STUDIES OF THE EVOLUTION OF PARASITIC PLANTS IN THE GENUS *STRIGA*, USING SYSTEMATIC, POPULATION GENETICS, AND GENOMIC APPROACHES

by

MATTHEW CHARLES ESTEP

(Under the Direction of Jeffrey L. Bennetzen)

ABSTRACT

Understanding the evolutionary histories of groups of organisms is a major focus of evolutionary biology. Three powerful molecular approaches, phylogenetics, population genetics and genomics, can be used to reveal evolutionary relationships at the gene, genome and population levels. These tools were used to investigate *Striga*, a genus of parasitic plants. Six chloroplast loci were employed to construct a hypothesis of evolutionary relationships within the genus. The resulting hypothesis depicts three distinct clades, with the only non-grass parasite (*S. gesneriodies*) sharing a very recent common ancestor with the grass parasite *S. aspera*. Sample sequencing of the nuclear genome revealed fourteen repetitive elements that are influencing genome size and arrangement in *Striga*. These include a DNA transposon, three satellite repeats, and ten retroelements. Genome size values suggest that polyploidization is also contributing to genome evolution within *Striga*. Twelve microsatellite markers were developed to investigate population structure within *S. hermonthica*, one of the few allogamous species within the genus and the greatest threat to agriculture. These markers were then applied to accessions collected in Mali to reveal a large amount of genetic diversity that is broadly distributed across populations with little genetic differentiation and large amounts of gene flow. Some population structure was apparent, but could not be attributed to "isolation by distance" or host species, suggesting other geo-ecological variables may be acting to differentiate Northern populations from Southern populations. Together these analyses offer a valuable understanding of species evolution, genome evolution, and population structure within this clade of agriculturally important parasites and will help direct future work in combating witchweeds.

INDEX WORDS: agricultural weed, DNA, genetic diversity, genome evolution, genome size, parasite, plant host, phylogenetics, polyploidization, repetitive DNA, sample sequence analysis, *Striga*, systematics, witchweed

STUDIES OF THE EVOLUTION OF PARASITIC PLANTS IN THE GENUS *STRIGA*, USING SYSTEMATIC, POPULATION GENETICS, AND GENOMIC APPROACHES

by

MATTHEW CHARLES ESTEP

BS, Appalachian State University, 1999

MS, Appalachian State University, 2002

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010

STUDIES OF THE EVOLUTION OF PARASITIC PLANTS IN THE GENUS *STRIGA*, USING SYSTEMATIC, POPULATION GENETICS, AND GENOMIC APPROACHES

by

MATTHEW CHARLES ESTEP

Major Professor:	Jeffrey Bennetzen
Committee:	John Burke Katrien Devos Jim Leebens-Mack Susan Wessler

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2010

DEDICATION

To Mike Estep, my friend, teacher and father.

ACKNOWLEDGEMENTS

After spending seven years in Athens (2 years as a lab manager and 5 years as a student), there are many people to thank and I will likely neglect a few by accident. First and foremost, I would like to thank my family, particularly Sarah, who endured all of the late nights, early mornings and weekends in the lab, my parents for their encouragement and understanding when I missed important family events, and my sisters for their encouragement. I would also like to thank Jeremy and Megan Debarry, fellow students and the greatest of friends. We spent many hours smoking cigarettes and discussing science, your encouragement and friendship mean a great deal to me.

My committee also deserves great thanks. They each brought a unique set of skills and a distinct viewpoint to each meeting and forced me to think and re-think my conclusions on many occasions. In particular Jeff and Katrien were wonderful mentors who afforded me many great research opportunities and whose lab members were always available for consultation and advice. Their high standards and critical minds demanded that I continually improve and always pushed the science forward. I will truly miss being a part of their respective research teams.

The plant biology department also deserves thanks for supplying a strong infrastructure for research and employs many wonderful researchers. My friend and colleague, Mike Boyd deserves great thanks for running the greenhouse facility and the diverse teaching collections housed within. Mike was always interested and willing to growing something new (even parasitic plants) and his passion for the diversity of plants

should always be admired. He taught me a great deal about horticulture, an important skill for any plant biologist.

I would also like to acknowledge and thank Susan Watkins for reminding me of important deadlines and making sure I stayed on track with paper work and degree requirements. Susan made dealing with the bureaucracy of UGA a manageable task and truly saved me hours of time and frustration.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

General biology of *Striga*:

Striga, commonly known as witchweed, is a genus of parasitic plants distributed across Africa, southern Europe, Australia, and tropical portions of Asia including China, India, and Indonesia. Natural low-density populations can be found in tropical grassland habitats across much of Africa (Mohamed *et al.*, 2001). African species have received the most scientific attention, focusing mainly on the agronomically important species *S. hermonthica, S. asiatica, S. forbesii, S. aspera,* and *S. gesnerioides*. The latter only parasitizes legumes, an important family of dicots, while most species within the genus are Poaceae (grass) parasites. Although there is extensive information concerning the parasitic lifestyle of this genus, comprehensive studies employing molecular markers have not been published on the systematics of the genus and, hence, the relationships of *Striga* species are not well understood (Musselman, 1987). Studies of chloroplast genome evolution within the family Orobanchaceae (containing *Striga*) have revealed the loss of many loci involved with photosynthesis in some holo-parasitic species (Wolfe *et al.*, 1992). However, the nuclear genomes of *Striga* and other Orobanchaceae species have not been investigated in any detail. A few studies have been conducted on the genetic diversity and population structure of *S. asiatica*, *S. gesnerioides*, and *S. hermonthica* using somewhat antiquated molecular techniques (Werth *et al.*, 1984; Bharathalakshmi *et al.*, 1990; Shawe & Ingrouille, 1993; Kuiper *et al.*, 1996; Olivier *et al.*, 1998; Koyama, 2000; Botanga *et al.*, 2002; Gethi *et al.*, 2005; Botanga & Timko, 2006).

Striga has intrigued plant biologists for much of the last century, due to its parasitic habit. The life cycle can be broken down into five major stages; germination, host contact, haustorium formation, growth, and reproduction. One of the most fascinating aspects of the *Striga* lifecycle is the chemical signal required for seed germination. This signal (strigolactones) are produced by most plants and has recently been shown to act within the plant as a hormone influencing branch architecture and is the same signal recognized by arbuscular mycorrhizal (AM) fungi (Bouwmeester *et al.*, 2007; Gomez-Roldan *et al.*, 2008). Approximately 80% of all land plants have a symbiotic relationship with AM fungi (Paszkowski, 2006a). This beneficial relationship allows nutrient starved plants to exchange carbon and water for phosphorus and nitrogen. Plants initiate this symbiosis by exuding a class of secondary metabolites, the strigolactones, into the rhizosphere (Bouwmeester *et al.*, 2007). These signals are products of the carotenoid pathway and provide the AM fungi with a chemical map of the host plant's root system (Matusova *et al.*, 2005). *Striga* seeds have evolved to recognize these signals, which induce germination and chemotropic growth of *Striga* seedlings towards the roots of a possible host. *Striga* is an obligate parasite and must attach to a host plant before stored nutrients become exhausted, usually within the first five days following germination. During this period, *Striga* enzymes release hydrogen peroxide into the surrounding rhizosphere. Hydrogen peroxide interacts with phenolic compounds found in the cell wall of possible hosts, initiating the release of quinones from the host that act as a second chemical signal. Attachment to the host root is accomplished via a modified root structure known as a haustorium. The quinone molecules released from the degraded cell wall of the host initiate *Striga* haustorium formation (Keyes *et al.*, 2001). Once attachment is complete, the plant spends 4-8 weeks developing underground; flower stalks then emerge and are pollinated by insects for most species. These fecund plants end their three-month life cycle by releasing thousands of small seeds (0.2-0.4 mm) with no other obvious modifications for dispersal (Krause & Weber, 1990).

Current phylogenetic placement and species taxonomy:

The parasitic life cycle, including the need for host-derived chemical signals to initiate germination, is not limited to the genus *Striga*, but appears to have evolved prior to the radiation of the family Orobanchaceae (Paszkowski, 2006b). This family of root parasites contains species that range from facultative to obligate parasites, including both hemi-parasitic (having chlorophyll at some life stage) and holo-parasitic (achlorophyllous) members (Nickrent, 2007). *Striga* has traditionally been placed within the family Scrophulariaceae, but was recently moved to Orobanchaceae (along with several other traditional Scrophulariaceae genera) based on a series of systematic investigations using a variety of chloroplast and nuclear loci (dePamphilis, 1995; dePamphilis *et al.*, 1997; Young *et al.*, 1999; Young & dePamphilis, 2000; Olmstead *et al.*, 2001; Wolfe, 2005). The unification of Orobanchaceae makes this family the largest of the 19 described parasitic plant families, containing 89 genera with 2061 described species (Nickrent, 2007). Although species sampling within *Striga* has been sparse, placement of the genus within this family is well supported as a monophyletic clade, sister to the genera *Buchnera* and *Cycnium* (dePamphilis *et al.*, 1997; Manen *et al.*, 2004; Bennett & Mathews, 2006).

A detailed taxonomy of the genus *Striga* describes 28 species that can be found in Africa, with some species containing several sub-species designations (Mohamed *et al.*, 2001). Older publications suggested as many as 40 species, including those found in Southern Europe and Asia, but further field studies and taxonomic work needs to be completed to fully describe the genus worldwide (Musselman, 1987). Musselman and colleagues also used their expertise in morphological characters to describe relationships within the genus (Mohamed, 1994). A total of 26 morphological characters (14 plant architecture characteristics and 12 inflorescence traits) were scored as plesiomorphic (basal) or apomorphic (derived) based on comparisons to other closely related genera within Orobanchaceae. The resulting phylogenetic hypothesis places *S. gesnerioides* and three other species in a basal clade with twelve symplesiomorphic (shared ancestral) morphological characters. Several species then independently branch from the phylogeny, including *S. junodi*, *S. masuria* and a small clade with three species (*S. hirsute*, *S. lutea*, *S. pubiflora*). The phylogeny terminates with a single large clade that contains two branches; one including *S. hermonthica* and *S. aspera* (not depicted as sister species) along with eight other species, while the second branch includes *S. asiatica*, *S. forbesii* and seven other species. No statistical procedures were implemented in this study to assign support to any of the branching patterns observed. Moreover, in later publications, an "*S. asiatica* complex" is described, including *S. asiatica*, *S. hirsute*, *S. lutea*, and *S. elegans*, that does not conform to their original hypothesis with no explanation for this grouping except "morphological similarities" (Mohamed & Musselman, 2008). At the same time, several publications have argued for *S. hermonthica* and *S. aspera* to be considered sister species, based on hybridization experiments and morphological similarities (Kuiper *et al.*, 1996; Aigbokhan *et al.*, 2000). The above-described hypotheses based on morphological characters does not place them as sister taxa. These inconsistencies need to be addressed using modern molecular techniques.

Chromosome number in *Striga:*

A traditional approach to understanding genome evolution has been to count chromosomes in meiotic (haploid) or mitotic cells (diploid). Usually the best tissue for these chromosome counts has been a root tip, but shoot tips (apical meristem) or flower organs (pollen mother cell and anthers) can be used. The highly transient haustorium (attachment organ) of parasitic plants replaces the root tip, making this tissue's use for cytogenetics impractical. Hence, other tissues have been used to estimate chromosome numbers in several *Striga* species (Musselman *et al.*, 1991; Iwo *et al.*, 1993; Aigbokhan *et al.*, 2000). Haploid chromosome numbers were reported as 20 for *S. asiatica* and 30 or 40 for *S. hermonthica* depending on location, using anther squashes from *Striga* flowers (Iwo *et al.*, 1993). A second study using pollen mother cells reported haploid chromosome numbers of 27 for *S. asiatica* and 32 for *S. hermonthica* (Musselman *et al.*, 1991). The most recent chromosome counts using shoot tips reported haploid chromosome numbers of 18 for *S. asiatica* and 19 for *S. hermonthica* (Aigbokhan *et al.*, 2000). Chromosome counts were also reported for *S. aspera* (N=30), *S. gesnerioides* (N=20), and *S. forbesii* (N=22) using anther squashes (Iwo *et al.*, 1993). The reported chromosome counts are widely variable between the three investigations and no consensus was obtained in any of the studies. Therefore, it seems prudent to assume that chromosome number and ploidy are not understood within the genus.

Chloroplast genome evolution in *Striga*:

The chloroplast genome contains genes necessary for a plant's autotrophic life style. Studies of hemi- and holo-parasitic plants have shown that many loci within the chloroplast genome have been lost or are evolving under relaxed selective pressures, while some remain under purifying selection (Wolfe *et al.*, 1992; Wolfe & dePamphilis, 1998; Young & dePamphilis, 2000; Young & dePamphilis, 2005). One holo-parasite, *Epifagus virginiana,* has a chloroplast genome (~70kb) approximately half the size of a normal photosynthetic plant (~150kb), having lost all of the loci involved in photosynthesis, while retaining some of the 59 loci involved in gene expression (Wolfe *et al.*, 1992). Substitution rate studies indicate that the *rbc*L locus is evolving under relaxed selective pressures in many non-photosynthetic Orobanchaceae lineages (Wolfe & dePamphilis, 1998; Young & dePamphilis, 2005). In contrast, the *mat*K locus, known for its high rate of nucleotide substitution, has experienced a statistically significant level of amino acid conservation across the same lineages (Young & dePamphilis, 2000). These studies suggest that the change from autotrophic to a parasitic lifestyle has reshaped the chloroplast genome in these lineages. Loci no longer required are jettisoned, shrinking the genome. They also illuminate the fact that loci in the chloroplast have necessary functional roles in plants and other plastid-containing organisms beyond photosynthesis (Barbrook *et al.*, 2006). How nuclear genomes are affected by the parasitic lifestyle remains to be investigated. Before this question can be addressed, studies of the structural composition of parasitic plant genomes need to be conducted.

Review of population genetic studies in *Striga*:

Interest in the genus *Striga* is usually limited to the agriculturally important species. Several studies have been conducted to investigate the genetic diversity and population structure of the three most devastating weeds (*S. hermonthica*, *S. asiatica*, and *S. gesnerioides*) (Werth *et al.*, 1984; Bharathalakshmi *et al.*, 1990; Shawe & Ingrouille, 1993; Kuiper *et al.*, 1996; Olivier *et al.*, 1998; Koyama, 2000; Botanga *et al.*, 2002; Gethi *et al.*, 2005; Botanga & Timko, 2006). These studies used either allozyme or AFLP markers.

Five studies employed allozyme markers, which allow heterozygosity to be observed in *S. asiatica* and *S. hermonthica* (Werth *et al.*, 1984; Bharathalakshmi *et al.*, 1990). One study of the genetic diversity within a population of *S. asiatica* accidentally introduced into the United States found that all of the individuals were monomorphic for all 32 loci analysed. This suggested a single introduction and demonstrated the ability of autogamous *Striga* species to invade new habitats from great distance (Africa to the US) (Werth *et al.*, 1984). Two genetic diversity studies of *S. hermonthica* using either nine or two loci, found high genetic diversity with slight effects for isolation by distance and no effect by host species (Bharathalakshmi *et al.*, 1990; Olivier *et al.*, 1998). The remaining two studies used larger numbers of loci (14-41) and collected populations in west and eastern Africa that were found to exhibit high genetic diversity and isolation by distance effects (Kuiper *et al.*, 1996; Koyama, 2000).

Four published investigations used AFLP markers, where genetic distance can be measured by the presence or absence of bands, to pursue population studies of *S. hermonthica*, *S. gesnerioides*, and *S. asiatica* (Botanga *et al.*, 2002; Gethi *et al.*, 2005; Botanga & Timko, 2006; Ali *et al.*, 2009). The examinations of *S. asiatica* and *S. gesnerioides* indicated greater genetic distances between populations than within populations. This was an expected result for these inbreeding species, where one to a few individuals can start a new population (founder effect) and genetic drift allows them to diversify (Mayr, 1954; Botanga *et al.*, 2002; Gethi *et al.*, 2005). Botanga and Timko (2006) also demonstrated the genetic uniformity of an introduced population of *S. gesnerioides* in Florida. The studies that examined *S. hermonthica* found greater genetic distances within populations, rather than between populations (Gethi *et al.*, 2005; Ali *et al.*, 2009). One issue with AFLPs, which are scored only as dominant (presence/absence) markers, is the limited amount of analysis that can be conducted on the collected data, with none of these studies being able to calculate standard population genetic diversity statistics like allelic diversity (richness), effective allelic diversity (evenness), expected heterozygosity (gene diversity), observed heterozygosity, or F-statistics.

All of the previous studies, combined, suggest that the autogamous species (*S. asiatica* and *S. gesnerioides*) have low genetic diversity within populations and that geographic distance has an influence on genetic differentiation, most likely through genetic drift. The formation of host-specific strains/races was also documented. These studies also suggest that the allogamous species (*S. hermonthica*) has a great deal of genetic diversity and that gene flow between populations reduces genetic differentiation, with an isolation by distance effect only seen at great geographic distances (>1000 km). None of these studies convincingly demonstrated host-specific races of *S. hermonthica* (Ali *et al.*, 2009).

An investigation into the phylogenetic relationships within the genus *Striga* using herbarium preserved specimens

In chapter two, we describe our approach to characterizing relationships within the genus *Striga*, in hopes that understanding these evolutionary relationships will help in the fight against witchweeds worldwide. Due to restrictions implemented by the US government, fresh Striga tissue is difficult to obtain in the US, except for the five weedy species. In order to overcome this hurdle, we chose to explore the use of herbarium-preserved specimens as a source for DNA. While many plant systematic projects have used herbarium-stored specimens, it is generally considered a last resort and used sparingly, due to common DNA degradation (mutation) processes that occur after death (Kigawa *et al.*, 2003). The largest collection of *Striga* species in the US is maintained under the direction of Dr. Lytton Musselman at Old Dominion University in Norfolk, Virginia. This valuable resource was used as the main source of specimens for this research. The results of this investigation allow us to describe a hypothesis of relationships with the genus based on six chloroplast loci. This hypothesis is distinctly different from previous hypotheses of relationships based on morphological characters and should be further explored with laboratory-grown plants as a source for DNA.

An investigation into the nuclear genome structures of weedy species of *Striga*

Over the last decade, the genomics revolution has given scientists a new vantage point on how genomes evolve. Focusing on a global genomic approach rather than the traditional locus-by-locus approach of genetics, incorporating cladistic analyses vetted by systematists, using high throughput informatics applications, and the tremendous decrease in time and cost of sequencing have combined to propel this field into the forefront of modern science. One area of research that has benefited from genomics is our understanding of genome size change. Plant biologists are pioneers in this field. All of the major processes known to influence genome size have been identified. Other than shaping the genomes of plants, these processes may also drive speciation by generating nuclear incompatibility. In particular, the process of transposable element amplification has been shown to be the key factor increasing genome size and illegitimate recombination has been found to be the major process that shrinks genome size (Bennetzen, 2007).

Compared to the genomes of other flowering plant species, the maize genome is of less than medium size at ~2400 Mb with a complex organization of genes surrounded by “seas” of nested LTR retrotransposons (LRPs) (SanMiguel *et al.*, 1996). Several studies have concluded that repetitive DNA composes 60-80% of the maize genome and that class I transposable elements (retroelements) constitute the largest portion of the repetitive DNA (Meyers *et al.*, 2001; Whitelaw *et al.*, 2003; Messing *et al.*, 2004; Baucom *et al.*, 2009). Investigations into the repeat content of the diploid species of *Zea* have also shown that while all species examined contain the same families of LRPs, there are large variations in the abundance of individual families between genomes (Estep and Bennetzen, unpublished). By comparing orthologous regions in sorghum, rice, and maize, several studies have shown that the presence or absence of repetitive elements constitutes the major differences between these genomes in the regions studied (Ilic *et al.*, 2003; Lai *et al.*, 2004; Ma *et al.*, 2005). One study comparing the genomes of species in the Oryza genus has shown that the activity of three LRPs are responsible for the genome size increase seen in O. australiensis (wild relative of cultivated rice) in comparison to other species in the genus (Piegu *et al.*, 2006). A more recent investigation into genome size evolution within the genus *Gossipium* has shown lineage-specific amplifications of LRPs similar to that observed in *Zea.* In addition, one family of *gypsy*-like LRP (Gorge3) has undergone a massive amplification in the two species with the largest genomes (Hawkins *et al.*, 2006). Taken together, these studies indicate that LRPs are one of the most labile structures in plant genomes and also show that sample sequencing techniques (either BAC end or shotgun sequences) are a valuable tool in describing the structural composition of genomes. In Chapter 3, we use a sample sequencing approach to describe the highly repetitive structures within the nuclear genomes of the five agricultural weeds within the genus (*S. hermonthica*, *S. asiatica*, *S. gesnerioides*, *S. forbesii*, and *S. aspera*). This comparative analysis begins to shed light on the highly repetitive portion of the genome and how these structures have changed in abundance across distinct lineages of the genus.

An investigation of genetic diversity and population structure in *Striga hermonthica*

S. hermonthica is a devastating agricultural weed that parasitizes grain crops, such as sorghum (*Sorghum bicolor*) and millet (*Pennisetum glaucum*) throughout Sub-Saharan Africa. In chapter four, I describe my development and use of a new set of neutrally-evolving microsatellite markers for *S. hermonthica,* discovered in sequence data generated in chapter 3. These markers are highly variable and can be used by multiple researchers, so that future studies can be directly compared. In chapter 5, these markers were used to investigate the genetic diversity and population structure of 11 populations in finer detail than previous studies. Our results show that *S. hermonthica* is genetically diverse and that gene flow between populations is apparently high enough to reduce the effects of genetic drift, resulting in an even distribution of diversity with little genetic differentiation across the country of Mali.

CHAPTER 2

PHYLOGENETIC ANALYSIS OF THE GENUS *STRIGA*, USING A MIXTURE OF FRESH AND HERBARIUM SPECIMENS.[1]

[1] Estep, M. and Bennetzen, J. To be submitted to *Plant Systematics and Evolution.*

Abstract:

The use of herbarium tissue as a source for phylogenetic analysis is explored in the genus *Striga* (Orobanchaceae), a group of parasitic plants found mostly in Africa and parts of Asia, that contains five agriculturally important species. Using sequence data from six chloroplast loci arrayed in a super-matrix, a phylogenetic hypothesis was developed. The resulting hypothesis is distinctly different from a hypothesis based on morphological characters and includes three clades within the genus each containing at least one agricultural weedy species.

Introduction:

Striga is a genus of mostly hemi-parasitic plants within the family Orobanchaceae (Olmstead *et al.*, 2001). Most of the species are found in Africa with a reduced number also found in southern Europe and parts of Asia (Mohamed *et al.*, 2001). The genus contains twenty-eight described species with several sub-species designations and likely contains several undescribed members (Mohamed *et al.*, 2001). Five of the twenty-eight described species are known as agricultural weeds, and are therefore commonly referred to as witchweeds. The "weedy" species can be found in agricultural fields of subsistence farmers parasitizing sorghum, pearl millet, rice, maize, and other grain crops (*S. asiatica, S. aspera*, *S. forbesii* and *S. hermonthica*) or on cowpea and other legumes (*S. gesnerioides*), particularly in sub-Saharan Africa (Musselman, 1987). In these habitats the *Striga* populations can reach densities of thousands of individuals per hectare (Van Mourik, 2007). The remaining 23 described species can be found in natural grassland habitats in much smaller populations (Mohamed *et al.*, 2001).

Naturalized populations of at least two species, *S. asiatica* and *S. gesnerioides* have been found within the United States, likely introduced on contaminated plants or seeds from Africa (Musselman, 1987; Botanga & Timko, 2006). The USDA characterizes this genus as a “noxious pest”, restricting the importation of plant material and requiring special permitting and expensive containment facilities for research purposes. While this designation acts to protect agriculture within the US, it has also restricted researchers’ ability to fully investigate the genus.

In this paper we describe our approach to determine the relatedness of species within the genus *Striga*, in hopes that understanding these evolutionary relationships will help in the fight against witchweeds worldwide. Due to the restrictions implemented by the US government and the absence of a current nursery with many live *Striga* species or the seed of many Striga species, fresh tissue was obtained only for the five weedy species. In order to overcome this hurdle, we chose to explore the use of herbarium-preserved specimens as a source of DNA. While many plant systematic projects have used herbarium-stored specimens, it is used only when necessary, due to common DNA degradation (mutation) processes that occur after tissue death (Kigawa *et al.*, 2003). The largest collection of *Striga* herbarium samples in the US is maintained under the direction of Dr. Lytton Musselman at Old Dominion University in Norfolk, Virginia. This valuable resource was used as the main source of specimens for this research.

Methods:

Plant tissue sources:

Herbarium tissue samples were collected from the Old Dominion University (ODU) Department of Biological Sciences herbarium, or New York Botanical Gardens (NYBG) herbarium. All passport information was transcribed from each individual sample, including locality of collection, collection year, species identification, collector, and accession numbers (Table 2-1). Some freshly grown, quick frozen tissue was available from the University of Virginia (UVA) quarantine facility (APHIS Plant Protection and Quarantine Permit No. 70902-P) for the five species that are considered agricultural weeds (Table 1). Live specimens of *Buchnera sp.* and *Aureolaria sp.* were sampled from the Plant Biology teaching collection at the University of Georgia (UGA) (maintained and provided by Mike Boyd).

Preparation of DNA samples:

Tissue was pulverized in 2ml tubes with 2 metal beads using a TissueLyzer (Qiagen, Valencia, CA). 500ul of extraction buffer (1M Tris, 1.4M NaCl, 100mM EDTA, 2% CTAB, 2% PEG-8000, pH 9.5) was added to each ground sample and incubated at 65C for 30 minutes. Samples were then chilled on ice for 5 minutes and 1 volume of chloroform:isoamyl alcohol (24:1) (CIA) was added. Samples were mixed for 5 minutes followed by centrifugation at 12,000g for 10 minutes. The aqueous layer was transferred to a clean 1.5ml tube and the CIA extraction was repeated a second time. DNA was precipitated by adding 1 volume of isopropanol and centrifugation at 12,000g for 10 minutes. The DNA pellet was washed with 70% ethanol and air dried for 5 minutes. The DNA was then re-suspended overnight in 200 ul of 1xTE buffer (10mM Tris, 5mM EDTA, pH 8.0).

PCR conditions and loci amplified:

Six genetic loci, known to have utility in phylogenetic reconstruction, were chosen from the chloroplast genome (Olmstead & Palmer, 1994; Shaw *et al.*, 2005). Two of these loci are protein encoding; *Rbcl*-ribulose-bisphosphate carboxylase, *MatK*-MaturaseK. The remaining four loci are non-coding (inter-genic regions) with primers anchored in conserved coding regions; *trnT*(UGU) - *trnL*(UAA), *trnD*(GUC)-*trnT*(GGU), *rpoB*- *trnC*(GCA), and *trnH*(GUG)-*psbA* (Table 2-2.). PCR reactions were 10 ul. Each reaction consisted of 10 ng of template DNA, 0.6 U of *Taq*, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTP's (each), and 2.5 mM of both primers. A touchdown PCR program was used on a MJ Research PTC-200 Peltier Thermocycler, consisting of an initial denaturation cycle of 94 °C for 5 min; 10 cycles at 94 °C for 45 sec, 68 °C (-2 °C per cycle) for 5 min, elongation at 72 °C for 1 min; 5 cycles at 94 °C for 45 sec, 58 °C for 2 min, elongation at 72 °C for 1 min; 25 cycles of 94 °C for 45 s, 50 °C for 2 min, elongation at 72 °C for 1 min; and completing with a final 30 min elongation at 72 °C. The resulting PCR products were gel extracted using the PureLink gel extraction kit, following the manufacturer's protocol (Invitrogen, Carlsbad CA). Cleaned PCR products were directly sequenced in 10ul reactions using the BigDye Terminator v3.1

cycle sequencing kit following the manufacturer's protocol (Applied Biosystems, Foster City, CA). Each sequencing reaction consisted of 5ng of PCR product, 2.5 mM forward or reverse primer, 0.5M Betaine and the remaining kit components. A PCR program was used on a MJ Research PTC-200 Peltier Thermocycler, consisting of 44 cycles with a denaturation step of 96 °C for 1 min, an annealing step of 53 °C for 30 sec, and an elongation step at 60 °C for 4 min. The resulting sequencing product was cleaned with Sephadex (G-50) Columns and separated on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA). Resultant chromatograms were imported to Sequencher (4.7) and trimmed based on the quality of the sequence read.

A second group of primers was designed for each locus from aligned sequences generated using the original primer set and DNA generated from frozen tissue (samples F52-F60). These primers were intended to amplify 100-300bp PCR products from herbarium tissue isolated DNA and were positioned around informative characters for phylogenetic reconstruction. These include *rbcl_2* (internal region of *Rbcl*), *MatK_2* (3' end of MatK), *trnT*(UGU) - *trnL*(UAA)_2 (3' end of *trnT*(UGU) - *trnL*(UAA)), *trnT*(UGU) - *trnL*(UAA)_3 (internal region of *trnT*(UGU) - *trnL*(UAA)), *trnD*(GUC)-*trnT*(GGU)_2 (internal region of *trnD*(GUC)-*trnT*(GGU)), *trnD*(GUC)-*trnT*(GGU)_3 (internal region of *trnD*(GUC)-*trnT*(GGU)), *trnD*(GUC)-*trnT*(GGU)_4 (3' region of *trnD*(GUC)-*trnT*(GGU)), *rpoB*- *trnC*(GCA)_2 (internal region of *rpoB*- *trnC*(GCA)), *rpoB*- *trnC*(GCA)_3 (internal region of *rpoB*- *trnC*(GCA)), *trnH*(GUG)-*psbA*_2 (5' region of *trnH*(GUG)-*psbA*), and *trnH*(GUG)-*psbA*_3 (3' region of *trnH*(GUG)-*psbA*) (Table 2-2.). The same PCR and sequencing conditions and procedures were used except that herbarium DNA samples were serially diluted (1:1, 1:10, 1:20, 1:50, 1:100, and 1:200) and at least threes attempt were made to amplify products from each dilution.

<u>Sequence alignment:</u>

Once each sequence was trimmed and cleaned, a series of Blast analyses were performed using BLASTN, with the database option set to "other – nucleotide collection (nr/nt)" and restricted to flowering plants (taxid: 3398) to identify and remove contaminated sequences. Sequences were then aligned by locus with the alignment algorithm Muscle implemented in SeaView (Galtier *et al.*, 1996). Each alignment was inspected and trimmed so that the majority of sequences overlapped. Individual alignments were concatenated into one super-matrix. The super-matrix was then edited to produce three different datasets with varying numbers of taxa based on the amount of missing data (Missing 1 locus or less, Missing 1-4 loci, and Missing more than 4 loci).

<u>Phylogenetic analysis:</u>

Each individual locus and the three forms of the super-matrix were analyzed in PAUP (4.0b10), using a full heuristic search with the optimality criterion set to parsimony, 1000 bootstrap replicates, and the *Buchnera sp.* and *Aureolaria sp.* sequences set as out-groups to evaluate relationships within the genus *Striga* (Swafford, 2002). The three super-matrices were further analyzed by a Maximum-Likelihood statistic implemented in RAxML (Stamatakis, 2006). Individual alignments were uploaded to the

CIPRES Portal and converted to the Phylip-relaxed format with NCLconverter (Miller *et al.*, 2010). The newly formatted alignment was entered into RAxML and run using the default options for a nucleotide alignment with bootstrap analysis, and the *Buchnera sp.* and *Aureolaria sp.* sequences set as out-groups.

Pairwise comparisons of herbarium and frozen tissue sequences:

All sequences generated from herbarium samples that also had a frozen tissue sample of the same species were compared to evaluate the effect of herbarium storage on sequence quality and accuracy. In each alignment, differences were counted and grouped by the type of base pair change (Transitions (Type-1 or Type-2), Transversions) (Table 2-3).

Results:

A total of 35 tissue samples were collected from frozen or herbarium tissue (Table 2-1). Full-length sequences were produced for all of the fresh-frozen tissue samples using published primers (Table 2-2). An alignment of 4326 bp was generated for these ten samples and used to design *Striga*-specific primers in polymorphic regions that were suitable for amplification of herbarium tissue DNA. Using these *Striga*-specific primers, at least one of the six chloroplast loci was amplified and sequenced from 32 of the collected samples (22 herbarium samples, 10 frozen samples). This included 19 *Striga* species and two out-group taxa. Sequences were not generated for three of the herbarium sampled taxa because PCR products could not be obtained (*S. gesnerioides* (H6), *S.*

macrantha (H15), and *S. linearifolis* (H18)). All sequences for a given PCR product were aligned and trimmed to produce nine separate alignments ranging in length from 97bp – 510bp. These include *rbcl_2* (243bp), *MatK_2* (510bp), *trnT*(UGU) - *trnL*(UAA)_2 (173bp), *trnD*(GUC)-*trnT*(GGU)_2 (97 bp), *trnD*(GUC)-*trnT*(GGU)_3 (371bp), *rpoB*-*trnC*(GCA)_2 (145bp), *rpoB*- *trnC*(GCA)_3 (165bp), *trnH*(GUG)-*psbA*_2 (277bp), and *trnH*(GUG)-*psbA*_3 (220bp). The individual alignments for each sample were then concatenated to generate an alignment of 2199bp. Attempts were made to amplify all nine PCR products from each sample, but the recalcitrant nature of the herbarium samples to PCR amplification caused some taxa to contain missing data.

The super-matrix (9 amplified regions aligned) generated is missing 34% of the alignment regions with many of the herbarium samples containing missing regions. Therefore the super-matrix was reduced to include taxa with at least five (56%) of the nine regions and missing 11% of the total alignment regions (middle-matrix). The mid-matrix contained a total of 21 samples representing 11 *Striga* species and two out-group species. The super-matrix was further reduced to only include taxa missing no more than one region (97% of regions present) and missing 3% of the total aligned regions (small-matrix). The small-matrix contained a total of 14 samples, representing eight *Striga* species and two out-group species.

Phylogenetic analysis:

Each of the individual locus alignments and the three forms of the super-matrix were analyzed for phylogenetic relationships using maximum parsimony (MP) performed in PAUP. All characters were coded as unordered, given equal weight, and gaps were treated as missing data. The two non-*Striga* taxa (*Buchnera americana* and *Aureolaria virginica*) were designated as out-groups. None of the single-region alignments produced well-supported phylogenies (data not shown). Therefore a maximum likelihood analysis was only conducted on the three forms of the super-matrix.

The small-matrix includes a total of 1855 constant characters, 257 variable parsimony-uninformative characters and 87 parsimony-informative characters. The maximum parsimony (MP) analysis of the small-matrix generated a well-supported phylogeny with a tree length of 390 steps, a consistency index (CI) of 0.93, a retention index (RI) of 0.85, and a homoplasy index (HI) of 0.06. There are four distinct terminal clades within the phylogeny (Fig 2-1). A basal clade (1) that includes *S. forbesii* and *S. pubiflora*, a second clade (2) including *S. asiatica*, *S. elegans*, and *S. lutea*, a third clade (3) that contains *S. aspera* and *S. gesnerioides*, and a fourth clade (4) containing *S. hermonthica*. Each of the four terminal clades is supported with a minimum bootstrap value of 98. The two clades that include *S. hermonthica*, *S. aspera*, and *S. gesnerioides* unite to form a large internal clade (5) supported with a bootstrap value of 99. This large clade is again included in a larger internal clade (6) also containing the *S. asiatica*, *S. elegans*, and *S. lutea* taxa, with a bootstrap value of 75. The maximum likelihood (ML) analysis produced a similar tree maintaining the four terminal clades identified in the parsimony analysis (Fig 2-2). There are two minor exceptions when comparing the MP and ML trees. The first is reduced bootstrap support for the largest internal clade (6) from 75 (MP) to 55 (ML) that includes *S. hermonthica*, *S. aspera*, *S. gesnerioides*, *S. asiatica*, *S. elegans*, and *S. lutea*. The second is a rearrangement of the relationships within the *S.*

asiatica, *S. elegans*, and *S. lutea* clade. It should also be noted that branch lengths of herbarium-sampled taxa are longer than those of frozen tissue samples (especially when comparing the same species).

The middle-matrix includes a total of 1736 constant characters, 334 variable parsimony-uninformative characters and 129 parsimony-informative characters. The MP analysis of the middle-matrix generated a well-supported phylogeny with a tree length of 597 steps, a CI of 0.86, a RI of 0.75, and a HI of 0.13. The resulting phylogeny includes the same four distinct clades and internal branching pattern as the small-matrix (Fig 2-3). The bootstrap support for these four clades is similar to the small-matrix, but slightly reduced. The first terminal clade (1) includes the same species as the small-matrix trees, but also includes *S. angloensis*, *S. latericea*, *S. masuria*, and a duplicate of *S. forbesii* (H24). The second (2), third (3), and forth (4) internal clades include the same species as in the small-matrix, but duplicate taxa were added. The ML analysis produces a similar tree and bootstrap support (Fig 2-4). There are some minor differences in the ML tree with respect to placement of taxa within the four major terminal clades and some reduction of bootstrap support. It is again important to note that branch lengths of some herbarium-sampled taxa are longer than would be expected.

The total-matrix includes a total of 1462 constant characters, 491 variable parsimony-uninformative characters and 246 parsimony-informative characters. The MP analysis did not generate a well-supported phylogeny with a tree length of 1151 steps, a CI of 0.81, a RI of 0.62, and a HI of 0.18. In both the MP and ML analyses, bootstrap support is greatly reduced and many defined clades are reduced to basal polytomies (Figs.

2-5 and 2-6). Exceptionally long branch lengths were generated for several of the herbarium-sampled taxa (Fig 2-6).

Effect of herbarium storage on sequence quality and accuracy:

A total of seven herbarium-sampled taxa sequences were compared to frozen tissue-sampled taxa sequences of the same species across the entire super-matrix (Table 2-3). The number of transitions (purines-purines or pyrimidines-pyrimidines) were recorded and ranged from 8-23 in different comparisons. These numbers were transformed into percent of total aligned bases and ranged from 0.55% - 3.7%. The number of transversions (purines – pyrimidines or vise versa) were recorded and ranged from 4-30 in different comparisons. These numbers were also transformed into percent of total aligned bases and ranged from 0.68% - 3.58%. The total number of transitions and transversions combined ranged between 1.36% and 5.35%. In most cases the number of transitions and transversions were approximately equivalent (based on percent of total aligned). In the *S. gesnerioides* (H7/F60) comparison, the number of transitions was approximately two times higher than transversions. The opposite pattern is seen in the *S. forbesii* comparison (H24/F59), where transversions are two times higher than transitions. Within the collection of herbarium specimens, the year of collection ranged from 1913 (*S. masuria* H19) to 2001 (*S. lutea* H13). The majority (76%) of herbarium specimens were collected in a period of six years, beginning in 1984 and ending in 1989. All of the herbarium specimens used in comparisons with frozen samples were collected within this active collection period, except *S. hermonthica* (H10) collected in 1978 (Table 2-1). An

XY scatter plot was generated for time since collection (21-32 years) and the percent of transitions/transversions (1.36-5.35%), yielding no significant correlations (data not shown). It was later discovered that all herbarium samples from ODU were probably treated for insect pests with methyl bromide, a known DNA mutagen (Kigawa *et al.*, 2003).

Discussion:

The impact that some species of *Striga* have on the local economies of subsistence agriculture in the developing world is extreme, while other species have remained parasites only of wild hosts. While many researchers have chosen to focus on the five weedy species, we have attempted to explore the evolutionary relationships within the genus. In this analysis, we have included as many distinct species of *Striga* as could be obtained without the time and expense of traveling to Africa and assembling new *Striga* collections. Instead, we chose to use mostly herbarium samples as a source for DNA.

Herbarium samples of *Striga* species are generally black to dark brown in color. When ground tissue is added to extraction buffer, a diverse set of colors were observed (red, purple, or black) and this trait was variable between and within taxa. Similar results were obtained with frozen tissues. The main difference observed between DNA extraction from the herbarium and fresh-frozen tissues was in the degree to which the color-causing compounds could be removed with polyethylene glycol and organic extraction from fresh-frozen tissue. The DNA from frozen samples was of high molecular weight and easily amplified with primers for products up to 2000bp. The herbarium samples yielded DNAs that were degraded, identified as a smear ranging from ~50-300bp in size upon agarose gel analysis (data not shown). Therefore, we designed primers that amplified 100-300bp products, strategically placed to capture the phylogenetic signal identified in an alignment of sequences from frozen tissue. Once we had tested the new primers on DNA from frozen samples, we discovered that the color-causing compounds also act as an inhibitor to PCR amplification. The best solution we found for this inhibitor was a series of dilutions, wherein a lower DNA (and thus contaminant) concentration often produced success in the PCR step.

The super-matrix of aligned sequences contained a good deal of missing data (34%). While missing data may seem like a weakness, it has been shown that the factors determining phylogenetic placement are the characters, which are present in the data matrix, and not the characters that are missing (Wiens, 2006). We resolved this issue by removing taxa with varying amounts of missing data and analyzed three data sets independently. We found that super-matrices containing between 3-11% missing data generated phylogenetic trees with the same internal branching pattern and clade formation. We noticed that the consistency index, retention index, and bootstrap support of the middle-matrix (11% missing) were reduced in comparison to the small-matrix (3% missing). The homoplasy index and tree length were both increased in the same comparison. The phylogenetic trees generated from the total super-matrix (34% missing data) followed the same trend of reduced support seen in the above comparison. This suggests that missing data was influencing our ability to construct an accurate phylogenetic tree, but did not cause inaccuracy in the trees that were generated.

We also noticed in the ML trees that branch lengths of the herbarium taxa were greater than those seen with DNA from fresh-frozen tissues (Figs. 2-2 and 2-4). Long branch lengths were also observed for the *Aureolaria* species used as an out-group. The long branch lengths in the out-group can be explained by genetic distance, but those from herbarium specimens are more likely to be an outcome of DNA degradation during storage. Using DNA sequences generated from frozen tissue as a control, we compared herbarium specimens from the same species to identify the cause of extended branch lengths. This analysis revealed that between 1.3 – 5.35% of the total aligned base pairs were different between individuals of the same species. This result suggests that the herbarium samples are accumulating mutations after death. These differences were further analyzed to investigate the frequency of cytosine deamination, a type of mutation reported as common in ancient DNA samples (Binladen *et al.*, 2006). Our analysis shows no distinct bias towards a particular type of mutation, but instead suggests that both transitions and transversions were equally likely. This is of particular interest, because the random after-death mutations act as noise in our analysis and, if a particular type of mutation was identified as age-related, those characters could be removed or ignored during analysis. It is important to note that, while these DNA sequence changes act as noise, the true phylogenetic signal remains within our alignments. This conclusion is supported by the high bootstrap support obtained for clades when a single species was sampled with both tissue types (*S. hermonthica*, *S. asiatica*, and *S. forbesii*) (Fig 2-4).

The results of this analysis allow us to hypothesize that the genus *Striga* has three distinct clades. The first is a basal lineage that contains the largest number of sampled species: *S. forbesii*, *S. angloensis*, *S. latericea*, *S. masuria*, and *S. pubiflora*. A second clade containing at least three species, *S. asiatica*, *S. elegans*, and *S. lutea.* A third clade includes *S. hermonthica*, *S. gesnerioides*, and *S. aspera*. This result argues that weediness (life as an agricultural pest) has evolved at least 3 times within this genus, leading to *S. forbesii*, to *S. asiatica*, and to the *S.hermonthica/S.gesnerioides/S.aspera* clade (Fig 2-7).

A detailed treatment of the taxonomy within the genus has been described and a few of the relationships have been proposed (Mohamed, 1994; Mohamed *et al.*, 1996; Mohamed *et al.*, 2001). Most of this knowledge was derived from a cladistic analysis of morphological characters or field observations (Safa *et al.*, 1984; Mohamed, 1994; Kuiper *et al.*, 1996). Mohamed and coworkers (1996) hypothesized a *S. asiatica* clade that includes *S. elegans*, *S. lutea*, and *S. asiatica*, and our study agrees with this model. We would also add to this clade *S. hirsute,* which was placed elsewhere in the same study. The remaining clusters described in the morphological analysis (with many more species analyzed) are not supported by the molecular analysis presented here. The most notable of these disagreements is the *S. hermonthica*, *S. aspera*, and *S. gesnerioides* clade (Fig 2-7).

Traditionally, the species *S. hermonthica* and *S. aspera* have been treated as sister taxa, with some researchers proposing that they might represent a single species because of their ability to hybridize (Kuiper *et al.*, 1996; Aigbokhan *et al.*, 2000). *S. gesnerioides* has also traditionally been separated from all other *Striga* species and considered a basal lineage within the genus, based on host preference and its tendency towards holo-parasitism (loss of chlorophyll) and reduced morphological characters (Musselman, 1987; Mohamed *et al.*, 1996). Our analysis suggests that *S. aspera* and *S. gesnerioides* are sister taxa and that *S. hermonthica* is sister to this clade. This discovery indicates that the switch to a very different host species in the progenitor of *S. gesnerioidies* was concurrent with a very rapid change in morphological characteristics and lifestyle, perhaps caused by adaptation to a dramatically different host.

The predicted recent derivation and hyper-evolution of the *S. gesnerioides* lineage will likely be contentious and should be further explored. One fascinating approach would be to cross *S. gesnerioides* with *S. aspera*, similar to work published on the *S. hermonthica*/ *S. aspera* relationship (Aigbokhan *et al.*, 2000). The resulting F1 plants could be self-pollinated to generate a mapping population that could be useful in dissecting the genetic loci involved in host preference and the numerous other morphological, physiological and developmental novelties that differentiate these two closely-related species.

This analysis, based mostly on herbarium tissue, highlights the need for a seed collection from all *Striga* species that could be utilized to generate high quality DNA for a larger and more detailed analysis of evolutionary relationships within *Striga*. In this larger analysis, whole plastid genome sequencing would greatly increase the number of character states and would allow missing data (gene loss) to act as informative characters. The use of nuclear DNAs will also be vital, to help confirm the phylogenies based only on organellar DNA, and to help generate a molecular clock that will allow assignment of dates to the divergence process.

Table 2-1. List of species sampled with collection information.

Taxon	Sample ID	Collector	Location	Collection date	Herbarium
S. angloensis	H25	R.G.N. Young	Angola	1937	ODU - 1261
S. angustifolia	H21	L. J. Musselman	India	1985	ODU - 7077
S. asiatica	H2	Safa & Musselman	Burkina Faso	1987	ODU
S. aspera	H1	Musselman & Mohamed	Mali	1988	ODU
S. brachycalyx	H3	Safa & Musselman	Burkina Faso	1987	ODU
S. densiflora	H4	L. J. Musselman	India	1985	ODU
S. densiflora	H23	L. J. Musselman	India	1985	ODU
S. elegans	H20	N. K. Hughes	Zimbabwe	NA	ODU
S. elegans	H5	N. K. Hughes	Zimabwe	1985	ODU - SA30
S. forbesii	H24	D. Knepper	Zimbabwe	1987	ODU
S. gesnerioides on Ipomea	H6	Musselman & Mohamed	Mali	1988	ODU
S. gesnerioides on native plants	H7	D. Knepper	Nigeria	1989	ODU
S. gesnerioides on Tephrosia	H8	Safa & Musselman	Burkina Faso	1987	ODU
S. hermonthica on Sorghum	H9	D. Knepper	Nigeria	1989	ODU
S. hermonthica on Sorghum	H10	Musselman & Mansfield	Nigeria	1978	ODU
S. junodii	H12	Zietsman & Bronkhorst	South Africa	2000	NYBG - 764786
S. klingii	H14	L. J. Musselman	Cameroon	1988	ODU
S. latericea	H16	L. J. Musselman	Ethiopia	1985	ODU
S. ledermanii = S. bilabiata ledermanii	H17	L. J. Musselman	Cameroon	1988	ODU
S. linearifolia = S. bilabiata linearifolia	H18	Safa & Musselman	Burkina Faso	1987	ODU
S. lutea	H13	F. Bradley et al.	Gabon	2001	ODU - 992
S. macrantha	H15	L. J. Musselman	Nigeria	1978	ODU
S. masuria	H19	G Forrest	China	1913	ODU
S. passargei	H22	Safa & Musselman	Burkina Faso	1987	ODU
S. pubiflora	H11	Tanner	Tanzania	1956	NYBG
Aureolaria sp	F57	M. Estep	Georgia, USA	2008	UGA
Buchnera sp.	F56	J. McNeal	Florida, USA	NA	UGA
S. asiatica	F58	Unknown	Tanzania	1992	UVA
S. aspera	F52	Unknown	Mali	1984	UVA
S. forbesii	F59	Unknown	Zimbabwe	1984	UVA
S. gesnerioides	454	Unknown	Benin	2008	UVA
S. gesnerioides	F60	Unknown	Benin	2008	UVA
S. hermonthica	F53	Unknown	Ivory Coast	NA	UVA
S. hermonthica	F54	Unknown	Cameroon	NA	UVA
S. hermonthica	F55	Unknown	Sudan	1995	UVA

Table 2-2. Primer sequences used to amplify PCR products and for sequencing.

Locus	Forward (5'-3')	Reverse (5'-3')
	Full length PCR product primers	
Rbcl	5’ATGTCACCACAAACAGAAACTAAAGC	5’ CGCAGTAAATCAACAAAGCCCA
MatK	5’CTTGTTTTGRCTNTATCGCACTATG	5’CTTTTGTGTTTCCGAGCYAAAGTT
trnT$_{\text{(UGU)}}$ - *trnL*$_{\text{(UAA)}}$	5’CATTACAAATGCGATGCTCT	5’ GGGGATAGAGGGACTTGAAC
trnD$_{\text{(GUC)}}$-*trnT*$_{\text{(GGU)}}$	5’ACCAATTGAACTACAATCCC	5’ CTACCACTGAGTTAAAAGGG
rpoB- *trnC*$_{\text{(GCA)}}$	5’CACCCRGATTYGAACTGGGG	5’ CKACAAAAYCCYTCRAATTG
trnH$_{\text{(GUG)}}$-*psbA*	5’CGCGCATGGTGGATTCACAATCC	5’ GTTATGCATGAACGTAATGCTC
	Herbarium PCR product primers	
rbcl_2	5'CTTGGCAGCATTCCGAGTA	5'AGCACGCAAGGCTTTGAATC
MatK_2	5'CTTTCAAGCTTGCGAATGAA	5’CTTTTGTGTTTCCGAGCYAAAGTT
trnT$_{\text{(UGU)}}$ - *trnL*$_{\text{(UAA)}}$_2	5'CCCCTTACGAATGAAAGCTG	5’ GGGGATAGAGGGACTTGAAC
trnT$_{\text{(UGU)}}$ - *trnL*$_{\text{(UAA)}}$_3	5'AGAACCGCTTCCATTGAGTC	5'TCGGTAGACGCTACGGACTT
trnD$_{\text{(GUC)}}$-*trnT*$_{\text{(GGU)}}$_2	5'AATCCCAGGGGACTAAAGGA	5'CGTTGGCAATATGTCTACGC
trnD$_{\text{(GUC)}}$-*trnT*$_{\text{(GGU)}}$_3	5'TTGCCAACGAATTTACAGTCC	5'CCTGGGGGTAGGGTACTACG
trnD$_{\text{(GUC)}}$-*trnT*$_{\text{(GGU)}}$_4	5'CCGAGGGATCTTTCCGTTT	5’ CTACCACTGAGTTAAAAGGG
rpoB- *trnC*$_{\text{(GCA)}}$_2	5'GCCGCCAAAATGATACAAAA	5'GCGGGAGAGTGTTTTTAGCA
rpoB- *trnC*$_{\text{(GCA)}}$_3	5'TGTCAACCCCAGAACAGAAA	5'TCGAAGCTGATTTGAAGAAGG
trnH$_{\text{(GUG)}}$-*psbA*_2	5’CGCGCATGGTGGATTCACAATCC	5'TCCTCGACTTTTTGATTTCCA
trnH$_{\text{(GUG)}}$-*psbA*_3	5'TGGAAATCAAAAAGTCGAGGA	5’ GTTATGCATGAACGTAATGCTC

Table 2-3. Pairwise comparisons of herbarium tissue samples and frozen tissue samples. The species name is followed by the sample ID.

	***S. asiatica* H2/F58**		***S. aspera* H1/F52**		***S. forbesii* H24/F59**		***S. gesnerioides* H7/F60**		***S. gesnerioides* H8/F60**		***S. hermonthica* H9/H53-55**		***S. hermonthica* H10/H53-55**	
	# of bp	% of total bp	# of bp	% of total bp	# of bp	% of total bp	# of bp	% of total bp	# of bp	% of total bp	# of bp	% of total bp	# of bp	% of total bp
Transitions	21	**1.7**	10	**0.55**	8	**1.51**	9	**3.7**	9	**1.55**	23	**1.51**	9	**0.68**
Type-1	12	0.95	7	0.39	2	0.38	5	2.06	6	1.03	7	0.46	5	0.38
Type-2	9	0.72	3	0.17	6	1.13	4	1.64	3	0.52	15	0.98	4	0.3
Transversions	28	**2.2**	15	**0.83**	19	**3.58**	4	**1.64**	9	**1.55**	30	**1.96**	9	**0.68**
Total	49	**3.9**	25	**1.38**	27	**5.09**	13	**5.35**	18	**3.1**	53	**3.47**	18	**1.36**

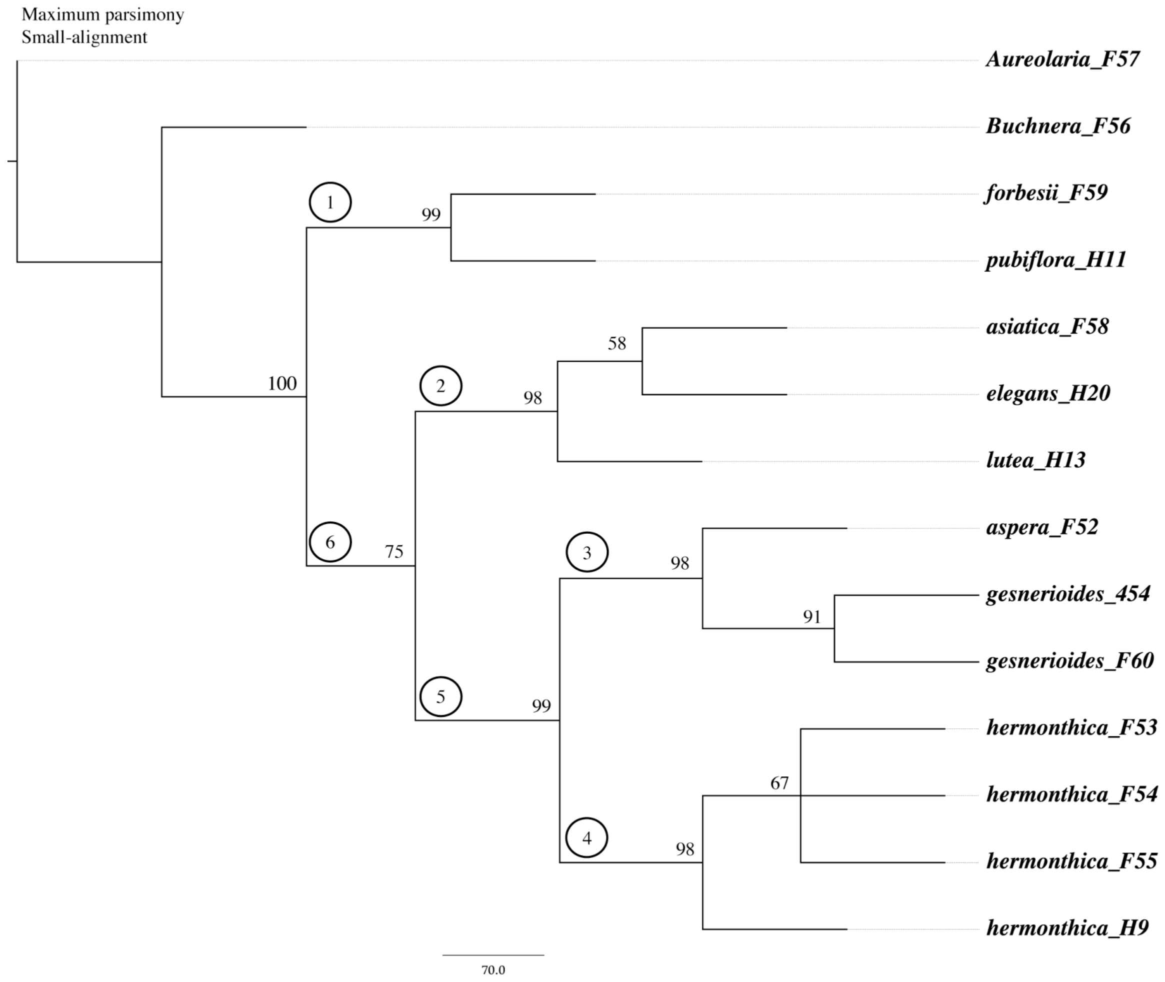


Figure 2-1. Maximum parsimony analysis of small-alignment. Bootstrap statistics are reported at branch nodes. Numbered circles are points of interest (see text)

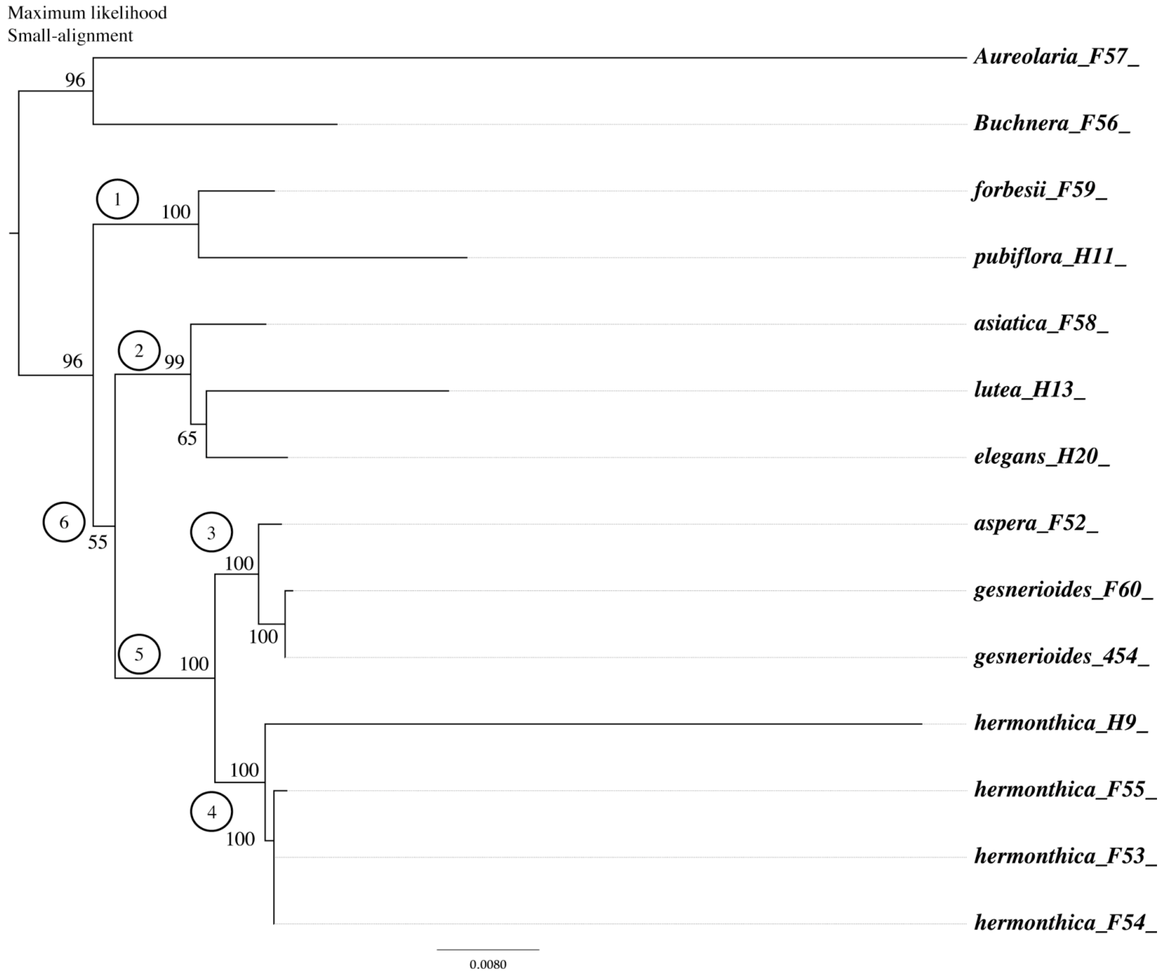


Figure 2-2. Maximum likelihood analysis of small-alignment. Bootstrap statistics are reported at branch nodes. Numbered circles are points of interest (see text).

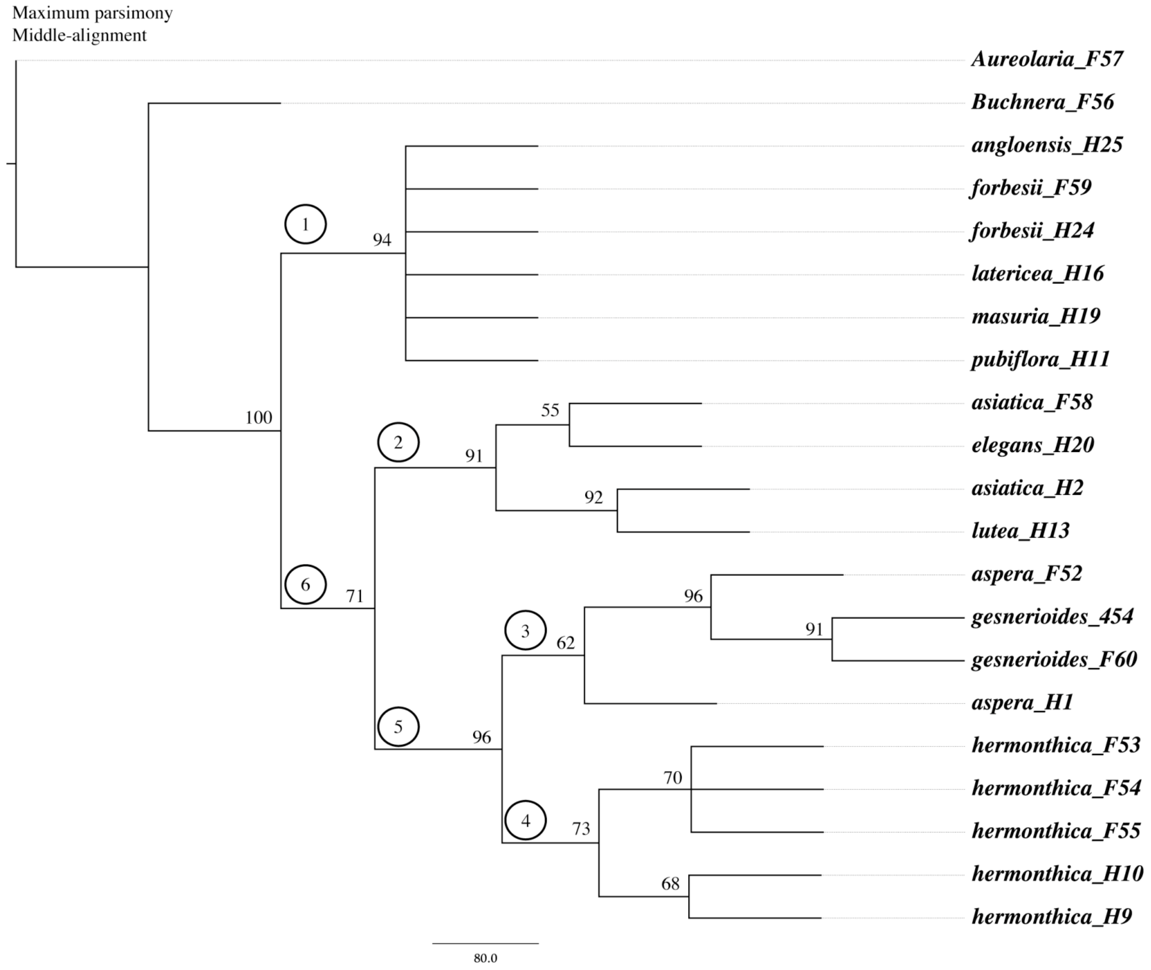


Figure 2-3. Maximum parsimony analysis of middle-alignment. Bootstrap statistics are reported at branch nodes. Numbered circles are points of interest (see text).

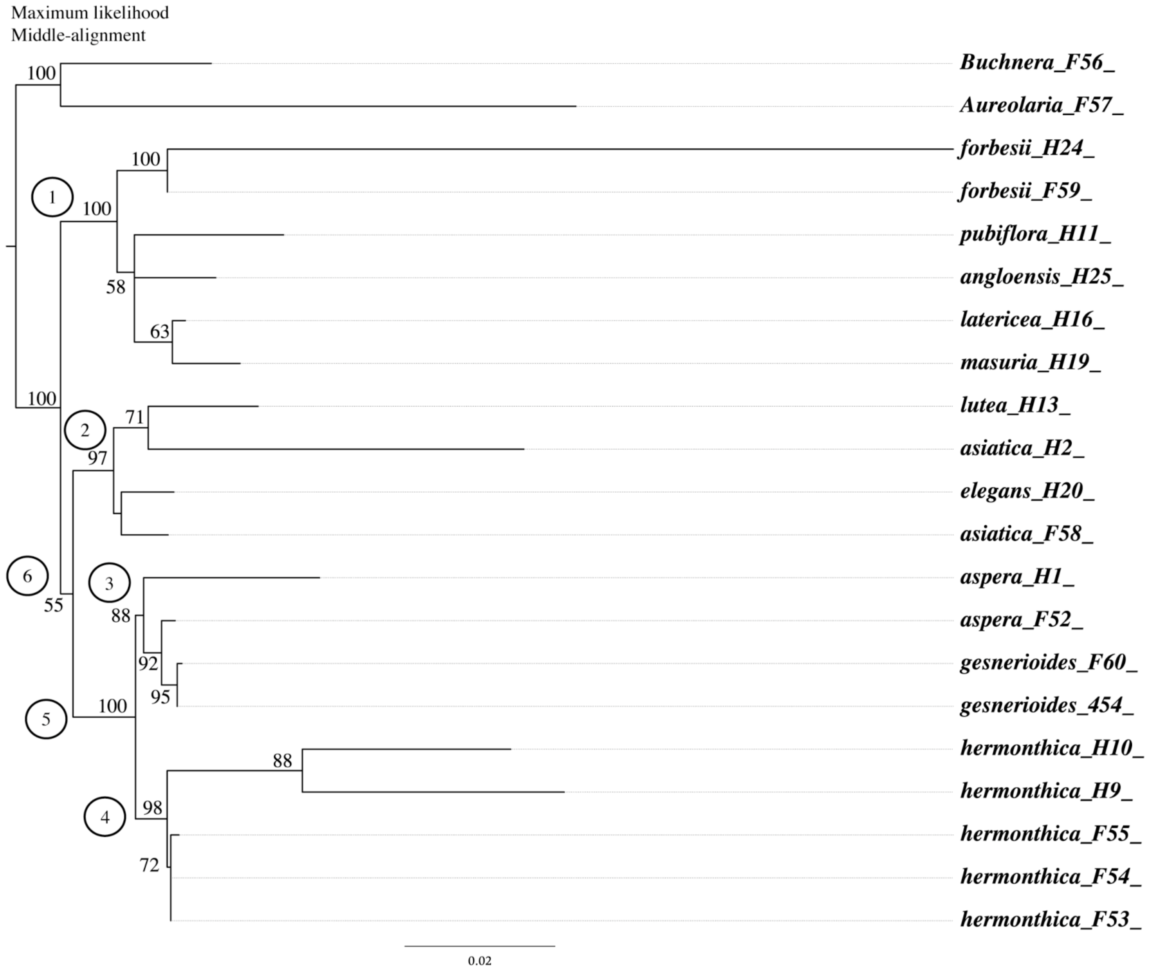


Figure 2-4. Maximum likelihood analysis of middle-alignment. Bootstrap statistics are reported at branch nodes. Numbered circles are points of interest (see text).

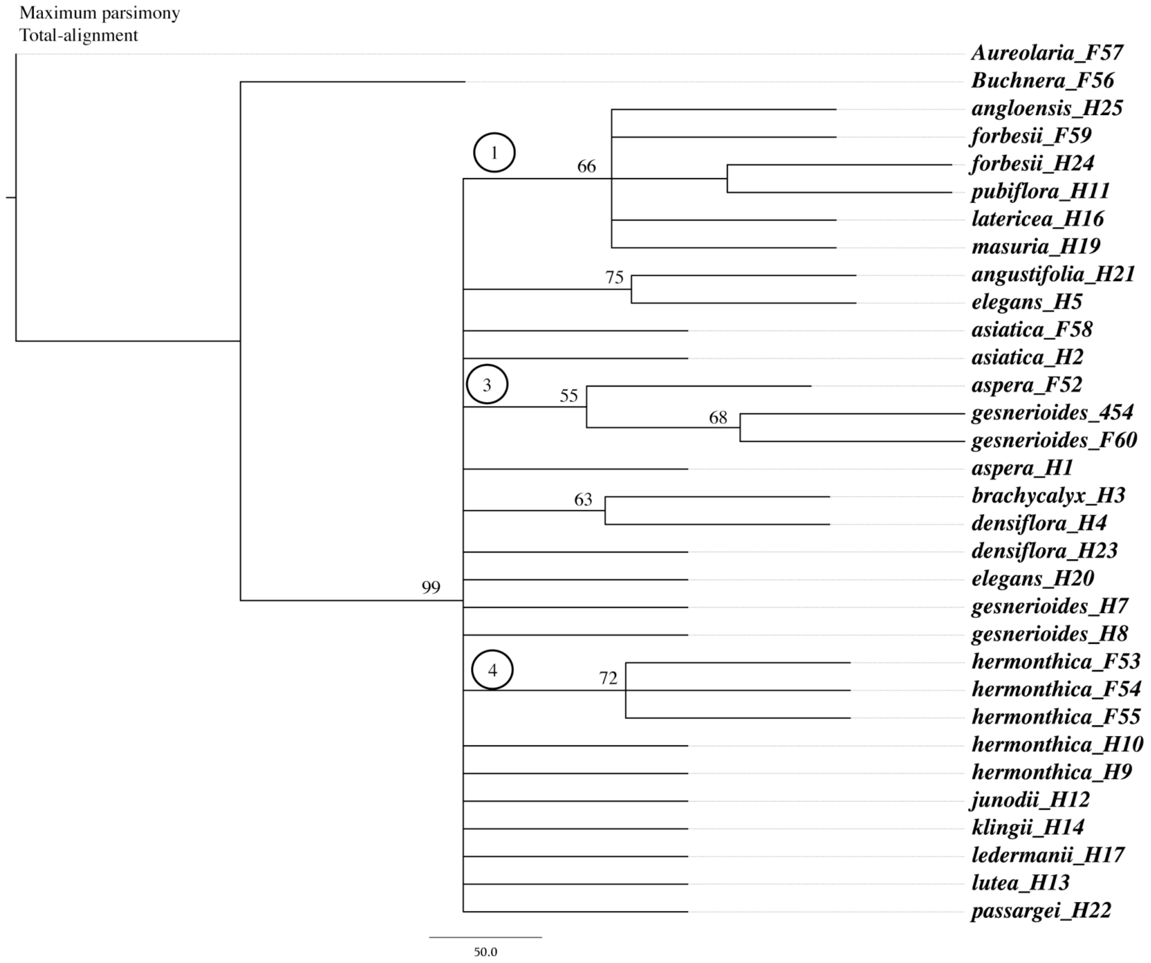


Figure 2-5. Maximum parsimony analysis of total-alignment. Bootstrap statistics are reported at branch nodes. Numbered circles are points of interest (see text).

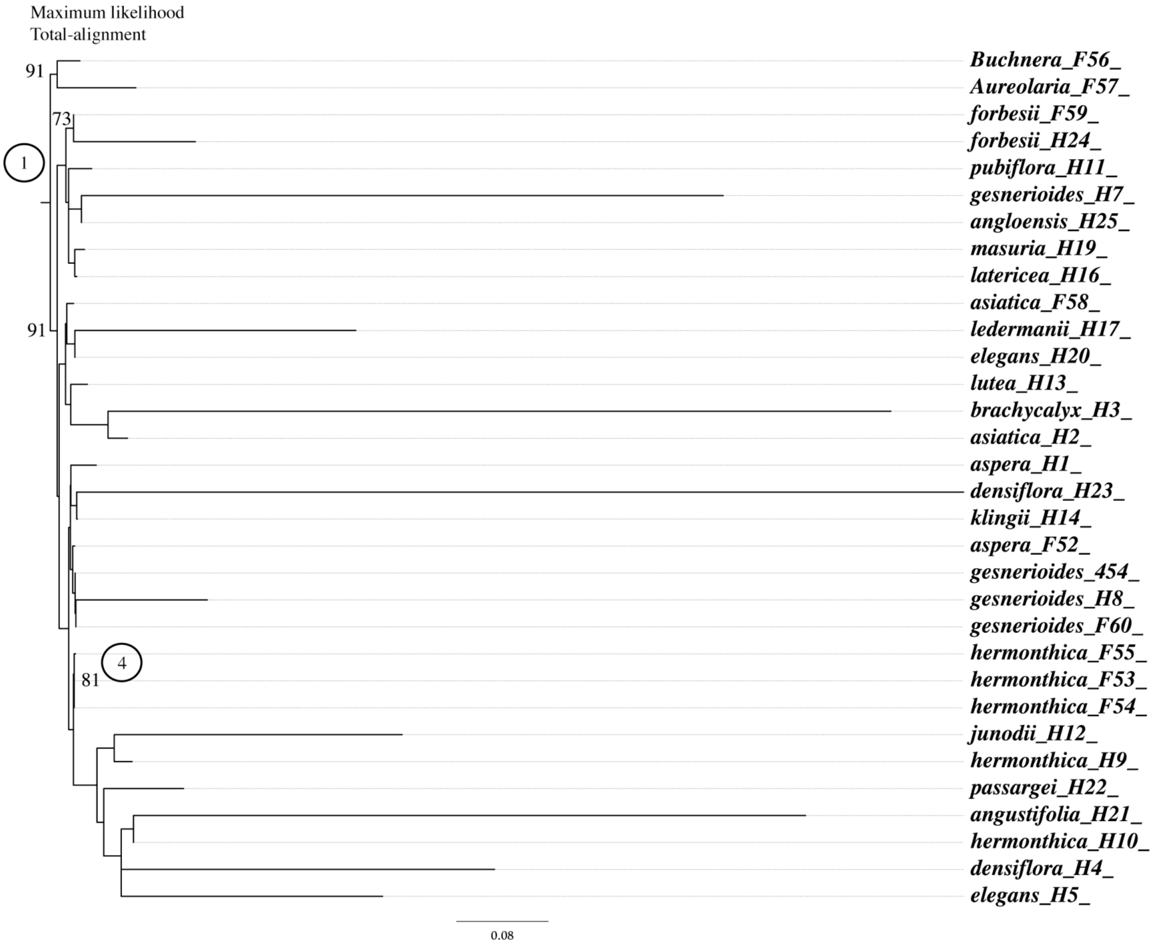


Figure 2-6. Maximum likelihood analysis of total-alignment. Bootstrap statistics are reported at branch nodes. Numbered circles are points of interest (see text).

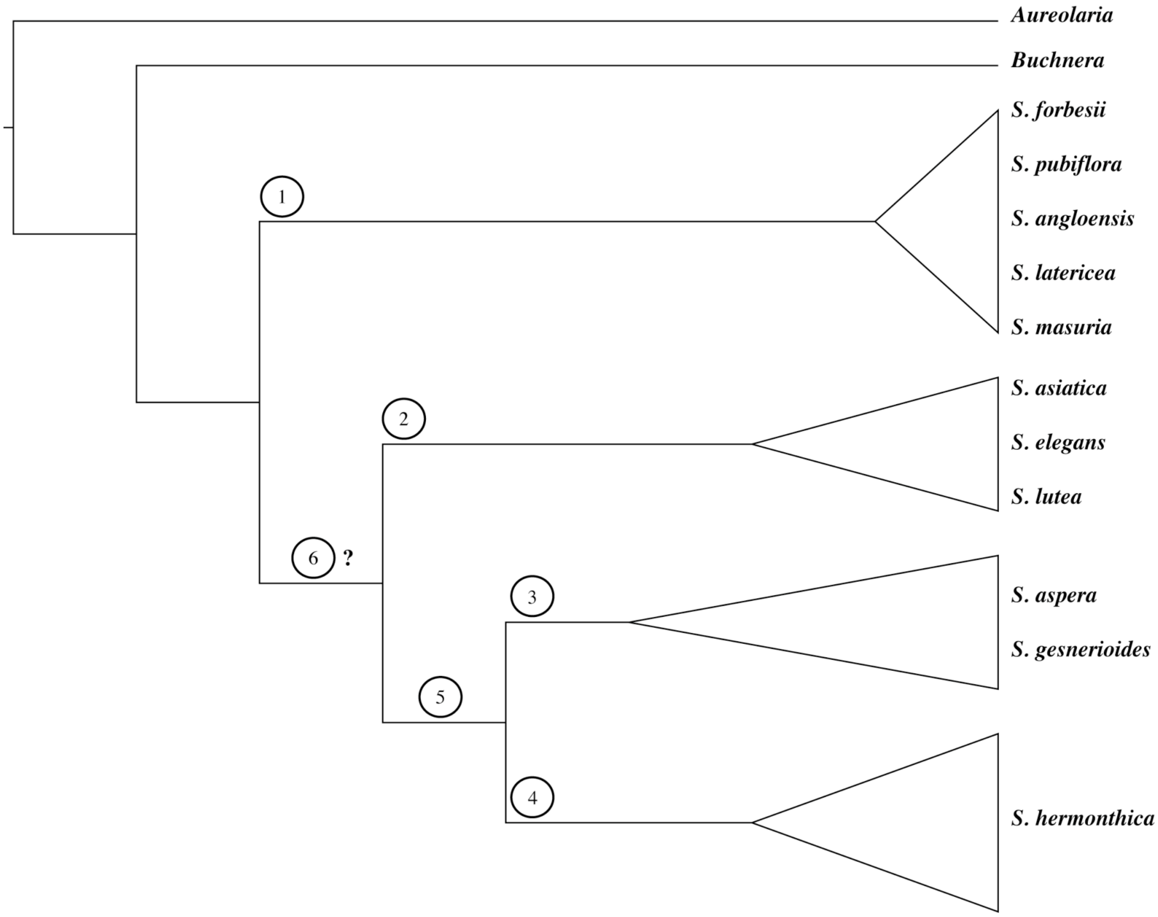


Figure 2-7: Phylogenetic hypothesis of the genus *Striga* based on six chloroplast loci.

CHAPTER 3

EFFICIENT GENOMIC CHARACTERIZATION FOR PARASITIC WEEDS OF THE GENUS *STRIGA* BY SAMPLE SEQUENCE ANALYSIS[2]

[2] Estep, M., Gowda, B., Huang, K., Timko, M., Bennetzen, J. To be submitted to *New Phytologist*

Summary:

- Sample sequence analysis (SSA) on randomly selected clones can describe the basic properties of complex plant genomes with a small data set and SSA-adapted annotation procedures.
- Phylogenetic analysis employing six chloroplast loci indicated the ancestral relationships of the five most agriculturally-important *Striga* species, with the unexpected result that the one legume parasite (*S. gesnerioides*) was found to be more closely related to some of the grass parasites than many of the grass parasites are to each other.
- Generation of ~2200 reads each for seven *Striga* DNA samples allowed identification of the highly repetitive DNA content in these genomes. Genome sizes were determined by flow sorting, and the values of 615 Mb (*S. asiatica*), 1227 Mb (*S. gesnerioides*), 1425 Mb (*S. hermonthica*) and 2460 Mb (*S. forbesii*) suggest a ploidy series.
- The fourteen most abundant repeats in these *Striga* species were identified and partially assembled. Annotation indicated that they represent eight long terminal repeat (LTR) retrotransposon families, three tandem satellite repeats, one LINE retroelement, and one DNA transposon.
- These repeats were differentially abundant in each species, with the LTR retrotransposons and the satellites most responsible for variation in genome size. Each species had some repetitive elements that were more abundant and some less abundant than the other *Striga* species examined, indicating that no single element or any unilateral growth or decrease trend in genome behavior was responsible for variation in genome size and composition. In addition, 140 candidate microsatellite markers were identified that can be used for further characterization of *Striga* genomes and genetics.

Key words: Genome size, Polyploidization, Repetitive DNA, Sample sequence analysis, *Striga*, Transposable elements.

Introduction:

Striga, commonly known as witchweed, is a genus of parasitic plants with an old world distribution. Natural low-density populations infecting a diverse group of grass species can be found in tropical grasslands across much of Africa and parts of Asia (Mohamed *et al.*, 2001). Most research has focused on the five African species that are agronomically important: *S. hermonthica, S. asiatica, S. forbesii, S. aspera,* and *S. gesnerioides*. The latter parasitizes legumes, while all other species within the genus are grass parasites. In agricultural environments, these five species are found in high-density populations parasitizing several staple crops, including maize, sorghum, pearl millet, and cowpea.

In Sub-Saharan Africa*, Striga* is the single most important biological limitation to food production (Ejeta, 2007b). Despite its devastating effects on crop yields and subsistence farmer livelihoods, relatively little is known about the genetics, development, biochemistry or physiology of this parasitic genus. Most studies have focused on *Striga* ecology (rhizosphere communication, seed bank dynamics, haustorium formation) or control (host resistance and chemical, abiotic or biotic control methods). One of the prominent gaps in *Striga* characterization is understanding the phylogenetic relationships of the 28 described African species and the less-well-defined Asian species. The base ploidy level and the evolutionary role of polyploidization, if any, within the genus are not known. A genetic map has not been developed for any *Striga* species. One recent study has initiated rigorous investigations of genetic population structure in *S. gesnerioides* (Botanga & Timko, 2006).

Genetic characterization of any wild plant is a challenging exercise, but particularly so with species that are obligate parasites, have no genetic maps, have few identified sequences, and have limited potential for cytogenetic analysis due to the absence of extensive root tissues. Hence, innovative technologies for efficient genome characterization need to be developed for *Striga*. One approach for genome characterization and genetic tool development is by sample sequence analysis (SSA) of randomly selected genomic DNA (Brenner *et al.*, 1993). This technique has been employed to explore genome composition in hexaploid wheat using the sequences of randomly-selected BAC clones (Devos *et al.*, 2005), in the genus *Oryza* using BAC end sequences (Piegu *et al.*, 2006) and within *Gossypium* and *Pisum* using genomic shotgun sequence data (Hawkins *et al.*, 2006; Macas *et al.*, 2007). All of these studies have concluded that transposable elements (TEs) constitute a large portion of plant nuclear genomes and can greatly influence genome size. Satellite (tandem) repeats have also been shown to occur in large and variable numbers that affect genome size.

In this study, we have used SSA to investigate and compare the genomes of several agronomically-important *Striga* species. Our results indicate general similarities and some dramatic differences in genome composition among the species analyzed. We have also uncovered sequence variability that will be useful for future characterization of *Striga* genomes and studies of inheritance within each species.

Materials and Methods:

Seven isolates representing five *Striga* species were employed in this analysis. These isolates were originally collected from multiple African countries over a period of several years (Table 1). For purposes of DNA extractions, these seven isolates were grown on maize (*S. asiatica* (M), *S. aspera*, *S. forbesii* and *S. hermonthica* (M)), sorghum (*S. asiatica* (S) and *S. hermonthica* (S)), or cowpea (*S. gesnerioides*) for approximately 60 days in a quarantine facility (APHIS Plant Protection and Quarantine Permit No. 70902-P) at the University of Virginia. The same host species from the original seed collections were used. Vegetative above-ground stems were collected into liquid nitrogen and shipped on dry ice to the University of Georgia for nucleic acid extraction.

Phylogenetic analysis of selected *Striga* species:

Approximately 0.1 gm of frozen tissue was ground to a fine powder with liquid nitrogen in a 1.5 ml eppendorf tube. Total genomic DNA was extracted using a modified CTAB procedure (Doyle & Doyle, 1987). Published sequences were employed to design primers to amplify and subsequently sequence six loci from the chloroplast genome of each *Striga* species (Shaw *et al.*, 2005). *Buchnera americana* was similarly analyzed and used as an out-group. The sequenced DNAs included two coding loci (*rbcL* and *matK*) and four intergenic regions (*TrnT*$_{\text{(UGU)}}$ – *TrnL*$_{\text{(UAA)}}$, *TrnD*$_{\text{(GUC)}}$ – *TrnT*$_{\text{(GGU)}}$, *rpoB* – *TrnC*$_{\text{(GCA)}}$, and *TrnH*$_{\text{(GUG)}}$ – *psbA*). Sequences were aligned with SeaView and then concatenated into a single super-matrix of 4326 bp (Galtier *et al.*, 1996). A parsimony analysis was conducted using PAUP with 1000 bootstrap replicates to evaluate relationships between the studied *Striga* species (Swafford, 2002).

Genome size estimates:

Genome size was measured by flow cytometry. Fresh leaf tissue was shipped to The Flow Cytometry and Imaging Core laboratory at the Virginia Mason Research Center (Seattle, WA). Genome size measurements were repeated a minimum of five times with *Oryza sativa* (genotype TP-109) as a size standard (Arumuganathan *et al.*, 1999; Temsch & Greilhuber, 2000).

DNA isolation:

Approximately 20 gm of frozen tissue were ground to a fine powder with a mortar and pestle and immediately suspended in a sucrose extraction buffer (SEB) following option Y as described in Peterson *et al.* (2000). The suspension was filtered through two layers of cheesecloth and miracloth to remove particulates, and 10% Triton X-100 (v/v) in SEB solution was then added to the filtered suspension at a 1:20 volume ratio. The Triton X-100 selectively lyses plastid membranes, while leaving nuclear membranes intact (Gurtubay *et al.*, 1980). Nuclei were isolated from the suspension using a series of centrifugations at 650xg for 15 min. Nuclei were then resuspended in 5 ml of Buffer AP1 from the DNeasy plant maxi-kit (Qiagen, Valencia, CA), and manufacturer's instructions were followed to isolate plastid-free DNA.

Library construction:

For each sample, ten ug of DNA were sheared using a Hydroshear (GeneMachines, San Carlos, CA) set on speed code 14, for 20 cycles, to obtain DNA fragmented in the size range of 3-5 kb. The sheared fragments were converted into blunt-ended molecules using mung bean nuclease (New England Biolabs, Beverly, MA), size selected (3-5 kb) on a 1% agarose gel, dephosphorylated using shrimp alkaline phosphatase (Roche, Indianapolis, IN), and A-tailed using *Taq* DNA polymerase (Roche, Indianapolis, IN) plus dATP (Roche, Indianapolis, IN). The modified fragments were then cloned into a Topo-4 cloning vector and transformed into ElectroMAX DH10B cells (Invitrogen, Carlsbad, CA). For each library, three 384-well plates of clones were randomly chosen, plasmid DNA was prepared and the inserts sequenced in both direction using BigDye terminator v3.1 chemistry (Applied Biosystems, Foster City, CA). The chromatographs obtained from the 3730 sequencer (Applied Biosystems, Foster City, CA) were base called with Phred (Ewing *et al.*, 1998). Low quality, vector, chloroplast, and mitochondrial sequences were identified with Phred and Cross_match and removed from the data sets before further analyses.

454 DNA Library preparation (*S. gesnerioides*):

Library preparation for 454 sequencing procedures were carried out at the Interdisciplinary Center for Biotechnology Research (University of Florida). Briefly, *Striga gesnerioides* race SG4z genomic DNA was fractionated into 400-800 bp fragments and the ends of the fragments were polished (blunted) using a combination of T4 DNA polymerase and T4 polynucleotide kinase. Short adaptor oligonucleotides were ligated onto the ends of the polished DNA fragments to provide priming sequences for both amplification and sequencing of the sample-library fragments. Adaptor B contains a 5'-biotin tag that enables immobilization of the library onto streptavidin-coated paramagnetic beads and washing to remove ligation products that will not participate in downstream DNA sequencing reactions. After nick repair, the non-biotinylated strand is released from the bead-DNA duplex by denaturation and used as a single-stranded template DNA (sstDNA) library. The sstDNA library was assessed for quality (proper size range) and the optimal amount (DNA copies per bead) needed for emulsion PCR (mPCR™) determined empirically by testing a series of sstDNA to capture bead ratios by emPCR and estimation of the fraction of recovered DNA-containing beads. A large volume production emPCR was then conducted to produce a sufficient number of DNA capture beads—populated with clonally amplified sstDNA—to load onto the 454 GS FLX.

454 Full scale sequencing:

Full scale 454 sequencing was carried out at the Genomics Core Facility (Department of Biology, University of Virginia). For full scale sequencing the populated DNA capture beads were added to the DNA Bead Incubation Mix (containing DNA polymerase) and layered with Enzyme Beads (containing sulfurylase and luciferase) onto the PicoTiterPlate™ (PTP) device with 2 large regions. The loaded PTP device was placed into the Genome Sequencer (GS) FLX instrument and run using the Roche XLR70 Titanium sequencing chemistry kit.

454 Post-Run Analysis:

Flow-data were assembled using the GSAssembler package (Software Version 2.0.00.22) as described in the Genome Sequencer Data Analysis Software Manual

(Version October 2008). Total raw and keypass wells for both regions were 1,147,687 and 1,127,168. The total number of sequences that passed all quality filters for both regions of the PTP was 820,210 with an average read length of 374 nucleotides, representing 307,208,175 bases.

Sampling 454 data:

A total of 30,000 sequence reads were extracted from the total data set by randomly choosing 10,000 reads, three times with a perl script. Chloroplast and mitochondrial sequences were identified with Cross_match and removed from the data sets before further analyses. Sequence annotation procedures (described below) were implemented on all three 454 data sets (~30,000 reads) with no observed differences, therefore only results from one data set are reported.

Sequence identification and reconstruction:

Each of the seven libraries was subjected to an “all versus all” BLASTN using an expect value of 1x10e-5 or smaller as an acceptance threshold, to assess the repetitive nature of the samples. Sequences were grouped into 3 categories; single copy, between two and ten copies, and eleven or more copies within the data set. Repeats in each of the libraries were also annotated/identified by comparing the libraries to known repetitive sequences using BLASTN, TBLASTX, and TIGR’s repeat database version 8.0 (Ouyang & Buell, 2004). Extended repeat pseudomolecules were also constructed for each of the libraries using the AAARF software package (DeBarry *et al.*, 2008). The 18 largest pseudomolecules (> 3 kb) built with AAARF were structurally annotated using blast searches (BLASTN and TBLASTX) against several PFAM databases (Finn *et al.*, 2008). These databases included those for retroelement domains such as GAG (PF00077), reverse transcriptase (PF00078), integrase (PF00665), RNaseH (PF00075), primer binding sites (Wilhelm *et al.*, 1994), and tRNA priming sites. PFAM's transposase (PF00872) database was also used. Known TEs were sought using the TIGR plant repeat database (Ouyang & Buell, 2004)and the MIPS REdat database (Spannagl *et al.*, 2007). These sequences were also subjected to analysis with Tandem Repeat Finder (Benson, 1999). All of the BLAST searches for each of the pseudomolecules assembled by AAARF were visualized simultaneously with the Apollo genome annotation browser and hand curated to identify class and family designations for each pseudomolecule. A *Striga*-specific repeat database was constructed, using the pseudomolecules identified in this analysis, to produce a more accurate annotation of each library.

Annotation and 95% confidence intervals:

Many TEs share regions of homology (see above) and therefore must be identified or annotated by a competitive process. All of the sample sequences were annotated using BLASTN and the repeat database constructed from the identified pseudomolecules. An Expect value of 1x10e-10 or lower was used as the criterion for accepting an annotation. The annotated sequences were then transformed into a percent of total base pairs for each repeat in each genome sample by dividing the repeat length in a single sequence by the total of that sequence read length. This approach allows us to precisely count each bp identified in our samples as a repeat. The transformed data on percent genome occupation were then subjected to a bootstrap analysis with replacement using SAS with 1000 permutations. The values produced in the bootstrap statistic were graphed to display differences between the genomes as a mean with a 95% confidence interval for each species or genotype.

Phylogenetic analysis of retroelements:

All of the nucleotide sequences were compared by TBLASTX to the PFAM database (PF00078, version 22.0), representing all known reverse transcriptase (RVT) sequences associated with retroelements in flowering plants. Sequences with homology were translated into all six possible reading frames and BLASTX was used to identify the correct translation. These sequences and fourteen RVT sequence from *Oryza*, *Sorghum*, and *Zea* were then aligned with ClustalX and SeaView (Galtier *et al.*, 1996), removing un-alignable sequences and removing any overhanging base pairs so that all sequence comparisons contained the same number of amino acid residues. The final alignment was then used to construct a neighbor-joining (NJ) tree with bootstrap values representing 1000 replicates.

Simple sequence repeats (SSRs, otherwise known as microsatellites) were identified within the two *S. hermonthica* samples using the Perl Script SSRIT (Temnykh *et al.*, 2001). Primer pairs were designed to amplify products between 100 and 300 bp for those sequences identified with SSRs using Primer3 (Rozen & Skaletsky, 2000).

Results:

Phylogenetic analysis of the agriculturally-important *Striga* species was performed in PAUP using parsimony. A single most parsimonious tree was generated with 356 steps (Fig 3-1). This tree has a consistence index (CI) of 0.9719, a homoplasy index (HI) of 0.0281, a retention index (RI) of 0.8936, and strong bootstrap support. The tree indicates that *S. forbesii* is a basal lineage within the genus and that *S. aspera*, *S. gesnerioides*, and *S. hermonthica* form the most derived clade. This analysis also suggests that *S. aspera* and *S. gesnerioides* are more closely related to each other (sister) than either is to *S. hermonthica*.

Flow cytometric estimates of genome size were generated from fresh plant material using *O. sativa* (genotype TP-109) as a standard. Genome sizes ranged from 0.605 pg/1C in *S. asiatica* (S) to 2.51 pg/1C in *S. forbesii* (Table 3-1). These values were converted into Mb estimates by multiplying by 980 Mb/pg, resulting in values ranging from ~590 Mb to ~2460 Mb. No significant differences were observed between the two *S. asiatica* or the two *S. hermonthica* genome size estimates. While each species was sampled with about the same number of DNA sequences in the subsequent analyses (except *S. gesnerioides*), the approximate four-fold range in genome size means that the percent of each genome sampled varied from ~0.08% in *S. forbesii* to ~0.3% in *S. asiatica* and *S. gesnerioides* (Table 3-1).

The average read length for all of the Sanger sequences generated was ~850 bp, with the *S. aspera* library having the smallest average read length of ~803 bp and the *S. hermonthica* (M) library having the largest average read length of ~880 bp. The average read length for the *S. gesnerioides* 454 library was ~377 bp. After removing vector, chloroplast, mitochondrial, and bacterial sequences, ~2200 sequences were generated from each Sanger sequencing library to produce ~1.8 Mb to ~2.0 Mb of nucleotide sequence per sample. About 10,000 sequences were sampled from the 454 library to produce ~3.7Mb of nucleotide sequence (Table 3-1). The guanine plus cytosine (GC) content was calculated for each library, ranging from 36.8% in *S. asiatica* (S) to 40.6% in *S. gesnerioides*. GC content was not significantly different between the two samples of *S. asiatica* (grown on maize [M] and grown on sorghum [S]) or the two samples of *S. hermonthica* (M & S).

To determine the repetitiveness of each genome sampled, sequences from a given library were compared against themselves by an all-by-all BLAST, and the total number of identified homologies ("hits") for each sequence was recorded. Sequences were grouped into three categories based on the number of hits counted: single copy, repetitive (2-10 copies), and highly repetitive (>10 copies). The groupings were displayed graphically by arranging the seven data sets in phylogenetic order (Fig. 3-2).

By the criteria employed, approximately 60% of the two *S. asiatica* samples were classified as single copy in the data set, with the remainder being repetitive (~30%) or highly repetitive (~10%). The largest genome sampled, *S. forbesii,* showed equal (~33%) distribution of single copy, repetitive, and highly repetitive sequences. The remaining four samples ranged from 40-56% single copy sequences, 25-40% repetitive and 18-25% highly repetitive sequences. It is worth noting that the low degree of sampling for each genome means that most sequences with copy numbers of less than several hundred per genome would be found only once in the sequenced sample (and thus be called "single copy"), while all those sequences found in the repetitive and highly repetitive categories are mostly from repeats that are found in copy numbers in the thousands or tens of thousands, respectively, in the haploid nuclear genome.

A total of 337 pseudomolecules were generated from the seven data sets using the AAARF program for *de novo* repeat assembly (DeBarry *et al.*, 2008). The AAARF software was used multiple times on each data set, using a range of parameters that can be adjusted in the program header. The settings employed were MinBlastIdentity ranging from 80-89%, Blast_e ranging from e-10 – e-25, BL2Seq-e ranging from e-3 – e-10, and Min_Extend_Len ranging from 0-10. These settings were tested to investigate reproducibility and to build the longest possible pseudomolecules. Eighteen pseudomolecules >3 kb were constructed. These eighteen were subjected to structural annotation to identify domains commonly found in TEs. Thirteen pseudomolecules were identified as likely TEs (one LINE-like retroelement, five *gypsy*-like LTR retrotransposons, six *copia*-like LTR retrotransposons, and one DNA transposon). The remaining five pseudomolecules did not exhibit homology to common structural domains found in TEs. The Tandem Repeats Finder software indicated that these five pseudomolecules are comprised of tandemly repeated sequences (Benson, 1999). Comparing the results from each of the eighteen pseudomolecule annotations revealed that several pseudomolecules built from separate data sets represented the same TE or tandem repeat, leaving a total of fourteen novel repetitive sequences (Table 3-2). None of the ten retroelements identified appear to be full-length pseudomolecules due to the lack of a GAG domain in most or an integrase domain in some, and the inability to identify at least one LTR for the eight *copia*-like or *gypsy*-like LTR retrotransposons.

The fourteen repeats were then used to investigate each of the sequences from all seven samples using BLASTN. The total length (bp) of the homology to a repeat was divided by the total length of the query sequence to identify the percent of each sequence composed of a repeat. In order to generate a prediction of the degree of accuracy in these estimates of total genome repeat content, transformed values for a given repeat were resampled 1000 times using bootstrap statistics to identify the mean and 95% confidence intervals for each repeat found in each library. These statistics were graphed for the different classes of TEs and tandem repeats, and for representative individual repeats, to investigate the variation between samples (Figure 3-3, A & B).

The percentage of each sample composed of identified LTR retrotransposons ranged from ~2.3% in the *S. asiatica* samples to ~11.7% in the *S. gesnerioides* sample. DNA transposons ranged from <1% in the *S. asiatica* samples to ~2% in the *S. forbesii* sample, and identified tandem repeats ranged from ~0.5% in the *S. gesnerioides* sample to ~9.4% in one *S. hermonthica* (M) sample (Fig. 3-3A). The two most repetitive *copia* and *gypsy* retrotransposons and the LINE retroelement were graphed to display the differences between samples (Fig. 3-3B). The LTR retrotransposon identified as Wico (SRC1) ranged from ~1.8% in the *S. asiatica* samples to ~9.8% of total DNA in the *S. forbesii* sample. The LTR retrotransposon identified as SRC4 ranged from ~0.5% in the *S. aspera* sample to ~3.8% of total DNA in the *S. forbesii* sample. The LTR retrotransposon identified as Wigy (SRG1) ranged from less than 1% of the DNA in the *S. asiatica* samples to ~3% in the *S. aspera* sample. Another *gypsy* retrotransposon (SRG5) ranged from undetected in the *S. asiatica* samples to ~2.1% in the *S. hermonthica* (M) sample. The only repeat identified as a LINE (SRL1) was fairly constant in genome representation, ranging from just under 1% in the *S. asiatica* and *S. forbesii* samples to ~1% in *S. hermonthica*. The satellite repeat identified as SS1 ranged from undetected in *S. forbesii* to ~9.3% in the *S. hermonthica* (M) sample. In total, 4691 sequences were annotated as one of the fourteen repeats identified using the AAARF software, and this amounted to about 20% of the total input sequence data.

In an attempt to investigate whether the LTR retrotransposons identified in our repeat analysis and pseudomolecule reconstruction actually represent distinct clades, phylogenetic trees were constructed using reverse transcriptase (RVT) sequence homologies identified within the total data set. Several known retroelements from other species were used to orient the phylogenetic analysis, and to search for evidence of horizontal transfer for these most abundant repeats. Each of the 186 RVT sequences was translated into the correct reading frame and two alignments were generated. The first alignment contained 46 sequences that aligned from the start codon of RVT for ~176 amino acid residues and were all annotated as Wico (data not shown). The second alignment contained a total of 97 sequences aligned across 171 amino acids, and this represented the bulk of RVT sequences that could be aligned within our data set (Fig 3-4). Each of the sequences within this alignment was annotated with the *Striga* repeat database generated in this project after the phylogenetic tree was constructed. The resulting annotations were mapped onto the phylogeny. Eight of the repeats identified as retrotransposons are identified on the phylogeny as a monophyletic bootstrap-supported clade, except SRG5, which only contained homology to the GAG domain. It is important to note that Wico is underrepresented in this phylogeny because 46 sequences representing this clade could not be aligned with those used to generate this tree.

In the un-rooted phylogenetic tree shown in Figure 3-4, sequences (including fourteen elements from various grass species) identified as *copia*-like and *gypsy*-like form two distinct clades that are sister to each other, while the LINE-like sequences are grouped into several clades that all appear basal to the *copia*/*gypsy* clade. These results are in complete agreement with the general observation that LINEs are the most ancient and diverse class of retroelements, and that the two superfamilies of LTR retrotransposons, *gypsy*-like and *copia*-like, were derived from a single ancestral LTR retrotransposon (Xiong & Eickbush, 1990).

A total of 489 sequences were identified as potentially useful SSRs, of which 287 were removed because they represented the same repeat based on monomer length and composition. A further 62 were removed because the simple sequence repeat was identified near the end of a sequence such that sufficient unique sequence was not available for primer design. Fifty of the remaining 140 SSRs were used to design primer pairs that were used to amplify products from eight control DNA samples of *S. hermonthica*. Forty seven (94%) of the attempted PCR reactions produced appropriately sized products based on gel electrophoretic analysis, and polymorphism was frequent (data not shown).

Discussion:

The parasitic lifestyle is found in only ~1% of flowering plant species (Kuijt, 1969). While many parasitic plants are not weeds, some like the weedy *Striga* species have adapted to cultivated host plants with the reduced genetic diversity that is routinely associated with domestication and agriculture. The leap from a low density population inhabiting tropical grasslands to a high density population inhabiting agricultural lands has occurred at least once and possibly several times within this single genus over the relatively short period of time since agriculture began (i.e., the last 10,000 yrs). The shift in lifestyle and the rapid breakdown of host resistance observed in field trials of newly developed *Striga*-resistant cultivars suggest that some species within the genus are evolving rapidly. It is expected that aspects of the predicted recent and rapid evolution of *Striga* would be manifested in the structure and/or composition of the weeds' nuclear genomes.

The first inspection of *Striga* genomes, provided here by sample sequence analysis, suggests that the nuclear DNA content of several agriculturally-important species is fairly typical when compared to other studied plant genomes. Repeats were found to be quite abundant, and most of these were LTR retrotransposons as in other large-genome angiosperms. The analysis uncovered distinctive properties of the genomes of each of these species because both the paired *S. asiatica* samples (from very different collection locations, Tanzania and Zimbabwe) and the paired *S. hermonthica* samples (from Sudan and Nigeria) exhibited high species-specific similarity in all genome properties analyzed (size, GC content, repeat types, repeat abundances).

Phylogenetic analysis reveals that *S. aspera* and *S. gesnerioides* are sister species and that *S. hermonthica* is closely related to both. This observation is unexpected because *S. aspera* and *S. hermonthica* are morphologically very similar and are believed to have the ability to cross hybridize, while *S. gesnerioides* is morphologically distinct (Aigbokhan *et al.*, 2000). While the genus contains ~28 species, only five of them are considered agricultural weeds. Our results suggest three distinct clades within the genus.

The most derived clade includes *S. aspera*, *S. gesnerioides*, and *S. hermonthica,* while the two other weedy species (*S. asiatica* and *S. forbesii*) represent independent lineages.

An approximate four-fold difference in genome size was observed within this sampling of the genus *Striga*, suggesting that the processes of polyploidization and/or TE amplification are influencing genome evolution. The genome size values suggest a ploidy series with two rounds of polyploidization: a diploid species having a genome size of ~600 Mb (*S. asiatica*), a tetraploid species with a genome size of ~1400 Mb (*S. hermonthica*), and an octoploid species with a genome size of ~ 2400 Mb (*S. forbesii*). Several rounds of polyploidization have been observed and inferred to drive speciation in the evolution of a closely related genus, *Orobanche* (Schneeweiss *et al.*, 2004).

In possible disagreement with the polyploidization hypothesis are the values calculated for percent GC content, total repetitiveness, and phylogenetic position (Table 3-1, Fig. 3-1, and Fig. 3-2). Both GC content and total repetitiveness values were found to increase at larger genome sizes, with an ~3% variation in GC content and an ~25% variation in repeat content across the species sampled. These increases would not be expected if simple polyploidy was the only cause of genome size variation, thus suggesting that some other process(es) may be shaping these genomes. Phylogenetic position does not dispute the polyploidization hypothesis, but suggests independent polyploidization events within the *S. forbesii* lineage unrelated to polyploidization in the more derived clades within the genus.

Using *de novo* repeat assembly and annotation approaches, fourteen different families of repetitive DNA were identified in the *Striga* genomes examined here. These repeats are the most abundant sequences in these *Striga* genomes, representing approximately 10.7%, 15.9%, 16.1%, 18.5%, and 19% of the *S. asiatica*, *S. forbesii, S. gesnerioides*, *S. aspera*, and *S. hermonthica* nuclear genomes, respectively. Each repeat belongs to a class or group of repetitive sequences found commonly in plant genomes. All of these most abundant TEs appear related to their closest relatives in other dicot species, suggesting that none of them originated as horizontal transfers from their grass hosts. Comparing the percent content of individual repeats across species allows us to identify differences between the genomes sampled. As an internal control, both the *S. asiatica* and *S. hermonthica* samples represent genotypes collected from different host species and different geographic locations. No significant differences were observed between either of the paired samples. Although these results suggest that the analysis provides definitive species-specific descriptions of repeat content, the data set was not deep enough to produce full-length pseudomolecules for any of the LTR retrotransposons. Hence, the observed contributions of these repeats are a minimal estimate of their true contributions. Moreover, most or all of those repeats with copy numbers less than a few hundred are expected to be missed by this analysis. In the ~2400 Mb maize genome, these low-copy-number repeats, commonly LTR retrotransposons, comprise about 16% of the nuclear DNA (SanMiguel & Bennetzen, 1998; Baucom *et al.*, 2009). Thus, the overall repeat composition of all of the studied genomes is very likely to be >50%, in line with comparable results for other angiosperm genomes (Bennetzen, 2000; Kazazian, 2004).

The low genomic redundancy (0.08% to 0.3% coverage) of the sample sequence data also limited the completeness and detail available for examining the phylogentic relatedness of the discovered LTR retrotransposons. However, next generation sequencing technologies will allow more comprehensive sampling, and thus, more complete discovery of repeats in any genome (Mardis, 2008). The AAARF program for the discovery and assembly of repeats into pseudomolecules has proven competent to function with short sequence reads (DeBarry *et al.*, 2008), so more pseudomolecules should be available for future phylogenetic tree assembly.

Examining the percent composition differences between samples and different repeat groups (i.e., retroelements, DNA transposons, and tandem/satellite repeats) revealed that each of these repeat classes is differentially influencing genome evolution in these species (Fig. 3-3A). TEs are more abundant in the larger genomes. Satellite repeats display a different pattern, with variation that does not correlate with genome size (~8% in the small to mid-sized *S. asiatica*, *S. aspera* and *S. hermonthica* genomes, ~1% in *S. gesnerioides* and ~3% in the largest genome, *S. forbesii*). The *S. forbesii* genome contains about the same number of Mb of tandem/satellite repeat as *S. asiatica*, however, since the *S. forbesii* genome is about four times the size of the *S. asiatica* genome they appear reduced in % composition comparisons. The data indicated that LTR retrotransposons and satellite repeats are the major contributors to the variation in genome size. However, the differences in observed repeat content in the species studied cannot fully explain the observed overall genome size variation. From this analysis, the "single copy" components of *S. asiatica*, *S. hermonthica*, *S. gesnerioides*, and *S. forbesii* are ~370 Mb, ~640 Mb, ~687 Mb and ~810 Mb, respectively, once again suggesting a ploidy series. It is not clear whether this might be a 2X, 4X, 6X series or only a 2X and 4X series, with species that differ in further DNA variations (e.g., segmental duplications, aneuploidies, other repeat amplifications) not identified in this study.

Further examination of LTR retrotransposon composition by sub-class (i.e., *copia* and *gypsy*) indicates that *gypsy*-like retrotransposons are relatively more abundant in *S. aspera*, *S. gesnerioides*, and *S. hermonthica* samples than in the larger *S. forbesii* genome while *copia*-like retrotransposons are more abundant in the *S. forbesii* sample than in the mid-sized genome samples. Similar results have been observed in other plant species, and cases in cotton (*Gossypium* species) and *Oryza australiensis* suggest that amplification of one or a few LTR retrotransposons families have been responsible for most of the nuclear genome growth seen in particular lineages (Hawkins *et al.*, 2006; Piegu *et al.*, 2006). However, the situation in *Striga* is similar to that in the genus *Zea* where LTR retrotransposons and satellite repeats of several different families are differentially amplified across several studied genomes (Estep & Bennetzen, unpublished). As in the case of *Zea*, the larger *Striga* genomes like *S. forbesii* contain more of some repeats but less of others compared to their smaller genome relatives. Overall, a greater quantity of repeats accounts for part of the increased genome size of the larger-genome *Striga* species, but this is not explained by a single element family or an exclusive genome growth trend in one direction. Hence, it appears that perhaps random but certainly stochastic activation of different TE and satellite repeats accounts for much of the genome size variation in the genus *Striga*.

Examining each individual repeat suggests that Wico (SRC1) had the largest effect on differences seen within the *copia* superfamily and that Wigy (SRG1) and SRG5 had the largest effect within the *gypsy* superfamily. Wigy comprises significantly more of the *S. aspera* sample (~3.1%) than the *S. forbesii* sample (~1.5%, ~40 Mb), but neither is significantly different from the *S. hermonthica* samples (~2.2%, ~30 Mb). SRG5 on the other hand, constitutes significantly less of the *S. forbesii* sample (~0.2%, ~5 Mb) than any of the *S. aspera* (~1.6%), *S. gesnerioides* (~1.6%, ~20 Mb), and *S. hermonthica* samples (~1.9%, ~27 Mb).

Based on length, percent GC content, and the lack of identifiable coding potential, the three tandem/satellite repeats identified in this study are similar to other satellite repeats found in centromeres and heterochromatin knob structures of several plant systems (Plohl *et al.*, 2008). The repeat SS1 constitutes ~8% of most of the genome samples, similar in percent sample composition in some maize lines to the amount of 180 bp satellite repeat responsible for knobs in *Zea* (Peacock *et al.*, 1981). Although no SS1 was detected in *S. forbesii*, it has additional copies of SS2 and SS3 that make its overall satellite repeat content (~2.5% of the genome, ~62 Mb) very similar to that found in *S. asiatica* (~8.1% of the genome, ~50 Mb).

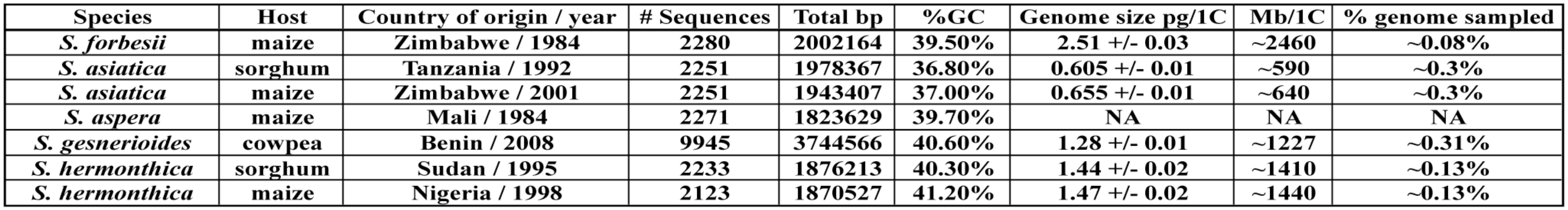
No other satellite repeat has been found in *Zea* that reaches anywhere near the abundance of the knob repeat (Estep & Bennetzen, unpublished). It is likely that SS1 is arrayed in large blocks found in heterochromatic regions within *Striga* genomes, similar to knobs observed in maize. In maize, the uniform heterochromatin staining of knob satellites made them useful to identify and track chromosomes during meiosis and mitosis (Rhoades & McClintock, 1935). Further investigations revealed that, in specific genetic backgrounds, those chromosomes with the greater numbers of knob repeat were preferentially passed on to offspring, but only through the female gametophyte. This process of distorted segregation was later termed meiotic drive (Rhoades & Dempsey, 1966). Meiotic drive can help explain the selfish over-accumulation of such sequences over time, and thus their abundance in many *Striga* genomes.

Beyond providing a global perspective on the forces influencing genome evolution, sample sequence analysis can also quickly identify molecular markers for future studies. Plant genomes contain numerous SSRs. These repeat sequences are commonly found to be highly polymorphic within a species and thus allow the development of genetic markers that can be used to describe the relationships between individuals and populations. A total of 140 candidate SSR markers were identified from the sample sequences of *S. hermonthica* and *S. aspera* using the freely available Perl script SSRIT (Temnykh *et al.*, 2001). These markers are now being used to characterize the diversity and population structure in *Striga* populations collected from several locations around the world.

Conclusions:

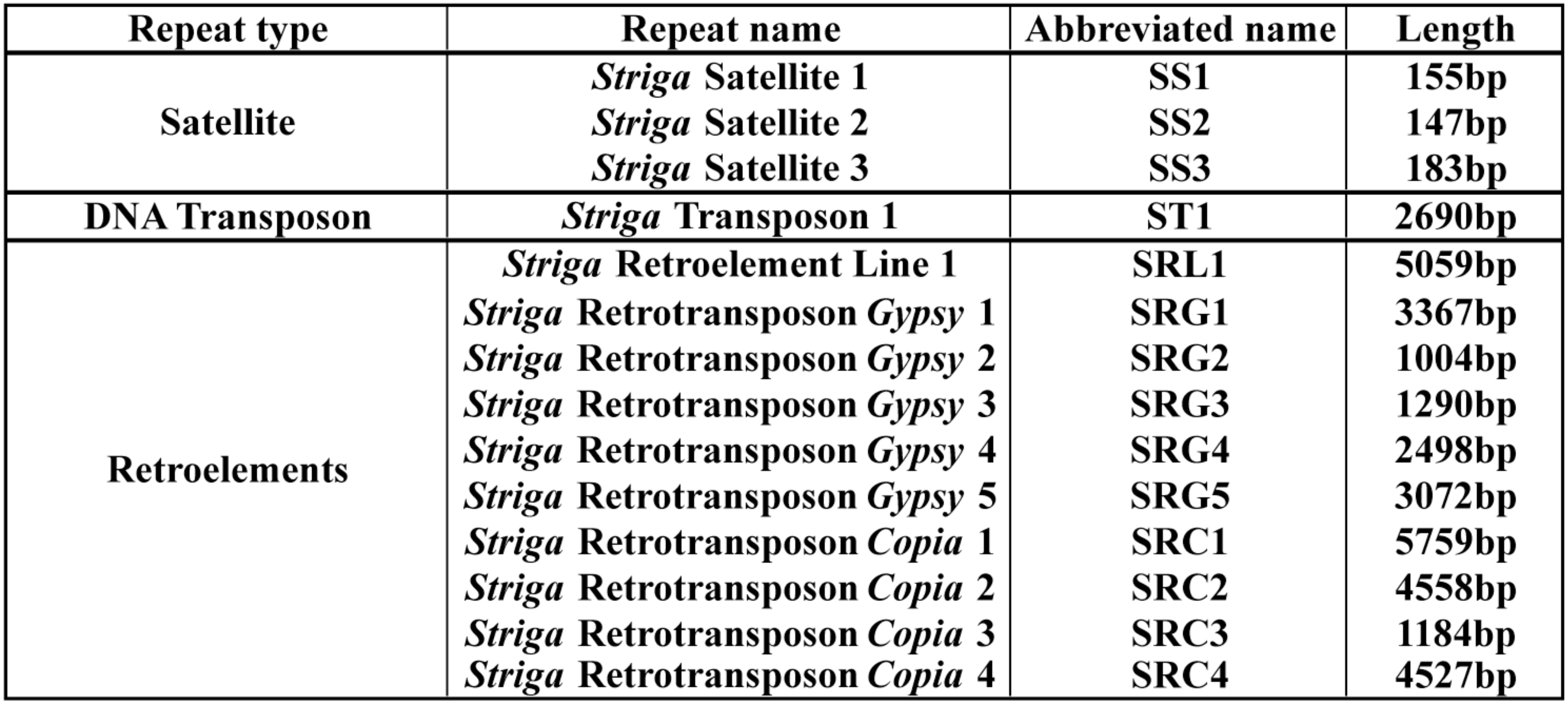
In this study, sample sequence analysis was applied to an important but challenging/understudied set of plant species to investigate global processes that can shape nuclear genomes. These modes of genome variation include polyploidization, TE amplification, and meiotic drive. *Striga* species were found to have typically complex angiosperm genomes, with LTR retrotransposons and satellite repeats the major players in genomic size variation and overall diversity. Evidence of recurrent polyploidy was also uncovered. At the same time, SSR markers were developed that will allow further investigation of evolutionary processes at the population level.

Table 3-1: Identified genome characteristics of the investigated *Striga* species.



Species	Host	Country of origin / year	# Sequences	Total bp	%GC	Genome size pg/1C	Mb/1C	% genome sampled
S. forbesii	**maize**	**Zimbabwe / 1984**	**2280**	**2002164**	**39.50%**	**2.51 +/- 0.03**	**~2460**	**~0.08%**
S. asiatica	**sorghum**	**Tanzania / 1992**	**2251**	**1978367**	**36.80%**	**0.605 +/- 0.01**	**~590**	**~0.3%**
S. asiatica	**maize**	**Zimbabwe / 2001**	**2251**	**1943407**	**37.00%**	**0.655 +/- 0.01**	**~640**	**~0.3%**
S. aspera	**maize**	**Mali / 1984**	**2271**	**1823629**	**39.70%**	**NA**	**NA**	**NA**
S. gesnerioides	**cowpea**	**Benin / 2008**	**9945**	**3744566**	**40.60%**	**1.28 +/- 0.01**	**~1227**	**~0.31%**
S. hermonthica	**sorghum**	**Sudan / 1995**	**2233**	**1876213**	**40.30%**	**1.44 +/- 0.02**	**~1410**	**~0.13%**
S. hermonthica	**maize**	**Nigeria / 1998**	**2123**	**1870527**	**41.20%**	**1.47 +/- 0.02**	**~1440**	**~0.13%**

Table 3-2: Newly identified *Striga* transposable elements and satellite repeats.



Repeat type	Repeat name	Abbreviated name	Length
Satellite	***Striga* Satellite 1**	**SS1**	**155bp**
	***Striga* Satellite 2**	**SS2**	**147bp**
	***Striga* Satellite 3**	**SS3**	**183bp**
DNA Transposon	***Striga* Transposon 1**	**ST1**	**2690bp**
Retroelements	***Striga* Retroelement Line 1**	**SRL1**	**5059bp**
	***Striga* Retrotransposon *Gypsy* 1**	**SRG1**	**3367bp**
	***Striga* Retrotransposon *Gypsy* 2**	**SRG2**	**1004bp**
	***Striga* Retrotransposon *Gypsy* 3**	**SRG3**	**1290bp**
	***Striga* Retrotransposon *Gypsy* 4**	**SRG4**	**2498bp**
	***Striga* Retrotransposon *Gypsy* 5**	**SRG5**	**3072bp**
	***Striga* Retrotransposon *Copia* 1**	**SRC1**	**5759bp**
	***Striga* Retrotransposon *Copia* 2**	**SRC2**	**4558bp**
	***Striga* Retrotransposon *Copia* 3**	**SRC3**	**1184bp**
	***Striga* Retrotransposon *Copia* 4**	**SRC4**	**4527bp**

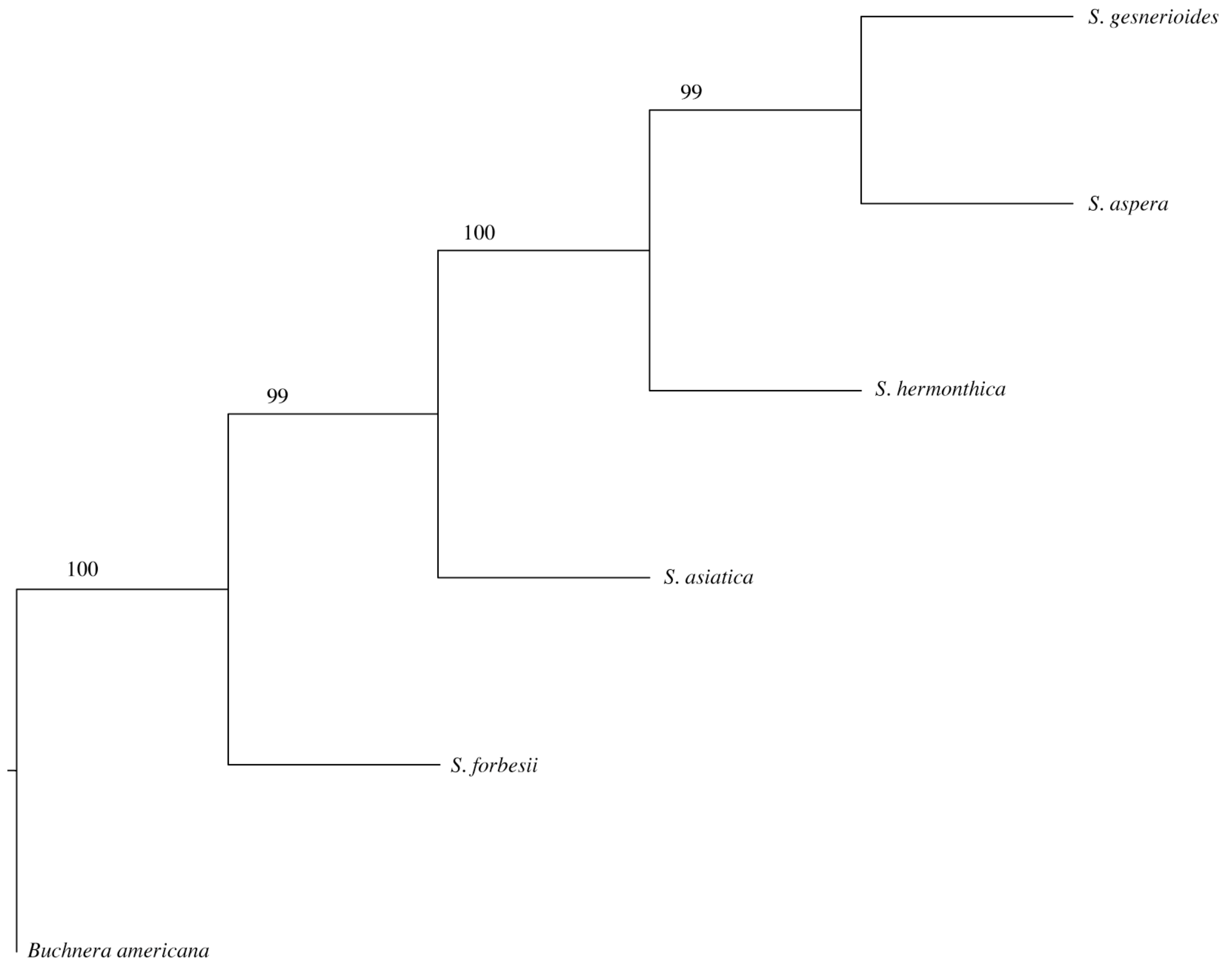


Figure 3-1: Phylogenetic tree of the agriculturally important *Striga* species based on six chloroplast loci (see Methods). Bootstrap values from 1000 replicates are indicated at nodes.

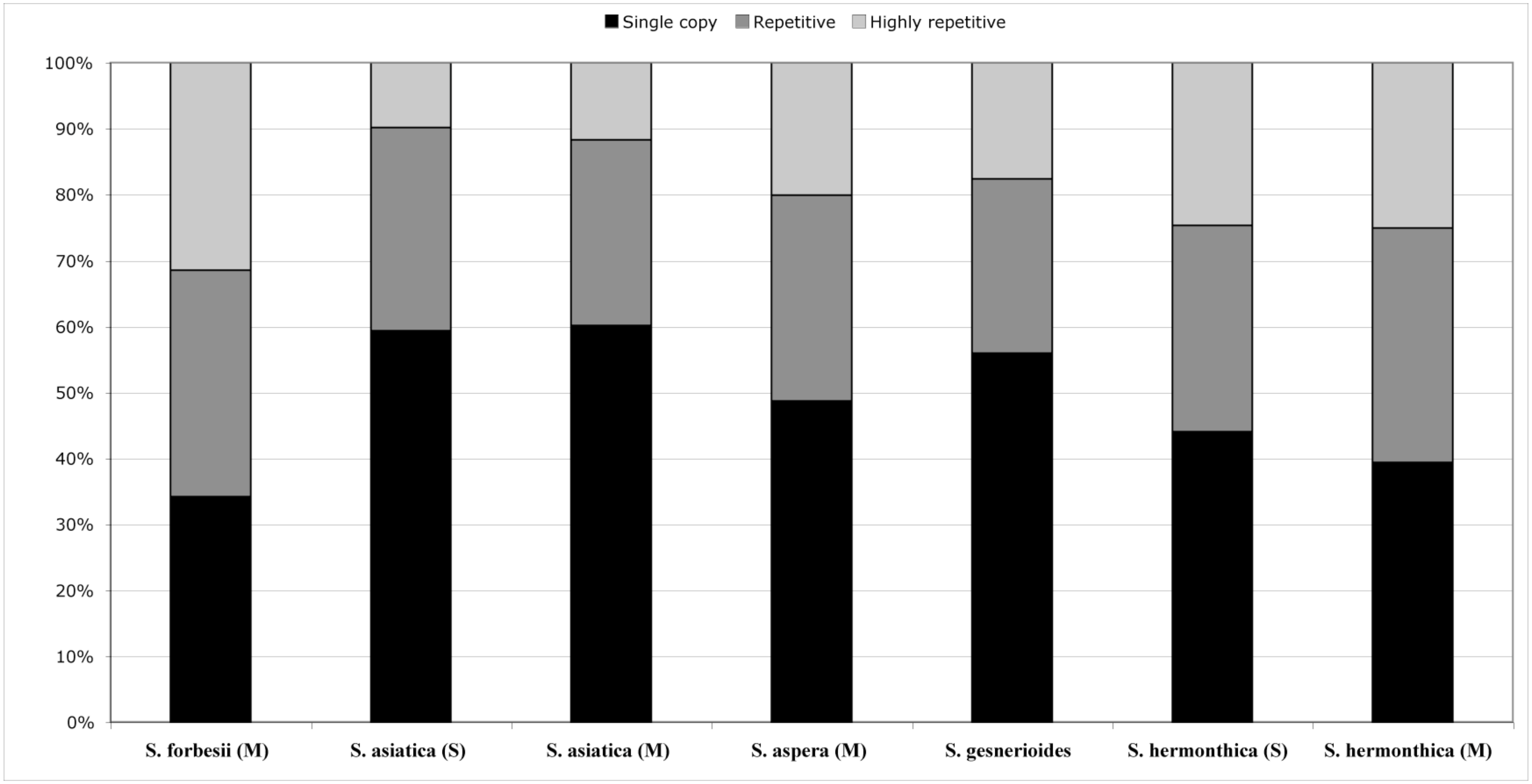


Figure 3-2: Graphical representation of repetitiveness for each *Striga* isolate sampled. Individual samples were compared to themselves by an all-by-all BLAST and the total numbers of homologies were counted. Sequences were grouped into three categories (single copy, repetitive (2-10 copies), and highly repetitive (>10 copies)) based on their copy number within each data set.

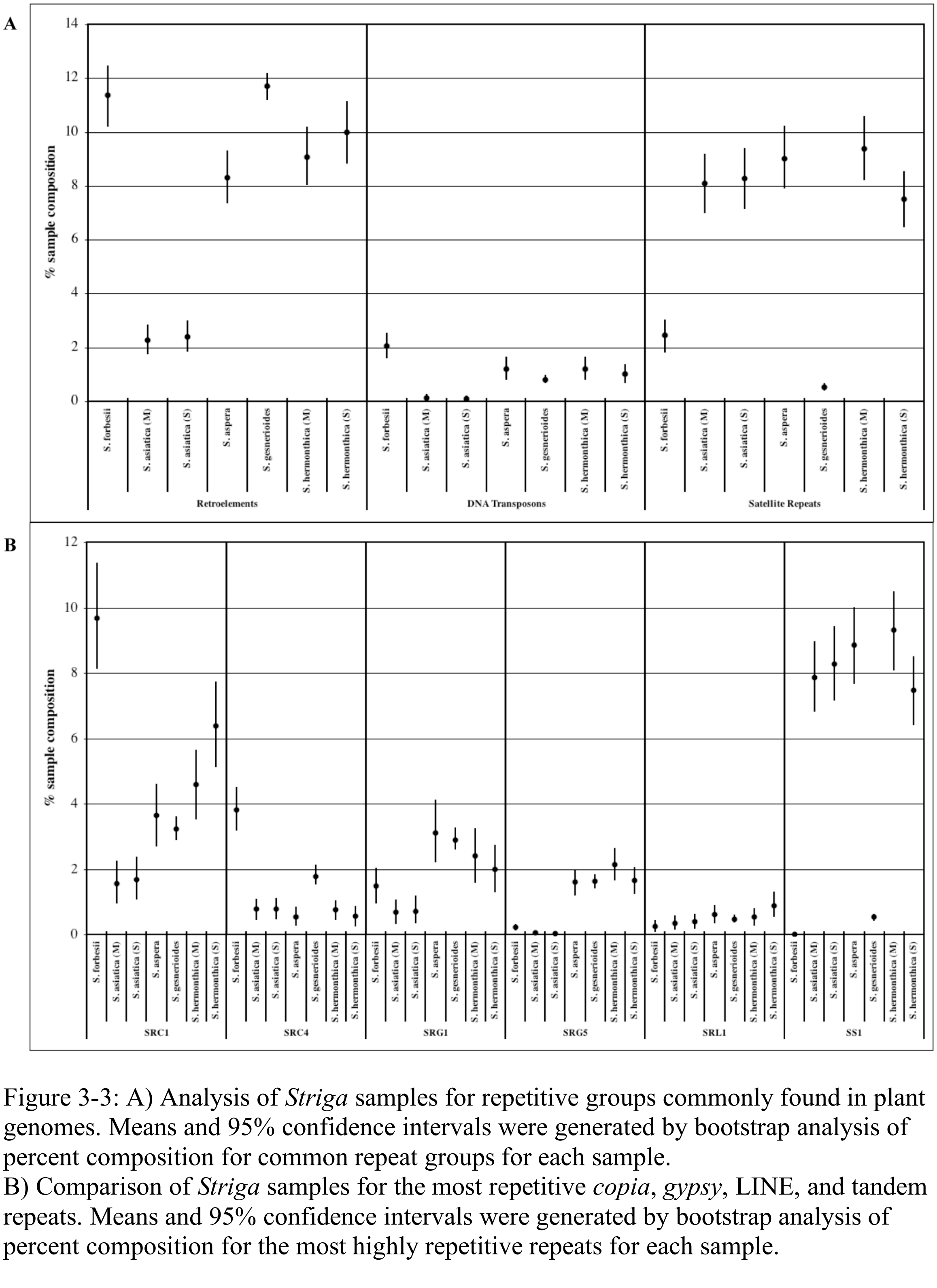


Figure 3-3: A) Analysis of *Striga* samples for repetitive groups commonly found in plant genomes. Means and 95% confidence intervals were generated by bootstrap analysis of percent composition for common repeat groups for each sample.

B) Comparison of *Striga* samples for the most repetitive *copia*, *gypsy*, LINE, and tandem repeats. Means and 95% confidence intervals were generated by bootstrap analysis of percent composition for the most highly repetitive repeats for each sample.

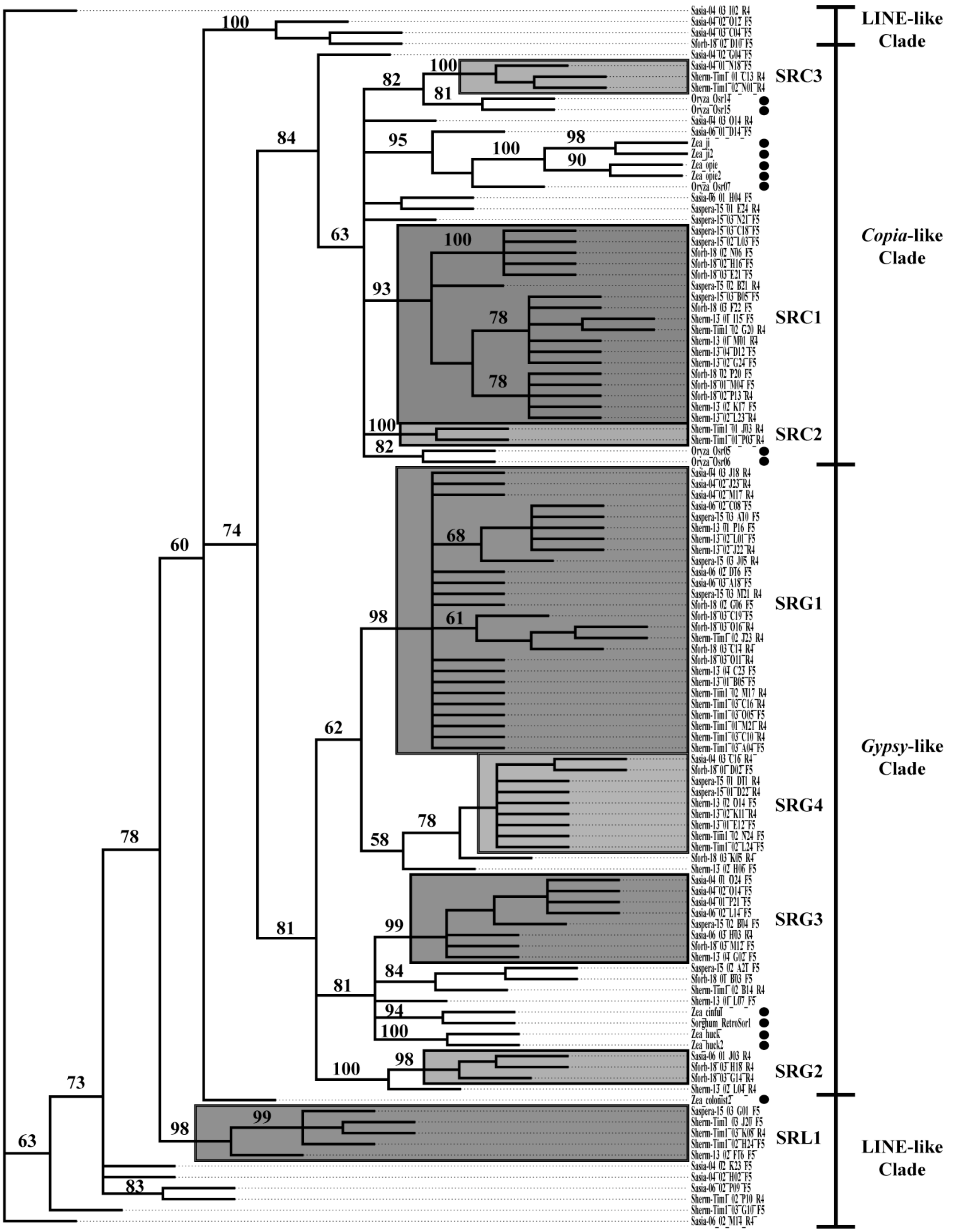


Figure 3-4: Neighbor-joining analysis of *Striga* reverse transcriptase sequences. Unrooted Neighbor-joining analysis of 97 *Striga*, 8 *Zea*, 5 *Oryza*, and 1 *Sorghum* reverse transcriptase sequences provides support for the monophyly of the LTR retrotransposons identified with AAARF. Each retroelement is boxed and shaded within the tree and identified to the right of taxon names. Bootstrap values >50 are indicated at supported nodes. A closed circle indicates grass retroelements that were also analyzed on this tree.

CHAPTER 4

DEVELOPMENT OF MICROSATELLITE MARKERS FOR CHARACTERIZING DIVERSITY IN A PARASITIC WITCHWEED, *STRIGA HERMONTHICA* (OROBANCHACEAE).[3]

[3] Estep, M., Van Mourik, T., Muth, P., Guindo, D., Parzies, H., Koita, O., Weltzien, E., Bennetzen, J. Accepted by *Molecular Ecology Resources*.
Reprinted here with permission of publisher, 07/06/2010.

Abstract:

Twelve polymorphic microsatellite loci were identified in the root parasite witchweed (*Striga hermonthica* (Del.) Benth.). These twelve loci exhibited 4-16 alleles each across a population of 32 *S. hermonthica* accessions, resulting in a total of 114 alleles. These are the first co-dominant markers developed within the genus *Striga* and can be used for investigations of inheritance, genetic diversity or population genetics in *S. hermonthica*, including our planned studies of witchweed populations across Mali and sub-Saharan Africa.

Main:

Witchweeds of the genus *Striga* belong to the parasitic plant family Orobanchaceae and exhibit a natural distribution across much of Africa, Mediterranean Europe and south Asia (Mohamed *et al.*, 2001). The genus is composed of 28 species that are all annuals and display diverse reproductive strategies ranging from fully inbreeding (*S. gesnerioides*) to obligate out-crossing (*S. hermonthica*). Natural low-density populations of *Striga* species can be found in tropical grasslands across much of Africa and parts of Asia. Five agronomically important species: *S. hermonthica, S. asiatica, S. forbesii, S. aspera,* and *S. gesnerioides,* have adapted to parasitizing several crop species and are problematic weeds for subsistence farmers.

Striga hermonthica, giant witchweed, is one of the most significant biological constraints on agriculture in sub-Saharan Africa (Ejeta, 2007b). It is typically found in agricultural fields parasitizing a diverse group of grass crops, including *Sorghum bicolor* (sorghum), *Pennisetum glaucum* (pearl millet), *Zea mays* (maize), and *Oryza sativa* (rice). Its population densities can reach thousands, sometimes millions, of individuals per hectare and more than one hundred thousand seeds per m^2 in the soil (Van Mourik, 2007). An individual plant can produce more than a hundred thousand seeds per individual plant in a single season (Van Mourik, 2007). These tiny seeds are mainly dispersed by anthropogenic activity, but wind, water and forage animals have also been shown to play a dispersal role (Berner *et al.*, 1994). A seed bank study in the USA has shown germination viability up to 12 years after burial (Bebawi *et al.*, 1984). Other studies in Africa have shown considerable portions of the total seeds (>40%) either dying or germinating in any given growing season (Oswald & Ransom, 2001; Van Mourik, 2007).

Microsatellites, otherwise known as simple sequence repeats (SSRs), with two, three or four nucleotide motifs were identified from a group of genome survey sequences (GenBank entries FI774410-FI778765) from *S. hermonthica* total genomic DNA with the Perl Script SSRIT (Temnykh *et al.*, 2001). Fifty primer pairs were designed to amplify products between 100 and 300 bp for these SSRs using the program Primer3 (Rozen & Skaletsky, 2000). An M13 universal priming site was added to the 5’ end of the forward primer. Using a three primer system with a universal fluorescent labeled primer (VIC, FAM, NED, PET) (Schuelke, 2000), each primer pair was tested on DNA from a panel of thirty-two individual *S. hermonthica* plants and scored for reproducibility, polymorphism, and the presence of no more than one or two distinct products per individual. Total genomic DNA was extracted from dried leaf tissue for these thirty-two plants using a modified CTAB extraction (Doyle & Doyle, 1987). PCR reactions were 10 ul. Each reaction consisted of 5-10 ng of template DNA, 0.6 U of *Taq*, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTP's (each), 1.25 mM forward primer, 1.25 mM fluorescently labeled M13 primer, and 2.5 mM reverse primer. A touchdown PCR program was used on a MJ Research PTC-200 Peliter Thermocycler, consisting of an initial denaturation cycle of 94 °C for 5 min; 10 cycles at 94 °C for 45 s, 68 °C (-2 °C per cycle) for 5 min, elongation at 72 °C for 1 min; 5 cycles at 94 °C for 45 s, 58 °C for 2 min, elongation at 72 °C for 1 min; 25 cycles of 94 °C for 45 s, 50 °C for 2 min, elongation at 72 °C for 1 min; and completing with a final 30 min elongation at 72 °C. Reactions with different fluorescent labels were then multiplexed with a LIZ 500 standard and separated on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA). Resultant chromatograms were scored using ABI GeneMapper software (version 4.0). The software package Micro-Checker was used to test for null alleles, stuttering, and large allele drop out (Oosterhout *et al.*, 2004).

Twelve polymorphic microsatellite markers were chosen from the set of fifty and were further characterized using GENALEX and PowerMarker software (Liu & Muse, 2005; Peakall & Smouse, 2006). A total of 114 alleles were identified with a range per locus of 4-16 alleles and an average of 9.5 alleles/locus (Table 4-1). Hardy-Weinberg equilibrium (HWE) tests suggest that each locus is evolving as expected under HWE assumptions, except microsatellite marker SH1012 that has a marginal significance. Tests for gametic disequilibrium revealed no significant linkage between markers. Micro-checker suggested the possibility of null alleles for SH1029 and SH1042, based on the excess of homozygotes. No stuttering or large allele drop out was detected for any of the twelve markers.

These twelve loci are reliable co-dominant markers that can be used to investigate the diversity of *S. hermonthica* populations and used to examine biological questions such as genetic diversity, gene flow between populations and regions, and genotyping samples of seed banks germinated in the laboratory.

Table 4-1: Characterization of 12 microsatellite markers and polymorphism [number of alleles (A), observed (H_O) and expected (H_E) heterozygosity, Hardy-Weinberg equilibrium probability (HWE), polymorphism information content (PIC)] detected in 30 accessions of *Striga hermonthica*

	Core		Product	(n=30)					GenBank
Locus	sequence	Primer sequence (5'-3')	length (bp)	A	H_O	H_E	HWE	PIC	accession no.
SH1005	$(TG)_{12}$	FAM-CGATCGCCTCCTGGATACTA TCGGAAAAATTGCGAAAAAC	224-229	4	0.700	0.585	0.213	0.532	FI776044
SH1008	$(ATTT)_8$	NED-CCGTGACCTCGATGAAGATT CCGCAACGTAAATTCCAAGT	293-306	4	0.633	0.619	0.312	0.542	FI774622
SH1009	$(TC)_{25}$	VIC-GCATCCAGATAAGGCTGCTT TGGGTTGTGTTGAGTGAGTGA	223-255	15	0.733	0.834	0.311	0.820	FI775250
SH1012	$(ACAT)_5$	PET-TGGATAAGGCCTTTTGTGAGA GCAACAGCCCATTTGAGTCT	303-314	6	0.667	0.802	0.0455*	0.772	FI775982
SH1014	$(TTC)_7$	FAM-AGGGACATTATGCAGCCAAC CGCATGACGAACAAGAAGAA	176-184	5	0.533	0.640	0.835	0.570	FI775983
SH1016	$(TACA)_{17}$	VIC-GATTTGGATATCGCGGTTGT TTCTGGCGATGAAAATGACA	250-322	13	0.833	0.834	0.731	0.818	FI775081
SH1029	$(CA)_{10}$	NED-CTTATAGCCCGCATGCAATC CCCCTCCGTTCAGTTCAGTA	258-276	16	0.700	0.901	0.162	0.893	FI776123
SH1030	$(AG)_{17}$	PET-TGTCTCGTTCCGTCCTCTCT GCAATGCAGGTAGCCTCCTA	231-262	15	0.767	0.849	0.477	0.837	FI775072
SH1032	$(GT)_{14}$	PET-TATCGAGTCGGGAAAGATGC CATTCCACACCCACTACACG	287-293	4	0.733	0.632	0.111	0.573	FI774526
SH1038	$(TGA)_9$	FAM-TCAGTGGTGCAGGTTAACGA CTGCAGCATGGAAGTTCGTA	231-283	16	0.867	0.909	0.511	0.903	FI775577
SH1041	$(TAA)_8$	NED-GGAGTGGCCAGGATCATTTA TCCCCGGGCTCTTAGTTAAT	159-201	11	0.767	0.803	0.659	0.781	FI775578
SH1042	$(TATG)_5$	VIC-CTCATTCCCTCGCTTTCTTG TTTCTGCGTTTTGTTTTGGA	274-284	5	0.567	0.767	0.163	0.729	FI775002
			Mean	9.5	0.708	0.765		0.731	

CHAPTER 5

GENETIC DIVERSITY AND POPULATION STRUCTURE OF *STRIGA HERMONTHICA*, PARASITES OF SORGHUM AND MILLET IN MALI[4]

[4] Estep, M., Van Mourik, T., Muth, P., Guindo, D., Parzies, H., Koita, O., Weltzien, E., Bennetzen, J. To be submitted to *Molecular Ecology*

Abstract:

Eleven populations of *Striga hermonthica*, collected in Mali, were investigated with twelve microsatellite markers to reveal a large amount of genetic diversity that is broadly distributed across populations with little genetic differentiation and large amounts of gene flow. Some population structure was apparent, but could not be attributed to "isolation by distance" or host species, suggesting other geo-ecological variables may be acting to differentiate Northern populations from Southern populations.

Introduction:

Striga hermonthica is a devastating agricultural weed that parasitizes grain crops, such as sorghum (*Sorghum bicolor*) and millet (*Pennisetum glaucum*) throughout Sub-Saharan Africa. This species belongs to a genus of parasitic plants that contains at least two other agricultural weeds with similarly dramatic effects on staple crop production (*S. asiatica* and *S. gesnerioides*). Together, these species of witchweed are the most important biological limitation to food production in Africa (Ejeta, 2007b).

The initial stages of the *Striga*::host interaction can be broken down into four key steps: seed germination, host attachment, haustorium formation, and penetration of the root vascular system (Yoshida & Shirasu, 2009). *Striga* seeds have evolved to recognize the chemical signals, strigolactones, which are involved in attracting arbusculuar mycorrhizal fungi which have beneficial interactions with the plant host (Akiyama K *et al.*, 2005; Matusova *et al.*, 2005). Several different strigolactones can induce germination and chemotropic growth of *Striga* seedlings towards the roots of a possible host. *Striga* is an obligate parasite and must attach to a host plant before nutrient stores in the tiny seed are exhausted. Firm attachment to the host root is accomplished via a structure known as a haustorium (Keyes *et al.*, 2001). Once attachment is complete, *Striga* forms a connection with the host vascular system and begins parasitizing the host (Bar-Nun *et al.*, 2008).

Efforts to identify crops with resistance to *Striga* have yielded very mixed results, with essentially no success to date with the hosts maize and pearl millet, ephemeral success with cowpea for resistance to the legume parasite *S. gesnerioides* (Timko *et al.*, 2007) and only partial success with sorghum (Ejeta, 2007a), the host species that is most likely to have co-evolved with *S. hermonthica* (Musselman, 1987). In order to better understand the diversity of *Striga* species, and whether it impinges on host resistance, several studies have been conducted to examine population level genetic diversity and identify the existence of races or structure within *Striga* populations (reviewed in (Mohamed *et al.*, 2007)). Both *S. asiatica* and *S. gesnerioides* are mainly autogamous (self-fertilization) and genetic diversity analyses have shown distinct races of both species across their ranges (Shawe & Ingrouille, 1993; Botanga *et al.*, 2002; Botanga & Timko, 2006).

S. hermonthica is an obligate out-crossing species (Safa *et al.*, 1984), so it is expected to show less differentiation between populations and greater diversity within populations than seen in related autogamous species (Hamrick, 1982). In agreement with this prediction, multiple genetic diversity studies using allozymes, Randomly Amplified Polymorphic DNA (RAPD), or Amplified Fragment Length Polymorphism (AFLP) markers have not convincingly demonstrated the existence of races in *S. hermonthica*, although extensive genetic diversity was observed (Bharathalakshmi *et al.*, 1990; Kuiper *et al.*, 1996; Olivier *et al.*, 1998; Koyama, 2000; Gethi *et al.*, 2005). In those studies where *S. hermonthica* was collected from different countries (on different sides of the continent) a geographic distance effect was noted. None of these studies demonstrated a genetic component to host specificity. However, the number of populations investigated, the number of loci analyzed, or the type of marker employed limited all of these studies.

The most powerful approach to characterizing genetic diversity in *S. hermonthica* would employ a robust set of reproducible, neutrally evolving, and co-dominant markers. Simple Sequence Repeat (SSR) markers with these properties have recently been developed for *Striga* (Estep *et al.*, 2010). In addition to robust markers, an optimal study of genetic diversity in *S. hermonthica* would employ populations collected in multiple years, parasitizing multiple staple crops (e.g., sorghum, pearl millet, maize, rice), in multiple agro-ecosystems (e.g., Sahel grasslands, Sudan savanna), and from across the species range at both macro and micro scales. In this manuscript, results from a first collection year are presented to describe the diversity of *S. hermonthica* across a broad swath of environments and agricultural zones in Mali, a nation that is dramatically impacted by *Striga* parasitism.

Methods:

Sampling of *S. hermonthica* populations:

Individual plants were collected from four regions (designated 3000, 4000, 5000, & 6000) of southern Mali (Fig 5-1). A total of eleven populations (agricultural fields) were chosen late in the growing season when *S. hermonthica* was in flower and easily identified. Each field chosen was under cultivation by farmers growing sorghum (*Sorghum bicolor*) or millet (*Pennisetum glaucum*) as a subsistence crop (Table 5-1). A population consisted of a minimum of 20 individuals collected along two linear transects, at 90 degree angles to each other, in a single agricultural field (~100m x ~100m). Leaves from individual plants were air-dried or stored in silica gel and most were shipped to the University of Hohenheim, Germany for DNA extraction. Some DNA extractions were carried out at the University of Bamako in Mali. Each region consisted of two populations within close proximity and one population ~10-20 km away, except region 6000 where only two populations were collected.

Region 3000 (Segou) supplied 3 populations (3700, 3800, 3900) to this study and is the farthest north and east within the study area. This region is near the southern boarder of the Sahel grassland ecosystem and receives ~600mm of rain per year. Population 3700 was collected in a sorghum field near the village of Souara in the Tominian district (13.12.466 N, 04.36.133 W). Population 3800 was collected in a pearl millet field ~ 100m from population 3700 (13.12.586 N, 04.35.928 W). Population 3900 was collected in a pearl millet field near the village of Madiama in the Djenne district (13.46.912 N, 04.23.691 W).

Region 4000 (Kati) provided 3 populations (4100, 4200, 4300) and is the farthest south within the study area. This region is part of the Sudan savanna and receives 800-1000mm of rain per year. Population 4100 was collected in a sorghum field near the village of Farabana in the Kangaba district (12.27.705 N, 08.07.487 W). Population 4200 was collected in a sorghum field near the village of Sindala in the Kati district (12.23.065

N, 08.18.013 W). Population 4300 was also collected in a sorghum field, ~100m from population 4200 (12.23.028 N, 08.17.991 W).

Region 5000 (Dioila) contributed 3 populations (5400, 5500, 5600). This region is also part of the Sudan savanna and receives 800-1000mm of rain per year. Population 5400 was collected in a sorghum field near the village of Wakoro in the Dioila district (12.35.540 N, 06.42.483 W). Populations 5500 and 5600 were collected in sorghum fields near the village of Tonga in the Dioila district within ~100m of each other (12.39.160 N, 06.45.909 W and 12.39.211 N, 06.45.916 W, respectively).

Region 6000 (Kayes) supplied 2 populations (6700, 6800). This region is the farthest west within the study area. This region is not part of the Sahel or Sudan ecosystems and is in a mountainous region that receives >1200mm of rain per year. Both populations were collected in sorghum fields near the village of Sagabari in the Sagabari district, within ~ 200m of each other (12.35.467 N, 09.50.178 W and 12.35.438 N, 09.50.240 W, respectively).

DNA extraction and marker amplification:

In Mali, DNA isolation was performed using Plant DNAzol Reagent following the manufacturer's protocol (Invitrogen, Carlsbad, CA). Once samples were precipitated, washed, and air-dried, they were shipped to the University of Georgia, USA and re-hydrated in 100ul of 1x TE buffer (100mM Tris-HCl, 10 mM EDTA @ pH 8.0). Samples shipped to the University of Hohenheim were homogenized using a TissueLyzer (Qiagen, Valencia, CA.) and total genomic DNA was extracted using a modified CTAB protocol (Doyle, Doyle, 1987).

A set of twelve neutral, non-coding, and co-dominant microsatellite markers was used to access the genetic diversity and genetic structure of the collected individuals (Estep *et al.*, 2010). PCR reactions were 10 ul, using a three-primer system with an M13 universal fluorescent-labeled primer (VIC, FAM, NED, PET) (Schuelke, 2000). Each reaction consisted of 5-10 ng of template DNA, 0.6 U of *Taq*, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTP's (each), 1.25 mM forward primer, 1.25 mM fluorescently labeled M13 primer, and 2.5 mM reverse primer. A touchdown PCR program was used on a MJ Research PTC-200 Peltier Thermocycler, consisting of an initial denaturation cycle of 94 °C for 5 min; 10 cycles at 94 °C for 45 s, 68 °C (-2 °C per cycle) for 5 min, elongation at 72 °C for 1 min; 5 cycles at 94 °C for 45 s, 58 °C for 2 min, elongation at 72 °C for 1 min; 25 cycles of 94 °C for 45 s, 50 °C for 2 min, elongation at 72 °C for 1 min; and completing with a final 30 min elongation at 72 °C. Reactions with different fluorescent labels were then multiplexed with a LIZ 500 standard and separated on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA). Resultant chromatograms were scored using ABI GeneMapper software (version 4.0).

Data analysis:

Descriptive population genetic statistics were calculated using GenAlEx and Genepop (Peakall & Smouse, 2006; Rousset, 2008). These included allelic diversity (richness), effective allelic diversity (evenness), expected heterozygosity (gene diversity), observed heterozygosity, and the fixation index. An Analysis of Molecular Variance (AMOVA) was conducted to estimate F-statistics and to estimate the number of migrants (Nm). Nei's standard genetic distance was calculated between pairs of populations for use in a principal coordinates analysis (PCA). Geographic distance was estimated from location coordinates (Table 5-1) and used in conjunction with the Nei's genetic distance matrix to test for isolation by distance (Mantel's test). A second Mantel's test was conducted using the same matrix for geographic distance and a second matrix of pairwise comparisons of the number of migrants (Nm). The same matrices were used in random permutation tests to test for significance of each Mantel's test result.

An analysis of population structure and individual population assignments was conducted with the raw data in the program STRUCTURE (Pritchard *et al.*, 2000). Predefined numbers of populations (K) ranged from 1-11 and an initial burn-in period of 20,000 replicates and 50,000 Markov Chain Monte Carlo (MCMC) iterations were used. Five independent simulations were run for each K value. To identify the number of populations that best reflect the structure of our sample, the average K value was calculated from the five runs and Delta K was calculated as in Evanno *et al.* (2005).

Results:

An average of 250 individuals were scored for each of the twelve microsatellite markers and a total of 181 alleles were identified within our sample data (Table 5-2). Values for the mean allelic diversity (richness) ranged from 7.0 in population 6800 to 10 in population 3900 (Table 5-2). Mean effective allelic diversity (evenness) values ranged from 4.0 in population 3700 to 5.9 in population 3900. There is no significant difference in the values for richness or evenness between populations. Values for the mean expected heterozygosity (gene diversity) ranged from 0.687 in population 4300 to 0.748 in population 3900 and no significant differences were observed between populations (Table 5-3). Similarly, values for the mean observed heterozygosity ranged from 0.689 in population 3800 to 0.783 in population 3700 with no significant differences between populations. The mean fixation index was calculated to show in which direction populations were trending out of Hardy-Weinberg proportions with seven populations appearing to have excess heterozygosity (negative values) and the remaining four populations indicating inbreeding or excessive homozygosity (positive values) in comparison to gene diversity (Table 5-3). The number of private alleles was also counted indicating that region 3000 exhibited 22, region 5000 yielded 20, region 4000 had 6, and region 6000 yielded 2.

An AMOVA was conducted to estimate the Rst (analogue of Fst) using a stepwise mutation model specific for microsatellite data (Slatkin, 1995). The resulting Rst value was 0.048 and Ris and Rit were both 0.991. A majority of the variance (95%) can be explained by within population variation. The remaining variance can be explained by among region variance (2%) and among population variance (3%). The number of migrants was also calculated from the Rst value using the equation Nm = (1-Rst)/4*Rst, resulting in Nm = 4.921.

A principal coordinates analysis was conducted using pairwise comparisons of Nei's standard genetic distance to identify major patterns within our data set (Fig 5-2). Coordinate one (x-axis) explains 33.29% of the variance in our data and splits the four collection regions into two distinct groups. Group one (on the left) contains individuals from region 4000 and region 5000. Group two (on the right) contains individuals from region 3000 and region 6000. The second coordinate (y-axis) explains 15.66% of the variance and does not appear to further divide the two groups obtained from coordinate one. While the two major groupings are distinct, one to a few individuals from region 3000 and 6000 appear to be placed within or near group one (Fig 5-2).

The program STRUCTURE was used to identify population structure within the data set (Fig 5-3). The predefined K=2 simulations had the highest value for Delta K among all predefined K values ranging from 1 to 11 (data not shown). This analysis grouped regions 4000 and 5000 as one cluster and grouped regions 3000 and 6000 as a second group. It is important to note that several individuals within these two groups exhibit admixture between the two groups.

A Mantel's test was performed with two data matrices, the first was a pairwise geographic distance matrix and the second was a pairwise Nei's standard genetic distance matrix (Fig 5-4). The resulting R^2 value was 0.0975 and was found not to be significant (P=0.06) based on 99 random permutations of the two original data matrices. A second Mantel's test was performed with the same geographic distance matrix and a pairwise matrix of the number of migrants. The resulting R^2 value was 0.2388 and was found to be significant (P=0.01), based on 99 random permutations of the two data matrices.

Discussion:

In this study, populations selected from Mali were characterized with 12 microsatellite markers to describe the genetic diversity and structure of *S. hermonthica* in this region. While this is not the first study of genetic diversity within *S. hermonthica* (Bharathalakshmi *et al.*, 1990; Kuiper *et al.*, 1996; Olivier *et al.*, 1998; Koyama, 2000; Gethi *et al.*, 2005; Ali *et al.*, 2009; Yoshida *et al.*, 2010), it is the first using reproducible, neutrally evolving, and co-dominant markers.

The analyses show that *S. hermonthica* is rich in allelic variation that is fairly evenly distributed among populations over a large geographic range (> 600 km) within multiple ecosystems. It also demonstrates a high level of genetic diversity among populations, with an average gene diversity of 0.715 (range of 0.687-0.748) and a relatively low level of genetic differentiation (Rst = 0.048). The observed values of heterozygosity and the fixation indices demonstrate that 63% (seven) of the populations appear to have an excess of heterozygous individuals. This can be explained by negative assortative mating that would be expected for an allogamous (obligate out-crossing) species such as *S. hermonthica* (Safa *et al.*, 1984). All three populations from region 5000 have an excess of homozygous individuals, suggesting an undetected null allele specific to this region.

The numbers of private alleles found in each region are not evenly distributed, with populations in the eastern portion of the study area (region 3000 and 5000) having 3-10 times the number of private alleles than those in the western portion (region 4000 and 6000). These values can be interpreted as a proxy for the age of a population, due to the high rate of mutation at many microsatellite loci. This would suggest that populations in the Eastern portion of the country have occupied this area for a longer period of time, while those in the Western portion are more recent arrivals. This result is interesting because we did not observe a reduction in genetic diversity that would likely accompany a founder effect.

While there is a great deal of evenly distributed genetic diversity within the populations studied, the AMOVA analysis indicates little genetic differentiation (Rst = 0.048) has occurred. Most (95%) of the diversity can be explained by allelic variation within populations. This result also suggests a large amount of gene flow among populations, based on the large number of migrants per generation (Nm=4.921). A small number of migrants per generation is enough gene flow to reverse the process of drift, which would allow the populations to differentiate over time. Anthropogenic activity, in the form of trading contaminated crop seeds, has been shown to be frequent in subsistence agricultural systems (Berner *et al.*, 1994). Other forms of dispersal like wind, water, and forage animals have also been shown to play a dispersal role in *S. hermonthica*, but the geographic distance examined in this study likely reduces their role.

A PCA was conducted to identify possible differences between populations based on genetic distance. An unexpected result was observed, where the two regions that were most geographically separated were found to group together (regions 3000 and 6000). This observation was further analyzed using an admixture test with the program STRUCTURE where individuals without population designation are grouped based on shared genotypes. Similar results were obtained from the STRUCTURE analysis, suggesting that two distinct “strains” or “races” (for lack of a better word) of *S. hermonthica* exist within the study area.

Two of the three populations from region 3000 were collected in pearl millet fields, while the remaining population was collected from a sorghum field. We saw no differences between these three populations or between the two populations from region 6000 (both from sorghum fields) that group together as one race in the PCA and STRUCTURE analyses. This observation suggests that while distinct races may exist, we cannot infer an association with a host species. It is important to point out that region 6000 is distinctly different from any of the other collection localities. It is difficult to imagine a locally adapted variety of a subsistence crop (sorghum or pearl millet) that would thrive in both the 3000 region with less than 600mm of rain per year and the 6000 region with more than 1000mm of rain per year. It has been reported that pearl millet-adapted strains of *S. hermonthica* exist and it has been experimentally shown that those strains designated as specific to pearl millet have no problems parasitizing sorghum cultivars while the reverse (sorghum adapted strains parasitizing pearl millet) is not true (Parker & Reid, 1979). This suggests that *S. hermonthica* "races" may be adapted to ecosystems (Sahel grasslands and Sudan Savanna) or some other geo-ecological variable, rather than (or in addition to) adaptation to hosts. This would help explain why, in the northern portions of the study area (even at great geographic distance and on different hosts), one "race" of *S. hermonthica* is found, while a second "race" is found in the southern regions, on the same host species. Sorghum and pearl millet are grown in both the Sahel ecosystem and the Sudan ecosystem, but the majority of pearl millet is grown in the Sahel and the majority of sorghum is grown in the Sudan. Further population collections and geo-ecological data collections are planned to address these questions.

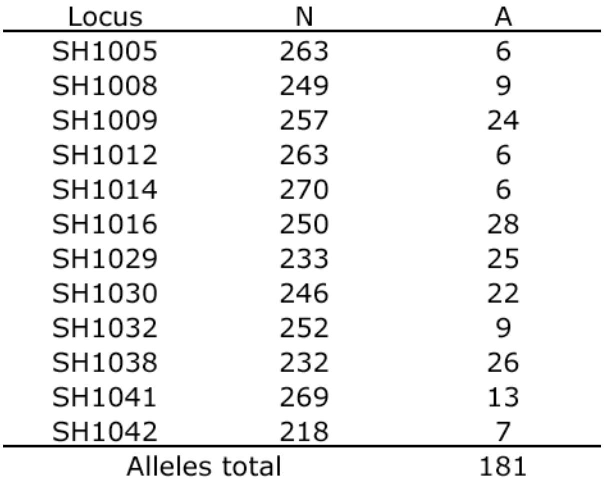
To better understand what processes are differentiating the "races" identified, we tested the hypothesis of isolation by distance with a Mantel's test. We did not observe a significant relationship between genetic distance and geographic distance, arguing that geographic distance is not responsible, at least by itself, for the differentiation of the two "races" identified. We also tested for a relationship between the amount of gene flow (Nm) and geographic distance. In this analysis we did find a significant negative relationship, suggesting that gene flow is occurring at short geographic distances but is greatly reduced at larger geographic distances. This result seems obvious, but affects the interpretation of long distance dispersal from the older 3000 region populations to the younger 6000 region populations and leads us to further argue that the identified "races" are possibly geo-ecosystem specific regardless of host.

Plant breeders are working to produce genotypes of sorghum and pearl millet that are resistant to *Striga* parasitism (Ejeta, 2007a). In many plant::pathogen interactions, resistance tends to be specific to a particular pathogen race, thus having a major impact on the durability of the resistance as individual races wax and wane in the agricultural environment. Hence, it would be very useful to identify any possible *S. hermonthica* genotypes (races) that exhibit different parasitic qualities. These genotypes, once discovered and characterized, can then be used to identify individual resistance genes in crop (host) germplasms and can be used to pyramid multiple resistance genes into a targeted crop plant. In order to fully characterize the existence of "races" and the factors driving their formation, further collections of *S. hermonthica* populations and their hosts are needed. Our working hypothesis argues for a Northern "race" in areas where pearl millet is the dominant crop and a Southern "race" in areas where sorghum is the dominant crop. These "races" may not be host species-specific, but instead may be under selective pressures from other environmental factors, yet to be identified, and/or could be host genotype-specific. The amount of precipitation, mean temperature, soil quality, available pollinators, or lengths of growing season are major environmental differences in the ecosystems that were sampled. Many of the same environmental factors are certainly driving the farmers' choice of host genotype in any given region. This analysis demonstrates extensive gene flow working to homogenize populations, likely caused by the small seed size, impressive parasite fecundity and the probable exchange of *Striga*-contaminated host seeds. These results also suggest that field screening for resistant varieties in Mali would be most appropriate if conducted in both Northern and Southern Malian environments.

Table 5-1: Collection sites with location and host crop.

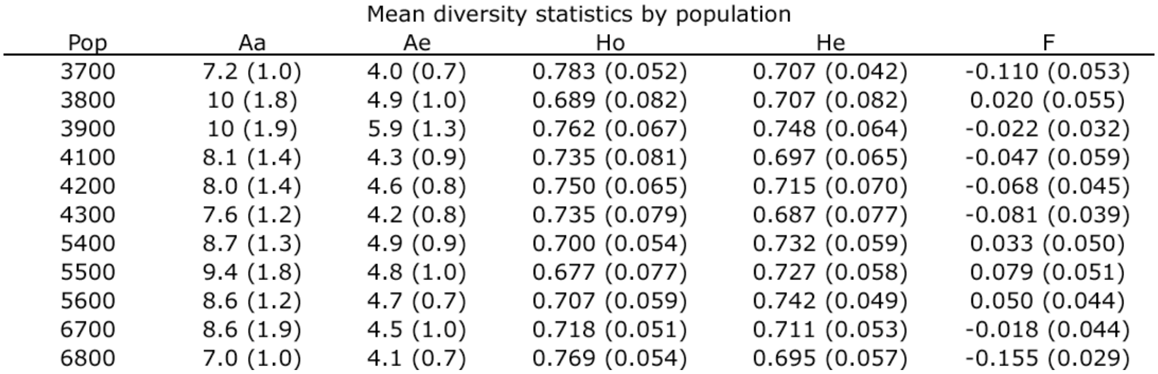
Region	Population	Latitude/Longitude	Host
3000	3700	13.12.466 N, 04.36.133 W	sorghum
	3800	13.12.586 N, 04.35.928 W	millet
	3900	13.46.912 N, 04.23.691 W	millet
4000	4100	12.27.705 N, 08.07.487 W	sorghum
	4200	12.23.065 N, 08.18.013 W	sorghum
	4300	12.23.028 N, 08.17.991 W	sorghum
5000	5400	12.35.540 N, 06.42.483 W	sorghum
	5500	12.39.160 N, 06.45.909 W	sorghum
	5600	12.39.211 N, 06.45.916 W	sorghum
6000	6700	12.35.467 N, 09.50.178 W	sorghum
	6800	12.35.438 N, 09.50.240 W	sorghum

Table 5-2: Total number of individuals (N) scored for each microsatellite marker and the total number of alleles (A) identified.



Locus	N	A
SH1005	263	6
SH1008	249	9
SH1009	257	24
SH1012	263	6
SH1014	270	6
SH1016	250	28
SH1029	233	25
SH1030	246	22
SH1032	252	9
SH1038	232	26
SH1041	269	13
SH1042	218	7
Alleles total		181

Table 5-3: Mean descriptive population genetic statistics for each population followed by standard deviation values. Allelic diversity (Aa), Effective allelic diversity (Ae), Observed heterozygosity (Ho), Expected heterozygosity (He), and Fixation index (F).



Mean diversity statistics by population

Pop	Aa	Ae	Ho	He	F
3700	7.2 (1.0)	4.0 (0.7)	0.783 (0.052)	0.707 (0.042)	-0.110 (0.053)
3800	10 (1.8)	4.9 (1.0)	0.689 (0.082)	0.707 (0.082)	0.020 (0.055)
3900	10 (1.9)	5.9 (1.3)	0.762 (0.067)	0.748 (0.064)	-0.022 (0.032)
4100	8.1 (1.4)	4.3 (0.9)	0.735 (0.081)	0.697 (0.065)	-0.047 (0.059)
4200	8.0 (1.4)	4.6 (0.8)	0.750 (0.065)	0.715 (0.070)	-0.068 (0.045)
4300	7.6 (1.2)	4.2 (0.8)	0.735 (0.079)	0.687 (0.077)	-0.081 (0.039)
5400	8.7 (1.3)	4.9 (0.9)	0.700 (0.054)	0.732 (0.059)	0.033 (0.050)
5500	9.4 (1.8)	4.8 (1.0)	0.677 (0.077)	0.727 (0.058)	0.079 (0.051)
5600	8.6 (1.2)	4.7 (0.7)	0.707 (0.059)	0.742 (0.049)	0.050 (0.044)
6700	8.6 (1.9)	4.5 (1.0)	0.718 (0.051)	0.711 (0.053)	-0.018 (0.044)
6800	7.0 (1.0)	4.1 (0.7)	0.769 (0.054)	0.695 (0.057)	-0.155 (0.029)

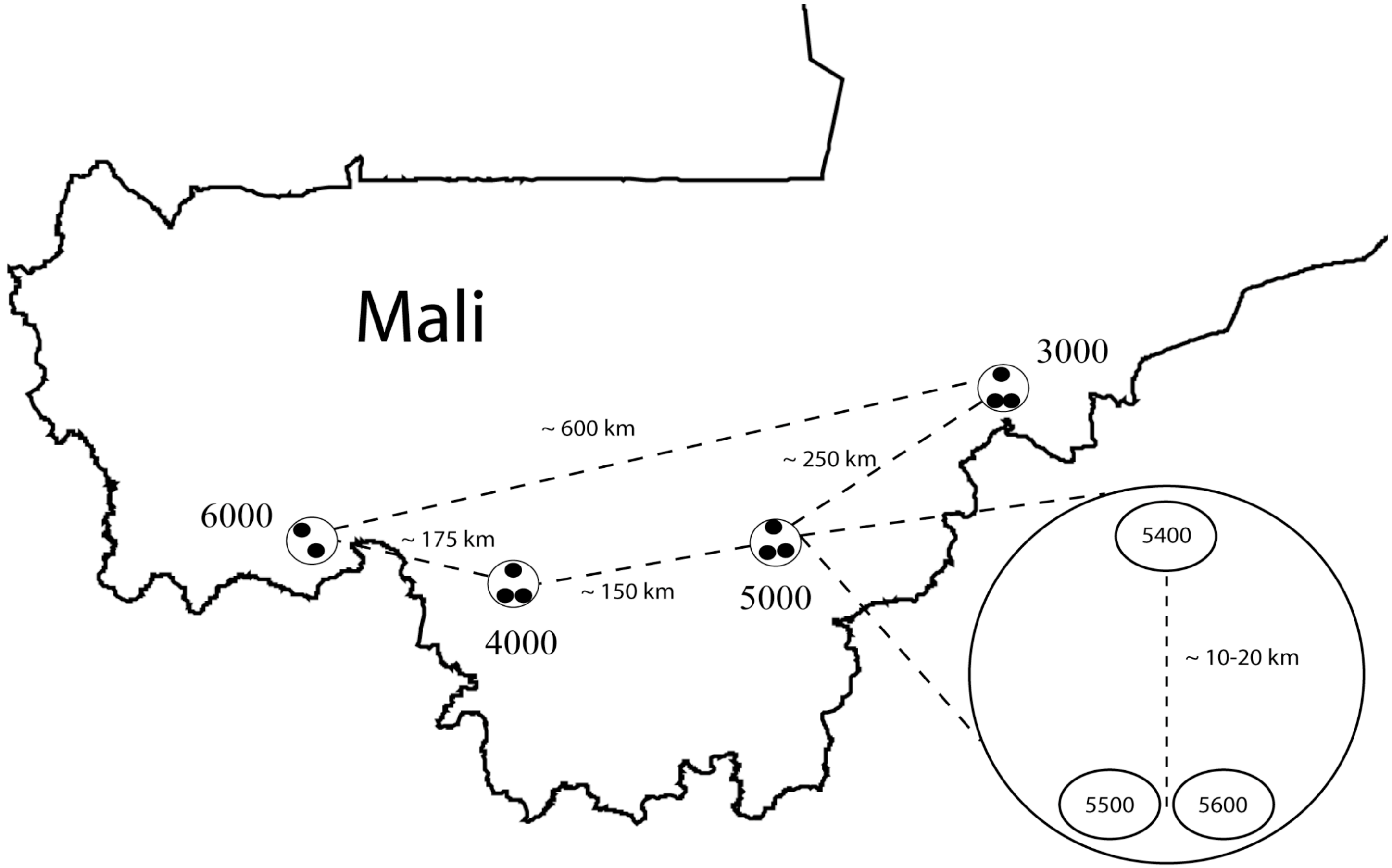


Figure 5-1. Map of collection sites with regional locations and approximate geographic distances.

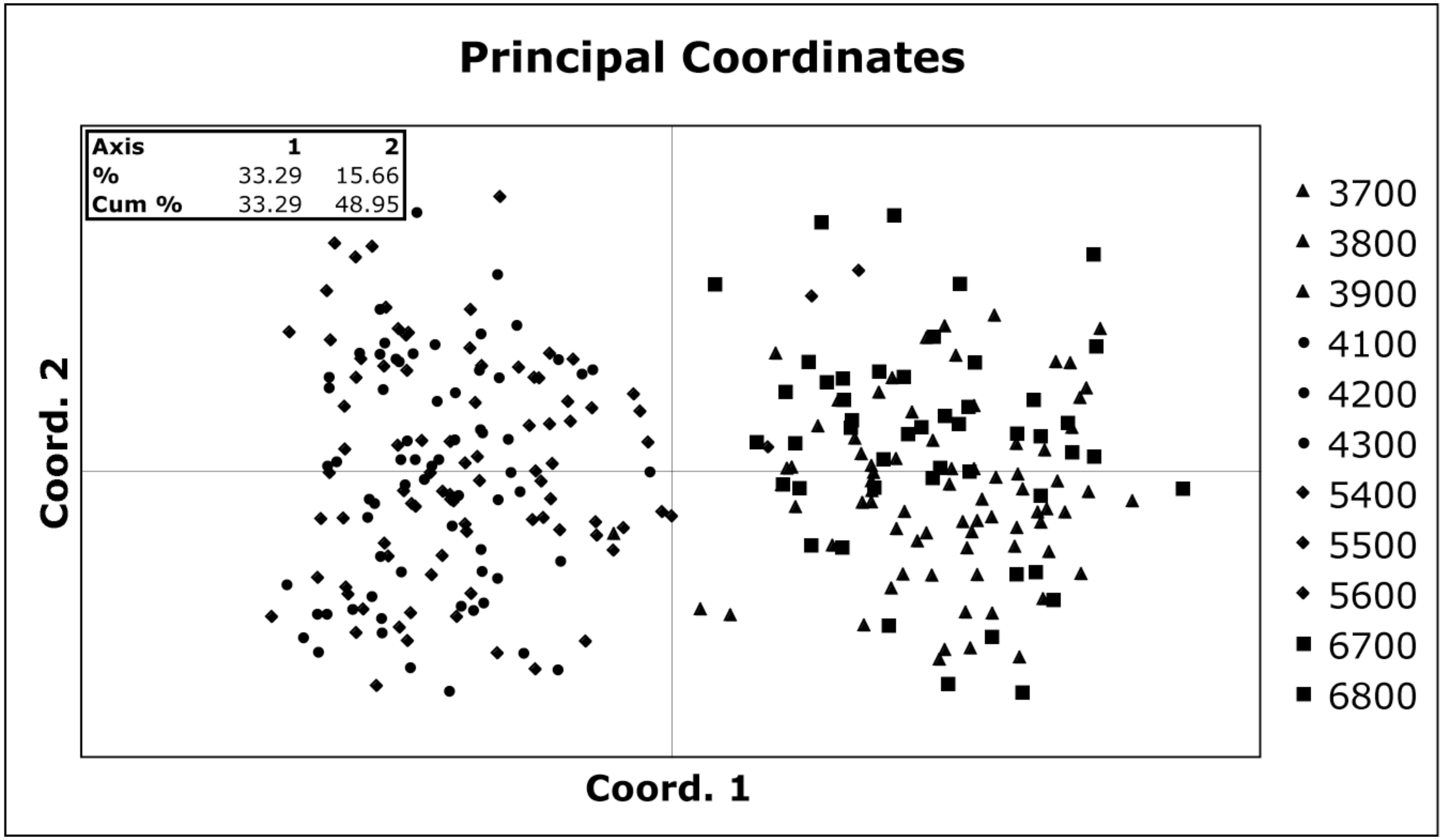


Figure. 5-2. A principal coordinates analysis of pairwise genetic distance between populations.

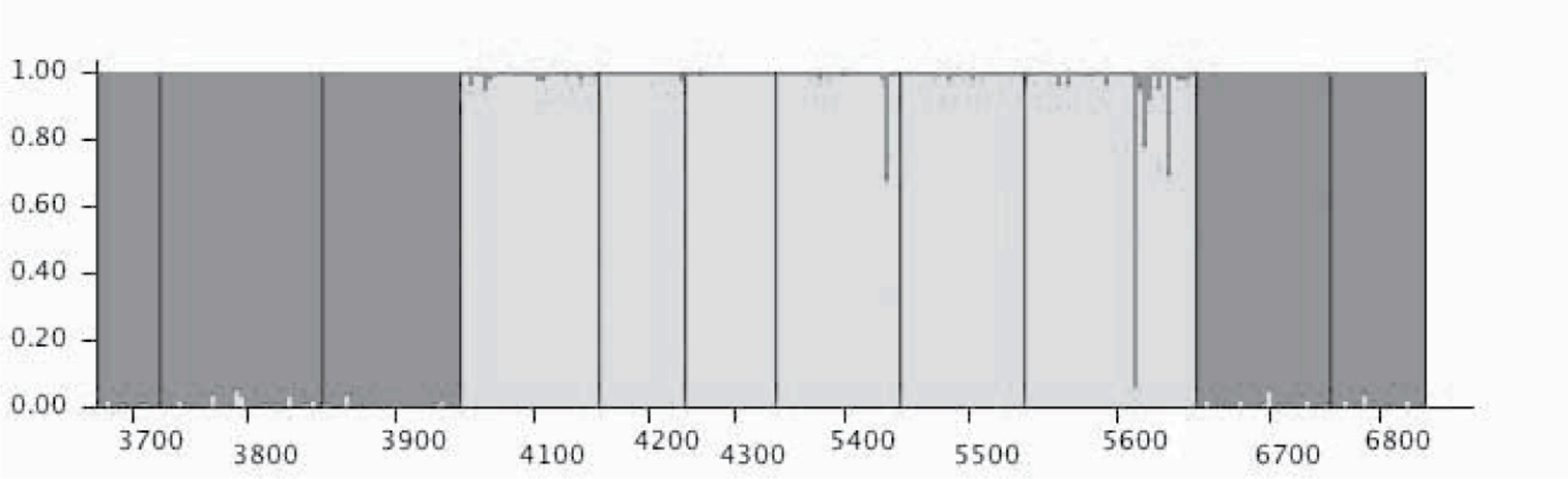


Figure 5-3. Results of population structure analysis with K=2.

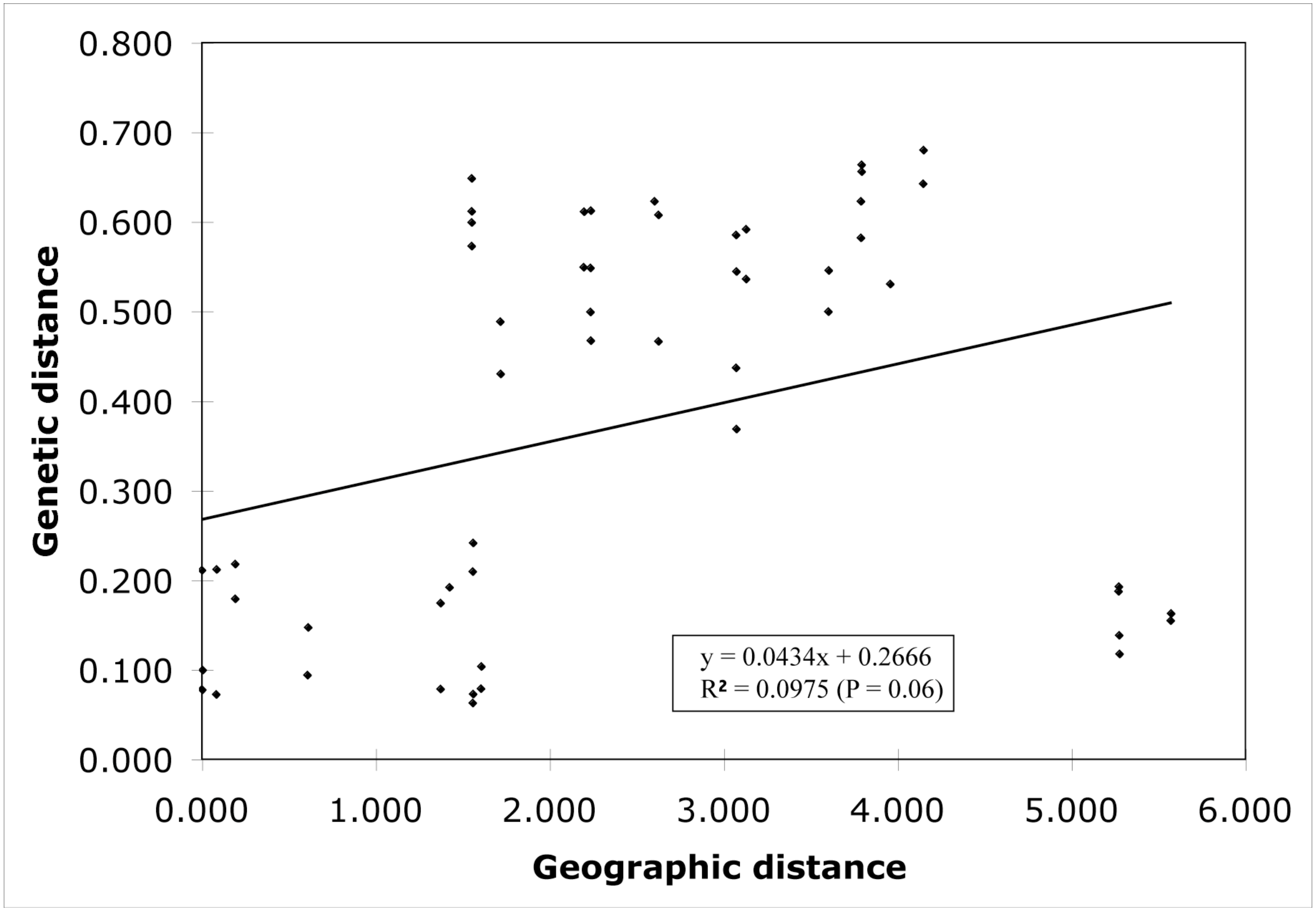


Figure 5-4. Graph of Mantel test for relationship between genetic distance and geographic distance.

CHAPTER 6

CONCLUDING REMARKS

The devastating effect of Striga parasitism on subsistence farming in much of Africa is a complicated problem caused by cultural practices, poor quality soils and very limited agricultural inputs. The accidental introduction of a single genotype of *S. asiatica* to the US more than 50 years ago (Garris & Wells, 1956), still not eradicated despite a USDA facility devoted exclusively to that task, has proven how persistent witchweed can be. The genus, including the three most invasive species (*S. asiatica*, *S. gesnerioides*, and *S. hermonthica*) share many of the same biological properties, but each exhibit slightly different approaches in life history traits like host preference, reproduction, and geographic range that have undoubtedly acted to diversify the genus.

The phylogenetic analysis presented in chapter 2 describes three distinct lineages within the genus with each containing at least one agricultural weed. This result argues that the shift in life histories associated with the agricultural weeds is not unique and that any of the *Striga* species has the ability to switch hosts (e.g., natural grassland or legume species to agricultural crop) to become an agricultural problem. Of course, all agricultural weedy species started out as natural species and some may have co-evolved with their current host. For instance, it is believed that both *S. hermonthica* and *Sorghum bicolor* arose in Eastern Africa (present day Ethiopia and Sudan), and have dramatically co-expanded their ranges over the last few thousand years with the spread of sorghum agriculture (Musselman & Hepper, 1986). This model suggests that *S. hermonthica* genetic diversity is most likely to radiate from this same location, and our preliminary results with S. *hermonthica* from Eastern and Western Mali agree with this prediction. In this regard, it will be particularly interesting to investigate the genetic properties of *S. hermonthica* found in the Nuba Mountains region of Sudan on wild grasses. If these are not the product of over-whelming gene flow from "domesticated" *S. hermonthica* found on nearby crops, then one might expect to find a high frequency of unique *Striga* gene alleles that differentiate populations.

This "multiple domestication model" for different *Striga* species suggests that additional *Striga*::host relationships may develop as different crops are moved into proximity with *Striga*. An example of this process has been the severe parasitism on maize (Rich & Ejeta, 2008). With no history of *Striga* exposure before the 16th century, it is not surprising that natural sources of resistance in maize have not yet become a significant part of the germplasm in African maize production. A similar story may also be true for *S. hermonthica* on pearl millet, which is believed to have been first domesticated in West Africa (Oumar *et al.*, 2008). Exposure of pearl millet to this parasitic weed may be a product of only the last few thousand years, when sorghum agriculture moved south and west from the horn of Africa, thus explaining a similar lack of identified strong resistance to *Striga* in the *P. glaucum* germplasm (Wilson *et al.*, 2000). As expected of a parasite that has not yet co-adapted with its host, the losses of yield due to *S. hermonthica* on pearl millet are routinely much more severe than seen in adjacent sorghum fields, but this may be a partly due to the cultural practices that place sorghum on the better soils that are provided with better farmer inputs.

As the agricultural weedy species continue to expand their ranges, mostly due to anthropogenic activity, the likelihood of two traditionally allopatric species hybridizing becomes a real possibility. Our results also indicate that the shift from a grass host to a legume host is a derived character state for *S. gesnerioides*, rather than a basal character state as previously assumed (Mohamed, 1994). This result should stimulate research in *S. gesnerioides* and may be used to further dissect and identify the loci that control host preference if F1 hybrids and mapping populations can be generated, similar to studies performed with *S. aspera* and *S. hermonthica* (Aigbokhan *et al.*, 2000).

The analysis of nuclear genome size coupled with sample sequence analysis (chapter 3) provided the first examination of nuclear genome structure within this family of parasitic plants (Orobanchaceae). Our results show that the nuclear genomes of parasitic plants are similar to other plant genomes investigated to date, with large amounts of repetitive sequences. These included the class 1 (retroelements), class 2 (DNA transposons), and satellite repeats, all commonly found in non-parasitic plant genomes. Different sources of the parasites, even though from geographic sources that were separated by more than 2000 miles, yielded the same basic genome composition, indicating that this is not a terribly variable trait within a species. All of the most-abundant repeats were most similar to repeats from other dicot species (not from hosts like sorghum) (data not shown), indicating that they were not likely to be the products of recent horizontal transfer events. We also showed that none of these repetitive sequences was solely responsible for changes in genome size within the genus, but acted together to influence genome size. The great reduction in amounts of SS1 (satellite repeat) observed in *S. gesnerioides* suggests that some genome restructuring recently occurred in the single clade containing *S. aspera*, *S. gesnerioides*, and *S. hermonthica*. This may have been the result of a host switch or could be the result of a meiotic drive system active in *S. aspera* and *S. hermonthica*, which was lost in *S. gesnerioides*. This observation may reduce the ability for hybridization experiments between *S. gesnerioides* and other *Striga* species as previously suggested. However, it should be noted that dramatic differences in satellite DNA content are not unusual even within a species, as observed in *Zea mays* germplasm (McClintock, 1978), and these have no demonstrated effect on mating potential.

The analysis of genome size demonstrated a four-fold variation in nuclear DNA content between the species examined, suggesting that polyploidization has occurred within the genus. Further analysis is needed on this point, to confirm possible polyploid states. One simple approach would be to investigate the number of alleles of genes present in each species. For instance, if a presumptive tetraploid species consistently exhibited twice as many different copies of each nuclear gene investigated, then it would suggest true tetraploidy. This approach would be complicated by paralogy and by the high level of heterozygosity in the allogamous species, but investigating a large number of different low-copy-number genes should allow distinction between orthologs, paralogs and homeologs. This issue of polyploidy is important because it may have driven some of the speciation events.

One of the most elusive attributes of *S. hermonthica* has been experimentally documenting host specialization at the genetic level. Many researchers have described morphologically distinct races of *S. hermonthica* on different crops or in different regions (reviewed in Musselman, 1987), but no genetic evidence has been generated that support this observation. Previous investigations into genetic diversity and population structure within *S. hermonthica* have all been severely limited by the choice of marker system, the number of loci analyzed, and/or the number of populations investigated (Bharathalakshmi *et al.*, 1990; Kuiper *et al.*, 1996; Olivier *et al.*, 1998; Koyama, 2000; Gethi *et al.*, 2005). In order to rectify these problems, a set of 12 microsatellite markers were developed and described in chapter 4. These markers are co-dominant, neutrally evolving, reproducible, and are highly variable. Their use allows a fine scale analysis of genetic diversity and population structure. One benefit of these markers is that future analyses of *S. hermonthica* diversity can be directly compared, if these markers are employed. This is of great potential value because the range of *S. hermonthica* is so vast that a single study cannot easily include all populations. Moreover, these markers will allow the rapid evaluation and comparison to previously examined populations of any new morphological variant or new virulent form that arises. The possibility also exists to migrate these markers to other closely related species like *S. gesnerioides* or *S. aspera*.

In chapter 5, we discuss the genetic analysis of diversity and structure in *S. hermonthica* and confirm the high level of genetic diversity previously described (Bharathalakshmi *et al.*, 1990). We also show a small amount of genetic differentiation between Northern and Southern populations and that populations in the Eastern region of Mali are likely older than populations in the Western portion of Mali. These results suggest that *S. hermonthica* is expanding its range from East to West and that host genotype or other geo-ecological variables are acting as selection factors to differentiate Northern and Southern populations. An analysis of gene flow suggests that “isolation by distance” is not a major force shaping population structure in this species as in other *Striga* species (*S. asiatica* and *S. gesnerioides*).

Future work in *Striga:*

The phylogenetic hypothesis of relationships within *Striga* described in chapter 2 was generated from less than ideal starting material. A larger sampling of *Striga* species is needed to further understand relationships within the genus. This would be best accomplished with freshly grown plant material, and should include at least one nuclear locus. Moreover, these data should be analyzed to attempt to determine the relative timing of the speciation events, using the standard molecular clock approach (Gaut, 1998). Given the possibility of rapid genetic change, especially in the plastid genome, for this parasitic genus, these results should be especially interesting. In particular, a comparison of rates of sequence change in the chloroplast DNA to that in the nuclear DNA might be most informative. Collection of silica-dried plant material or seeds will require a significant investment in and commitment to field collection.

Further investigations into the nuclear genome of *Striga* and other Orobanchaceae species are underway. These include an EST sequencing project in *S. hermonthica* (Yoshida *et al.*, 2010) and another EST sequencing project including three Orobanchaceae genera; *Triphysaria*, *Orobanche*, and *Striga* (Timko 2010, pers. comm.). The *Striga* repeat database described in chapter 3 will be a useful resource to begin the annotation of the first sequenced *Striga* genome.

A second field season of *S. hermonthica* population collections in and around Mali from our collaboration with the Van Mourik lab is now complete. In this series of collections, great effort was taken to expand the range to the western limit of the species in Senegal, as well as further collections in the Northern and Southern regions of Mali and farther east into Niger. Populations parasitizing maize, sorghum, pearl millet, and rice were also included to further dissect genetic differentiation possibly associated with host preference. This study, along with a broader collection across Africa and Asia, will provide extensive new insights into the genetic composition of *S. hermonthica*, and its origins.

Recent work has determined that strigolactones (a primary *Striga* germination stimulant) serve as plant hormones to both inhibit vegetative branching and to encourage arbuscular mycorrhizal (AM) association (Akiyama *et al.*, 2005; Gomez-Roldan *et al.*, 2008). This observation suggests how some *Striga* species were "domesticated" into agricultural weeds. One farmer-selected trait that all cereal crops and many other crops share in common is a reduction in branching. This trait is intentionally selected during domestication, making a stronger plant that can maintain a single large seed head for harvest. Farmers could have unknowingly selected plants that exuded large amounts of strigolactones by selecting plants that exhibited a reduced number of branches. This would produce plants that not only have reduced branches, but also are more susceptible to *Striga* infection because they strongly stimulate them to germinate. The increased amount of strigolactones would also attract the symbiotic AM fungi, likely producing an added benefit of reduced phosphate starvation. In Sub-Saharan Africa, where sorghum and pearl millet were domesticated, this might have led to increased *Striga* parasitism as an unavoidable outcome of improved crop yield, especially on poor (e.g, phosphate deficient) soils. This hypothesis would help explain why *Striga* attacks many of our grain crops, regardless of where they were domesticated. This model suggests that *Striga* may not have been a major problem in early sorghum domestication, but instead became a problem after selection for less branches and better performance on poor soils, perhaps in an environment where Striga was not yet abundant.

BIBLIOGRAPHY

1. **Aigbokhan E, Berner D, Musselman L, Mignouna H. 2000.** Evaluation of Variability in *Striga Aspera*, *Striga Hermonthica* and Their Hybrids Using Morphological Characters and Random Amplifed Polymorphic DNA Markers. *Weed Research* **40**: 375-386.

2. **Akiyama K, Matsuzaki H. 2005.** Plant Sesquiterpenes Induce Hyphal Branching in Arbuscular Mycorrhizal Fungi. *Nature* **435**(7043): 824-827.

3. **Ali R, El-Hussein A, Mohamed K, Babiker A. 2009.** Specificity and Genetic Relatedness among *Striga Hermonthica* Strains in Sudan. *Life Science International Journal* **3**(3): 1159-1166.

4. **Arumuganathan K, Tallury SP, Fraser ML, Bruneau AH, Qu R. 1999.** Nuclear DNA Content of Thirteen Turfgrass Species by Flow Cytometry. *Crop science* **39**(5): 1518-1521.

5. **Bar-Nun N, Sachs T, Mayer AM. 2008.** A Role for Iaa in the Infection of Arabidopsis Thaliana by Orobanche Aegyptiaca. *Ann Bot* **101**(2): 261-265.

6. **Barbrook AC, Howe CJ, Purton S. 2006.** Why Are Plastid Genomes Retained in Non-Photosynthetic Organisms? *Trends in Plant Science* **11**(2): 101-108.

7. **Baucom R, Estill J, Chaparro C, Upshaw N, Jogi A, Deragon J, Westerman R, Sanmiguel P, Bennetzen J. 2009.** Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *PLoS Genetics* **5**(11).

8. **Bebawi F, Eplee R, Harris C, Norris R. 1984.** Longevity of Witchweed (*Striga asiatica*) Seed. *Weed Science* **32**: 494-497.

9. **Bennett JR, Mathews S. 2006.** Phylogeny of the Parasitic Plant Family Orobanchaceae Inferred from Phytochrome A. *American journal of Botany* **93**(7): 1039-1051.

10. **Bennetzen JL. 2000.** Transposable Elements Contributions to Plant Gene and Genome Evolution. *Plant Molecular Biology* **42**(1): 251-269.

11. **Bennetzen JL. 2007.** Patterns in Grass Genome Evolution. *Current Opinion in Plant Biology* **10**(2): 176-181.

12. **Benson G. 1999.** Tandem Repeats Finder: A Program to Analyze DNA Sequence. *Nucleic Acids Research* **27**(2): 573-580.

13. **Berner D, Cardwell K, Faturoti B, Ikie F, Williams O. 1994.** Relative Roles of Wind, Crop Seeds, and Cattle in Dispersal of *Striga* Spp. *Plant Disease* **78**: 402-406.

14. **Bharathalakshmi C, Werth R Musselman L. 1990.** A Study of Genetic Diversity among Host-Specific Populations of the Witchweed *Striga Hermonthica* (Del.) Benth. (Scrophulariaceae) in Africa. *Plant Systematics and Evolution* **172**(1-4): 1-12.

15. **Binladen J, Wiuf C, Thomas M, Gilbert P, Bunce M, Barnett R, Larson G, Greenwood A, Haile J, Ho S, Hansen A, Willerslev E. 2006.** Assessing the Fidelity of Ancient DNA Sequences Amplified from Nuclear Genes. *Genetics* **172**(2): 733-741.

16. **Botanga C, Kling J, Berner D, Timko. M. 2002.** Genetic Variability of *Striga asiatica* (L.) Kuntz Based on AFLP Analysis and Host-Parasite Interaction. *Euphytica* **128**: 375-388.

17. **Botanga CJ, Timko MP. 2006.** Phenetic Relationships Among Different Races of *Striga gesnerioides* (Willd.) Vatke from West Africa. *Genome* **49**(11): 1351-1365.

18. **Bouwmeester HJ, Roux C, Lopez-Raez JA, Becard G. 2007.** Rhizosphere Communication of Plants, Parasitic Plants and Am Fungi. *Trends Plant Sci* **12**(5): 224-230.

19. **Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh BS. 1993.** Characterization of the Pufferfish (Fugu) Genome as a Compact Model Vertebrate Genome. *Nature* **366**: 265-268.

20. **DeBarry JD, Liu R, Bennetzen JL. 2008.** Discovery and Assembly of Repeat Family Pseudomolecules from Sparse Genomics Sequence Data Using the Assisted

Automated Assembler of Repeat Families (AAARF) Algorithm. *BMC Bioinformatics* **13**(9): 235.

21. **dePamphilis CW. 1995.** *Genes and Genomes*. pp. 177-205. In Parasitic Plants. Edited by Malcolm C. Press and Jonathan D. Graves London: Chapman & Hall.

22. **dePamphilis CW, Young ND, Wolfe AD. 1997.** Evolution of Plastid Gene Rps2 in a Lineage of Hemiparasitic and Holoparasitic Plants: Many Losses of Photosynthesis and Complex Patterns of Rate Variation. *Proc Natl Acad Sci U S A* **94**(14): 7367-7372.

23. **Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL. B. 2005.** Analysis and Mapping of Randomly Chosen Bacterial Artificial Chromosome Clones from Hexaploid Bread Wheat. *Proceedings of the National Academy of Sciences, USA* **102**: 19243-19248.

24. **Doyle J, Doyle J. 1987.** A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf Tissue. *Phytochemical Bulletin* **19**: 11-15.

25. **Ejeta G. 2007a.** Breeding for Striga Resistance in Sorghum: Exploitation of an Intricate Host–Parasite Biology. *Crop science* **47**: 216-227.

26. **Ejeta G. 2007b.** *The Striga Scourge in Africa: A Growing Pandemic*. In: G. EjetaJ. Gressel eds. *Integrating New Technologies for Striga Control: Towards Ending the Witch-Hunt*, 71-84. Singapore: World Scientific Publishing Co. Pte. Ltd.

27. **Estep M, Van Mourik T, Muth P, Guindo D, Parzies H, Koita O, Weltzien E, Bennetzen J. 2010.** Development of Microsatellite Markers for Characterizing Diversity in a Parasitic Witchweed, *Striga hermonthica* (Orobanchaceae). *Molecular Ecology Notes* (In press).

28. **Evanno G, Regnaut S, Goudet J. 2005.** Detecting the Number of Clusters of Individuals Using the Software Structure: A Simulation Study. *Molecular Ecology* **14**(8): 2611-2620.

29. **Ewing B, Hillier L, Wendl MC, Green P. 1998.** Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research* **8**(3): 175-185.

30. **Finn RD, Tate J, Mistry J, Coggill PC, Sammut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. 2008.** The Pfam Protein Families Database:. *Nucleic Acids Research* **36**(Database Issue): D281-288.

31. **Galtier N, Gouy M, Gautier C. 1996.** Seaview and Phylo_Win, Two Graphic Tools for Sequence Alignment and Molecular Phylogeny. *Computational Applications in Biological Sciences* **12**: 543-548.

32. **Garris H, Wells J. 1956.** Parasitic Herbaceous Annual Associated with Corn Disease in North Carolina. *Plant Disease Reporter* **40**: 837-839.

33. **Gaut B. 1998.** Molecular Clocks and Nucleotide Substitution Rates in Higher Plants. *Evolutionary biology* **30**: 93-120.

34. **Gethi G, Smith M, Mitchell SKresovich S. 2005.** Genetic Diversity of *Striga hermonthica* and *Striga asiatica* Populations in Kenya. *Weed Research* **45**(1): 64-73.

35. **Gomez-Roldan V, Fermas S, Brewer PB, Puech-Pagès V, Dun EA, Pillot JP, Letisse F, Matusova R, Danoun S, Portais JC, Bouwmeester H, Bécard G, Beveridge CA, Rameau CSF. R. 2008.** Strigolactones Inhibition of Shoot Branching. *Nature* **11**(455(7210)): 189-194.

36. **Gurtubay JIG, Goni FM, Gomez-Fernandez JC, Otamendi JJ, Macarulla JM. 1980.** Triton X-100 Solubilization of Mitochondrial Inner and Outer Membranes. *Journal of Bioenergetics and biomembranes* **12**(1-2): 47-70.

37. **Hamrick J. 1982.** Plant Population Genetics and Evolution. *American journal of Botany* **69**(10): 1685-1693.

38. **Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006.** Differential Lineage-Specific Amplification of Transposable Elements Is Responsible for Genome Size Variation in *Gossypium*. *Genome Research* **16**(10): 1252-1261.

39. **Ilic K, SanMiguel PJ, Bennetzen JL. 2003.** A Complex History of Rearrangement in an Orthologous Region of the Maize, Sorghum, and Rice Genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(21): 12265-12270.

40. **Iwo G, Husaini S, Olaniyan G. 1993.** Cytological Observations and Distribution of *Striga* Species in Central Part of Nigeria. *Feddes Repertorium* **104**(7-8): 497-501.

41. **Kazazian HH. 2004.** Mobile Elements: Drivers of Genome Evolution. *Science* **303**(5664): 1626-1632.

42. **Keyes WJ, Taylor JV, Apkarian RP, Lynn DG. 2001.** Dancing Together. Social Controls in Parasitic Plant Development. *Plant Physiol* **127**(4): 1508-1512.

43. **Kigawa R, Nochide H, Kimura H, Miura S. 2003.** Effects of Various Fumigants, Thermal Methods and Corbon Dioxide Treatment on DNA Extraction and Amplification: A Case Study on Freeze-Dried Mushroom and Freeze-Dried Muscle Specimens *Collection Forum* **18**(1-2): 74-89.

44. **Koyama M 2000**. Genetic Variability of *Striga hermonthica* and Effects of Resistant Cultivars on Striga Population Dynamics.In. *Breeding for Striga resistance in cereals*. Margraf Verlag, Weikersheim, Germany. 247-260.

45. **Krause D, Weber HC. 1990.** SEM Observations on Seeds of *Striga* Spp and *Buchnera americana* (Scrophulariaceae). *Plant Systematics and Evolution* **170**(3-4): 257-263.

46. **Kuijt J. 1969.** *The Biology of Parasitic Flowering Plants*: University of California.

47. **Kuiper E, Koevoets P, Verkleij J 1996**. Genetic Variability of *Striga aspera* and *Striga hermonthica*. Are They Really Two Distinct Species?In. *Sixth international Parasitic Weeds Symposium*. Cordoba, Spain. 123-134.

48. **Lai JS, Ma JX, Swigonova Z, Ramakrishna W, Linton E, Llaca V, Tanyolac B, Park YJ, Jeong Y, Bennetzen JL, Messing J. 2004.** Gene Loss and Movement in the Maize Genome. *Genome Research* **14**(10A): 1924-1931.

49. **Liu K, Muse S. 2005.** Powermarker: An Integrated Analysis Environment for Genetic Marker Analysis. *Bioinformatics* **21**(9): 2128-2129.

50. **Ma JX, SanMiguel P, Lai JS, Messing J, Bennetzen JL. 2005.** DNA Rearrangement in Orthologous Orp Regions of the Maize, Rice and Sorghum Genomes. *Genetics* **170**(3): 1209-1220.

51. **Macas J, Neumann P, Navratilova A. 2007.** Repetitve DNA in the Pea (*Pisum sativum* L.) Genome: Comprehensive Characterization Using 454 Sequencing and Comparison to Soybean and Medicago Truncatula. *BMC Genomics* **8**(427).

52. **Manen JF, Habashi C, Jeanmonod D, Park JM, Schneeweiss GM. 2004.** Phylogeny and Intraspecific Variability of Holoparasitic *Orobanche* (Orobanchaceae) Inferred from Plastid Rbcl Sequences. *Mol Phylogenet Evol* **33**(2): 482-500.

53. **Mardis ER. 2008.** Next-Generation DNA Sequenceing Methods. *Annual Review of Genomics and Human Genetics* **9**: 387-402.

54. **Matusova R, Rani K, Verstappen FW, Franssen MC, Beale MH, Bouwmeester HJ. 2005.** The Strigolactone Germination Stimulants of the Plant-Parasitic *Striga* and *Orobanche* Spp. Are Derived from the Carotenoid Pathway. *Plant Physiol* **139**(2): 920-934.

55. **Mayr E 1954.** Change of Genetic Environment and Evolution. In: J. Huxley ed. *Evolution as a Process*. London: Allen and Unwin.

56. **McClintock B 1978.** Significance of Chromosome Constitutions in Tracing the Origin and Migration of Races of Maize in the Americas. In: D. Walden ed. *Maize Breeding and Genetics*. New York: John Wiley & Sons, 159-184.

57. **Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei FS, Fuks G, Soderlund CA, Mayer KFX, Wing RA. 2004.** Sequence Composition and Genome Organization of Maize. *Proceedings of the National Academy of Sciences of the United States of America* **101**(40): 14349-14354.

58. **Meyers BC, Tingley SV, Morgante M. 2001.** Abundance, Distribution, and Transcriptional Activity of Repetitive Elements in the Maize Genome. *Genome Research* **11**(10): 1660-1676.

59. **Miller M, Holder M, Vos R, Midford P, Liebowitz T, Chan L, Hoover P, Warnow T 2010**. The Cipres Portals.In.

60. **Mohamed K. 1994.** *Biosystematics and Diversification in the Genus Striga Lour. (Scrophulariaceae) in Africa.* Dissertation, Old Dominion University Norfolk, VA.

61. **Mohamed K, Bolin J, Musselman L, Peterson A 2007.** Genetic Diversity of *Striga* and Implications for Control and Modeling Future Distributions. In: G. EjetaJ. Gressel eds. *Integrating New Technologies for Striga Control: Towards Ending the Witch-Hunt*, 71-84.

62. **Mohamed K, Musselman L 2008**. Taxonomy of Agronomically Important *Striga* and *Orobanche* Species.In R. Labrada. *Progress on Farmer Training in Parasitic Weed Management. Plant Production & Protection Division, Food and Agricultural Organization of the United Nations (FAO)*. Rome. 7-14.

63. **Mohamed K, Musselman L, Aigbokhan E, Berner D 1996**. Evolution and Taxonomy of Agronomically Important *Striga* Species.In M. MorenoJ. Cubero. *Advances in parasitic plant research*. 54-73.

64. **Mohamed KI, Musselman LJ, Riches CR. 2001.** The Genus *Striga* (Scrophulariaceae) in Africa. *Annals of the Missouri Botanical Garden* **88**(1): 60-103.

65. **Musselman L, Bharathalaksmi, Safa S, Knepper D, Mohamed K, White C 1991**. Recent Research on the Biology of *Striga asiatica, S. gesnerioides and S. hermonthica*.In S. Kim. *Combating Striga in Africa. Proceedings of the international workshop Organ IITA, ICRISAT IDRC*. Ibadan, Nigeria: International Institute of Tropical Agriculture. 31-41.

66. **Musselman L, Hepper F. 1986.** The Witchweeds (*Striga*, Scrophulariaceae) of the Sudan Republic. *Kew Bulletin* **41**(1): 205-221.

67. **Musselman LJ. 1987.** *Parasitic Weeds in Agriculture - Striga*. Baco Raton: CRC Press, Inc.

68. **Nickrent DL 2007**. Parasitic Plant Connection.In. http://www.parasiticplants.siu.edu/

69. **Olivier A, Glaszmann J, Lanaud C, Leroux G. 1998.** Population Structure, Genetic Diversity and Host Specificity of the Parasitic Weed *Striga hermonthica* (Scrophulariaceae) in Sahel. *Plant Systematics and Evolution* **209**: 33-45.

70. **Olmstead RG, dePamphilis CW, Wolfe AD, Young ND, Elisons WJ, Reeves PA. 2001.** Disintegration of the Scrophulariaceae. *Am J Bot* **88**(2): 348-361.

71. **Olmstead RG, Palmer JD. 1994.** Chloroplast DNA Systematics - a Review of Methods and Data-Analysis. *American journal of Botany* **81**(9): 1205-1224.

72. **Oosterhout C, Hutchinson W, Wills D, Shipley P. 2004.** Micro-Checker: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data. *Molecular Ecology Notes* **4**: 535-538.

73. **Oswald A, Ransom J. 2001.** *Striga* Control and Improved Farm Productivity Using Crop Rotation. *Crop protection* **20**(2): 113-120.

74. **Oumar I, Mariac C, Pham J, Vigouroux Y. 2008.** Phylogeny and Orign of Pearl Millet (*Pennisetum glaucum* [L.] R. Br) as Revealed by Microsatellite Loci. *Theoretical and Applied Genetics* **117**: 489-497.

75. **Ouyang S, Buell CR. 2004.** The Tigr Plant Repeat Databases: A Collective Resource for the Identification of Repetitive Sequences in Plants. *Nucleic Acids Research* **32**(Database issue): D360-363.

76. **Parker C, Reid D 1979**. Host Specificity In *Striga* species — Some Preliminary Observations.In L. Musselman, A. WorshamR. Epplee. *Proceedings of the second symposium on parasitic weeds*. Raleigh, N.C.: North Carolina State University. 79-90.

77. **Paszkowski U. 2006a.** A Journey through Signaling in Arbuscular Mycorrhizal Symbioses 2006. *New Phytol* **172**(1): 35-46.

78. **Paszkowski U. 2006b.** Mutualism and Parasitism: The Yin and Yang of Plant Symbioses. *Curr Opin Plant Biol* **9**(4): 364-370.

79. **Peacock WJ, Dennis ES, Rhoades MM, Pryor AJ. 1981.** Highly Repeated DNA Sequence Limited to Knob Heterochromatin in Maize. *Proceedings from the National Acedemy of Sciences* **78**: 4490-4494.

80. **Peakall R, Smouse PE. 2006.** Genalex 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. *Molecular Ecology Notes*(6): 288-295.

81. **Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. 2006.** Doubling Genome Size without

Polyploidization: Dynamics of Retrotransposition-Driven Genomic Expansions in *Oryza australiensis*, a Wild Relative of Rice. *Genome Research* **16**(10): 1262-1269.

82. **Plohl M, Luchetti A, Mestrovic N, Mantovani B. 2008.** Satellite DNAs between Selfishness and Functionality: Structure, Genomics, and Evolution of Tandem Repeats in Centromere (Hetero)Chromatin. *Gene* **409**: 72-82.

83. **Pritchard JK, Stephens M, Donnelly P. 2000.** Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**(2): 945-959.

84. **Rhoades M, McClintock B. 1935.** The Cytogenetics of Maize. *Botanical reviews* **1**: 292-325.

85. **Rhoades MM, Dempsey E. 1966.** The Effect of Abnormal 10 on Preferential Segregation and Crossing over in Maize. *Genetics* **53**: 989-1020.

86. **Rich P, Ejeta G. 2008.** Towards Effective Resistance to Striga in African Maize. *Plant Signal Behavior* **3**(9): 618-621.

87. **Rousset F. 2008.** Genepop'007: A Complete Reimplementation of the Genepop Software for Windows and Linux. *Mol. Ecol. Resources* **8**: 103-106.

88. **Rozen S, Skaletsky H, eds. 2000.** *Primer3 on the Www for General Users and for Biologist Programmers.* Bioinformatics Methods and Protocols: Methods in Molecular Biology. Totowa, NJ: Humana Press.

89. **Safa S, Jones B, Musselman L. 1984.** Mechanisms Favoring Outbreeding in *Striga hermonthica* (Scrophulariaceae). *New Phytologist* **96**: 299-305.

90. **SanMiguel P, Bennetzen JL. 1998.** Evidence That a Recent Increase in Maize Genome Size Was Caused by the Massive Amplification of Intergene Retrotransposons. *Annals of Botany* **82**: 37-44.

91. **SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, MelakeBerhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL. 1996.** Nested Retrotransposons in the Intergenic Regions of the Maize Genome. *Science* **274**(5288): 765-768.

92. **Schneeweiss GM, Palomeque T, Colwell AE, Weiss-Schneeweiss H. 2004.** Chromosome Numbers and Karyotype Evolution in Holoparasitic *Orobanche* (Orobanchaceae) and Related Genera. *American journal of Botany* **91**(3): 439-448.

93. **Schuelke M. 2000.** An Economic Method for the Fluorescent Labeling of PCR Fragments. *Nature Biotechnology* **18**: 233-234.

94. **Shaw J, Lickey EB, Beck JT, Farmer SB, Liu WS, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL. 2005.** The Tortoise and the Hare II: Relative Utility of 21 Noncoding Chloroplast DNA Sequences for Phylogenetic Analysis. *American journal of Botany* **92**(1): 142-166.

95. **Shawe K, Ingrouille M. 1993.** Isozyme Analysis Demonstrates Host Selection of Parasitic Pathotypes in the Association between Cowpea and *S. gesnerioides*. *Brighton Crop Protection Conference-weeds 2*: 919-924.

96. **Slatkin M. 1995.** A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. *Genetics* **139**: 457.

97. **Spannagl M, Noubibou O, Haase D, Yang L, Gundlach H, Hindemitt T, Klee K, Haberer G, Schoof H, Mayer K. 2007.** Mipsplantsdb—Plant Database Resource for Integrative and Comparative Plant Genome Research. *Nucleic Acids Research* **35**(Database Isuue): D834 - D840.

98. **Stamatakis A. 2006.** Raxml-Vi-Hpc: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models *Bioinformatics* **22**(21): 2688-2690.

99. **Swafford D 2002**. Paup 4.0: Phylogenetic Analysis Using Parsimony and Other Methods. .In. Sunderland, MA.: Sinauer Associates.

100. **Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. 2001.** Computational and Experimental Analysis of Microsatellites in Rice (Oryza Sativa L.): Frequency, Length Variation, Transposon Associations, and Genetic Marker Potential. *Genome Research* **11**: 1441-1452.

101. **Temsch EM, Greilhuber J. 2000.** Genome Size Variation in *Arachis hypogaea* and *A. monticola* Re-Evaluated. *Genome* **43**(3): 449-451.

102. **Timko M, Gowda B, Ouedraogo J, Ousmane B. 2007.** *Molecular Markers for Analysis of Resistance to Striga gesnerioides in Cowpea.* . In: G. EjetaJ. Gressel eds. *Integrating New Technologies for Striga Control: Towards Ending the Witch-Hunt*, 71-84. Singapore: World Scientific Publishing Co. Pte. Ltd.

103. **Van Mourik T. 2007.** *Striga Hermonthica Seed Bank Dynamics: Process Quantification and Modeling.* Dissertation Wageningen University.

104. **Werth C, Riopel J, Gillespie N. 1984.** Genetic Uniformaity in an Introduced Population of Witchweed (*Striga asiatica*) in the United States. *Weed Science* **32**(5): 645-648.

105. **Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, SanMiguel P, Lakey N, Bedell J, Yuan Y, Budiman MA, Resnick A, Van Aken S, Utterback T, Riedmuller S, Williams M, Feldblyum T, Schubert K, Beachy R, Fraser CM, Quackenbush J. 2003.** Enrichment of Gene-Coding Sequences in Maize by Genome Filtration. *Science* **302**(5653): 2118-2120.

106. **Wiens J. 2006.** Missing Data and the Design of Phylogenetic Analysis. *Journal of biomedical informatics* **39**: 34-42.

107. **Wilhelm M, Wilhelm FX, Keith G, Agoutin B, Heyman T. 1994.** Yeast Ty1 Retrotransposon: The Minus-Strand Primer Binding Site and a Cis-Acting Domain of the Ty1 Rna Are Both Important for Packaging of Primer Trna inside Virus-Like Particles *Nucleic Acids Research* **22**: 4560-4565.

108. **Wilson J, Hess D, Hanna W. 2000.** Resistance to *Striga hermonthica* in Wild Accessions of the Primary Gene Pool of *Pennisetum glaucum*. *Phytopathology* **90**: 1169-1172.

109. **Wolfe AD. 2005.** ISSR Techniques for Evolutionary Biology. *Methods Enzymol* **395**: 134-144.

110. **Wolfe AD, dePamphilis CW. 1998.** The Effect of Relaxed Functional Constraints on the Photosynthetic Gene Rbcl in Photosynthetic and Nonphotosynthetic Parasitic Plants. *Molecular Biology and Evolution* **15**(10): 1243-1258.

111. **Wolfe KH, Morden CW, Palmer JD. 1992.** Function and Evolution of a Minimal Plastid Genome from a Nonphotosynthetic Parasitic Plant. *Proceedings of the National Academy of Sciences of the United States of America* **89**(22): 10648-10652.

112. **Xiong Y, Eickbush T. 1990.** Origin and Evolution of Retroelements Based Upon Their Reverse Transcriptase Sequences. *European Molecular Biology Organization* **9**(10): 3353-3362.

113. **Yoshida S, Ishida J, Kamal N, Abdelbagi A, Namba S, Shirasu K. 2010.** A Full-Length Enriched CDNA Library and Expressed Sequence Tag Analysis of the Parasitic Weed, *Striga hermonthica*. *BMC Plant Biology* **10**.

114. **Yoshida S, Shirasu K. 2009.** Multiple Layers of Incompatibility to the Parasitic Witchweed, *Striga hermonthica*. *New Phytologist* **183**(1): 180-189.

115. **Young ND, dePamphilis CW. 2000.** Purifying Selection Detected in the Plastid Gene Matk and Flanking Ribozyme Regions within a Group II Intron of Nonphotosynthetic Plants. *Mol Biol Evol* **17**(12): 1933-1941.

116. **Young ND, dePamphilis CW. 2005.** Rate Variation in Parasitic Plants: Correlated and Uncorrelated Patterns among Plastid Genes of Different Function. *BMC Evol Biol* **5**(1): 16.

117. **Young ND, Steiner KE, dePamphilis CW. 1999.** The Evolution of Parasitism in Scrophulariaceae/Orobanchaceae: Plastid Gene Sequences Refute an Evolutionary Transition Series. *Annals of the Missouri Botanical Garden* **86**(4): 876-893.