University.

Computer time from the Center for High Performance Computing at Utah State University is acknowledged. The computational resources were provided through the National Science Foundation under Grant No. CTS-0321170 with matching funds provided by Utah State University.

I also want to thank the friends I have made during my stay in Utah: Ellen Klinger, Mark Ellis, Craig Faulhaber, Debi Fisk, Jake Davidson, Ryan O'Donnell, Maure Smith, Dave Brown, Sid Wright, and the Strange, Scoville, Spaulding, Latta, Wilson, Burningham, and Duffy families. You all have encouraged me and helped me pass the time, each in your own special ways.

Joshua P. Der
CONTENTS

Page

ABSTRACT ........................................................................................................................................ ii

DEDICATION ..................................................................................................................................... iv

ACKNOWLEDGMENTS .................................................................................................................... v

LIST OF TABLES .............................................................................................................................. viii

LIST OF FIGURES ........................................................................................................................... ix

CHAPTER

1. INTRODUCTION. 1 ..............................................................................................................................

2. GLOBAL CHLOROPLAST PHYLOGENY AND BIOGEOGRAPHY OF BRACKEN (Pteridium: Dennstaedtiaceae) ............................................................. 7

3. COMPLETE CHLOROPLAST GENOME SEQUENCE AND HIGH LEVELS OF RNA EDITING IN THE FERN PTERIDUM AQUILINUM DETERMINED BY MASSIVELY PARALLEL PYROSEQUENCING .......................................................... 32

4. DE NOVO CHARACTERIZATION OF THE GAMETOPHYTE TRANSCRIPTOME IN BRACKEN FERN, PTERIDUM AQUILINUM ....................... 49

5. CONCLUSIONS .............................................................................................................................. 79

APPENDICES

A. PERMISSION TO REPRINT CHAPTER 2 .................................................................................. 82

B. VOUCHER INFORMATION FOR CHAPTER 2 ............................................................................. 88

C. ADDITIONAL FILES FOR CHAPTER 4 ...................................................................................... 91

CURRICULUM VITAE ....................................................................................................................... 92
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-1</td>
<td>Transcriptome sequence statistics</td>
</tr>
<tr>
<td>4-2</td>
<td>Assembly summary statistics</td>
</tr>
<tr>
<td>4-3</td>
<td>Transcriptome coverage estimates</td>
</tr>
<tr>
<td>4-4</td>
<td>Taxonomic distribution of unigene blastx hits in the nr database</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Bayesian phylogeny</td>
<td>17</td>
</tr>
<tr>
<td>2-2</td>
<td>Biogeography of bracken</td>
<td>19</td>
</tr>
<tr>
<td>3-1</td>
<td>Gene map of the <em>Pteridium aquilinum</em> chloroplast genome</td>
<td>39</td>
</tr>
<tr>
<td>3-2</td>
<td>Chloroplast genome sequence coverage</td>
<td>41</td>
</tr>
<tr>
<td>3-3</td>
<td>Density of RNA editing sites</td>
<td>43</td>
</tr>
<tr>
<td>4-1</td>
<td>Overview of <em>P. aquilinum</em> transcriptome sequencing and assembly</td>
<td>54</td>
</tr>
<tr>
<td>4-2</td>
<td>Unigene accumulation curve</td>
<td>58</td>
</tr>
<tr>
<td>4-3</td>
<td>Distribution of GO-slim functional categories</td>
<td>61</td>
</tr>
<tr>
<td>4-4</td>
<td>Unigene overlap with <em>Arabidopsis, Physcomitrella,</em> and <em>Selaginella</em></td>
<td>63</td>
</tr>
</tbody>
</table>
Evolutionary studies are increasingly looking to molecular genetic data to inform inference because of the wealth of historical information archived in the genome. Early evolutionary genetic studies relied on markers such as DNA restriction sites or various types of DNA fragment profiles, so called "genetic fingerprints" (Avise, 2004). Direct access to the historical record in the genome has been enabled by methods for sequencing DNA (Hillis et al., 1996; Soltis et al., 1998), allowing us to connect an organism’s genotype with variation in form and function in cells, tissues, and whole organisms. New methods in molecular biology are rapidly advancing our knowledge of gene and genome function using tools such as gene expression microarrays, genetic knockout mutants, and CHip- and Methyl-seq enrichment procedures. Despite the advances these methods are driving in biotechnology, it is critical to place these discoveries within an evolutionary context to fully comprehend these complex systems.

Evolutionary studies, especially those using genetic data, often face a trade-off between resources devoted to sampling breadth (i.e. number of units to be surveyed) and sampling depth (i.e. the quantity of data collected about each unit). For example, when inferring the evolutionary relationships among members of a group, a researcher must decide how many individuals/taxa to sample, often at the expense of the number of character traits that can be collected and analyzed. How this balance shifts will largely be determined by the research objectives, the scope and scale of the project, and the available resources to conduct the research.
As the cost of DNA sequencing drops with new advances in high-throughput technologies (Shendure and Ji, 2008; Ansorge, 2009), the trade-off between sampling breadth and depth is becoming less important for classic types of evolutionary studies. However, high-throughput DNA sequencing is enabling new kinds of evolutionary analysis on a whole-genome scale (Rokas and Abbot, 2009). Ironically, studies using massive throughput sequencing technologies continue to push the limits of what is possible within the bounds of current technology and available resources, bringing the trade-off between sampling breadth and depth to a whole new scale.

Often, different sampling strategies are required to examine evolutionary processes in different aspects of a study system. These alternative sampling strategies can be complementary, and yield results that provide insight for interpreting the results from other components of the research program. This dissertation seeks to highlight the reciprocal illumination garnered by utilizing dramatically different approaches to broadly and deeply sample genetic data to infer historical and contemporary evolutionary processes. I present three studies, which use traditional, and emergent, high-throughput DNA sequencing approaches to sample genetic information to examine evolution in the bracken fern, *Pteridium*.

*Pteridium* is a cosmopolitan fern genus comprised of several closely related species, which are well differentiated from other genera in the family Dennstaedtiaceae (Tryon, 1941). Bracken is notorious as a weed in open fields and is toxic to people and livestock (Holm et al., 1997). Despite its toxicity, bracken is eaten as a delicacy in several parts of the world and, due to its often high local abundance and large coarse stature, is sometimes used as thatching or packing material. Because bracken is common, easily cultured and manipulated, and can impose large economic
consequences, it has become one of the most intensively studied ferns. Bracken has been used as a model system for the study of the fern life cycle (Webster and Steeves, 1958; Gottlieb, 1959; Takahashi, 1961; DeMaggo and Raghavan, 1972; Sheffield and Bell, 1981; Sheffield, 1984, 1985, 2008; Wynn et al., 2000), gametophyte development and the pheromonal mechanism of sex determination (Steeves et al., 1955; Näf, 1958; Whittier, 1962; Sobota and Partanen, 1966; Bell and Duckett, 1976; Ashcroft and Sheffield, 2000; Robertson, 2002; Schneller, 2008), cyanogenesis (Cooper-Driver and Swain, 1970), carcinogenesis (Potter and Baird, 2000; Alonso-Amelot and Avendano, 2002; Beniston and Campo, 2003), invasion ecology (Schneider, 2004; Rodrigues da Silva and Matos, 2006; Ghorbani et al., 2007), and climate change (Pakeman and Marrs, 1996).

Chapter 2 sets the stage for all subsequent evolutionary work in the genus *Pteridium* by providing a sound phylogenetic and biogeographic framework from which I can appropriately sample taxa for detailed studies. I adopted a broad and shallow sampling strategy to examine ancient historical events in *Pteridium*, which led to speciation and the spread of bracken around the world. This study examined chloroplast DNA variation within *Pteridium* in a global phylogenetic context to ascertain evolutionary patterns of divergence, dispersal and colonization. These data were also used to examine the maternal ancestry of two tetraploid taxa hypothesized to be of hybrid origin.

Chapter 3 utilized a focused and deep sampling strategy to obtain the complete chloroplast genome sequence of *Pteridium aquilinum* ssp. *aquilinum* which, combined with deep RNA sequencing, enabled a detailed examination of RNA editing in the chloroplast transcriptome. The complete nucleotide sequence of the chloroplast genome
enabled us to take a comparative genomic approach to assess structural and sequence level changes in the chloroplast genomes among leptosporangiate ferns.

Chapter 4 examined the same deep RNA sequencing data set presented in Chapter 3, but takes advantage of a physical sample preparation procedure (cDNA normalization) and a shift in perspective to broadly sample expressed genes in the very large nuclear genome of Pteridium (nearly 10 Gb, three times the size of the human genome). These sequences were derived from the gametophyte stage of the plant life cycle, providing a fertile data set for investigation into the evolution of life history and reproduction in ferns. These data also provide the first systems-level view of genome composition in ferns, which I examined within a comparative evolutionary genomics framework. I also bioinformatically mined these sequences to infer the functional composition of genes expressed in this life stage.

Together these three studies illustrate the complementarity of different sampling strategies to develop a research program in evolutionary genetics and genomics in a non-model organism.

LITERATURE CITED


ABSTRACT

Bracken ferns (genus *Pteridium*) represent an ancient species complex with a natural worldwide distribution. *Pteridium* has historically been treated as comprising a single species, but recent treatments have recognized several related species. Phenotypic plasticity, geographically structured morphological variation, and geographically biased sampling have all contributed to taxonomic confusion in the genus. We sampled bracken specimens worldwide and used variable regions of the chloroplast genome to investigate phylogeography and reticulate evolution within the genus. Our results distinguish two major clades within *Pteridium*, a primarily northern hemisphere Laurasian/African clade, which includes all taxa currently assigned to *P. aquilinum*, and a primarily southern hemisphere Austral/South American clade, which includes *P. esculentum* and *P. arachnoideum*. All European accessions of *P. aquilinum* subsp. *aquilinum* appear in a monophyletic group and are nested within a clade containing the African *P. aquilinum* taxa (*P. aquilinum* subsp. *capense* and *P. aquilinum* subsp. *centrall-africanum*). Our results allow us to hypothesize the maternal progenitors of two allotetraploid bracken species, *P. caudatum* and *P. semihastatum*. We also discuss the biogeography of bracken in the context of the chloroplast phylogeny. Our

---

study is one of the first to take a worldwide perspective in addressing variation in a broadly distributed species complex.

INTRODUCTION

Describing any biological characteristic of organisms requires adequate sampling so that statements are universally valid. Thus, one should consider the range of variation within the taxon (be it taxonomic, morphological, physiological, ecological, genetic, or geographic) to accurately describe diversity within the group (Hillis, 1998). To capture such variation, one must sample sufficiently to understand both the limits and typical ranges of variation. Such sampling means collecting and examining representatives across the distribution of the taxon. In general, sampling across a taxon’s range is easier to do for lower taxonomic levels because successively smaller clades encompass increasingly narrow ranges. However, sampling across geographic ranges becomes logistically challenging for species or species-complexes with worldwide (or nearly so) distributions.

Examples of widespread species for which a global perspective has been taken include highly mobile birds (Burg and Croxall, 2004), invertebrates (Lee, 2000; Boyer et al., 2007), and fungi (Hibbett, 2001; Banke and McDonald, 2005). Within green plants, worldwide distributions are common among “invasive” species, i.e., those with recent (less than 500 years), human-mediated, widespread distributions. Examples include aquatic weeds (Barrett, 1989) among others (Holm et al., 1997). At least one study has examined evolutionary history in a widespread invasive species, Cardamine flexuosa (Lihová et al., 2006), but few examples, if any, have examined the phylogenetic structure of a species complex with an ancient (preagricultural; 10,000 years) worldwide
distribution (but see Bakker et al., 1995). For a species to occupy such a broad range, it must have a wide ecological amplitude (i.e., ecological valence) or be able to adapt quickly to a wide range of local environmental conditions without speciation. Additionally, to achieve a wide distribution, a species must have a high dispersal and colonization ability, or be an old taxon with a broad ancestral distribution, or both. The genus *Pteridium* Gled. ex Scop. (bracken fern, Dennstaedtiaceae) represents one such taxon, having achieved a natural worldwide distribution and occupying diverse habitats (Page, 1976, 1986; Holm et al., 1997). Fossil evidence indicates that bracken had achieved a worldwide distribution by the Oligocene, ~23.8 mya (reviewed by Page, 1976), and several lines of evidence indicate that bracken can disperse and establish following long distance dispersal by spores (Punetha, 1991; Rumsey et al., 1991).

*Pteridium* is often treated as a monotypic genus after Tryon (1941), but most contemporary systematists recognize the genus as a species complex in need of taxonomic revision (Page, 1976; Brownsey, 1989; Page and Mill, 1995; Thomson, 2000b). Bracken ferns are easily recognized and well differentiated from other genera in Dennstaedtiaceae, and many infrageneric taxa within *Pteridium* have high levels of geographically based morphological structure. However, great confusion in defining infrageneric taxa has resulted from the fact that bracken has high levels of phenotypic plasticity, few diagnostic morphological characters, and the presence of intermediate phenotypes where different morphological forms come into contact, demonstrating that reproductive barriers are incomplete (Page, 1976). These factors, coupled with local taxonomic judgments based on geographically biased sampling, have led to a large number of local forms being described as new species, subspecies, or varieties, resulting in a multiplicity of names (Tryon, 1941; Page, 1976; Thomson, 2000a). The
genus is distributed worldwide and is notorious as a weed because of its exceptional ability to grow rhizomatously in dense patches, overgrowing open fields and pasture (Tryon, 1941; Holm et al., 1997).

Bracken ferns have a long and complex taxonomic history. The first bracken species were described by Linnaeus in the genus *Pteris* L. (Linnaeus, 1753). Later authors followed this generic circumscription, but Agardh (1839) was the first to examine specimens worldwide and set the brackens apart as section *Ornithopteris* J.Agardh (Tryon, 1941; Brownsey, 1989). Later, Hooker (1858), in a comprehensive treatment of *Pteris*, subsumed all brackens as varieties of *Pteris aquilina* L. Various authors segregated the brackens from *Pteris*, but it was not until Kuhn (1879) defined *Pteridium aquilinum* (L.) Kuhn that the brackens were widely accepted as a distinct genus (Tryon, 1941). The last global revision of the genus was Tryon’s (1941) monograph, in which he reduced more than 135 previously named variants into a single species, with two subspecies containing 12 varieties. Subsequent authors have continued to modify the taxonomy of bracken, but most works reflect a geographically limited perspective (Brownsey, 1989; Page and Mill, 1995; Wolf et al., 1995; Speer, 2000; Thomson and Alonso-Amelot, 2002; Gureyeva and Page, 2005; Thomson et al., 2005, 2008). Recent evidence suggests that two *Pteridium* taxa have allopolyploid hybrid origins, and these two tetraploids are currently recognized as segregate species (Tan and Thomson, 1990b; Thomson, 2000a, b; Thomson and Alonso-Amelot, 2002).

Recent work has reexamined the systematic utility of morphological characters (Thomson and Martin, 1996), characterized the structure of the chloroplast genome (Tan and Thomson, 1990a; Tan, 1991), and used genetic fingerprinting to examine evolutionary history in *Pteridium* (Thomson, 2000a, b). Much of this work contributes
toward a taxonomic revision of the genus. The typification of *Pteridium aquilinum* (L.) Kuhn has been revisited, and nomenclature within ‘latiusculum’ morphotypes (N. America, Europe and northeast Asia) has been clarified (Thomson, 2004; Thomson et al., 2008).

Global perspectives on phylogeography can be inferred using chloroplast DNA sequence variation to detect and reconstruct historical evolutionary events, including population demographics, migration, colonization, and both ancient and contemporary hybridization events (Rieseberg and Soltis, 1991; McCauley, 1995; Rieseberg et al., 1996; Ennos et al., 1999). Variation in chloroplast sequence data has been used to reconstruct phylogenetic relationships among plants as diverse and ancient as land plants to variation among recently diverged populations within a single species. *Pteridium* is a well-defined and evolutionarily isolated genus comprising several closely related taxa. Examining patterns of chloroplast variation presents an excellent opportunity to infer patterns of widespread dispersal, colonization, and divergent evolutionary history on a global scale. This study represents one of the only studies of worldwide variation in a terrestrial plant.

The objectives of this present study are threefold. First, we examine chloroplast DNA variation within *Pteridium* in a global phylogenetic context to ascertain evolutionary patterns of divergence, dispersal and colonization. Second, we examine maternal ancestry in the two tetraploid taxa hypothesized to be of hybrid origin. Third, we establish a clear framework for developing hypotheses of evolutionary history that can be tested with comparative nuclear sequence data. These results also provide important information necessary for taxonomic revision of infrageneric taxa in *Pteridium*. 
MATERIALS AND METHODS

**Taxonomic sampling**—

Sampling was designed to cover the range of morphological and geographical diversity within *Pteridium*, representing nearly all currently recognized species, and most infraspecific taxa, with the exception of *P. aquilinum* subsp. *feei* (W. Schaffn. ex Fée) J. A. Thomson, Mickel & Mehltreter, endemic to central Mexico. Determination of specimens used in this study follow Thomson and Alonso-Amelot (2002), Thomson (2004), and Thomson et al. (2005, 2008). In addition to the materials used here, a large series of specimens from major herbaria has been examined in the course of the taxonomic revisions listed. We sampled 77 bracken specimens, most of which were included previously in the taxonomic studies described. Tissue samples were collected from either wild sources or from sporophytes grown in a common garden derived from known wild sources and propagated from rhizome segments or mass spore sowings (Thomson, 2000a). Complete voucher information with geographic sources and GenBank accession numbers is provided in Appendix B.

**DNA extraction, PCR amplification, and sequencing**—

Total genomic DNA was extracted from tissue that had been silica-dried, freeze-dried, or pickled in CTAB/NaCl/ascorbate (Thomson, 2002). The chloroplast markers *trnS*<sup>GGA</sup>–*rpS4* spacer+gene and *rpL16* intron were amplified in 25 µL polymerase chain reactions (PCR) using the fern-specific primers published in Small et al. (2005). Each PCR reaction contained 25–50 ng of template DNA, 1× Promega PCR buffer (Promega, Madison, Wisconsin, USA), 1.5 mM MgCl<sub>2</sub>, 0.20 mM of each dNTP, 0.25 µM each of the forward and reverse primers, and ~12.5 U Taq polymerase. Most PCR reactions were
completed under a standard temperature cycle procedure beginning with a 2 min
denaturation step at 94°C; followed by 30 cycles of 94°C, 48°C and 72°C each for 1 min;
finishing with a 7 min elongation step at 72°C. Problematic samples were amplified using
a slow ramp thermal cycle protocol that began with a 2 min denaturation step at 95°C,
followed by 30 cycles in which the sample was denatured at 95°C for 1 min, dropped to
45°C for 1 min, then slowly raised to 68°C at a rate of 0.2°C/sec and held at 68°C for 4
min. After cycling, reactions were held at 68°C for 10 min to complete elongation of PCR
products. PCR products were directly cycle-sequenced in both directions using the PCR
primers and ABI BigDye Terminator version 3.1 chemistry (Applied Biosystems, Foster
City, California, USA).

**Sequence alignment and phylogenetic analyses**—
Sequences were checked against electropherograms and manually edited, if
necessary, using the program 4Peaks version 1.7 (Griekspoor and Groothuis, 2006).
Sequences were manually aligned using the program Se-Al version 2.0a11 (Rambaut,
2002). The aligned concatenated data matrix is available in TreeBase (http://
www.treebase.org; study number S2235). Maximum parsimony (MP) analyses were
performed on the two-marker concatenated data set in the program PAUP* version
4.0b10 (Swofford, 2002). All nucleotide site characters were unordered and equally
weighted, treating gaps as “missing” data. Heuristic MP searches used tree-bisection-
reconnection (TBR) branch swapping on starting trees generated from 10,000 random
stepwise addition sequence replicates, holding 10 trees at each addition step. All the
most parsimonious trees were saved, and the strict consensus of these was calculated.
MP bootstrap support was assessed from 10,000 bootstrap replicates using heuristic
searches with TBR branch swapping on starting trees generated from 10 random stepwise addition sequence replicates and saving all the most parsimonious trees.

Models of DNA sequence evolution used in Bayesian phylogenetic inference (BI) were selected for each chloroplast marker using the second order Akaike information criterion (AICc) implemented in the program MrModelTest version 2.2 (Nylander, 2004), using likelihood scores estimated in PAUP* for the neighbor-joining tree under alternative models of evolution. The total number of alignment sites in each data partition was used as the sample size for AICc calculations (Posada and Buckley, 2004). Akaike weights and evidence ratios were examined to help guide selection of the best-fit model of molecular evolution for each partition (Burnham and Anderson, 2002).

Bayesian analysis was performed in the parallel version of the program MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003; Altekar et al., 2004). Data were partitioned for the two chloroplast markers, allowing model parameters for each partition to vary independently while linking topology and branch lengths across partitions. Two independent runs, with six chains each, were conducted simultaneously for 10,000,000 generations. Parameter estimates were sampled every 1000 generations. The average standard deviation of split frequencies and the potential scale reduction factor (PSRF) were calculated after discarding the first 25 percent of the samples (2.5 million generations) as burn-in to assess topological and parameter convergence (respectively) between the two runs. BI clade credibility values (i.e., posterior probabilities for clades) and tree posterior probabilities were also calculated after the first 2.5 million generations were discarded as burn-in.

Alignment of two outgroup taxa (*Dennstaedtia davalloides* (R. Br.) T. Moore and *Hypolepis muelleri* N. A. Wakef.) with the *Pteridium* sequences was not possible due to a
high level of divergence (both sequence substitution saturation and ambiguous indels). Inclusion of these taxa in preliminary phylogenetic analyses contributed to the destabilization of clades within *Pteridium*. Because of this, the midpoint rooting method implemented in PAUP* was used to root the *Pteridium* phylogeny in both MP and BI analyses. While this approach is not ideal, it has yielded acceptable results in other taxa when appropriate data are not available (Schuettpelz and Hoot, 2006). The monophyly of *Pteridium* and the midpoint root was confirmed by outgroup phylogenetic analysis of *rbcL* sequences from a subset of our sampled *Pteridium* taxa aligned with additional outgroup taxa from Dennstaedtiaceae (data not shown).

**Biogeographic analysis**—

Broad geographic regions were mapped onto the phylogeny using the parsimony criterion to elucidate major historical biogeographic events. Geographic areas were coded as six unordered character states (Hawaii, North/South America, Europe, Asia/India, Africa, Austral) and traced onto the tree using the program Mesquite version 2.5 (Maddison and Maddison, 2008). Specimens in cultivation were coded from the geographic area of their source.

**RESULTS**

**Sequence characteristics**—

The aligned *trnS–rpS4* spacer+gene and *rpL16* intron data matrices included 1018 and 753 nucleotide sites with 29 and 25 variable sites, respectively. Of those variable sites, 22 were parsimony-informative in *trnS–rpS4* spacer+gene and 21 sites were parsimony-informative in *rpL16* intron. There were two 5-bp repeat sequence indels located 32 nucleotides apart in the *trnS–rpS4* spacer. Individuals either lacked
one or both of the repeat sequences, resulting in three observed indel haplotypes. These haplotypes have been reported previously (Thomson et al., 2008), and our study is consistent with those findings, so we adopt their nomenclature here (haplotypes A, B, and C). This study newly establishes the haplotypes for *Pteridium aquilinum* subsp. *decompositum* (Gaudich.) Lamoureux ex J. A. Thomson, *P. caudatum* (L.) Maxon, and *P. semihastatum* (Wall. ex J. Agardh) S. B. Andrews, and extends coverage of *P. a.* subsp. *pinetorum* (C. N. Page & R. R. Mill) J. A. Thomson to eastern Europe. Haplotype B (the presence of the first repeat sequence, GTTTT) was observed in *P. a.* subsp. *aquilinum* from Europe and both *P. a.* subsp. *centrali-africanum* Hieron. and *P. a.* subsp. *capense* (Thunb.) C. Chr. in Africa, while haplotype C (the presence of the second repeat sequence, AGTCT) was observed in *P. arachnoideum* (Kaulf.) Maxon in Central and South America, *P. esculentum* (G. Forst.) Cockayne in Australia, New Zealand, and New Caledonia, *P. semihastatum* in Australia and Malaysia, and a single accession of *P. caudatum* from Costa Rica. Haplotype A (the absence of both repeat sequences) was observed in the remaining taxa (i.e., *P. aquilinum* from Asia and North America and the remaining *P. caudatum* accessions). Comparison of bracken haplotypes with those found in *Dennstaedtia davalliodes* and *Hypolepis muelleri* reveals that haplotype C is likely to be plesiomorphic. Parsimony character reconstruction of these indels on our phylogeny indicates that the second repeat was lost in the common ancestor of *P. aquilinum*, and the first repeat was gained in the common ancestor of the African and European *P. aquilinum* subspecies. The evolution of these indels is mapped on the Bayesian phylogeny (Figure 2-1).
Figure 2-1. Bayesian phylogeny. Phylogram inferred from the concatenated *trnS-rpS4* spacer+gene and *rPL16* intron alignment. Branch lengths are proportional to the number of nucleotide substitutions/site. Support values from maximum parsimony bootstrap analysis (BS) and Bayesian posterior probabilities for clades (PP) are given above branches (BS/PP). Clades receiving less than 50% BS or 0.5 PP are indicated with a dash (—).

P. caudatum, 238 HECR, Costa Rica
P. caudatum, 274 QCOL, Colombia
P. caudatum, MD2-2VENZ, Venezuela
P. a. ssp. latiusculum, 147 WMCH, USA, Michigan
P. a. ssp. pseudocaudatum, 169 FLUS, USA, Florida
P. a. ssp. pseudocaudatum, 203 HFLA, USA, Florida
P. a. ssp. pseudocaudatum, Der 68, USA, Florida
P. a. ssp. latiusculum, 148 BRMN, USA, Maine
P. a. ssp. pubescens, Der 67, USA, California
P. a. ssp. pubescens, 325 OWUS, USA, Washington
P. a. ssp. pubescens, 100 AOUUS, USA, Oregon
P. a. ssp. latiusculum, 143 YCCM, USA, Massachusetts
P. a. ssp. decompositum, 293 MAHI, USA, Hawaii
P. a. ssp. pinetorum, 164 KUKR, Ukraine
P. a. ssp. pinetorum, NnR, Siberia
P. a. ssp. japonicum, 029 TTWN, Taiwan
P. a. ssp. japonicum, 071 AUP, Japan
P. a. ssp. japonicum, 085 YSCH, Taiwan
P. a. ssp. japonicum, 096 GNCH, China
P. a. ssp. japonicum, 164 MOGJ, Japan
P. a. ssp. japonicum, 280 CHUP, Japan
P. a. ssp. japonicum, 316 SHUP, Japan
P. a. ssp. japonicum, 305 YGIN, India

Loss of second repeat
-> Haplotype A

Gain of first repeat
-> Haplotype B

Ancestral haplotype (presence of second repeat, absence of first repeat)
-> Haplotype C

0.3 substitutions/site

P. arachnoideum, ME2-1VNZA, Venezuela
P. esculentum, 387 CRDQ, Australia, Queensland
P. esculentum, 401 RNYA, Australia, New South Wales
P. caudatum, 323 CCCR, Costa Rica
P. caudatum, 274 QCOL, Colombia
P. caudatum, 278 KMAL, Malaysia
P. caulifer, Wolf 638, New Zealand
P. esculentum, 332 SVNZ, New Zealand
P. esculentum, 324 HNZN, New Zealand
P. esculentum, 213 CRAN, Australia, New South Wales
P. esculentum, 083 WANA, Australia, Western Australia
P. esculentum, 127 KEWA, Australia, Western Australia
P. arachnoideum, 137 SPBR, Brazil, São Paulo
**Phylogenetic analyses**—

Maximum parsimony (MP) analysis of the concatenated data set resulted in 12 most parsimonious trees with a length of 59 (consistency index, CI = 0.9153; retention index, RI = 0.9892). The best fit model used in Bayesian Inference (BI) was GTR+G and HKY+G for *trnS–rpS4* spacer+gene and *rpL16* intron, respectively. The AICc weight of the best fit model was 1.4 times greater than the next-best model for *trnS–rpS4* spacer +gene and 1.3 times greater for *rpL16* intron. The topology of the MP strict consensus tree (Figure 2-2A) is congruent with the BI phylogeny (Figure 2-1). MP bootstrap support (BS) and BI posterior probabilities of clades (PP, i.e., clade credibility values) are reported on the tree for supported nodes (Figure 2-1). Two fully supported clades are resolved at the base of the phylogeny, separating *P. arachnoideum*, *P. esculentum*, and *P. semihastatum* from *P. aquilinum*. There is a basal polytomy within the *P. aquilinum* clade. A single accession of the tetraploid species *P. caudatum* is grouped with *P. semihastatum* in the former clade, while the remaining *P. caudatum* accessions are grouped with subspecies *pseudocaudatum* (Clute) Hultén in the *P. aquilinum* clade.

**Biogeographic analysis**—

Mapping geographic areas onto the phylogeny revealed a number of biogeographic patterns (Figure 2-2A). The Austral taxon *P. esculentum* is found in a clade with *P. arachnoideum* from South and Central America and *P. semihastatum* from Southeast Asia and Australia, forming a group with a post-Gondwanan (after Africa split away) southern continental distribution referred to here as the Austral/South American clade. Within the *P. aquilinum* clade, basal biogeographic patterns are unresolved, but the European *P. aquilinum* subsp. *aquilinum* accessions emerge from within the African brackens and *P. aquilinum* subsp. *decompositum* in Hawaii is in a clade with all of the
Figure 2.2. Biogeography of bracken. (A) Broad geographic distribution of bracken specimens traced onto the maximum parsimony phylogeny using parsimony-based character reconstruction. Geographic areas were coded as six unordered character states: Hawaii, North/South America, Europe, Asia/India, Africa, Austral. (B) Global map of sampled bracken specimens, color-coded consistently with the geographic area character states in (A).
sampled accessions of *P. aquilinum* subsp. *pubescens* (Underw.) J. A. Thomson, Mickel & Mehltreter from western North America. The locations of specimens included in this study are indicated on a global map (Figure 2-2B).

**DISCUSSION**

*Phylogenetic relationships among bracken taxa*—

*Pteridium* species form two distinct and fully supported basal clades (100% maximum parsimony bootstrap—BS, 1.0 Bayesian posterior probability—PP). The first clade contains *P. esculentum* from Australia, New Caledonia, and New Zealand and *P. arachnoideum* from the neotropics, but neither of these species form distinct monophyletic groups. The tetraploid species *P. semihastatum* and a single accession of the tetraploid species *P. caudatum* from Costa Rica are included in this first main clade and together are monophyletic (85% BS, 1.0 PP). This tetraploid group is derived from a clade containing an accession of *P. esculentum* from New Caledonia (55% BS, 0.99 PP). The second main clade in *Pteridium* includes all of the *P. aquilinum* accessions we sampled and three of the four *P. caudatum* accessions we included in our analysis (Costa Rica and northern South America). This second main *Pteridium* clade (the *P. aquilinum* clade) is fully supported as monophyletic (100% BS, 1.0 PP), but phylogenetic relationships of lineages within this group are not resolved. The three *P. caudatum* accessions in the *P. aquilinum* clade are supported as monophyletic with 63% BS and 0.99 PP and are included in a polytomy with all of the *P. aquilinum* subsp. *pseudocaudatum* accessions we sampled (Florida, USA) and a *P. a.* subsp. *latiusculum* (Desv.) Hultén accession from Michigan, USA (64% BS, 0.98 PP). *Pteridium a.* subsp. *latiusculum* accessions are scattered throughout the *P. aquilinum* clade and therefore is
not monophyletic in our analyses. This phylogenetic pattern in *P. a.* subsp. *latiusculum* may be explained if this taxon is primarily defined by plesiomorphies (morphologically) and/or if it is susceptible to widespread hybridization (Speer et al., 1998). Together, *P. a.* subsp. *japonicum* (Nakai) Á. Löve & D. Löve from coastal Asia and *P. a.* subsp. *pinetorum* from the boreal Asian interior form a clade supported by 63% BS and 0.98 PP, but neither subspecies forms a monophyletic group. *Pteridium a.* subsp. *decompositum* (Gaudich.) Lamoureux ex J. A. Thomson from Hawaii and *P. a.* subsp. *pubescens* from western North America and a single *P. a.* subsp. *latiusculum* accession from northeastern North America form a polytomy supported by 86% BS and 1.0 PP. This pattern suggests a trade wind dispersal route from North America to Hawaii consistent with other wind-dispersed fern taxa (Geiger et al., 2007). *Pteridium a.* subsp. *wightianum* (J. Agardh) W. C. Shieh from the Himalayas of northern India (isolate YGIN) emerges from the aquilinum clade polytomy, while the remaining *P. a.* subsp. *wightianum* accessions from Sri Lanka and Southeast Asia form a monophyletic group supported by 87% BS and 1.0 PP. Isolate YGIN was also distinguished from other *P. a.* subsp. *wightianum* by both a distinctive nuclear genome marker and morphology in an earlier study (Thomson, 2000a). All the *P. a.* subsp. *aquilinum* accessions we sampled (Europe) are monophyletic (62% BS, 1.0 PP), but emerge from a paraphyletic grade of African *P. aquilinum* subsp. *capense*. Also emerging from *P. a.* subsp. *capense* is *P. a.* subsp. *centrali-africanum*, which is supported as monophyletic with 87% BS and 1.0 PP.

We emphasize here that apparent paraphyly at the level of species or below in our data set should not be used to recircumscribe taxa for two reasons. First, we have examined only variation in chloroplast DNA. Because of lineage sorting and recombination, we have no reason to expect exactly the same patterns for nuclear
genes (Soltis et al., 1992). Second, some models of speciation predict a high frequency of paraphyly among recently diverged species (Rieseberg and Brouillet, 1994; Funk and Omland, 2003).

**Hybridization and the origin of the tetraploid species**—

Although the majority of *Pteridium* taxa are at the same ploidy level (2N = 104; Page, 1976), recent evidence suggests that *P. semihastatum* and *P. caudatum* are tetraploid taxa (4N = 208; Tan and Thomson, 1990b; Thomson, 2000a, b; Thomson and Alonso-Amelot, 2002). Allopolyploidy appears to be a common speciation process in plants, and it often involves a triploid intermediate (Ramsey and Schemske, 1998). Two main types of evidence for allopolyploid speciation can be gathered. Nuclear markers, preferably fixed for different alleles in the different parents, can show additive effects in allotetraploids, and simultaneously reveal both parents, whereas uniparentally inherited markers will reveal one parent and whether there have been multiple origins (Soltis and Soltis, 1993). Although we have no evidence for the mechanism of inheritance of chloroplast DNA in *Pteridium*, several studies have demonstrated maternal inheritance in other species of ferns (Gastony and Yatskievych, 1992; Vogel et al., 1998).

*Pteridium caudatum* and *P. semihastatum* have been shown to be tetraploid bracken species based on DNA c-values, spore size, guard cell length, and morphology of the cells in the false indusium (Thomson, 2000a, b; Thomson and Alonso-Amelot, 2002). Hypothesized allopolyploid origins have been inferred from intermediate morphology and additive AP-PCR DNA genotypes of these two species (Thomson, 2000a; Thomson and Alonso-Amelot, 2002). These studies have suggested putative progenitors of *P. caudatum* to be *P. aquilinum* subsp. *pubescens* in the north and *P. arachnoideum* in the south (Thomson and Alonso-Amelot, 2002). These data have
similarly been used to infer the putative progenitors of *P. semihastatum* in Southeast Asia and northern Australasia as *P. aquilinum* subsp. *wightianum* (=*P. revolutum* sensu Brownsey, 1989) and *P. esculentum* (Thomson, 2000a; Thomson and Alonso-Amelot, 2002). Our data based on the phylogeny of chloroplast sequences support the maternal parentage of *P. semihastatum* to be *P. esculentum*, while *P. caudatum* emerges from two parts of the phylogeny (with North American *P. aquilinum* subsp. *semicaudatum/latusculum* and embedded in the Austral/South American clade, which includes *P. arachnoideum, P. esculentum*, and *P. semihastatum*). Two possible explanations for the phylogenetic placement of our *P. caudatum* accessions seem plausible: (1) reciprocal and multiple hybrid origins of this species or (2) a single hybrid origin of *P. caudatum* with the maternal parent as *P. a. subsp. pseudocaudatum/latusculum* with subsequent introgression of the tetraploid *P. semihastatum* chloroplast genomes. Additionally, the Austral/South American parent of *P. caudatum* may be an unsampled haplotype of the co-occurring *P. arachnoideum* from South America or long-distance hybridization from an Austral member of that clade. Long-distance hybridization in bracken has been documented between eastern North America and Scotland (Rumsey et al., 1991). To evaluate these hypotheses, we need nuclear genetic data to determine parentage and additivity of genomic elements in the allotetraploid species. Nuclear sequence data promise to be especially useful in elucidating specific phylogenetic relationships among the various bracken taxa (Schuettpelz et al., 2008).

**Phylogeography of bracken**—

*Pteridium* occurs throughout the world, except in hot and cold desert regions. Range-wide sampling is necessary to adequately assess variation within the genus and delineate taxa. Only one such study has been accomplished in *Pteridium*: that of Tryon
(1941). While the varieties in Tryon’s treatment are often geographically based, he only briefly discussed the geographic ranges of each taxon and instead focused on the morphological differences between the varieties. Page (1976) expanded and updated the ecological and phytogeographic information known for each of Tryon’s varieties, but emphasized that the adoption of Tryon’s nomenclature was done for practical reasons rather than validation of these taxa, and that an updated worldwide revision of the genus was needed. While our sampling strategy attempts to cover the range of variation in *Pteridium* worldwide, our study would benefit from additional sampling from Madagascar, South and Central America, the Caribbean, and parts of North America.

Recent phylogenetic analyses of ferns, which have included extensive taxonomic sampling and have used fossils for molecular date calibrations, have estimated divergence dates between *Pteridium* and *Dennstaedtia* to be about 114 mya (Schneider et al., 2004) and 92 mya (Pryer et al., 2004). These estimates represent upper ages for the evolution of *Pteridium* and correspond to the Cretaceous period. Fragments of fossils attributable to *Pteridium* are known from the Tertiary period as far back as the Oligocene (33.7–23.8 mya) from Europe to Australia, suggesting that even by this time, bracken may have achieved a widespread distribution (Page, 1976). During the time when bracken likely originated and the earliest divergences may have occurred (from 100 to 30 mya), Africa and India separated from a southern landmass that included Australia, Antarctica, and South America (Scotese, 2001). This separation may correspond to the basal divergence among brackens, splitting the Austral/South American clade from the *P. aquilinum* clade of African and Laurasian affinity. The polytomy in the *P. aquilinum* clade may represent a rapid radiation of African and Indian bracken as they moved north, colliding with a warm temperate and paratropical Laurasia by the end of the Paleocene...
(about 50 mya). Alternatively, the polytomy in the *P. aquilinum* clade may result from insufficient data to distinguish phylogenetic relationships among lineages in the clade.

As continents have come to their modern global locations, and with the onset of the current ice age, the Quaternary glaciation beginning 2.5 mya, there have been periods of extensive ice sheets and cold, dry steppe coming into Europe and North America and likely driving bracken into southern refugia. These repeated glacial periods may be invoked to explain the phylogeographic pattern whereby the European brackens are derived from within the African taxa if Africa served as a source for colonization of Europe during recent geological times. The strong geographic structure observed at large (continental) scales among bracken taxa is consistent with patterns expected under a vicariance model, with range expansion across land bridges and only occasional long-distance dispersal resulting in hybrid forms. This evidence contradicts the common assumption in ferns that repeated long-distance dispersal of spores will erode biogeographic patterns resulting from vicariance in widespread species via constant gene flow (Tryon, 1985; Wolf et al., 2001). Recent empirical evidence in *Ceratopteris* Brongn. shows that genetic reproductive barriers can accumulate quickly in allopatry, thereby restricting gene flow and potentially maintaining the signature of vicariance (Nakazato et al., 2007).

**Conclusions**—

This study highlights the importance of range-wide perspectives when undertaking evolutionary and monographic studies. Any taxonomic revision of *Pteridium* must reflect a global perspective and should include information from diverse areas of research including morphology, cytology, genetics, and reproductive biology. To evaluate the biogeographic and hybrid origin hypotheses presented here, we will need a better-
resolved species phylogeny with increased population-level sampling that incorporates nuclear data and fossils for date calibration. These data might also allow us to examine additional reticulation and lineage sorting in the evolutionary history of Pteridium as well as evaluate contemporary and historical gene flow.

LITERATURE CITED


N YLANDER, J. A. A. 2004. MrModeltest 2.2. Program distributed by the author, Evolutionary Biology Centre, Uppsala University, website: http://www.abc.se/~nylander/ [accessed 17 April 2007].


CHAPTER 3

COMPLETE CHLOROPLAST GENOME SEQUENCE AND HIGH LEVELS OF RNA EDITING IN THE FERN *PTERIDUM AQUILINUM* DETERMINED BY MASSIVELY PARALLEL PYROSEQUENCING¹

ABSTRACT

RNA editing is the post-transcriptional modification of RNA molecules relative to their encoding genomic DNA sequences. RNA editing in land plant organelles occurs in the form of pyrimidine exchanges. In general, levels of RNA editing are higher in mitochondrial genomes than in chloroplast genomes, and in seed-free vascular plants and hornworts relative to other lineages of land plants (i.e. liverworts, mosses, and seed plants). To date, genome-wide chloroplast RNA editing has been systematically examined in only seed-free plants: a fern (*Adiantum capillus-veneris*) and a hornwort (*Anthoceros formosae*). The extremely high levels of RNA editing observed in chloroplast transcripts of ferns and hornworts relative to seed plants (up to 25 times higher) raises a number of interesting questions on the evolution of RNA editing in plastomes. We present a genome-wide analysis of chloroplast RNA editing in a second fern species. We use a novel, rapid method to determine the complete chloroplast genome sequence and identify RNA editing sites using second-generation high-throughput shotgun DNA sequencing technologies. This study also represents the first application of RNA-seq to examine RNA editing in a chloroplast transcriptome. The complete circular mapping chloroplast genome sequence of *Pteridium aquilinum* is 152,362 bp long and contains a gene set and gene order identical to *Adiantum capillus-

¹ Coauthors: Aaron M. Duffy, Matt Kusner, Chen Gu, Paul Overvoorde, and Paul G. Wolf.
veneris. There are 117 different genes, including 84 protein coding genes, 4 ribosomal RNA genes, and 29 transfer RNA genes. We experimentally detected 551 unique C to U RNA editing sites and 300 U to C RNA editing sites in the chloroplast genome, 2.5 times that detected in Adiantum and approaching the level of RNA editing observed in Anthoceros. RNA editing events have been observed in protein coding genes, tRNA, rRNA, introns, and intergenic regions. This work will enable a reexamination of the evolution of chloroplast RNA editing by surveying RNA editing in the complete plastid transcriptome of a second fern. Our approach should also be widely applicable to studying RNA editing in the chloroplasts of other plant lineages.

INTRODUCTION

Chloroplasts play a critical role in the function of plant cells as the center for photosynthesis and starch metabolism. Derived from a cyanobacterial-like ancestor and incorporated into a eukaryotic genetic environment, the chloroplast genome has evolved a complex system of RNA metabolism that combines features of the ancestral endosymbiont with those of its eukaryotic host (1). Such a complex system of RNA processing may function to regulate plastid gene expression (2, 3), to increase molecular diversity (4, 5), and to functionally preserve the photosynthetic machinery (6). One of the most striking components of chloroplast RNA metabolism is the occurrence of RNA editing to alter the nucleotide sequence of mature RNA molecules relative to their encoding genomic DNA, thus providing an exception to the “central dogma” of molecular biology (7, 8).

RNA editing has evolved independently in at least seven lineages of eukaryotes, including animals, fungi, plasmodial slime molds, and land plants, among others (9). The
evolution of RNA editing in plants apparently occurred during the transition to land and is observed within both organellar genomes (10-12) and the nuclear genome (13), although evidence suggests that it has been secondarily lost in marchantiid liverworts (14-17). Rates of RNA editing are typically higher in mitochondria than in chloroplasts (14, 18), with the extreme case observed in Isoetes, where 1,420 RNA editing sites have been identified in the mitochondrial genome (19). The chloroplast genomes of hornworts and ferns also experience high levels of RNA editing, on the order of several hundred sites (20, 21), whereas only 30-40 RNA editing sites typically occur in the chloroplast genomes of seed plants (22). The distribution of RNA editing sites (both phylogenetic and specific sites) is highly dynamic (23), suggesting that RNA editing sites evolve rapidly. The exceptionally low rates of RNA editing in seed plant plastomes relative to bryophytes and seed-free vascular plants is generally attributed to the steady independent loss of ancestral RNA editing sites in different lineages (24, 25).

Given the importance of chloroplasts in plant cell function, the global identification of RNA editing sites in chloroplast transcripts is required to understand the precise coding potential of the chloroplast genome and will substantially enhance our understanding of gene expression and function in this important organelle. Because RNA editing has been systematically examined in the chloroplast genome of Adiantum capillus-veneris (21), by examining RNA editing in a second fern species, we can examine evolutionary patterns of RNA editing in ferns relative to seed plants, where a dramatic reduction in the level of RNA editing has occurred. These data can be examined in light of several hypotheses regarding the evolution of RNA editing in plant organellar genomes (6, 25-27).
The classic approach to surveying RNA editing in complete organellar genomes involves laborious PCR amplification and sequencing of cDNAs for each gene. This requires *a priori* sequence information about each predicted transcript to develop primers and is dependent on an accurate annotation of the genome. Additionally, without cloning and sequencing multiple amplicons for each target transcript, it is impossible to determine the extent of incomplete RNA editing.

Recent developments in high throughput sequencing technologies have revolutionized the collection and analysis of genome-scale sequence data. Two independent research groups have reported alternative approaches using second-generation sequencing technologies for the rapid determination of complete chloroplast genome sequences (28, 29). Additionally, several recent studies have used deep sequencing of cDNA to study RNA editing (13, 30-33), but only one of them systematically identified novel RNA editing sites in a plant organellar genome (32). We utilized Roche 454 pyrosequencing to determine the complete chloroplast genome sequence of *Pteridium aquilinum* and identify novel RNA editing sites. Our approach differs from previous studies using second-generation sequencing to determine chloroplast genome sequences in that we did not first isolate chloroplast DNA from nuclear and mitochondrial DNA. Alternatively, we used bioinformatic methods to extract the chloroplast DNA sequence *in silico*. This approach allowed us to avoid the difficult and laborious sample preparation protocols needed to isolate chloroplast DNA or to PCR amplify the complete genome in overlapping fragments. At the same time, we obtained valuable sequence information from the nuclear and mitochondrial genomes in this non-model organism. By also sequencing cDNA, we were able to obtain expressed sequence information for the chloroplast transcriptome without relying on any prior
information about the chloroplast genome. These data, used in conjunction with the plastome sequence, enabled not only an analysis of RNA editing but also provided experimental evidence for intron splice sites and exon boundaries that were used in annotation of the genome sequence.

MATERIALS AND METHODS

Chloroplast genome sequencing, assembly, and annotation

Total genomic DNA was extracted from fronds of Pteridium aquilinum ssp. aquilinum (source: Sheffield 48, Manchester, UK) using a bulk CTAB protocol. Genomic DNA was further purified by centrifugation on a CsCl gradient, where two nucleic acid fractions were isolated separately (34). Purified DNA from each fraction was sequenced separately and pooled on the Roche 454 GS-FLX platform using standard and Titanium chemistries. In total, 711,178 reads were obtained that represented 216.19 Mbp of genomic sequence data. Whole genome shotgun DNA sequence reads collected in the course of this study have been deposited in the NCBI Sequence Read Archive (submission no. SRA020058). The complete chloroplast genome sequence was extracted and assembled using a combination of de novo and reference-guided bioinformatic approaches. Reads were assembled de novo using MIRA (35). Putative chloroplast contigs were identified with NCBI blastn by querying the complete chloroplast genome sequences of Adiantum capillus-veneris (Genbank accession: AY178864), Alsophila spinulosa (Genbank accession: FJ556581), and Angiopteris evecta (Genbank accession: DQ821119) against the assembly, using an e-value threshold of 1e-4. Contigs returning significant BLAST hits were parsed out of the complete de novo assembly using a custom Perl script. Additionally, several reference-guided assemblies were
generated with the YASRA pipeline (36) using the chloroplast genomes of *Adiantum capillus-veneris*, *Alsophila spinulosa*, and *Angiopteris evecta* as reference sequences at various levels of sequence similarity stringency. Putative chloroplast contigs from both *de novo* and reference-guided approaches were collected and assembled to generate a version 1 draft *Pteridium* chloroplast genome sequence. The original DNA sequence reads were then mapped to the draft genome sequence in Consed v.19 (37) and DOGMA (38) was used to identify potentially erroneous frame shifts in protein coding sequences. A corrected version 2 draft sequence (with and without the second inverted repeat) was then used as the scaffold for a new reference-guided assembly using the Roche GS Mapper v.2.3 software. The resulting high quality, read-supported contigs were used to produce the final complete chloroplast genome sequence in the standard circular orientation typically presented for chloroplast genomes, beginning with the large single copy region (LSC) followed by the first inverted repeat (IRB), the small single copy region (SSC), and the second inverted repeat (IRA). The Mauve v.2.2.0 plugin (39) for Geneious Pro v.5.0.1 (40) was used to generate a whole genome alignment for the complete *Adiantum*, *Alsophila*, and *Pteridium* chloroplast genomes. This alignment with annotations for *Adiantum* and *Alsophila* was used to annotate the *Pteridium* chloroplast genome sequence. A circular gene map for the annotated sequence was generated in OGDRAW (41).

**Transcriptome sequencing and analysis**

Total RNA was isolated from *Pteridium aquilinum* ssp. *aquilinum* gametophytes (source: Wolf 84, Norwich, UK) grown in culture on an agarose mineral nutrient media using the Spectrum Plant Total RNA kit (Sigma), incorporating an on-column DNase treatment (Qiagen) during extraction. Total RNA was sent to the Center for Genomics
and Bioinformatics at Indiana University (IU CGB) where a normalized transcriptome (cDNA) library optimized for Roche 454 GS-FLX Titanium sequencing was prepared (K Mockaitis, unpublished, IU CGB). This library was sequenced using the standard GS-FLX Titanium protocol on 3 regions of a 4-region PicoTitrePlate. Transcriptome sequence data are available in the NCBI Sequence Read Archive (SRA012887). Transcriptome reads were mapped to the chloroplast genome using the Roche GS Mapper v2.3 software specifying cDNA reads and a genomic reference sequence. RNA editing sites were identified by parsing the results file reporting differences between reads and the reference sequence. Additional predicted RNA editing sites were added to protein coding genes during annotation to correct start and stop codons necessary for proper translation of the gene. Intron splice sites were verified by mapping transcriptome reads to the chloroplast genome using GMAP to report intron splice donor and acceptor positions (42, 43).

RESULTS

Complete chloroplast genome

The complete circular-mapping chloroplast genome sequence of *Pteridium aquilinum* is 152,362 bp (Figure 3-1) and arranged in the typical quadripartite plant plastid genome structure consisting of a large single copy region (LSC; 84,335 bp) and a small single copy region (SSC; 21,259 bp), separated by two copies of a large inverted repeat sequence (IRA and IRB; 23,384 bp). The complete annotated nucleotide sequence has been deposited in Genbank (accession no. HM535629). In mapping whole genome shotgun sequence reads to the final chloroplast sequence, there was an average of 34.5x read support for each position (Figure 3-2), with the coverage at
Figure 3-1: Gene map of the *Pteridium aquilinum* chloroplast genome. Genes on the inside of the circle are transcribed clockwise and those on the outside are transcribed counterclockwise. Genes with introns are indicated with an asterisk. Genes are color coded according to protein complexes or functional categories indicated in the legend. Note that *rps12* is trans-spliced with two copies of the 3' end and that exon 2 of *ndhB* is located in the inverted repeat resulting in an orphan copy of this exon.
individual sites ranging from 4 to 88 reads. When the IR is considered only once, a total of 117 different genes have been detected and annotated, including 84 protein coding genes, 4 ribosomal RNAs, and 29 transfer RNAs (Figure 3-1). The full gene complement and gene order observed in the *Pteridium* chloroplast genome is identical to *Adiantum capillus-veneris*, but differs from *Alsophila spinulosa* in the absence of *psaM*. There are 63,270 G or C nucleotides in the sequence, representing 41.5% nucleotide GC content across the whole genome. Nucleotide GC content is positively correlated with sequence coverage and reaches a local maxima in the ribosomal RNA genes located in the IR (Figure 3-2). There are 17 genes with introns and one gene is trans-spliced (*rps12*). An orphan copy of *ndhB* exon 2 is located in the inverted repeat. GMAP identified 51 intron splice-junction site pairs, which were supported by 83 transcript reads.

**Analysis of RNA editing**

Of 725,312 transcriptome reads totaling 260 Mbp of sequence data, 20,291 reads (2.8%) mapped uniquely to the chloroplast genome, covering 80.67% of the sequence. We experimentally detected 851 unique RNA editing sites (counting the inverted repeat only once), including 551 C to U sites and 300 U to C sites. There are 233 editing sites located in the inverted repeat region resulting in a total of 1,114 sites in the 152,362 bp genome (0.73%). The density of these RNA editing sites along the length of the chloroplast genome is shown in Figure 3-3. There are 660 editing sites in protein coding sequence. Among these, RNA editing modifies the start codon of 26 proteins, removes 37 premature stop codons, and generates 6 stop codons. We detected 15 editing sites in tRNA genes and 168 editing sites in rRNA genes: 111 in rrn23, 52 in rrn16, 5 in rrn4.5, 0 in rrn5. There are 12 RNA editing sites in the introns of 6 genes and 80 editing sites in intergenic regions.
Figure 3-2: Chloroplast genome sequence coverage. The per-base read depth coverage of Roche 454 DNA sequence reads mapped to the chloroplast genome is indicated by the filled-line graph in red. Mean x-fold coverage across the complete genome is 34.54x. To summarize rapid changes in coverage for adjacent regions of the chloroplast genome, the average local coverage for each site was calculated using a 5 kb sliding window (black line, left axis). Average percent GC content was also calculated on a 5 kb sliding window and overlaid on the figure (blue line, right axis). Overall, 41.5% of the genome is comprised of G or C nucleotides. The location of the large single copy region (LSC), small single copy region (SSC), and inverted repeat regions (IRB and IRA) have been mapped across the length of the sequence.
DISCUSSION

The nucleotide sequence and gene map for the *Pteridium* chloroplast genome presented in this study is consistent with the coarse physical restriction and gene map produced by Tan (44) and Stein et al. (45). By sequencing the chloroplast genome of a basal polypod (*Pteridium*), we are able to resolve the evolution of several structural differences between the *Adiantum* and *Alsophila* chloroplast genomes. For example, the inverted repeat region in both *Adiantum* and *Pteridium* has expanded into the SSC to include functional fragments of the 5’ end of *chlL* and the 3’ end of *ndhF* in a short overlapping sequence that is 24 to 34 bp long. Additionally, the 565 bp repetitive region in *Alsophila* reported by Gao (46) is also absent from *Pteridium*. Other features are unique to only one of the polypod species, such as a 2 kb long segment between *ndhC* and *trnV-UAC* that is only present in *Pteridium*. The clade derived from the common ancestor of *Pteridium* and *Adiantum* includes over 9,000 species and represents approximately 80% of fern species diversity (47, 48). This fact makes the chloroplast genome sequence of *Pteridium* a valuable resource for developing chloroplast DNA sequence markers and studying chloroplast evolution in the vast majority of ferns.

A previous study of RNA editing in the transcript of *accD* in *Pteridium* identified 14 RNA editing sites, including the creation of a functional start and stop codon, the repair of a stop codon, and two U to C edits (49). Our data confirm all but two of the C to U edits they report. Based on the nucleotide mismatches in our chloroplast transcript assembly, we were able to predict the nature (C to U or U to C) and orientation (forward or reverse strand) of each putative RNA editing site. However, in a small number of cases, our strand predictions did not match the orientation of predicted transcripts based on the annotation. This could result from four possible scenarios: first, a pyrimidine
Figure 3-3: Density of RNA editing sites. The density probability function of RNA editing sites across the length of the chloroplast genome sequence using a 2.5 kb smoothing bandwidth for the kernel density estimator. The area under the curve sums to one. Individual RNA editing sites are shown as a rug along the bottom of the plot and the location of the LSC, SSC, and IR regions is indicated by short vertical bars.
exchange in an antisense RNA translated from the opposite DNA strand; second, we observe a low frequency of non-standard RNA editing events in the chloroplast transcriptome (e.g. G to A); third, transcript paralogs from the nuclear or mitochondrial genomes are incorrectly mapping to the chloroplast genome; and fourth, we may be detecting SNP mismatches between the strain used to generate the chloroplast genome and that used for transcriptome sequencing (both individuals are sourced from Great Britain, approximately 200 km apart).

This study is the first to use a bioinformatic approach to isolate the chloroplast genome from a sample of whole genome shotgun DNA sequencing on a second-generation high-throughput sequencing platform. We are also the first to use second-generation sequencing to survey RNA editing across the complete plastid genome of any species. These approaches should be broadly applicable to rapid plastid genome sequencing and analyses of RNA editing in other species. In this study, we report the second complete chloroplast genome sequence for a polypodiaceous fern and identified the highest level of RNA editing in a vascular plant chloroplast genome. We also report an exceptional level of RNA editing in ribosomal RNA genes and intergenic regions, previously unobserved in any plant chloroplast genome.

REFERENCES


CHAPTER 4

DE NOVO CHARACTERIZATION OF THE GAMETOPHYTE TRANSCRIPTOME IN BRACKEN FERN, PTERIDIUM AQUILINUM

ABSTRACT

Background

Because of their phylogenetic position and unique characteristics of their biology and life cycle, ferns represent an important lineage for studying the evolution of land plants. Large and complex genomes in ferns combined with the absence of economically important species have been a barrier to the development of genomic resources. However, high throughput sequencing technologies are now being widely applied to non-model species. We leveraged the Roche 454 GS-FLX Titanium pyrosequencing platform to sequence the gametophyte transcriptome of bracken fern (Pteridium aquilinum) to develop genomic resources for evolutionary studies.

Results

Quality and adapter trimmed reads totaling 254 Mbp were assembled de novo into 38,889 unigenes with a mean length of 685.8 bp and a total assembly size of 26.7 Mbp with an average read-depth coverage of 8.5x. Estimates of transcriptome coverage suggest that these sequences cover 87% of the complete transcriptome and have tagged all of the transcripts. Of the 38,889 unigenes in this data set, 53.6% had blastx hits in the NCBI non-redundant protein database (nr) and the longest open reading frame in 30.7% of the unigenes had positive domain matches in InterProScan searches. We were able to assign 41.4% of the unigenes with a GO functional annotation and

---

14.2% with an enzyme code annotation. Enzyme codes were used to retrieve and color KEGG pathway maps. A comparative genomics approach revealed a substantial proportion of genes expressed in bracken gametophytes to be shared across the genomes of *Arabidopsis*, *Selaginella* and *Physcomitrella*, as well as identifying a substantial number of potentially novel genes. We identified 2,320 perfect SSR loci, providing an extensive resource that can be explored and developed for population genomics and genetic mapping studies.

**Conclusions**

This study is the first comprehensive transcriptome analysis for a fern and represents an important scientific resource for comparative evolutionary and functional genomics studies in land plants. We demonstrate the utility of high-throughput sequencing of a normalized cDNA library for *de novo* transcriptome characterization and gene discovery in a non-model organism.

**BACKGROUND**

As the sister lineage to seed plants, ferns (i.e. monilophytes) represent a critical clade for comparative evolutionary studies in land plants [1, 2]. In contrast to seed plants, ferns typically retain the ancestral condition for a suite of life history traits (e.g. the lack of secondary growth, homospory, motile sperm, and independent free-living gametophyte and sporophyte generations). Ferns are thus an important outgroup for studying the evolution of wood, seeds, pollen, flowers, and fruit among other economically important characteristics of seed plants, as well as the evolution of development in these complex structures and the expansion of gene families associated with seed plant evolution (e.g. transcription associated proteins). For reasons not yet
fully understood, ferns typically have much higher chromosome numbers and larger
genomes than seed plants [1, 3, 4]. Understanding the factors that influence these
differences and their evolutionary consequences will require developing genomic
resources in ferns to provide comparative context to understand the evolution of these
traits [3-5]. Additionally, because ferns have evolved and maintained free-living and
photosynthetic gametophyte and sporophyte life stages, they are an ideal group for
studies of both life cycle evolution in land plants and genome function in separate
haploid and diploid phases.

While taxonomic sampling of plants in genome-scale projects has expanded
dramatically with a decrease in the cost of DNA sequencing, the development of
genomic resources in ferns has lagged far behind those of other plant groups. This
deficit has primarily been attributed to technical and economic barriers associated with
the complex and large genomes in ferns, and is compounded by the limited agronomic
value of most ferns [4]. To date, genomic information in ferns is limited to a genetic
linkage map [6] and a modest expressed sequence tag (EST) data set comprised of
about 5,000 Sanger sequences [7] for Ceratopteris richardii, a data set for Adiantum
capillus-veneris which includes just over 30,500 ESTs [8, 9], and fewer than 500 ESTs
for Pteridium aquilinum [9].

With the introduction of cost efficient and massively-parallel high-throughput
sequencing technologies, genome scale studies in non-model organisms are being
actively pursued for gene discovery, expression profiling, SNP and SSR marker
development, and studies in functional, comparative, and evolutionary genomics in taxa
where little or no previous genomic information exists [10-24]. We chose the Roche 454
Life Sciences GS-FLX Titanium technology to sequence a normalized cDNA library for
the gametophyte generation of the bracken fern, *Pteridium aquilinum* (L.) Kuhn.

*Pteridium* (family: Dennstaedtiaceae) is a cosmopolitan fern genus comprised of
several closely related species which are well differentiated from other genera in the
family. *Pteridium aquilinum* is the most widespread of the bracken species and is
distributed throughout the northern hemisphere and Africa [25]. Bracken is notorious as
a weed in open fields and is toxic to people and livestock. Despite its toxicity, bracken is
eaten as a delicacy in several parts of the world, and due to its often high local
abundance and large coarse stature, is sometimes used as thatching or packing
material. Because bracken is common, easily cultured and manipulated, and can have a
major economic impact, it has become one of the most intensively studied fern species.
Bracken has been used as a model system for the study of the fern life cycle [26-34],
gametophyte development and the pheromonal mechanism of sex determination
[35-42], cyanogenesis [43], carcinogenesis [44-46], invasion ecology [47-49], and
climate change [50].

This study was conceived to develop an extensive expressed sequence resource
in ferns for evolutionary and functional genomics, with the long term objective of
examining the role that alternation of generations plays in genome evolution. We present
the first comprehensive transcriptome characterization for a fern gametophyte, including
an assessment of transcriptome coverage, gene family and functional classification,
SSR identification, and a comparative analysis of gene sets across land plants.
Sequencing and *de novo* assembly

Raw Roche 454 GS-FLX Titanium reads were quality- and adapter-trimmed and size-selected to yield 681,722 cleaned reads with a mean length of 372.6 bp and 254 Mbp of total sequence data (Table 4-1, Figure 4-1A). Reads were first assembled in MIRA v2.9.46 [51] and the resulting assembly was passed through a second assembly step in CAP3 [52] to join additional contigs (Table 4-2). The resulting final assembly consisted of 38,889 unigenes (i.e. unique gene sequences; assembled contigs + singletons) with a mean length of 685.8 bp, which summed to a total assembly length of 26.7 Mbp (Table 4-1, Figure 4-1B). The average read-depth coverage for the final unigene assembly was 8.50x (Table 4-2). The distribution of unigene coverage was highly left-skewed toward low coverage with an extremely long tail (maximum coverage just under 1800x; Figure 1C,D). The steep decline in read-depth coverage suggests that cDNA normalization was effective and is typical for a normalized library [15].

<table>
<thead>
<tr>
<th>Table 4-1 - Transcriptome sequence statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw reads</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Number of sequences</td>
</tr>
<tr>
<td>Mean length (bp)</td>
</tr>
<tr>
<td>Standard deviation (bp)</td>
</tr>
<tr>
<td>Mode length (bp)</td>
</tr>
<tr>
<td>Median length (bp)</td>
</tr>
<tr>
<td>Range in length (bp)</td>
</tr>
<tr>
<td>Total length (Mbp)</td>
</tr>
</tbody>
</table>

Summary statistics for sequence data at different stages of processing.
Figure 4-1 - Overview of *P. aquilinum* transcriptome sequencing and assembly.

(A) A histogram of the filter passed and adapter/quality trimmed Roche 454 GS-FLX Titanium read lengths. (B) A histogram of unigene lengths for the final unigene set after the 2-step assembly. Note that the longest unigene is 4,489 bp and the x-axis has been truncated at 3 kb. (C) A histogram of the average read-depth coverage for unigenes. The steep decline in coverage observed here is typical of normalized libraries [15]. Coverage values between 30x and 1800x have been binned (see the vertical axis in Figure 1D). (D) A density scatterplot showing the relationship between unigene length and coverage. Points with a higher local density are darker.
Table 4-2 - Assembly summary statistics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Primary assembly (MIRA)</th>
<th>Secondary assembly (CAP3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reads assembled</td>
<td>622074</td>
<td>622074</td>
</tr>
<tr>
<td>Number of reads discarded</td>
<td>59648</td>
<td>0</td>
</tr>
<tr>
<td>Number of singletons</td>
<td>638</td>
<td>183</td>
</tr>
<tr>
<td>Number of primary contigs</td>
<td>50020</td>
<td>32801</td>
</tr>
<tr>
<td>Number of secondary contigs</td>
<td>0</td>
<td>5905</td>
</tr>
<tr>
<td>Number of unigenes</td>
<td>50658</td>
<td>38889</td>
</tr>
<tr>
<td>Mean unigene length (bp)</td>
<td>637.70</td>
<td>685.76</td>
</tr>
<tr>
<td>Largest unigene length (bp)</td>
<td>4489</td>
<td>4897</td>
</tr>
<tr>
<td>Total assembly length (Mbp)</td>
<td>32.30</td>
<td>26.67</td>
</tr>
<tr>
<td>Average unigene coverage</td>
<td>5.49</td>
<td>5.99</td>
</tr>
<tr>
<td>Average assembly coverage</td>
<td>7.02</td>
<td>8.50</td>
</tr>
</tbody>
</table>

A comparison of the primary and secondary assemblies. Secondary assembly was used to join additional contigs and reduce redundancy in the final unigene set. The number of singletons and primary contigs in the secondary assembly are the number of sequences for each class retained from the primary assembly. Two methods were used to calculate coverage for each assembly. “Average unigene coverage” is the arithmetic mean of the average read-depth coverage for each unigene. “Average assembly coverage” is the weighted mean of the coverage for each unigene, weighted by unigene length (equivalent to the sum of the read lengths divided by the sum of the unigene lengths).
Transcriptome coverage and data quality

Because information about the actual size and composition of the transcriptome is often unknown, we utilized a simulation-based tool, ESTcalc [53], to estimate the expected depth and breadth of transcriptome coverage for this data set. The model for transcriptome coverage backing ESTcalc was parameterized using the well characterized *Arabidopsis thaliana* transcriptome and several “next-generation” sequencing runs using both normalized and non-normalized cDNA libraries [53]. Using the results from their simulations (retrieved using ESTcalc), our dataset is expected to have sequenced 87% of the transcriptome and to have tagged 100% of the genes, 70% of which are expected to be sequenced to 90% of their length (Table 4-3). Consistent with these estimates, we were able to identify 338 of 357 (94.7%) single copy eukaryotic ultra-conserved orthologs (UCOs [54]) and 155 of 163 (95.1%) shared single copy tribes from *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, and *Oryza sativa* [55]. These gene sets represent a conserved subset of genes expected to be present in eukaryotic and plant transcriptomes, respectively, and can be used as a proxy for overall gene detection and sampling breadth. As a final measure of gene detection in this data set, we utilized a bootstrapped data resampling approach using the distribution of reads in our final assembly (see methods section) to generate a unigene accumulation curve which plots the number of unigenes detected as a function of sequencing effort (Figure 4-2). Using this method, we estimate that on average 90%, 95% and 99% of the unigenes were tagged after approximately 323,445; 402,951; and 521,184 reads were sampled (Figure 4-2). On average, it took 1,357.2 reads to detect each of the last ten unigenes.
### Table 4-3 - Transcriptome coverage estimates: ESTcalc

<table>
<thead>
<tr>
<th>Input Parameters</th>
<th>ESTcalc estimate</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of technologies</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Technology</td>
<td>454 GS-FLX</td>
<td>454 GS-FLX (Titanium)</td>
</tr>
<tr>
<td>Library type</td>
<td>normalized</td>
<td>normalized</td>
</tr>
<tr>
<td>MB/Plate</td>
<td>254</td>
<td>254.0076</td>
</tr>
<tr>
<td>Reads/Plate</td>
<td>681722</td>
<td>681722</td>
</tr>
<tr>
<td>BP/Read (mean)</td>
<td>372.6</td>
<td>372.6</td>
</tr>
</tbody>
</table>

### Output Estimates

<table>
<thead>
<tr>
<th></th>
<th>ESTcalc estimate</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Assembled Sequence (MB)</td>
<td>26.2</td>
<td>26.67</td>
</tr>
<tr>
<td>Unigene count</td>
<td>32044</td>
<td>38889</td>
</tr>
<tr>
<td>Mean unigene length (bp)</td>
<td>819</td>
<td>685.80</td>
</tr>
<tr>
<td>Mean unigene length (longest unigene per gene, bp)</td>
<td>1143</td>
<td>—</td>
</tr>
<tr>
<td>Singleton yield (%)</td>
<td>19</td>
<td>0.0047</td>
</tr>
<tr>
<td>Percent transcriptome (%)</td>
<td>87</td>
<td>—</td>
</tr>
<tr>
<td>Percent of genes tagged (%)</td>
<td>100</td>
<td>—</td>
</tr>
<tr>
<td>Percent of genes with 90% coverage (%)</td>
<td>69.8</td>
<td>—</td>
</tr>
<tr>
<td>Percent of genes with 90% coverage by largest unigene (%)</td>
<td>56.4</td>
<td>—</td>
</tr>
<tr>
<td>Percent of genes with 100% coverage (%)</td>
<td>23.7</td>
<td>—</td>
</tr>
<tr>
<td>Percent of genes with 100% coverage by largest unigene (%)</td>
<td>22.2</td>
<td>—</td>
</tr>
</tbody>
</table>

Estimates of transcriptome coverage based on simulations modelled using the *Arabidopsis thalliana* floral transcriptome.
Figure 4-2 - Unigene accumulation curve
The mean number of unigenes detected as a function of the number of reads sampled. The complete set of reads in the 2-step assembly were shuffled and drawn at random for 1000 bootstrap replicates.

To identify potential contaminant sequences in the sample or sequencing library, we examined the taxonomic distribution of blastx hits for each unigene searched in the NCBI nr database. We examined both the taxonomic classification of the best hit as well extracted from the PlantTribes2.0 database ([56] and CWD, unpublished). These two as the lowest common ancestor (LCA) assignment for each unigene using MEGAN v.3.7.2 [57]. Of the 22,716 unigenes with a blast hit, only 1% had a best hit to an organism outside of the green plants and 0.5% of the unigenes with a blast hit received an LCA assigned taxon which is not within, or a super set of, land plants (Table 4-4). We also examined the unigene set for potential genomic DNA contamination by screening unigenes for blastn hits to the complete chloroplast genome sequence of *Pteridium aquilinum* (JPD, unpublished). None of the chloroplast sequences identified in the transcriptome were longer than 5 kb or contained more than five adjacent genes and thus can reasonably be considered putative transcripts [58, 59]. That we did not detect
Table 4-4 - Taxonomic distribution of unigene blastx hits in the nr database

<table>
<thead>
<tr>
<th>Taxonomic category</th>
<th>Best blastx hit</th>
<th>Lowest common ancestor for blastx hits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of</td>
<td>Percent of</td>
</tr>
<tr>
<td></td>
<td>unigenes</td>
<td>unigenes with hits</td>
</tr>
<tr>
<td>Eukaryotes</td>
<td>22,685</td>
<td>99.9%</td>
</tr>
<tr>
<td>Green plants</td>
<td>22,468</td>
<td>98.9%</td>
</tr>
<tr>
<td>“Green algae”</td>
<td>104</td>
<td>0.5%</td>
</tr>
<tr>
<td>Land plants</td>
<td>22,364</td>
<td>98.5%</td>
</tr>
<tr>
<td>“Bryophytes”</td>
<td>375</td>
<td>1.7%</td>
</tr>
<tr>
<td>Vascular plants</td>
<td>21,989</td>
<td>96.8%</td>
</tr>
<tr>
<td>Lycophytes</td>
<td>60</td>
<td>0.3%</td>
</tr>
<tr>
<td>Ferns</td>
<td>371</td>
<td>1.6%</td>
</tr>
<tr>
<td>Seed plants</td>
<td>21,558</td>
<td>94.9%</td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>4,817</td>
<td>21.2%</td>
</tr>
<tr>
<td>Angiosperms</td>
<td>16,741</td>
<td>73.7%</td>
</tr>
<tr>
<td>Animals</td>
<td>185</td>
<td>0.8%</td>
</tr>
<tr>
<td>Fungi</td>
<td>2</td>
<td>0.0%</td>
</tr>
<tr>
<td>Other eukaryotes</td>
<td>30</td>
<td>0.1%</td>
</tr>
<tr>
<td>Bacteria</td>
<td>22</td>
<td>0.1%</td>
</tr>
<tr>
<td>Artificial sequences,</td>
<td>9</td>
<td>0.0%</td>
</tr>
<tr>
<td>hits don’t pass threshold,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>or taxon not assigned</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unigenes were searched in the NCBI nr protein database using blastx with an e-value threshold of 1e-10, keeping the best ten hits. Of the 38,889 unigenes, 22716 (58.4%) had a positive hit. The lowest common ancestor (LCA) assignment for a sequence was calculated using the LCA algorithm implemented in MEGAN v3.7.2 [57] based on at least three blastx hits with a bitscore greater than 75 and within 10% of the best bitscore. Note: the predicted proteins from Selaginella moellendorfii are not currently included in the nr database and thus are not reflected in these results.
any long fragments of chloroplast DNA in the transcriptome assembly gives us confidence that our DNase treatment during RNA extraction was effective and the resulting cDNA library used in sequencing is free of contaminant genomic DNA.

**Functional annotation**

Unigenes were annotated with gene ontology (GO) terms, enzyme codes, and conserved protein domain functions using the Blast2GO suite [60-62]. Unigenes were first interrogated against the NCBI nr protein database using a blastx e-value threshold of 1e-10, keeping the top 20 high scoring alignments longer than 33 residues, resulting in 22,716 unigenes (58.4%) with positive blast hits. The longest open reading frame (ORF) from 11,940 unigenes (30.70%) had positive matches to conserved protein domains using InterProScan (IPS) searches implemented in Blast2GO. These results (nr blastx and IPS) were used to assign 64,578 GO terms to 15,855 unigenes (Appendix C). These GO terms were used to map 6,978 enzyme codes to 5,554 unigenes. Enzyme codes were then used then to retrieve and color 141 KEGG pathway maps. The full set of GO terms for each unigene was also mapped to the plant GO-slim vocabulary to examine the broad classification of gene functions represented in the transcriptome (Figure 4-3).

**Comparative genomics**

Unigenes were classified into 7,126 tribe (inflation level 3) and 9,548 orthogroup MCL clusters (Appendix C) in the PlantTribes2.0 gene family database on the basis of the best blastx hit to the inferred protein models of the ten complete plant genomes included in PlantTribes2.0 ([56] and CWD, unpublished). To evaluate the level of gene overlap between the *Pteridium* gametophyte transcriptome and other land plants, we
Figure 4-3 - Distribution of plant GO-slim functional categories

The relative proportion of plant GO-slim terms represented by more than 150 unigenes for the three major categories in the GO vocabulary (biological process, molecular function, and cellular component).
examined overlap in both PlantTribes2.0 orthogroup cluster membership and blastx hits for predicted proteins in *Physcomitrella patens*, *Selaginella moelendorfii*, and *Arabidopsis thaliana* (Figure 4-4).

**SSR identification**

A total of 2,320 perfect di-, tri-, and tetranucleotide simple sequence repeats (SSRs) longer than 9, 6, and 5 repeats, respectively, were identified in 1,990 unigenes using the SSR identification tool (SSRIT) [63]. Additionally, primers for 424 potentially amplifiable non-compound SSRs were chosen using Primer3 [64] as implemented in msatCommander [65] (Appendix C). Since this RNA was extracted from gametophytes derived from spores collected from a single diploid sporophyte, we are unable to determine the level of variation present at these SSR loci in natural populations.

**DISCUSSION**

We have used high-throughput sequencing data to characterize the gametophyte transcriptome of *Pteridium aquilinum*, a species for which very little genomic data are available. These data also represent an 865-fold increase over the expressed sequence data previously available for *Pteridium* in Genbank [9].

**Assembly quality**

Because contaminant adapter/primer sequences, polyA/T repeats, and low complexity end sequences can substantially compromise *de novo* assembly and can be difficult to completely remove (KM Dlugosch, personal communication), we aggressively filtered and trimmed the reads beyond the default instrument-level processing routines at the cost of sequence information loss (approximately 11.6 Mbp were removed,
Figure 4-4 - Unigene overlap with *Arabidopsis, Physcomitrella, and Selaginella*

(A) PlantTribes2.0 Orthogroups: Unigenes were assigned to Tribe- and OrthoMCL clusters derived from PlantTribes2.0 based on the best blast hit for each unigene. The presence of genes from *Arabidopsis thaliana, Physcomitrella patens,* and *Selaginella moellendorfii* in each OrthoGroup was evaluated. Of 38,889 total unigenes, 15,374 unigenes did not have a positive blast hit and thus were not assigned to an OrthoGroup cluster. 

(B) Blastx: The complete unigene set was queried against the complete set of predicted proteins in the genomes of *Arabidopsis thaliana, Physcomitrella patens,* and *Selaginella moellendorfii* using an e-value cut off of 1e-5. Unigenes with positive hits in more than one proteome are shown in the intersect for those species. Of 38,889 total unigenes, 13,992 unigenes did not have a positive blast hit.
representing 4.4% of the raw data). Considering the sheer quantity and depth of sequencing produced by next-generation sequencing platforms, we deemed this an acceptable level of loss to improve accuracy in the assembly. We also used a two-step assembly strategy to maximize the information content of our final unigene sequence set. We adopted this approach because MIRA is able to handle the large number of reads produced by next-generation sequencing technologies and utilizes a multi-pass strategy to identify and correct sequencing and assembly errors to produce a highly accurate assembly, but is sensitive to uneven sequencing depth of coverage and allelic diversity, resulting in a high number of redundant contigs. CAP3 is a proven and efficient DNA sequence assembler that can be used to join highly similar overlapping sequences, but is unable to handle the large number of reads produced by these new massively-parallel sequencing platforms. By combining these two assembly tools, we were able to join contigs and singletons that failed to assemble in MIRA to reduce sequence-level redundancy in our final unigene data set.

In examining the taxonomic distribution of nr blastx hits for the unigenes, we identified only a small proportion of sequences with best blast hits or LCA assignments outside of the green plants. When we examine these hits in greater detail, we find that many of them only align to short conserved domains, are hypothetical proteins of unknown function from model organisms, or are genes which are conserved across broad taxonomic levels, such as cytochrome P450, alpha-tubulin and dynein proteins. Additionally, because no other fern genomes have been sequenced, some of these sequences may represent novel fern genes. Thus the evidence indicates that there is very little contamination in these data.
Transcriptome coverage

While the simulations that underlay ESTcalc are based on the well characterized Arabidopsis thaliana floral transcriptome (approximately 18,000 genes with transcripts averaging 1,500 bp long) and assume perfect cDNA normalization and sequence assembly, Wall et al. [53] show that their results were highly predictive for empirical datasets from diverse eukaryotic species and tissues, making their simulations useful as a null model for predicting transcriptome coverage in other organisms. However, the predicted singleton rate for the level of sequencing we achieved is much higher than observed in our two-step assembly (19% versus 0.005%). This is likely an artifact of our primary assembly, in which MIRA discarded reads that were not assembled into any contig and thus were not carried forward into our final assembly. This default behavior of MIRA (-OUT:sssip=no), can be switched on in future unigene builds if singletons are desired. Other possible explanations for this deviation are that the cDNA library preparation [17, 66] or the longer read lengths obtained in this study resulted in a read to transcript profile that does not fit the ESTcalc models well. Finally, the transcriptome of Pteridium aquilinum may encompass fewer genes than Arabidopsis, resulting in a higher than predicted level of transcriptome coverage for our sequencing level and subsequently more efficient recruitment of singletons into contigs during assembly. Possible over-assembly seems an unlikely explanation for the deficit of singletons we observed, because of the stringent percent similarity parameters used for both stages of our two-step assembly (94% and 95% for MIRA and CAP3, respectively). Furthermore, the final number of unigenes in our final assembly remains higher, and the mean unigene length is shorter, than predicted by ESTcalc. Consistent with the ESTcalc estimate that we have tagged 100% of the transcripts present in this library, and based
on our unigene accumulation curve, the rate of new unigene detection for this cDNA library has declined to the point that any additional sequencing is unlikely to detect new genes, but may however serve to condense and join non-overlapping contigs in our assembly.

**Functional annotation**

The distribution of GO functional categories represented in the gametophyte transcriptome reflects the typical distribution of expressed plant genes for the biological process, cellular component and molecular function GO categories (Figure 4-3). Most of the unigenes annotated with a cellular component are localized to plastids or mitochondria, but a large number of them are also targeted for ribosomes or the plasma membrane. The molecular function of unigenes is heavily dominated by binding nucleic acids or proteins and metabolic activity, including hydrolase and kinase activity. The biological processes represented include all of the major cellular processes from transport and cellular organization to transcription, translation, and metabolism. Visual examination of the annotated/colored KEGG maps (not shown) indicates that we have captured all of the genes required for glycolysis and the citrate cycle, a large number of genes involved in nucleic and amino acid metabolism, chlorophyll biosynthesis, and plant hormone biosynthesis, including gibberellin, abscisic acid, strigolactone, cytokinin, brassinosteroid, and auxin.

**Comparative genomics**

The PlantTribes2.0 database contains an objective classification system for plant genes and gene families [56]. By identifying similar sequences in this classification system, we assigned unigene sequences with putative gene family identities. The most
abundant of these gene families present in the unigene set was the pentatricopeptide repeat protein (PPR) family, with over 900 unigenes classified as PPR proteins. We were also able to identify 65 unigenes classified in the MADS-box transcription factor family. Using this classification, gene sequences from *Pteridium* can be extracted for gene families of interest for use in studies of gene family evolution or phylogenomics. The overlap in orthogroup membership and blast hits for proteins in *Arabidopsis thaliana*, *Selaginella moellendorffii*, and *Physcomitrella patens* is similar (Figure 4A/B), but some striking differences can be observed. In both the PlantTribes and blast-based venn diagrams, most of the unigenes which were identified in *Arabidopsis*, *Selaginella*, or *Physcomitrella* are also shared across all three species. In the PlantTribes classification, most of the genes are shared with *Arabidopsis* (21,649 unigenes), which is the most closely related species included in this comparison, while slightly less and approximately equal gene representation is shared with *Selaginella* and *Physcomitrella* (19,649 and 19,485 unigenes, respectively). This is in contrast to the blast-based examination of gene set overlap in which *Arabidopsis* again has the greatest number of unigenes with hits (23,148 unigenes), but *Physcomitrella* has hits with 6,122 more unigenes than *Selaginella* (Figure 4-4). At first this seems counterintuitive because *Pteridium* shares a more recent common ancestor with *Selaginella* than with *Physcomitrella*, but both *Pteridium* and *Physcomitrella* have maintained a homosporous life cycle with a large independent gametophyte stage. If genes required in a free living gametophyte are maintained in *Pteridium* and *Physcomitrella*, but are free to be lost or evolve new functions in *Selaginella*, we expect to see a higher degree of similarity between *Pteridium* and *Physcomitrella* for a certain subset of genes.
This study is the first comprehensive sequencing effort and analysis of gene function in the transcriptome of a fern and represents the most extensive expressed sequence resource available in ferns to date, nearly 16 times more data than exists for *Adiantum capillus-veneris*. These data are an important new scientific resource for comparative evolutionary studies in land plants and will be of great value for studies of genome structure and function in ferns. These data can be used to develop microarrays for gene expression assays or serve as a reference transcriptome for RNA-seq experiments in *Pteridium*. As additional genome-scale projects in diverse plants are undertaken, these data will be of immense value in representing ferns, the sister clade to seed plants, in comparative genomic analyses.

**METHODS**

**Gametophyte culture, library preparation, and sequencing**

*Pteridium aquilinum* ssp. *aquilinum* spores (collection number: Wolf 83; sourced from a single sporophyte individual collected in Norwich, UK) were sown onto sterile agar nutrient media containing Bold’s macronutrients and Nitch’s micronutrients (prepared as described by [67]) and grown under white light. Whole gametophytes including both vegetative and sexually mature male, female, and hermaphoditic individuals of various ages (up to 9 months from germination) were flash frozen in liquid nitrogen and ground to a fine powder. Total RNA was isolated using the Sigma Spectrum Plant Total RNA Kit, incorporating on-column DNase I (Qiagen) digestion during extraction to remove traces of genomic DNA. Total RNA was concentrated by precipitating in 2.5 M ammonium acetate and 70 % ethanol, then resolubilizing the RNA
pellet in RNase-free water to approximately 500 ng/µL. Total RNA was quantified and its quality verified using an Agilent Bioanalyzer 2100. Total RNA was sent to the Center for Genomics and Bioinformatics at Indiana University, Bloomington (IU CGB), where a normalized transcriptome (cDNA) library optimized for Roche 454 GS-FLX Titanium sequencing was prepared using custom methods (K Mockaitis, unpublished, available upon request). This library was sequenced using the GS-FLX Titanium process on 3 regions of a 4-region PicoTitre plate, according to manufacturer instructions.

Sequence preprocessing, transcriptome assembly, and coverage assessment

Sequence reads generated in this study were deposited in the NCBI sequence read archive (SRA012887). Raw sequence reads that passed instrument software quality filters were trimmed of custom oligonucleotide adapter sequences (Justin Choi, unpublished, IU CGB). The resulting sequences were further processed with SeqClean [68] and SnoWhite v1.0.3 [69] to remove low quality, short, and contaminant sequences, and to aggressively trim polyA/T sequences. Cleaned reads were assembled \textit{de novo} in MIRA v2.9.46 [51, 70] using a minimum percent identity of 94% to align reads. This primary assembly was passed through a secondary assembly step in CAP3 (95% identity, 25 bp overlap) [52] to reduce redundancy in the final assembly and join additional contigs. The final set of unigene sequences analyzed in this paper is included with the reads in the Sequence Read Archive. Custom perl scripts were used to extract summary information about the reads and assemblies (Tables 4-1 and 4-2; scripts available by request from JPD). The average unigene coverage was calculated as the arithmetic mean of the average read-depth coverage for each unigene, whereas the average read-depth coverage for the assembly as a whole takes into account read and
contig lengths, and is equivalent to the per-base read-depth coverage across all of the unigenes.

We utilized a web-based tool, ESTcalc [53], to estimate the predicted level of transcriptome coverage for our data set. Input parameters to ESTcalc require that we specify the sequencing technology used (or a combination of technologies) and either the total sequencing level (Mbp), or the number of reads and an estimate of read lengths. We used the best approximation for sequencing technology available (454 GS-FLX) and the empirical values observed for the cleaned sequence data (254 Mbp or 681,722 reads with an average of 372.6 bp/read) to obtain our estimates. The estimates reported were identical whether we parameterized on total sequence or supplied read length information as well.

To determine the number of eukaryotic ultra-conserved orthologs (UCOs [54]) we captured in the *Pteridium* transcriptome data set, we queried a list of 357 UCO coding sequences from *Arabidopsis* (A Kozik, unpublished) into the unigene set with an e-value threshold of 1e-10 using NCBI tblastx. These blast results were then parsed to determine then number of UCOs with a positive hit that returned an amino acid alignment greater than 30 residues long.

We assessed the changing rate of new gene detection as a function of sampling effort (unigene accumulation curve, Figure 4-2) using a bootstrapped random sampling protocol implemented in a custom perl script (available from JPD on request). This script uses the empirical distribution of read number per unigene in our final assembly to randomly sample reads one at a time and tracks the total number of unigenes detected at each step. Because the order of sampling can impact the shape of this curve, we computed 1,000 replicate random sample orders and calculated the mean number of
unigenes detected after each draw. To evaluate the level of variation in the number of unigenes detected, we also calculated the 95% confidence interval on the number of unigenes. Using this curve, we then estimated the number of reads it took to capture an average of 90%, 95%, and 99% of the unigenes and the average number of additional reads required to detect each of the last 10 unigenes.

To evaluate for the presence of potential contaminating sequences, we examined the taxonomic distribution of blastx hits for the unigene set in the NCBI nr protein database using an e-value threshold of 1e-10. The top 10 blast hits for each unigene were kept and examined in MEGAN v.3.7.2 [57]. MEGAN is a tool built for the examination of metagenomic data sets and provides a number of useful functions to explore the information content of large blast results. The blast results for each unigene were mapped onto the NCBI taxonomy tree by examining just the best hit (lowest e-value) or by using the lowest common ancestor (LCA) algorithm [57]. LCA was determined using at least three blast hits with a bitscore greater than 75 and within 10% of the top bitscore for that unigene.

**Functional annotation**

The same blast search used to examine the taxonomic distribution of blast hits was used to identify putative homologous proteins and annotate each sequence with gene ontology (GO) terms using Blast2GO [60-62]. Blast2GO was also used to automatically handle InterProScan (IPS) searches to identify conserved protein domains in translations of the longest ORF in each unigene. Any GO terms associated with IPS hits were then merged into the blast-based GO annotation. GO terms were then used to map enzyme codes to each sequence. Enzyme codes were then used to automatically color and retrieve KEGG pathway maps [71, 72]. As a final step in examining a broad
functional representation of the gametophyte transcriptome, GO terms were mapped to
the reduced GO-slim ontology and visualized and explored with directed acyclic graphs
(not shown) and summarized with filtered pie charts including GO categories
represented by at least 150 sequences (Figure 4-3).

**Comparative genomics**

Unigenes were classified into tribe- and orthoMCL clusters in the PlantTribes2.0
database using a custom perl pipeline (dePamphilis lab, unpublished) which queries
each unigene against the complete inferred protein set from ten plant species that have
complete sequenced genomes, using a blastx e-value threshold of 1e-10. Unigenes
were assigned to clusters based on the best blast hit. Species used for blast searches
and gene clustering in the PlantTribes2.0 database include: *Chlamydomonas reinhardtii,*
*Physcomitrella patens, Selaginella moellendorfii, Oryza sativa, Sorghum bicolor, Vitis
vinifera, Populus trichocarpa, Medicago truncatula, Carica papaya, and Arabidopsis
thaliana*. Meta-information about each assigned cluster was extracted from the database
for each unigene and was included in the pipeline output file. Simple text-based
searches examined this information to retrieve gene family names and putative gene
family functional data. The shared single copy tribes for *Arabidopsis, Vitis, Populus,* and
*Oryza* were identified in the PlantTribes2.0 database and the number of these tribes
detected in the unigene set was determined by examining the pipeline output file.
Orthogroup assignments for *Pteridium* unigenes were examined for cluster membership
by *Selaginella, Physcomitrella,* and *Arabidopsis* to generate a venn diagram showing
putative gene level overlap (Figure 4-4A). Unigenes were also directly queried against
each of these proteomes using a blastx e-value threshold of 1e-10 to examine the
distribution of similar proteins in these three species. Venn diagrams were generated to graphically illustrate the overlap of unigenes for each proteome (Figure 4-4B).

REFERENCES


24. Zagrobelny M, Scheibye-Alsing K, Jensen NB, Møller BL, Gorodkin J, Bak S: **454 pyrosequencing based transcriptome analysis of *Zygaena filipendulae* with focus on genes involved in biosynthesis of cyanogenic glucosides.** *BMC Genomics* 2009, **10**:574.


32. Takahashi C: **Chromosome study on induced apospory in the bracken fern.** *La Kromosomo* 1961, **48**:1602-1605.

33. Webster BD, Steeves TA: **Morphogenesis in *Pteridium aquilinum* (L.) Kuhn - general morphology and growth habit.** *Phytomorphology* 1958, **8**:30-41.


42. Whittier DP: The origin and development of apogamous structures in the gametophyte of _Pteridium_ in sterile culture. _Phytomorphology_ 1962, **12**:10-20.


45. Beniston RG, Campo MS: Quercetin elevates p27(Kip1) and arrests both primary and HPV16 E6/E7 transformed human keratinocytes in G1. _Oncogene_ 2003, **22**:5504-5514.


48. Rodrigues Da Silva UDS, Matos DMDS: The invasion of _Pteridium aquilinum_ and the impoverishment of the seed bank in fire prone areas of Brazilian Atlantic forest. _Biodivers Conserv_ 2006, **15**:3035-3043.


57. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17**:377-386.


64. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.


68. The Gene Indices Project software tools: SeqClean [http://compbio.dfci.harvard.edu/tgi/software/]

69. SnoWhite: A cleaning pipeline for cDNA sequences [http://kdlugosch.net/software/]


CHAPTER 5

CONCLUSIONS

This dissertation takes a multifaceted approach to examine evolution in bracken, from a classic systematics and biogeographic study, through complete sequencing of the chloroplast genome, to surveying genes expressed in the reproductive life stage of ferns. I examine evolutionary events at scales from species divergence and hybridization within subspecies all the way to comparisons across major lineages of land plants.

In chapter 2, I reviewed the state of systematic knowledge within *Pteridium* and constructed a comprehensive phylogenetic framework for the genus, sampling bracken specimens across its geographic range. My results enabled specific biogeographic and hybrid origin hypotheses to be proposed and set the stage for further work on the systematics and evolution in bracken. This study also highlighted the importance of range-wide perspectives when undertaking evolutionary and monographic studies.

In chapter 3, I sequenced the complete chloroplast genome of *Pteridium aquilinum* and performed a genome-wide survey of chloroplast RNA editing. This study is the first reported use of Roche 454 sequencing and bioinformatic extraction/assembly of a complete chloroplast genome and the first study to utilize second-generation sequencing technologies to analyze RNA editing in a plastid genome. The chloroplast genome of bracken contains 117 different genes, including 84 protein coding genes, 4 ribosomal RNA genes, and 29 transfer RNA genes. A total of 551 unique C to U RNA editing sites and 300 U to C RNA editing sites in the chloroplast genome were experimentally detected. This is the highest known rate of RNA editing ever detected in a vascular plant, over 2.5 times that detected in *Adiantum capillus-veneris* and over 20 times that seen in seed plants.
Chapter 4 leveraged high-throughput transcriptome sequencing to survey genes expressed in the gametophyte life stage in *Pteridium*. This study represents the first comprehensive sequencing effort and analysis of gene function in the transcriptome of a fern and represents the most extensive expressed sequence resource available in ferns to date. These data are an important new scientific resource for comparative evolutionary studies in land plants and will be of great value for studies of genome structure and function in ferns. These data can be used to develop microarrays for gene expression assays or serve as a reference transcriptome for RNA-seq experiments in *Pteridium*. As additional genome-scale projects in diverse plants are undertaken, these data will be of immense value in representing ferns, the sister clade to seed plants, in comparative genomic analyses.

In this dissertation I have presented three studies which examine genome structure, function, and evolution in bracken. By shifting my sampling strategy dependent on specific research objectives and questions presented in each study I have been able to examine evolution in bracken at different scales, from a reconstruction of phylogenetic relationships and ancient evolutionary events to an examination of chloroplast sequence composition and chloroplast RNA metabolism, and to a broad survey of expressed genes in one life stage of the fern life cycle. These studies together form the foundation for a research program in bracken evolutionary genetics and genomics which shows potential for a great many projects in the future.
APPENDIX A

PERMISSION TO REPRINT CHAPTER 2

Step 3: Order Confirmation

Thank you for your order! A confirmation for your order will be sent to your account email address. If you have questions about your order, you can call us at 978-646-2600, M-F between 8:00 AM and 6:00 PM (Eastern), or write to us at info@copyright.com.

Confirmation Number: 2762099
Order Date: 05/17/2010
Order Total: $3.00

If you pay by credit card, your order will be finalized and your card will be charged within 24 hours. If you pay by invoice, you can change or cancel your order until the invoice is generated.

Payment Information
Joshua Der
josh.der@usu.edu
+1 (435)7604907
Payment Method: CC ending in 2486

Order Details
AMERICAN JOURNAL OF BOTANY : THE JOURNAL FOR ALL PLANT BIOLOGISTS
Order detail ID: 42389398
Permission Status: Granted

Permission type: Republish into a book, journal, newsletter...
Requested use: Dissertation
Republish title: GENOMIC PERSPECTIVES ON EVOLUTION IN BRACKEN FERN
Republishing organization: UTAH STATE UNIVERSITY
Organization status: Non-profit 501(c)(3)
Publication year: 2009
Publication Type: Journal
Publisher: BOTANICAL SOCIETY OF AMERICA
Rightsholder: BOTANICAL SOCIETY OF AMERICA INC
Author/Editor: Joshua P. Der, John A. Thomson, Jeran K. Stratford, and Paul G. Wolf

Type of content: Full article/chapter
Description of requested content: Global Chloroplast Phylogeny and Biogeography of Bracken (Pteridium: Dennstaedtiaceae)
Page range(s): Vol. 96, pp. 1041-1049
Translating to: No Translation
Requested content’s publication date: 05/01/2009
Your reference: JOSHUA P. DER, DISSERTATION, CHAPTER 2.

$3.00
Confirmation Number: 2762099

Special Rightsholder Terms & Conditions
The following terms & conditions apply to the specific publication under which they are listed

There are no special terms.

Copyright Clearance Center Terms and Conditions
The following terms & conditions apply to your entire order.

For uses traditionally reported under Copyright Clearance Center’s (CCC) pay-per-use service to photocopy for general business use, library reserves, ILL/document delivery, etc., (formerly TRS), the applicable terms and conditions follow immediately below in Section A. For republication or electronic republication uses (including e-mail, intranet, extranet and Internet postings), the applicable terms and conditions follow in Section B.

Section A
Pay-per-use service to photocopy for general business use, library reserves, ILL/document delivery, etc., (formerly TRS): Rights to Copy

Scope of Authorization. You may reproduce chapters and articles from the publications included in CCC’s pay-per-use service by reporting those copies and paying the rightsholders’ fees through CCC. Reproduction by photocopy machine is covered by this pay-per-use system of authorizations, as is the facsimile transmission of copies (“faxing”, “telecopying” or equivalent technology).

Limitations on Use. Authorizations are limited (a) to the use of the person or entity reporting to CCC and (b) to the use of a specific customer of the person or entity reporting to CCC unless otherwise indicated in the permissions policy statement printed in the publication. Other limitations may be imposed by individual publications, as described in their own masthead or permission policy statements.

Uses excluded:
- electronic storage of any reproduction (whether in plain-text, PDF, or any other format) other than on a transitory basis
- the input of works into computerized databases
- reproduction of an entire publication (cover-to-cover)
- reproduction for resale to anyone other than a specific customer, republication in a different form

Please obtain authorizations for these uses through other CCC services or directly from the rightsholder.

If you have any comments or questions about this pay-per-use service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to info@copyright.com.

Section B
All Other Pay-Per-Use Services, (formerly DPS and RLS): Terms and Conditions

1. Licenses are granted by Copyright Clearance Center, Inc. ("CCC"), as agent for the rightsholder identified on the Order Confirmation (the "Rightsholder"), and are for republication or electronic reproduction of a copyrighted work as described in detail on the Order Confirmation (the "Work"). "Reproduction", as used herein, generally means the inclusion of the Work, in whole or in part, in a new work which is also described on the Order Confirmation. "Electronic reproduction", as used herein, generally means e-mail use (including instant messaging or other electronic transmission to a defined group of recipients) or posting on an intranet, extranet or Internet site. CCC provides this license through its pay-per-use services. User shall be deemed to have accepted and agreed to all of these terms and conditions if User republishes or makes any electronic reproduction of the Work in any fashion.

2. All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set...
forth on that document includes only those rights expressly set forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. While User may exercise the rights licensed immediately upon issuance of the Order Confirmation, the license is automatically revoked and is null and void, as if it had never been issued, if complete payment for the license is not received on a timely basis either from User directly or through a payment agent, such as a credit card company. In the event that the material for which a republication (rather than an electronic reproduction) license is sought includes third party materials (such as photographs, illustrations, graphs and similar materials), User is responsible for identifying, and seeking separate licenses (under this service or otherwise) for, any of such third party materials which are identified anywhere in the work as included in the Work by permission; without a separate license, such third party materials may not be used. In the event that the material for which an electronic reproduction license is sought includes third party materials (such as photographs, illustrations, graphs and similar materials) which are identified as included in a Work by permission, such third party materials may not be reproduced except in the context of the Work. Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is "one-time", (ii) is non-exclusive and non-transferable, and (iii) is limited to use completed within 30 days for any use on the Internet, 60 days for any use on an intranet or extranet and one year for any other use, all as measured from the "republication date" as identified in the Order Confirmation, if any, and otherwise from the date of the Order Confirmation. Upon completion of the licensed use, or at the end of the period identified in the previous sentence (if earlier), User shall immediately cease any new use of the Work(s) and shall destroy or (particularly in the case of electronic republications) render inaccessible (such as by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User's stock at the end of such period). All rights not expressly granted are reserved; any license granted is further limited as set forth in any restrictions included in the Order Confirmation and/or in these terms and conditions.

3. Use of proper copyright notice for a Work is required as a condition of any license granted hereunder. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: "Used [or, as appropriate, "Reproduced" or "Republished"] with permission of [Rightsholder's name], [Work's title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc." Such notice must be placed immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link), in a reasonably legible font size, unless, in the case of a republication license, the new work containing the republished Work contains a substantial number of credits or copyright notices and they are all printed in a single, easily located place. Failure to include the required notice results in loss to the Rightsholder and CCC, and User hereby agrees to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.

4. User may not make or permit any alterations to the Work, unless expressly set forth in the Order Confirmation (after request by User and approval by Rightsholder). No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties' rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder.

5. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.

6. User hereby indemnifies and agrees to defend the Rightsholder and CCC, and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.

7. LIMITATION OF LIABILITY. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.

8. WARRANTIES. THE WORK(S) AND RIGHT(S) ARE PROVIDED "AS IS". CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

9. The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder, subject to the following exception. The exception: in cases where a person is specifically engaged by a second person to provide reproductions of Works to (and/or to obtain licenses on behalf of) the second person and EITHER (i) the first person expressly notifies the second person of its obligation not to redistribute or otherwise use the material without authorization, OR (ii) the second person is identified in
the Order Confirmation as the publisher of the material to include the Work licensed under this service, then the license may be transferred to the second person and both the first person and the second person shall be jointly deemed to be the "User" under the Order Confirmation and these terms and conditions for that transaction. An organizational User and its principals, employees, agents and affiliates are jointly and severally liable for the performance of all payments and other obligations hereunder. No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties.

10. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

11. Any failure by User to pay any amount when due, or any use by User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 10 days of notice thereof shall result in immediate termination of such license without further notice. Invoices are due and payable on "net 30" terms. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder’s ordinary license price therefore; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder’s ordinary license price for the most closely analogous licensable use plus Rightsholder’s and/or CCC's costs and expenses incurred in collecting such payment.

12. The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court.

If you have any comments or questions about this pay-per-use service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to info@copyright.com.
17 May 2010

Joshua P. Der
Department of Biology
Utah State University
5305 Old Main Hill
Logan, UT 84322-5305
(435) 760-4907

Dear Jeran K. Stratford,

I am in the process of preparing my dissertation in the Biology department at Utah State University, and will be completing my degree in Summer of 2010. Utah State University requires all co-authors on published chapters to grant permission to use published materials in theses or dissertations, and I am requesting your permission to include the journal article co-authored by you as shown below.

Please indicate your approval of this request by signing in the space provided below.

If you have any questions, please contact me.

Thank you for your cooperation,

Joshua P. Der

I hereby give permission to Joshua P. Der to reprint the following material in his dissertation:


Signed

Date
Permission to use published material in Thesis or Dissertation

I hereby give permission to Joshua P. Der to reprint the following material in his PhD dissertation for the Biology department at Utah State University:


I do so with pleasure, and wish to acknowledge Mr Der's predominant role in all aspects of both the laboratory and data analyses involved, as well as the writing and presentation of the results for publication.

[Signature]

A. Thomson

(John A. Thomson MSc MAgSc PhD)
Professor Emeritus, University of Sydney
Honorary Research Associate, National Herbarium of NSW

Email  
John.Thomson@rbgsyd.nsw.gov.au
john@accsoft.com.au
pteridium@bigpond.com

Date  18 May 2010
APPENDIX B

VOUCHER INFORMATION FOR CHAPTER 2

Taxa, vouchers, localities and GenBank accession numbers for bracken specimens used in Chapter 2. Voucher specimens are deposited in the following herbaria: NSW = National Herbarium of New South Wales, Australia; UTC = Intermountain Herbarium, Utah State University, USA.

**Taxon** — Voucher: Collector (Herbarium accession #) Sample isolate ID; Locality; Latitude; Longitude; GenBank accessions: trnS-rpS4 spacer+gene; rpL16 intron.

**Pteridium aquilinum subsp. aquilinum** — T. Reichstein (NSW 420392) 017 FZSW; Switzerland, Filzbach; 47.12; 9.10; FJ177240; FJ177159; R. Prelli (NSW 420391) 031 BRFR; France, Britagne, Erquy; 48.38; 2.27; FJ177241; FJ177160; M. E. Gillham (NSW 420404) 065 CLYW; United Kingdom, Wales, Neath Valley; 51.65; -3.80; FJ177242; FJ177161; A. F. Dyer (NSW 420393) 099 EDSC; United Kingdom, Scotland, Pentland Hills; 55.5; -3.25; FJ177243; FJ177162; R. Viane (NSW 420405) 103ASSP; Spain, Asturias Province, Villanueva; 43.37; -5.83; FJ177244; FJ177163; M. E. Gillham (NSW 420396) 106 OXLN; United Kingdom, England, London, Oxshott Common; 51.38; -0.42; FJ177245; FJ177164; F. Piccoli (NSW 420390) 188 CORF; France, Corsica, Haute Corse; 42.57; 9.30; FJ177246; FJ177165; E. Sheffield (NSW 420397) 218 ACAW; United Kingdom, Wales, Caerphilly Mt.; 51.35; -0.42; FJ177247; FJ177166; F. Piccoli (NSW 420394) 202 RVIT; Italy, Ravenna; 44.42; 12.2; FJ177248; FJ177167; C. N. Page (NSW 420408) 218 ACAW; United Kingdom, Wales, Caerphilly Mt.; 51.35; -0.42; FJ177249; FJ177168; B. Capezedo & A. E. Salvo (NSW 420374) 226 LBSP; Spain, Cadiz, Los Barrios; 36.18; -5.50; FJ177250; FJ177169; M.E. Gillham (NSW 420376) 307 BRGW; United Kingdom, Wales, Bridgend; 32.65; -16.88; FJ177251; FJ177170; V. Korzhenevsky (NSW 627463) TAUR; Ukraine, Caucasus Mountains, Crimea; 44.98; 34.62; FJ177288; FJ177207.

**Pteridium aquilinum subsp. capense** (Thunb.) C. Chr. — C. S. McMaster (NSW 617745) 110 KUFZ; Zambia, Kundalilla Falls; -12.45; 30.13; FJ177256; FJ177175; V. Rashbrook (NSW 203322) 191 GSAF; Republic of South Africa, Grahamstown; -33.19; 26.32; FJ177257; FJ177176; Watling (NSW 20334) 228 MCAM; Cameroon, Mundamba; 4.42; 8.87; FJ177258; FJ177177; P. J. Myerscough (NSW 20333) 353 BBSA; Republic of South Africa, Betty’s Bay, Silver Sands; -34.35; 18.9; FJ177260; FJ177179; B. E. Okoli (NSW 203328) 368 NDNG; Nigeria, Niger Delta; 4.83; 6.25; FJ177262; FJ177181; A. C. Chikuni (NSW 617743) 503CAPM/ZOMA; Malawi, Zomba Plateau; -15.3; 35.32; FJ177263; FJ177182.

**Pteridium aquilinum subsp. centrali-africanum** Hieron. — C. S. McMaster (NSW 617746) 112 ZAMB/CH3Z; Zambia, Mkushi, Chilongoma Hills; -14.38; 29.53; FJ177264; FJ177183; C. S. McMaster (NSW 617747) AKFZ; Zambia, Kundalilla Falls; -12.45; 30.13; FJ177265; FJ177184; C. S. McMaster (NSW 617757) CAF4; Zambia, Mkushi, Chilongoma Hills; -14.38; 29.53; FJ177266; FJ177185.

**Pteridium aquilinum subsp. decompositum** (Gaudich.)
Lamoureux ex J. A. Thomson — C. W. Smith (NSW 420357) 292 MAHI; USA, HI, Maui, Haleakala National Park; 20.75; -156.17 FJ177268; FJ177187. Pteridium aquilinum subsp. japonicum (Nakai) Á. Löve & D. Löve — S.-M. Chew (NSW 420347) 029 TTTN; Taiwan, Nankang, Taipei; 21.02; 121.02; FJ177271; FJ177190; Y.-J. Zhang (NSW 420336) 071 AIJP; Japan, Honshu, Aichi Prefecture, Mt. Ogura; 36.17; 138.25; FJ177273; FJ177192; T. Kumashiro (NSW 420348) 280 CHJP; Japan, Honshu, Chiba Prefecture; 35.60; 140.10; FJ177275; FJ177194; T. Miyata

Pteridium aquilinum subsp. latiusculum (Desv.) Hultén — E. Klekowski (NSW 420315) 143 YCCM; USA, MA, Yarmouth; 41.70; -70.23; FJ177277; FJ177196; W. H. Wagner Jr (NSW 420310) 147 WMCH; USA, MI, Waterloo; 42.15; -84.24; FJ177278; FJ177197; D. Barrington (UTC 249671) DSB 2287; USA, VT, Colchester, Niquette Bay State Park; 44.58; -73.20; FJ177280; FJ177199; P. Wolf

Pteridium aquilinum subsp. pinetorum (C. N. Page & R. R. Mill) J. A. Thomson — V. Bogatyr (NSW 420402) 164 KUKR; Ukraine, Kiev; 50.43; 30.5; FJ177235; FJ177154; E. A. Ershova & A. I. Shmakov (NSW 807731) N2nR; Russia, Siberia, Altai Plain; 53.35; 83.73; FJ177237; FJ177156; O. N. Perestoronina (NSW 807736) RU4K; Russia, Kirov Region; 58.60; 49.65; FJ177238; FJ177157. Pteridium aquilinum subsp. pseudocaudatum (Clute) Hultén — E. Sheffield (NSW 420317) 169 FLUS; USA, FL, Cape Kennedy; 28.47; -80.47; FJ177282; FJ177201; P. G. Wolf & M. D. Windham (NSW 420316) 203 HFLA; USA, FL, Hawthorne; 29.55; -82.08; FJ177283; FJ177202; P. Soltis (UTC 249632) Der 68; USA, FL, Paynes Prairie State Preserve; 29.52; -82.30; FJ177299; FJ177218. Pteridium aquilinum subsp. pubescens (Underw.) J. A. Thomson, Mickel & Mehltreter — D. Barrington (NSW 419173) 100 AOUS; USA, OR, Ashland; 42.23; -122.73; FJ177284; FJ177203; J. Schneller (NSW 420312) 325 OWUS; USA, WA, Olympic Peninsula; 47.55; -124.24; FJ177285; FJ177204; J. Der (UTC 249629) JPD 66; Canada, British Columbia, Vancouver; 49.26; -123.26; FJ177297; FJ177216; J. Der (UTC 247719) JPD 67; USA, CA, Feather River Canyon; 39.74; -120.71; FJ177298; FJ177217. Pteridium aquilinum subsp. wightianum (J. Agardh) W. C. Shieh — R. D. E Jayesekara (NSW 419562) 001 AMSL; Sri Lanka, Ambewela; 7.03; 80.59; FJ177289; FJ177208; J. V. Pancho (NSW 420371) 007 KLPH; Philippines, Luzon, Kinabuhayan; 14.25; 121.5; FJ177290; FJ177209; T. Partomihardjo (NSW 419534) 068 SERI; Indonesia, Molucca Island, Seram; 3.10; 129.05; FJ177291; FJ177210; R. Kiew (NSW 420251) 182 PMAL; Malaysia, Pahang; 4.00; 105.00; FJ177292; FJ177211; S. P. Khullar (NSW 419570) 305 YGIN; India, Garhwal, Yamunotri Hills; 30.92; 78.47; FJ177293; FJ177212; J. A. Thomson (NSW 420257) 354 WFNQ; Australia, N. Queensland, Wallaman Falls; -18.55; 145.80; FJ177294; FJ177213; M. D. Dassayanake (NSW 419571) 362 HKSL; Sri Lanka, Hakgala; 6.55; 80.48; FJ177295; FJ177214; S. J. Moore (NSW 705077) 416 TREV;
Taiwan, Taitung Hsieh; 23.27; 120.96; FJ177296; FJ177215. 

**Pteridium arachnoideum** (Kauff.) Maxon — B. Perez-Garcia (NSW 420308) 144 RMEX; Mexico, Molango, Rio Malila; 20.48; -98.44; FJ177221; FJ177140; M. G. E. Noronha & P. G. Windisch (NSW 420303) 317 SPBR; Brazil, Sao Paulo; -22.42; -49.00; FJ177219; FJ177138; M. E. Alonso-Amelot (NSW 505771) ME2-1VNZA; Venezuela, Mérida, Cerro La Bandera; 8.42; -69.03; FJ177220; FJ177139. 

**Pteridium caudatum** (L.) Maxon — J. Villalobos-Salazar & M. Firenczi (NSW 420247) 238 HECR; Costa Rica, Heredia; 10.00; -84.08; FJ177222; FJ177141; J. Lenne (NSW 420295) 274 QCOL; Colombia, Valle, Quisquina; 3.60; -76.48; FJ177223; FJ177142; A. C. Jermy & T. G. Walker (NSW 420305) 323 COCR; Costa Rica, Cordillera; 9.15; -83.8; FJ177224; FJ177143; M. E. Alonso-Amelot (NSW 505770) MD2-2VENZ; Venezuela, Mérida, La Hechicera; 8.42; -71.03; FJ177225; FJ177144. 

**Pteridium esculentum** (G. Forst.) Cockayne — J. A. Thomson (NSW 420273) 083 WAWA; Australia, WA, Waroona; -32.51; 115.59; FJ177226; FJ177145; J. A. Thomson (NSW 420274) 127 KEWA; Australia, WA, Kenton; -34.57; 117.01; FJ177227; FJ177146; J. A. Thomson (NSW 420848) 213 CRAN; Australia, NSW, Craven; -32.15; 151.95; FJ177228; FJ177147; J. A. Thomson (NSW 420264) 275 KONC; New Caledonia, Mt. Koghi; -22.17; 166.50; FJ177229; FJ177148; C. Surman (NSW 420267) 324 HNNZ; New Zealand, South Island, Nelson, Hira Forest; -41.18; 173.17; FJ177230; FJ177149; C. Surman (NSW 420289) 332 SVNZ; New Zealand, North Island, Wellington, Stokes Valley; -41.07; 175.08; FJ177231; FJ177150; J. A. Thomson (NSW 420879) 387 CRDO; Australia, N. Qld, Credition; -21.20; 148.53; FJ177232; FJ177151; J. A. Thomson (NSW 420269) 401 RYNA; Australia, NSW, Sydney, Ryde; -33.47; 151.05; FJ177233; FJ177152; P. Wolf (UTC 250545) PGW 638; New Zealand, North Island, Ruakura; -37.78; 175.31; FJ177234; FJ177153. 

**Pteridium semihastatum** (Wall. ex J. Agardh) S. B. Andrews — P. Brocklehurst & G. Wightman (NSW 420865) 251 PFNT; Australia, NT, Petherick's Forest Park; -13.06; 130.24; FJ177286; FJ177205; Mujamil & T. J. Ho (NSW 419554) 278 KMAL; Malaysia, Selangor, Kapar; 3.07; 101.24; FJ177287; FJ177206.
APPENDIX C

ADDITIONAL FILES FOR CHAPTER 4

Additional files for analyses performed in Chapter 4 can be downloaded from:
https://docs.google.com/leaf?id=0BwOnBtQxjHP0MjhkYjdiZTAzMzYWNC00N2M4LWlyZmEtMGQzYTdlMjBjOWFi&hl=en

Additional file 1 – Unigene functional annotations from Blast2GO
Spreadsheet file (csv) with unigene GO annotations assigned by Blast2GO. Each annotation is on a line, unigenes with multiple GO terms have more than one line. Five columns correspond to: unigene ID, functional description based on blast hits, GO ID, GO term, and GO category (P: biological process; C: cellular component; F: molecular function).
File available at: https://docs.google.com/leaf?id=0BwOnBtQxjHP0ZDdhZjI2MTMtoDBlZC00MjA1LWJmOTctNmRtMDI5MW15ZmVt&hl=en

Additional file 2 – PlantTribes2.0 gene family classification
Spreadsheet file (csv) with putative tribe and orthogroup assignments (stringency level 3) for each unigene with cluster membership counts from the ten proteomes included in PlantTribes2.0. Functional and gene family descriptors for clusters are primarily inherited from the Arabidopsis thaliana genes included in the cluster.
File available at: https://docs.google.com/leaf?id=0BwOnBtQxjHP0YTFmMzc1MjAtOTdldZl00NY4LTlhODktMjZhZmEwZWVmZThj&hl=en

Additional file 3 – Primer sequences and details for SSR loci
Spreadsheet file (csv) with primer sequences for potentially amplifiable SSR loci selected using MSATCOMMANDER.
File available at: https://docs.google.com/leaf?id=0BwOnBtQxjHP0ZGJmNzl2ODMtM2U1Ny00ZTRlLTkxODItOGFhYjM4N2M3YmYw&hl=en
Career objective

Faculty position at a liberal arts or state university that emphasises research and education.

Professional address

Department of Biology
Utah State University
5305 Old Main Hill
Logan, UT 84322-5305

Education

Advisor: Dr. Paul G. Wolf

Advisor: Dr. Daniel L. Nickrent

Advisor: Dr. Michael R. Mesler

Peer-Reviewed Publications


Mark W. Ellis, Jessie M. Roper, Rochelle Gainer, Joshua P. Der and Paul G. Wolf. 2009. The taxonomic designation of Eriogonum corymbosum var. nilesii (Polygonaceae) is supported by AFLP and cpDNA analyses. Systematic Botany 34 (4): 693-703. doi: 10.1600/036364409790139736

1 Recipient of SIUC Alumni Association Outstanding Masters Thesis Award and Midwestern Association of Graduate Schools Distinguished Masters Thesis Award.
Peer-Reviewed Publications (continued)


**Grants**

2008 Center for Integrated Biosystems Student Research Grant, Utah State University. “Developing genomic resources for both phases of the plant life cycle.” $6,948.50.

**Awards and Special Recognition**

2006 Botanical Society of America. Pteridological Section Student Travel Award.

2006 Midwestern Association of Graduate Schools. Distinguished Master’s Thesis Award.

2005 Southern Illinois University, Carbondale Alumni Association. Outstanding Master’s Thesis Award.

2004 National Academy of Sciences. Travel award to attend the Arthur Sackler Colloquium on Systematics and the Origin of Species, On Ernst Mayr’s 100th Anniversary.


2003 Humboldt State University. Outstanding Student of the Year: Outstanding Contribution to a University Organization.

**Invited Seminars**

Research Reports


Published Abstracts


Additional Presentations


Teaching Experience

Utah State University, 2005-present
- Plant Taxonomy (1 semester)
- General Biology 1 Lab (1 semester)
- General Biology 2 Lab (4 semesters)
- Genetics Lab (3 semesters)
- Biological Discovery Lab (1 semester)
- Genetics Lecture (5 guest lectures)

Southern Illinois University, Carbondale, 2003-2005
- Plant Diversity Lecture (2 guest lectures)
- Plants and Society Lab (1 semester)
- Plant Diversity Lab (1 semester)

Service and Leadership Experience

2005 – 10  Graduate Representative, Curriculum Committee, Department of Biology, Utah State University.

2004 – 05  President, Plant Biology Graduate Student Organization, Southern Illinois University, Carbondale.

2004 – 05  Treasurer, Midwestern Ecology and Evolution Conference, Southern Illinois University, Carbondale.
Service and Leadership Experience (continued)

2004 – 05 Graduate Representative, Curriculum Committee, Department of Plant Biology, Southern Illinois University, Carbondale.

2001 – 02 Student Representative, Sexual Assault Advisory Committee, Humboldt State University.

2000 – 02 Outdoor Youth Counselor, Leadership Education Adventure Program, Youth Educational Services, Humboldt State University.

2000 – 01 Vice President, National Residence Hall Honorary, Humboldt Chapter, Humboldt State University.

1999 – 2000 Living Group Advisor / Resident Assistant, Department of Housing, Humboldt State University.

1998 – 99 President, Residence Hall Association, Humboldt State University.

References

Dr. Paul G. Wolf
Department of Biology
Utah State University
5305 Old Main Hill
Logan, UT 84322
(435) 797-4034
wolf@biology.usu.edu

Dr. Daniel L. Nickrent
Department of Plant Biology
Southern Illinois University, Carbondale
Carbondale, IL 62901-6509
(618) 453-3223
nickrent@plant.siu.edu