Genomic changes associated with transitions to heterospory and genome downsizing in land plants

Rijan R. Dhakal

Follow this and additional works at: https://louis.uah.edu/uah-theses

Recommended Citation
https://louis.uah.edu/uah-theses/391

This Thesis is brought to you for free and open access by the UAH Electronic Theses and Dissertations at LOUIS. It has been accepted for inclusion in Theses by an authorized administrator of LOUIS.
Genomic changes associated with transitions to heterospory and genome downsizing in land plants

Rijan R. Dhakal

A THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Biological sciences in The Department of Biology to The Graduate School of The University of Alabama in Huntsville

December 2022

Approved by:
Dr. Paul Wolf, Research Advisor/Committee Chair
Dr. Paul Wolf, Committee Member
Dr. Alex Harkess, Committee Member
Dr. Jerome Baudry, Committee Member
Dr. Paul Wolf, Department Chair
Dr. Rainer Steinwandt, College Dean
Dr. Jon Hakkila, Graduate Dean
ABSTRACT

Genomic changes associated with transitions to heterospory and genome downsizing in land plants

Rijan R. Dhakal

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Biological sciences

Biology
The University of Alabama in Huntsville
December 2022

Green plants, ubiquitous and essential in the biosphere, have two reproductive life cycles: heterosporous and homosporous. Heterosporous plants are generally seen more often in daily life and tend to be the source of staple foods. Heterosporous reproduction has evolved at least 11 times from ancestral homosporous lineages. Heterospory, presumed to have selective advantages, has a perplexing correlation with chromosome numbers. Here I have used novel data, tools, and methodology to discover 19, so far unknown, gene families, with the potential to explain the correlation between the reproductive life cycle in green plants and chromosome numbers. The discovery of these gene families provides a potential avenue for gain-of-function functional genomics research. Such studies could provide further insights that could help explain the evolution of traits in green plants, which are a crucial part of the surrounding ecosystems.
Acknowledgements

I would like to thank Dr. Paul Wolf (University of Alabama in Huntsville), Dr. Alex Harkess (Auburn University), Dr. Jerome Baudry (University of Alabama in Huntsville) and Dr. Norman Wickett (Clemson University) for guiding me throughout the research that has shaped up to be this thesis.

I am grateful to my advisors for providing an academic environment which was challenging and novel to me while he was easy to approach for help throughout.

I would also like to extend my gratitude to NSF grant DEB 1911459, which provided the funds for this project.

I would also like to thank the Alabama Supercomputing Authority (ASC) for generously providing the computer environment we needed to see this research completed. I would especially like to thank Dr. David Young at the ASC for being completely helpful throughout this process.

I am also thankful to the maintainers of sophisticated, peer-reviewed, free and open source software libraries/packages that made this project at all possible. The software packages used in this project have more than 2 decades of cutting-edge research behind them and it is a near miracle that software this advanced it freely available to researchers.
# Table of Contents

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

Acknowledgements ................................. iii

Table of Contents .................................... v

List of Figures ....................................... vi

List of Tables ....................................... vii

Chapter 1. Introduction ............................ 1

1.1 Background ....................................... 1

Chapter 2. Research Objectives .................. 9

Chapter 3. Research Questions .................... 11

Chapter 4. Methods ................................. 12

4.1 Sample Selection ............................... 13

4.2 Obtaining genomic information ............... 13

4.3 Orthology inference ............................ 14

4.4 Species tree inference .......................... 15
4.5 Gene family expansion and contraction inference 15
4.6 Selection analysis 16
4.7 Multiple sequence alignment and codon alignment 16
  4.7.1 Gene tree generation 17
  4.7.2 Lineage specific selection analysis 17

Chapter 5. Results 19
5.1 Copy number variation analysis 19
5.2 Families with significant changes on each node 19
5.3 Selection analysis data 22

Chapter 6. Discussion 25
6.1 Copy number variation analysis 25
6.2 Selection analysis 26
6.3 Limitations of the study 27

References 31
## List of Figures

1.1 The homosporous life cycle in green plants. ..................................... 2  
1.2 The heterosporous life cycle in green plants. ................................. 3  
1.3 The land plant phylogeny and the multiple origins of heterospory. ........ 7  
1.4 The average number of chromosomes in homosporous and heterosporous plants [16]. ................................................................. 8  

4.1 An illustration of the binary, rooted, ultrametric tree used as the base species tree. ................................................................. 18  

5.1 Numbers of significant gene family copy number contractions or expansions in the three heterosporous clades. ......................... 23  
5.2 An example of what the visualization of branch-site selection analysis looks like. ................................................................. 24
## List of Tables

4.1 The distribution of the numbers of samples across their sources. . 14

5.1 Identity and most common associated function of the families that were expanding significantly on the seed plant node. . . . . . . . 20

5.2 Identity and most common associated function of the families that were expanding significantly on the heterosporous lycophyta node. 21

5.3 Identity and most common associated function of the families that were expanding significantly on the heterosporous lycophyta node. 21
Chapter 1. Introduction

1.1 Background

The evolution of sexual reproduction is a hallmark of eukaryotic life; various genomic forces led to the evolution of sexual reproduction during the evolutionary trajectory that led to the last eukaryotic common ancestor and its salient features separating it from its prokaryotic ancestors. Once the fundamental features of sexual reproduction were initiated with the last common eukaryotic ancestor, there have been ample pre-zygotic and post-zygotic variations among eukaryotes in terms of their reproductive lifecycles [9]. Within eukaryotic life, embryophytes (land plants) have a reproduction life cycle built around the alternation of generations. Alternations of generations, while being a sexual reproductive lifecycle, works differently than animal reproduction.

In animals, egg and sperm cells form by meiosis, whereas this occurs via mitosis in plants. Within alternation of generations, land plants have an alternation of multicellular diploid and haploid phases. The haploid phase arises from a spore, which in plants is the product of meiosis. Spores can either be the same size (homospory; figure 1.1) or two distinct sizes (heterospory). In heterosporous species, the smaller microspore germinates to form a multicellular male gametophyte, which produces a sperm cell through mitosis; the larger megaspore germi-
nates into a multicellular female gametophyte that produces an egg cell (Figure 1.2). Spores of homosporous species germinate and produce potentially bisexual gametophytes, able to bear both egg and sperm on the same individual. However, it is not uncommon for homosporous plants such as mosses to have separate sexes.

**Figure 1.1:** The homosporous life cycle in green plants.

Extant heterosporous plants consist of three lineages: heterosporous ferns, all seed plants, and heterosporous lycophytes. Most plant species are heterosporous angiosperms (flowering plants), whereas the most common homosporous species
are bryophytes, ferns, and lycophytes (the club mosses). All other land plants, including homosporous ferns and lycophytes, are homosporous. There have been at least 11 independent transitions to heterospory from the ancestral condition of homospory in vascular plants, but only three of these transitions are extant (Figure 1.2). The repeated evolution of heterospory represents convergence in vascular plant lineages [3]. Heterospory completely negates the possibility of gametophytic self-fertilization and “forces” mitotic (gametophytic) out-crossing in
land plants. This out-crossing has been proposed as the selective advantage behind heterospory [28]. Modern terrestrial vegetation is dominated by seed plants, and heterospory was an essential prerequisite to evolution of the seed [27]. The fundamentally different modes of reproduction makes the transition from homospory to heterospory a non-trivial one; the evolution of heterospory has been labeled as the most significant iterative innovation in the evolution of vascular plants [3].

Whereas heterospory has played a crucial role in the evolution of land plants, it is not the only mystery associated with the evolution of heterospory. All the lineages that have seen the independent evolution of heterospory have also coincided with a significant drop in chromosome numbers (figure 1.3). The association between spore type and chromosome number was first reported by Klekowski and Baker [17], who noted initially that on average, ferns have n = 57 chromosomes, while the mean angiosperm chromosome number is n = 13. A more recent meta-analysis of plant chromosome counts [16] substantiated the previous analysis, and the significant differences between heterosporous and homosporous plants remain, with means of 2n = 115 for homosporous plants and 2n = 27.24 for heterosporous plants (Figure 1.4). On average, homosporous plants have 4 times more chromosomes that heterosporous plants, and this difference holds even without the extreme examples of chromosome number, such as the homosporous fern Ophioglossum reticulatum with 2n = 1440 [15], higher than any other known eukaryote.
Earlier approaches to explaining the difference in chromosome numbers between heterosporous and homosporous plants focused on understanding why homosporous plants accumulate chromosomes faster [11]. The once-dominant theory was that homosporous plants primarily reproduce via gametophytic selfing, the fusion of gametes produced by mitosis from the same gametophyte (parent). Gametophytic selfing produces completely homozygous zygotes/offspring and would necessitate polyploidy-based redundancy to avoid genetic load, therefore leading to selection for larger genomes [12]. However, the tendency of polyploids to act as genetic diploids countered the once prominent gametophytic selfing hypothesis [10]. The rejection of the gametophytic selfing-based hypothesis presented by Klekowski [11] coincided with a shift from morphology and cytology based exploration to molecular-based studies of plant phylogenetics and evolution. As phylogenetic research of homosporous plants began to incorporate genomic methods, information from gene copy number patterns indicated that homosporous plants have had lower rates of paleopolyploidy than heterosporous plants despite having more significant chromosome numbers today [6]; [5]; [21]. Instead, compared to angiosperms, high chromosome numbers in homosporous plants seem to result from higher retention of chromosomes from the fewer rounds of polyploidy [2]; [20]. It seems that heterosporous plants go through higher rates of fractionation, i.e. the potentially permissive loss of duplicate gene and regulatory elements resulting from the relaxed effect of purifying selection on duplicated genomic elements. Higher retention/lower rates of fractionation, rather than increase in chromosome numbers, suggests that homosporous lineages are not outliers that
accumulate chromosomes faster than non-homosporous lineages. Instead, it sug-
gests that heterosporous lineages have perhaps undergone higher rates of pale-
opolyploidy and genome downsizing via reduction in chromosome numbers [2];
[6]; [18]; [19]; [34]; [4].
Figure 1.3: The land plant phylogeny and the multiple origins of heterospory.
Figure 1.4: The average number of chromosomes in homosporous and heterosporous plants [16].
Chapter 2. Research Objectives

Despite the crucial role played by heterospory in the evolution of terrestrial plant life, the mechanisms that underlie transitions to heterospory have remained unclear [16]. In addition, the correlation between heterospory and the drop in chromosome numbers remains unexplored. Why are changes in modes of reproduction and chromosome numbers associated? Are there parallels between the genetic factors such as expansion and contractions of Copy Number Variation (CNVs) in gene families, or selection on specific genes that underlie the transitions? Are the CNVs and selection in the gene families similar, opposite, or completely unrelated among plant species that arise from the nodes leading to heterospory? Convergent evolution in gene families within independently derived heterosporous lineages could provide evidence for the mechanisms by which heterospory evolved in land plants. Such changes may include expansions or contractions of gene families or changes in the rates of nucleotide substitution that reflect selection on specific gene family members. I hypothesize that the independent origins of heterospory involve selection and/or copy number variation on the same gene families/pathways that could suggest neofunctionalization i.e. when a gene copy takes on a completely new metabolic or biological role or subfunctionalization i.e. diversification of roles of gene copies. Models of nucleotide
substitution rates may detect trends in the selection of specific gene families that underlie transitions to heterospory, or other genomic changes undergoing similar selection rates could be behind the recurrent evolution of heterosporous plants. Associating changes in gene copy number or selection with these transitions will not explain the causation behind the transitions to heterospory; that is a task for gain-of-phenotype research. However, it will improve our capacity to circumscribe more specific hypotheses to test for potential causes behind the correlation between the transition to heterospory and a reduction in chromosome number. This strictly exploratory study aims to find potential gene families that affect both reproduction and meiosis (and by extension chromosome numbers), two traits that are directly linked by the reproductive life cycle wherein spores are produced by meiosis. Although this study is exploratory in nature, as more homosporous species become genetically transformable, the discovery of potential candidate gene families showing signs of selection on heterosporous nodes could provide novel insights through gene editing and functional validation.
Chapter 3. Research Questions

Working under the hypothesis that similar genomic factors be associated with both a reduction in chromosome number and the evolution of heterosporous lineages, I will explore the following questions:

- Have any gene families undergone a significant contraction or expansion in copy number on lineages leading to heterosporous clades? If so, is there anything about putative gene function that provides a clue to the association of heterospory and chromosome number?

- Have any gene families undergone distinct and parallel patterns of selection on lineages leading to heterosporous clades? If so, is there anything about putative gene function that provides a clue to the association of heterospory and chromosome number?

Exploration and comparison of genomic factors that underlie evolutionarily related heterosporous and homosporous lineages could be a source of novel phylogenetic insight into a possible association between the disparate traits of chromosome number and spore production.
Chapter 4. Methods

I focus on quantifying and qualifying genomic changes taking place along specific branches of phylogenies. The foundational data for this research comes from homology inference, i.e. the discovery of genes sets/families across species with shared ancestry. With in genomic homology, there are two major kinds of relationships: Orthology i.e. genes connected by a speciation event and paralogy, i.e. genes connected by a gene duplication events, mostly within species but possible across species [13]. The initial goal is to determine if any gene family (orthogroup i.e. an orthologous gene family) has undergone a significant expansion or contraction in copy number in the ancestors of each of the three heterosporous lineages. I will next use an analysis of selection to determine if and gene families have undergone an unusual amount of directional selection on these same three lineages. Results from either of these could help provide clues, in the form of possible gene function, as to why genome downsizing might be associated with heterospory. The quantification of the genomic changes required the reconstruction of copy number variation in a phylogenetic context and testing branches of interest on gene phylogenies for natural selection, calculated based on the ratio of non-synonymous substitutions against synonymous substitutions.
4.1 Sample Selection

Selection of taxa was guided and constrained by several factors. Pilot tests suggested that I could handle about 100 transcriptomes computationally for downstream analysis such as homology inference, gene family reconstruction in a phylogenetic context, heuristics based gene-tree inference and selection analysis. Because the goal was to explore unique changes on heterosporous lineages, I needed to ensure that each of the three extant heterosporous lineages were represented. This necessarily required reducing seed plants to a skeleton of samples, whereas heterosporous ferns and lycophytes were represented as much as possible because there are fewer species and fewer available transcriptomes. I also ensured that homosporous taxa were sampled extensively to provide phylogenetic context, including evenly throughout the homosporous lycophytes and ferns. In addition, I included outgroup samples from the three main bryophyte lineages. The complete list of transcriptomes used is supplemental materials.

4.2 Obtaining genomic information

The transcripts were obtained from three different sources: [25], [20], and [26]. The following table shows the distribution of species counts across species:

Automation was used when retrieving transcriptome files from their respective repositories to minimize human error. I accessed files that were outside of the onekp using manual steps.
Table 4.1: The distribution of the numbers of samples across their sources.

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of species from source</th>
</tr>
</thead>
<tbody>
<tr>
<td>[25]</td>
<td>102</td>
</tr>
<tr>
<td>[26]</td>
<td>4</td>
</tr>
<tr>
<td>[20]</td>
<td>5</td>
</tr>
</tbody>
</table>

4.3 Orthology inference

Given the goal of exploring similarities and differences across a list of taxon, homology inference was the preliminary step. Homology inference is the identification of genes with shared ancestry within and across species. This inference is primarily based on sequence similarity, using OrthoFinder [7] version 2.5.4. OrthoFinder was run with default configurations. A total of 30888 orthogroups were identified. Out of 2091844 genes, OrthoFinder assigned 2054931 (98.2%) genes to orthogroups against 36913 (1.2%) genes it could not assign to orthogroups. Only 613 (1.97%) orthogroups contained genes of all species involved against 15117 (48.5%) orthogroups containing two or fewer species.
See supplemental data for the complete details on the output of the orthofinder run.

4.4 Species tree inference

The species tree for the species list was generated using STAG and STRIDE. STAG generates an unrooted species tree that accounts for multi-copy gene families, and STRIDE can root the unrooted species tree. See supplemental material to see the final tree in newick format. The tree generated using this method, when illustrated with ggtree, looks as follows:

4.5 Gene family expansion and contraction inference

CAFE5 [22] was used to generate preliminary data on gene family expansions and contractions. CAFE5 first estimates a global rate of change of evolution ‘lambda’ and then uses that lambda in its implementation of maximum likelihood estimation to reconstruct the evolution of gene families throughout the phylogeny. CAFE5 was initially used with default configurations against all gene families generated by Orthofinder and the species tree generated by STAG and STRIDE. The exhaustive table of counts for all orthogroups failed to initialize with CAFE5’s inference model. With empirical testing, it was found that the issue was that the difference between the smallest count and the largest count for some of the gene families was too large for CAFE5’s statistical model. After a few rounds of testing, I found that, for this specific dataset, CAFE5 could not use any gene family where the difference between the smallest and the largest count
was over 68. For this dataset, gene families with differences between the counts greater than 68 distort the lambda or the global rate of change of evolution. The distortion to the global lambda makes it look like all the gene families are rapidly evolving and leads to mathematically illegible outcomes within the calculations that CAFE5 uses to reconstruct the numbers within gene families. Once families with differences larger than 68 were filtered out, CAFE5 could generate standard output.

See supplemental material to see the whole of CAFE5’s output.

4.6 Selection analysis

The following methods and steps were involved in the selection analysis of the data:

4.7 Multiple sequence alignment and codon alignment

The methods for selection analysis implemented within this required codon-based data for execution. Therefore, I converted the protein data within gene families into codon alignments. The peptide sequences within each gene family were first aligned using MAFFT [14]. I used pal2nal.pl[32] to convert the peptide sequence alignments generated by MAFFT, alongside their respective CDS sequences, into codon alignments.
4.7.1 Gene tree generation

I used IQ-tree2 [23] to generate gene phylogenies for the codon alignments of each gene family. The gene phylogenies were generated with the following configurations:

- A General Time reversible (GTR) DNA model [33].
- Empirical codon frequencies calculated from the data (which is the default for a GTR DNA model, as implemented in IQ-tree2).
- A proportion of sites were allowed invariably to account for rate heterogeneity across sites.
- IQtree2 was configured to use 100 non-parametric bootstrap replicates. I used nonparametric bootstrap to speed up tree inference.

Within IQtree-2 version 2.1.2, the above configurations can be re-created using the flags GTR+F+I -b 100.

4.7.2 Lineage specific selection analysis

Because my goal is to ask if there are significant genomic changes specifically on the heterosporous lineages, I used aBSREL[31] to test the branches/lineages in the gene phylogenies for selection. One of the many natural selection models implemented inside HyPhy, aBSREL is an ”improved” reiteration of branch-site selection models that are used to test for selection, not on specific sites but instead along a proportion of sites along branches.
Figure 4.1: An illustration of the binary, rooted, ultrametric tree used as the base species tree.
Chapter 5. Results

5.1 Copy number variation analysis

Focusing on the maximum-likelihood based data produced by CAFE5, I quantified copy number variation at all the nodes of the phylogeny. Of the 221 nodes on the phylogeny, three were the most important to us, each leading up to the 3 extant heterosporous lineages:

- Seed plants
- Heterosporous lycophyta
- Heterosporous ferns

5.2 Families with significant changes on each node

There were 9 gene families that significantly expanded on the node leading to seed plants, 6 to heterosporous lycophyta and 6 to heterosporous ferns. According to the results produced by CAFE, there are no families where significant change (expansion or contraction) took place in all three of our nodes of interest. However, several changes occurred on one or two of the three lineages.

The following venn-diagram summarizes the dynamic between the nodes further:
Table 5.1: Identity and most common associated function of the families that were expanding significantly on the seed plant node.

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Most common function of A. thaliana genes in the family</th>
</tr>
</thead>
<tbody>
<tr>
<td>OG0000053</td>
<td>F-box family gene most commonly used in leaf morphology determination</td>
</tr>
<tr>
<td>OG0000080</td>
<td>Mitochondrial transcription factor</td>
</tr>
<tr>
<td>OG0000107</td>
<td>Terpenoid superfamily protein</td>
</tr>
<tr>
<td>OG0000125</td>
<td>MATE efflux family protein</td>
</tr>
<tr>
<td>OG0000159</td>
<td>A member of the glycoprotein family</td>
</tr>
<tr>
<td>OG0000191</td>
<td>A member of the receptor-like protein kinase family</td>
</tr>
<tr>
<td>OG0000245</td>
<td>{Missing Arabidopsis genes}</td>
</tr>
<tr>
<td>OG0000422</td>
<td>Mixed Lineage Kinase Domain-like protein</td>
</tr>
<tr>
<td>OG0000619</td>
<td>Floral-property control gene with non-mendelian segregation.</td>
</tr>
</tbody>
</table>

According to the gene family reconstruction by CAFE5, I detected 19 gene families with significant expansion and contraction leading to the three heterosporous nodes. The node leading to seed plants had 9 families with significant changes. The Arabidopsis thaliana genes present within these gene families had the following functions:

There were 6 gene families that experienced significant change on the nodes leading to the heterosporous lycophyta. The Arabidopsis thaliana genes present within these gene families had the following functions:
Table 5.2: Identity and most common associated function of the families that were expanding significantly on the heterosporous lycophyta node.

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Most common function of A. thaliana genes in the family</th>
</tr>
</thead>
<tbody>
<tr>
<td>OG0000019</td>
<td>Beta glucosidase gene family.</td>
</tr>
<tr>
<td>OG0000026</td>
<td>A member of the protein kinase superfamily.</td>
</tr>
<tr>
<td>OG0000174</td>
<td>Encodes a protein with likely histone demethylation activity. Its mutation can lead to instability in morphological integrity.</td>
</tr>
<tr>
<td>OG0000563</td>
<td>A member of the lactone oxidase family.</td>
</tr>
<tr>
<td>OG0000619</td>
<td>Floral-property control gene with non-mendelian segregation.</td>
</tr>
<tr>
<td>OG0001580</td>
<td>A member of the Knotted1-like homeobox gene family.</td>
</tr>
</tbody>
</table>

Table 5.3: Identity and most common associated function of the families that were expanding significantly on the heterosporous lycophyta node.

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Most common function of A. thaliana genes in the family</th>
</tr>
</thead>
<tbody>
<tr>
<td>OG0000277</td>
<td>Adenine nucleotide alpha hydrolases-like superfamily protein.</td>
</tr>
<tr>
<td>OG0000861</td>
<td>A member of the PHO1 family. It is involved in inorganic phosphate transport and homeostasis.</td>
</tr>
<tr>
<td>OG0000895</td>
<td>A member of the Cysteine proteinases superfamily.</td>
</tr>
<tr>
<td>OG0000906</td>
<td>Ovate family protein.</td>
</tr>
<tr>
<td>OG0001580</td>
<td>A member of the Knotted1-like homeobox gene family.</td>
</tr>
<tr>
<td>OG0002379</td>
<td>DNA polymerase epsilon catalytic subunit.</td>
</tr>
</tbody>
</table>

There were 6 gene families that experienced significant change on the nodes leading to the heterosporous ferns. The Arabidopsis thaliana genes present within these gene families had the following functions:

There was 1 homologous family that saw a statistically significant expansion at both the nodes leading up to both seed plants and heterosporous lycophyta. The name assigned (arbitrarily) to this family was OG0000619. The Arabidopsis thaliana genes present in the homolog either are Glucose-methanol-
choline (GMC) oxidoreductase family proteins or hothead genes with observed patterns of non-Mendelian inheritance [29]. Whereas the causation behind the non-Mendelian inheritance of these Arabidopsis thaliana genes is unclear, non-Mendelian inheritance can trace its roots back to factors in meiosis [1], which are related to chromosome numbers. There was 1 homologous family that saw a statistically significant expansion at both the nodes leading up to both heterosporous lycophyta and heterosporous ferns. The name assigned (arbitrarily) to this family was OG00001580. The Arabidopsis thaliana genes present in the homolog were Class I knotted1-like homeobox gene family members (together with other KNAT genes). Homeobox genes are highly conserved across all of eukaryotic life and have functions in early morphological development. Homeobox mutants can lead to erratic changes (homeosis) in the anatomy [24].

5.3 Selection analysis data

There were 3115 homologs with at least 106 species present (out of the 111). From a conservative perspective (estimate?), those 3115 homologous families are viable for selection analysis without losing phylogenetic analogy to the exhaustive species list. The nature of the output data generated by the methods I used for natural selection analysis is too “noisy” for manual and visual analysis. Attached is an example of a “small” orthogroup turned into a gene phylogeny can look like:

Within the orthogroups that have been compiled for selection analysis do not so far reveal any signs of selection that favor the heterosporous nodes over the homosporous nodes.
Figure 5.1: Numbers of significant gene family copy number contractions or expansions in the three heterosporous clades.
Figure 5.2: An example of what the visualization of branch-site selection analysis looks like.
Chapter 6. Discussion

6.1 Copy number variation analysis

It has been over 50 years since Klekowski and Baker (1966) first pointed out the association between high chromosome numbers in homosporous plants. In a phylogenetic context, it now appears that chromosome number reduction has occurred on the lineages leading to heterospory. However, the scientific community still lacks a plausible explanation for this association of two seemingly unrelated traits. The research goal was to look for signatures of evolution that could provide hints about gene families that are playing a significant role in the evolutionary dynamic of the correlation between heterosporous spore production. Signatures of evolution from gene family expansion or contraction analysis reveal a limited number of parallel changes on heterosporous lineages. There were no gene families with significant changes leading to all three nodes. For pairwise intersections, there was one gene with significant changes leading to heterosporous lycophytes and heterosporous ferns and one gene family that had significant changes leading to heterosporous lycophytes and seed plants. There were no intersections between seed plants and heterosporous ferns. OG0001580 is the one gene family that was shared between heterosporous ferns and heterosporous lycophyta. The Arabidopsis thaliana genes present in this homologous family are members of the
homeobox gene family. The homeobox genes are known to have large effects in early morphological development and mutations can lead to some severe homeosis [8]. OG0000619 was the one gene family that was shared between seed plants and heterosporous lycophyta. The Arabidopsis thaliana genes present in this homologous family are known as Hothead genes responsible for floral structure and do not follow Mendelian segregation. Hothead genes have roles in determining floral morphology- something closely related to spore production, and Hothead genes in Arabidopsis thaliana have been observed to undergo non-Mendelian inheritance [29] which could trace its causes to meiosis. There is some history of research into the non-Mendelian nature of these specific genes archived in The Arabidopsis Information Resource (TAIR). Future work should explore these genes and the factors involved in more detail. Objectively, the genes of this gene family are homeobox genes and hothead genes only in Arabidopsis thaliana, and could have completely different molecular and biological roles in other species. Albeit the significant expansion of genes on two lineages leading to heterosporous warrant further investigation into their biological and molecular function across species.

### 6.2 Selection analysis

The study of trait evolution across families of green plants using natural selection data in a phylogenetic context requires access to more computational power to generate a greater density of data than is currently available. This should not be an issue given the steady progress in data generation being made by the computational systems this project has access to.
6.3 Limitations of the study

It is possible for the gene family reconstruction or the associated functions of the gene families by CAFE5 to be spurious and entirely random. I used CAFE5 to generate data for thousands of families on a large species tree with a deep phylogeny. However, this sample set was pushing the limits of the algorithms implemented within CAFE5. Based on direct communications with the authors of CAFE5, apparently, the tree was “too large and too deep”. For the reconstruction of gene families, for the 9461 families that were successfully reconstructed, 40 of the largest families could not be compiled. Those 40 homologous families accounted for some tens of thousands of genes, and while there is no promise of good data, it was a non-negligible portion of the total data that simply could not be processed. The difficulty CAFE5 had in reconstructing gene families, given the size and the depth of the phylogeny, reduces confidence in the output and warrants some further investigation with some alternative approaches.

Some solutions to the size and depth issues that can simultaneously provide some hints about spuriousness in the results:

- Compromise the generous phylogenetic background data and use a smaller and shallower species tree. The tree can be made shallower by giving up outgroups but it is currently unclear how effective that can be.

- Use multiple lists of taxa with analogous phylogenetic outlines and compare the outcomes between each other. The comparisons between outputs neither
exactly provide specificity nor exactly provide specificity, but if there is a
difference at all, then it brings the status of the reconstruction into question.

- Use parsimony based approaches for gene family reconstruction as opposed
to maximum likelihood estimation altogether. Our implementation of MLE
in this project was thanks to the robust history of MLE’s capacity to gen-
erate data on smaller datasets.

Depending on quality, the transcriptome of a species can be an incomplete
representation of the genes present in that species. One metric of transcriptome
quality is BUSCO [30], which is a method to gauge the quality of a genome
using single copy orthologous genes, where the completeness of a genome can
be gauged by comparison against a standardized database. BUSCO produces a
percentage-based quality score that reflects the completeness of any specific sam-
ple of sequencing data. The percentage scores reflect the completeness of the
sample sequencing data. While the taxon list was fairly exhaustive and represen-
tative of the phylogenetic context, the BUSCO scores for the samples ranged from
the low 70s to the high 90s in terms of percentage. It is possible that, within that
range of transcriptome quality scores across the list of taxon, some of the gene
families as found by my runs of homology inference, do not reflect copy number
variation in the gene families with complete accuracy. The range of quality of the
transcriptomes has a potential to affect CNV and by extension the reconstruction
families across the phylogeny. We can only gauge the actual size of the impact
from the range of BUSCO scores with a completely different list of samples. A
“simple” solution to this is to use taxa where all species have BUSCO scores in
the high 90s. I built the research in this project around exploring a potential relationship between spore production and chromosome numbers, but these two traits could be unrelated. A way to gauge the existence of a potential relationship between spore production and chromosome numbers would be to expand the current experiment design to include heterosporous nodes alongside other nodes where there has been a marked decrease in chromosome numbers without the evolution of heterospory. The study of homology from the perspective of gene families/orthogroups provides an objective avenue to gauge the similarities and differences between species, families, lineages and nodes. This comes back to the fact that gene families can be identified, labeled/tagged and easily parsed for insights. While objective and easy to implement, looking at phylogenetics with well-set gene families misses the possibility of catching similarities in functions shared across or between two or more than two different gene families. In other words, it is easy to parse information tagged using gene families, but parsing information by the function of gene families across different gene families requires a different approach, which so far does not have a standardized approach. While a standardized approach is lacking, manual parsing of the results in this project is possible, given the small number of gene families to be investigated. As such, a manual investigation of the significantly expanding gene families will be a part of the expansion of this project. The research objectives, built around the hypothesis that the evolution of correlated traits in heterosporous lineages probably share a similar history of evolution, were to search for similarity in signatures of evolution. Within this sample set and methods, while significantly smaller than
initially expected, there were two gene families that warrant further investigations into their potential role in floral morphology/spore production and related roles in meiosis. When I started this project, there were no phylogenetic and genomic insights on the evolution of heterospory and its correlation with drops in chromosome numbers from the perspective of gene families. The exploratory nature of this study means that the insights collected here require further investigation with the use of functional genomics. Nevertheless, this project has provided some novel insights that make it possible to approach functional genomics at all.
References


