Summer 2016

# Resolving Gnetum Evolutionary History

Angela McFadden
*Central Washington University*, mcfaddena@cwu.edu

# RESOLVING *GNETUM* EVOLUTIONARY HISTORY

---

A Thesis

Presented to

The Graduate Faculty

Central Washington University

---

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Biology

---

by

Angela McFadden

July 2016

CENTRAL WASHINGTON UNIVERSITY

Graduate Studies

We hereby approve the thesis of

Angela McFadden

Candidate for the degree of Master of Science

APPROVED FOR THE GRADUATE FACULTY

_____ _____
Dr. Linda Raubeson, Committee Chair

_____ _____
Dr. Jim Johnson

_____ _____
Dr. Jennifer Dechaine

_____ _____
Dean of Graduate Studies

ABSTRACT

RESOLVING *GNETUM* EVOLUTIONARY HISTORY

by

Angela McFadden

2016

*Gnetum* are non-flowering seed plants of the tropics, indigenous to South America, Africa, and Asia. This group of about 40 species is fascinating to botanists because it shares distinctive morphological characteristics with flowering plants, such as broad leaves, woody stems, and flower-like strobili. There are still questions surrounding the relationships within the genus of *Gnetum*. With that in mind, I focused my work on generating phylogenetic hypotheses, using two molecular data sets: a concatenation of over 60 different chloroplast genes (66,815 base pairs), and the whole chloroplast genome (128,772 base pairs). This allowed me to compare the two phylogenies and assess whether adding non-coding regions increase phylogenetic resolution. Statistical tests determined that the data were sufficient to answer questions about deep splits, and to resolve the branches within the genus. I used each of the two data sets to infer Maximum Parsimony and Maximum Likelihood phylogenetic hypotheses for 18 species of *Gnetum*. Confidence levels for most nodes were very high, and trees show clades consistent with biogeography. My bootstrap results suggest that the South American clade may not be the earliest diverging lineage, although statistical tests support the South American clade at the base of the *Gnetum* tree.

# ACKNOWLEDGEMENTS

I would like to thank the Gymnosperm Tree of Life and NSF (DEB 0629607) for funding. Thank you Rich Cronn for preparing the first batch of libraries and Sarah Mathews for providing plant materials and funding the second round of sequencing for this project. Thank you to Jeffry Mower and Wenhu Guo for providing the mtDNA draft sequence, and Volker Knoop and Monika Polsakiewicz for preparing the DNA for the mitochondrial reference.

I would also like to thank my committee, Dr. Jim Johnson and Dr. Jenny Dechaine, but especially my chair, Dr. Linda Raubeson, who not only gave me helpful feedback and edits, but worked with me to design and finish this project. She dealt with my ups and downs throughout this process with patience and grace. You have been a wonderful advisor and friend.

I would like to thank my friends, even though I can't mention everyone, there are a few special people I want to thank: Katie, Stefani, and Sara, thank you for being so supportive all the time, you are wonderful friends who made me believe I could handle this. Maria, thank you for forcing me get help when I needed it. Chris and Ayla, thanks for letting me couch surf for months while I finished up! Dave, thank you for reading and re-reading my manuscript, and listening to me practice my talks, you have been so supportive and motivating the last 2 years, and I can't thank you enough.

Last, but certainly not least, to my family: You have all been so wonderfully supportive, but I would especially like to thank my parents, Tammy and Calvin, and my big brother Joe. You can't imagine how much your help has meant to me. Love you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

This research investigates the phylogenetic relationships of the species of the genus *Gnetum. Gnetum* is a genus of gymnosperms, non-flowering seed plants, interesting because of its angiosperm-like morphology, and suspected Gondwanan vicariance (where the modern distribution of the genus is due to the breakup of the supercontinent of Gondwana). I used chloroplast genome sequences and chloroplast protein coding genes to generate phylogenetic hypotheses using Maximum Likelihood and Maximum Parsimony approaches. I have compared my results to two earlier studies of *Gnetum,* each based on much less DNA data.

## *PHYLOGENETICS AND ITS VALUE*

Phylogenetics is the inference of the evolutionary history and relationships of groups of organisms. This inference can be accomplished in a few different ways. Comparing shared morphological and physiological characteristics has been a standard phylogenetic methodology since before the term "phylogenetics" was used to describe the process. The comparison of morphological characteristics to determine evolutionary relationships is, often, the only way to conduct phylogenetic studies in relation to fossils, which are often devoid of DNA. Today, baring work with fossils, phylogenetic analyses conducted using molecular data is more common. This is, in part, due to the large amount of data that can be gained from DNA. Thousands of characters can be analyzed for common ancestry, which can lead to powerfully significant results.

It has been well established that having a solid phylogenetic framework is important for many biological fields (Soltis & Soltis, 2003). These fields include, but are not limited to ecology, evolutionary biology, and bioprospecting. Phylogenetics contributes to ecology mainly in terms of conservation. With the rapid extinction of species in recent years, a focus on maintaining ecological diversity has increased. Saving all extant species is improbable, but by having a complete phylogenetic framework, we may be able to save enough species from each clade or taxonomic group that a genus may retain enough diversity to survive and, perhaps, produce future speciation (Vane-Wright et al., 1991).

Phylogenetics contributes to evolutionary biology by helping us to understand the connections between living species and their ancestors. By understanding how genetic changes formed extant species, we can make predictions about how species may evolve in the future. In addition, knowing the evolution of a group or genus can lead to discovery of shared characteristics, including nitrogen-fixing symbioses, chemical defense mechanisms, and other complex pathways (Soltis & Soltis, 2003).

Bioprospecting is another common use for phylogeny. A well-resolved phylogeny guides workers looking for bioactive compounds useful for drugs. For example, taxol, a chemical compound extracted from *Taxus brevifolia* by Wani et al. (1971), was found to be an effective cancer-fighting agent. As only small amounts could be obtained from the thick bark of *Taxus brevifolia*, large-scale production of this drug was hindered (Kingston, 2016). Fortunately, in 1982, Chauvière et al., determined that the species most

closely related to *T. brevifolia*, the more common *T. baccata*, has a precursor to the taxol compound, and mass production of the, still popular, drug began.

### *USING CHLOROPLAST DNA AS A PHYLOGENETIC MARKER*

Chloroplast DNA (cpDNA) is commonly used for plant systematics, because: 1) it is the easiest genome with which to work; 2) it is present in high copy number in the leaf tissue; 3) the gene order within cpDNA is highly conserved in land plants (Palmer et al., 1988) and so PCR and alignment of the resulting sequences is straight forward; 4) thousands of published cpDNA sequences are available for comparison; and, 5) the genome is small enough to be tractable, yet large enough to contain many variable characters.

The cpDNA is isolated from the chloroplast, the organelle within plant cells that performs photosynthesis. The chloroplast was once a free-living bacterium that was engulfed by an ancestor of the modern plant cell. Because of its bacterial ancestry, cpDNA is circular and separate from the nuclear genome of the cell. Most chloroplast genomes are comprised of four sections: the large single copy region (LSC), the small single copy region (SSC) and two sections that are separated by the SSC and LSC, the inverted repeats (IR) (Raubeson et al., 2005). The IRs of chloroplast genomes range in size from 10 kilobase (kb), as in *Osmunda cinnamomea*, to 25 kb, as in most angiosperms (Palmer et al., 1986), and were found to be approximately 19kb in *Gnetum* (Hsu et al., 2015, Mao et al., 2015).

The chloroplast genome, on average, is between 120 and 160 kb (Raubeson et al., 2005), but the cpDNA of Gnetophytes (*Welwitschia*, *Ephedra*, and *Gnetum*) is even

smaller and more compact than other land plants (McCoy et al., 2008; Wu et al., 2009). *Welwitschia* has the largest cpDNA of the Gnetophytes, with a total of ~118,000-119,000 (McCoy et al., 2008; Wu et al., 2009). The smallest of the Gnetophyte chloroplasts seem to be found in *Ephedra*. Wu et al. (2009) published the complete cpDNA of *Ephedra equiseta* with a total of 109,518 bp. In the same study, *Gnetum parvifolium* was found to be 114, 914 bp. Other complete *Gnetum* plastids (including *G. ula, G, montanum, G. gnemon*, and *G. parvifolium*) have revealed similarly small plastids, with sizes ranging from 113,249-115,022 bp (Hsu et al., 2015; Mao et al., 2015; Zhu et al., 2015; Wu et al., 2009).

*INTRODUCTION TO GNETUM*

*Gnetum* are indigenous to South America, Africa, and Asia. The vegetative morphology of *Gnetum* are characterized by its broad leaves and woody stems. Most of the species are lianas, though some take the form of trees or shrubs. All 41 recognized species (WCSP, 2011) are dioecious and, along with the rest of the Gnetophytes (*Welwitschia* and *Ephedra*) undergo double fertilization, making Gnetophytes unique among non-flowering seed plants (Uno et al., 2001; Friedman & Carmichael, 1996).

In West and Central Africa, these plants, specifically *Gnetum africanum* and *G. buchholzianum*, play a large role in the economy. Individuals harvest the edible leaves to process, dry, and sell at markets. A socio-economic study found that in Cameroon and Nigeria especially, the economy was essentially sustained by the sale of these plants (Fuashi et al., 2010). In the same study, they described harvesting techniques. As with many of the *Gnetum* species, *G. africanum* and *G. buchholzianum* are lianas, or climbing

vines. In order to harvest the edible leaves, people will either pull or cut down the vine, sometimes going so far as cutting down the tree that the vine occupies. These harvesting methods have led to concerns about the conservation of species in the wild. To help relieve pressure on the wild populations, researchers are encouraging cultivation of *G. africanum* and *G. buchholzianum* (Ndam et al., 2001).

Other studies focus on the uses of *Gnetum* as a source of medical treatment. One study by Tan et al. (2013) found that *Gnetum gnemon*, a wide-spread Asian species, long used as a source of food, also has medicinal uses. Extracts were found to contain anti-quorum sensing (QS) properties. Quorum sensing is a bacterial form of communication which aids in colony formation and conjugation. The ability to disrupt this process means that studying the chemical components of these, and other similar specimens, could be vital in the fight against antibiotic resistant strains of bacteria. Species most closely related to *G. gnemon* should be investigated for similar anti-microbial properties. Phylogenetic studies, like mine, help to resolve those species relationships.

### *THE PLACEMENT OF GNETUM AMONG OTHER SEED PLANTS*

*Gnetum* belongs to the Gnetophytes, a monophyletic group of non-flowering seed plants that includes only three genera – *Ephedra*, *Welwitschia* and *Gnetum*. Although occasional debates about the naturalness of the group have arisen, most data supports the monophyly of the three genera .The morphological similarities (especially in the cones and ovules) between the three genera were first detailed by Hooker (1863). Later, cladistic studies based on morphology have supported monophyly (Doyle, 1996; Friedman & Carmichael, 1996; Price, 1996) as have studies using molecular data (Chase

et al., 1993; Price, 1996, and many others). In most of these studies, *Gnetum* and *Welwitschia* have been found to be sister taxa, with *Ephedra* as a basal clade. Thus, *Welwitschia* is the most appropriate outgroup taxon for a *Gnetum* phylogenetic analysis.

The placement of the Gnetophytes within the seed plants is contentious. Although I will not be addressing this question in my research, it is worth discussing here as it does influence some phylogenetic work within *Gnetum*. There are four major hypotheses of gnetalean relationships relative to the other extant seed plants: Anthophyte, Gnetales-Sister, Gnetifer, and Gnepine (Figure 1). The Anthophyte hypothesis states that Gnetophytes are a sister group to Angiosperms. Gnetales-Sister refers to a relationship where the Gnetophytes are a sister to the four other extant seed plant lines. Gnetifer refers to Gnetophytes and conifers being sisters, and Gnepine refers to Gnetophytes and the conifer family Pinaceae (separate from other conifers) as sisters.

In his description of the genus *Gnetum*, Hooker (1863) pointed out the similarity in appearance of the strobilus of *Gnetum* to flowers of some angiosperms. Scalariform pitting was also found in the vessels of both Angiosperms and species of *Gnetum* (Muhamad & Statler, 1982). Furthermore, *Ephedra* and *Gnetum* share the specialized reproductive mechanism, double fertilization, with flowering plants, although there are slight differences in the process. As double fertilization is only seen in Angiosperms and Gnetophytes, it could be seen as a homologous trait (Friedman & Carmichael, 1996). This in combination with vasculature of the leaves and the fleshy ("berry-like") seeds, made botanists wonder whether there could be the phylogenetic link between

Gnetophytes and angiosperms. The idea of a *Gnetum*/Angiosperm sister relationship became known as the Anthophyte hypothesis.

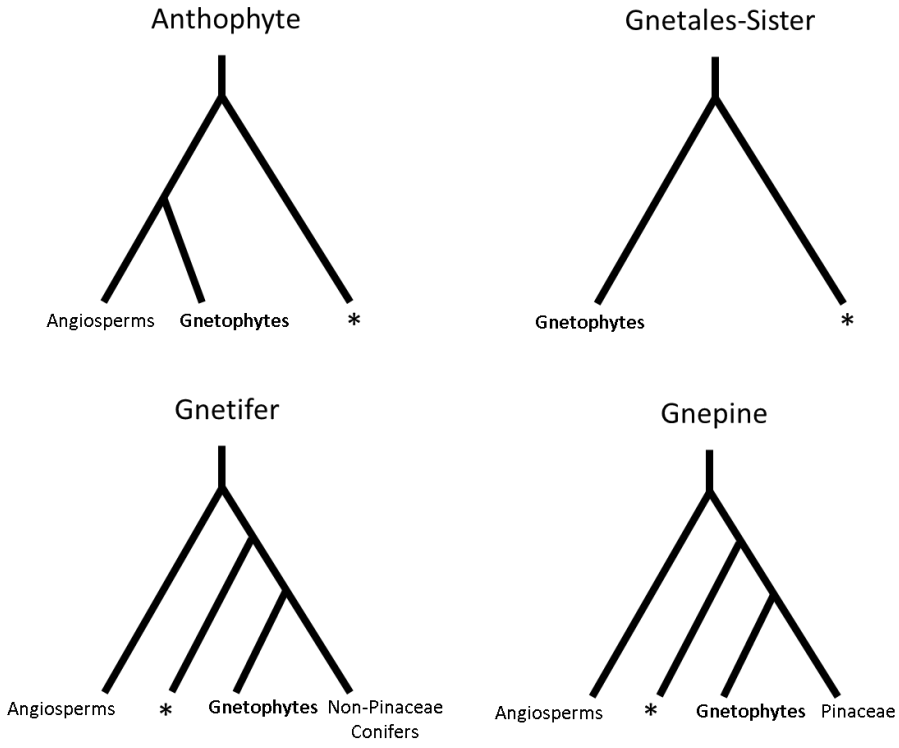



Figure 1 Gnetophytes' placement in relation to other seed plants. The symbol (*) indicates all other seed plant genera not specifically mentioned.

There are morphological traits that explicitly contradict the Anthophyte hypothesis, however. For example, *Gnetum* species have circular pitting, similar to the structure found in conifer tracheids, suggesting a Gnetifer or Gne-Pine relationship. Later, molecular studies were conducted, and those, in the main, did not support the Anthophyte hypothesis (see review in Magallón & Sanderson, 2002 and Burleigh & Mathews, 2004).

Multiple molecular studies using various methods have lent support to one of the latter three hypotheses. Early molecular studies commonly supported Gnetales-Sister, whereas later studies were more likely to support the Gnetifer and Gnepine relationships. To identify which problems may have contributed to this confusion (*e.g.* disparate sampling sizes, rates of nucleotide changes across sites, choice of criterion) and address them, Burleigh and Mathews (2004) examined molecular data from multiple sources. Their analyses supported the Gnepine hypothesis. Though this was their best-supported topology, they state that the rare genomic changes that support a Gnetifer relationship should be further explored. Though the knowledge of Gnetophytes' relationships to other seed plants does not pertain directly to my thesis, the data from this study will contribute to work that is addressing this issue, through the Gymnosperm Tree of Life (GToL) project.

### *PREVIOUS WORK ON THE RELATIONSHIPS WITHIN GNETUM*

Markgraf's 1930 monograph of *Gnetum* is the current standard for the classification. In this monograph, he described his belief that the lianas evolved from the arborescent species. He described two sections of the genus: Section *Gnetum* and Section *Cylindrostachys* Markgr. The sections do not solely follow geographic distribution, but were based on a combination of geographic distribution and morphological similarities, such as the color of the leaves when dried and the shape of male spikes, between taxa (Table 1). Section *Gnetum* contains the African and S. American species, but also the arborescent species, which are found in Asia.

Table 1 Classification of *Gnetum* Species and Their Inclusion in Phylogenetic Studies

| **Markgraf, 1929** | **Geographical Distribution** | **Won & Renner, 2006** | **Hou et al., 2015** | **This Study** |
|---|---|---|---|---|
| **Section *Gnemonomorphi* Markgr. (*Gnetum* Sect. *Gnetum*)** | | | | |
| **Subsection *Eugnemones* Markgr.** | Asia | | | |
| **1. *G. gnemon* L.** | | 1. Yes | 1. Yes | 1. Yes |
| **2. *G. costatum* K. Sch.** | | 2. Yes | 2. Yes | |
| **Subsection *Micrognemones* Markgr.** | Africa | | | |
| **3. *G. africanum*Welw.** | | 3. Yes | 3. Yes | 3. Yes |
| **4. *G. buchholzianum* Engl.** | | | 4. Yes | 4. Yes |
| **Subsection *Araeognemones* Markgr.** | S. America | | | |
| **5. *G. camporum* Makgr.** | | | 5. Yes | |
| **6. *G. nodiflorum* Brongn.** | | 6. Yes | 6. Yes | 6. Yes |
| **7. *G. leyboldii* Tul.** | | 7. Yes* | 7. Yes | 7. Yes |
| **8. *G. paniculatum* Spruce ex Bentham** | | 8. Yes | 8. Yes* | |
| **9. *G. schwackeanum* Taub.** | | 9. Yes | 9. Yes | |
| **10. *G. urens* (Aubl.) Bl.** | | 10. Yes | 10. Yes | 10. Yes |
| **11. *G. venosum* Spruce ex Bentham** | | | | |
| ***Gnetum* Section *Cylindrostachys* Markgr.** | | | | |
| **Subsection *Stipitati* Markgr.** | Asia | | | |
| **12. *G. arobreum* Foxw.** | | | | |
| **13. *G. contractum* Markgr** | | | | |
| **14. *G. gracilipes* Cheng** | | | | |
| **15. *G. latifolium* Bl.** | | 15. Yes | 15. Yes | 15. Yes |
| **16. *G. montanum* Margr.** | | | 16. Yes | 16. Yes |
| **17. *G. oblongum* Markgr.** | | | | |
| **18. *G. pendulum* Cheng** | | | | 18. Yes |
| **19. *G. tenuifolium* Ridl.** | | 19. Yes | 19. Yes | |
| **20. *G. ula* Brongn.** | | 20. Yes | 20. Yes* | 20. Yes |
| **Subsection *Sessiles* Markgr.** | Asia | | | |
| **21. *G. acutum* Markgr.** | | 21. Yes | | 21. Yes |
| **22. *G. bosavicum* Markgr.** | | | | |
| **23. *G. cuspidatum* Bl.** | | 23. Yes | 23. Yes | |
| **24. *G. diminutum* Markgr.** | | 24. Yes | 24. Yes | 24. Yes |
| **25. *G. gnemonoides* Brongn.** | | 25. Yes | 25. Yes | 25. Yes |
| **26. *G. cleistostachyum* Cheng** | | | | |
| **27. *G. globosum* Markgr.** | | | | |
| **28. *G. hainanense* Cheng** | | 28. Yes | 28. Yes | 28. Yes |
| ***29. G. klossii* Merrill** | | 29. Yes | 29. Yes | 29. Yes |
| ***30. G. leptostachyum* Bl** | | | 30. Yes | |
| ***31. G. lofuense* Cheng** | | | 31. Yes | |
| **32. *G. loerzingii* Markgr.** | | | | |
| **33. *G. macrostachyum* Hooker** | | 33. Yes | | |
| **34. *G. microcarpum* Bl.** | | 34. Yes | 34. Yes | |
| **35. *G. neglectum* Bl.** | | 35. Yes | 35. Yes | |

| | | | |
|---|---|---|---|
| **36. *G. oxycarpum* Ridl.** | | | |
| **37. *G. parvifolium* Cheng ex Chun** | 37. Yes | 37. Yes | 37. Yes* |
| **38. *G. raya* Markgr.** | 38. Yes | 38. Yes | 38. Yes |
| **39. *G. ridleyi* Gamble** | | | |

**This table contains the classification and geographical distributions of *Gnetum*, in the phylogenetic framework established by Markgraf (1929), as seen in Price (1996).**
**A number followed by a "Yes" gives reference to which taxa (as listed in the Classification column) are used in the three topologies of *Gnetum* discussed in this manuscript, those being: Won and Renner (2006), Hou et al. (2015) and my study.**
**An asterisk (*) denotes that a species was included in the phylogeny under an accepted synonym of the species.**

Won and Renner (2006) published a molecular phylogeny for the species of *Gnetum* (Figure 2); however, the main focus of this study was not the phylogenetic resolution of the genus. Their goal was to apply current dating techniques to a genus that (at the time) had no fossil record (possible *Gnetum* fossils have since been recovered). In order to apply a molecular clock analysis, they needed a topology within the genus, but their calibration fossils were outside the genus. Their *Gnetum* phylogeny, therefore, was inferred from their analysis of two different data sets.

For the first analysis, Won and Renner (2006) wanted to include *Welwitschia*, *Ephedra*, and other seed plants as outgroups, but due to an inability to align all these sequences across the Gnetophytes, they performed outgroup-based analyses using only two (*rbcL* and *matK*) chloroplast genes, which were conservative enough to meet the requirements for the alignment, but were too uninformative to resolve the branches within *Gnetum.* To resolve those braches, they conducted a second analysis, containing only *Gnetum* species, inferred from six loci: three chloroplast loci, two nuclear loci, and one mitochondrial. Based on the topology of their first two-gene phylogeny, Won and

Renner (2006) rooted their main *Gnetum* species tree (constructed with a concatenation of the six loci stated above), using the S. American taxa as a functional outgroup.

Won and Renner applied a rate analysis to determine the ages of the nodes in their topology. Based on their dating results, Won and Renner's data supported a divergence time that suggested *Gnetum's* radiation was not Gondwanan, and thus the species underwent two distinct intercontinental dispersal events in the Eocene and Oligocene epochs; one dispersal event occurring from South America to Africa, and a second event from Africa to Asia. This work raises interesting points about the distribution of *Gnetum*, but it also highlights the need for further study.

Hou et al. (2015) built on the work of Won and Renner (2006) by including taxa that were not present in the Won and Renner data set, such as the second African taxa *G. buchholzianum*, and adding multiple accessions of many species to address delimitation, which was missing in Won and Renner's 2006 publication. Hou et al. also indicated that the lack of support for some of the nodes in Won and Renner's study, particularly in regards to the deep divergences, supported the need for additional study. With the above issues as a basis for further work, Hou et al. inferred a phylogeny using a total of 27 species, and implemented their own dating study.

Hou et al. (2015) used five genetic markers, a mixture of three nuclear genes (18S, 26S, and nrITS) and two chloroplast genes (*rbcL* and *matK*), which are two of the plastid genes used by Won and Renner (2006), giving them a total of 7,605 base pairs (bp). The resulting phylogeny showed three major clades, grouped by continent (S. America, Asia, and Africa), and the same basal split between the South American clade

and the rest of the genus (S. American-basal hypothesis) that was seen in Won and Renner (2006), with the African and Asian clades being sisters.

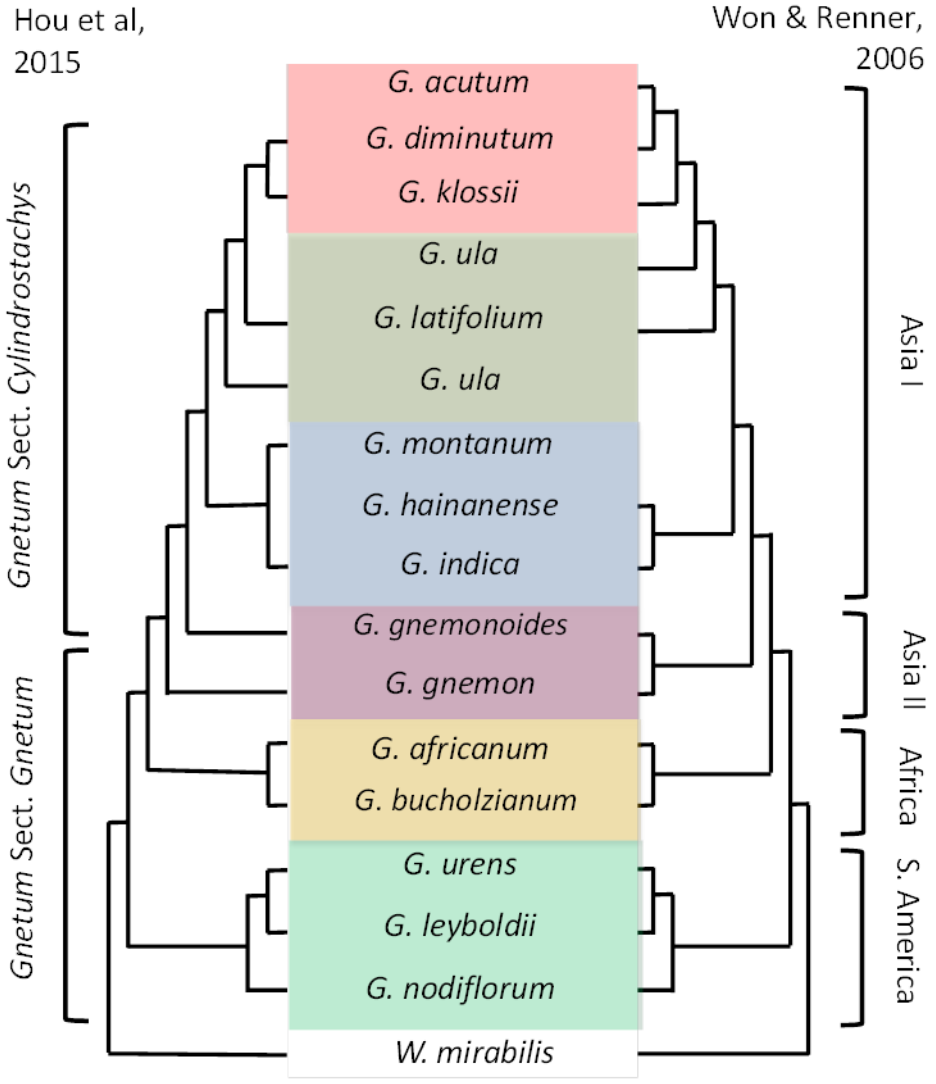



Figure 2 Topologies from previous *Gnetum* studies (Hou et al., 2015, Won & Renner, 2006); branch lengths are not drawn to scale. Sections are labeled according to categories determined by the original researchers. Colors are a reference to the clades found in the topologies of this study and are consistent throughout. Only those species included in the current study are shown.

The major difference seen in the topology of Hou et al. (2015) was in *Gnetum* Sect. *Cylindrostachys*. Hou et al. (2015) supported a monophyly of *Gnetum* Sect. *Cylindrostachys*, whereas Won and Renner (2006) did not. This difference can be noted in Figure 2, where *G. gnemonoides* is as the sister to *G. gnemon* in Won and Renner's (2006) topology, but *G. gnemonoides* is sister the all the other members of section *Gnetum* Sect. *Cylindrostachys* in the Hou et al. (2015) topology.

The dating study of Hou et al. (2015) returned very different results than the Won and Renner analysis. With the help of two possible *Gnetum* fossils described in 2009 and 2010 (Guo et al., 2009, Rydin & Friis 2010), Hou et al. (2015) hypothesized an earlier divergence, in the Cenozoic era. With the earliest divergence occurring in the Cenozoic, Gondwanan vicariance cannot be ruled out as a possible explanation for the current distribution of *Gnetum*.

*JUSTIFICATION*

When I began this work, the only phylogeny of the genus *Gnetum* was inferred by Won and Renner (2006). Due to their interest in dating techniques rather than phylogenetic resolution, the researchers did not fully explore their phylogenetic hypothesis in their 2006 study. The genus level phylogeny of *Gnetum* was based on 8,957 bp from six loci, two of which are the highly conserved, and therefore less informative, *rbcL* and *matK*. In addition, only two of these loci were used to test their basal clade. This two gene data set supported the S. American taxa as the basal clade; this clade was then used to root the species level topology. Also, Won and Renner did not perform a Maximum Likelihood (ML) bootstrap analysis nor did they perform any additional

statistical analyses, such as a Shimodara-Hasawaga (SH) or Shimodara Approximately Unbiased Test (AU). The results of the dating analysis support *Gnetum's* earliest divergence occurring relatively recently, in the Oligocene epoch or later. This would mean that two distinct long distance dispersal events over salt water must have occurred. This result was in contrast to the common thinking that *Gnetum* was one of a limited number of examples for Gondwanan vicariance. The limited data set, lack of statistical testing of their basal clade, and Oligocene divergence timeline all justified further research into the genus *Gnetum.*

As I was working on my project, Hou et al. (2015) improved upon the work of Won and Renner by adding additional taxa and multiple representatives of each species. Only the resolution of the Asian species (the shortest branches of the phylogeny) benefited from the inclusion of multiple accessions. Though Hou et al. (2015) improved upon the number of species and accessions, they inferred their phylogeny from slightly fewer nucleotides, and also lacked ML bootstrapping and additional statistical tests. The resulting topology (Figure 2) shows some differences from that of Won and Renner, specifically in relation to which species are most closely related to *G. gnemon*, which is of special interest due to its anti-microbial compound. In Hou et al., *G. gnemonoides* is not even within the same monophyletic group as *G. gnemon*, although the two are reported as being part of the same larger clade, referred to as Asia I, by Won and Renner.

The Hou et al., dating analysis returned a very different result from that of Won and Renner, 2006. Hou et al. estimated a Cretaceous basal divergence, approximately 50 million years earlier than the Won and Renner Oligocene estimate. If the Cretaceous

divergence is correct, then oceanic long distance dispersal would not have been required; Gondwanan vicariance could be a possibility for *Gnetum*. In addition to needing better statistical testing, differences in the results of the two studies, led to a need for a larger data set to resolve the issue.

The results of my research should further our phylogenetic understanding of *Gnetum*. Both of the prior studies described used fewer than 10,000 bp to estimate their phylogenies, I have generated a data set with over 100,000 bp. I have used this data set to create a well-supported phylogeny, with multiple species from each of the major clades, and well supported P values when additional statistical tests were applied to the major clades. My robust data set not only helped to add substantive statistical support for the various branches of *Gnetum* clades, but will further refine the distribution timeline once a rate analysis is applied.

# CHAPTER II

# METHODS

## *DNA EXTRACTION*

Members of the Raubeson Lab and Gymnosperm Tree of Life group had obtained raw sequence data for nine species of *Gnetum* before I started work on the *Gnetum* phylogeny. I worked to add additional species to the study. I obtained aliquots of isolated DNAs and 74 dried herbarium samples from Sarah Mathews at CSIRO, Canberra, Australia. After sorting out samples for which I already had libraries (Table 2), I performed a total of 22 total genomic DNA extractions from the dried tissue. The herbarium samples were a challenge to work with, but fresh material could not be obtained. Degradation was one of the issues faced. Some of the DNA was already degraded to the point where the bulk of the fragments were 100 to 200 bp long, or smaller, and therefore, not ideal for library construction. On any samples with visible herbarium glue, the glue was removed from the leaves before beginning the extraction.

My extraction process consisted of adding 0.3 g of dried plant leaf tissue to a microcentrifuge tube. Liquid nitrogen was added to the tube, and plastic pestles were Liquid nitrogen was added to the tube, and plastic pestles were used to crush the plant tissue into a powder. From this point, the extraction occurred as per the Nucleon PhytoPure DNA Extraction Kit instructions (GE Healthcare UK Limited, Buckinghamshire, UK). As the extractions seldom yielded a precipitate that would have been easily removed from the surrounding liquid, additional steps, which included a

centrifuge and washing step, were performed on the extractions before being re-suspended in a buffer solution.

Once I completed the extraction process, 2 µl of DNA was run on an agarose gel to determine the fragment sizes, and the Qubit fluorometer 2.0 (Invitrogen, Carlsbad, CA, USA) was used per manufacturer's instructions to measure the concentration of each sample, a total of 114 in all. I conducted these tests on the 22 extractions, and also on the 92 *Gnetum* DNA aliquots provided by Sarah Mathews. Seventeen of those 114 samples had concentrations high enough to fit the input requirements (1 µg of DNA in 50 µl) continued to the TruSeq library construction steps.

Table 2. List of Species and the Source of Materials for This Study.

| **Species** | **Extraction** | **Library Preparation** | **Source** |
|---|---|---|---|
| ***G. acutum* Markgr.** | B | A | K.M. Wong 599 12 |
| ***G. africanum* Welw.** | A | A | S. Lisowski 47293 |
| ***G. buchholzianum* Engl.** | A | A | 749 K 0006076 |
| ***G. diminutum* Markgr.** | B | A | SM 13-100 |
| ***G. gnemon* L (Reference)** | - | - | Genbank NC_026301 |
| ***G. gnemon* L** | B | A | SM 13-82 |
| ***G. gnemonoides* Brongn.** | B | B | SM 09-64a |
| ***G. hainanense* Cheng** | B | B | SM 09-65b |
| ***G. indicum* (Lour,) Merr.** | B | B | DS 19791010 |
| ***G. klossii* Merrill ex Markgr.** | B | B | SM 09-66b |
| ***G. latifolium* Bl.** | B | B | SM 09-67a |
| ***G. leyboldii* Tul** | A | A | Won 513 |
| ***G. montanum* Markgr.** | B | B | NYBG 224/84-A |
| ***G. nodiflorum* Brongn.** | B | B | SM 09-69a |
| ***G. pendulum* Cheng** | A | A | SM 13-118 |
| ***G. ula* Brongn.** | B | B | SM 09-71a |
| ***G. urens* Bl.** | B | B | DS s.n. |
| ***W. mirabilis* (cp genome)** | - | - | Genbank NC_010654.1 |

Under the Extraction and Library Preparation headers, A = completed by Angela McFadden, B = received sample or data from another source.
SM = Sarah Mathews, NYBG = New York Botanical Gardens DS = Dennis Stevenson. The characters following the individual or institution refer to a reference number from the source location.

*LIBRARY CONSTRUCTION*

For the library construction, the TruSeq Library Preparation High Throughput Kit (Illumina Inc., San Diego, CA, USA) was used, but as we had only 17 samples to prepare, the TruSeq protocol for Low-Throughput was used per the manufacturer's instructions, except where indicated. To create DNA fragments of the size indicated in the protocols, samples were sheered via ultrasonic sonication, using the Sonic Wave Digital Ultrasonic Jewelry Cleaner, for a duration of 90-180 seconds. The sizes were confirmed using electrophoresis on an agarose gel, and compared against a size marker. This process was repeated until the fragments were concentrated around 400 to 500 bp.

The next step in library preparation cleaved overhangs on the DNA to create blunt ends, and then added phosphate groups to the 5' ends, which is required for ligation. After the blunting step, the MinElute PCR Purification Kit (Qiagen, Hilden, Germany) was used, per the instructions, except that I eluted 15 µl with EB (MinElute PCR Purification Kit) in the final step which allowed for the correct amount of liquid to continue to the adenylation of the 3' ends, which adds an adenine (A) to the 3' ends of the DNA fragments.

Once adenylation was complete, I added the appropriate adapter to each sample. The adapters are specific identifying sequences that, once ligated to our samples, allowed us to have targets for sequencing, and allowed for multiplexing. For five of the samples (*G. diminutum, G. pendulum, G. gnemon, G. buchholzianum,* and *G. africanum*) I ligated the adapters per the instructions (incubation for 10 min at 30°). In an attempt to make ligation more effective, the other species (*G. leyboldii, G. acutum, G. arboretum, and G.*

*schwackeanum*) had a slightly altered ligation protocol. This protocol consisted of adding the kit components, but instead of using the provided Resuspension Buffer, a 1% PEG solution was added (Pfeiffer & Zimmerman, 1983). In addition to the PEG, the ligation times were altered. The samples were ligated at 4°, overnight. For *G. leyboldii* only, the DNA and adapters were heated to 65° for five minutes before ligation.

After ligation was stopped, a cleaning step was performed using the MinElute PCR Purification Kit protocol per the instructions, except, that the DNA was eluted in 20 µl of EB. The ligated samples were then subjected to size separation via TAE agarose electrophoresis to separate the unincorporated adapter sequences. A brief ethidium bromide (EtBr) stain allowed viewing on a transilluminator, and the desired DNA, at around 500 bp, was cut from the 2% TAE agarose gel.

Polymerase Chain Reaction (PCR) was used to multiply the library fragments. In an attempt to produce more of the desired product and reduce the amount of primer-dimer in the PCR products, I began altering the amount of reagents used. I used less primer cocktail (kit) and replaced the volume with resuspension buffer (kit). This provided some improvement, but I still saw bands of primer dimer. To further purify the PCR product, I began extracting the DNA from the agarose using the Qiagen Gel Extraction protocols (per instructions).

Once completed, some of the PCR products had a cleaning step performed on them. After viewing on an agarose gel, I determined that further size selection was required (due to the copious amounts of primer-dimer I was still receiving). To do this, I loaded the PCR products on 2% TAE gels and proceeded with a second size selection,

removing the sections of the lanes between 400 and 1000 bp and removed the DNA from the agarose using QIAquick Gel Extraction protocol (Qiagen, Hilden, Germany), per the QIAquick Gel Extraction instructions. Samples obtained from like species were combined (as an example, more than one *G. buchholzianum* library was completed, using the same adapter, and so all *G. buchholzianum* samples were combined into one aliquot), and the concentration of each of the samples was determined using the Qubit fluorimeter 2.0, the target concentration being 1 µg.

Using this information, I pooled the completed libraries, which were then concentrated using a MinElute column. The combined libraries were eluted in 15 µl of Elution Buffer (EB). The pooled and multiplexed libraries were sent to Harvard University for single-end Illumina sequencing. Harvard returned the de-multiplexed data.

*BIOINFORMATICS*

The millions of 101 bp single-end sequencing reads that were returned from Harvard were then trimmed by quality as well as the standard removal of adapter artifacts. I used Trimmomatic 0.3 (Bolger et al., 2014) for this task. I also trimmed by removing the low quality nucleotides at the start and end of each read. Next, I used a sliding window trimming technique to ensure quality throughout the entirety of each read. Lastly, I removed reads less than 50 bp long, and those that scored below an average Phred score quality of 20 (Figure 3). After removing all Illumina artifacts, I further refined the data by using a variety of trimming methods. The remaining reads (Table 3) had very high quality scores (32-40 Phred quality).

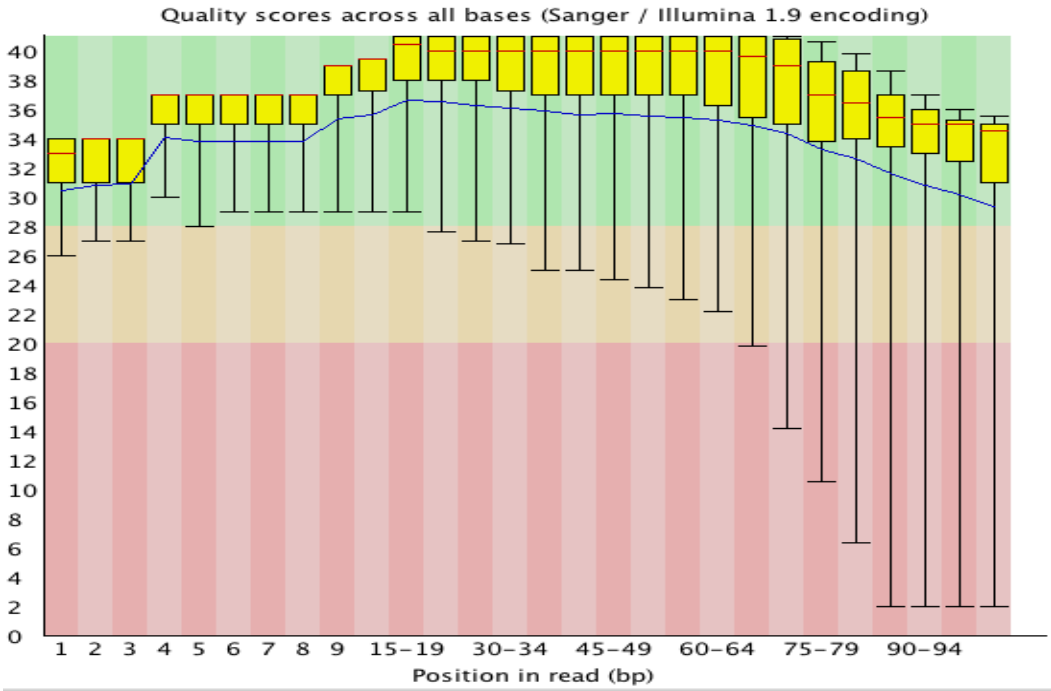
I assembled the remaining reads against the *G. gnemon* chloroplast sequence using the Geneious 8.0.5 (Kearse et al., 2012) "Map to Reference" function, which uses a combination of reference-guided and *de novo* assembly. The chloroplast sequence is the complete plastome sequence from the National Center for Biotechnology Information (NCBI) (NC_026301).

In addition to basing phylogenetic analysis on the complete plastome genome, I performed analyses using specific chloroplast genes (Table 4). After assembly using the *G. gnemon* plastome complete sequence from NCBI, the chloroplast genes for each species were attained by using a transfer of annotation feature in Geneious 8.0.5 (Kearse et al., 2012). Each of the subsequent assemblies were submitted to MAFFT's alignment program, and reformatted into Nexus files. Once aligned, the phylogenetic analyses could be conducted.

*PHYLOGENETIC ANALYSIS*

The alignment of these genome sequences was the basis of phylogenetic analyses using Maximum Parsimony (MP) and Maximum Likelihood (ML). Two data matrices were constructed using chloroplast protein-coding genes (cp Genes) and the whole chloroplast plastid (plastome). I performed ML and MP analyses on each data matrix, determining the best-fit tree and a bootstrap tree after a bootstrap analysis. I conducted MP analyses using default settings of a heuristic search using a program called Phylogenetic Analysis Using Parsimony (PAUP) 4.0a146 (Swofford, 2002), and I conducted the ML analyses using RaxML in the Cipres Science Gateway (Stamatakis, 2014).

**a**




**b**

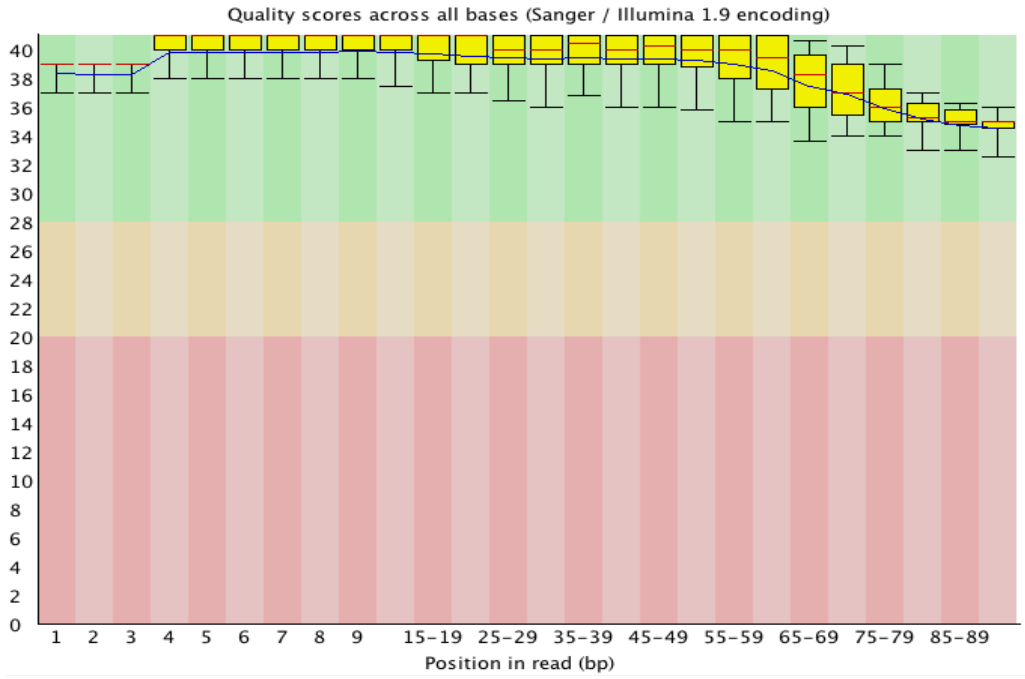



Figure 3 Graphical depictions of the effects of the trimming treatments on the reads, generated using FastQC (Andrews, 2011). The y-axis shows the Phred quality score, in relation to the x-axis, which shows the position in the read (bp).
Figure 3a shows the data before treatment. Figure 3b shows data after treatment. In Figure 3b, all the low quality reads have been removed and the average scores of all reads in the resulting concatenation are very high.

Table 3 Read Data

| Species | Raw # of Reads | # Reads After Trim | # Reads Assembled |
|---|---|---|---|
| ***G. acutum*** | 3,225,383 | 779,043 | 5,649 |
| ***G. africanum*** | 18,610,739 | 9,49 2,420 | 730,694 |
| ***G. buchholzianum*** | 23,518,745 | 7,272,462 | 154,835 |
| ***G. diminutum*** | 11,019,839 | 8,227,335 | 42,099 |
| ***G. gnemon*** | 4,399,310 | 3,042,329 | 52,867 |
| ***G. gnemonoides*** | 17,049,077 | 14,125,155 | 72,442 |
| ***G. hainanense*** | 16,395,361 | 13,492,545 | 170,968 |
| ***G. indica*** | 20,691,222 | 18,010,519 | 181,165 |
| ***G. klossii*** | 20,396,975 | 17,535,400 | 58,874 |
| ***G. latifolium*** | 17,153,368 | 14,004,513 | 127,068 |
| ***G. leyboldii*** | 15,349,715 | 914,011 | 18,604 |
| ***G. montanum*** | 21,443,037 | 18,558,950 | 123,852 |
| ***G. nodiflorum*** | 12,209,610 | 9,373,225 | 147,853 |
| ***G. pendulum*** | 7,085,919 | 4,964,411 | 45,872 |
| ***G. ula*** | 18,182,223 | 15,015,087 | 92,673 |
| ***G. urens*** | 14,528,036 | 11,985,025 | 220,534 |

Table 4 List of Chloroplast Genes Used in cp Genes Data Set.

| **Chloroplast Genes** | | | | | | | |
|---|---|---|---|---|---|---|---|
| *atpA* | *infA* | *psaA* | *psbD* | *psbM* | *rbl20* | *rps3* | *rps18* |
| *atpB* | *matK* | *psaB* | *psbE* | *psbN* | *rbl22* | *rps4* | *rps19* |
| *atpE* | *petA* | *psaC* | *psbF* | *psbT* | *rbl33* | *rps7* | *ycf1* |
| *atpF* | *petB* | *psaI* | *psbH* | *psbZ* | *rpoA* | *rps8* | *ycf2* |
| *atpH* | *petD* | *psaJ* | *psbI* | *rbsL* | *rpoB* | *rps11* | *ycf3* |
| *atpI* | *petG* | *psbA* | *psbJ* | *rbl12* | *rpoC1* | *rps12* | *ycf4* |
| *ccsA* | *petL* | *psbB* | *psbK* | *rbl14* | *rpoC2* | *rps14* | *ycf12* |
| *cemA* | *petN* | *psbC* | *psbL* | *rbl16* | *rps2* | | |

To test the relationships between the major clades of *Gnetum*, I conducted additional likelihood-based statistical tests using PAUP Version 4.0a149 (Swofford, 2002). The tree statements used for these tests leave the branches of the various clades unresolved except for the specific node or relationship pattern being tested. I conducted the one-tailed Shimodaira-Hasegawa test (SH) and a Shimodaira Approximately Unbiased Test (AU). I conducted these tests were conducted using the default settings in PAUP Version 4.0a149 (Model = HKY85). The hypotheses (Table 5) tested by the SH and AU tests were:

1. To test the monophyly of the currently accepted sections (*Gnetum* and *Cylindrostachys*), which are not strictly based on geographical distribution, we tested two possibilities:
    a. The statistical likelihood that the species in the southern hemisphere (South America and Africa) are sister to Asian species (rather than grouped by currently recognized sections).
    b. The currently recognized sections of the genus (Section *Gnetum* and Section *Cylindrostachys* Markgr.) are phylogenetically valid.
2. To test the statistical likelihood of each of the basal-clade hypotheses, we tested:
    a. South American taxa (subsection *Araeognemones* Markgr.) as sister to the rest of the genus.
    b. African species (subsection *Micrognemones* Markgr.) is sister to the rest of the genus.

c. The arborescent species (subsection *Eugnemones* Markgr.), represented by *G. gnemon* in this study, is sister to the rest of the genus.

3. To test the statistical likelihood of monophyly for each of the subsections (that were not already included in prior models) we tested:

a. The subsection *Stipitati* Markgr. is monophyletic.

b. The subsection *Sessiles* Markgr. is monophyletic.

*ADDITIONAL ANALYSES: MITOCHONDRIA*

In addition to the chloroplast data, I attempted to assemble mitochondrial genome sequence from the total genomic pool of Illumina reads. To do so, I obtained a draft version (Guo & Mower, unpublished) of the *G. gnemon* mitochondrial genome and approximately 20 gene sequences and attempted assemblies against these two references using the same methods used for chloroplast sequences. The number of reads that assembled to the mitochondrial references was significantly lower than those that assembled for the plastome for each species, and the nature of the assemblies was problematic. I saw hundreds of reads that were assembling to sections a few hundred bp's long, and then long stretches where the coverage was at or near zero (Figure 4). This was especially noticible with the second set of libraries generated from herbarium specimens instead of fresh material (Table 2). Attempts were made to clarify the situation, such as using higher stringency settings, assembling only reads that did not assemble to the chloroplast (eliminating thousands of reads from the pool), but ultimately the decision was made to move forward with the chloroplast data here and wait on the mitochondrial

DNA analysis until such time as the mitochondrial draft could be finished, or new tools were developed for differentiating Plastome reads from mtDNA reads, which were both beyond the scope of this thesis.

The *Welwitschia* mitochondrial genome (NC_029130) was published as I was completing my work and I did try assembling my reads using that reference. The assembly to the complete *Welwitschia* mitochondrial genome did work better, but it still had some of the same issues, plus some new ones. The *Welwitschia* mitochondrial genome is almost 1 million base pairs long (978,846 bp), and fewer than 60,000 base pairs assembled for any *Gnetum* species. The areas that did assemble were mostly mitochondrial specific genes, suggesting *Gnetum* mtDNA truly is being assembled, but that the intergenic regions were too divergent from those of *Welwitschia* to assemble. A finished *Gnetum* mitochondrial genome, once available, might be a reference that would produce usable phylogenetic data for both mtDNA genes and non-coding sequence.

The difficult nature of extracting DNA from dried herbarium leaf samples could have contributed to the issues faced with the mitochondrial data. Perhaps I was able to obtain enough chloroplast data, due to the high concentration of chloroplast in the leaves, but was not able to retrieve enough mitochondrial DNA to have a workable dataset. However, the problem could lie in the reference itself, being a draft sequence, it could be an issue of dirty data or artifacts.

This analyses on mitochondrial DNA may be subject to complications due to another issue as well: intergenomic and horizontal gene transfer (IGT and HGT, respectively). IGT is relatively common. There have been many cases showing transfer of

genetic material between the chloroplast and mitochondrial genomes (Stern & Palmer 1984, Zhang et al., 2012, Rodriguez-Moreno et al., 2011, Alverson et al., 2010, Alverson et al., 2011) and even between the nuclear and mitochondrial genomes (Rodriguez-Moreno et al., 2011, Alverson et al., 2010, Alverson et al., 2011). HGT is the transfer of genes between species. This has been seen in many instances between bacteria and plants (Synanen & Kado 2002, Brown 2003) and found frequently in the mitochondrial of genomes angiosperms (Puerta 2014). Won and Renner (2003) observed a transfer between *Gnetum* and the angiosperms *Pagamea* and *Petunia*. The mechanism for this is still unknown (bacterial, viral, or insect vectors are all possibilities). The potential problem IGT and HGT pose for this analysis is that there are so many sequences in the mtDNA that are similar to cpDNA. Since there is such an abundance of reads in the Illumina pool from cpDNA (due to the high numbers of chloroplast in the leaves), when the reads are assembled to the mitochondrial reference, the chloroplast reads will assemble to these cp-like areas. This would lead to areas where the consensus sequences being produced are not mitochondrial, but chloroplast assemblies.

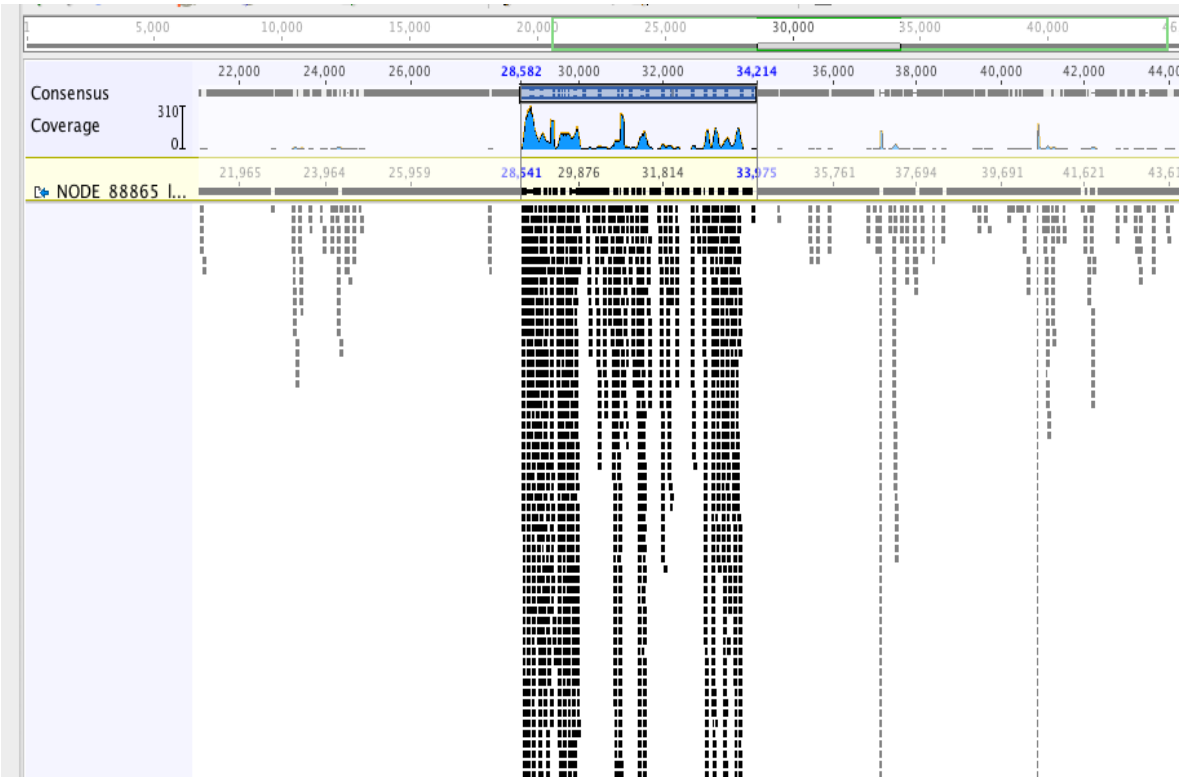


Figure 4 Reads of *G. diminutum* assembling to the mitochondrial draft. Some areas have very high coverage, with many reads assembling against a short section, while other areas are left with no coverage.

Table 5 Statistical Test Results

| **Tree Statement** | **Data Set** | **-lnL** | **Diff -ln** | **SH** | **wtd-SH** | **AU** |
|---|---|---|---|---|---|---|
| **(S. America)** | Plastome (With Welwitschia) | 446778.4975 | (best) | | | |
| **(S. America & Africa) (Asia)** | Plastome (With Welwitschia) | 449607.6494 | 2829.152 | 0.0000* | 0.0000* | ~0* |
| **(Section *Gnetum*) (Section *Cylindrostachys)*** | Plastome (With Welwitschia) | 452873.2835 | 6094.786 | 0.0000* | 0.0000* | ~0* |
| **(Africa)** | Plastome (With Welwitschia) | 459360.0644 | 12581.57 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Sessiles*)** | Plastome (With Welwitschia) | 469081.0989 | 22302.6 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Stipitati*)** | Plastome (With Welwitschia) | 469571.2538 | 22792.76 | 0.0000* | 0.0000* | ~0* |
| **(Arborescent *Gnetum*)** | Plastome (With Welwitschia) | 469792.4117 | 23013.91 | 0.0000* | 0.0000* | ~0* |
| **(S. America)** | Plastome | 342810.44049 | (best) | | | |
| **(S. America & Africa) (Asia)** | Plastome | 344795.25339 | 1984.81290 | 0.0000* | 0.0000* | ~0* |
| **(Section *Gnetum*) (Section *Cylindrostachys)*** | Plastome | 347790.44700 | 4980.00651 | 0.0000* | 0.0000* | ~0* |
| **(Africa)** | Plastome | 352881.60332 | 10071.16283 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Sessiles*)** | Plastome | 356847.39830 | 14036.95781 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Stipitati*)** | Plastome | 356907.06403 | 14096.62354 | 0.0000* | 0.0000* | ~0* |
| **(Arborescent *Gnetum*)** | Plastome | 357127.88500 | 14317.44451 | 0.0000* | 0.0000* | ~0* |
| **(S. America)** | cp Genes (With Welwitschia) | 229889.67378 | (best) | | | |
| **(S. America & Africa) (Asia)** | cp Genes (With Welwitschia) | 231905.42037 | 2015.74660 | 0.0000* | 0.0000* | ~0* |
| **(Section *Gnetum*) (Section *Cylindrostachys)*** | cp Genes (With Welwitschia) | 234314.84856 | 4425.17478 | 0.0000* | 0.0000* | ~0* |
| **(Africa)** | cp Genes (With Welwitschia) | 237598.69200 | 7709.01823 | 0.0000* | 0.0000* | ~0* |
| **(Arborescent *Gnetum*)** | cp Genes (With Welwitschia) | 242476.14895 | 12586.47518 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Sessiles*)** | cp Genes (With Welwitschia) | 242530.08272 | 12640.40895 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Stipitati*)** | cp Genes (With Welwitschia) | 242620.59139 | 12730.91761 | 0.0000* | 0.0000* | ~0* |
| **(S. America)** | cp Genes | 189809.03226 | (best) | | | |
| **(S. America & Africa) (Asia)** | cp Genes | 191857.03872 | 2048.00646 | 0.0000* | 0.0000* | ~0* |
| **(Section *Gnetum*) (Section *Cylindrostachys)*** | cp Genes | 194186.20939 | 4377.17713 | 0.0000* | 0.0000* | ~0* |
| **(Africa)** | cp Genes | 197424.48152 | 7615.44926 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Sessiles*)** | cp Genes | 201888.47183 | 12079.43957 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Stipitati*)** | cp Genes | 201943.19357 | 12134.16131 | 0.0000* | 0.0000* | ~0* |
| **(Arborescent *Gnetum*)** | cp Genes | 202112.36340 | 12303.33114 | 0.0000* | 0.0000* | ~0* |
| **(S. America)** | Won & Renner (Outgroups) | 51106.56419 | (best) | | | |
| **(S. America & Africa) (Asia)** | Won & Renner (Outgroups) | 51219.48837 | 112.92418 | 0.2855 | 0.0145* | 0.0074* |
| **(Section *Gnetum*) (Section *Cylindrostachys)*** | Won & Renner (Outgroups) | 51421.08494 | 314.52075 | 0.0077* | 0.0121* | 0.0042* |
| **(Arborescent *Gnetum*)** | Won & Renner (Outgroups) | 51826.91841 | 720.35421 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Sessiles*)** | Won & Renner (Outgroups) | 51895.34914 | 788.78495 | 0.0000* | 0.0000* | ~0* |
| **(Africa)** | Won & Renner (Outgroups) | 51945.08033 | 838.51613 | 0.0000* | 0.0000* | ~0* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **(Subsection *Stipitati*)** | Won & Renner | (Outgroups) | 52025.74707 | 919.18288 | 0.0000* | 0.0000* | ~0* |
| **(S. America & Africa) (Asia)** | Won & Renner | | 43982.29545 | (best) | | | |
| **(Section *Gnetum*) (Section *Cylindrostachys)*** | Won & Renner | | 44260.27689 | 277.98144 | 0.2520 | 0.2116 | 0.0735 |
| **(S. America)** | Won & Renner | | 45858.01988 | 1875.72443 | 0.0000* | 0.0000* | ~0* |
| **(Arborescent *Gnetum*)** | Won & Renner | | 46524.58645 | 2542.29100 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Sessiles*)** | Won & Renner | | 48340.45848 | 4358.16303 | 0.0000* | 0.0000* | ~0* |
| **(Subsection *Stipitati*)** | Won & Renner | | 48935.89845 | 4953.60300 | 0.0000* | 0.0000* | ~0* |
| **(Africa)** | Won & Renner | | 49623.58301 | 5641.28756 | 0.0000* | 0.0000* | ~0* |

- **Values for tests are p values for Null hypothesis of no difference between trees. *p < 0.05**
- **Tree Statement shows the arrangement of taxa in the test, ordered from best supported to least supported. Taxa within the (brackets) were set against the rest of the genus, or against the second set of (brackets) when present.**
- **Won and Renner data obtained from TreeBase. Won and Renner (outgroups) = *matK* and *rbcL*, aligned to various seed plants, Won and Renner (*Gnetum*) = 6 loci, all only *Gnetum* genus included.**

# CHAPTER III

# RESULTS

I used two data sets: the mostly complete plastome sequences and the protein coding genes extracted from the plastomes (Table 4). The total number of informative characters in the plastome data set was 9,731 bp and in the cp Genes data set was 5,089 bp (Table 6). For each, MP and ML approaches were used to infer phylogenetic hypotheses. The resulting trees were rooted using the *Welwitschsia mirabilis* (accession number NC_029130.1) Plastome or cp Genes. The topologies of each of the resultant chloroplast-based phylogenies were similar for both data sets and methods, and the confidence levels for most nodes were very high (Figure 5).

Table 6 Phylogenetic Character Information

| **Data Matrix** | **Total Characters** | **Constant Characters** | **Autapomorphic Characters** | **Informative Characters** |
|---|---|---|---|---|
| **Plastome** | 128,772 | 98,178 | 21,863 | 9,731 |
| **cp Genes** | 66,815 | 50,504 | 11,222 | 5,089 |

## *TOPOLOGIES*

My topologies are mostly congruent with those hypothesized in previous studies; the South American taxa diverged first, and there are two major clades within the Asian species that were classified as Asia I and Asia II by Won and Renner (2006) (Figure 2). When the ML best fit tree was compared to the phylogenies of Hou et al. (2015) and Won and Renner (2006) (Figure 2), the topologies of shared species were consistent, save the placement of *G. ula*, which I found to be a sister of *G. latifolium*. In the study by Won and Renner (2006) *G. ula* was included as a member of their *cuspidatum* clade

(represented by *G. diminutum, G. klossii*, and *G. acutum* in this study). In the study by Hou et al. (2015) *G. ula* (syn *G. edule*) was found to be sister to a clade containing *G. latifolium*, but also *G. neglectum*, *G. leptostachyum*, and others.

The *G. diminutum, G. klossii*, and *G. acutum* clade (Figure 5a) is well defined and well supported. My results support a *G. diminutum /G. klossii* node (100/100/69 for Plastome ML, Plastome MP, and cp Genes ML, respectively). There is a disagreement that should be noted, in the cp Genes matrix MP, the tree supports a *G. acutum/G. klossii* sister relationship, though the bootstrap value is low, at 57 (Figure 5a).

The relationship between *G. montanum, G. indica*, and *G. hainanense* had been poorly resolved in previous work, and where resolution was available, the support values were relatively low. The phylogeny here shows fully resolved, well supported (with bootstrap values of 100) branching in the ML and MP analyses of the plastome matrix. The ML and MP best-fit trees inferred from the cp Genes data set showed no resolution in this clade, the bootstrap analysis supported a branching pattern that had a different sister relationship (Figure 5b), though the support values are not nearly as high (59/77 vs 100/100). *G. pendulum*, which to this point had not appeared in any other phylogeny, also belongs in this group. It is shown to be a sister to the *G. montanum/G. hainanense/G. indica* clade.

One notable discrepancy is observed between the S. American and African taxa of the chloroplast phylogenies. The node placing the S. American taxa as sister to the rest of the genus (S. American-basal hypothesis) is found in the best-fit analyses for both whole Plastome and the cp Genes (Figure 5c) under both the ML and MP models. All of the

bootstrap analyses, however, supported an alternative basal clade, with the African taxa the sister to the rest of the genus (African-basal hypothesis), often with BS support values of 100.

*STATISTICAL RESULTS*

The SH and AU statistical tests, using PAUP Version 4.0a149 (Swofford, 2002), were applied to my Plastome and cp Genes data matrices to test the major clades. The placement of the South American taxa (subsection *Araeognemones* Markgr.) as sister to the rest of the genus, as was found in my ML and MP trees, was significantly better than any other alternative (Table 5). The hypotheses that either the arborescent species or the African taxa were the first clade had significantly less support in my data (Table 5). The same tests were applied to the two Won and Renner (2006) data sets obtained from TreeBase (https://treebase.org/treebase-web/search/study/summary.html?id=1548). The statistical results from the Won and Renner (2006) data sets also showed support for the S. American-basal hypothesis, but the support values were not as high as with my data (Table 5)

*Gnetum* Section *Cylindrostachys* Markgr. is supported as monophyletic in my study, as well as in Won and Renner (2006) and Hou et al. (2015). *Gnetum* Section *Gnetum*, on the other hand, is only supported in Hou et al. (2015). In both this study and the study conducted by Won and Renner (2006), this section is separated geographically, with *G. gnemon* being more closely related to G. gnemonoides and the other Asian species. This result was also highly supported by my data in the AU and SH tests (Table 5).

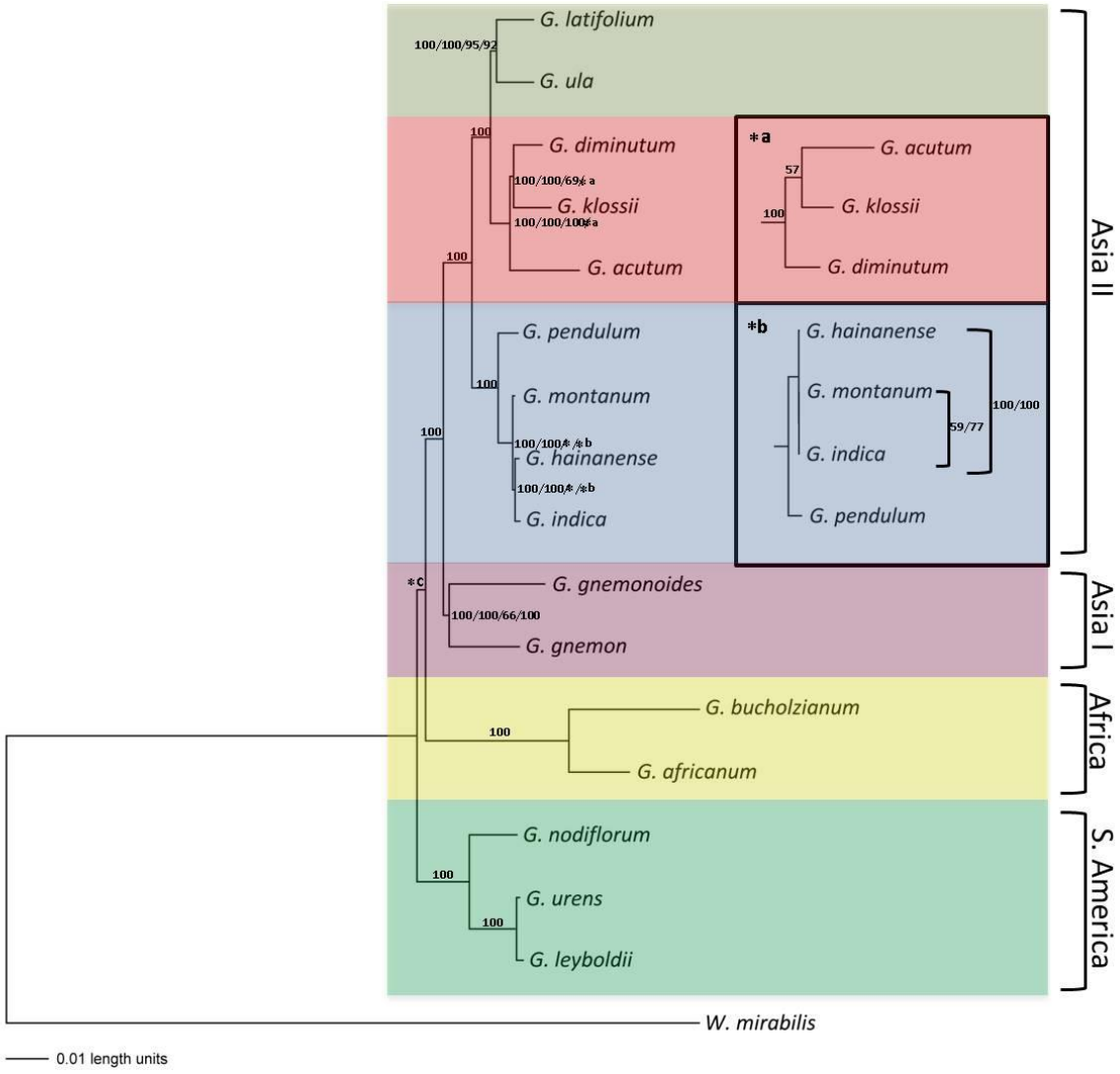



Figure 5 The phylogeny as represented by the Plastome ML best-fit tree. Scores from each bootstrap analysis is shown on the branches (Plastome ML/Plastome MP/cp Genes ML/cp Genes MP). Colors are a reference to the clade, and are consistent throughout figures, the geographical identifiers reference the clades found by Won and Renner (2006). When only one number is present at a particular node, the BS value on that node identical in all analyses. An asterisk (✱) denotes a node that was not recovered in all analyses. Variant (✱a) shows the branching pattern from the cp Genes MP analyses. Section (✱b) shows a branching pattern from the cp Genes ML best-fit tree. In the best-fit trees (MP and ML) for the cp Genes, the relationship between *G. montanum*, *G. indica*, and *G. hainanense* was unresolved. The bootstrap values for ML and MP are included to the right of the branches. The third asterisk, (✱c), signifies a disagreement on whether the S. American or African taxa are basal. The best-fit analyses supported the topology as displayed above. The bootstrap analyses all supported the opposite, with values of 100/100/58/100.

*PLASTOME VS GENES*

There were primarily two large data sets used for this study. Though the cp Genes data set was slightly less than half the size of the whole plastome, it was still significantly larger than the data sets used to infer phylogenies in either of the other two studies (Table 6). When comparing these two data sets, the most noticeable thing is the number of informative characters (Table 6). The plastome data set has almost double the number of informative characters that are seen in the genes matrix. The result of this is seen in the ML and MP genes topology. For both the ML and MP of the cp Genes matrix, there was no resolution in the *G. hainanense, G. indica,* and *G. montanum* (excepting the bootstrap analysis, which had relatively low values (59/77). In the plastome-inferred topologies, however, there was resolution in this clade, with high bootstrap values (100/100) (Figure 5b).

In the SH and AU tests (Table 5), the plastome and cp Genes data both return a best-fit topology well differentiated from other possible phylogenies. The p values reject the null hypothesis, that there is no difference between my tested tree statements, and support the S. American-basal hypothesis. I tested the possible hypotheses, both with and without a *Welwitschia mirabilis* outgroup, using each of my data sets. I also compared the possible hypotheses using the Won and Renner (2006) data. The difference between their smaller data sets, and my larger one, is clear. The p values of my tests, in all configurations, are below 0.05. The Won and Renner data is unable to statistically differentiate between the best supported and the second best tree statements. When *Welwitschia* is included with the Won and Renner *rbcL* and *matK* matrix, the South

American basal tree is best supported and the Southern Hemisphere-Northern Hemisphere tree is second best. The second best tree is not significantly less likely given these data under the SH test, but raises to statistical significance with the AU test. When the full Won and Renner gene set is used without any outgroup taxa, the Hemispheres tree (South American and African taxa as sisters) is the best tree, but it cannot be statistically discriminated from the Markgraf section hypothesis under either the SH or AU tests.

# CHAPTER IV

# DISCUSSION

My results show that species relationships in *Gnetum* mainly correspond to the geographic distribution of the taxa. My phylogenetic hypotheses, based on MP and ML analyses of two large chloroplast datasets, are mostly congruent with the results of previous workers, specifically Won and Renner (2006) and Hou et al. (2015). There are, however, questions that have come to light. The hypothesis that the S. American taxa are the basal clade and sister to the other species in the genus is examined here. ML and MP trees, as well as the majority of the SH and AU values support this, though bootstrapping methods support the African-basal hypothesis. If the African taxa were basal, this would mean a shift in the ideas of where *Gnetum* originated and how the species dispersed across the tropics.

## *PHYLOGENIES*

The South American taxa (Subsection *Araeognemones* Markgr.) were favored as the basal lineage in the topologies of both Won and Renner (2006) and Hou et al. (2015). This configuration was therefore the basis of their dating analyses. My MP and ML best-fit trees also support this topology, as did the AU test. However, my bootstrap results suggest that there might be reasonable alternative hypotheses. Both bootstrap trees (ML and MP) supported the African taxa (subsection *Micrognemones* Markgr.) at the base of the tree (cp genes ML boot 58, all others 100). Bootstrapping is a method of resampling the sample with replacement (Chernick and LaBudde, 2011). Bootstrapping allows for many plausible trees to be described, and provides a value based on the number of times

a specific node is found. With many plausible trees, it is unsurprising that we would see variation. It is surprising, however, that we see the African-basal hypothesis receiving such strong bootstrap support, especially when the other statistical tests support the S. American-basal hypothesis (Table 5). Tests such as bootstrapping and the SH test do have a known selection bias. The value of comparing many trees is overshadowed by the possibility of overconfidence in an incorrect topology (Shimodaira, 2002). The AU test was created as means of reducing this bias, and has been shown to be effective (Shimodaira, 2002). In this instance, though, both the SH and AU tests support the S. American taxa as the basal lineage (Table 5).

One possible reason for the discrepancy between the two basal hypotheses is Long Branch Attraction (LBA). Though we used ML analysis, which is supposed to be less sensitive to LBA than other methods (Bergsten 2005), it has been shown to be susceptible on alignments 100,000 bp or more (Kück et al., 2012). Outgroup taxa almost always present as long branches (Bergsten 2005). *Gnetum* are, unfortunately, very divergent from their closest extant relative, *Welwitschia mirabilis* (McCoy et al., 2008). This means, just by chance and not due to shared evolutionary history, there are likely many similarities between the mutations in the 'long branch' leading to *Welwitschia*, and the next longest branch. As both the African and S. American taxa might be expected to have "long branches," additional testing should be done to determine if LBA is distorting the *Gnetum* tree. This could be done using separate partition analysis, long branch extraction, or methodological disconcordance.

The Plastome data set should also be examined for bias due to the inclusion of both copies of the inverted repeat (IR) in the assembled plastome sequence; this has the effect of weighting (2X) the phylogenetically informative characters located there. The repeated section may be biasing the bootstrap in the direction of the African-basal hypothesis, although the characters are similarly weighted in the best-fit analyses. Also, the fact that we are seeing the same pattern with the cp Genes data set, which would not be susceptible to the IR bias, seems to negate this as an important factor. Still, it could certainly be contributing to the degree of certainty we see in the bootstrap values of the Plastome data set. The bootstrap value for the African-basal node in the cp Genes ML is 58, whereas in the Plastome ML it is 100.

If the African-basal hypothesis is true, the implications are far reaching. The current dating analyses (Won & Renner, 2006; Hou et al., 2015) place South America as the "epicenter" of the genus and the distribution radiated out from there. The idea is that from South America, the species then moved to Africa, then Asia. Won and Renner (2006) stated a rather young timeline for this radiation (Oligocene and Miocene), whereas Hou et al. (2015) estimated an older (late Mesozoic) radiation. If the major clades split more recently (Won and Renner scenario), *Gnetum* would need to disperse long distances over salt water. If major clades split earlier (Hou et al. dating), their divergences could coincide with continental movement. Still dispersal for short distances over water to explain the more recent radiations in Asia and S. America would be necessary. Dispersal through water, by fish, is supported by a study on the *Gnetum* species *G. venosum.* (Kubitski 1985, as cited in Hou et al., 2015). Mechanisms for oceanic Gnetum dispersal,

though, are uncertain. Clearly, determining the precise relationship of the major clades and the dates of the divergences is critical to understanding *Gnetum* biogeographic history.

### *CODING AND NON-CODING REGIONS*

I anticipated that the differences between the cp Genes data set (coding regions only) and the plastome data set (coding and non-coding regions combined) would be clear. Coding regions, in general, are more conservative than non-coding regions. It follows then, that the coding and non-coding regions would be evolving at different rates. The anticipated result are that the plastome data set would have more phylogenetically informative characters, the plastome data set would provide a more resolved topology, and the plastome data set would be more able to differentiate between the tree hypothesis (AU and SH tests), than the cp Genes data set.

The plastome data did indeed have almost double the number of informative characters found in the genes (Table 6). Also, the plastome data set allowed for resolution of the *G. hainanense, G. montanum,* and *G. indica* relationship that were not resolved in the ML or MP of the cp Genes data set. There was some resolution of this area in the bootstrap analysis of the genes, though the values were some of the lowest in this study (59 and 77 for ML and MP, respectively). Other than the additional resolution in the plastome data set, the Plastome and cp Genes matrices returned very similar topologies (Figure 5). Indeed, both data sets seemed equally able to differentiate between the major hypotheses, with statistical significance (Table 5). This makes a good case for large data sets in general, as the Won and Renner (2006) data that was analyzed could not

differentiate between the all possible hypotheses, but both of my data sets were able to do so.

### *REMAINING WORK*

Running further analyses is suggested. Adding additional outgroups may help to mitigate the LBA issues. Additionally, it would be benefitial to find partitions in the data, test with the Swofford-Olsen-Waddell-Hillis (SOWH), and use additional models, such as the General Time Reversible (GTR + gamma + I) model to investigate support for the basal clade. A rate, or dating analysis will be helpful in addressing the biogeographic questions about these phylogenies.

### *CONCLUSIONS*

In conclusion, I inferred phylogenetic hypotheses that support a deep New World-Old World split within the genus *Gnetum*. This seems to be the consensus among molecular studies, and should be considered when the infrageneric taxonomy is revised. Specifically, the arborescent species should be grouped with the other Asian species, not with the African and S. American taxa. My analyses also give rise to new questions over the basal lineage of this genus, whether or not the South American clade is the earliest diverging lineages. The answer to this question will either support or refute the current ideas of *Gnetum's* biogeographic history.

REFERENCES CITED

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, & Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Molecular Biology and Evolution. 27(6):1436–1448. http://doi.org/10.1093/molbev/msq029

Alverson AJ, Rice DW, Dickinson S, Barry K, & Palmer JD. 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. The Plant Cell. 23(7):2499–2513. http://doi.org/10.1105/tpc.111.087189

Andrews S. Babraham Bioinformatics. 2011. FastQC: a quality control tool for high throughput sequence data. Available online at: www.Bioinformatics.babraham.ac.uk/projects/fastqc

Bakker FT, Culham A, Pankhurst CE, Gibby M. 2000. Mitochondrial and chloroplast DNA-based phylogeny of *Pelagonium* (Geraniaceae).Am J Bot. 87(4): 727-734.

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21:163–193

Biye EH, Balkwill K, & Cron GV. 2014. A clarification of *Gnetum* L. (Gnetaceae) in Africa and the description of two new species. Plant SystEvol. 300:263-272

Bolger AM, Lohse M, & Usadel B. 2014, Bioinformatics, BTU170.

Brown JR. 2003. Ancient horizontal gene transfer. Nat Rev Genet. 4(2)121-132. doi:10.1038/nrg1000

Carlquist S. 1996. Wood, bark, and stem anatomy of Gnetales: a summary. Int J Plant Sci. 157(6 Suppl.):S58-S76

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang Q-Y, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH, Graham SW, Spencer CHB, Dayanandan S, and Albert VA. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Annals of the Missouri Botanical Garden. 80(3): 528-80.

Chauvière G, Guènard D, Pascard C, Picot F, Potier P, and Prangé T. 1982. Taxagifine: new taxane derivative from *Taxus baccata* L. (taxaceae). J. Chem. Soc., Chem. Commun. 495-496. doi: 10.1039/C39820000495

Chernick MR and LaBudde RA. 2011. An introduction to bootstrap methods with applications to R. John Wiley & Sons Inc. Hoboken, NJ.

Clifton SW, Minx P, Fauron CM-R, Gibson M, Allen JO, Sun H, Thompson M, Barbazuk WB, Kanuganti S, Tayloe C, Meyer L, Wilson RK, Newton KJ. (2004). Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiology. 136(3):3486–3503. http://doi.org/10.1104/pp.104.044602

Crosby K & Smith DR. 2012. Does the mode of plastome inheritance influence plastome genome architecture?. PLOS one 7(9):1-8.

Doyle JA. 1996. Seed plant phylogeny and the relationships of Gnetales. Int J Plant Sci. 157(6 Suppl.):S3-S39.

Friedman I. 1996. Introduction to biology and evolution of the Gnetales. Int J Plant Sci. 157(6 Suppl.):S1-S2

Friedman I & Carmichael JS. 1996. Double fertilization in Gnetales: implications for understanding reproductive diversification among seed plants. Int J Plant Sci. 157(6 Suppl.):S77-S94.

Fuashi NA, Popoola L, Mosua IS, Wehmbazeyi NF, Louis NN, & Elah EM. 2010. Harvesting and marketing *Gnetum* species (Engl) in Cameroon and Nigeria. J Ecol Nat Environ. 2(9):187-193.

Guo SX, Sha JG, Bian LZ, & Qiu YL. 2009. Male spike strobiles with *Gnetum* affinity from the early Cretaceous in western Liaoning, northeast China. J SystEvol. 47(2):93-102. doi: 10.1111/j.1759-6831.2009.00007.x

Hagemann R. 2004. The sexual inheritance of plant organelles. In: Daniel H, Chase C (eds). Molecular Biology and Biotechnology of Plant Organelles.Springer, Dordrecht. 93-113

Hooker JD. 1863. On *Welwitschia*, a new genus of Gnetaceae. Trans Linn Soc. 24:1-48

Hou C, Humphreys AM, Thureborn O, & Rydin C. 2015. New insights into evolutionary history of *Gnetum* (Gnetales).Taxon. 64(2):239-253. doi:http://dx.doi.org/10.12705/642.12

Hsu CY, Wu CS, Surveswaran S, & Chaw SM. 2015. The complete plastome sequence of *Gnetum Ula* (Gnetales: Gnetaceae). Mitochondrial DNA. doi:10.3109/19401736.2015.1079874

Katoh K & Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013 Apr;30(4):772-80. PubMed PMID:23329690.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, & Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 28(12), 1647-1649.

Kingston DG. 2016. Paclitaxel (TaxolTM): discovery and development. 9:245-247. IN: ASP History. American Society of Pharmacognosy.

Kleine T, Maier UG, & Leister D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol. 60:115- 38; PMID:19014347; http://dx.doi.org/10.1146/ annurev.arplant.043008.092119

Kück P, Mayer C, Wägele J-W, Misof B. 2012. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. PLOS ONE 7(5): e36593. doi:10.1371/journal.pone.0036593

Mao J, Zhou F, Liu T, Wu Z, Zhong T, Liu C, Wei Q, Chen J, & Huang S. 2015. The complete chloroplast genome of *Gnetum* montanum and sequence analysis. Mitochondrial DNA. http://dx.doi.ord/10.3109/19401736.2015.1127368

McCoy SR, Kuehl JV, Boore JL, & Raubeson LA. 2008. The complete plastome genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. BMC Evol Biol. 8:130 doi:10.1186/1471-2148-8-130

Morgensen HL. 1996. The hows and whys of cytoplasmic inheritance in seed plants. Amer J Bot. 83(3):383-404.

Muhammad AF & Sattler R. 1982. Vessel structure of *Gnetum* and the origin of angiosperms. Amer J Bot. 69(6):1004-1021

Ndam N, Nkefor JP, & Blackmore P. 2001. Domestication of *Gnetum africanum* and *G. buchholzianum* (Gnetaceae), over-exploited wild forest vegetables of the Central African Region. SystGeogr Pl. 71:739-745

Palmer JD. 1992. Mitochondrial DNA in plant systematics. *In* P.S. Soltis, DE Soltis and JJ Doyle [eds.], Molecular systematics of plants, 36-49. Chapman and Hall, New York, New York, USA.

Palmer JD, Jansen RK, Michaels H, Manhart J, & Chase M. 1988. Chloroplast DNA variation and plant phylogeny.Annals of the Missouri Botanical Garden. 75:1180-1206.

Palmer JD & Stein DB. 1986. Conservation of chloroplast genome structure among vascular plants. Current Genetics. 10:823-833.

Price RA. 1996. Systematics of the Gnetales: a review of morphological and molecular evidence. Int J Plant Sci. 157(6 Suppl.):S40-S49.

Pring DR & Lonsdale DM. 1985. Molecular biology of higher plant mitochondrial DNA. Int Rev Cytol. 97:1-46

Raubeson LA & Jansen RK. 2005. Chloroplast genomes of plants. IN: Henry RJ. Plant diversity and evolution: Genotypic and phenotypic variation in higher plants. CAB International. Oxfordshire, UK.

Rodríguez-Moreno L, González VM, Benjak A, Martí MC, Puigdomènech P, Aranda MA, & Garcia-Mas J. 2011. Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. BMC Genomics. 201112:424. doi: 10.1186/1471-2164-12-424

Rydin C & Friis EM. 2010. A new early Cretaceous relative of Gnetales: *Siphonospermumsimplex*gen. Et sp. Nov. from the Yixian Formation of northeast China. B. M. C. Evol. Biol. 10:183. doi: 10.1186/1471-2148-10-183

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51(3):492-508.

Soltis DE & Soltis PS. 2003. The role of phylogenetics in comparative genetics. Plant Physiol. 132(4): 1790-1800. doi:10.1104/pp.103.022509

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312-1313. doi: 10.1093/bioinformatics/btu033

Stern DB & Palmer JD. 1984. Extensive and widespread homologies between mitochondrial DNA and chloroplast DNA in plants. PNAS. 81(7):1946-1950

Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Syvanen M & Kado CI. (eds) 2002. Horizontal gene transfer. Elsevier. 2nd Edition. 437-445

Timmis JN, Ayliffe MA, Huang CY, & Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123-35. PMID:14735123; http://dx.doi. org/10.1038/nrg1271

Uno G, Storey R, & Moore R. 2001. Principles of botany.McGraw-Hill. New York NY 10020

Vane-Wright RI, Humphries CJ, Williams PH. 1991. What to protect? – Systematics and the agony of choice. Biol Conserv. 55:235-254

Wani MC, Taylor HL, Wall ME, Coggon P, and McPhail AT. 1971. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia. Journal of the American Chemical Society, 93(9), pp.2325-2327.

Ward BL, Anderson RS, & Bendich AJ. 1981. The mitochondrial genome is large and variable in a family of plants (cucurbitaceae). Cell. 25(3):793-803

WCSP 2011.World checklist of selected plant families. Facilitated by the Royal Botanic Gardens, KEW. http://apps.kew.org/wcsp/ (accessed June 2016)

Won H & Renner SS. 2003. Horizontal gene transfer from flowering plants to *Gnetum*. PNAS. 100(19):10824-10829

Won H & Renner SS. 2006. Dating dispersal and radiation in the Gymnosperm *Gnetum* (Gnetales) - clock calibration when outgroup relationships are uncertain. Syst Biol. 55(4):610-62

Zhang, T., Fang, Y., Wang, X., Deng, X., Zhang, X., Hu, S., & Yu, J. (2012). The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. PLOS ONE, 7(1), e30531. http://doi.org/10.1371/journal.pone.0030531

Zhu A, Guo W, Gupta1 S, Fan W, & Mower JP.2016. Evolutionary dynamics of the plastome inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol. 209:1747-56. doi:10.1111/nph.13743