UNEQUAL RECOMBINATION AND OTHER REARRANGEMENTS IN PLANT NUCLEAR AND CHLOROPLAST GENOMES

by

DANIEL CHRISTOPHER FRAILEY

(Under the Direction of Jeffrey L. Bennetzen)

ABSTRACT

Plant genomes vary tremendously in terms of size and chromosome structure. One factor that plays a major role in genome evolution is homologous recombination. Both meiotic and mitotic recombination increase genetic diversity by rearranging combinations of alleles, creating new alleles and altering the copy numbers of genes by unequal recombination. Previous studies identified unequal recombination events between two disease resistance gene homologues from sorghum, *Pc* A and *Pc* C. Interestingly, most of the recombination events occurred in the least conserved part of the gene, which was the domain responsible for pathogen recognition. Changes here have the potential to allow evolution of new resistance specificities. We created transgenic maize lines containing tandem *Pc* A and *Pc* C genes, and then used a PCR assay to screen maize pollen for unequal recombination. We identified 23 unequal recombination events, yielding a rate of 1 per ~7700 pollen grains. DNA sequence analysis indicated that, unlike in sorghum, all unequal recombined *Pc* products had crossovers occurring in the conserved regions of the genes.

Transposons are responsible for most plant genome variation. For instance, LTR retrotransposon amplification can rapidly increase a species' genome size. This increase can be partially counteracted by unequal recombination between LTRs, deleting the internal sequence

and leaving behind a solo LTR. We used the same techniques as for the *Pc* locus to identify a total of 23 unequal recombination events from 10 LTR retrotransposons. We found a positive correlation between recombination and LTR sequence identity and a negative correlation with DNA methylation.

We also assembled and analyzed the chloroplast genomes from five species of parasitic plants. In most angiosperms, the chloroplast genome, or plastome, is highly conserved. We found several rearrangements and gene deletions in four of the five species. We also found large increases in plastome size. Plastomes contain two virtually identical inverted repeats separating a large and small single copy region. The plastome size increase was due to expansion of these repeats into the single copy regions.

INDEX WORDS: unequal recombination, NBS-LRR genes, LTR retrotransposons, parasitic plants, plastomes

UNEQUAL RECOMBINATION AND OTHER REARRANGEMENTS IN PLANT NUCLEAR AND CHLOROPLAST GENOMES

by

DANIEL CHRISTOPHER FRAILEY

BA, University of Delaware, 2012

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

2017

UNEQUAL RECOMBINATION AND OTHER REARRANGEMENTS IN PLANT NUCLEAR AND CHLOROPLAST GENOMES

by

DANIEL CHRISTOPHER FRAILEY

Major Professor:	Jeffrey L. Bennetzen
Committee:	Kelly Dawe
	Michael McEachern
	Robert J. Schmitz
	Shavannor Smith

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2017

DEDICATION

This dissertation is dedicated to my parents in thanks of all their support.

ACKNOWLEDGEMENTS

I want to thank my parents and my family for all their support and encouragement throughout my time as graduate student. I also want to thank my friends here in Athens for making my time here enjoyable and fun. I want to thank my advisor, Jeffrey Bennetzen, for all his help and guidance with research, writing papers, and giving presentations, without which I would have never been able to complete my dissertation, and also for providing me with many opportunities to attend conferences. I want to thank my committee members who all provided me with invaluable advice on my dissertation project. I would like to thank the other graduate students and professors in the Genetics and Plant Biology departments who were always willing to give help. I also thank all the members of my lab both for their making the lab a great place to work as well as their advice on research, and especially Srinivasa Chaluvadi for his patience and guidance, who was always willing to help and without whom I would never have finished.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Plant genomes tend to have higher rates of evolution rates than other eukaryotes and vary greatly in terms of genome size and chromosome structure (Kejnovsky et al. 2009, Murat et al. 2012). This project looked at two factors that play a major role in genome evolution and variation: recombination and transposable elements. Meiotic recombination increases genetic diversity by rearranging allele combinations on a chromatid, as well as creating new recombinant alleles. Unequal recombination causes various types of structural variations, including duplications, deletions, inversions, and large chromosome rearrangements (Gaut et al. 2007). Transposable elements can also cause structural changes when they break chromosomes or serve as homologous sites for unequal recombination. Particularly important are the affects transposon insertions can have on the expression of nearby genes (Naito et al. 2009, Hollister et al. 2011, Hirsch and Springer 2017). Additionally, they are the main factor responsible for the massive variation of genome size seen in plants (Bennetzen 2000, Bennetzen 2002, Kidwell 2002). The third part of this project was the assembly and analysis of the chloroplast genomes from five species of parasitic plants. Chloroplast genomes are easy to study due to their small size, and the chloroplast genomes of parasitic plants provide an excellent study for how rapidly chromosome rearrangements can cause genomes to diverge.

Homologous recombination

Homologous recombination (HR) has two main functions: to repair broken DNA during somatic growth (leading to "mitotic recombination") and to increase genetic variation during meiosis. There are several key differences between mitotic and meiotic recombination. Mitotic recombination is one of several pathways that repairs double-strand breaks (DSBs). Meiotic recombination is under more developmental control, and begins with the creation of programmed DSBs at pachytene in prophase I of meiosis. In the process, linkage between adjacent alleles can be broken to create new combinations and increase genetic diversity in the meiotic products. Specific sites where meiotic recombination occurs are not precisely defined, but it happens at much higher frequencies in certain regions known as hotspots. Meiotic recombination is required for proper segregation during meiosis and occurs at a much higher rate per cell than mitotic recombination (Li et al. 2007, Osman et al. 2011, Krejci et al. 2012). Another difference is that meiotic recombination preferentially occurs between two homologous chromosomes, while mitotic recombination mostly occurs between two sister chromatids. In maize, 94% of meiotic recombination events at the *a1* locus occurred between homologous chromatids (Yandeau-Nelson et al. 2006).

Early in prophase I, Spo11 (sporulation specific protein 11) generates DSBs. Spo11 shares homology with the TOP6A type II topoisomerase from the archaeon *Sulpholobus shibatae* (Bergerat et al. 1997, Keeney et al. 1997). Arabidopsis contains three Spo11 paralogs. Two of them, Spo11-1 and Spo11-2, are required for meiosis (Hartung & Puchta 2000, 2001; Grelon et al., 2001; Stacey et al., 2006). They both contain catalytically active tyrosine residues that are necessary for DSB formation, and the proteins probably function as a heterodimer (Hartung et al. 2007). Spo11 has not been studied functionally in maize, but homologs of all three *Spo11* genes

have been identified (Sidhu et al. 2017). After DSB formation, Spo11 remains covalently attached to the 5' ends of the DNA on either side of the break. It is removed as the DNA ends are resected to generate 3' overhangs. Rad51 and Dmc1 interact with the ssDNA to form nucleoprotein filaments. The filament on one side of the break invades the homologous duplex DNA of one of the non-sister chromatids to form a single-end invasion intermediate. The other end of the displaced DNA strand forms a D-loop. At this point, recombination can be resolved in one of two ways. The D-loop can be resolved via synthesis-dependent strand annealing to form a non-crossover (NCO). Alternatively, the displaced strand can extend as the invading strand polymerizes resulting in second end-capture of the 3' end on the other side of the DSB. The broken DNA strands are ligated to form a double Holliday junction (dHJ) that can create a crossover (CO) when resolved (Osman et al. 2011).

COs are the reciprocal exchange of genetic material between the recombining strands, while NCOs lack this reciprocal exchange at distal sites. In most studied plants, only a small subset of the DSBs are ultimately resolved as COs. In maize, ~500 DSBs are formed and only 20 are resolved as COs (Pawlowski *et al*. 2003, Anderson *et al*. 2003). In Arabidopsis, out of 100-250 DSBs per meiosis around 8-12 are resolved as COs (Sanchez-Moran et al. 2007, Varas et al. 2015). At least one CO per chromosome arm is required for proper segregation during meiosis. This is known as the obligate CO (Jones, 1984; Jones & Franklin, 2006; Shinohara et al., 2008). There are two classes of COs that are formed via different pathways. Class I crossovers comprise the majority of crossovers, and are also responsible for establishing the obligate CO. Class I COs display CO interference, in which the presence of one CO prevents other COs from forming nearby (Jones and Franklin 2006). Class II COs make up around 15% of the COs (de los Santos et al., 2003; Borner et al., 2004, Higgins et al., 2004). These COs do not display interference and

instead have a more random distribution (Copenhaver *et al*. 2002, Higgins et al. 2004). Class I and class II COs are formed through two different pathways requiring different sets of proteins (Borner et al., 2004). In yeast, class I COs depend on the ZMM proteins (Zip1 [Zipper 1], Zip2, Zip3, Zip4, Msh4 [MutS homolog 4], Msh5, and Mer3 [meiotic recombination 3]). Orthologs for *Msh4*, *Msh5*, *Mer3*, and several *Zip* genes have been identified in Arabidopsis (Higgins et al., 2004, 2005 and 2008, Chen et al. 2005, Mercier et al. 2005, Chelysheva et al., 2007). Class II COs in budding yeast depend on the Mus81 (the methyl methanesulfonate-and UV-sensitive 81 protein)-Mms4 (methyl methanesulfonate-sensitivity 4 protein) heterodimer (de los Santos et al., 2003), while in Arabidopsis they depend on the Eme1 (essential meiotic endonuclease 1)-Mus81 heterodimer (Geuting et al., 2009). In maize, homologs for the *Msh4* and *Mus81* genes have been identified (Sidhu et al. 2017).

Recombination hotspots are regions of up to a few kb (usually 1-2 kb) where recombination occurs at a higher rate than the genome average. In yeast, specific sequence motifs have been associated with hotspots (Wahls and Davidson 2012). In humans, mice, and some other mammals, a PR domain zinc finger protein, PRDM9, binds to specific sequence motifs and determines where recombination occurs (Berg et al. 2010, Baudat et al. 2013). Plants have no discovered sequence motifs, and no known functional or structural homolog of PRDM9. However, there are several features that that are correlated with recombination frequency (Melamed-Bessudo et al. 2016).

Chromosomal position seems to have a large effect on recombination. In all studied plants, recombination occurs at higher rates in gene dense regions. In wheat (Choulet *et al*. 2014), barley (The International Barley Genome Sequencing Consortium 2012), sorghum (Paterson et al. 2009), and rice (Si et al. 2015), recombination increases with distance from the

centromere. Other species such as maize (McMullen et al. 2009, Rodgers-Melnick et al. 2015). and tomato (Stephan and Langley 1998) show a similar increase, followed by a decrease again next to the telomeres. Arabidopsis, on the other hand, shows no such biased distribution, but this may be partly a technical outcome of incomplete assembly of centromeric regions in this species (Drouaud et al. 2006). In all plants studied, most genes are found far from the centromeres. However, gene density alone cannot account for the higher recombination rates seen at the ends of chromosomes. The distribution of maize genes accounts for only about 50% of the variation in recombination rates (Anderson et al. 2006). In wheat, if the terminal region of a chromosome is deleted, recombination rates increase in the new terminal segment (Jones *et al*. 2002, Qi et al. 2002). Hence, gene density and chromosomal position both seem to influence recombination rates at the ends of chromosomes.

Because most studies of the distribution of recombination events in plant genomes involve analysis only of CO outcomes, it is not clear whether NCO outcomes have the same bias. In rice centromeres, for instance, a high ratio of solo LTRs to intact LTR retrotransposons has been argued to be an outcome of a very high rate of unequal recombination in this region (Ma and Bennetzen, 2006), even though centromeres are traditionally viewed as an extreme cold spot for meiotic CO events.

Sequence homology is another factor that influences recombination. Regions with higher sequence identity between the recombining strands have much higher recombination rates than those with lower sequence identity. In Arabidopsis, a single nucleotide mismatch can reduce recombination 3-fold. Additional mismatches have a lesser effect, and recombination rates tend to level off at around 20%. Location of the mismatches within the recombining template does not seem to affect recombination rates (Opperman et al. 2004). Length of the homologous templates

also influences recombination rates, with longer templates having higher recombination rates, often in a linear relationship. In yeast, as few as 30 bp have been observed to recombine with each other (Haber 2000). In plants, a few hundred bp are capable of recombination (Puchta and Hohn 1991), but the minimum length required is not yet known.

DNA methylation also seems to have an effect on recombination. In the fungus *Ascobolus immersus*, DNA methylation across a 7.5 kb hotspot reduced recombination 50-fold when one of the recombining homologs was methylated, and several hundredfold when both homologs were methylated (Maloisel and Rossignol 1998). In Arabidopsis, hotspots tend to have low amounts of DNA methylation (Choi et al. 2013; Rodgers-Melnick et al. 2015), and DNA methylation is capable of suppressing recombination at euchromatic hotspots (Yelina et al. 2015). Arabidopsis mutant lines *MET1* (methyltransferase 1) and *DDM1* (deficient in DNA methylation 1), both containing mutations in genes important to the maintenance of DNA methylation, showed an increase in meiotic recombination in the chromosome arms (Mirouze et al. 2012; Melamed-Bessudo and Levy 2012).

Unequal recombination occurs between two homologous sequences that are non-allelic repeats. It creates additional variation by causing structural changes. Unequal recombination between two tandem gene duplicates can cause either a gene duplication or deletion, resulting in copy number variation within plant populations. About 65% of annotated plant genes have a duplicate copy (Panchy et al. 2016). Although many of these originated from polyploidy events, otherwise known as whole genome duplication, polyploidy does not explain all duplications. In Arabidopsis, the number of duplicated genes originating from local duplications is comparable to the number arising from polyploidy events. Intrachromatid recombination between repeats can result in deletion or inversion of the DNA between the repeats. Unequal recombination between

different chromosomes can result in large chromosomal rearrangements like translocations (Gaut et al. 2007). Polyploidy events and recombination together can further increase genome rearrangements. Evidence suggests that recombination between homeologous chromosomes is suppressed in polyploids. The Pairing homologous1 (*Ph1*) locus in wheat and the *PrBn* locus in *Brassica* were both found to suppress homeologous recombination (Feldman 1993; Grandont et al. 2013). However, homeologous recombination does still occur occasionally (Chalhoub et al. 2014). Recent autopolyploids and allopolyploids have been seen to have increased rates of recombination in several studied species. Autotetraploids and allotetraploids in Arabidopsis have an increased recombination rate by 25-50% (Pecinka et al. 2011), allotetraploids in *Gossypium* have an increase by more than 50% (Brubaker et al. 1999), and allotriploids and allotetraploids in *Brassica* have an increase by 1.4-1.7 fold (Leflon et al. 2010).

Structural and functional evolution of NBS-LLR resistance genes in plants

Plants are exposed to a diverse array of constantly evolving pathogens, and require an adaptable defense system capable of responding to a wide range of threats. Plants have the additional challenge of lacking a powerful, multidirectional circulatory system to transport defense cells throughout the organism. Each individual plant cell needs to be able to recognize and respond to pathogenic threats. Plants have evolved a pathogen response system that fulfills both of these requirements and consists of up to hundreds of disease resistance (R) genes in a single plant (Jones and Dangl 2006). An R gene recognizes a pathogen by the presence of the pathogen's *avr* gene product. *Avr* genes are defined as any gene that produces a protein or secondary product that is recognized by a plant R gene. Often, *avr* genes encode "effectors" that aid the pathogen in infection (Flor 1971). Once a pathogen is recognized, the R gene induces

several defense mechanisms, including the hypersensitive response that ultimately results in programmed cell death. This effectively starves the pathogen by depriving it of nutrients (Belkhadir et al. 2004).

The most abundant R genes are the NBS-LRR genes. These genes encode proteins consisting of a nucleotide binding site (NBS) domain and a leucine rich repeat (LRR) region, in addition to some variable N-terminal and C-terminal domains (Dangl and Jones 2001, McHale et al. 2006). The NBS domain is involved in pathway signaling and induces the hypersensitive response after pathogen recognition. NBS domains tend to be conserved, even among different species, and contain several highly conserved motifs, including the P-loop, kinase-2, and GLPL motifs (Tan and Wu 2012, Staskawicz et al. 1995). The LRR region consists of multiple LRR motif repeats and is responsible for pathogen recognition (Ellis et al. 1999). In some cases, LRR repeats recognize pathogens by binding directly to the pathogen's effector (Jia et al. 2000, Deslandes et al. 2003). Alternatively, they can recognize pathogens indirectly through what is known as the guard and decoy hypothesis. In the guard model, the LRR region interacts with a target of the effector and recognizes alterations in the target made by the pathogen. In the decoy model, the LRR region interacts with a structural mimic of the pathogen's target. The effector will attack the mimic instead of its intended target and trigger recognition by the NBS-LRR gene (Dodds and Rathjen 2010). The LRR region of an NBS-LRR gene is generally specific to a single or limited number of pathogen strains. LRR regions are extremely diverse, in both nucleotide sequence and LRR repeat number, to provide the plant with a high potential number of disease resistance specificities (McHale et al. 2006).

NBS-LRR proteins activate multiple responses following pathogen recognition, including the hypersensitive response that results in programmed cell death. This is an effective

mechanism to deal with biotrophic pathogens that require healthy plant tissue for nutrient acquisition (Morel and Dangl 1997). Some pathogens have found a way to use the hypersensitive response against the plant. These necrotrophic pathogens are unable to invade healthy plant tissue. They release toxins that can be recognized by an R gene, which in turn invokes the hypersensitive response. This aids the pathogen by killing the plant cells, resulting in necrotic tissues that can be invaded by the pathogen. Many pathogens that are unable to penetrate living plant cell defenses themselves use this strategy (Lorang et al. 2007, Nagy et al. 2007; Faris et al. 2010).

NBS-LRR genes are located within the genome as either single genes or within clusters of NBS-LRR genes (Hulbert et al. 2001). Some of these clusters contain closely related NBS-LRR genes that originated from tandem duplication events. Others contain a mixture of less closely related NBS-LRR genes that originated from ectopic duplications, transposition, and segmental duplications (McDowell and Simon 2006). The proportion of NBS-LRR genes that exist in clusters varies among species. 38.2% of all NBS-LRR genes are contained in eight clusters in *L. japonicus* (Li et al. 2010), 50% of NBS-LRR genes are clustered in rice (Leister 2004) and 51% are clustered in *B. distachyon* (Tan and Wu 2012). Potato and *M. truncatula* have higher percentages of their NBS-LRR genes in clusters, at 73% and 80%, respectively (Jupe et al. 2012, Ameline-Torregrosa et al. 2008). These clusters allow for unequal recombination to occur between these genes. Recombination between alleles and genes has the potential to create new resistance specificities (reviewed in Michelmore and Meyers 1998, Young 2000). This has been observed in maize at the *Rp1* locus. The *Rp1* locus contains multiple NBS-LRR genes and recombination between them has been shown to be able to create new resistance specificities

(Hulbert and Bennetzen 1991, Richter et al. 1995). Interallelic equal recombination can also lead to new resistance specificities (Noél et al. 1999, Parniske and Jones 1999).

The *Pc* locus of sorghum

Periconia circinata is a fungal necrotroph that causes Milo disease in sorghum (Leukal 1948). The *Pc* locus in sorghum recognizes peritoxins produced by this pathogen and triggers the hypersensitive response, resulting in infection. Milo disease was a major problem in the United States in the 1930s, but soon disappeared, despite the continued presence of *P. circinata*. This was due to sorghum evolving from the dominant susceptible *Pc* phenotype to the recessive resistant *pc* phenotype. This mutation occurred naturally in approximately 1 out of every 8000 gametes. The *pc* phenotype is present in almost all sorghum lines grown in the United States and has no apparent fitness costs. The *Pc* phenotype persists in Africa, and presumably protects against a biotrophic pathogen not present in the United States (Schertz and Tai 1969).

Nagy et al. (2007) mapped the *Pc* locus in the inbred sorghum line BTx623 to a .9-cM region on the short arm of chromosome 9. They further localized it to a 110 kb segment which they then sequenced. They discovered 12 genes and analyzed each one in susceptible (*Pc/Pc*) sorghum and 13 resistant (*pc/*pc) offspring. All genes were identical except for three tandemly duplicated NBS-LRR genes that comprised the *Pc* locus. Rearrangements in these three genes were present in all 13 resistant mutants.

The three NBS-LRR genes were called *Pc* A, *Pc* B, and *Pc* C. The three genes are predicted to encode proteins that are 1277, 1194, and 1203 amino acids, respectively. *Pc* B and *Pc* C are 3606 and 3632 bp, respectively, while *Pc* A is considerably larger at 5113 bp due to an intron in the 3' end of the gene. *Pc* A and *Pc* B are separated by 12,638 bp of intergenic

sequence, while *Pc* B and *Pc* C are separated by 13,713 bp. *Pc* A and *Pc* C have especially high homology while *Pc* B has undergone more sequence divergence. The intergenic region between *Pc* A and *Pc* B also shares high homology to the intergenic region between *Pc* B and *Pc* C, including large retrotransposon insertions that share 92% sequence identity. One of the retrotransposons and either *Pc* A or *Pc* C were probably produced in a single duplication event. There are multiple stretches of several hundred bp regions of 100% identity where unequal homologous recombination could potentially occur in the intergenic regions. The 5' ends of *Pc* A and *Pc* C from 870 bp upstream of the genes to 1239 bp within the genes have 100% sequence identity. This region contains the conserved NBS domains. The 3' ends, containing the LRR region, are more divergent and contain polymorphisms throughout (Nagy and Bennetzen 2008).

The 13 resistant individuals mentioned earlier, called M1-M13, were further studied and characterized. M1-M7 contained a single paralog that was identical with *Pc* A in the 5' region and identical to *Pc* C in the 3' end, indicating unequal recombination between the A and C paralogs. Three mutants, M8-M10, contained a single paralog that was identical to *Pc* C. These were likely unequal recombination events between the 5' ends of *Pc* A and *Pc* C. Since the first 1238 bp are identical between A and C, recombination events that occurred here would produce a gene identical to *Pc* C. One mutant, M11, contained *Pc* A and *Pc* C, but was missing entirely *Pc* B. This could have been caused by an unequal recombination event between the A-B and B-C intergenic regions. Mutant M12 contained complete versions of *Pc* A and *Pc* C, but a truncated version of *Pc* B. *Pc* B had an internal deletion of 468 bp in the LRR region. Mutant M13 contained *Pc* A and a recombinant *Pc* B and *Pc* C paralog. In all 13 cases, *Pc* B was missing either completely or partially, indicating that *Pc* B is the gene responsible for causing susceptibility to *Periconia circinata*.

Any unequal recombination event that occurred within the flanking regions of *Pc* B would have deleted it and resulted in resistance to Milo disease in the gamete. However, 10 out of 12 recombination events occurred between *Pc* A and *Pc* C, and 7 of these occurred between the LRR regions, specifically within a 560 bp hotspot. This hotspot is the least conserved region in the gene. Generally, recombination is strongly biased towards regions of high sequence identity (Opperman et al. 2004, Puchta and Hohn 1991), so this result was surprising. Because it is this part of the gene that is responsible for pathogen recognition, recombination events here would be the most likely to generate new resistance specificities. Therefore, changes here are predicted to be the most beneficial to the plant in terms of potentially evolving resistance to new diseases. Hence, all of these results suggested that this unequal recombination hotspot in the LRR domains of the *Pc* genes might be the first case of site directed recombination detected in plants, and thus was worthy of further investigation.

LTR Retrotransposons

Plant genome sizes vary enormously (Kejnovsky et al. 2009, Murat et al. 2012). Within the dicots, which arose ~150 million years ago, genome sizes range >2000-fold from ~60 Mb in *Genlisea aurea* (Greihuber et al. 2006) to ~150 Gb in *Paris japonica* (Pellicer et al. 2010). This is a massive difference compared to mammals, which arose around the same time (Warren et al. 2008), but whose genome sizes only range from 1.6 Gbp in Carriker's round-eared bat (Smith et al. 2013) to ~8 Gbp in red viscacha rat (Gallardo et al. 1999). The grasses diverged only ~70 million years ago and their genome sizes vary 55-fold (Bennet and Smith 1991). Even closely related species can vary in their genome size. Despite this massive range, genome size does not correlate with biological complexity in higher eukaryotes and there is relatively little variation in the amount of gene content. This was known as the C-value paradox (Thomas 1971). It has since

been discovered that the variation in genome size is due to variation in the nongenic regions. Transposable elements, or transposons, make up the majority of these nongenic regions (Bennetzen 2000, Bennetzen 2002, Kidwell 2002). At the low end, at least 15-20% of the ~140 Mb Arabidopsis genome consists of transposons (Arabidopsis Genome Initiative 2000, Liu 2005). In many plant species, transposons make up the majority of the genome. Transposons are estimated to comprise approximately 85% of the ~2.4 Gb maize genome (Schnable et al. 2009), while in wheat it is likely that >90% of the genome is repetitive (Gill et al. 1996, Sandhu and Gill 2002, Li et al. 2004). Transposable element amplification can rapidly increase genome size and it is often lineage specific. Within the same genus, the genome of *Oryza australiensis* is more than twice the size as the genome of *O. sativa* due primarily to an additional ~400 Mb from three retrotransposon families (Piegu et al. 2006). An Australian clade of cotton, *Gossypium*, is almost 3 times the size of the American clade due to lineage-specific amplification of transposons (Hawkins et al. 2006). In the most dramatic case, the diploid teosinte, *Zea luxurians*, has more than doubled its transposon content in the less than 2 million years since its divergence from a shared lineage with *Zea mays* due to the amplification of more than 2 Gb of new transposons (Estep et al. 2013).

Transposon activity sometimes increases following a polyploidy event, further increasing genome size (Wendel et al. 2016). Transposable elements also make up the majority of the structural genomic diversity in angiosperms. (Kidwell and Lisch 1997, Bennetzen 2000). The Poaceae genomes are conserved in gene content and order (Ahn and Tanksley 1993; Barakat et al. 1997), while there is no correspondence between the transposons of intergenic regions (SanMiguel et al. 1996; Bennetzen et al. 1998; Tikhonov et al. 1999). The intergenic regions are

nearly 100% variable even among different haplotypes of maize (Fu and Dooner 2002, Song and Messing 2003; Brunner et al. 2005; Yao and Schnable 2005).

Transposons are mobile genetic elements that can amplify themselves during transposition. Transposons are divided into Class 1 and Class II elements. Class I elements consist of retrotransposons that replicate through reverse transcription of an RNA intermediate. Class II elements consist of DNA transposons that transpose without an RNA intermediate in a 'cut-and-paste' mechanism (Craig et al. 2002, Wessler 2006). Retrotransposons can be further divided into two groups, the long terminal repeat (LTR) retrotransposons and the non-LTR retrotransposons, which include the long interspersed elements (LINEs) and the short interspersed elements (SINEs) (Eickbush and Malik 2002).

LTR retrotransposons are named after their two flanking direct repeat sequences, the 5'LTR and 3'LTR. LTR retrotransposons range from 2 kb to 18 kb, and LTRs range from around 100 bp to several kb. The LTRs regulate transcription and are essential to produce intermediates in transposition. LTR retrotransposons contain several genes in their internal sequence that encode proteins that are responsible for transposition. LTR retrotransposons contain genes similar to the retrovirus *gag* and *pol* genes (Finnegan 2012). A few LTR retrotransposons contain a third gene similar to the *env* gene from retroviruses (Malik et al. 2000). Although this gene is important for retroviruses, it is usually non-functional in LTR retrotransposons (Song et al. 1994, Volff 2006, Miguel et al. 2008). LTR retrotransposons are further classified based on sequence and gene order into four groups: Ty1-copia-like, Ty3-gypsy-like, DIRS, or BEL-Pao-like. Most LTR retrotransposons fall into the Ty1-copia-like and Ty3-gypsy-like groups (Lynch 2007). These groups have been found in animals, fungi, protists, and plants, while BEL-Pao have so far only been found in metazoan genomes (Chaux and Wagner 2011).

The increase in genome size caused by LTR retrotransposon amplification can be partially counteracted by unequal recombination. The LTRs of an LTR retrotransposon can recombine with each other, deleting the internal sequence and one of the LTRs. The remaining solo LTR is made up of the 5' end of one LTR and the 3' end of the second LTR. Solo LTRs have been observed in many species (Shepherd et al., 1984; Sentry and Smyth, 1989; SanMiguel et al., 1996; Noma et al., 1997; Chen et al., 1998; Han et al., 2000, Ma et al. 2004). At the *Rar1* locus in barley, there is evidence of both intra-element and inter-element recombination (Shirasu et al. 2000). Most solo LTRs, however, are believed to have originated through intra-element recombination (Devos et al. 2002, Vitte and Panaud 2003).

When an LTR retrotransposon copies and inserts itself in the genome, both LTRs in the newly inserted retrotransposon are completely identical to each other. Mutations can then cause them to diverge. The age of an LTR retrotransposon can be estimated by the amount of sequence similarity between its two LTRs (SanMiguel et al. 1998). This divergence should also affect the frequency of solo LTR generation, and such a result was predicted in Baucom et al. (2009). These authors observed that intact LTR retrotransposons were rapidly removed from gene-rich regions, suggesting that these recombination-rich regions might more rapidly generate solo LTRs. Further studies on the issue of solo LTR generation by unequal recombination are warranted, and are the subject of a chapter of this dissertation.

Parasitic Plant Plastomes

The primary role of the chloroplast is photosynthesis, with additional functions in synthesis of pigments, starch, lipids, amino acids, and sulfur compounds. The chloroplast contains its own genome, the plastome (Palmer 1985, Wicke et al. 2011). The plastome of

vascular plants is believed to have a single origin from a cyanobacterium-like organism more than a billion years ago. The model is that a mitochondrion-containing eukaryote formed an endosymbiotic relationship with a cyanobacterium, and this became the ancestor of vascular plants (McFadden and van Dooren 2004, Keeling 2010). Since its origin, these models propose that there has been significant transfer of chloroplast sequence to both the nuclear and mitochondrial genomes. This also caused the chloroplast to become subject to regulation by nuclear genes. Modern plant plastomes contain an estimated 5-10% of the genes that were contained in the ancestral cyanobacterial genomes. It is estimated that at least 2000 Arabidopsis nuclear genes originated from the cyanobacterial genome (Martin et al. 2002). This transfer of chloroplast genes into the nucleus is ongoing today. Most of the transferred DNA drifts into functional and structural loss, but some of it inserts at a location where it can be transcribed (Kleine et al. 2009).

Plastomes are under strong selective pressure in most plants and therefore tend to be highly conserved in terms of gene content and order, nucleotide substitution rates, structure, and size (Raubeson and Jansen 2005). Most plastomes range from 120 to 160 kb (Palmer 1985). Originally, plastomes were thought to be arranged as single circular molecules. Although a percent of plastomes do have this structure, the majority are present as concatemers of two or more molecules, in either circular or linear configurations (Wicke et al. 2011). It is not yet known how these concatemers form or break apart. Plastomes have a quadripartite structure consisting of a large single copy (LSC) region and a small single copy region (SSC) separated by two virtually identical inverted repeat (IR_A and IR_B) regions (Kolodner and Tewari 1979). Recombination between the two inverted repeats plays a role in stabilizing the plastome (Maréchal et al. 2009). In angiosperms, LSCs are typically 80-90 kb, SSCs 16-27 kb, and

inverted repeats 20-30 kb (Wicke et al. 2011). The IRs include the ribosomal RNA genes and a number of other genes. The plastome usually contains 16S, 23S, and 5S rRNA genes and 27-31 tRNA genes, enough for translation of all amino acids, and at least three of the four subunits of a prokaryotic-type RNA polymerase (*rpoB, C1, C2*). The plastome also contains the majority of the genes for the polypeptides of Photosystem I, Photosystem II, the cytochrome b_6f complex, and ATP synthase (Stoebe and Kowallik 1999, Green 2011). Genes that function in photosynthesis (*atp* [ATPase], *ndh* [NADH dehydrogenase], *pet* [cytochrome b6/f], *psa* [photosystem I], *psb* [photosystem II], *ccsA* [cytochrome c biogenesis, *cemA* [heme-binding], *ycf3/4* [yeast cadmium factor 3-4], *rbcL* [RuBisCO large subunit), transcription, transcript maturation or translation (*rpo, matK* [maturase K], *rpl* [ribosomal protein L], *rps* [ribosomal protein S], *infA* [translation initiation factor 1]) and other pathways (*accD* [acetyl-CoA carboxylase beta subunit, *clpP* [ATP-dependent Clp protease], *ycf1, ycf2*) evolve at a lower rate than nuclear genes (Wolfe 1987). The number of different protein-coding genes and tRNAs are similar in most species, with much of the size difference due to the number of genes contained within the IRs (Green 2011). *Pelargonium hotorum* has one of the largest plastomes due to a three-fold increase in IR length, but still contains the normal number of different protein-coding genes and 29 tRNA genes (Chumley et al. 2006).

Unlike most species of plants, parasitic plants often contain highly divergent plastomes (dePamphilis and Palmer 1990, Wolfe et al. 1992b, Nickrent et al. 1997, Funk et al. 2007, McNeal et al. 2007, Krause 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016). Non-photosynthetic plants that obtain all their nutrients from the host plant are called holoparasites, while hemiparasites are photosynthetic plants that obtain some nutrients from the host and the rest from photosynthesis. Only around 10% of parasitic plants are holoparasites.

Hemiparasites can be further broken down into obligate parasites or facultative parasites. Obligate parasites require a host plant during at least part of their life cycle. Facultative parasites are able to complete their entire life cycle without a host plant, but can parasitize if one is available. Holoparasitic plant plastomes have lower selective pressure due to no reliance on photosynthesis. This can lead to higher nucleotide substitution rates, and functional gene loss through pseudogenization or physical loss through gene deletion, resulting in a smaller plastome size. This phenomenon is also observed, to a lesser degree, in hemiparasites that require photosynthesis (Wolfe et al. 1992b, Delavault et al. 1996, Funk et al. 2007, McNeal et al. 2007, Wickett et al. 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016). In some cases, an entire copy of one of the inverted repeats has been lost (Downie and Palmer 1992, Wicke et al. 2013, Wicke et al. 2016). Since the repeats play a stabilizing role in plastome structure, this leads to further destabilization and rearrangements in the plastome (Perry and Wolfe 2002). Although size reduction and gene loss are common in most non-photosynthetic parasites studied so far, the rate and extent of plastome reduction is lineage specific. Housekeeping genes can also be lost, suggesting that gene function alone does not explain which genes are lost (Wimpee et al. 1991, Wolfe 1992a and b, Colwell 1994, Funk 2007, McNeal et al. 2007, Delannoy et al. 2011). The likelihood that a gene is lost is influenced by its size, its association with essential genes, and its function (Lohan 1998 and Wolfe, Wicke et al. 2013). The genes that are retained are under purifying selection despite changes in evolutionary rates (Young and dePamphilis 2005).

The Orobanchaceae are the largest family of parasitic plants. They contain a single non-parasitic genus, *Lindenbergia*, that contains about a dozen species (Olmstead et al. 2001, Bennet and Matthews 2006). The Orobanchaceae include holoparasites and both facultative and obligate hemiparasites (McNeal et al. 2013). In this study, we assembled and annotated the plastomes of

five species of hemiparasitic Orobanchaceae to investigate the dynamics of these genomes, and, we discovered numerous cases of gene loss and other forms of rearrangement.

CHAPTER 2

UNEQUAL MEIOTIC RECOMBINATION BETWEEN SORGHUM NBS-LRR GENES IN TRANSGENIC MAIZE[1]

[1] Frailey, D.C., Chaluvadi, S.R., Bennetzen, J.L. To be submitted to *G3-Genes Genomics Genetics*.

Abstract

Most plants contain hundreds of NBS-LRR genes that condition race-specific resistance to a wide array of pathogens. Unequal recombination within NBS-LRR gene clusters has been shown to alter disease resistance gene composition and/or specificity, including at the *Pc* locus of sorghum. Among the three NBS-LRR genes at *Pc*, a ~560 bp hotspot for unequal homologous recombination was found to be located at the least conserved part of the gene, within the LRR region. Unequal recombination at this site would have a particularly high potential for creating new resistance gene specificities, and thus optimize the potential for new pathogen recognition and resistance. In this study, we created transgenic maize lines containing the *Pc* genes to identify and characterize unequal recombination events between *Pc* paralogs. We ran PCR on DNA extracted from pools of maize pollen to identify a total of 23 independent unequal recombination events. We calculated an unequal recombination rate of 1/5420 gametes, similar to the rate of ~1/8000 seen for natural instability at this locus in sorghum. However, unlike the sorghum observations, all 23 unequal recombination events were sited in the more conserved 5' ends of the genes, and none in the LRR regions. This lack of specificity to the LRR region was also observed for mitotic unequal recombination events in rice and maize.

Introduction

Plants need many disease resistance (R) genes in order to recognize and defend against a diverse range of pathogens in their environments. The most abundant type of R genes are the NBS-LRR genes. These genes encode proteins consisting of a nucleotide binding site (NBS) domain and a leucine rich repeat (LRR) region, in addition to some variable N-terminal and C-terminal domains (Dangl and Jones 2001, McHale et al. 2006). The NBS domain is responsible

for the first step in a signaling cascade that initiates the plant disease response. The NBS domains tend to be highly conserved (Tan and Wu 2012, Staskawicz et al. 1995), as expected for their conserved function. The LRR region consists of multiple LRR motifs and is responsible for pathogen recognition (Ellis et al. 1999). The LRR domains are quite variable and exhibit evidence of diversifying selection. Changes in the LRR region can cause the encoded protein to recognize and initiate resistance against a new pathogen (McHale et al. 2006).

NBS-LRR proteins activate multiple responses following pathogen recognition, including the hypersensitive response that results in programmed cell death localized to the site of infection. This is an effective mechanism to deal with biotrophic pathogens that require living plant tissue to obtain nutrients (Morel and Dangl 1997). Some pathogens have found a way to use the hypersensitive response against the plant. These necrotrophic pathogens are unable to invade healthy plant tissue. Each releases a toxin that is recognized by a specific NBS-LRR gene, which in turn invokes the hypersensitive response, and the resultant necrotic tissue now allows access by the pathogen (Lorang et al. 2007, Nagy et al. 2007; Faris et al. 2010).

Plants need to continuously evolve new disease resistance genes to fight rapidly evolving pathogens. One method that can generate new resistance alleles is recombination, particularly unequal recombination. If recombination is sited in the LRR regions, interallelic equal recombination has been shown to create new resistance specificities (Noél et al. 1999, Parniske and Jones 1999). Unequal recombination can duplicate or delete NBS-LRR genes (Anderson et al. 1996, Parniske et al. 1997**)**, which are often found in NBS-LRR gene clusters. Unequal recombination has also been shown to create novel resistance specificities (Sudupak et al. 1993, Sun et al. 2001, Webb et al. 2002**)**.

Unequal recombination was observed by Nagy et al. (2008) in the *Pc* locus in *Sorghum bicolor*. The *Pc* locus consists of three NBS-LRR paralogs, called *Pc* A, *Pc* B, and *Pc* C. *Pc* B and *Pc* C are both around 3.6 kb, while *Pc* A has an intron near the 3' end that increases gene size to ~5 kb. All three paralogs are tandem duplications arranged in the same orientation, with *Pc* B between *Pc* A and *Pc* C. There is ~12 kb of intergenic DNA between *Pc* A and *Pc* B and a slightly diverged copy of this same intergenic DNA is between *Pc* B and *Pc* C. All three paralogs share a relatively high amount of homology, but *Pc* A and *Pc* C are especially similar to each other. The first 1900 bp of *Pc* A and *Pc* C are identical, except a single SNP at bp 1239. The rest of the genes (2664 bp in *Pc* A and 1709 bp in *Pc* C) have SNPs throughout. The conserved portion of the gene contains the NBS domain, and the more divergent half contains the LRR region. The alignment between *Pc* A and *Pc* C is shown in Figure 2.1.

Pc B recognizes the peritoxin producted by fungus *Periconia circinata*. *P. circinata* is a necrotroph, so plants containing the *Pc* B gene are susceptible to infection. Nagy et al. 2007 identified 13 resistant sorghum seedlings, each individually derived by natural locus instability. One of these had an internal deletion in *Pc* B and the other 12 had *Pc* B completely missing due to unequal recombination. Ten of these *Pc* B loss events were the result of unequal recombination between *Pc* A and *Pc* C (Figure 2.1), while the other two retained both *Pc* A and *Pc* C, suggesting that these unequal recombination were sited in the repeated intergenic DNA that flanks *Pc* B (not shown).

In 7 of the 10 unequal recombinations between *Pc* A and *Pc* C, the events were sited within the 3' end of the gene, specifically within a 560 bp hotspot. The other 3 events occurred within the first 1238 bp of the 5' end of the gene. The 560 bp hotspot is in the least conserved part of the gene (Figure 2.2). Recombination is generally biased towards regions of high

homology (Opperman et al. 2004; Puchta and Hohn 1991), so the fact that most of the recombination events occurred within the least conserved region was unexpected. The hotspot was contained within the LRR domain, where changes should be the most beneficial for evolving new disease resistances. Hence, this suggests that a novel form of site-directed recombination was present at Pc, leading not to a precise site of recombination as in site-directed recombination at the yeast mating type locus (Haber 2012) or chordate immunoglobin gene creation (Dudley et al. 2005), but rather a bias towards a general region where new pathogen recognition specificity could be created (Nagy et al. 2008).

In order to test whether there was something intrinsic to *Pc* locus structure that sited unequal recombination to the LRR regions, a construct containing the yeast *ade2* gene, flanked by *Pc* A and *Pc* C, was transformed into haploid *ade2- S. cerevisiae*. Mitotic unequal recombinations that yielded *ade2-* colonies or colony sectors were detected, and analyzed for the site of recombination between *Pc* A and *Pc* C. Several dozen recombination events were analyzed and found to have occurred at the 5' homologous ends of the genes (J. Hawkins and J. Bennetzen, unpub. obs.). Hence, the LRR region bias observed in sorghum apparently involves some specific aspect of sorghum recombinational biology and/or genome structure.

Understanding unequal recombination at the *Pc* locus of sorghum is expected to provide a wealth of novel insights into disease resistance gene evolution in plants. Hence, we describe herein a transgenic strategy for identifying the inducing factors, rates and outcomes of unequal recombination in both mitotic and meiotic tissues at *Pc* in a close relative of sorghum, *Zea mays* (maize). Although unequal recombination events were detected and analyzed, our results indicate that *Pc* unequal recombination is quite different in its specificity in transgenic maize than it is in the natural sorghum genome environment.

Methods

Experimental Pc *construct*

The experimental construct consisted of several genes and gene fragments. The proposed, and achieved, construct structure was, in order, Ubi - *Pc* A – mTFP –Nos - *Pc* C – GFP - Nos Ubi stands for maize ubiquitin promoter, mTFP is monomeric teal fluorescent protein, Nos is the nos terminator sequence, and GFP is green fluorescent protein. In this construct, TFP was fused with the *Pc* A protein, thus yielding teal fluorescence driven by the strong Ubi promoter. Unequal recombination would be observed as loss of the TFP fluorescence but a great increase in GFP fluorescence now driven by the Ubi promoter. To create the experimental construct, we PCR amplified *Pc* A, mTFP, and Nos. Each PCR reaction consisted of 5-20ng template DNA, 10 µL Q5 Reaction Buffer, 10 µL betaine, 2.5 µL DMSO, 1 µL 10mM dNTPs, 2.5µL forward primer, 2.5µL reverse primer, 0.5 µL Q5 High-Fidelity DNA Polymerase, and 20 µL water. Touchdown PCR was performed on each reaction on an MJ Research PTC-200 Peltier Thermal Cycler. The program used was 98°C for 30sec; 10 cycles consisting of denaturation at 98°C for 20 sec, annealing at 68°C - 1°C each cycle for 15sec, elongation at 72°C for .5 to 2.5 min depending on template length; 25 cycles consisting of denaturation at 98°C for 15 sec, 58°C for 15 sec, elongation at 72°C for .5 to 2.5 min depending on template length; and a final elongation step at 72°C for 10 min. The primers added either a restriction site or a gateway cloning site to the ends of each gene. An attB1 Gateway cloning site was added to the 5’ end of *Pc* A. A *Kpn*I restriction site was added to the to the 3’ end of *Pc* A and the 5’ end of mTFP. A *Bam*H1 site was added to the 3’ end of mTFP and the 5’ end of Nos. The 3’ end of Nos contained an attB5r Gateway cloning sequence. Each PCR product was gel extracted using the QIAGen Purification Kit.

NEB *Bam*H1 and NEB *Kpn*I were used to digest the ends of each gene. We used Promega T4 ligase to ligate each the three sequences together. The final product was amplified with PCR using the 5' *Pc* A primer and the 3' Nos primer. The NEB PCR Cloning Kit was used to clone this fragment into NEB 10-beta Competent *E. coli*. Plasmids were isolated using the QIAGen Miniprep kit and sent to Macrogen for sequencing for confirmation. We used the BP Clonase II enzyme mix from Thermo Fisher Scientific to transfer our construct into a pDONR221p1-p5 vector.

We used the same PCR settings above to amplify *Pc* C, GFP, and Nos. These primers contained either overlapping sequence of the adjacent gene or overlapping sequence for the pUC19 vector. The 5' primer for *Pc* C also added an attB5 Gateway cloning site between the vector and gene sequence, and the 3' primer for Nos added an attB2 sequence between Nos and pUC 19. We used the Gibson assembly Cloning Kit to ligate the *Pc* C, GFP, and Nos together and into the pUC19 vector (Yanisch Perron et al. 1985). The BP Clonase II reaction was used to transfer *Pc* C-GFP-Nos into the pDONR 221p5-p2 vector.

We used an LR Clonase II Plus enzyme mix with the pDONR221p1-p5 vector (containing *Pc* A-mTFP-Nos) and the pDONR221p5-p2 vector (containing *Pc* C-GFP-Nos) to get both constructs into the pANIC 6D vector and the pANIC 10A vector (Mann et al. 2012). Both vectors are identical except in their selection marker. pANIC 10A contains hygromycin resistance which works well with rice and pANIC 6D contains bar resistance which works well with maize. Both vectors contained the maize ubiquitin promoter upstream of the *Pc* A gene. These two final vectors were sent to the Iowa Plant Transformation facility for *Agrobacterium*-mediated transformation into Hi Type II hybrid maize, a hybrid line of B73 and A188, and Japonica cv. Nipponbare rice. In this case, as in all other constructs, junctions were PCR

amplified and Sanger sequenced to determine whether the assembled sequences were present in the appropriate structure.

On occasion, some fusions with GFP or other fluorescent protein genes do not yield functional proteins (Wang and Chong 2003, Snapp 2010). Hence, prior to sending any construct for stable transformation into maize or rice, we employed a transient root assay system to investigate the quality of the constructs and whether the PC-FP fusions yielded a fluorescent phenotype. Ubi – *Pc* A – RFP was agroinfiltrated in rice roots and produced strong RFP expression. *Pc* C – GFP was similarly tested using the entcup2 promoter and similarly showed strong GFP expression. Hence, our analyses indicated that the constructs would yield the desired phenotypes, so we proceeded to generation stable constructs in transgenic plants.

The positive control construct consisted of *Pc* A-GFP-Nos. This construct was generated in the same way as the *Pc* A-mTFP-Nos fragment. We used an LR Clonase II reaction to insert this into the pANIC 10A vector and sent the vector to the Iowa Plant Transformation facility for *Agrobacterium*-mediated transformation into Japonica cv. Nipponbare rice.

Confirming presence of constructs

We extracted DNA from transgenic maize and rice lines using CTAB (Clarke 2009) and set up PCR using primers for *Pc* A, *Pc* C, mTFP, and GFP, using the touchdown PCR program described above. We sent the PCR products to Macrogen (Rocksville, MD) for Sanger sequencing to confirm each line contained the construct and the sequence was correct. We extracted RNA using the Invitrogen RNA extraction protocol with TRIzol. We then generated cDNA using the Quanta Biosciences qScript^{TM} Flex cDNA Synthesis Kit of mTFP and GFP. We

set up PCR using the touchdown PCR program described above with primers internal to mTFP and GFP to confirm both genes were being transcribed.

Somatic recombination

Despite detection of mTFP and GFP-homologous RNA, we were unable to detect any fluorescence in stably transformed tissues. We instead used a PCR assay to find several somatic recombination events in maize leaves. We divided leaves into 500mg sections and extracted DNA from each section using the CTAB DNA extraction protocol (Clarke 2009). We set up touchdown PCR using the above conditions and an elongation time of 2.5 min with the 5' primer located in the ubiquitin promoter and the 3' primer in GFP. All PCR products were run on a 1% agarose gel. Bands that were ~5 kb were excised and the DNA extracted using the QIAGen gel extraction kit. We sent the DNA to Macrogen for Sanger sequencing. Sequences were examined in Geneious 8.1.6 (Kearse et al. 2012) to determine where recombination had occurred.

Meiotic Recombination

Maize pollen was collected from individual plants from two different transgenic lines, 12t1 and 13t1, and measured out by mass into pools of pollen grains. We extracted DNA from each pool using the DNA extraction protocol described by Lopez-Sanchez 2005. Touchdown PCR was employed on total DNA extracted from individual pools of ~1400 pollen grains using the same conditions as described above, but with an elongation time of 2.5 min. I used a 5' primer with a binding site within the ubiquitin promoter and a 3' primer within GFP. Nested PCR was done on all PCR products with a second set of primers with ubiquitin and GFP binding sites. All PCR products ~ 5 kb in length were sent to Macrogen USA (Rockville, MD) for Sanger

sequencing to confirm that they contained a recombinant *Pc* locus. Sequences were examined in Geneious 8.1.6 to determine where recombination had occurred.

Results

Strategy and goals

There were two primary goals of this project. One was to develop a new approach in plants to study somatic and meiotic recombination. The second part of this study was to then use these techniques to study unequal recombination between the *Pc* genes in maize.

To create a fluorescent assay to detect somatic recombination, we first developed a construct that was transformed into rice and maize (Figure 2.3A). The first part of the construct consists of the maize ubiquitin promoter, the *Pc* A paralog, mTFP, and Nos terminator sequence. The maize ubiquitin promoter is active continuously at all developmental stages in all tissue types, so that this construct would be expressed throughout the plant. The stop codon has been removed from the end of the *Pc* A paralog so that a *Pc* A-mTFP fusion protein should be expressed from the ubiquitin promoter. mTFP is teal fluorescent protein, a more stable version of cerulean fluorescent protein (CFP). In addition, the Nos terminator sequence was added at the end of the mTFP sequence to guarantee that transcription and translation would stop after mTFP. The second part of the construct consists of the *Pc* C paralog, GFP, and the Nos terminator sequence. In most cells, the *Pc* A-mTFP fusion protein should be expressed, which we expected to produce teal fluorescence visible under a light with a wavelength of 458 nm. The *Pc* C-GFP fusion protein did not contain promoter, so would not be expressed. However, if unequal recombination occurs between *Pc* A and *Pc* C, then mTFP would be lost and instead there would be a *Pc* A/C hybrid protein fused to GFP (Figure 2.3B), now driven by the ubiquitin promoter.

Unequal recombination events were expected to be visible as fluorescent green under a light with a wavelength of 495 nm. We also created a positive control construct, which consisted of the ubiquitin promoter, *Pc* A, GFP, and Nos (Figure 2.3A). This was to test that GFP fluorescence was working. Unfortunately, neither mTFP nor GFP were visible in either construct, despite transient expression in rice roots of fusion *Pc* A – RFP and *Pc* C – GFP proteins showing strong fluorescence. This result remains unexplained, but led us to alter our strategy to recover unequal recombinants sited within this construct.

Instead of precise visual detection of unequal recombination as somatic sectors or meiotic gametes, we employed PCR on DNA derived from somatic tissues or gamete pools. Our PCR-based assay employed a forward primer with a binding site within the ubiquitin promoter and a reverse primer with a binding site within GFP. In any non-recombinant cells, this would yield an ~10 kb product that would likely not be amplified with the extension time employed. In any cells where the *Pc* genes had recombined, an ~5 kb product would be amplified. In this way, we were able to identify both mitotic and meiotic recombination events, determine where recombination had occurred within each one, and calculate the rate of meiotic recombination between these two genes.

Somatic unequal recombination in rice and maize

PCR and Sanger sequencing confirmed the constructs were present in both the maize and rice lines. cDNA analysis confirmed that mTFP and GFP were both being transcribed. We were unable to see any fluorescence, either teal fluorescence in the experimental transgenic lines or green fluorescence in the control lines. We tried multiple methods to see fluorescence, including

looking under a fluorescent microscope, chlorophyll bleaching by growing maize in the dark, and agroinfiltration of rice roots, but were still unable to see any fluorescence.

Since we could not use fluorescence to identify somatic recombination events, we instead used PCR on DNA extracted from ~500 mg leaf sections from our transgenic maize and rice lines. We found a total of 5 PCR products each that were ~5 kb in length from rice and maize tissue. Sequencing confirmed that they were the result of unequal recombination events between *Pc* A and *Pc* C. All five recombination events in both species occurred within the most 5' 1238 bp of the *Pc* genes, in the NBS region (Figure 2.2).

Meiotic recombination events in maize

We used the same PCR technique and primers as we did to identify somatic sectors on DNA extracted from pools of ~1400 pollen grains. We found 10 recombinant *Pc* genes out of 48 samples from maize line 12t1, and 13 recombinant *Pc* genes out of 41 samples from maize line 13t1, for a total of 23 recombinant *Pc* genes out of 89 samples. Nested PCR and Sanger sequencing performed by Macrogen confirmed they were recombinant events between *Pc* A and *Pc* C. We calculated a rate of 1 recombination event per 5540 pollen grains. Each recombination event was sequenced and all 23 events were within the first 1238 bp of the genes. This was the same pattern seen in somatic recombination in maize and rice.

Discussion

Our method of using fluorescence to detect somatic recombination events unfortunately did not work. We tried multiple methods to detect fluorescence but were unable to see either mTFP fluorescence or GFP fluorescence. Fusion fluorescent proteins sometimes do not fold

correctly, leading to a loss of fluorescent activity (reviewed in Snapp 2010). Since we could not use fluorescence to identify recombination events, we instead used PCR. By using primers flanking the *Pc* genes, we were able to identify recombination events based on the size of the PCR product. We were able to detect five somatic recombination events each from both maize and rice leaves. All ten events were sited in the highly conserved NBS regions. Because we do not know the time in development of the unequal events, or the exact number of cells with and without the unequal recombination in the leaf tissue samples, we are unable to determine the timing or calculate the rate of these events. The fluorescent protein assays would have made this possible, and allowed us to investigate environmental factors that affect the timing or rate, so we continue to pursue construct modifications that will make this possible.

Because maize produces great quantities of pollen that is easy to collect, compared to rice, we limited our meiotic analyses to maize. The PCR assay on pollen DNA was successful at detecting meiotic recombination events even though many pollen were pooled to meet the numerical demands of the project. We found a total of 23 unequal meiotic recombination events out of a total of ~124,600 pollen grains, for a rate of 1.8×10^{-4}, or 1 out of every ~5420 gametes. This was similar to the rate found in sorghum, which was 1.25×10^{-4} or 1 out of every 8000 gametes (Nagy et al. 2008). All 23 recombinant *Pc* genes were identical to *Pc* C. The first SNP between *Pc* A and *Pc* C is at the 1239^{th} bp, so recombination must have occurred somewhere within the first 1238 bp. Even though the recombinant *Pc* gene is identical to *Pc* C, we know that recombination occurred because the *Pc* gene is flanked by the ubiquitin promoter on the 5' end and GFP on the 3' end.

Our transgenic maize and the natural sorghum locus show very different patterns in recombination between the two *Pc* paralogs. One possibility is the process responsible for

directing recombination to the LRR region in sorghum is not present in maize. Sorghum and maize diverged from a common ancestor around 12 million years ago (Swigonová et al. 2004). This mechanism may not be present in maize, either because it evolved after maize and sorghum diverged, or it has been lost in maize after divergence. Maize has been extensively inbred since domestication, so a process that increased variability may have inadvertently been selected against during domestication.

Another possibility is that this mechanism does exist in maize, but is not acting on the transgenic *Pc* genes. The *Pc* construct was not directed to the Pc-orthologous region in maize, which appears to be contain the Rp3 disease resistance gene cluster (Nagy et al. 2007; Sanz-Alferez et al. 1995). Most of the maize genome consists of repetitive sequences that are suppressed for recombination. Another possibility is that the site-directed recombination seen in sorghum depends on specific epigenetic modifications. The newly inserted *Pc* genes in maize may not have acquired these modifications. It would be interesting to see in later generations of our transgenic maize lines if recombination patterns at the *Pc* genes shifted and resembled recombination in sorghum.

This experiment only looked at recombination between one pair of disease resistance genes. The *Pc* genes were chosen because they were recombining in sorghum within the LRR domain. This experiment could be repeated with other disease resistance gene clusters from maize, by a similar PCR strategy but without the use of transgenics, to see if maize genes recombine in the LRR regions. If so, this would suggest the reason we only saw recombination in the 5' region of the *Pc* genes was due to some incompatibility between the sorghum genes and maize mechanism for directing recombination to the LRR region, or that a natural NBS-LRR gene location is needed to see this type of site-directed recombination.

If other maize NBS-LRR genes only exhibit equal or unequal meiotic recombination within the 5' ends of the genes, this would suggest several possibilities. One, maize may not have the same mechanism as sorghum to direct recombination to the LRR region of NBS-LRR genes. Two, the reason we saw recombination at the 5' end of the *Pc* genes was due to the chromosomal insertion position. And, three, perhaps the newly inserted genes do not have the necessary established epigenetic marks to induce site-directed recombination.

Although the results from the *Pc* recombination were not as interesting as hoped, we did develop a successful method for detecting meiotic recombination in maize pollen. Meiotic recombination at any one specific locus is relatively rare, averaging less than 1/20,000 pollen per 1 kb of gene (Yandeau-Nelson et al. 2006). The advantage of this method is that we can screen thousands of meiotic products with a single PCR, and therefore find these rare events with relatively little time or effort. This technique could be used to study meiotic recombination in any plant that produces a sufficient amount of pollen.

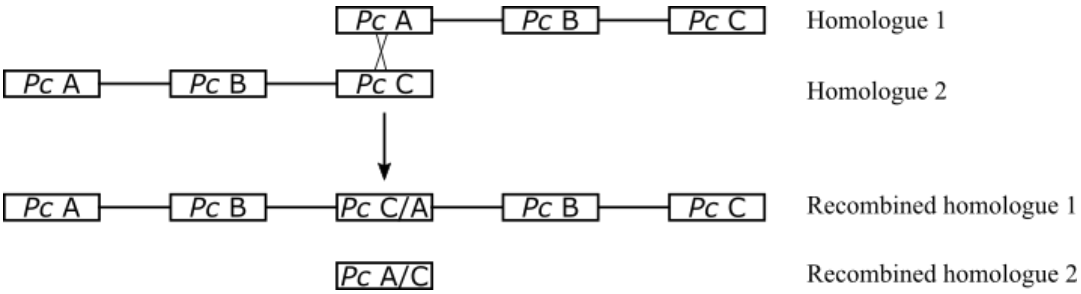


Figure 2.1. Unequal recombination between *Pc* A and *Pc* C. Unequal recombination between the *Pc* genes creates a recombinant *Pc* gene along with reciprocal duplication and deletion events. Recombined homologue 2 shows replacement of *Pc* A and *Pc* C with a recombinant chimera of the two genes, while *Pc* B is deleted so the plant receiving this chromatid will be resistant to infection by *Periconia circinata.*

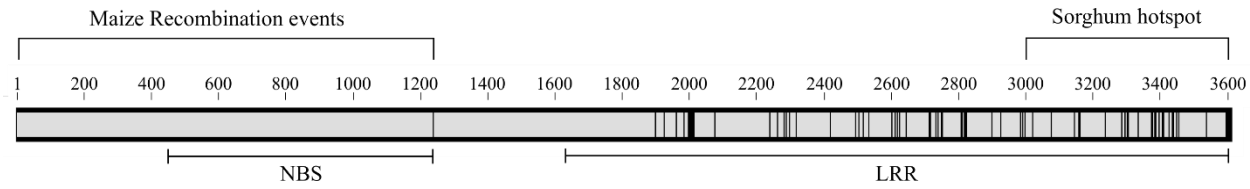
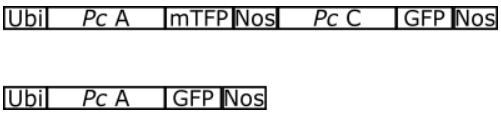


Figure 2.2. Alignment between *Pc* A and *Pc* C genes. The gray bar shows the alignment between *Pc* A and *Pc* C. Each black line represents a single nucleotide difference between the two genes. The NBS and LRR domains are indicated along the bottom. The 560 bp hotspot in sorghum and the region where all the maize recombination events are found are indicated along the top.

A)



B)

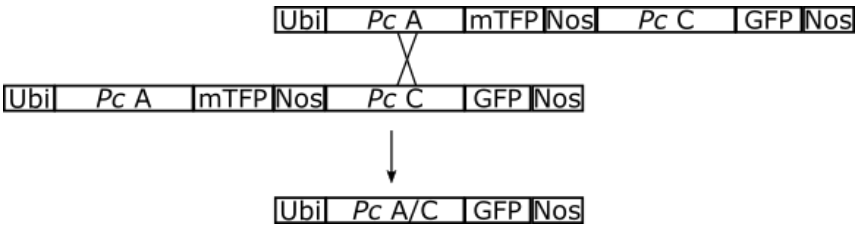


Figure 2.3. *Pc* constructs transformed into maize. A) Experimental construct (top) contains the maize ubiquitin promoter (Ubi), *Pc* A, mTFP, Nos, *Pc* C, GFP, and a second Nos terminator. The ubiquitin promoter causes expression of the *Pc* A-mTFP fusion protein. Nos prevents expression of *Pc* C-GFP. The control construct (bottom) contains the maize ubiquitin promoter, *Pc* A, GFP, and Nos. This construct was entered into transgenics to make sure that we could detect green fluorescence with our assay system.

B) Unequal recombination between *Pc* A and *Pc* C removes the 3' end of *Pc* A, mTFP, the first Nos, and the 5' end of *Pc* C. A new *Pc* A/C recombinant gene is created, and the ubiquitin promoter causes expression of the *Pc* A/C-GFP fusion protein.

CHAPTER 3

UNEQUAL MEIOTIC RECOMBINATION BETWEEN THE LONG TERMINAL REPEATS (LTRS) OF MAIZE RETROTRANSPOSONS[2]

[2] Frailey, D.C., Bennetzen, J.L. To be submitted to *PLOS Genetics.*

Abstract

Transposable elements (TEs) are responsible for the majority of structural variation in plant genomes. The LTR retrotransposons are the most abundant TEs in plant and are most important for genome size variation. Genome size increases can be partially counteracted through deletions caused by unequal recombination between the directly repeated LTRs in an LTR retrotransposon, generating solo LTRs, but neither the rates nor specificities of such events have ever been determined in plants. In this project, we looked for solo LTR generation at 50 LTR retrotransposons located near the ends of the short arm of chromosome 1 and the long arm of chromosome 9. We found a total of 23 such unequal recombination events, and all were associated with 10 LTR retrotransposons. We found that unequal recombination rate had a positive correlation with LTR sequence identity and a negative correlation with the amount of DNA methylation. We found no correlations with chromosomal position, LTR length, or the distance to the nearest genes, although this could be due to our sample not being sufficiently numerous or variable. We also mapped the crossover points for each recombinant LTR and found an apparently random distribution of crossover points.

Introduction

Plant genomes vary greatly in size. Dicot genomes range from 60 Mb (Greihuber et al. 2006) in *Genlisea aurea* to 150 Gb in *Paris japonica* (Pellicer et al. 2010). The grasses diverged only ~70 mya (Gaut 2002), yet their diploid genome sizes vary at least 40-fold (Bennet and Smith 1991). Despite this massive range, genome size does not correlate with biological complexity in higher eukaryotes. This observation has been called the C-value paradox, which was solved by the observation that TEs (especially retroelements) can rapidly proliferate to

occupy genome space without huge biological effects (Bennetzen et al. 1998; Tikhonov et al. 1999; Estep et al. 2013). Although Poaceae genomes are highly conserved in gene content and order (Ahn and Tanksley 1993; Barakat et al. 1997), there is little to no correspondence between the TEs in intergenic regions when comparing different grasses (Bennetzen et al. 1998; Tikhonov et al. 1999).

TEs are mobile genetic elements that can make extra copies through the act of transposition (Craig et al. 2002). These are inserted in new locations in the genome, causing an increase in genome size. Transposons can be Class 1 transposable elements, consisting of retrotransposons, or Class II elements, consisting of DNA transposons (Wessler 2006). Retrotransposons replicate through reverse transcription and an RNA intermediate. They can be divided into two groups, the long terminal repeat (LTR) retrotransposons and the non-LTR retrotransposons. LTR retrotransposons contain several genes in the in the region between the two LTRs, which encode proteins involved in transposition (Malik et al. 2000). LTR retrotransposons are subclassified into Ty1-*copia*-like, Ty3-*gypsy*-like, or BEL-Pao-like superfamilies based on sequence similarity and the order of their genes (Lynch 2007). Ty1-*copia*-like and Ty3-*gypsy*-like TEs have been found in animals, fungi, protists, and plants, while BEL-Pao have only been found in animal genomes so far (Chaux and Wagner 2011). The LTR retrotransposons are some of the most abundant elements in eukaryotic genomes and comprise the majority of most plant genomes. The maize genome contains 150,000 to 250,000 LTR retrotransposons (SanMiguel and Vitte 2009) and altogether they make up ~75% of the genome (Schnable et al. 2009).

The increase in genome size from LTR retrotransposon amplification can be partially counteracted by unequal recombination. The LTRs of an LTR retrotransposon can recombine

with each other, deleting the internal sequence and one of the LTRs. The remaining solo LTR is made up of the 5' end of one LTR and the 3' end of the second LTR. Solo LTRs have been observed in many species (Shepherd et al., 1984; Sentry and Smyth, 1989; SanMiguel et al., 1996; Noma et al., 1997; Chen et al., 1998; Han et al., 2000). In the Arabidopsis genome and at the *Rar1* locus in barley, there is evidence of both intra-element and inter-element recombination (Shirasu et al. 2000; Devos et al., 2002). Most solo LTRs, however, have originated through intra-element recombination (Devos et al., 2002; Vitte and Panaud 2003). When an LTR retrotransposon copies and inserts itself in the genome, both LTRs in the newly inserted retrotransposon are completely identical to each other. Mutational drift can then cause them to diverge. The age of an LTR retrotransposon can be estimated by the amount of sequence similarity between two LTRs (SanMiguel et al. 1998; Perlman and Boeke 2004).

The siting of meiotic recombination is usually not precisely defined, but there are several factors that influence recombination rates across the genome. One such factor is chromosomal position, with some regions of the genome having relatively high recombination rates (hotspots), and others having relatively low rates or no recombination. In plants, recombination rates are positively correlated to gene dense regions. In maize, recombination rates also increase as distance from the centromere increases, followed by a decrease again immediately next to the telomeres (McMullen et al. 2009, Rodgers-Melnick et al. 2015). Another factor that influences recombination is the amount of sequence identity between the recombining homologs, with more similar homologs having higher recombination rates than more divergent sequences. In some cases, even a small amount of changes can dramatically affect recombination rates. In maize, meiotic recombination at the *a1* locus was seven-fold lower between two alleles with 97% identity compared to two alleles with 99.9% identity (Yandeau-Nelson et al. 2006). The length of

the recombining sequences is also positively correlated with recombination rates. In plants, a few hundred bp are capable of recombining, although the minimum length of homology required for recombination is not yet known (Opperman et al. 2004; Puchta and Hohn 1991). DNA methylation is also associated with changes in recombination rates. Recombination hotspots are correlated with low amounts of DNA methylation, and methylation of a hotspot has been shown to severely reduce recombination in the fungus *Ascobolus immersus* (Maloisel and Rossignol 1998) as well as in Arabidopsis (Yelina et al. 2015).

In this study, we identified and characterized 23 recombination events between the LTRs of individual LTR retrotransposons. We chose 50 LTR retrotransposons across a 7.8 Mb region of the short arm of chromosome 1 and the 6.6 Mb region of the long arm of chromosome 9 that had been very accurately sequenced and annotated (Bruggmann et al. 2006). Both regions are located near the ends of the chromosome. To find recombination events, we ran PCR on DNA extracted from numerous pools of ~1400 pollen grains. We collected pollen from each maize plant and divided it into pools, with each pool containing pollen collected from only a single plant. We set up PCR using primers that would amplify the LTR if recombination had occurred between the two LTRs of an LTR retrotransposon in any one of the 1400 pollen grains. For each LTR retrotransposon, we ran several PCRs on a total of ~22,500 pollen grains. We compared the recombining LTR retrotransposons with the non-recombining retrotransposons to look for correlations of recombination rate with sequence identity, LTR length, chromosome position, methylation, and LTR distance from the nearest genes. We also sequenced each recombinant LTR to both confirm it was a recombinant LTR and to identify where the crossover point had occurred.

Methods

Identifying recombination events

We downloaded the MASiVEdb Sirevirus plant retrotransposon database (Bousios et al. 2012) and the MIPS Repeat Element Database (Nussbaumer et al. 2013) and aligned LTR retrotransposons to chromosome 1 and chromosome 9 of the AGPV3.21 maize B73 assembly using Geneious 8.1.6 (Kearse et al. 2012). LTR retrotransposons were chosen from a 7.8 Mb region of the short arm of chromosome 1 and a 6.6 Mb region of the long arm of chromosome 9. Both regions have been well sequenced and annotated by Bruggmann et al. 2006. We randomly chose 50 LTR retrotransposons, shown in Table 3.1, within these two regions based on whether we could design working primers. For each retrotransposon, we designed a total of 4 primers. Two of them, the junction primers, spanned the junctions at the two ends of the retrotransposon and faced inward. The other two, the control primers, were within each LTR and faced the junction primers. Figure 3.1 shows the location of the primers for each LTR retrotransposon. For each junction primer, we set up PCR with the corresponding control primer to test that the junction primer worked. Each PCR reaction consisted of 5-20ng template DNA, 10 μL Q5 Reaction Buffer, 10 μL betaine, 2.5 μL DMSO, 1 μL 10mM dNTPs, 2.5 μL forward primer, 2.5 μL reverse primer, .5 μL Q5 High-Fidelity DNA Polymerase, and 20 μL water. Touchdown PCR was performed on each reaction on an MJ Research PTC-200 Peltier Thermal Cycler. The program used was 98°C for 30sec; 10 cycles consisting of denaturation at 98°C for 20 sec, annealing at 68°C - 1°C each cycle for 15sec, elongation at 72°C for 0.5 min; 25 cycles consisting of denaturation at 98°C for 15 sec, 58°C for 15 sec, elongation at 72°C for 0.5 min, and a final elongation step at 72°C for 10 min. A third PCR test was done for each set of primers

using the two junction primers on standard B73 maize DNA to make sure they did not produce non-specific PCR products.

Maize pollen was collected from individual maize B73 plants and measured out by mass into pools. A total of 0.42 milligrams of pollen was used per pool, which our pollen counting procedure estimated to contain ~1400 pollen. We extracted DNA from each pool using the DNA extraction protocol described by Lopez-Sanchez 2005, which was modified to extract DNA from maize pollen using extraction buffer (200 mM Tris-HCl, 250 mM NaCl, 25 mM EDTA, 0.5% SDS). We set up Touchdown PCR on maize pollen DNA using the same conditions as described above using the junction and control primers. For each LTR retrotransposon, we set up PCR reactions on DNA from 16 different pollen pools. All PCR products were run on a 1% agarose gel. All PCR products that were the expected size of a solo LTR were excised and the DNA extracted using the QIAGen gel extraction kit. We sent the DNA to Macrogen USA (Rockville, MD) for Sanger sequencing to confirm whether they were recombinant LTRs.

LTR characteristics

Each LTR pair for each LTR retrotransposon was aligned using Clustal W (Larkin et al. 2007) to identify the SNPs and indels between the two LTRs. The number of identical sites and the pairwise % identity was recorded for each pair. The chromosomal position was recorded for each LTR retrotransposon and the distance from the nearest end of the chromosome was measured. The methylome of the B73 maize genome was obtained from the Plant Methylome DB (Gent et al. 2012) and the number of CG, CHG, and CHH sites for each LTR was recorded. The number of methylated CG, CHG, or CHH sites were divided by the total number of CG, CHG, or CHH, respectively, to calculate the percentage of each type of methylation. RLTR33

and RLTR 44 were not covered and were left out of this analysis. The B73 RefGen_v3 Gene Models track was downloaded from MaizeGDB and aligned to the chromosomes in Geneious 8.1.6. We measured the distance from each LTR retrotransposon to the nearest proximal and distal gene. For each characteristic, we found the mean and the median. We divided the LTR retrotransposons into two groups below and above the median and two groups below and above the mean. We compared the number of recombining LTR retrotransposons and the total number of recombination events between the two groups to look for correlations with recombination rates.

Results

We examined a total of 50 LTR retrotransposons for unequal recombination between the LTRs. For each LTR retrotransposon, we ran PCR on a total of ~22,500 pollen grains. We found a total of 23 unequal recombination events, all associated with 10 LTR retrotransposons. For these 10 elements, the number of events ranged from 1 to 4 out of 16 pollen samples. The number of unequal recombination events detected for each of these 10 LTR retrotransposons is shown in Table 3.2, yielding a median of 2 and a mean of 2.3. If the low number of unequal recombination events was exclusively an outcome of 50 equally recombinagenic TEs that yielded only rare events, then the median and mean in this group should have been 1. Because the median was 2, this indicates that the 10 LTR retrotransposons that yielded unequal recombination events were, at the very least, enriched for recombinagenic LTR retrotransposons that differed from the other 40 LTR retrotransposons that had not yielded unequal recombination outcomes.

The LTRs ranged in size from 576 bp to 1952 bp, with a median size of 1262 bp and a mean size of 1263 bp. A total of 25 elements had LTRs that were below both the median and the mean. Six of these yielded unequal recombination events, 17 in total. The 25 LTR retrotransposons that had LTRs above the median and the mean, included 4 that unequally recombined for a total of 6 recombination events. All LTR lengths are shown in Table 3.3 and Table 3.4.

The sequence identity between the LTR pairs ranged from 86.9% to 100%, with a median of 97.6% and a mean of 96.7%. The 25 LTR pairs below the median exhibited only one unequally recombining element with a total of 3 events. 25 LTR pairs were above the median, with 9 of them unequally recombining for a total of 20 events. The 22 LTR retrotransposons below the mean yielded one unequally recombining for a total of 3 recombination events. The 28 LTR retrotransposons above the mean had 9 unequally recombining for a total of 20 events. Table 3.5 shows the LTR pairwise identities for all LTR retrotransposons.

Twenty of the LTR retrotransposons were on the short arm of chromosome 1, with six of them unequally recombining for a total of 10 events. The other 30 LTR retrotransposons were on the long arm of chromosome 9, with 4 of them yielding a total of 13 unequal recombination events. The distance of the LTR retrotransposons from the nearest telomere ranged from 309,117 bp to 8,354,575 bp, with a median of 3,708,822 bp and a mean of 4,159,485 bp. The 25 LTR retrotransposons below the median exhibited 6 unequally recombining for a total of 15 events. The 25 LTR retrotransposons above the median had 4 unequally recombining for a total of 8 events. The 24 LTR retrotransposons below the mean in distance from the telomere yielded 6 unequally recombining for a total of 15 events. The 26 LTR retrotransposons above the mean

had 4 unequally recombining for a total of 8 events. Table 3.6 shows the chromosome number and position for each LTR retrotransposon investigated.

The distance to the nearest gene ranged from 0 bp to 59,475 bp with a median of 6485 bp and a mean of 10,841 bp. The 25 LTR retrotransposons below the median distance had 6 unequally recombining elements for a total of 16 events. The 25 LTR retrotransposons above the median distance from a gene had 4 unequally recombining for a total of 7 events. The 16 LTR retrotransposons that were above the mean distance from a gene had 3 unequally recombining, yielding a total of 6 events. The 34 LTR retrotransposons closer to a gene than the mean had 7 unequally recombining for a total of 17 events. We also looked at the average of the distances to the nearest gene on either side of the LTR retrotransposon. The average distances ranged from 0 bp to 116,003 bp, with a median of 22,949 bp and a mean of 27,065 bp. The 25 LTR retrotransposons below the median distance from a gene had 5 unequally recombining for a total of 13 events, while 25 LTR retrotransposons above the mean distance had 5 unequally recombining for a total of 10 events. A total of 28 LTR retrotransposons were below the mean distance from a gene, with 5 of them unequally recombining to yield a total of 13 events. The 22 LTR retrotransposons above the mean distance from a gene had 5 unequally recombining to yield 10 events. Two of the LTR retrotransposons, RLTR4 and RLTR43, were within an intron of annotated genes and were recorded as being 0 bp away from the nearest gene on either side. All distances of the LTR retrotransposons from the nearest gene are shown in Table 3.7.

We also individually measured CG, CHG, and CHH methylated sites, as shown in Tables 3.9 through 3.11. The percentage of methylated CG sites out of total CG sites ranged from 1.9% to 100%, with a median of 88.3% and a and a mean of 72.3%. The 24 LTR retrotransposons with below the median number of CG methylations per kb had 3 of them unequally recombining to

yield a total of 6 events. The 25 LTR retrotransposons above the median had 7 unequally recombining to yield 17 events. The 16 LTR retrotransposons below the mean yielded 3 unequally recombining for a total of 7 events. The 32 LTR retrotransposons above the mean had 7 unequally recombining to yield 16 events. The percentage of CHG sites ranged from 2.0% to 100%, with a median of 71.8% and a mean of 50.4%. The 24 LTR retrotransposons below the median had 4 unequally recombining for a total of 10 events. The 24 LTR retrotransposons above the median had 6 unequally recombining for a total of 13 events. The 20 LTR retrotransposons below the mean had 8 unequally recombining to yield 10 events. The 28 LTR retrotransposons above the mean had 2 unequally recombining for a total of 13 events. The percentage of CHH sites ranged from 0.0% to 40.0%, with a median of 3.9% and a mean of 7.2%. The 24 LTR retrotransposons below the median had 5 unequally recombining to yield 9 events. The 24 LTR retrotransposons above the median had 4 unequally recombining for a total of 15 events. The 33 LTR retrotransposons below the mean had 7 unequally recombining to yield 18 events. The 15 LTR retrotransposons above the mean had 3 unequally recombining for a total of 5 events. Tables 3.8-3.10 show the amount of methylation for each LTR.

All of these investigations of LTR retrotransposon properties relative to the frequency of unequal recombination events were subjected to p-value statistical analysis to see if any correlated. As shown in Table 3.11, only a high degree of LTR homology within an LTR retrotransposon (p value 0.0047) and exhibited significant correlation with a high frequency of unequal homologous recombination to yield solo LTRs.

The number of sequence variations between the LTRs of each LTR retrotransposon ranged from 1 to 31, and this could be used to indicate the approximate sites where unequal recombination events were resolved to create chimeric solo LTRs. Figure 3.2 shows the 5’ and 3’

LTRs with all variations between them indicated, as well as all recombinant LTRs and which version of each variant they contain. For instance, the SLTR39 alignment was 1251 bp and contained two sequence variations at bp 642 and 1029. There were four unequal recombinants. Two contained the most 5' variation from the 5' LTR, while the other two unequal recombinants contained the both 5' LTR variations. Figure 3.3 depicts a normalized LTR that shows the approximate unequal recombination positions for all of the 23 events detected in this study.

Discussion

Because DNA removal by illegitimate recombination is so rapid in genomes like maize that most TE sequences are removed within less than 2 million years (Ma et al. 2004), solo LTRs must be generated at a fairly high rate before the entire TE is obliterated by an accumulation of deletions that degenerate LTR homology. Also, it is known that the rate of solo LTR generation is variable across genomic regions. For instance, the ratio of solo LTRs to intact LTR retrotransposons is highest at the core of centromeres and in genic regions, suggesting that these locations are more recombinationally active (at least for unequal events) than gene-poor heterochromatin (Ma and Bennetzen 2006). Moreover, there are no results that suggest the process of unequal meiotic recombination involves any mechanistic components or biases that are not also pertinent to equal homologous recombination in meiosis, so our investigation of unequal recombination to generate solo LTRs should be informative vis-à-vis meiotic recombination per se.

In an investigation of 22,400 maize pollen for unequal recombination events that generated solo LTRs from any of 50 LTR retrotransposon, we found 10 unequally recombining LTR retrotransposons that generated 23 solo LTR events. Each one retained the flanking target

site duplications, indicating that this was standard unequal homologous recombination. These results provide an overall rate of 1/48,700 of unequal recombination in meiosis. If unequal recombination rates were equally low for all 50 LTR retrotransposons, then one would expect to only see 0 or 1 unequal recombination event per element. The fact that numerous LTR retrotransposons yielded 2 or more unequal events indicated that they are hotspots for this activity. If one measures the rate of unequal homologous recombination among the 10 LTR retrotransposons that yielded these events, then one calculates a frequency of 1/9740. This number is fairly similar to the 1/1000 to 1/10,000 rate seen for unequal recombination for tandemly repeated genes in multicellular eukaryotes (Sun et al. 2001, Yandeau-Nelson 2006, Nagy et al. 2007).

Given that we discovered two separable categories of LTR retrotransposons, one generating solo LTRs and one not doing so, it seemed appropriate to investigate whether any other genome properties differentiated these two cohorts. Undoubtedly, some of the non-recombining LTR retrotransposons would have yielded unequal recombinants if a greater number of pollen had been investigated, because one detected event is numerically quite similar to zero detected events. Hence, these two categories will have some noise contributed by a "non-recombining" set that actually contains some number of undetected unequal recombinations. But this false negative problem should only decrease the variance seen between the two cohorts, and thus any detected significant variance is likely to be an underestimate of the true variance.

We compared the unequally recombining LTR retrotransposons to the non-recombining elements with regard to sequence identity, LTR length, chromosomal position, distance to the nearest genes, and DNA methylation. Our results indicate that only a high degree of LTR

sequence identity and a low degree of DNA methylation correlated with high levels of unequal meiotic recombination.

Previous studies have shown a positive correlation between template length and recombination rates (Opperman et al. 2004, Puchta and Hohn 1991**)**. We found no apparent correlation between LTR size and recombination rates. However, 44 out of the 50 LTR retrotransposons we looked at contained LTRs from 1161 bp to 1413 bp. Only three LTR retrotransposons had much smaller LTRs, ranging from 576 bp to 643 bp, and only three had much larger LTRs, ranging from 1617 bp to 1952 bp. None of the LTR retrotransposons in these smaller and larger LTR groups unequally recombined, so we do feel that we have sufficient data to draw many conclusions about the effect of LTR size. We can conclude, however, that if any such effect were large and very sensitive to the ~3-fold size range we investigated, then we would have detected such an effect. If we had larger sample sizes or a broader range of LTR lengths, we would be more likely to detect any correlation between LTR length and unequal recombination rate.

There does seem to be a correlation with LTR sequence identity and unequal recombination. Nine out of the 10 recombining LTR pairs had a sequence identity of 98% or above, even though the majority of the LTR pairs had a sequence identity below 98%. Previous studies have shown that sequence divergence can have a strong effect on reducing recombination rates, with even a few SNP changes causing reductions in recombination rate (Opperman et al. 2004). Our results agree with previous studies to suggest that sequence divergence is an important factor that influences recombination. The one exception we see in our study is NLTR21, which has a sequence identity of only 90.8% but still produced 3 unequal

recombination events. The unusual properties of this LTR retrotransposon will be further discussed below.

Previous studies have shown that in maize, cross-over rates are at a minimum near the centromere and increase with distance from the centromere (McMullen et al. 2009, Rodgers-Melnick et al. 2015). However, we did not see any correlation between chromosomal position and unequal recombination rate. The unequally recombining LTR retrotransposons were spread throughout the regions we studied, with no apparent clustering. However, the LTR retrotransposons we studied spanned a 6.1 Mb region on chromosome 1 and a 6.5 Mb region on chromosome 9. These chromosomes are ~300 Mb and ~114 Mb, respectively, so we only examined a small portion of each chromosome, and each region was gene-rich and centromere-distal, suggesting an expected high rate of meiotic recombination overall. If we had looked at LTR retrotransposons from a broader range across the genome, we would likely have seen a correlation between chromosome position and unequal recombination rates.

We did not see any correlation between unequal recombination and the distance to nearest gene. We looked for a correlation with both the distance to the nearest gene and the average of the distances from the nearest genes on either side, but there was no apparent correlation between any of these measurements and unequal recombination rate. Once again, the fact that both investigated regions are relatively gene rich suggests that additional studies are needed comparing solo LTR generation in highly heterochromatic regions with euchromatic regions.

We did not see any apparent correlation between unequal recombination and the number of methylated sites. We looked at each type of methylation, CG, CHG, and CHH, individually. For each type, we calculated the percentage of methylated sites out of the total number of sites.

We found no statistically significant correlation between any of the types of methylation and unequal meiotic recombination.

In *S. cerevisiae*, meiotic gene conversion often shows polarity within genes, with a higher frequency at the 5' ends and decreasing frequency towards the 3' ends (Schultes and Szostak 1990, Malone et al. 1992). Conversely, a genome wide analysis in Arabidopsis found no such polarity (Sun et al. 2012). In maize, some genes show polarity (Xu et al. 1995, Patterson et al. 1995, Eggleston et al. 1995, Dooner and He 2014), while others have random gene conversion distributions (Dooner and Martinez-Ferez 1997, Okagaki and Weil 1997). For each LTR retrotransposon, we aligned the LTRs to identify all variations between them. We sequenced each recombinant LTR and saw whether each variant originated from the 5' LTR or the 3' LTR to identify the region where the crossover took place. We observed crossovers distributed across the LTRs with no obvious polarity or clustering towards any particular region of the LTRs. NLTR25, NLTR27 and RLTR3 were the only LTR retrotransposons with only one unequal recombination event. NLTR21 produced three recombinants that all looked identical to the 5' LTR and to each other. All other LTR retrotransposons produced multiple distinct chimeric LTRs that were distinguishable from the each other by their crossover site.

There are several unique factors about NLTR21 compared to the rest of the unequally recombining LTR retrotransposons. NLTR21 had both the lowest sequence identity out of all the unequally recombining LTRs. All three of the LTR recombinants look identical to the 5' LTR. This is the only LTR retrotransposon that showed this pattern in the unequal recombinants. We looked more closely at the region around NLTR21 to see if there was anything unique about its location that may explain these results. The structure of NLTR21 is shown in Figure 3.4. NLTR 21 has the most drastic difference between the distance from the nearest distance gene distal and

the distance from the nearest proximal gene. It is also the furthest of the unequally recombining LTR retrotransposon from the nearest gene at the 5' end, and is the closest of unequally recombining LTR retrotransposons to the nearest gene at the 3' end. The nearby gene is ~1 kb away from the 3' end of the LTR retrotransposon and was annotated as GRMZM2G703469, although the function of the gene is unknown. There are no other annotated transposons between NLTR21 and GRMZM2G703469. On the 5' end of NLTR21, there is a truncated 3' end of another LTR retrotransposon 30 bp away. This LTR retrotransposon contains 1679 bp of the internal sequence, including the *pol* gene, and an 1146 bp 3' LTR. Both the internal sequence and the LTR together share 82.1% sequence identity with the 3' end of NLTR21, while the LTR shares 76.5% identity with the 3' LTR and 74.4% identity with the 5' LTR. It is possible that the proximity of either the partial LTR retrotransposon, the gene, or a combination of both influences recombination in some way that gives us the unexpected recombination rate and pattern seen in NLTR21. Perhaps something about this structure leads to meiotic recombination initiated by SPO11 cleavage in DNA flanking NLTR21, with some bias towards resolving these events through Holliday structure resolution within the NLTR21 LTRs 5' end.

Solo LTRs have been observed in all investigated species that have LTR retrotransposons and their formation through unequal recombination must play a role in genome size reduction (Shepherd et al., 1984; Sentry and Smyth, 1989; SanMiguel et al., 1996; Noma et al., 1997; Chen et al., 1998; Han et al., 2000). We have observed unequal recombination within 10 out of 50 LTR retrotransposons. We found meiotic recombination rates as high as 1/5600. We only looked in a very limited region of the maize genome where the recombination rates were higher than the genome average (Bruggmann et al. 2006), so future directions could be to expand this project to look at LTR retrotransposons across the genome, once high quality assembly and annotation in

heterochromatic regions makes this possible. Such an analysis would likely identify differences in the de novo rates of unequal recombination at different chromosomal positions, given that accumulation differences have been observed across the genome by solo LTR to intact LTR retrotransposon analysis across the rice genome (Ma and Bennetzen 2006). Increasing the number of studied LTR retrotransposons could also uncover additional factors that influence recombination but not drastically enough to be able to detect with our sample sizes. It would also be interesting to find LTR retrotransposons with similar characteristics to NLTR21 and see if they have a similar recombination pattern to try to understand what is causing the high recombination rates and unusual pattern.

The biggest issue we had during this experiment was designing pairs of primers that each worked individually when tested with the control primers, but also did not produce non-specific products. The maize genome is estimated to be at least 85% repetitive sequence, so most LTR retrotransposons will be inserted into repetitive sequence. To solve this problem, we designed primers to start immediately adjacent to the LTRs and extend into the LTR, spanning the junction. This junction sequence was unique and greatly increased the specificity of our primers. It is likely that if this experiment were repeated in a smaller genome with less repetitive sequence, there would be an even greater primer success rate, allowing more LTR retrotransposons to be studied. This would also reduce the ratio between the solo LTRs we are trying to amplify and the total amount of DNA. This PCR method on pollen DNA could also be expanded to look at meiotic recombination between any two sequences, as long as the end junctions were unique and the recombinant product was not too large to be amplified by PCR.

Table 3.1. Names and source of all LTR retrotransposons studied in this project. List of all LTR retrotransposons we studied, indicating whether they came from the MASiVEdb Sirevirus plant retrotransposon database or the MIPS Repeat Element Database, and the sequence ID in that database.

	Database	**Database ID**
MLTR7	MIPS	Zm9L_4L_WD
MLTR80	MIPS	zeon_ac145224-1
NLTR12	MIPS	Zm9L_81L
NLTR13	MASiVEdb	Zmay_Chr_1-P-2673200
NLTR20	MIPS	Zm9L_81L
NLTR21	MASiVEdb	Zmay_Chr_1-D-3538149
NLTR24	MASiVEdb	Zmay_Chr_1-P-4275244
NLTR25	MASiVEdb	Zmay_Chr_1-P-4367568
NLTR27	MASiVEdb	Zmay_Chr_1-P-4822448
NLTR28	MASiVEdb	Zmay_Chr_1-D-4965358
NLTR35	MASiVEdb	Zmay_Chr_1-P-5595616
NLTR36	MASiVEdb	Zmay_Chr_1-D-6639380
NLTR39	MASiVEdb	Zmay_Chr_1-P-6786589
NLTR4	MASiVEdb	Zmay_Chr_1-D-2175066
NLTR43	MASiVEdb	Zmay_Chr_1-D-7181174
NLTR5	MASiVEdb	Zmay_Chr_1-D-2190109
NLTR50	MASiVEdb	Zmay_Chr_1-P-8359358
NLTR58	MIPS	huck_5BL-1
NLTR6	MASiVEdb	Zmay_Chr_1-D-2267727
NLTR8	MIPS	opie_af546189-1
RLTR11	MASiVEdb	Zmay_Chr_9-P-150518637
RLTR19	MASiVEdb	Zmay_Chr_9-D-151736534
RLTR22	MASiVEdb	Zmay_Chr_9-D-149831652
RLTR23	MASiVEdb	Zmay_Chr_9-P-156357636
RLTR27	MASiVEdb	Zmay_Chr_9-P-155391998
RLTR29	MASiVEdb	Zmay_Chr_9-P-155251649
RLTR3	MASiVEdb	Zmay_Chr_1-D-3235572
RLTR32	MASiVEdb	Zmay_Chr_9-D-154441284
RLTR33	MIPS	Zm9L_41L
RLTR34	MASiVEdb	Zmay_Chr_9-D-154053450
RLTR35	MASiVEdb	Zmay_Chr_9-P-153862016
RLTR36	MASiVEdb	Zmay_Chr_9-D-153797812
RLTR39	MASiVEdb	Zmay_Chr_9-P-153088838
RLTR4	MIPS	Zm1S_124L

RLTR41	MASiVEdb	Zmay_Chr_9-P-153001352
RLTR43	MASiVEdb	Zmay_Chr_9-D-152499324
RLTR44	MASiVEdb	Zmay_Chr_9-P-152286774
RLTR46	MASiVEdb	Zmay_Chr_9-D-152030984
RLTR5	MASiVEdb	Zmay_Chr_9-D-149765654
RLTR6	MASiVEdb	Zmay_Chr_9-D-149930281
RLTR8	MASiVEdb	Zmay_Chr_9-D-150360950
RLTR9	MASiVEdb	Zmay_Chr_9-D-150374768
SLTR12	MASiVEdb	Zmay_Chr_9-P-150625161
SLTR15	MASiVEdb	Zmay_Chr_9-D-151222862
SLTR16	MASiVEdb	Zmay_Chr_9-P-151342135
SLTR34	MASiVEdb	Zmay_Chr_9-P-153387751
SLTR39	MASiVEdb	Zmay_Chr_9-P-154354996
SLTR41	MASiVEdb	Zmay_Chr_9-P-154488558
SLTR42	MASiVEdb	Zmay_Chr_9-D-154503974
SLTR54	MASiVEdb	Zmay_Chr_9-D-156441589

Table 3.2. List of unequally recombining LTR retrotransposons and the number of unequal recombination events.

LTR retrotransposon	Unequal recombination events detected
NLTR13	2
NLTR21	3
NLTR24	2
NLTR25	1
NLTR27	1
SLTR39	4
RLTR3	1
RLTR6	4
RLTR32	3
RLTR36	2

Table 3.3. 5’ LTR lengths of each investigated LTR retrotransposon. Table shows the lengths of the LTRs of each LTR retrotransposon sorted by the smallest to the largest 5’ LTR. LTR retrotransposons that unequally recombined are highlighted in yellow.

	Unequal recombination event number	Length 5' LTR (bp)	Length 3' LTR (bp)
RLTR4	0	586	576
MLTR7	0	643	643
MLTR80	0	643	643
RLTR43	0	1188	1161
NLTR20	0	1194	1103
NLTR43	0	1207	1211
NLTR35	0	1209	1183
NLTR21	3	1213	1210
RLTR22	0	1218	1211
NLTR8	0	1219	1273
RLTR27	0	1220	1224
RLTR46	0	1224	1263
RLTR41	0	1228	1246
NLTR24	2	1234	1235
RLTR6	4	1237	1236
NLTR6	0	1239	1258
RLTR32	3	1241	1240
RLTR44	0	1242	1243
SLTR39	4	1251	1251
SLTR12	0	1253	1235
SLTR34	0	1256	1257
RLTR29	0	1257	1218
NLTR36	0	1258	1257
RLTR3	1	1260	1259
RLTR9	0	1264	1204
SLTR16	0	1264	1259
SLTR42	0	1271	1251
NLTR39	0	1271	1271
NLTR50	0	1275	1287
NLTR5	0	1277	1301
NLTR4	0	1289	1307
RLTR23	0	1291	1274
RLTR5	0	1294	1332
RLTR34	0	1308	1327

RLTR36	2	1308	1327
SLTR41	0	1308	1351
NLTR13	2	1314	1333
RLTR11	0	1319	1362
SLTR15	0	1321	1363
NLTR27	1	1323	1323
NLTR25	1	1323	1345
RLTR35	0	1325	1326
SLTR54	0	1325	1344
NLTR28	0	1333	1333
RLTR39	0	1352	1313
NLTR12	0	1396	1398
RLTR19	0	1413	1358
NLTR58	0	1617	1617
RLTR8	0	1686	1646
RLTR33	0	1952	1949
Median		1264	1261
Mean		1262.78	1262.74

Table 3.4. 3’ LTR lengths of each investigated LTR retrotransposon. Table shows the lengths of the LTRs of each LTR retrotransposon sorted by the smallest to the largest 3’ LTR. LTR retrotransposons that unequally recombined are highlighted in yellow.

	Unequal recombination event number	Length 3' LTR (bp)	Length 5' LTR (bp)
RLTR4	0	576	586
MLTR7	0	643	643
MLTR80	0	643	643
NLTR20	0	1103	1194
RLTR43	0	1161	1188
NLTR35	0	1183	1209
RLTR9	0	1204	1264
NLTR21	3	1210	1213
NLTR43	0	1211	1207
RLTR22	0	1211	1218
RLTR29	0	1218	1257
RLTR27	0	1224	1220
NLTR24	2	1235	1234
SLTR12	0	1235	1253
RLTR6	4	1236	1237
RLTR32	3	1240	1241
RLTR44	0	1243	1242
RLTR41	0	1246	1228
SLTR39	4	1251	1251
SLTR42	0	1251	1271
SLTR34	0	1257	1256
NLTR36	0	1257	1258
NLTR6	0	1258	1239
RLTR3	1	1259	1260
SLTR16	0	1259	1264
RLTR46	0	1263	1224
NLTR39	0	1271	1271
NLTR8	0	1273	1219
RLTR23	0	1274	1291
NLTR50	0	1287	1275
NLTR5	0	1301	1277
NLTR4	0	1307	1289
RLTR39	0	1313	1352
NLTR27	1	1323	1323

RLTR35	0	1326	1325
RLTR34	0	1327	1308
RLTR36	2	1327	1308
RLTR5	0	1332	1294
NLTR13	2	1333	1314
NLTR28	0	1333	1333
SLTR54	0	1344	1325
NLTR25	1	1345	1323
SLTR41	0	1351	1308
RLTR19	0	1358	1413
RLTR11	0	1362	1319
SLTR15	0	1363	1321
NLTR12	0	1398	1396
NLTR58	0	1617	1617
RLTR8	0	1646	1686
RLTR33	0	1949	1952
Median		1261	1264
Mean		1262.74	1262.78

Table 3.5. Pairwise identity between the LTRs for each LTR retrotransposon studied. Table shows the % pairwise identity between the two LTRs for each LTR retrotransposon. LTR retrotransposons that unequally recombined are highlighted in yellow.

	Unequal recombination event number	Pairwise % Identity
RLTR19	0	86.9
NLTR20	0	88.7
NLTR21	3	90.8
NLTR35	0	91.3
NLTR8	0	92.6
RLTR11	0	93
RLTR43	0	93.9
RLTR29	0	94.1
NLTR43	0	94.2
RLTR9	0	94.9
SLTR15	0	94.9
SLTR41	0	95.2
RLTR27	0	95.4
NLTR50	0	95.5
NLTR5	0	96.1
RLTR46	0	96.2
RLTR39	0	96.3
RLTR5	0	96.5
RLTR23	0	96.6
NLTR6	0	96.7
NLTR58	0	96.7
RLTR8	0	96.7
NLTR39	0	96.9
RLTR34	0	97.1
NLTR4	0	97.4
RLTR41	0	97.7
RLTR4	0	97.8
RLTR22	0	97.8
RLTR36	2	98
NLTR13	2	98
SLTR54	0	98.1
NLTR25	1	98.1
SLTR42	0	98.3
SLTR12	0	98.5

NLTR12	0	98.6
RLTR3	1	99
SLTR16	0	99
NLTR36	0	99.1
MLTR7	0	99.2
SLTR34	0	99.2
NLTR28	0	99.3
RLTR33	0	99.3
NLTR24	2	99.4
RLTR6	4	99.4
RLTR44	0	99.4
MLTR80	0	99.5
RLTR32	3	99.7
SLTR39	4	99.8
NLTR27	1	99.9
RLTR35	0	100
Median		97.55
Mean		96.734

Table 3.6. Chromosome and distance to the nearest telomere for each LTR retrotransposon studied. Table shows the chromosome number and the distance from the telomere for each LTR retrotransposon. LTR retrotransposons that unequally recombined are highlighted in yellow.

	Unequal recombination event number	Chromosome	Distance from telomere
SLTR54	0	Chr9	309,117
RLTR23	0	Chr9	393,070
RLTR27	0	Chr9	1,358,697
RLTR29	0	Chr9	1,499,909
NLTR4	0	Chr1	2,172,937
NLTR5	0	Chr1	2,187,980
SLTR42	0	Chr9	2,247,184
SLTR41	0	Chr9	2,262,600
NLTR6	0	Chr1	2,265,598
RLTR32	3	Chr9	2,309,874
SLTR39	4	Chr9	2396,162
NLTR8	0	Chr1	2,403,460
RLTR33	0	Chr9	2,491,608
MLTR80	0	Chr9	2,538,154
NLTR12	0	Chr1	2,646,744
NLTR13	2	Chr1	2,671,071
RLTR34	0	Chr9	2,697,708
RLTR35	0	Chr9	2,889,142
NLTR58	0	Chr1	2,905,696
RLTR36	2	Chr9	2,953,346
RLTR3	1	Chr1	3,228,422
SLTR34	0	Chr9	3,363,407
NLTR20	0	Chr1	3,510,665
NLTR21	3	Chr1	3,531,805
RLTR39	0	Chr9	3,665,079
RLTR41	0	Chr9	3,752,565
RLTR43	0	Chr9	4,255,446
NLTR24	2	Chr1	4,268,900
NLTR25	1	Chr1	4,361,224
RLTR44	0	Chr9	4,467,996
RLTR46	0	Chr9	4,723,786
NLTR27	1	Chr1	4,816,104
NLTR28	0	Chr1	4,959,014
RLTR19	0	Chr9	5,018,136

SLTR16	0	Chr9	5,412,535
SLTR15	0	Chr9	5,531,808
NLTR35	0	Chr1	5,589,272
SLTR12	0	Chr9	6,129,409
RLTR11	0	Chr9	6,235,934
RLTR9	0	Chr9	6,379,805
RLTR8	0	Chr9	6,393,620
NLTR36	0	Chr1	6,633,036
MLTR7	0	Chr9	6,779,644
NLTR39	0	Chr1	6,780,245
RLTR6	4	Chr9	6,824,287
RLTR22	0	Chr9	6,922,918
RLTR5	0	Chr9	6,988,916
NLTR43	0	Chr1	7,174,730
RLTR4	0	Chr1	8,320,893
NLTR50	0	Chr1	8,354,575
Median			3,708,822
Mean			4,159,484.66

Table 3.7. The distance to the nearest gene, the nearest distal gene, the nearest proximal gene, and the average of the distances. Table shows the distance to the nearest gene, to the nearest distal gene upstream of the 5' LTR, the distance to the nearest proximal gene downstream of the 3' LTR, and the average distance between the nearest distal and proximal gene. The table is sorted by distance to the nearest gene starting with the smallest difference. RLTR4 and RLTR43 are annotated as being inside the introns of genes. LTR retrotransposons that unequally recombined are highlighted in yellow.

	Unequal recomb event number	Nearest Gene	Nearest distal gene	Nearest proximal gene	Average distance of nearest genes
RLTR4	0	0	0	0	0
RLTR43	0	0	0	0	0
RLTR46	0	417	417	2636	1526.5
NLTR8	0	547	547	1621	1084
MLTR7	0	952	952	1852	1402
NLTR21	3	1023	67528	1023	34275.5
NLTR58	0	1612	11443	1612	6527.5
RLTR39	0	1786	1786	45336	1786
NLTR39	0	1787	13184	1787	13184
RLTR32	3	1944	1944	3430	2687
SLTR39	4	2147	13748	2147	7947.5
NLTR12	0	2392	104189	2392	53290.5
RLTR22	0	2421	2421	53591	28006
NLTR43	0	2441	12650	2441	7545.5
NLTR13	2	2601	2601	13001	2601
RLTR19	0	2743	2743	14110	8426.5
SLTR41	0	2806	42286	2806	42286
SLTR54	0	2909	2909	121401	62155
NLTR50	0	3010	50527	3010	50527
NLTR24	2	3913	3913	22613	3913
RLTR35	0	4429	11181	4429	11181
RLTR36	2	4927	13095	4927	9011
NLTR35	0	5082	21172	5082	21172

NLTR4	0	6395	6395	23388	14891.5
NLTR28	0	6445	87732	6445	47088.5
RLTR8	0	6524	6524	46235	26379.5
NLTR5	0	8447	21438	8447	14942.5
RLTR44	0	8484	8484	14940	8484
RLTR33	0	8777	32672	8777	32672
NLTR25	1	9076	61321	9076	61321
RLTR11	0	9728	116003	9728	116003
RLTR5	0	9837	13833	9837	11835
SLTR12	0	10008	116213	10008	63110.5
SLTR16	0	10704	11973	10704	11338.5
SLTR15	0	14712	65157	14712	39934.5
SLTR34	0	15205	51642	15205	33423.5
RLTR27	0	15784	35785	15784	35785
RLTR29	0	17007	17007	60508	17007
SLTR42	0	18222	18222	27354	22788
RLTR34	0	18857	27363	18857	23110
RLTR9	0	20339	20339	35310	27824.5
MLTR80	0	21953	113612	21953	67782.5
NLTR20	0	22734	22734	46388	22734
RLTR6	4	23058	33625	23058	28341.5
RLTR3	1	24581	34451	24581	29516
RLTR41	0	25473	25473	33441	25473
NLTR6	0	28837	41813	28837	35325
NLTR27	1	30077	30077	50300	30077
RLTR23	0	39413	39413	81055	39413
NLTR36	0	59475	59475	132712	96093.5
Median		6484.5	19280.5	11852.5	22949
Mean		10840.82	30000.24	21977.74	27064.57

Table 3.8. The percentage CG methylation for each LTR retrotransposon. Table shows the percentage of CG methylation per for each LTR retrotransposon. LTR retrotransposons that unequally recombined are highlighted in yellow.

Name	Unequal Recombination events	% methylated CG
SLTR41	0	1.9
RLTR46	0	3.1
NLTR35	0	3.7
RLTR9	0	3.9
NLTR21	3	8.2
NLTR50	0	12.0
RLTR27	0	27.6
SLTR42	0	28.1
RLTR3	1	47.8
SLTR34	0	50.2
RLTR19	0	53.6
NLTR43	0	53.7
NLTR36	2	54.8
NLTR28	0	60.3
RLTR35	0	62.3
MLTR80	0	68.4
NLTR8	0	78.6
SLTR54	0	82.1
RLTR8	0	82.8
NLTR39	0	85.2
RLTR29	0	86.7
NLTR20	0	86.7
RLTR34	0	87.4
SLTR15	0	88.2
NLTR27	1	88.5
RLTR41	0	88.6
RLTR6	4	90.3
NLTR58	0	90.4
RLTR4	0	90.9
RLTR36	0	91.1
RLTR5	0	91.7
SLTR16	0	91.8

NLTR5	0	92.0
RLTR32	3	92.6
RLTR39	0	92.8
RLTR23	0	93.0
NLTR12	0	93.6
NLTR24	2	94.0
NLTR4	0	94.5
SLTR39	4	94.8
RLTR22	0	94.9
NLTR13	2	94.9
NLTR6	0	96.3
NLTR25	1	98.5
MLTR7	0	98.6
RLTR11	0	99.4
SLTR12	0	100.0
RLTR43	0	100.0
Median		88.3
Mean		72.3

Table 3.9. The percentage CHG methylation for each LTR retrotransposon. Table shows the percentage of CHG methylation per for each LTR retrotransposon. LTR retrotransposons that unequally recombined are highlighted in yellow.

Name	Unequal Recombination events	Percent methylated CHG
RLTR3	1	2.0
SLTR41	0	2.4
NLTR35	0	5.0
RLTR9	0	5.8
RLTR46	0	6.3
NLTR21	3	11.7
RLTR27	0	13.2
NLTR50	0	25.0
RLTR34	0	30.2
SLTR42	0	34.3
SLTR34	0	38.7
RLTR35	0	40.9
SLTR54	0	43.6
NLTR36	2	44.1
RLTR6	4	48.4
NLTR8	0	53.4
RLTR19	0	55.2
RLTR8	0	56.4
NLTR28	0	59.3
RLTR43	0	60.0
NLTR43	0	61.9
MLTR80	0	67.9
MLTR7	0	69.7
RLTR11	0	70.6
NLTR39	0	73.1
RLTR4	0	73.2
SLTR15	0	75.3
RLTR23	0	75.7
NLTR20	0	75.8
NLTR58	0	79.5
NLTR13	2	80.3
RLTR32	3	80.8
RLTR36	0	80.9

RLTR5	0	82.4
RLTR41	0	83.4
NLTR12	0	83.4
SLTR16	0	83.9
NLTR27	1	84.0
NLTR24	2	84.4
RLTR39	0	84.4
NLTR6	0	85.1
NLTR5	0	85.7
RLTR22	0	86.9
NLTR4	0	88.0
SLTR39	4	88.2
RLTR29	0	90.0
NLTR25	1	90.4
SLTR12	0	100.0
Median		71.8
Mean		60.4

Table 3.10. The percentage CHH methylation for each LTR retrotransposon. Table shows the percentage of CHH methylation per for each LTR retrotransposon. LTR retrotransposons that unequally recombined are highlighted in yellow.

Name	Unequal Recombination events	% methylated CHH
RLTR43	0	0.0
NLTR50	0	0.4
RLTR23	0	0.4
RLTR3	1	0.4
NLTR25	1	0.6
RLTR4	0	0.7
RLTR34	0	0.7
SLTR54	0	0.7
MLTR7	0	1.1
SLTR39	4	1.3
SLTR15	0	1.3
SLTR41	0	1.6
NLTR20	0	2.4
NLTR58	0	2.6
NLTR13	2	3.0
NLTR5	0	3.1
NLTR27	1	3.1
RLTR29	0	3.2
MLTR80	0	3.3
NLTR6	0	3.4
RLTR27	0	3.7
RLTR46	0	3.8
NLTR8	0	3.8
NLTR35	0	3.8
NLTR43	0	4.0
NLTR24	2	4.1
RLTR6	4	4.3
RLTR41	0	4.5
RLTR39	0	4.6
NLTR4	0	5.3
RLTR32	3	5.9
SLTR16	0	6.1

SLTR34	0	6.5
RLTR9	0	6.6
RLTR36	0	6.8
RLTR22	0	6.9
NLTR39	0	9.0
NLTR21	3	9.7
NLTR12	0	11.3
RLTR5	0	11.3
RLTR8	0	13.7
SLTR42	0	16.1
NLTR36	2	19.3
NLTR28	0	19.6
RLTR11	0	20.7
RLTR35	0	22.8
RLTR19	0	36.5
SLTR12	0	40.0
Median		3.9
Mean		7.2

Table 3.11. P-values for the correlation of recombination and each studied characteristic. For each characteristic (LTR length, LTR sequence homology, distance of the retrotransposon from the chromosome end, distance to the nearest gene, average of the distances of the retrotransposon to the nearest proximal and distal gene, and the amount of CG, CHG, and CHH methylation), the table shows the number of unequal recombination (UR) below and above the median, the p-value calculated from the sorting the retrotransposons by median, the number of unequally recombining elements below and above the mean, the total number of elements below and above the mean, and the p-value calculated by sorting the retrotransposons my mean. Significant p-values are indicated with *.

	Unequally recombining elements below median	Unequally recombining elements above median	Median p-value	Unequally recombining elements below mean	Unequally recombining elements above mean	Elements below mean	Elements above mean	Mean p-value
LTR length	6	4	0.48	6	4	24	26	0.48
Sequence homology	1	9	0.0047*	1	9	22	28	.015*
Distance from end	6	4	0.48	6	4	24	26	0.4
Nearest gene distance	6	4	0.48	7	3	34	16	0.88
Average gene distance	5	5	1	5	5	28	22	0.18
CG methylation	3	7	0.16	3	7	16	32	0.8
CHG methylation	4	6	0.48	4	6	20	28	0.9
CHH methylation	5	5	1	8	2	33	15	0.38

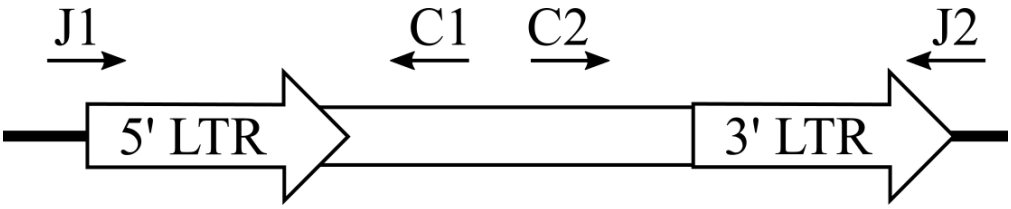


Figure 3.1. Primers used to study LTR unequal recombination. An example of a retrotransposon is shown above with the LTRs labeled and the adjacent sequence indicated by a black line. The two junction primers used to identify recombination events are labeled as J1 and J2 and span the junction between the LTRs and the sequence adjacent to the retrotransposon. The two control primers are labeled as C1 and C2 and were used to test the junction primers.

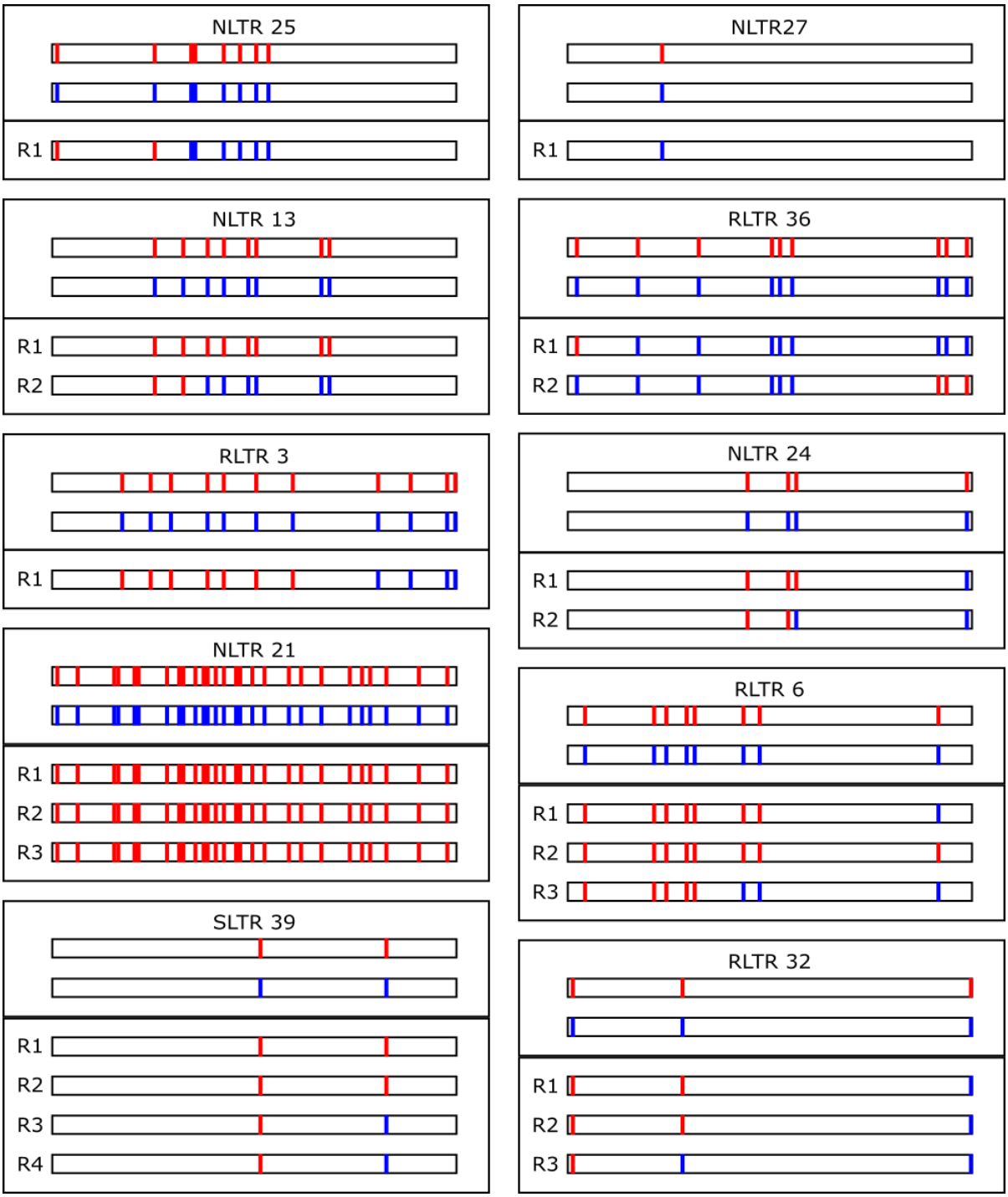


Figure 3.2. Solo LTR outcomes from unequally recombined LTRs. Each separate box shows the 5' and 3' LTRs and all recombinants for each of the 10 unequally recombined LTR retrotransposons. The top bar in the top section of each box is the 5' LTR, while the second bar is the 3' LTR. Red and blue bands are variations between each pair of LTRs. The bottom section

of each box shows all unequal recombinant products (Labeled R1, R2, etc…). Red bands indicate the variation at that point originated from the 5’ LTR, while blue bands indicate the variation that originated from the 3’ LTR.

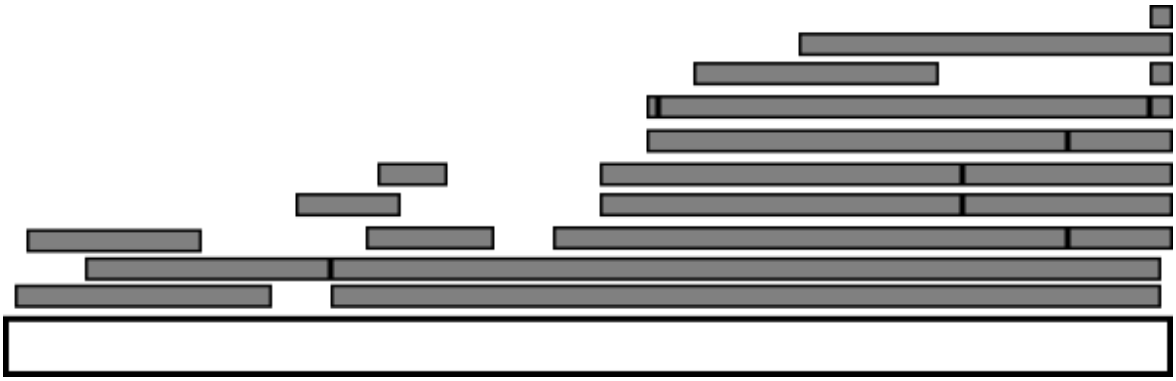


Figure 3.3. Normalized LTR showing all unequal recombination locations. The white bar shows the normalized LTR of all unequally recombining LTRs in this study. The gray boxes above the white bar show the range where each CO took place.

Figure 3.4. NLTR21 and the surrounding region. NLTR21 is annotated in yellow with the 5' and 3' LTRs annotated in light blue. The truncated LTR retrotransposon is shown in green, with the LTR annotated in dark blue and the *pol* gene annotated in orange. The predicted gene GRMZM2G703469 is annotated in red.

CHAPTER 4

GENE LOSS AND GENOME REARRANGEMENT IN THE PLASTIDS OF FIVE HEMIPARASITES IN THE FAMILY OROBANCHACEAE[3]

[3] Frailey, D.C., Chaluvadi, S.R., Vaughn, J.N., Coatney, C.G., Bennetzen, J.L. Submitted to *BMC Plant Biology*, 10/3/2017.

Abstract

The chloroplast genomes (plastome) of most plants are highly conserved in structure, gene content, and gene order. Parasitic plants, including those that are fully photosynthetic, often contain plastome rearrangements. These most notably include gene deletions that result in a smaller plastome size. The nature of gene loss and genome structural rearrangement has not been extensively studied in parasitic plants, so their role in the adaptation of these parasites is not fully understood.

De novo sequencing, assembly and annotation of the chloroplast genomes of five photosynthetic parasites from the family Orobanchaceae were employed to investigate plastome dynamics. Four had major structural rearrangements, including gene duplications and gene losses, that differentiated the taxa. The facultative parasite *Aureolaria virginica* had the most similar genome content to its close non-parasitic relative, *Lindenbergia philippensis*, with similar genome size and organization, and no differences in gene content. In contrast, the facultative parasite *Buchnera americana* and three obligate parasites in the genus *Striga* all had enlargements of their plastomes, primarily caused by expansion within the large inverted repeats (IRs) that are a standard plastome feature. Some of these IR increases were shared by multiple investigated species, but others were unique to particular lineages. Gene deletions and pseudogenization were also both shared and lineage-specific, with the particularly frequent loss of the *ndh* genes involved in electron recycling.

Five new plastid genomes were fully assembled and compared. The results indicate that plastome instability is common in parasitic plants, even those that retain the need to perform essential plastid functions like photosynthesis. Gene losses were slow and not identical across taxa, suggesting that different lineages had different uses or needs for some of their plastome

gene content, including genes involved in some aspects of photosynthesis. Recent repeat region extensions, some unique to terminal species branches, were observed after the divergence of the *Buchnera/Striga* clade, suggesting that this otherwise rare event has some special value for gene duplication in this lineage.

Keywords: Chloroplast, Chromosome rearrangement, Gene deletion, Parasite, Plastome, *Striga*

Introduction

Beyond its essential roles in photosynthesis, the chloroplast is also the specialized site for synthesis of some pigments, lipids, amino acids, and sulfur compounds (Palmer 1985, Wicke et al. 2011). The chloroplast contains its own genome, the plastome. The plastome in vascular plants is believed to have a single endocytotic origin, more than a billion years ago, from a cyanobacterium-like organism (Palmer 2000, McFadden and van Dooren 2004, Keeling 2010). Over time, most genes of the cyanobacterium have either been lost or have migrated to the nuclear and/or mitochondrial genomes, so that modern plant plastomes contain approximately 10% of the ancestral endosymbiont genes (Martin et al. 2002).

Plastomes are under strong selective pressure in most plants and therefore tend to be highly conserved in terms of gene content and order, nucleotide substitution rates, structure, and size (Raubeson and Jansen 2005). Most plastomes range between 120 to 160 kb in size (Palmer 1985). Originally, plastomes were thought to be arranged as single circular molecules. Although a percent of plastomes do have this structure, they are also present in dimers and higher order concatemers, in both circular and linear configurations (Wicke et al. 2011). Plastomes have a

quadripartite structure consisting of a large single copy (LSC) region and a small single copy region (SSC) separated by two virtually identical inverted repeat (IR_A and IR_B) regions (Kolodner and Tewari 1979). Recombination between the two inverted repeats plays a role in stabilizing the plastome (Maréchal et al. 2009). In angiosperms, LSCs are typically 80-90 kb, SSCs 16-27 kb, and inverted repeats 20-30 kb (Wicke et al. 2011). The plastome contains genes encoding proteins involved in photosynthesis as well as the other biochemical pathways carried out in the chloroplast. It also encodes 30 tRNAs and several other structural rRNAs (Palmer 1991, Sugiura 1992, Bock 2007).

Unlike most species of plants, parasitic plants often contain highly divergent plastomes (dePamphilis and Palmer 1990, Wolfe et al. 1992b, Nickrent et al. 1997, Funk et al. 2007, McNeal et al. 2007, Krause 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016). Non-photosynthetic plants that obtain all their nutrients from the host plant are called holoparasites, while hemiparasites are photosynthetic plants that obtain some nutrients from the host and the rest from photosynthesis. Hemiparasites can be further broken down into obligate parasites that require a host plant during at least part of their life cycle or facultative parasites that are capable of completing their entire life cycle without a host plant, but can parasitize if a host is available. Parasitic plant plastomes can have lower selective pressure due to less reliance on photosynthesis. This can lead to higher nucleotide substitution rates and rearrangements. It can also lead to functional gene loss through pseudogenization or physical loss through gene deletion, resulting in a smaller plastome size (Wolfe et al. 1992b, Delavault et al. 1996, Funk et al. 2007, McNeal et al. 2007, Wickett et al. 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016). In some cases, an entire copy of one of the inverted repeats has been lost (Downie and Palmer 1992, Wicke et al. 2013, Wicke et al. 2016). Since the repeats play a

stabilizing role in plastome structure, this leads to further destabilization and rearrangements in the plastome (Perry and Wolfe 2002).

The Orobanchaceae are the largest family of parasitic plants. They contain a single non-parasitic genus, *Lindenbergia* (Olmstead et al. 2001), in addition to holoparasites and both facultative and obligate hemiparasites (McNeal et al. 2013). In this study, we looked at two species of facultative hemiparasites, *Aureolaria virginica* and *Buchnera americana*, and three species of obligate hemiparasites all in the genus *Striga*: *S. hermonthica, S. forbesii*, and *S. aspera*. The plastome of the autotroph *Lindenbergia philippensis* had already been fully sequenced and annotated by Wicke et al. 2013, so it provided a valuable outgroup to compare to our parasitic species. Despite the fact that all of our studied species rely on photosynthesis, we discovered numerous cases of gene loss and other forms of rearrangement in their plastomes.

Methods

Sample collection, DNA extraction and DNA sequencing

DNA preparations from leaf tissue of *Striga* species were the same as those previously described (Estep et al. 2012). *B. americana* and *A. virginica* leaf tissues were collected from live specimens in the Plant Biology Teaching collection at the University of Georgia. High molecular weight DNAs were isolated from these samples by a modified cetyltrimethylammonium bromide procedure (Doyle and Doyle 1987) and were used for Illumina library preparation with the NEBNext® Ultra™ II DNA Library Prep Kit as per manufacturer's instructions. Illumina MiSeq PE250 runs were performed at the Georgia Genomics Facility.

Plastome assembly

Plastid-homologous sequences were selected from the total cellular DNA reads of *Striga hermonthica* and then assembled with a *de novo* assembly pipeline, both as previously described (Vaughn et al. 2014), to create ten contigs. To fill the ten gaps in this assembly, primers were designed for the end of each contig and PCR was set up for each primer pair. Each PCR reaction consisted of 5-20 ng template DNA, 10 μL Q5 Reaction Buffer, 10 μL betaine, 2.5 μL DMSO, 1 μL 10mM dNTPs, 2.5 μL forward primer, 2.5 μL reverse primer, .5 μL Q5 High-Fidelity DNA Polymerase, and 20 μL water. Touchdown PCR was performed on each reaction on an MJ Research PTC-200 Peltier Thermal Cycler. The program used was 98°C for 30 sec; 10 cycles consisting of denaturation at 98°C for 20 sec, annealing at 68°C - 1°C each cycle for 15 sec, elongation at 72°C for 1 min; 25 cycles consisting of denaturation at 98°C for 15 sec, 58°C for 15 sec, elongation at 72°C for 1 min; and a final elongation step at 72°C for 10 min. PCR products were sent to Macrogen USA (Rockville, MD) for Sanger sequencing. Sequences were manually assembled to the original contig ends in order to fill gaps.

The assembled *S. hermonthica* plastome was aligned to the *Lindenbergia philippensis* plastome using Mauve 2.3.1 (Darling et al. 2014). All junctions to predicted rearrangements in the *S. hermonthica* plastome were PCR amplified and Sanger sequenced (Macrogen USA) to confirm that the assembly was correct.

Reads for *S. forbessi, S. aspera, A. virginica,* and *B. americana* were mapped against the *S. hermonthica* plastome. All reads that mapped to the plastome were assembled *de novo* using Geneious 8.1.6 (Kearse et al. 2012). All reads were then mapped to the assembled contigs to complete the assembly.

The repeat region of each assembly was identified by the region of the plastome having approximately two times the read coverage compared to the rest of the plastome. In the case of *S. hermonthica*, we ran PCR on all the junctions between the repeats and the LSC and SSC. Sequencing of these PCR products confirmed the assembly of these junctions.

Plastome annotation and alignment

The genes from the annotated *L. philippenis* plastome were downloaded from NCBI along with the plastome genes from all plant species from Chloroplast Genome DB (Cui et al. 2006). We ran BLAST on all genes against each assembled plastome. All genes identified in the plastome were then mapped to the plastome using Geneious. Reads were also mapped to each gene individually to confirm that the sequence assembly was correct and that there were no genes missing in our assemblies that had reads map to them. Individual genes were aligned across all five parasitic species as well as *L. philippensis* using Clustal W (Larkin et al. 2007) to identify mutations, insertions, and deletions. Genes that contained one or more frameshift mutations or premature stop codons were considered potential pseudogenes.

The plastomes were aligned using Mauve 2.3.1. The phylogenetic tree was generated with sequence data from all genes shared by all five species. These genes were concatenated into a single sequence, which were aligned using Clustal W in Mega 7.02 (Kumar et al. 2016). *L. philippensis* was included as an outgroup. Bayesian phylogenetic analyses were done on MrBayes 3.2.1 (Ronquist and Huelsenback 2003) using Generalized Time Reversible + Γ + I model of evolution. We ran an initial 2.0×10^7 trees with sampling every 10,000 generations. The first 5.0×10^6 trees were not included in the final analysis. The remaining trees were used to

generate a consensus tree. The same data were used to generate a phylogenetic tree using MEGA [35], and this generated nearly identical results, so only the MrBayes tree is presented.

Results

Plastome assembly

Illumina sequence analysis of total DNA from leaves provided sufficient plastome-homologous sequence data for full genome assembly of all five studied species. All of the initial contiguous sequences were confirmed in their order, and gaps filled, by PCR and Sanger sequencing of the PCR product. Hence, complete assemblies were obtained, and any predicted rearrangements by comparisons of these species genomes were also confirmed by PCR and Sanger sequencing.

Our assemblies indicated that the plastome of *A. virginica* is 153,547 bp, close in size to the 155,103 bp plastome of *L. philippensis.* The lengths of the LSC, SSC, and IR (84,317 bp, 17,168 bp, and 26,031 bp, respectively) in *A. virginica* were all similar in size to those in *L. philippensis* (85,584 bp, 17,885 bp, and 25,812 bp, respectively). The plastomes from the other four species (*S. hermonthica*, *S. aspera*, S. *forbesii and B. americana)* showed major differences in both total plastome size and the size of each region.

Repeat expansion

The three *Striga* species and *B. americana* all exhibited an increase in plastome size, ranging from 166,596 bp in *B. americana* to 190,233 bp in *S. forbesii*. The increase in plastome size correlates with an increase of the IR lengths, ranging from 43,864 bp in *B. americana* to 63,240 bp in *S. forbesii*. Conversely, the LSC and SSC regions in all four species showed a

reduction in size. SSC lengths ranged from 3,377 bp in *Buchnera* to 11,191 bp in *S. forbesii*, while the LSC lengths ranged from 51,628 bp in *S. hermonthica* to 75,491 bp in *B. americana.* All plastome component sizes are shown in Table 4.1.

In *B. americana*, repeat expansion has occurred in both the LSC and SSC regions. Approximately 8 kb of LSC sequence flanking the repeat in *L. philippensis* is now included in the repeat region in *B. americana*. This region contains the genes *trnH*, *psbA*, *matK*, *trnK*, *rps16*, *trnQ*, and *psbK*. The expansion also occurred into both flanking regions of the SSC. On one end of the SSC, *ndhF* has been lost in *B. americana* while the adjacent 6 kb region is now part of the repeat region. *ndhE*, *psaC*, *ndhD*, *ccsA*, *trnL*, and *rpl32* are all contained within this region. On the other end of the SSC, an approximately 5 kb region containing *rpl32* and *ycf1* has been shifted into the repeat in *B. americana*.

The three *Striga* species show an almost identical pattern of IR expansion, with the majority of expansion occurring in the LSC region. The two LSC-flanking regions as well as an internal region are now part of the repeat in all three *Striga* species. An approximately 12.5 kb region flanking the IR in *L. philippensis* is part of the IR in *Striga.* This region contains trn*H*, *psbA*, *matK*, *trnK*, *rps16*, *trnQ*, *psbK*, *psbI*, *trnS*, *trnG*, *trnR*, and *atpA*. On the other end of the LSC, the IR region has expanded into approximately 18.8 kb of adjacent sequence. This region contains 15 genes: *clpP*, *psbB*, *psbT*, *psbN*, *psbH*, *rpoA*, *rps11*, *rpl36*, *infA*, *rps8*, *rpl14*, *rpl16*, *rps3*, *rpl22*, and *rps19*. The internal region is 5.5 kb and contains *ycf4*, *cemA*, *petA*, *psbJ*, *psbL*, *psbF*, and *psbE*. In *L. philippensis*, there are 61 kb between this region and the IR on one side and 18 kb between this region and the IR on the other side. Approximately 4.5 kb of the SSC region adjacent to the repeat is now part of the repeat region in all three *Striga* species. This region contains the *ycf1* gene. All details of repeat expansion are shown in Table 4.1.

GC content

GC content across the entire plastome ranged from 37.7% in *Buchnera* to 38.4% in *Aureolaria*. Protein-coding genes made up between 56.7% of the plastome in *A. virginica* to 58.3% in *S. hermonthica*. tRNA and structural RNA-coding sequence contributed from 4.8% in *S. forbesii* to 5.9% in *Aureolaria* of these plastomes. GC content in protein-coding genes ranged from 37.8% in *Buchnera* to 38.1% in *A. virginica*. GC content in structural RNAs ranged from 55.2% in *Buchnera* to 55.5% in the *Striga* species.

Overall gene content

No tRNA or rRNA genes have been lost in any of the five species. All of the plastomes encode 30 different tRNAs and 4 different rRNAs. Compared to the outgroup *Lindenbergia*, *Aureolaria* has not lost any protein-coding genes and contains a total of 79 single copy genes and six more that are contained within the repeat region and therefore present in two copies. *Buchnera* has a total of 77 single copy genes, plus 17 in the repeat regions, with 7 single copy genes being possible pseudogenes and 2 genes in the repeats being possible pseudogenes. *S. forbesii* contains 79 single copy genes, plus 35 contained within the repeats, including 13 potential pseudogenes in the single copy region and 1 potential pseudogene in the repeat regions. *S. hermonthica* has 76 single copy protein-coding genes, 35 in the repeats and 9 potential pseudogenes in the single copy regions and 1 potential pseudogene in the repeat regions. *S. aspera* contains 77 single copy genes, plus 35 in the repeat regions, with 11 potential pseudogenes pseudogenes in the repeat region and 1 potential pseudogene in the repeat regions. (Table 4.2).

Gene losses

In *L. philippensis* and *A. virginica*, *ndhA* is a 2.2 kb gene in the SSC containing two exons, 539 bp and 553 bp, separated by a 1.1 kb intron. *S. hermonthica* is missing the second exon of *ndhA*, while *S. forbesii* is missing the first exon. *B. americana* and *S. aspera* have both exons, but have a predicted frameshift mutations and premature stop codons in the first exon. All *ndh* genes are present with the structure of functional genes in *A. virginica*. *B. americana* and the three *Striga* species all have one or more stop codons or frameshifts in *ndhB,* starting at amino acid 446 in *B. americana* and 448 in *S. forbesii*. *S. hermonthica* and *S. aspera* both have a frameshift at amino acid 24, suggesting a shared event in their common ancestor, followed by multiple frameshifts and stop codons. *ndhC* is present in *B. americana* but contains a stop codon at amino acid 82. *S. forbesii* contains *ndhC* with a frameshift starting at amino acid 23 followed by multiple stop codons. *ndhC* is missing from both *S. hermonthica* and *S. aspera*. All four species contain multiple frameshifts and stop codons throughout *ndhD* starting at amino acid 36 in *B. americana,* 25 in *S. forbesii*, 4 in *S. hermonthica* and 40 in *S. aspera*. *S. forbesii* contains a stop codon in *ndhE* at amino acid 101. *B. americana* has a stop codon at amino acid 11. *S. hermonthica* has a stop codon at amino acid 40. *B. americana, S. hermonthica,* and *S. aspera* all contain 5' frameshift mutations, at amino acid 34 in *B. americana* and amino acid 32 in the *Striga* species, but all are independent events. All three species have additional stop codons following the beginning of the frameshift. *ndhF* is missing from *B. americana* and contains a frameshift at amino acid 20 in *S. forbesii* followed by multiple stop codons. *S. hermonthica* and *S. aspera* both contain a stop codon at the third amino acid followed by additional stop codons and frameshifts. *ndhG* contains a frameshift and multiple stop codons starting at amino acid 63 in *B. americana* and amino acid 3 in *S. forbesii*. *ndhG* is missing from *S. hermonthica* and *S.*

aspera. *ndhH* is missing from *B. americana*, *S. hermonthica*, and *S. aspera*. *ndhH* is present in *S. forbesii* but contains frameshifts and multiple stop codons starting at amino acid 11. *ndhI* contains a stop codon in *B. americana* at amino acid 90 and in *S. forbesii* at amino acid 17, followed by a frameshift and additional stop codons. *S. hermonthica* contains a frameshift starting at amino acid 47 and *S. aspera* contains a frameshift starting at amino acid 42, both followed by multiple stop codons. *ndhJ* contains stop codons starting at amino acid 140 in *B. americana*. In *S. forbesii, ndhJ* contains a frameshift at amino acid 71 followed by multiple stop codons. *ndhJ* in *S. hermonthica* contains stop codons starting at amino acid 102. *ndhJ* contains a frameshift in *S. aspera* starting at amino acid 44 followed by stop codons. *ndhK* contains a stop codon at amino acid 120 in *B. americana* and amino acid 14 in *S. forbesii*. Both species contain additional stop codons and *S. forbesii* contains a frameshift. *ndhK* in both *S. hermonthica* and *S. aspera* contain a frameshift at amino acid 4 followed by stop codons. Pertinent frameshifts and stop codons are shown in Figure 4.1. Table 4.3 shows mutations shared by two or more species at the identical amino acid that resulted in frameshifts or stop codons.

The *accD* gene is present in all five species. In *A. virginica*, it appears functional with very few differences in comparison to *L. philippensis*. In the other four species, there are numerous predicted mutations including multiple frameshifts. However, the 3' end of this gene is considerably more conserved than the 5' end of the gene**.** It's possible that *accD* has a conserved protein domain encoded by the 3' end of the gene. *ycf1* and *ycf2* also seem to have multiple frameshift mutations in all species, except *ycf2* in *A. virginica*.

Plastome alignments and phylogenetic analysis

All six plastomes were aligned with Mauve 2.3.1 and are shown in Figure 4.2A. *A. virginica* is largely identical to *L. philippensis*, except that the SSC is in an inverted orientation. All other plastomes contained multiple rearrangements. Within the *Striga* genus, *S. hermonthica* and *S. aspera* have no rearrangements between them, while *S. forbesii* has a single rearrangement within the repeat region. At a higher resolution, Figure 4.2B shows the example of the Mauve alignment of just *S. hermonthica* and *A. virginica.*

The timing of gene loss events

In many cases, a specific nucleotide location of a sequence change that is predicted to inactivate a gene was shared by more than one species. Hence, it is most likely that such events occurred in a common ancestor of the species that share the event. Because these deletions, frameshifts and/or stop codon creations are predicted to inactivate the gene, then it is appropriate to plot the lineage and relative time of inactivation of each gene onto a phylogenetic tree of the compared species. The phylogenetic tree, with *L. philippensis* as an outgroup, is shown in Figure 4.3, with predicted shared gene inactivations indicated as gene names along specific branches.

Discussion

Plastome conservation and change

Plastomes perform common and vital functions in virtually all vascular plants, and thus are very powerful tools for phylogenetic analysis. The parasitic plants that are non-photosynthetic lose strong selection for the many plastome genes involved in photosynthesis, but are expected to retain coding capacity for genes that are involved in other essential chloroplast

roles. In previous studies of non-photosynthetic plants (dePamphilis and Palmer 1990, Wolfe et al. 1992b, Krause 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016), most or all photosynthetic genes are seen to be pseudogenized or lost. This process appears to be caused by random mutation, and is thus thought to be a gradual process that will only be completed at some distant time after the parasitic/non-photosynthetic lifestyle is established. Our observations, akin to those observed by Wicke et al. 2016 in a broader but less detailed survey of parasitic plants, indicate that some chloroplast genes are lost even in fully photosynthetic parasites, suggesting that chloroplast roles other than generation of photosynthate may now be provided by the host. Parasitism has evolved independently at least 12 to 13 times within angiosperms (Westwood et al. 2010). Parasitism within Orobanchaceae is believed to have evolved once and is estimated to have occurred 32-64 mya (Bremer et al. 2004, Wolfe et al. 2005, Naumann et al. 2013). It is interesting that these different plastomes have come to somewhat different final gene contents and that many sequences are still present despite their loss of apparent function. This suggests that the gene loss process is very slow, especially compared to nuclear DNA loss (Wolfe et al. 1987), and raises the possibility that the different adaptations of these different species has led to different plastome gene loss outcomes. For instance, most *ndh* genes were lost early in the lineage leading to the genus *Striga*, but *ndhE*, *ndhH* and *ndhJ* have only been lost in terminal branches, suggesting a possible transitional role for these genes even when the other *ndh* genes were no longer functional. Further analysis of the nature of plastome genome change should provide additional insights into which genetic (and, thus, physiological) functions are lost, in what lineages and at what evolutionary rates and times.

Plastome size

Most plant plastomes range in size from 120 to 160 kb. *L. philippensis* falls within this range. The *A. virginica* plastome is slightly smaller than that of *L. philippensis*, but still falls within the typical range. The LSC, SSC, and IR of *A. virginica* are all similar in size to those of *L. philippensis* and fall within the general size range for angiosperms. *B. americana* and the three *Striga* species, however, all show interesting size difference in both whole plastome length and the lengths of their LSCs, SSCs, and IRs. All have a larger plastome than *L. philippensis* and above the range generally seen in plants. This was unexpected as most parasitic species studied so far have smaller plastome sizes (Wolfe et al. 1992b, Delavault et al. 1996, Funk et al. 2007, McNeal et al. 2007, Wickett et al. 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016). The increase in size can be explained by an increase in the length of the inverted repeats. All five species have an increased inverted repeat size, although *Aureolaria*'s is close to *Lindenbergia* (26,031 bp compared to 25,812 bp) and falls within the typical range of angiosperm inverted repeat size. The other four species have considerably larger repeats above the 20 to 30 kb size generally seen in angiosperms, ranging from 43,864 bp in *Buchnera* to 63,240 bp in *S. forbesii*. Small changes of a few hundred bp in inverted repeat size are common in land plants (Goulding et al. 1996, Wang et al. 2008, Downie and Jansen 2015), but larger changes as seen here are rare **(**Goulding et al. 1996, Wang et al. 2008, Downie and Jansen 2015, Knox and Palmer 1999, Dugas et al. 2015, Zhu et al. 2016). The largest known repeat belongs to *Pelargonium hotorum*, at 75,741 bp (Chumley et al. 2006). Slight IR expansion has been seen in the parasitic plastomes of *Schwalbea americana* and *Cistanche phelypaea*. However, these expansions were much smaller in size, encompassing 2 genes and 3 genes (Wicke et al. 2013), respectively, compared to the expansions we saw in *Striga* and *B. americana*.

The repeat size increase that we observed is due to an expansion into the single copy regions. All five parasitic species have smaller SSC and LSC regions compared to *L. philippensis*. The *Striga* expansions differ dramatically from the expansion in *B. americana.* In *B. americana*, most of the repeat expansion has been into the SSC region. In the three *Striga* species, there has been some repeat expansion into the SSC region, but most of the expansion has been into the LSC region. There are two similarities in the expansion patterns between the *Striga* species and *Buchnera.* The first is the 8 kb in the LSC region adjacent to the IR (spanning from *trn*H to *psb*K). This region is contained within the repeat of all four species, although the *Striga* IRs extend a further 4.5 kb region into this region. It is possible that the initial 8 kb expansion occurred before the species diverged and was followed by an additional 4.5 kb expansion in the *Striga* lineage. The second similarity is the expansion into the SSC that contains the *ycf*1 gene. The *B. americana* IR includes an additional 0.5 kb containing *rpl*32, suggesting that this was an additional expansion after the lineages diverged. Other than these two similarities, the rest of the repeat expansions are unique to either *B. americana* or the *Striga* species and most likely occurred after the two lineages diverged. These several independent repeat expansions indicate either selection for this phenomenon or a mechanistic anomaly in these lineages that makes this outcome unusually likely, compared to other studied plants. Functionally, these increases in the duplicated regions do not appear to have created any new genes, but the newly duplicated genes will have a higher level of potential expression simply due to their higher copy number. Hence, selection for a higher expression level on one or more of the genes newly added to the expanded IR is a possible driving force for this expansion.

S. aspera and *S. hermonthica* show no rearrangements between them or differences between the repeat expansions. There is one major structural difference between the plastomes of

these two and that of *S. forbesii*. Within the repeat, a 2.5 kb region containing *cemA* and *petA* is located in a different part of the region in an inverse orientation. One possibility is that repeat expansion occurred independently in the *S. forbesii* lineage and the shared S. *hermonthica/aspera* lineage after the two lineages diverged. The other possibility is that the repeat expansion occurred before the common ancestor of these three species diverged, followed by lineage-specific rearrangement(s) after divergence.

Missing genes

The NAD(P)H-dehydrogenase complex is involved in one of the multiple pathways that recycles electrons around photosystem I (Casano et al. 2000, Nixon 2000). The NAD(P)H complex functions under stressful conditions and may be essential for photosynthesis under conditions of highly variable light intensities (Martín et al. 2009, Martín and Sabater 2010). Eleven subunits are encoded by the plastome: *ndhA, B, C, D, E, F, G, H, I, J,* and *K* (Friedrich et al. 1995). One or more *ndh* genes have been lost several times in land plants and many of them are commonly missing in parasitic plants (McCoy et al. 2008, Wu et al. 2009, McNeal et al. 2007, Funk et al. 2007, dePamphilis and Palmer 1990, Chang et al. 2006, Wu et al. 2010, Wicke et al. 2013, Wicke et al. 2016, Wickett et al. 2008, Barret et al. 2014).

*Ycf*1 was the first plastid-encoded protein identified whose presence was shown to be essential for the survival of green plants. *Ycf*1 protein is an essential component of the translocon that is responsible for chloroplast protein transport in green plants (de Vries et al. 2015). However, no *ycf*1 homolog was detected among Poales, except *Typha latifolia* (Guisinger et al. 2010). In our five studied species, *ycf*1 was observed to be highly divergent, with predicted frameshift mutations. The gene *ycf*2 still has an unknown function (Logacheva et al. 2016) and

has a functional structure in *A. virginica*, but contains multiple frameshift mutations in *B. americana* and the three *Striga* species. Both *ycf*1 and *ycf*2 have high substitution rates in most of the land plant, including non-parasites, and may have become pseudogenes in some lineages (Oliver et al. 2010, Wolfe et al. 2010). The 5' ends of both genes tend to be relatively conserved, while the remaining portions of the genes are more divergent. We observed a similar pattern with the most conserved regions of both genes being in the 5' end. Complete losses of both *ycf*1 and *ycf*2 have occurred in some monocots (Downie et al. 1994).

AccD encodes the beta subunit of the multimeric acetyl-CoA carboxylase, which mediates the conversion of acetyl-CoA to malonyl-CoA during fatty acid synthesis. The *accD* gene has been lost several times in angiosperms (Jansen et al. 2007), where its function is taken over by nuclear copies (Nakkaew et al. 2008). The 3' region of the *acc*D gene in all five of our studied species is considerably more conserved than the 5' end. Higher divergence or truncation of the 5' end of the *acc*D gene also has been observed in several species (Wicke et al. 2013, Samigullin et al. 2016). We do not have a nuclear genome sequence for any of the species investigated in this manuscript, but we predict that a nuclear *accD* homologue is now functional in the *B. americana* and *Striga* lineages. We also predict that the 3' end of the *accD* gene may have a separate, plastid-specific, function.

Plastome alignments

A. virginica has no major genic rearrangements compared to *L. philippensis* except an inverted SSC region. The SSC region is frequently flipped in plastomes, with opposite SSC orientations often present in a 50:50 ratio in plant cells (Palmer 1985). Comparisons between *A. virginica* and *B. americana* are interesting as they are both facultative hemiparasites. Because

they are both capable of living without a host, thereby depending heavily on photosynthesis, it is expected that they would have a relatively small number of rearrangements compared to autotrophic plant plastomes. This is the case with *A. virginica.* However, *B. americana* has multiple rearrangements compared to *A. virginica* and *L. philippensis*. The three *Striga* species also have several rearrangements compared to *A. virginica* and *L. philippensis*, as well as compared to *B. americana*. There are no rearrangements between *S. hermonthica* and *S. aspera*, and a single rearrangement between these two and *S. forbesii*. These patterns all fall within an appropriate phylogenetic order that predicts their time and lineage of origin.

The phylogenetic tree agrees with existing phylogenetic analysis as far as the relationship between *L. philippensis*, *A. virginica*, *B. americana*, and the *Striga* genus (Bennet and Mathews 2006), as well as the relationship of the three *Striga* species to each other with *S. hermonthica* and *S. aspera* being more closely related to each other than to *S. forbesii* (Estep et al. 2012).

While we were assembling these plastomes, an *S. hermonthica* plastome assembly was published in Wicke et al. 2016. We compared our assembly to theirs and found very few differences, except the inverted repeats. Their assembly was missing one of the inverted repeat copies, while ours has both copies. We took several steps to confirm the presence of both copies in our assembly. We took out one of the copies of the repeat and mapped the reads to this sequence (containing the LSC, one repeat, and the SSC). The repeat had approximately 2x coverage compared to the two single copy regions. We also PCR amplified and sequenced the junctions between the repeat and the single copy regions. If there were two copies of the repeat, then there should be two IR-LSC junctions and two IR-SC junctions. All four of these junctions produced a band when amplified with PCR and the sequence confirmed our assembly. If there was only one copy of the repeat, then there should be a single IR-LSC junction, one IR-SSC

junction, and the LSC-SSC junction. We attempted PCR with multiple primer pairs to try to amplify the LSC-SSC junction, but none produced an amplification product. Each of the individual primers worked, because they produced a product when paired with a primer contained in the IR. Our recent discussions with S. Wicke (pers. comm.) find an agreement that the *S. hermonthica* she is now studying has two copies of the IR in its plastome. Hence, it is certain that this is a feature of the *S. hermonthica* plastome, but comparison to the first published manuscript (Wicke et al. 2016) suggests that IR copy number might be a variable within this species.

Plastome instability and parasitic plant function

Although all five of the plant species investigated in this study are obligate photosynthetic organisms, plastome instability was found to be extensive in four species. These results are in agreement with previous observations for other parasitic plants (dePamphilis and Palmer 1990, Wolfe et al. 1992b, Nickrent et al. 1997, Funk et al. 2007, McNeal et al. 2007, Krause 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016), and has also been seen more rarely in non-parasites (Palmer 1987, Downie and Palmer 1992, Tsudzuki et al. 1992, Naumann et al. 2013, Plunkett and Downie 2000, Chumley et al. 2006, Daniell et al. 2006, Guisinger et al. 2010, Wolfe et al. 2010, Grewe et al. 2009, Guo et al. 2014). Hence, plastome instability is a source of genomic variation that provides the raw material for natural selection. In any green plant, this selection is expected to lead to overall conservation of gene content, with exceptions for any chloroplast function that can be provided by the environment directly. For parasitic plants, the host plant is a primary environmental contributor, potentially for photosynthate and many of the numerous small molecules that are generated by pathways

localized to the chloroplast. The *ndh* pathway is lost in many lineages, suggesting that its role in photosynthesis is often dispensable. Given the rapid and near-complete loss of this pathway independently in so many lineages, it suggests that *ndh* may actually be selected against in some plant lineages. Future studies that return this pathway to a species from which it has been naturally lost would provide an excellent opportunity to investigate the lineages and conditions under which the *ndh* gene array might benefit or debilitate a plant.

In this same vein, all of the gene changes observed in this study could be investigated for their role in plant fitness. Expression levels of the genes that differ in their copy number across these taxa, for instance due to their inclusion or lack of inclusion in the IRs, could suggest a role for the generation and retention of such rearrangements across many plant lineages. Natural variation for plastome structure should be tested across more Orobanchaceae species, both to acquire better ideas of the lineages and rates of such rearrangements, but also to provide the raw material for comparisons of the functional outcomes of these rearrangements. With the ability to transform chloroplasts now available in many plant species (Svab et al. 1990, Sikdar et al. 1998, Sidorov et al. 1999, Wani et al. 2010), investigations of natural plastome variation and its role in organismal function could be directly compared to experiments that test these gene effects in engineered transgenics.

Conclusions

We sequenced and assembled the chloroplast genomes from five species of hemiparasitic plants from the Orobanchaceae. We compared the assemblies to the available plastome assembly from *Lindenbergia philippensis*, an autotroph from the Orobanchaceae. *Aureolaria virginica* was almost identical to *L. philippensis* in terms of plastome structure, gene content, and gene order.

B. americana, S. forbesii, S. hermonthica, and *S. aspera* all showed a high number of rearrangements and both physical and functional gene loss. The most interesting result was the increase in the plastome size in these four species, since most parasites have plastomes with decreased sizes. The increase was due to the repeat region expanding into the single copy regions, although what effect this has on the plastome is not yet known. *B. americana, S. hermonthica,* and *S. aspera* all had missing *ndh* genes, while these three species and additionally *S. forbesii* had multiple *ndh* genes that are potentially nonfunctional due to predicted frameshift mutations and stop codons. No other genes were missing from any of the species, but *accD, ycf1,* and *ycf2* all had high levels of divergence with frameshifts in *B. americana, S. forbesii, S. hermonthica,* and *S. aspera*, as well as stop codons in *ycf1* and *ycf2* in *S. hermonthica* and *S. aspera*.

Table 4.1. The sizes of components in five assembled Orobanchaceae plastomes. The table shows the size of the IR, how much the IR has expanded compared to *L. philippensis*, the amount of the expanded repeat sequence that came from the LSC and from the SSC, the number of originally single-copy genes now found in the IR, and the number of genes originally found in the LSC and SSC now contained in the IR.

	Plastome Length (bp)	LSC Length (bp)	SSC Length (bp)	IR Length (bp)		Expansion of IR (bp)	IR from LSC (bp)	IR from SSC (bp)	IR Genes from LSC	IR Genes from SSC
A. virginica	153547	84317	17168	26031		-	-	-	-	-
B. americana	166596	75491	3377	43864		18744	8050	10694	7	8
S. forbessii	190233	52563	11191	63240		37218	32681	4536	34	1
S. hermonthica	186418	51628	9884	62453		36799	32299	4500	34	1
S. aspera	185932	51706	10504	61861		36201	31986	4215	34	1

Table 4.2. The number of genes in the assembled Orobanchaceae plastomes. The table shows the total number of genes that are predicted to encode proteins, the number of unique coding genes, the number of genes that are potentially pseudogenes, and the number of types of tRNA and rRNA genes. The most leftward column shows the total number of predicted genes, including pseudogenes, while the number in parentheses is the number of these genes that are duplicated because of their presence in the IR.

	Total Protein-encoding Genes (Duplicated)	Potential Pseudogenes	tRNA	rRNA
A. virginica	91 (12)	0	30	4
B. americana	111 (17)	9	30	4
S. forbessii	149 (70)	14	30	4
S. hermonthica	146 (70)	10	30	4
S. aspera	147 (70)	12	30	4

Table 4.3. List of shared frameshift mutations and stop codons. List of frameshift mutations (FS) and stop codons (SC) that are shared by two or more species and the amino acid position in each species. FS and SC in the same box are the result of the identical nucleotide mutation in each species. Stop codons created by frameshift mutations are not included. *S. forbessii* is missing the first exon of *ndhA*, which contains the shared stop codon at the 20th amino acid in the other three species. *ndhE* does not have any shared mutations, but *B. americana*, *S. hermonthica*, and *S. aspera* have independent insertions at the same aligned amino acid position (35 in *B. americana* and 33 in the two *Striga* species).

Gene	Species	Amino Acid	SC or FS
ndhA	*B. americana*	19	SC
	S. hermonthica	20	SC
	S. aspera	20	SC
ndhB	*S. hermonthica*	24	FS
	S. aspera	24	FS
	S. hermonthica	118	SC + FS
	S. aspera	118	SC + FS
ndhD	*S. hermonthica*	94	FS
	S. aspera	94	FS
	S. hermonthica	116	FS
	S. aspera	116	FS
	S. hermonthica	152	FS
	S. aspera	152	FS
	S. forbessii	269	FS
accD	*B. americana*	57	FS
	S. forbessii	12	FS
	S. hermonthica	12	FS
	S. aspera	59	FS
	B. americana	92	FS
	S. forbessii	47	FS
	S. hermonthica	47	FS
	S. aspera	129	FS
ycf1	*B. americana*	251	FS
	S. forbessii	256	FS
	S. hermonthica	258	FS
	S. aspera	254	FS
	B. americana	360	FS
	S. forbessii	386	FS

	S. hermonthica	269	FS
	S. aspera	269	FS
	S. hermonthica	319	FS
	S. aspera	317	FS
	S. hermonthica	358	FS
	S. aspera	356	FS
	S. hermonthica	372	FS
	S. aspera	370	FS
	S. hermonthica	391	FS
	S. aspera	389	FS
	S. hermonthica	425	FS
	S. aspera	424	FS
ndhF	*S. hermonthica*	3	SC
	S. aspera	3	SC
	S. forbessii	20	SC
	S. hermonthica	20	SC
	S. aspera	20	SC
	S. hermonthica	81	FS
	S. aspera	91	FS
	S. forbessii	617	FS
	S. aspera	593	FS
ndhI	*B. americana*	90	SC
	S. forbessii	89	SC
	S. hermonthica	90	SC
	S. aspera	90	SC
	S. hermonthica	112	FS
	S. aspera	112	FS
	S. hermonthica	392	FS
	S. aspera	391	FS
	S. forbessii	467	FS
	S. hermonthica	491	FS
	S. aspera	475	FS
	B. americana	674	FS
	S. forbessii	648	FS
	S. hermonthica	661	FS
	S. aspera	660	FS
	S. hermonthica	1288	FS
	S. aspera	1287	FS
	S. hermonthica	1452	SC
	S. aspera	1451	SC
	S. hermonthica	1462	FS
	S. aspera	1461	FS
ycf2	*S. hermonthica*	482	FS
	S. aspera	489	FS
	B. americana	848	FS
	S. forbessii	1024	FS
	S. hermonthica	987	FS
	S. aspera	994	FS

ndhK	*S. hermonthica*	3	FS
	S. aspera	3	FS

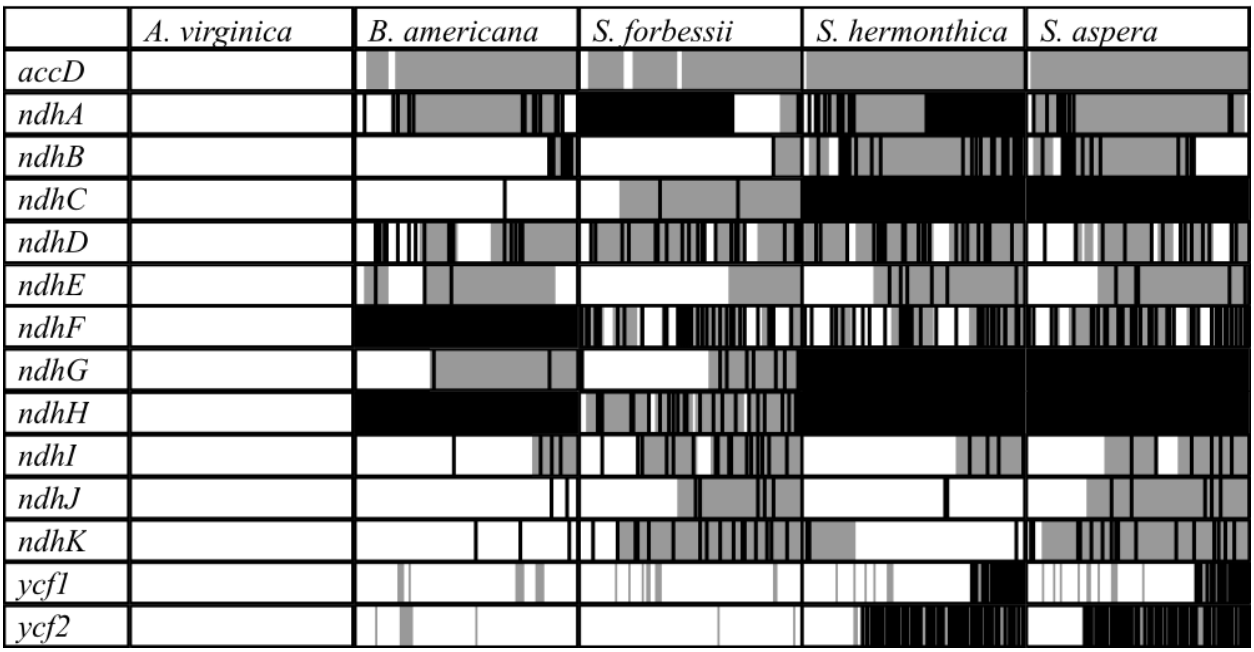
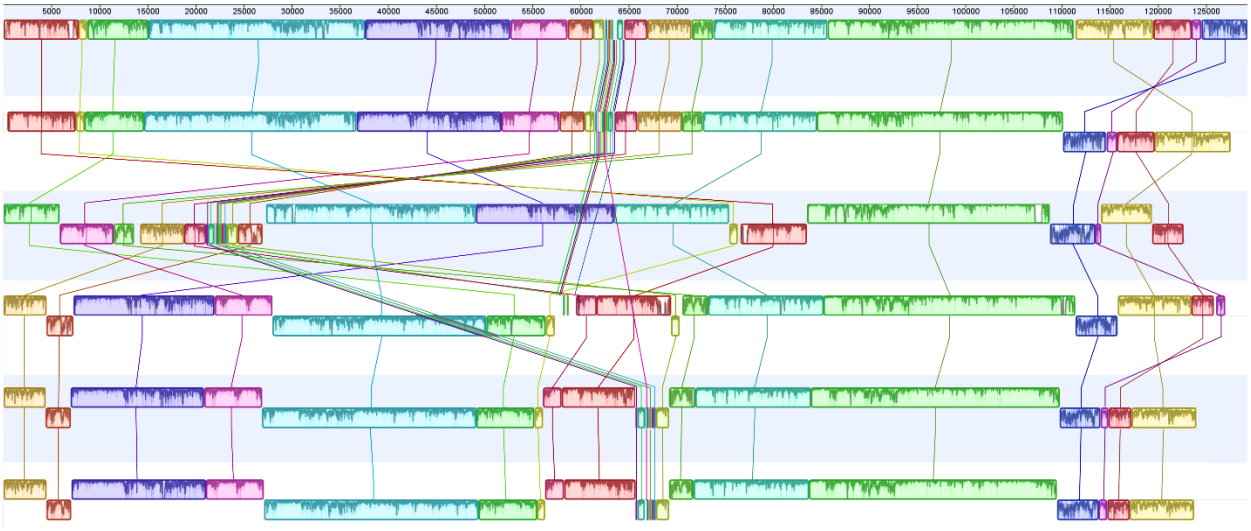


Figure 4.1. Stop codons and frameshifts in potential pseudogenes. Genes that are missing or predicted as non-functional in at least one of the five species. Solid white bars indicate the gene is present and appears functional in that species, while bars with gray shading or black lines may be pseudogenes. The structure of the protein-encoding portion of the gene is indicated within each box, with the 5' end being at the far left and the 3' end at the far right. The size of each box is 100% of the protein-encoding region, so they are not corrected for the different sizes of the coding regions of each gene. Gray shading indicates portions of the genes with a shifted frame, while black lines indicate stop codons. Solid black bars are genes or gene segments that are completely missing from that species. For instance, the figure shows that the 5' end of *ndhA* is missing in *S. forbessii* and the 3' end is missing in *S. hermonthica*, but the multiple stop codons and frameshifts in the 5' end in *S. hermonthica* should guarantee that it is a non-functional gene.

A)



B)

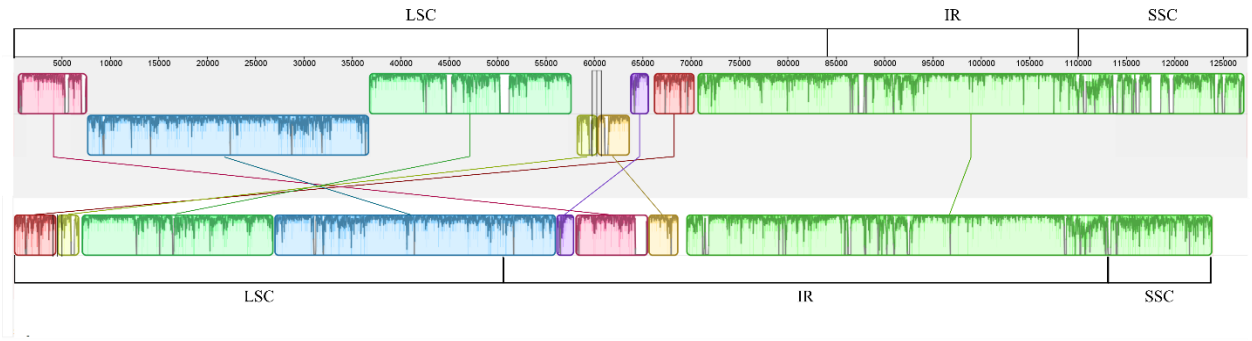


Figure 4.2. Mauve alignment of plastomes. A. Alignment done in Mauve 2.3.1 showing plastomes with one copy of the IR taken out. The order of plastomes, from top to bottom, is *L. philippensis*, *A. virginica*, *B.americana*, *S. forbessii*, *S. hermonthica*, and *S. aspera*. Each colored block is a region of collinear sequence among all six plastomes. Blocks on the top row are in the same orientation, while blocks on the bottom row are in inverse orientation.

B. Mauve alignment of *A. virginica* (top) and *S. hermonthica* (bottom). The blocks and colors do not correspond to Figure 2A. The LSC, IR, and SSC are indicated below each alignment.

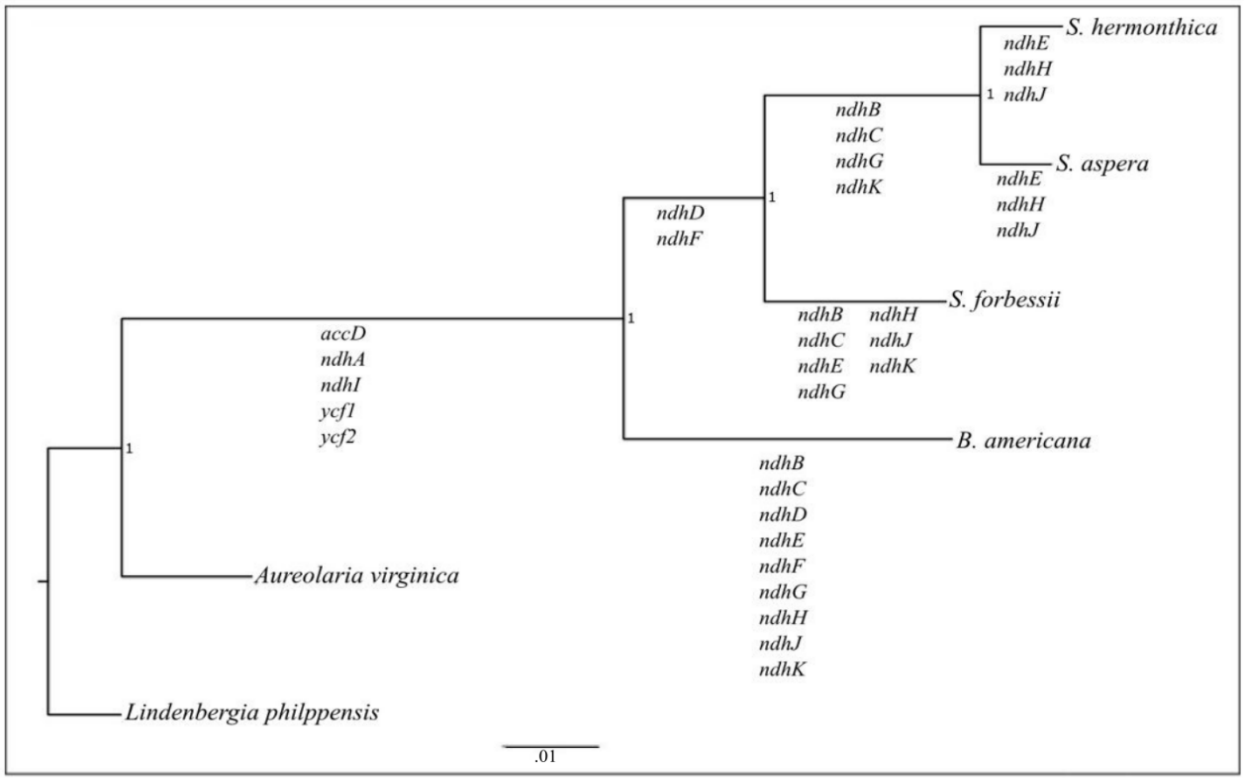


Figure 4.3. Phylogenetic tree of the plastomes. Phylogenetic tree of the investigated hemiparasite plastomes, decorated with the branch of specific inactivational events for the listed genes. These inactivational events included deletions, frameshifts and/or stop codon generation. This tree is from MrBayes (as describe in Materials and Methods), and the branch lengths are proportional to the number of substitutions per site with the scale shown at the bottom. The number 1 at each node is the posterior probability, the likelihood that the tree is correct.

CHAPTER 5

CONCLUSIONS

Recombination is one of the driving forces in the evolution of disease resistance (R) genes in plants (Michelmore and Meyers 1998, Young 2000). In chapter 1, I investigated unequal meiotic recombination in maize within a disease resistance gene cluster from sorghum, the *Pc* locus. The *Pc* locus consists of three paralogs, *Pc* A, *Pc* B, and *Pc* C. All three belong to the NBS-LRR class of R genes (Nagy et al. 2008). The NBS-LRR genes are the largest class of R genes and are named after their nucleotide binding site (NBS) and leucine rich repeat (LRR) domains. The NBS domain is responsible for downstream activation of the plant disease response and is conserved among NBS-LRR genes. The LRR domain is responsible for pathogen recognition. Since each gene recognizes different pathogens, each gene contains a different LRR domain, and they tend to be highly divergent (Dangl and Jones 2001, McHale et al. 2006). Nagy et al. 2008 found several unequal recombination events in sorghum between *Pc* A and *Pc* C. Surprisingly, 7 out of 10 of these recombination events occurred in the 3' end of the gene where there is the most sequence divergence, even though recombination is usually strongly biased towards regions of high sequence identity. Specifically, they occurred within a 560 bp hotspot located in the LRR domain. Recombination here would create a gene containing a new LRR domain, with the potential to provide resistance against a new pathogen. This unequal recombination pattern was not seen when the *Pc* locus was transformed into yeast, so sorghum seems to contain some mechanism directing recombination to the LRR domain. To study this

mechanism, we created transgenic maize and rice lines that contained *Pc* A and *Pc* C fluorescent fusion proteins. The lines were designed to express a *Pc* A-mTFP fusion protein and fluoresce teal. When unequal recombination occurred between the *Pc* genes, there would be a color change in the fluorescence from teal to green in those cells and their mitotic descendants. Unfortunately, we were unable to detect any fluorescence in any of our lines, possibly due to misfolding of the fluorescent protein leading to a loss of fluorescent activity (reviewed in Snapp 2010).

Because the fluorescence assay was not feasible, we developed a PCR-based technique that allowed us to identify meiotic recombination events in maize. By running PCR on DNA extracted from pools of pollen grains, we were able to screen thousands of meiotic products in a single PCR reaction. We found a total of 23 meiotic recombination events between *Pc* A and *Pc* C. All 23 recombined within the conserved 5' end of the gene, unlike the pattern seen in sorghum. Since we ran PCR on DNA samples extracted from a known number of pollen grains, we were able to calculate a rate of 1.3×10^{-4}, or 1 unequal recombination event every ~5549 pollen grains. This rate is similar to the 1/1000 to 1/10,000 seen for unequal recombination between tandemly repeated genes (Sun et al. 2001, Yandeau-Nelson 2006, Nagy et al. 2007).

There are several possible reasons for the different unequal recombination patterns between the *Pc* genes in maize compared to sorghum. The mechanism responsible for the LRR-sited recombination in sorghum may not be present in maize. Maize has been intensively inbred since domestication, and a process that increased R gene diversity may have been inadvertently selected against, perhaps because the same process that creates a new R gene could also remove an already useful R gene. Alternatively, the mechanism may exist in maize but not function at the transgenic *Pc* genes. It's possible that in the transgenic lines we looked at, the *Pc* genes were inserted into a part of the genome where the mechanism did not function. Another possibility is

that the mechanism depends on certain epigenetic marks. These marks may not be established at the transgenes. A future study could look at unequal meiotic recombination between more disease resistance genes, at both their natural sites and in transgenics, including genes originating from maize, to see if any show similar LRR-sited recombination. This same strategy could also be expanded to look at unequal recombination in other species, although this experiment only works well on species that produce a sufficient quantity of pollen.

Breeding for disease resistance in crops is restricted by the number of disease resistance specificities available in the germplasm. In many cases, resistance specificities for a disease may never have existed in a particular crop. This is an expected problem with microbial pathogens that have enormous populations and short generation times that allow them to rapidly evolve to escape recognition. It can also occur after introduction of a crop into a new environment, when it is exposed to pathogens it has never encountered in the past. Alternatively, many resistance specificities that once existed have been lost. R genes, including NBS-LRR genes, follow a rapid birth and death pattern of evolution. Resistance specificities can be lost through random mutations or unequal recombination if they are not under selection (that is, if the disease is not present) (Michelmore and Meyers 1998). Most crops have undergone a bottleneck at the beginning of their domestication, followed by intense inbreeding. Both of these processes reduce genetic variation and have resulted in an even greater loss of resistance specificities in crops when compared to their wild counterparts (Leberg 1992, Hyten et al. 2006, Charlesworth and Charlesworth 1987, Spielman et al. 2004).

Plants have already evolved a highly adaptable defense response against pathogens. However, in many cases this is not adequate when it comes to crops. For one, evolution of new resistance specificities is based on chance, so there is no guarantee that a plant will ever evolve a

resistance against a specific pathogen. This problem is amplified in crops, which have less genetic diversity in their gene pools from which to evolve new specificities, and their tendency toward monoculture makes epidemics particularly likely. Even if a plant can naturally evolve resistance to a new pathogen, this can take a very long time. When that plant is an important crop, reduced food productivity and human starvation may occur while waiting for a source of resistance to evolve and/or be identified. This problem is further complicated by resistance specificities that require interactions between multiple R genes (Ashikawa et al. 2008, Birker et al. 2009, Arora et al. 2013).

Understanding the mechanism and influencing factors behind LRR-directed recombination in NBS-LRR genes would be a huge step forward to breeding new disease resistant plants. If these conditions are identified, breeders might generate novel resistance alleles quickly and efficiently. Discovering specific genes that direct meiotic recombination to the LRR regions of NBS-LRR genes would be particularly helpful. These genes could be upregulated to increase chances of novel resistance specificities being generated in progeny.

The PCR technique we used on maize pollen is effective at finding recombination events at a single locus. We used this technique again in chapter 2 to study unequal meiotic recombination between LTRs, but it could be used to find recombination events at any locus in any plant that produces a sufficient amount of pollen. The benefits are that it allows us to rapidly and inexpensively screen thousands of meiotic products to find recombination events. It also allows quantification of recombination events so that a recombination rate can be calculated. This, in turn, allows for comparison of recombination rates between different sequences, genome locations, etc. to study how different factors may affect recombination.

In chapter 2, we studied unequal meiotic recombination between the LTRs of LTR retrotransposons. Transposons are responsible for the majority of the variation in plant genome size, as well as most of the structural variation. LTR retrotransposons, named after their two flanking direct repeats, are among the most abundant type of transposon in eukaryotes (Kidwell and Lisch 1997, Bennetzen 2000). In maize, it is estimated they make up ~75% of the genome (Schnable et al. 2009). LTR retrotransposon amplification can rapidly increase genome size. Some of this increase can be counteracted by unequal recombination between the LTRs to generate solo LTRs (Shepherd et al., 1984; Sentry and Smyth, 1989; SanMiguel et al., 1996; Noma et al., 1997; Chen et al., 1998; Han et al., 2000, Ma et al. 2004). This can occur between either the LTRs of the same element or the LTRs of two different elements, although the majority of events observed so far have been intra-element (Devos et al. 2002, Shirasu et al. 2000, Vitte and Panaud 2003). Unequal recombination between two LTRs deletes the internal retrotransposon sequence along with one of the LTR copies. The remaining solo LTR contains part of the 5' end of the 5' LTR and part of the 3'end of the 3' LTR. In this project, we chose 50 LTR retrotransposons located in the distal regions of the short arm of chromosome 1 and the long arm of chromosome 9. We used the same PCR technique on maize pollen DNA that was employed to find *Pc* recombinants. For each LTR retrotransposon, we ran PCR on DNA extracted from pools of maize pollen to identify any unequal recombination events that generated a solo LTR. We found a total of 23 such unequal recombination events, and they all arose from 10 of the 50 LTR retrotransposon. The median number of solo LTRs generated for each of the 10 unequally recombining elements was 2, indicating that there was a distinct separation between unequally recombining and non-unequally recombining sets of LTR retrotransposons.

In comparing these two sets, I found a significant positive correlation between recombination rate and LTR sequence similarity and a significant negative correlation between recombination rate and the amount of DNA methylation. I did not find any significant correlation between unequal recombination activity and LTR length, chromosomal position, or distance from the nearest genes. However, at least some of the lack of observed correlations is likely due to the smallish number of LTR retrotransposons that we studied. Most of the LTRs were between ~1150 to ~1400 bp, which is not a large size difference and could explain why we saw no correlation with LTR size and recombination. We only looked at LTR retrotransposons from a very small portion of the whole maize genome, all of it with "gene-rich" (i.e., euchromatic) characteristics from a maize cytogenetics standpoint (Bruggmann et al. 2006). If we had looked at LTR retrotransposons across the maize genome, we would likely have seen an effect of location on unequal recombination rates (as predicted in rice (Ma and Bennetzen, 2006). However, this was not feasible at this time because the assembly and annotation of the maize genome is not particularly good in the highly repetitive regions near centromeres or other heterochromatic domains (Schnable et al. 2009).

The biggest issue with this experiment was designing pairs of primers that each worked individually with the control primers, but did not amplifying non-specific sequences when used together. This was challenging because the majority of the maize genome is highly repetitive, such that most primers will have more than one potential binding site. We solved this problem by designing primers that spanned the junctions of each LTR retrotransposon insertion site, making each primer-binding site unique (Liu et al. 2007). This experiment would probably work even better in another species with a smaller genome, as long as it produced a sufficient amount of pollen. A smaller genome would have less repetitive DNA, making the primer binding even

more likely to be specific. It would also reduce the ratio of the target LTR retrotransposon to the total amount of DNA.

LTR retrotransposons play a huge role in genome evolution. Lineage specific transposon activity can cause drastic differences between even closely related species. The genome of *Oryza autraliensis* is more than twice the size of *O. sativa* due to the amplification of just three retrotransposon families (Piegu et al. 2006). Understanding the mechanisms and frequency by which transposons both amplify and are removed will help us better understand plant genome evolution in general. Recombination also plays a major role in genome evolution by increasing genetic diversity. Unequal recombination can cause gene duplications and deletions, inversions, and other large chromosomal rearrangements like translocations (Gaut et al. 2007). A larger scale version of this study could help identify and quantify the effects of various factors on recombination rates. A better understanding of the general factors influencing recombination combined with an understanding of the specific mechanisms regulating the LRR directed recombination seen at the *Pc* locus could greatly help in the development of new disease resistance genes. It would also help us understand the evolutionary history of the NBS-LRR genes, as well as any genes organized in tandem gene clusters.

In chapter 3, we sequenced and analyzed the chloroplast genomes from five species of parasitic plants. The chloroplast genome, or plastome, of most plants tends to be highly conserved in terms of size, structure, gene content, and gene order (Raubeson and Jansen 2005). Parasitic plants, however, often have highly rearranged plastomes. Less reliance on photosynthesis reduces selectional constraints on photosynthetic genes, but this should be limited to non-photosynthetic parasites. Deletions of non-essential genes often result in smaller plastome sizes. Parasitic plants can be either hemiparasites or holoparasites. Hemiparasites acquire some

nutrients from the host and some from photosynthesis. Hemiparasites can be further classified as obligate hemiparasites, which require a host, or facultative hemiparasites, which can survive without a host. Holoparasites are non-photosynthetic and require a host for nutrients.

Plastome divergence generally increases as parasitic activity increases and photosynthetic activity decreases (dePamphilis and Palmer 1990, Wolfe et al. 1992b, Delavault et al. 1996, Funk et al. 2007, McNeal et al. 2007, Wickett et al. 2008, Wicke et al. 2013, Wicke et al. 2016, Samigullin et al. 2016). We assembled the full plastomes from five hemiparasites, all in the Orobanchaceae family: *Aureolaria virginica, Buchnera americana, Striga forbesii, Striga hermonthica,* and *Striga aspera*. *A. virginica* and *B. americana* are facultative hemiparasites, while the three *Striga* species are obligate hemiparasites. We compared them to the plastome of a closely related non-parasite, *Lindenbergia philippensis*, that was assembled by Wicke et al. 2013. *A. virginica* had no major structural changes or gene losses compared to *L. philippensis*. *B. americana* and the *Striga* species all had major structural changes and physical or functional gene losses. The most surprising result was a large increase in plastome size for these four species. Plastomes are arranged into a large single copy (LSC) region and a small single copy (SSC) region separated by two virtually identical inverted repeats (IRs). We found that the increase in plastome size was due to an expansion of the IRs into the two single copy regions.

We are still unsure what effect the increased repeat size has on parasitic plastome function or evolution. Genes contained within the repeat regions in the plastome have lower evolutionary rates than those outside, due to gene conversion between the repeat regions (Zhu et al. 2016). This could provide a mechanism to protect beneficial genes from accumulating mutations in an unstable plastome. There are no plastome assemblies from other species in either the *Aureolaria* genus or *Buchnera* genus. It would be interesting to see if all *Aureolaria* species

show conserved plastomes or if the plastomes have more severely diverged in some species. Similarly, it would be interesting to see if other *Buchnera* species are more or less conserved than *B. americana*. If there are differences in the species of either of these genera it could help elucidate the evolutionary history of the plastome. Additionally, the plastome assembly of a species somewhere between these two on the phylogenetic tree could provide further insight into plastome divergence and repeat expansion. Assembling plastomes of additional species similar to *B. americana* and the *Striga* species in terms of plastome divergence and repeat expansion size could help determine if there are any patterns to the repeat expansion or if it is completely random.

Parasitic plants are a major problem in many parts of the world. The *Striga* species alone cause devastating crop losses and threaten food supplies across Africa, often for poorer farmers in less developed countries (Ejeta 2007). Understanding the evolution of the plastome as plants become more parasitic could give us a deeper understanding of how plants evolve parasitism. Although the exact pattern of gene loss is often lineage-specific, certain genes are lost more commonly than others. The *ndh* genes are among the most common and often the first genes lost after a lineage assumes a parasitic lifestyle (McCoy et al. 2008, Wu et al. 2009, McNeal et al. 2007, Funk et al. 2007, dePamphilis and Palmer 1990, Chang et al. 2006, Wu et al. 2010, Wicke et al. 2013, Wicke et al. 2016, Wickett et al. 2008, Barret et al. 2014). We saw the same pattern in *B. americana* and the *Striga* species. Understanding the reasons why these and other commonly lost genes are non-essential to photosynthetic parasites could provide insight into the specific environments and conditions in which plants are most likely to evolve parasitism.

REFERENCES

Ahn S, Tanksley SD. 1993. Comparative linkage maps of the rice and the maize genomes. *Proc. Natl. Acad. Sci. USA.* 92:7980–7984.

Ali RAMA, El-Hussein AA, Mohamed KI, Babiker AGT. 2009. Specificity and genetic relatedness among *Striga hermonthica* strains in Sudan. *Life Sci. Int. J.* 3:1159-1166.

Ameline-Torregrosa C, Wang BB, O'Bleness MS. 2008. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. *Plant Physiol.* 146:5-21.

Anderson LK, Doyle GG, Bringham B, Carter J, Hooker KD, Lai A, Rice M, Stack SM. 2003. High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics.* 165:849-865.

Anderson LK, Lai A, Stack SM, Rizzon C, Gaut BS. 2006. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome Res.* 16:115-122.

Anderson PA, Okubara PA, Arroyo-Garcia R, Meyers BC, Michelmore RW. 1996. Molecular analysis of irradiation-induced and spontaneous mutants at a disease resistance locus in *Lactuca sativa*. *Mol. Gen. Genet.* 251:316-325.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis. *Nature*. 408:796-815.

Arora D, Gross T, Brueggeman R. 2013. Allele characterization of genes required for *rpg4*-mediated wheat stem rust resistance identifies *Rpg5* as the R gene. *Phytopathology*. 103:1153-1161.

Ashikawa I, Hayashi N, Yamane H, Kanamori H, Wu J, Matsumoto T, Ono K, Yano M. 2008. Two adjacent nucleotide-binding site–leucine-rice repeat class genes are required to confer Pikm-specific blast resistance *Genetics* 180:2267-2276.

Atera EA, Itoh K, Onyango JC. 2011. Evaluation of ecologies and severity of Striga weed on rice in sub-Saharan Africa. *Agric. Biol. J. of N. Am.* 2:752-760.

Barakat A, Carels N, Bernardi G. 1997. The distribution of genes in the genomes of Gramineae. *Proc. Natl. Acad. Sci. USA.* 94:6857–6861.

Barrett CF, Freudenstein JV, Li J, Mayfield-Jones DR, Perez L, Pires JC, Santos C. 2014. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids and implications for heterotrophic angiosperms. *Mol. Biol. Evol*. 31:3095–3112.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* 5:e1000732.

Baudat F, Imai Y, de Massy B. 2013. Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.* 14:794-806.

Belkhadir Y, Subramaniam R, Dangl J. 2004. Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Curr. Opin. Plant Biol*. 7:391-399.

Bennet JR, Mathews S. 2006. Phylogeny of the parasitic plant family Orobanchaceae inferred from phytochrome A. *Am. J. Bot*. 93:1039-1051.

Bennet MD, Smith JB. 1991. Nuclear DNA amounts in angiosperms. *Phil. Trans. R. Soc. Land B. Biol. Sci.* 334:309-345.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42:251-269.

Bennetzen JL. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115:29-36.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* 95:127-132.

Bennetzen JL, SanMiguel P, Chen M, Tikhonov A, Francki M, AvramovaZ. 1998. Grass Genomes. *Proc. Natl. Acad. Sci. USA.* 95:1975–1978.

Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hotspot activity and meiotic instability in humans. *Nat. Genet.* 42:859-863.

Bergerat A, deMassy B, Gadelle D, Varoutas PC, Nicolas A, Forterre P. 1997. An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature* 386: 414–417.

Berner DK, Cardwell KF, Faturoti BO, Ikie FO, Williams OA. 1994. Relative roles of wind, crop seeds, and cattle in dispersal of *Striga* spp. *Plant Dis.* 78:402-406.

Bharathalakshmi, Werth CR, Musselman LJ. 1990. A study of genetic diversity among host-specific populations of the witchweed *Striga hermonthica* (Scrophulariaceae) in Africa. *Plant Syst. Evol.* 172:1-12.

Birker D, Heidrich K, Takahara H, Narusaka M, Deslandes L, Narusaka Y, Reymond M, Parker JE, O'Connell R. 2009. A locus conferring resistance to *Colletotrichum higginsianum* is shared by four geographically distinct Arabidopsis accessions. *Plant J*. 60:602-613.

Bock R. 2007. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. Cell and molecular biology of plastids. Berlin Heidelberg: Springer. p. 29-63.

Borner GV, Kleckner N, Hunter N. 2004. Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. *Cell.* 117:29-45

Bousios A, Minga E, Kalitsou N, Pantermali M, Tsaballa A, Darzentas N. 2012. MASiVEdb: the Sirevirus Plant Retrotransposon Database. *BMC Genomics* 13:158.

Bremer K, Fijs EM, Bremer B. 2004. Molecular phylogenetic dating of asteroid flowering plants shows early Cretaceous diversification. *Syst. Biol*. 53:496-505.

Brubaker CL, Paterson AH, Wendel JF. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome.* 42:184-2003.

Bruggmann R, Bharti AK, Gundlach H, Lai J, Young S, Pontaroli AC, Wei F, Haberer G, Fuks G, Du C, et al. 2006. Uneven chromosome contraction and expansion in the maize genome. *Genome Res.* 16:1241-1251.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell.* 17:343–360.

Casano LM, Zapata JM, Marti M, Sabater B. 2000. Chlororespiration and poising of cyclic electron transport. *J. Biol. Chem.* 275:942–948.

Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. 2014. Early allopolyploid evolution in the post-neolithic *Brassica napus* oilseed genome. *Science.* 345:950-953.

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol. Biol. Evol*. 23:279–291.

Charlesworth D, Charlesworth B. 1987. Inbreeding depression and its evolutionary consequences. *Ann. Ecol. Syst.* 18:237-268.

Chelysheva L, Gendrot G, Vezon D, Doutriaux MP, Mercier R, Grelon M. 2007. Zip4⁄Spo22 is required for class I CO formation but not for synapsis completion in *Arabidopsis thaliana*. *PLoS Genet.* 3:802–813.

Chen CB, Zhang W, Timofejeva L, Gerardin Y, Ma H. 2005. The Arabidopsis ROCK-n-ROLLERS gene encodes a homolog of the yeast Atp-dependent DNA helicase Mer3 and is required for normal meiotic crossover formation. *Plant J.* 43:321–334.

Chen M, SanMiguel P, Bennetzen JL. 1998. Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. *Genetics* 148:435–443.

Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA, Copenhaver GP, Franklin FC, McVean G, Henderson IR. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat Genet*. 45:1327–1336.

Choulet F, Albert A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, et al. 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science* 2014, 345:1249721.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. 2006. The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol. Biol. Evol.* 23:2175-2190.

Clarke JD. 2009. Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor Protoc.* doi: 10.1101/pdb.prot5177.

Colwell AE. 1994. Genome Evolution in a Nonphotosynthetic Plant, *Conopholis americana*. PhD dissertation (St. Louis, WA: Washington University).

Copenhaver GP, Housworth EA, Stahl FW. 2002. Crossover interference in Arabidopsis. *Genetics.* 160:1631–1639.

Craig NL, Craigie R, Gellert M, Lambowitz AM. editors. 2002. Mobile DNA II. Washington: ASM Press.

Cui L, Veeraraghavan N, Richter A, Wall K, Jansen RK, Leebens-Mack J, Makalowska I, dePamphilis CW. 2006. ChloroplastDB: the chloroplast genome database. *Nucleic Acids Res.* 34:D692-696.

Dangl JL, Jones JDG. 2001. Plant pathogens and integrated defense response to infection. *Nature* 411:826-833.

Daniell H, Lee S, Grevich J, Saski C, Quesada-Vargas T, Guda C, Tomkins J, Jansen RK. 2006. Complete chloroplast genome sequences of *Solanum bulbocastanum, Solanum*

lycopersicum and comparative analyses with other Solanaceae genomes. *Theor. Appl. Genet.* 112:1503-1518.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Gen. Res.* 14:394-1403.

Day RN, Booker CF. 2008. Characterization of an improved donor fluorescent protein for Förster resonance energy transfer microscopy. *J. Biomed. Opt.* doi:10.1117/1.2939094.

Delannoy E, Fujii S, Colas des Francs-Small C, Brundrett M, Small I. 2011. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Mol. Biol. Evol*. 28:2077–2086.

de la Chaux N, Wagner A. 2011. BEL/Pao retrotransposons in metazoan genomes. *BMC Evol. Biol.* 11:54.

Delavault PM, Russo NM, Lusson NA, Thalouarn P. 1996. Organization of the reduced plastid genome of *Lathraea clandestina* an achlorophyllous parasitic plant. *Physiol. Plant* 96:674–682.

de los Santos T, Hunter N, Lee C, Larkin B, Loidl J, Hollingsworth NM. 2003. The Mus81 /Mms4 endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics.* 164:81-94.

dePamphilis CW, Palmer JD. 1990. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature.* 348:336-339.

Deslandes L, Olivier J, Peeters N, Feng DX, Khounlotham M, Boucher C, Somssich I, Genin S, Marco Y. 2003. Physical interaction between RRSI-R, a protein conferring resistance to

bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. *Proc. Natl. Acad. Sci. USA* 100:8024-8029.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* 12:1075-1079.

de Vries J, Sousa FL, Bölter B, Soll J, Gould SB. 2015. YCF1: a green TIC? *Plant Cell*. 27:1827-1833.

Dodds P, Rathjen JP. 2010. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* 11:539-548.

Dooner HK, He L. 2014. Polarized gene conversion at the *bz* locus of maize. *Proc. Natl. Acad. Sci. USA.* 111:13918-13923.

Dooner HK, Martinez-Ferez IM. 1997. Recombination occurs uniformly within the *bronze* gene, a meiotic recombination hotspot in the maize genome. *Plant Cell.* 9:1633-l 646.

Downie SR, Jansen RK. 2015. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat mitochondrial to plastid transfer of DNA and identification of highly divergent noncoding regions. *Syst Bot.* 40:336–351.

Downie SR, Katz-Downie DS, Wolfe KH, Calie PJ, Palmer JD. 1994. Structure and evolution of the largest chloroplast gene (ORF2280): internal plasticity and multiple gene loss during angiosperm evolution. *Curr. Genet.* 25:367–3781.

Downie SR, Palmer JD. 1992. Restriction site mapping of the chloroplast DNA inverted repeat: a molecular phylogeny of the Asteridae. *Ann. Mo. Bot. Gard*. 79:266-283.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull*. 19:11-15.

Drouaud J, Camilleri C, Bourguignon P, Canaguier A, Bérard A, Vezon D, Giancola S, Brunel D, Colot V, Prum B. et al. 2006. Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Res.* 16:106-114.

Dudley DD, Chaudhuri J, Bassing CH, Alt FW. 2005. Mechanism and control of V(D)J recombination versus class switch recombination: similarities and differences. *Adv. Immunol.* 86:43-112.

Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT, Hajrah NH, Alharbi NS, Al-Malki AL, Sabir JSM, Bailey CD. 2015. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. *Sci. Rep*. 5:16958.

Eggleston WB, Alleman M, Kermicle JL. 1995. Molecular organization and germinal instability of R-stippled maize. 141:347-360.

Eickbush TH, Malik HS. 2002. Origins and evolution of retrotransposons. In: Mobile DNA II. Craig NL, Craigie R, Gellert M, Lambowitz AM. editors. Herndon: ASM Press, 1111-1144.

Ejeta G. 2007. The *Striga* scourge in Africa: a growing pandemic. In: Ejeta G, Gressel J, editors. *Integrating new technologies for Striga control: towards ending the witch-hunt*. World Scientific Publishing Co. Pte. Ltd, Singapore. p. 71-84.

Ejeta G, Rich PJ, Mohamed A. 2007. Dissecting a complex trait to simpler components for effective breeding of sorghum with a high level of *Striga* resistance. In: Ejeta G, Gressel J, editors. *Integrating new technologies for Striga control: towards ending the witch-hunt.* World Scientific Publishing Co. Pte. Ltd, Singapore. p. 87-98.

Ellis JG, Lawrence GJ, Luck JE, Dodds PN. 1999. Identification of regions in alleles of the flax rust resistance gene *L* that determines differences in gene-for-gene specificity. *Plant Cell* 11:495-506.

Estep MC, DeBarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity (Edinb).* 110:194-204.

Estep MC, Gowda BS, Huang K, Timko MP, Bennetzen JL. 2012. Genomic characterization for parasitic weeds of the genus by sample sequence analysis. *Plant Genome*. 5:30-41.

Estep M, Van Mourik TA, Muth P, Guindo D, Parzies HK, Koita OA, Weltzien E, Bennetzen JL. 2010. Development of microsatellite markers for characterizing diversity in a parasitic witchweed, *Striga hermonthica* (Orobanchaceae). *Mol. Ecol. Resour.* 10:1098-1105.

Estep MC, Van Mourik TA, Muth P, Guindo D, Parzies HK, Koita OA, Weltzien E, Bennetzen JL. Genetic diversity of a parasitic weed, *Striga hermonthica*, on sorghum and pearl millet in Mali. *Trop. Plant Biol.* 4:91-98

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14:2611-2620.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics.* 131:479-491.

Faris JD, Zhang Z, Lu H, Lu S, Reddy L, Cloutier S, Fellers JP, Meinhardt SW, Rasmussen JB, Xu SS, Oliver RP, Simons KJ, Friesen TL. 2010. A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. *Proc. Natl. Acad. Sci. USA* 107:13544-13549.

Feldman M. 1993. Cytogenetic activity and mode of action of the pairing homeologous (Ph1) gene of wheat. *Crop Sci.* 33:894-897.

Finnegan DJ. Retrotransposons. *Curr. Biol.* 11:R432-R437.

Flor HH. 1971. Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* 9:275-296.

Friedrich T, Steinmüller K, Weiss H. 1995. The proton-pumping respiratory complex I of bacteria and mitochondria and its homologue in chloroplasts. *FEBS Lett.* 367:107–111.

Fu H, Dooner HK. 2002. Intraspecific violation of genetic collinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA*> 99:9573-9578.

Funk H, Berg S, Krupinska K, Maier U, Krause K. 2007. Complete DNA sequences of plastid genomes of two parasitic flowering plant species *Cuscuta reflexa* and *Cuscuta gronovii. BMC Plant Biol.* 7:45.

Gallardo MH, Bickham JW, Honeycut RL, Ojeda RA, Köhler N. 1999. Discovery of tetraploidy in a mammal. *Nature.* 401:341.

Gaut BS. 2002. Evolutionary dynamics of grass genomes. *New Phytol.* 154:15-28.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Gen.* 8:77-84.

Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, Dawe RK. 2012. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res*. 23:628-637.

Gethi, JG, Smith ME, Mitchell SE, Kresovich S. 2005. Genetic diversity of *Striga hermonthica* and *Striga asiatica* populations in Kenya. *Weed Res.* 45:64-73.

Geuting V, Kobbe D, Hartung F, Durr J, Focke M, Puchta H. 2009. Two distinct Mus81–Eme1 complexes from Arabidopsis process Holliday junctions. *Plant Physiol.* 150: 1062–1071.

Gill KS, Gill BS, Endo TR, Boyko EV. 1996. Identification and high-density mapping of gene rich regions in chromosome group 5 of wheat. *Genetics.* 143:1001–1012.

Goulding SE, Wolfe KH, Olmstead RG, Morden CW. 1996. Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet*. 252:195-206.

Grandont L, Jenczewski E, Lloyd A. 2013. Meiosis and its deviations in polyploid plants. *Cytogenet. Genome Res.* 140:171-184.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. *Plant J.* 66:34-44.

Greihuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol.* (Stuttg). 8:770-777.

Grelon M, Vezon D, Gendrot G, Pelletier G. 2001. AtSPO11-1 is necessary for efficient meiotic recombination in plants. *EMBO J.* 20: 589–600.

Grewe F, Viehoever P, Weisshaar B, Knoop V. 2009. A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte *Isoëtes engelmannii*. *Nucl. Acids Res.* 37:5093-5104.

Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK. 2010. Implications of the plastid genome sequence of typha (typhaceae, poales) for understanding genome evolution in poaceae. *J. Mol. Evol*. 70:146-166.

Guo W, Grewe F, Cobo-Clark A, Fan W, Duan Z, Adams RP, Schwarzbach AE, Mower JP. 2014. Predominant and substoichiometric isomers of the plastid genome coexist within *Juniperus* plants and have shifted multiple times during cupressophyte evolution. *Genome Biol. Evol.* 6:580-590.

Haber JE. 2000. Partners and pathways repairing a double-strand break. *Trends Genet.* 16:259-264.

Haber JE. 2012. Mating-type genes and *MAT* switching in *Saccharomyces cerevisiae*. *Genetics.* 191:33-64.

Hamrick JL. 1982. Plant population genetics and evolution. *Am. J. Bot.* 69: 1685-1693.

Han CG, Frank MJ, Ohtsubo H, Ohtsubo E. 2000. New transposable elements identified as insertions in rice transposon *Tnr1*. *Genes Genet. Syst.* 75:69–77.

Hartung F, Puchta H. 2000. Molecular characterization of two paralogous Spo11 homologues in *Arabidopsis thaliana. Nucleic Acids Res.* 28:1548-1554.

Hartung F, Puchta H. 2001. Molecular characterization of homologues of both subunits a (Spo11) and b of the archaebacterial topoisomerase 6 in plants. *Gene.* 271: 81–86.

Hartung F, Wurz-Wildersinn R, Fuchs J, Schubert I, Suer S, Puchta H. 2007. The catalytically active tyrosine residues of both SPO11-1 and SPO11-2 are required for meiotic double-strand break induction in Arabidopsis. *Plant Cell*. 19: 3090–3099.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.* 16:1252-1261.

Higgins JD, Armstrong SJ, Franklin FCH, Jones GH. 2004. The Arabidopsis MutS homolog AtMSH4 functions at an early step in recombination: evidence for two classes of recombination in Arabidopsis. *Gene. Dev.* 18: 2557–2570.

Higgins JD, Sanchez-Moran E, Armstrong SJ, Jones GH, Franklin FCH. 2005. The Arabidopsis synaptonemal complex protein ZYP1 is required for chromosome synapsis and normal fidelity of crossing over. *Genes Dev.* 19:2488–2500.

Higgins JD, Vignard J, Mercier R, Pugh AG, Franklin FCH, Jones GH. 2008. AtMSH5 partners AtMSH4 in the class I meiotic crossover pathway in *Arabidopsis thaliana*, but is not required for synapsis. *Plant J.* 55: 28–39.

Hirsch CD, Springer NM. 2017. Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta.* 1860:157-65.

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci USA.* 108:2322–2327.

Hulbert SH, Bennetzen JL. 1991. Recombination at the Rp1 locus of maize. *Mol. Gen. Genet.* 226:377-382.

Hulbert SH, Webb CA, Smith SM, Sun Q. 2001. Resistance gene complexes: Evolution and utilization. *Annu. Rev. Phytopathol.* 39:285-312.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Scie. USA.* 103:16666-16671.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack JH, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV, Boore JL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA.* 104:19369–19374.

Jia Y, McAdams SA, Bryan GT, Hershey HP, Valent B. 2000. Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO J.* 19:4004-4014.

Jones DG, Dangl JL. 2006. The plant immune system. *Nature* 444:323-329.

Jones GH. 1984. The control of chiasma distribution. *SEB Symposium.* 38: 293–320.

Jones GH, Franklin FC. 2006. Meiotic crossing-over: obligation and interference. *Cell.* 126: 246–248.

Jones L, Rybka K, Lukaszewski A. 2002. The effect of a deficiency and a deletion on recombination in chromosome 1BL in wheat. *Theor. Appl. Genet.* 104:1204-1208.

Jupe F, Pritchard L, Etherington GJ, MacKenzie K, Cock PJA, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JDG, et al. 2012. Identification and localization of the NB-LRR gene family within the potato genome. *BMC Genomics*. 13:75.

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

Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Phil. Trans. R. Soc. Lond. B. Biol. Sci*. 365:729-748.

Keeney S, Giroux CN, Kleckner N. 1997. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* 88: 375–384.

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol. Evol.* 24:572-582.

Kidwell MG. 2002. Transposable element and the evolution of genome size in eukaryotes. *Genetica* 115:49-63.

Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animal and plants. *Proc. Natl. Acad. Sci. USA.* 94:7704–7711.

Knox EB, Palmer JD. 1992. The chloroplast genome arrangement of *Lobelia thuliniana* (*Lobeliaceae*): Expansion of the inverted repeat in an ancestor of the *Campanulales*. *Plant Syst. Evol.* 214:49-64.

Kolodner R, Tewari KK. 1979. Inverted repeats in chloroplast DNA from higher plants. *P. Natl. Acad. Sci. USA.* 76:41-45.

Kountche BA, Hash CT, Dodo H, et al. 2013. Development of a pearl millet *Striga*-resistant genepool: Response to five cycles of recurrent selection under *Striga*-infested field conditions in West Africa. *Field Crop. Res.* 154:82-90.

Krause K. 2008. From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. *Curr. Genet.* 54:111-121.

Krejci L, Atlmannova V, Spirek M, Zhao X. 2012. Homologous recombination and its regulation. *Nucl. Acids Res.* 40:5795-5818.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0. *Mol. Biol. Evol.* 33:1870-1874.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

Leberg PL. 1992. Effects of population bottlenecks on genetic diversity as measured by allozyme electrophoresis. *Evolution*. 46:477-494.

Leflon M, Grandont L, Eber F, Huteau V, Coriton O, Chelysheva L, Jenczewski E, Chevre AM. 2010. Crossovers get a boost in *Brassica* allotriploid and allotetraploid hybrids. *Plant Cell.* 22:2253-2264.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.* 20:116-122.

Leukal RW. 1948. *Periconia circinata* and its relation to milo disease. *J. Agric. Res.* 77:201-222.

Li J, Hsia A, Schnable PS. 2007. Recent advances in plant recombination. *Curr. Opin. Plant Biol.* 10:131-135.

Li W, Zhang P, Fellers JP, Friebe B, Gill BS. 2004. Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J.* 40:500-511.

Li X, Cheng Y, Ma W, Zhao Y, Jiang H, Zhang M. 2010. Identification and characterization of NBS-encoding disease resistance genes in *Lotus japonicus*. *Plant Syst. Evol.* 289:101-110.

Liu R. 2005. The evolution of gene composition in angiosperms. *PhD Dissertation*. (Athens, GA: University of Georgia)

Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL. 2007. A GeneTrek analysis of the maize genome. *Proc. Natl. Acad. Sci. USA.* 104:11844-11849.

Logacheva MD, Schelkunov MI, Shtratnikova VY, Matveeva MV, Penin AA. 2016. Comparative analysis of plastid genomes of non-photosynthetic Ericaceae and their photosynthetic relatives. *Sci. Report.* doi: 10.1038/srep30042.

Lohan AJ, Wolfe KH. 1998. A subset of conserved tRNA genes in plastid DNA of nongreen plants. *Genetics*. 150:425–433.

Lopez-Sanchez H. 2005. Assessing corn pollen flow and outcross in seed and grain production fields. *Retrospective Theses and Dissertations*. 1578.

Lorang JM, Sweat TA, Wopert TJ. 2007. Plant disease susceptibility conferred by a “resistance” gene. *Proc. Natl. Acad. Sci. USA* 104:14861-14866.

Lynch M. 2007. Mobile Genetic Elements. In:. Lynch M. editor. The Origin of Genome Architecture Sunderland: Sinauer Associates; p. 151-192.

Ma J, Bennetzen JL. 2006. Recombination, rearrangement, reshuffling and divergence in a centromeric region of rice. *Proc. Natl. Acad. Sci. USA.* 103:383-388.

Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genome DNA loss in rice. *Genome Res.* 14:860-869.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* 10:1307-1308.

Maloisel L, Rossignol JL. 1998. Suppression of crossing-over by DNA methylation in *Ascobolus*. *Genes Dev.* 12:1381–1389.

Malone RE, Bullard S, Lundquist S, Kim S, Tarkowski T. 1992. A meiotic gene conversion gradient opposite to the direction of transcription. *Nature.* 359:154–155.

Mann DG, Lafayette PR, Abercombie LL, King ZR, Mazarei M, Halter MC, Poovaiah CR, Baxter H, Shen H, Dixon RA, et al. 2012. Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (Panicum virgatum L.) and other monocot species. *Plant Biotechnol. J.* 10:226-236.

Maréchal A, Parent J, Véronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. 2009. Whirly proteins maintain plastid genome stability in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA.* 106:14693-14698.

Martín M, Funk HT, Serrot PH, Poltnigg P, Sabater B. 2009. Functional characterization of the thylakoid Ndh complex phosphorylation by site-directed mutations in the *ndhF* gene. *BBA-Bioenergetics.* 1787:920-928.

Martín M, Sabater B. 2010. Plastid *ndh* genes in plant evolution. *Plant Physiol. Biochem.* 48:636-645.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA*. 99:12246-12251.

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

McDowell JM, Simon SA. 2006. Recent insights into R gene evolution. *Mol. Plant Pathol.* 7:437-448.

McFadden GI, van Dooren GG. 2004. Evolution: red algal genome affirms a common original of all plastids. *Curr. Biol.* 14:R514:516.

McHale L, Tan X, Koehl P, Michelmore RW. 2006. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* 7:212.

McMullen MD, Kresovich S, Sanchez Villeda H, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thomsberry J, Acharya C, Bottoms C, et al. 2009. Genetic properties of the maize nested association mapping population. *Science*. 325:737-740.

McNeal JR, Bennet JR, Wolfe AD, Mathews S. 2013. Phylogeny and origins of holoparasitism in Orobanchaceae. *Am. J. Bot*. 100:971-983.

McNeal JR, Kuehl J, Boore J, dePamphilis C. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol.* 7:57.

Melamed-Bessudo C, Levy AA. 2012. Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proc. Natl. Acad. Sci. USA.* 109:E981–E988.

Melamed-Bessudo C, Shilo S, Levy AA. 2016. Meiotic recombination and genome evolution in plants. *Curr. Opin. Plant Biol.* 30:82-87.

Mercier R, Jolivet S, Vezon D, Huppe E, Chelysheva L, Giovanni M, Nogue F, Doutriaux MP, Horlow C, Grelon M, et al. 2005. Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on Mer3, whereas the other one is not. *Curr. Biol.* 15:692–701.

Mercier R, Mézard C, Jenczewski E, Macaisne N, Grelon M. 2015. The molecular biology of meiosis in plants. *Annu. Rev. Plant Biol.* 66:297-327.

Michelmore RW, Meyers BC. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* 8:1113-1130.

Miguel C, Simoes M, Oliveira MM, Rocheta M. 2008. Envelope-like retrotransposons in the plant kingdom: evidence of their presence in gymnosperms (*Pinus pinaster*). *J. Mol. Evol*. 67:517-525.

Mirouze M, Lieberman-Lazarovich M, Aversano R, Bucher E, Nicolet J, Reinders J, Paszkowski J. 2012. Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proc. Natl. Acad. Sci. USA.* 109:5880–5885.

Mohamed KI, Bolin JF, Musselman LJ, Peterson AT. 2007. Genetic diversity of *Striga* and implication for control and modeling future distributions. In: Ejeta G, Gressel J, editors.

Integrating new technologies for Striga control: towards ending the witch-hunt. World Scientific Publishing Co. Pte. Ltd, Singapore. p. 71-84.

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

Morel JB, Dangl JL. 1997. The hypersensitive response and the induction of cell death in plants. *Cell. Death Differ.* 4:671-683.

Murat F, Van de Peer Y, Salse J. 2012. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol. Evol.* 4:917-928.

Nagy ED, Bennetzen JL. 2008. Pathogen corruption and site-directed recombination at a plant disease resistance gene cluster. *Genome Res.* 18:1918-1923.

Nagy ED, Lee T, Ramakrishna W, Xu Z, Klein PE, SanMiguel P, Cheng C, Li J, Devos KM, Schertz K, Dunkle L, Bennetzen JL. 2007. Fine mapping of the *Pc* locus of *Sorghum bicolor*, a gene controlling the reaction to a fungal pathogen and its host-selective toxin. *Theor. Appl. Genet*. 114:961-970.

Nakkaew A, Chotigeat W, Eksomtramage T, Phongdara A. 2008. Cloning and expression of a plastid-encoded subunit betacarboxyltransferase gene (*accD*) and a nuclear-encoded subunit biotin carboxylase of acetyl-CoA carboxylase from oil palm (*Elaeis guineensis* Jacq.). *Plant Sci*. 175:497–504.

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*. 461:1130-1134.

Naumann J, Salomo K, Der JP, Wafula EK, Bolin JF, Maass E, Frenzke L, Samain MS, Neinhuis C, dePamphilis CW, Wanke S. 2013. Single-copy nuclear genes place haustorial

Hydnoraceae within Piperales and reveal a Cretaceous origin of multiple parasitic angiosperm lineages. *Plos One.* doi: 10.1371/journal.pone.0079204.

Nickrent DL, Ouyang R, Joel D, dePamphilis CW. 1997. Do nonasterid holoparasitic flowering plants have plastid genomes? *Plant Mol. Biol.* 34:717-729.

Nixon PJ. 2000. Chlororespiration. *Philos. Trans. R. Soc. B. Biol. Sci.* 355:1541–1547.

Noél, L, Moores, TL, Van Der Biezen A, Parniske M, Daniels MJ. 1999. Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. *Plant Cell.* 11:2099-2111.

Noma K, Nakajima R, Ohtsubo H, Ohtsubo E. 1997. RIRE1, a retrotransposon from wild rice *Oryza australiensis*. *Genes Genet. Syst.* 72:131–140.

Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M. 2013. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* 41:D1144-1151.

Okagaki RJ, Weil CF. 1997. Analysis of recombination sites within the maize *waxy* locus. *Genetics*. 147:815-821.

Olivier A, Glaszmann JC, Lanaud C, Leroux GD. 1998. Population structure, genetic diversity and host specificity of the parasitic weed *Striga hermonthica* (Scrophulariaceae) in Sahel. *Plant Syst. Evol.* 209:33-45.

Oliver M, Murdock A, Mishler BD, Kuehl J, Boore J, Mandoli D, Everett K, Wolf PG, Duffy A, Karol KG. 2010. Chloroplast genome sequence of the moss *Tortula ruralis*: gene content polymorphism and structural arrangement relative to other green plant chloroplast genomes. *BMC Genomics*. 11:143.

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

Opperman R, Emmanuel E, Levy AA. 2004. The effect of sequence divergence on recombination between direct repeats in Arabidopsis. *Genetics.* 168:2207-2215.

Osman K, Higgins JD, Sanchez-Moran E, Armstrong SJ, Franklin CH. 2011. Pathways to meiotic recombination in *Arabidopsis thaliana*. *New Phytol.* 190: 523-544.

Palmer JD. 1985. Comparative organization of chloroplast genome. *Annu. Rev. Genet.* 19:325-354.

Palmer, JD. 1987. Chloroplast DNA evolution and biosystematics uses of chloroplast DNA variation. *Am. Nat.* 130:S6-S29.

Palmer JD. 1991. Plastid chromosomes: structure and evolution. In: Bogorad L, Vasil IK, editors. Cell culture and somatic genetics of plant, vol 7A. Molecular biology of plastids. San Diego: Academic Press; p. 5-53.

Palmer JD. 2000. Molecular evolution: a single birth of all plastids? *Nature.* 405:32-33.

Panchy N, Lehti-Shiu M, Shiu S. 2016. Evolution of gene duplication in plants. *Plant Physiol.* 171:2294-2316.

Parker C. 2013. The parasitic weeds of the Orobanchaceae. In: Joel, DM, Gressel, J., Musselman, LJ, editors. *Parasitic Orobanchaceae: Parasitic mechanisms and control strategies*. Springer-Verlag, Berlin, Heidelberg. p. 313-344.

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K< Wulff BBH, Jones JDG. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. *Cell.* 91:821-832.

Parniske M, Jones JDG. 1999. Recombination between diverged clusters of the tomato *Cf-9* plant disease resistance gene family. *Proc. Natl. Acad. Sci. USA.* 96:5850-5855.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009.The *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 457:551-556.

Patterson GI, Kubo KM, Shroyer T, Chandler VL. 1995. Sequences required for paramutation of the maize b gene map to a region containing the promoter and upstream sequences. *Genetics*. 140:1389-1406.

Pawlowski WP, Golubovskaya IN, Cande WZ. 2003. Altered nuclear distribution of recombination protein RAD51 in maize mutants suggests the involvement of RAD51 in meiotic homology recognition. *Plant Cell*. 15:1807-1816.

Peakall R, Smouse PE. 2012. GenAlEx 65: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics.* 28:2537-2539.

Pecinka A, Fang W, Rehmsmeier M, Levy AA, Mittelsten-Scheid O. 2011. Polyploidization increases meiotic recombination frequency in Arabidopsis. *BMC Biol.* 9:24.

Pellicer J, Fay M, Leitch I. 2010. The largest eukaryotic genome of them all. *Bot J. Linn. Soc.* 164:10-15.

Perry AS, Wolfe KH. 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J. Mol. Evol.* 55:501–508.

Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16:1262-1269.

Plunkett GM, Downie SR. 2000. Expansion and contraction of the chloroplast inverted repeat in Abiaceae subfamily Apioideae. *Syst. Bot*. 25:648-667.

Pritchard JK, Stephens M, Donnely P. 2000. Inference of population structure using multilocus genotype data. *Genetics.* 155:945-959.

Puchta H, Hohn B. 1991. A transient assay in plant cells reveals a positive correlation between extrachromosomal recombination and length of homologous overlap. *Nuc. Acids Res.* 19:2693-2700.

Qi LL, Friebe B, Gill BS. 2002. A strategy for enhancing recombination in proximal regions of chromosomes. *Chromosome Res.* 10:645-654.

Raubeson LA, Jansen RK. 2005. Chloroplast genomes of plants. In: Henry RJ, editor. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Wallingford, UK: CABI Publishing p. 45–68.

Richter TE, Pryor TJ, Bennetzen JL, Hulbert SH. 1995. New rust resistance specificities associated with recombination in the Rp1 complex in maize. *Genetics.* 141:373-381.

Rodgers-Melnick E, Bradbury PJ, Elshirea RJ, Glaubitza JC, Acharyaa CB, Mitchella SE, Lic C, Lic Y, Buckler ES. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Sci. USA.* 112:3823-3828.

Román B. 2013. Population diversity and dynamics of parasitic weeds. In: Joel, DM, Gressel, J., Musselman, LJ, editors. *Parasitic Orobanchaceae: Parasitic mechanisms and control strategies.* Springer-Verlag, Berlin, Heidelberg. p. 345-356.

Román B, Rubiales D, Torres AM, Cubero JI, Satovic Z. 2001. Genetic diversity in *Orobanche crenata* populations from southern Spain. *Theoret. Appl. Genet.* 103: 1108-1114.

Ronquist F, Huelsenback JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19:1572-1574.

Samigullin TH, Logacheva MD, Penin AA, Vallejo CM. 2016. Complete plastid genome of the recent holoparasites *Lathraea squamaria* reveals earliest stages of plastome reduction in Orobanchaceae. *Plos One* doi:10.1371/journal.pone.0150718.

Sanchez-Moran E, Santos JL, Jones GH, Franklin FCH. 2007. ASY1 mediates AtDMC1-dependent interhomolog recombination during meiosis in Arabidopsis. *Gene. Dev.* 21:2220-2233.

Sandhu D, Gill KS. 2002. Gene-containing regions of wheat and the other grass genomes. *Plant Physiol.* 128:803–811.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* 20:43-45.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science.* 274:765–768.

SanMiguel P, Vitte, C. 2009. The LTR-retrotransposons of maize. In: Bennetzen JL and Hake S, editors. Maize handbook – volume II: genetics and genomics. Springer, New York, pp. 307-327.

Sanz-Alferez S, Richter TE, Hulbert SH, Bennetzen JL. 1995. The *Rp3* disease resistance gene of maize: mapping and characterization of introgressed alleles. *Theor. Appl. Genet.* 91:25-32.

Schertz KF, Tai YP. 1969. Inheritance of reaction of *Sorghum bicolor* (L.) Moench to toxin produced by *Periconia circinata* (Mang.) sacc. *Crop Sci*. 9:621-624.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 326:1112-1115.

Schuelke M. 2000. An economic method for the florescent labeling of PCR fragments. *Nat Biotechno.* 18:233-234Schuermann D, Molinier J, Fritsch O, Hohn B. 2005. The dual nature of homologous recombination in plants. *Trends Genet.* 21:172-181.

Schultes NP, Szostak JW. 1990. Decreasing gradients of gene conversion on both sides of the initiation site for meiotic recombination at the *ARG4* locus in yeast. *Genetics.* 126:813-822.

Sentry JW, Smyth DR. 1989. An element with long terminal repeats and its variant arrangements in the genome of *Lilium henryi. Mol. Gen. Genet.* 215:349–354.

Shepherd NS, Schwarz-Sommer Z, Blumberg J, Gupt M, Wienand U, Saedler H. 1984. Similarity of the *Cin1* repetitive family of *Zea mays* to eukaryotic transposable elements. *Nature*. 307:185-187.

Shinohara M, Oh SD, Hunter N, Shinohara A. 2008. Crossover assurance and crossover interference are distinctly regulated by the ZMM proteins during yeast meiosis. *Nat. Genet.* 40: 299–309.

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* 10:908–915.

Si W, Yuan Y, Huang J, Zhang X, Zhang Y, Zhang Y, Tian D, Wang C, Yang Y, Yang S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytol.* 206:1491-1502.

Sidhu GK, Warzecha T, Pawlowski WP. 2017. Evolution of meiotic recombination genes in maize and teosinte. *BMC Genomics.* 18:106.

Sidorov VA, Kasten D, Pang SZ, Hajdukiewicz PTJ, Staub JM, Nehra NS. 1999. Stable chloroplast transformation in potato: use of green fluorescent protein as a plastid marker. *Plant J*. 19:209-216.

Sikdar SR, Serino G, Chaudhuri S, Maliga P. 1998. Plastid transformation in *Arabidopsis thaliana*. *Plant Cell Rep.* 18:20-24.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.

Smith JD, Bickham JW, Gregory TR. 2013. Patterns of genome size diversity in bats (order Chiroptera). *Genome.* 56:457-472.

Snapp E. 2010. Design and use of fluorescent fusion proteins in cell biology. *Curr. Protoc. Cell Biol.* doi: 10.1002/0471143030.cb2104s27

Song R, Messing J. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci. USA*. 100:9055–9060.

Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG. 1994. An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev*. 8:2046-2057.

Spielman D, Brook BW, Briscoe DA, Frankham R. 2004. Does inbreeding and loss of genetic diversity decrease disease resistance? *Conserv. Genet.* 5:439-448.

Stacey NJ, Kuromori T, Azumi Y, Roberts G, Breuer C, Wada T, Maxwell A, Roberts K, Sugimoto-Shirasu K. 2006. Arabidopsis SPO11- 2 functions with SPO11-1 in meiotic recombination. *Plant J.* 48: 206–216.

Staskawicz BJ, Ausubel FM, Baker BJ, Ellis JG, Jones DG. 1995. Molecular genetics of plant disease resistance. *Science* 268:661-667.

Stephan W, Langley CH. 1998. DNA polymorphism in Lycopersicon and crossing-over per physical length. *Genetics.* 150:1585-1593.

Stoebe B, Kowallik KV. 1999. Gene-cluster analysis in chloroplast genomics. *Trends Genet*. 15:344–347.

Sudupak MA, Bennetzen JL, Hulbert SH. 1993. Unequal exchange and meiotic instability of disease-resistance genes in the Rp1 region of maize. *Genetics*. 133:119-125.

Sugiura M. 1992. The chloroplast genome. *Plant Mol. Biol*. 19:149-168.

Sun Q, Collins NC, Ayliffe M, Smith SM, Drake J, Pryor T, Hulbert SH. 2001. Recombination between paralogues at the Rp1 rust resistance locus in maize. *Genetics* 158:423-428.

Sun Y, Ambrose JH, Haughey BS, Webster TD, Pierrie SN, Muñoz DF, Wellman EC, Cherian S, Lewis SM, Berchowitz LE, Copenhaver GP. 2012. Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in *Arabidopsis thaliana*. *PLoS Genet.* 8:e1002968.

Svab Z, Hajdukiewicz P, Maliga P. 1990. Stable transformation of plastids in higher plants. *Proc. Natl. Acad. Sci. USA.* 87:8526-8530.

Swigonová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res.* 14:1916-1923.

Tan S, Wu S. 2012. Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. *Comp. Funct. Genomics.* DOI: 10.1007/s11434-014-0155-3.

The International Barley Genome Sequencing Consortium. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature*. 491:711.

Thomas CA. 1971. The genetic organization of chromosomes. *A. Rev. Genet.* 5:237-256.

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein ND, Bennetzen JL, Avramova Z. 1999. Colinearity and its exceptions in orthologous *ADH* regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA.* 96:7409–7414.

Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, Sugiura M. 1992. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnQ trnK psbA trnI* and *trnH* and the absence of *rps16*. *Mol. Gen. Genet.* 232:206-214.

Varas J, Sanchez-Moran E, Copenhaver GP, Santos JL, Pradillo M. 2015. Analysis of the relationships between DNA double-strand breaks, synaptonemal complex and crossovers using the Atfas1-4 mutant. *PLoS Genet*. 11:e1005301.

Vaughn JN, Chaluvadi SR, Tushar, Rangan L, Bennetzen JL. 2014. Whole plastome sequences from five ginger species facilitate marker development and define limits to barcode methodology. *Plos One* doi:10.1371/journal.pone.0108581.

Vitte C, Panaud O. 2003. Formation of Solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol. Biol. Evol.* 20:528–540.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*. 28:913-922.

Wagner HW, Sefc KM. 1999. IDENTITY 1.0 Centre for Applied Genetics, University of Agricultural Sciences, Vienna.

Wahls WP, Davidson MK. 2012. New paradigms for conserved, multifactorial, *cis*-acting regulation of meiotic recombination. *Nucleic Acids Res.* 40:9983-9989.

Wang H, Chong S. 2003. Visualization of coupled protein folding and binding in bacteria and purification of the heterodimeric complex. *Proc. Natl. Acad. Sci. USA.* 100:478-483.

Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC. Evol. Biol*. 8:36.

Wani SH, Haider N, Kumar H, Singh NB. 2010. Plant plastid engineering. *Curr. Genomics* 11:500-512.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature.* 453:175-183.

Webb CA, Richter TE, Collins NC, Nicolas M, Trick HN, Pryor T, Hulbert SH. 2002. Genetic and molecular characterization of the maize rp3 rust resistance locus. *Genetics.* 162:381-394.

Welsh AB, Mohamed KI. 2011. Genetic diversity of *Striga hermonthica* populations in Ethiopia: evaluating the role of geography and host specificity in shaping population structure. *Int. J. Plant Sci.* 172:773-782.

Wendel JF, Jackson SA, Meyers BC, Wing, RA. 2016. Evolution of plant genome architecture. *Genome Biology.* 17:37.

Wessler S. 2006. Transposable elements and the evolution of eukaryotic genomes. *Proc. Natl. Sci. Acad. USA.* 103:17600-17601.

Westwood JH, Yoder JI, Timko MP, dePamphilis CW. 2010. The evolution of parasitism in plants. *Trends Plant Sci*. 15:227-235.

Wicke S, Müller KF, dePamphilis CW, Quandt D, Bellot S, Schneeweiss GM. 2016. Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. *P. Natl. A. Sci.* 113:9045-9050.

Wicke S, Müller KF, dePamphilis CW, Quandt D, Wickett NJ, Zhang Y, Renner SS, Schneeweiss GM. 2013 Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *Plant Cell.* 25:3711-3725.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content gene order gene function. *Plant Mol. Biol.* 76:273-297.

Wickett NJ, Zhang Y, Hansen SK, Roper JM, Kuehl JV, Plock SA, Wolf PG, dePamphilis CW, Boore JL, Goffinet B. 2008. Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. *Mol. Biol. Evol*. 25:393-401.

Wilson JP, Hess DE, Hanna WW. 2000. Resistance to *Striga hermonthica* in wild accessions of the primary gene pool of *Pennisetum glaucum*. *Phytopathology*. 90:1169-1172.

Wilson JP, Hess DE, Hanna WW, Kumar KA, and Gupta SC. 2004. *Pennisetum glaucum* subsp. *monodii* accessions with *Striga* resistance in West Africa. *Crop Prot.* 23:865-870.

Wimpee CF, Wrobel RL, Garvin DK. 1991. A divergent plastid genome in *Conopholis americana*, an achlorophyllous parasitic plant. *Plant Mol. Biol*. 17:161–166.

Wolf PG, Der J, Duffy A, Davidson J, Grusz A, Pryer KM. 2010. The evolution of chloroplast genes and genomes in ferns. *Plant Mol. Biol.* doi:10.1007/s11103-010-9706-4.

Wolfe AD, Randle CP, Liu L, Steiner KE. 2005. Phylogeny and biogeography of *Orobanchaceae*. *Folia Geobot.* 40:115-134.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci USA*. 84:9054-9058.

Wolfe KH, Morden CW, Ems SC, Palmer JD. 1992a. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: Loss or accelerated sequence evolution of tRNA and ribosomal protein genes. *J. Mol. Evol*. 35: 304–317.

Wolfe KH, Morden CW, Palmer JD. 1992b. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl. Acad. Sci. USA*. 89:10648–10652.

Wolfe PG, Roper JM, Duffy AM. 2010. The evolution of chloroplast genome structure in ferns. *Genome* 53:731-738.

Wright S. 1950. Genetical structure of populations. *Nature.* 166: 247-249.

Wu C, Lai Y, Lin C, Wang Y, Chaw S. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. *Mol. Phylogenet. Evol.* 52:115-124.

Wu F, Chan M, Liao D, Hsu C, Lee Y, Daniell H, Duvall M, Lin C. 2010. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biol*. 10:68.

Xu X, Hsia AP, Zhang L, Nikolau BJ, Schnable PS. 1995. Meiotic recombination break points resolve at higher rates at the 5’ end of a maize coding sequence. *Plant Cell.* 12:2151-2161.

Yandeau-Nelson MD, Xia Y, Li J, Neuffer MG, Schnable PS. 2006. Unequal sister chromatid and homolog recombination at a tandem duplication of the *a1* locus in maize. *Genetics.* 173:2211-2226.

Yanisch-Perron C, Vieira J, Messing J. 1985. Improved M13 phage cloning vectors and ... sequences of the M13mp18 and pUC19 vectors. *Gene*. 33:103–119.

Yao H, Schnable PS. 2005. Cis-effects on meiotic recombination across distinct *a1-sh2* intervals in a common Zea genetic background. *Genetics.* 170:1929–1944.

Yelina NE, Lambing C, Hardcastle TJ, Zhao X, Santos B, Henderson IR. 2015. DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. *Genes Dev*. 29:2183–2202.

Young ND. 2000. The genetic architecture of resistance. *Curr. Opin. Plant Biol.* 3:285-290.

Young ND, dePamphilis CW. 2005. Rate variation in parasitic plants: Correlated and un-correlated patterns among plastid genes of different function. *BMC Evol. Biol.* 5:16.

Zhang L, Yan L, Jiang J, Want Y, Jian Y, Yan T, Cao Y. 2014. The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*. *Virulence*. 15:655-664.

Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol*. 209:1747-1756.

APPENDIX A

CONTINENT-SPANNING GENE FLOW OF THE PARASITIC WITCHWEED, *STRIGA HERMONTHICA*[4]

[4] Frailey, D.C., Estep, M.C., Van Mourik, T.A., Chaluvadi, S.R., Hash, C.T., Abraha, N., Weltzien, E. Bennetzen, J.L. To be submitted to *Annals of Botany*.

Abstract

Parasitic weeds of the genus *Striga* are the most severe biotic limitation to cereal production in Africa, and *Striga hermonthica* is the most devastating of these witchweeds. As a step toward finding possible genetic approaches for controlling *Striga*, we investigated the nature of *S. hermonthica* genetic diversity on several crop species in a broad east-west transect across north-central Africa.

A total of 120 samples were collected from Eritrea, Mali, Niger and Senegal. Twelve simple sequence repeat (SSR) markers were employed to assess sequence polymorphism, on samples collected from pearl millet, sorghum, maize or rice hosts.

Key Results: As previously observed in Africa for this outcrossing species, genetic diversity and heterozygosity were high in all *S. hermonthica* samples collected.

Surprisingly, all samples were best explained as members of a single population with near-unlimited gene flow. That is, neither location of collection nor host-species were associated with any population distinctions.

These results indicate that, at least across this east-west dimension, *S. hermonthica* populations quickly move toward equilibrium, indicating that any virulence genes active against specific host resistance loci will be rapidly disseminated across the Sahelian *Striga* germplasm.

Introduction

The genus *Striga* contains over 40 species of parasitic plants, commonly known as witchweeds, that range across Africa and into the Arabian Peninsula and southern Asia. Eleven of these species are parasites of agricultural crops, and are particularly problematic in Africa. *Striga* infects many tropical and subtropical cereals, including sorghum, pearl millet, maize, rice,

and finger millet (Ejeta, 2007). *Striga gesnerioides* is the only *Striga* species that is a serious parasite on dicotyledonous crops, with devastating effects on cowpea production in Africa (Mohamed and Musselman, 2008; Parker, 2013). *Striga* infestation significantly reduces grain yield in areas of infestation, with losses often estimated above 50% in sub-Saharan Africa. It can also result in total crop loss when combined with adverse environmental conditions such as drought. Because *Striga* parasitizes the food crops of many poor farmers in developing countries of Africa and Asia, it has an enormous impact on food security. Estimates of the scale of infested regions in Africa range from 50 to 300 million hectares (Ejeta, 2007). In most of Africa, *S. hermonthica* is the single most damaging species of witchweed (Parker, 2013).

Striga hermonthica is an obligate outcrosser, and a single plant is capable of producing 50,000-200,000 seeds (Parker and Riches, 1993). Seed can remain dormant for years, leading to enormous seed reservoirs in the soil. These reserves provide both a source of ongoing crop infestation and a reservoir for *S. hermonthica* genetic diversity, thus optimizing the ability of this pest to rapidly adapt to new host plants or environmental conditions (Román, 2013).

S. hermonthica has the widest distribution of any of the *Striga* species. For all *Striga* species, wide dispersal of the tiny seed by wind and water is a major problem. For *S. hermonthica*, perhaps the most important factor contributing to seed dispersal is the exchange among farmers of crop seeds contaminated with witchweed seed (Ejeta, 2007). It has been estimated that, in Africa, 20-40% of all cereal seed is infested (Berner et al., 1994). Hence, high genetic diversity (Estep et al., 2011) and long-range dispersal through contaminated crop seeds lead to high gene flow across great distances (Román et al., 2001; Román, 2013). Thus, it is not surprising that one previous study indicated that *S. hermonthica* across Mali is essentially one

population, with a tiny amount of differentiation that was better explained by human vectoring of the seed than it was by ecological adaptation (Estep et al., 2011).

S. hermonthica parasitizes a range of crops, including sorghum, pearl millet, finger millet, maize, rice, tef millet and sugar cane. It is possible that different populations of *S. hermonthica* are adapted to different host species. If this is the case, then one expects portioning of the *S. hermonthica* genetic diversity across different hosts, a question that has not yet been thoroughly investigated.

A powerful method for studying genetic diversity is by use of DNA markers, and Simple Sequence Repeats (SSRs) have been particularly useful because they are co-dominant and often neutrally evolving. Such markers have already been developed for *Striga* (Estep et al., 2010) and used in a previous study to examine genetic diversity of *S. hermonthica* populations across Mali (Estep et al., 2011). Our study uses these markers to investigate the genetic diversity of populations of *S. hermonthica* collected from multiple host species and across a wider range of Africa, including Senegal, Mali, Niger, and a single population from Eritrea. The results of these investigations indicate an even higher level of gene flow than previously predicted.

Materials and Methods

Sample Collection

Plant, leaf, and stem tissues were collected from *S. hermonthica* growing in farmers' fields in Senegal, Mali, Niger and Eritrea. A total of 120 individuals from 14 populations were collected; including 58 samples from six populations from three regions in Mali, 35 samples from five populations from three regions in Senegal, 18 samples from two populations from one region in Niger, and nine samples from one population in Eritrea. Each *Striga* individual was

collected from a different parasitized host plant. Seven of the sampled populations were found in pearl millet fields, five populations were collected from sorghum fields, and one population each was collected from maize or rice fields. Figure A.1 shows where the Central and W. African samples were collected. Table A.1 shows how many samples from each population were acquired.

DNA extraction and marker amplification

Leaf, stem, and/or flower tissue was collected from each sample. The tissue was dried for storage until the DNA was extracted. DNA was extracted using the QIAgen DNeasy 96 Plant Kit. Each DNA sample was investigated at 12 SSR loci. The 12 SSRs employed were previously published (Estep et al., 2010) and are described in Table A.2. The PCR reactions employed M13-tailed universal fluorescent-labeled primers HEX, NED, and FAM (Schuelke, 2000), to facilitate automated analysis. Each PCR reaction consisted of 5-20ng template DNA, 2µL Q5 Reaction Buffer, 2µL Q5 High GC Enhancer, .25µL 10mM dNTPs, .05µL forward primer, .5µL reverse primer, .25µL M13 fluorescent-labeled primer, .1µL Q5 High-Fidelity DNA Polymerase, and 2.85µL water. Touchdown PCR was performed on each reaction on a PTC-200 Peltier Thermal Cycler (MJ Research, Ramsey, MN, USA). The program used was 98°C for 30sec followed by 10 cycles consisting of denaturation at 98°C for 20sec, annealing at 65°C (-1°C each cycle) for 15sec, elongation at 72°C for 25sec. Next, there were 38 cycles consisting of denaturation at 98°C for 15sec, 58°C for 15sec, elongation at 72°C for 25sec; and a final elongation step at 72°C for 10min. After PCR, 1500µL of formamide and 1µL of 500 ROX dye size standard was added to each sample. The samples were sent to the Georgia Genomics Facility where they were genotyped according to size using gel electrophoresis on an ABI 3730 sequencer (Applied

Biosystems, Foster City, CA, USA). The chromatograms were then scored with Softgenetics Genemarker version 2.4.

Data Analysis

GenAlEx (Peakall and Smouse, 2012), Identity (Wagner and Sefc, 1990), and Microsat 2.0 (http://genetics.stanford.edu/hpgl/projects/microsat/) were used to calculate various population genetic statistics. GenAlEx was used to calculate the number of individuals included per locus, Fst values (Wright, 1950), Rst values (Slatkin, 1995), mean number of alleles per population, mean effective number of alleles per population, mean expected heterozygosity per population, mean observed heterozygosity per population, and mean index of fixation per population. Identity and Microsat 2.0 were used to calculate mean expected heterozygosity per locus, mean observed heterozygosity per locus, and the frequency of null alleles per locus. GenAlEx was also used to perform an analysis of molecular variance (AMOVA) (Excoffier et al., 1992) and a principal coordinates analysis (PCoA).

STRUCTURE (Pritchard et al., 2000) was used to further look for evidence of population structure. An initial burn-in period of 50,000 replicates and 50,000 Markov Chain Monte Carlo (MCMC) iterations were used. The predefined number of populations (K) ranged from 2 to 21 and ten simulations were run for each K value. The average K value from the ten runs was taken and used by Structure Harvester (Evanno et al., 2005) to calculate the Delta K value. K = 14 had the highest Delta K value (19.1).

Results

Striga samples were collected from three regions of Mali, three regions of Senegal, and one region each of Niger and Eritrea from October to November of 2009. Leaf, stem, and/or flower tissue was collected for each sample.

All 120 samples belonging to the 14 populations were scored at each of the 12 SSR markers, with an average success rate of 114 samples per PCR locus. A combined total of 118 alleles were found for the 12 loci (Table A.2). The number of alleles per locus ranged from 6 to 14, with the average number of alleles per locus being 9.8. The mean number of alleles per locus ranged from 3.3 in the population Senegal_PM2 to 5.8 in population Mali_PM1. The mean number of effective alleles ranged from 2.4 in population Mali_PM3 to 3.6 in population Niger_PM. The mean expected heterozygosity ranged from .56 in population Mali_PM3 to .69 in population Niger_PM. The mean observed heterozygosity ranged from .63 in Mali_R to .84 in Niger_S. All 14 populations had a negative mean fixation index, calculated by Identity 1.0, indicating that there is a higher level of heterozygosity than expected. All values are shown in Table A.3.

An analysis of molecular variation (AMOVA) was done using GenAlEx to calculate the Rst, Ris, and Rit values (Figure A.2). Rit, Ris, and Rst are analogs of Fit, Fis, and Fst, respectively, that are adapted for microsatellite markers. Rit is the amount of inbreeding within individuals with respect to the total variation of all populations. Ris is the amount of inbreeding within individuals with respect to their individual populations, and Rst is the amount of inbreeding within a subpopulation with respect to the total variation of all populations. The Rst value for the populations was .016. The Ris value was .57 and the Rit value was .57. Hence, only

1% of the genetic difference was found to be due to among population variation, while 43% was due to within individual variation and 56% was due to among individual variation.

A principal coordinates analysis (PCoA) was performed to further look for variation that might correlate with host specificity or location of collection (Figure A.3). Coordinate 1 explained 9.9% of the variation, coordinate 2 explained 6.8% of the variation, and coordinate 3 (not shown) explained 5.5% of the variation. The total variation explained by these components was ~22%. The samples did not indicate any major clustering by host or location of collection. The SSR genotyping data for *Striga* accessions was also used to calculate pairwise chord genetic distances between all genotypes and thereby generate a dissimilarity matrix from which a minimum evolution tree was calculated. The relatedness tree based on chord genetic distances depicted in Figure A.4 did not show significant clustering based on country of collection and host species, but two clusters did show some enrichment for *Striga* accessions collected on pearl millet.

The program STRUCTURE was run to try to identify population structure among the samples. Structure Harvester was used to determine which K value was the best fit for the data. K = 14 had the highest Delta K value, but was not substantially better than several other K values, such as 3, 4 and 16 (Figure A.6A). STRUCTURE-produced figures were inspected for every K value between 2 and 20. The STRUCTURE-produced diagram for K=3 and K=14 are shown in Figure 5. Like all of the other STRUCTURE-produced plots at different K values, the STRUCTURE plot in Fig. A.6B exhibits a high level of admixture among populations, once again indicating that no clear population structure exists.

Discussion

This study examined genetic diversity among *Striga hermonthica* populations from Senegal, Niger, Mali, and Eritrea using 12 microsatellite markers. Overall, a high level of heterozygosity was found in all populations. This high level of heterozygosity was also observed in a smaller sample of exclusively Malian materials (Estep et al., 2011) and in Kenya (Gethi et al., 2005). Additional earlier studies did not use markers that allowed assessment of heterozygosity (Bharathalakshmi et al., 1990; Olivier et al., 1996; Ali et al., 2009; Welsh and Mohamed 2011).

The Rst value for these samples was relatively low. Rst is a measure of the proportion of among population variation compared to within population variation for samples analyzed with SSR markers. The observed Rst value of .016 indicates that there is very little among population variation, and that most genetic variation in *S. hermonthica* is within-population variation. In full agreement, AMOVA indicated that almost all variation was within populations or within individuals, with very little variation among populations. These results indicate that there is extensive gene flow among populations, even among those across large geographic distances. These results agree with a previous study that had found high gene flow between Malian populations of *S. hermonthica* (Estep et al., 2011). Our results are surprising, given that samples were collected from sorghum in Eritrea, a point ~4300 kilometers from the nearest other collection site, in Niger. We initially viewed the Eritrean samples as an outgroup for our otherwise sub-Saharan and West African collections, but their lack of distinction means that gene flow occurs even between these far-removed locations. It has been proposed that *S. hermonthica* co-evolved in the mountains of Ethiopia and the Nubian hills of central Sudan, along with its highly tolerant host, *Sorghum bicolor* (Ejeta, 2007; Atera and Itoh, 2011). If this were the case,

then one would expect the center of *Striga* diversity to be in central Sudan, with independent radiations north-east to Eritrea and south-west to Niger, Mali and eventually Senegal. Our data do not support this scenario, but rather suggest that gene flow between all of these locations is ongoing, perhaps erasing population distinctions that may have existed from the original movement of *S. hermonthica* onto African crops.

Outcrossing populations such as *S. hermonthica* tend to have higher heterozygosity and a resultant lower degree of population distinction than self-fertilizing species (Hamrick, 1982). In this regard, the investigated samples approach Hardy-Weinberg equilibrium to a degree that suggests nearly complete gene flow, otherwise known as panmixis. Because climatic differences and variation in agricultural practices are both much greater in the north-south transect than in the east-west transect, then southern Africa would be the most likely place to find a unique population structure for *S. hermonthica*.

S. hermonthica is rarely seen on the pearl millet (*Pennisetum glaucum* subsp. *violaceum*) that is grown East of the Nile River. Our data suggest that it is not any differences in *S. hermonthica* genetics that explains this pearl millet resistance, but rather it is due to resistance traits that are encoded in the genomes of East African pearl millet. Hence, this E. African germplasm should be a good source of possible resistance for W. African pearl millet, which is highly susceptible at all locations from Western Sudan to the Atlantic Ocean, and in southern Africa.

Our results using both PCoA and STRUCTURE revealed no apparent clustering of populations based on host. This result is primarily based only on sorghum and pearl millet comparisons, because our numbers of populations derived from maize and rice were low. Still, we would have been able to easily see if the maize or rice samples were highly distinctive, and

they were not. The absence of distinction between *S. hermonthica* collected on sorghum or pearl millet is a surprise, given the very different outcomes of parasitism on pearl millet (which is almost always highly susceptible west of the Nile River) and sorghum (which routinely exhibits reasonable tolerance or even resistance). Recent evidence of qualitative resistance to *Striga hermonthica* from wild pearl millet accessions collected in Senegal (Wilson et al., 2000, 2004) and quantitative resistance to *Striga hermonthica* in a breeding population based upon less-susceptible pearl millet landraces from Niger (Kountche et al., 2013) suggests additional sources of *Striga* resistance that have not yet been exploited for pearl millet improvement.

Given our results, it seems most likely that any *Striga* population distinction that might occur, based on response by and to the host, will be erased very rapidly by an enormous level of gene flow. Perhaps this obliteration of adaptation by gene flow would be most likely to not occur in regions where only sorghum or only pearl millet are grown, but if gene flow can rapidly occur between locations as remote as Eritrea and Senegal, then even mono-cultured areas may receive huge doses of *S. hermonthica* germplasm from regions of mixed culture or mono-culture with a different crop.

All previous studies have revealed high genetic diversity within populations of *S. hermonthica*, but little to moderate genetic diversity among populations of *S. hermonthica* (Bharathalakshmi et al., 1990; Olivier et al., 1996; Gethi et al., 2005; Ali et al., 2009; Estep et al., 2011; Welsh and Mohamed, 2011). Initial studies indicated that almost all discovered variation is due to geographic location and distance, with little to none being due to specialization in host plant (Bharathalakshmi et al., 1990; Olivier et al., 1996; Gethi et al., 2005). Conversely, a study of populations in Sudan indicated that there were some clear differences between populations adapted to pearl millet and those adapted to either sorghum or maize (Ali et al., 2009). However,

additional more recent studies of populations in Mali (Estep et al., 2011) and Ethiopia (Welsh and Mohamed, 2011) did not find any population structure due to host specificity. Our studies yielded results similar to all previous publications, except Ali et al., (2009), which stands as an exception to all previous and subsequent research on this question.

The absence of population structure in *S. hermonthica* that is based on host specificity suggests that any host specificity that exists within *S. hermonthica* populations is not based on the genetics of the parasite. Any apparent host specificity that is observed might be due to differences in agronomic practices or seed-based gene flow for different host plants. In Sahelian Mali, for instance, sorghum is grown on better soil and with higher inputs than the pearl millet in adjacent fields. Alternatively, there could be a genetic basis for host specificity, but manifested by a small number of parasite genes, thus requiring more markers to be detected.

Genetics studies done by Román et al., (2001) examined populations of the parasitic plant *Orobanche crenata* across southern Spain. This species is similar to *S. hermonthica* in seed production and dispersal. Both species produce large quantities of small seeds with long-range dispersal methods. Populations of *O. crenata* across southern Spain showed a similar pattern of genetic diversity as seen in this study. Almost all variation was within populations rather than among populations, despite the long distances separating populations. This result was attributed to high gene flow among populations, partially due to transfer of crop seeds among farmers that were contaminated with *O. crenata* seeds.

The lack of either host specificity or geographic differentiation observed in our study suggests that genetic resistance factors in the host(s) could be universally effective or ineffective at all locations in the regions of Africa that we investigated. This would be a great boon to crop breeders, as they would not need to test for effectiveness at multiple locations. However, there

are numerous anecdotal reports (Ejeta et al., 2007) of *Striga* resistance that broke down at a particular location, suggesting the influx of a new *S. hermonthica* race. Our data do not support such a prediction, suggesting instead that this may have been due to changes in host seed genetics, in agricultural practices or in the environment. For instance, Kountche et al., (2013) observed large host genotype × environment effects in screens of pearl millet full-sib progenies against *Striga hermonthica* populations (for which resistance was quantitatively inherited) at Cinzana, Mali and Sadoré, Niger. If environmental variables, rather than *Striga* seed genetics, are the major factor in host tolerance or resistance, then multi-location trials across multiple-seasons will still be needed to predict crop performance in the face of this devastating pathogen. If the arrival of a new *S. hermonthica* race is actually involved in some host resistance breakdown, then these witchweed samples should be collected and subjected to deep genetic analysis to find novel alleles that correlate with observed virulence and pathogenicity differences. Given the high gene flow and high heterozygosity properties, linkage disequilibrium is expected to be tiny in this parasite, so markers associated with different *S, hermonthica* virulence phenotypes will probably need to be discovered within the responsible genes themselves.

Table A.1. Information about *Striga hermonthica* populations. Country, region, district, and nearest village for each population, and the number of samples collected from each population.

Pop	New Pop	Host Pop	Country	Region	District	Village	# Samples
3700	Mali_PM1	Pearl Millet	Mali	Ségou	Tominian	Sindala	8
4400	Mali_PM1	Pearl Millet	Mali	Koulikoro	Siby/Sibi (Kati)	Sindala	6
10000	Mali_PM1	Pearl Millet	Mali	Koulikoro	Didieni (Kolokani)	Didieni	2
2600	Mali_PM2	Pearl Millet	Mali	Mopti	Bandiagara	Bandiagara	5
3900	Mali_PM2	Pearl Millet	Mali	Mopti	Madiama (Djenné)	Madiama	5
11100	Mali_PM3	Pearl Millet	Mali	Ségou	Cinzana (Ségou)	Sanogola	8
3100	Mali_S	Sorghum	Mali	Mopti	Madiama (Djenné)	Madiama	8
4500	Mali_M	Maize	Mali	Koulikoro	Siby/Sibi (Kati)	Siby	7
4600	Mali_R	Rice	Mali	Koulikoro	Siby/Sibi (Kati)	Siby	9
7100	Senegal_PM1	Pearl Millet	Senegal	Diourbel	Bambey	Ndialite	8
8400	Senegal_PM2	Pearl Millet	Senegal	Kaolack	Kaffrine	Ndianga	5
9700	Senegal_PM3	Pearl Millet	Senegal	Fatick	Founjoun/Foundiougne	Firdaussi	7
9800	Senegal_PM3	Pearl Millet	Senegal	Fatick	Founjoun/Foundiougne	Firdaussi	2
7200	Senegal_S1	Sorghum	Senegal	Diourbel	Bambey	Ndialite	7
8500	Senegal_S2	Sorghum	Senegal	Kaolack	Kaffrine	Ndianga	6
12200	Niger_PM	Pearl Millet	Niger	Dosso	Gaya	Bengou	1
12400	Niger_PM	Pearl Millet	Niger	Dosso	Gaya	Gaya	9
12100	Niger_S	Sorghum	Niger	Dosso	Gaya	Bengou	3
12300	Niger_S	Sorghum	Niger	Dosso	Gaya	Bengou	5
20100	Eritrea_S	Sorghum	Eritrea	-	-	-	9

Table A.2. Individuals and alleles per locus. Total number of individuals scored and the number of alleles identified at each locus.

Locus	Individuals	Alleles
SH1005	116	7
SH1008	109	10
SH1009	113	10
SH1012	115	13
SH1014	120	8
SH1016	118	12
SH1021	110	8
SH1024	118	6
SH1029	116	14
SH1030	94	13
SH1039	120	7
SH1042	119	10
Average	114	9.8

Table A.3. Alleles, heterozygosity, and fixation index for *S. hermonthica* populations. Mean number of alleles per locus (Na), mean number of effective alleles per locus (Ne), observed heterozygosity (Ho), expected heterozygosity (He), and fixation index (F) for each population

Population	Na	Ne	Ho	He	F
Mali_PM1	5.8	3.1	0.75	0.65	-0.17
Mali_PM2	5.0	3.1	0.71	0.63	-0.11
Mali_PM3	3.4	2.4	0.77	0.56	-0.38
Mali_S	4.5	3.0	0.65	0.61	-0.09
Mali_M	4.2	3.1	0.72	0.62	-0.18
Mali_R	4.4	2.7	0.63	0.57	-0.12
Senegal_PM1	4.7	3.0	0.79	0.64	-0.20
Senegal_PM2	3.3	2.4	0.79	0.56	-0.41
Senegal_PM3	4.1	2.9	0.81	0.61	-0.34
Senegal_S1	4.3	3.0	0.72	0.63	-0.19
Senegal_S2	3.8	2.8	0.72	0.60	-0.21
Niger_PM	5.3	3.6	0.82	0.69	-0.19
Niger_S	4.6	3.3	0.84	0.66	-0.31
Eritrea_S	4.3	2.9	0.79	0.61	-0.33
Average	4.2	2.8	0.73	0.60	-0.22

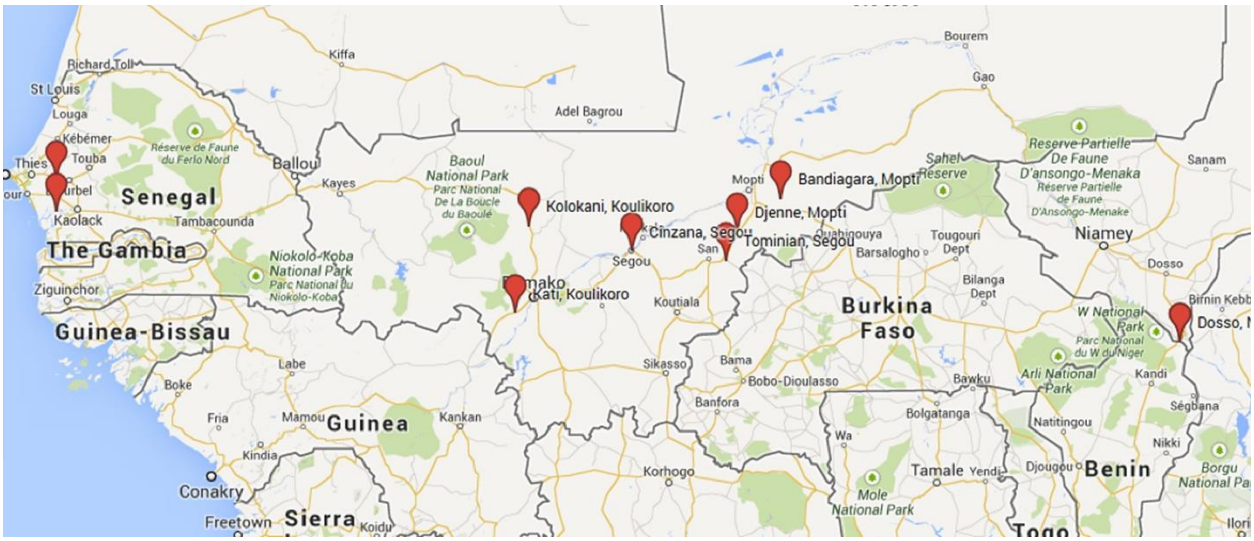


Figure A.1. Map of *S. hermonthica* populations. West and Central African collection sites are indicated by red markers.

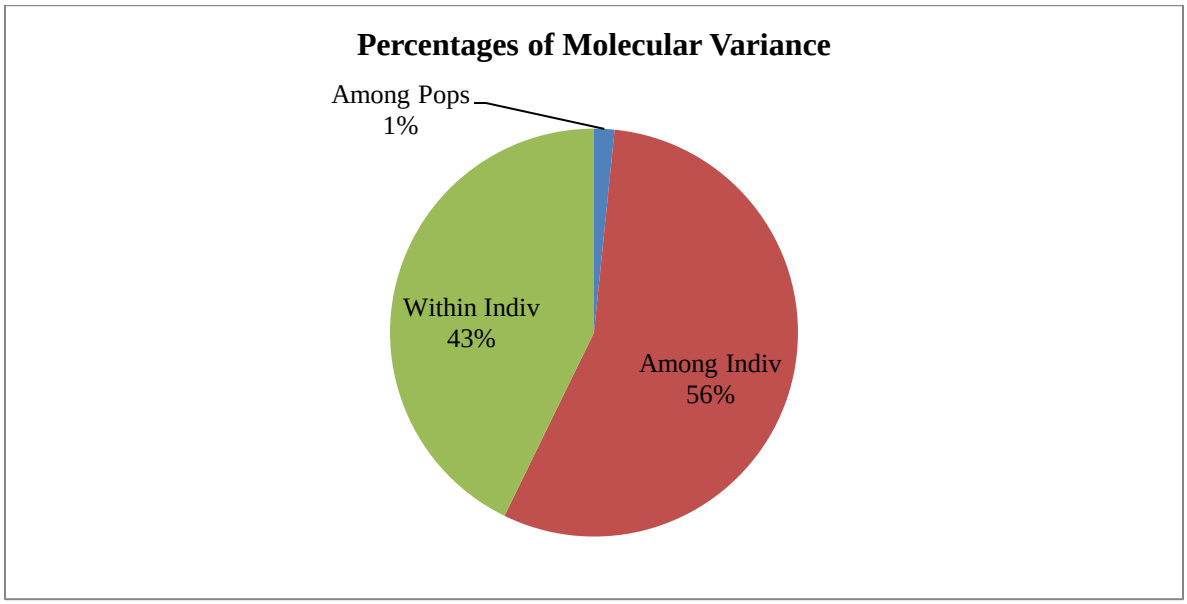


Figure A.2. AMOVA of *S. hermonthica* populations. Analysis of molecular variance using all 120 individuals from Mali, Senegal, Niger, and Eritrea.

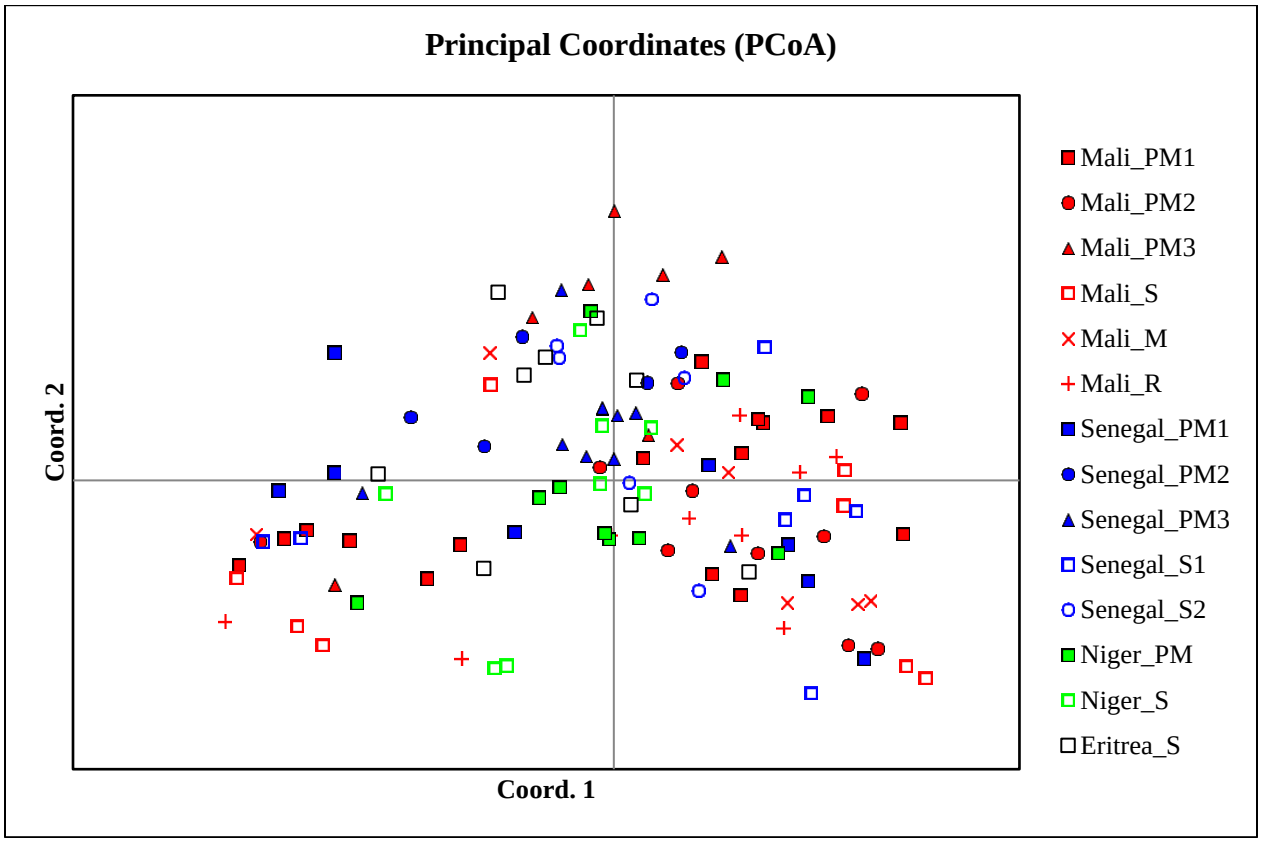


Figure A.3. Principle coordinates analysis of all 120 individuals. Countries are indicated by color: Mali (red), Senegal (blue), Niger (green), and Eritrea (black). Hosts are indicated by shape: pearl millet (solid shapes), sorghum (hollow shapes), maize (x), and rice (+).

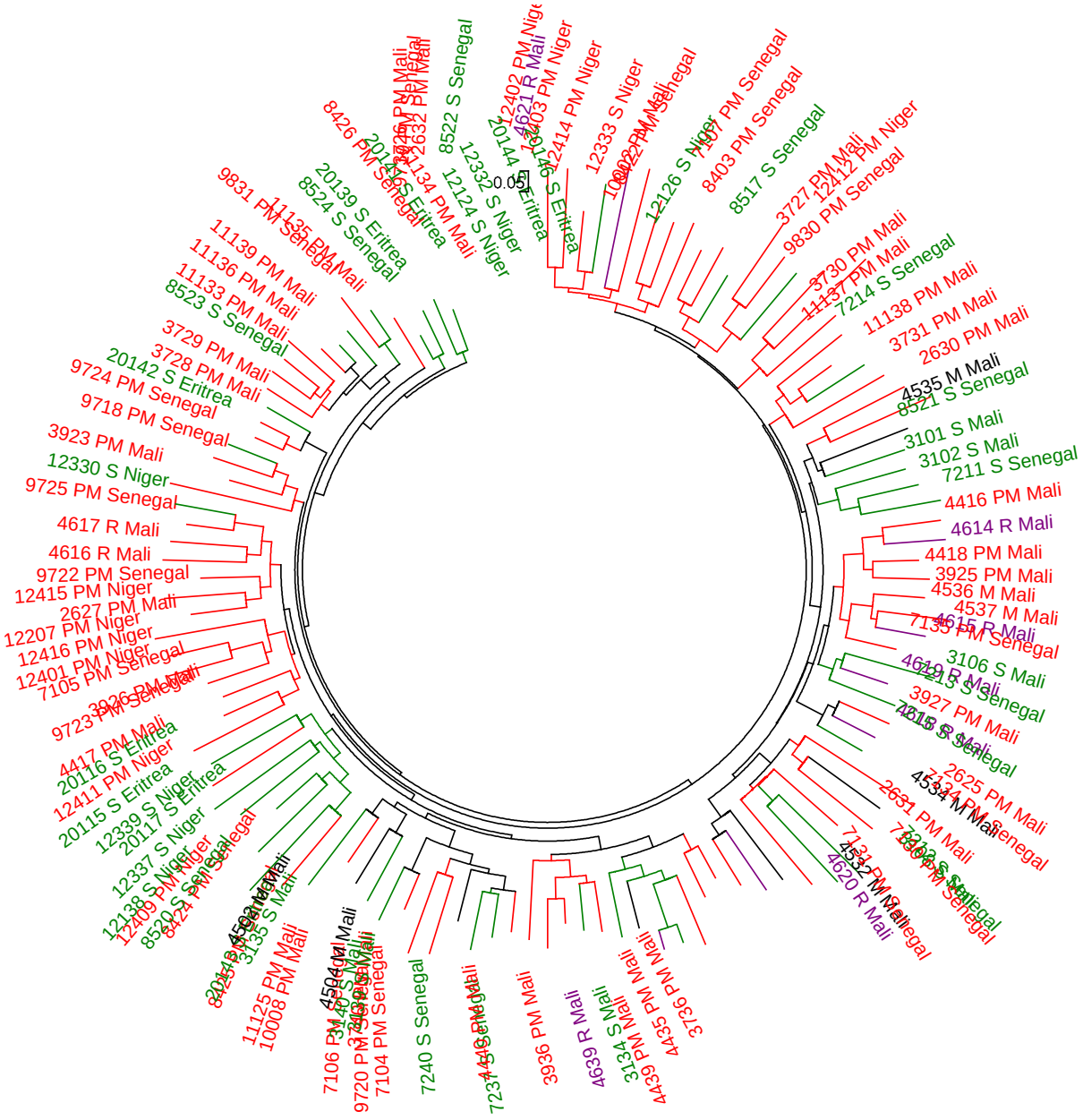


Figure A.4. Minimum evolution tree based on the chord genetic distances. (Slatkin, 1995). The branches of the tree are color-coded based on the host species of the given Striga accession. Host species are indicated by color: pearl millet (red), sorghum (green), rice (purple), and maize (black).

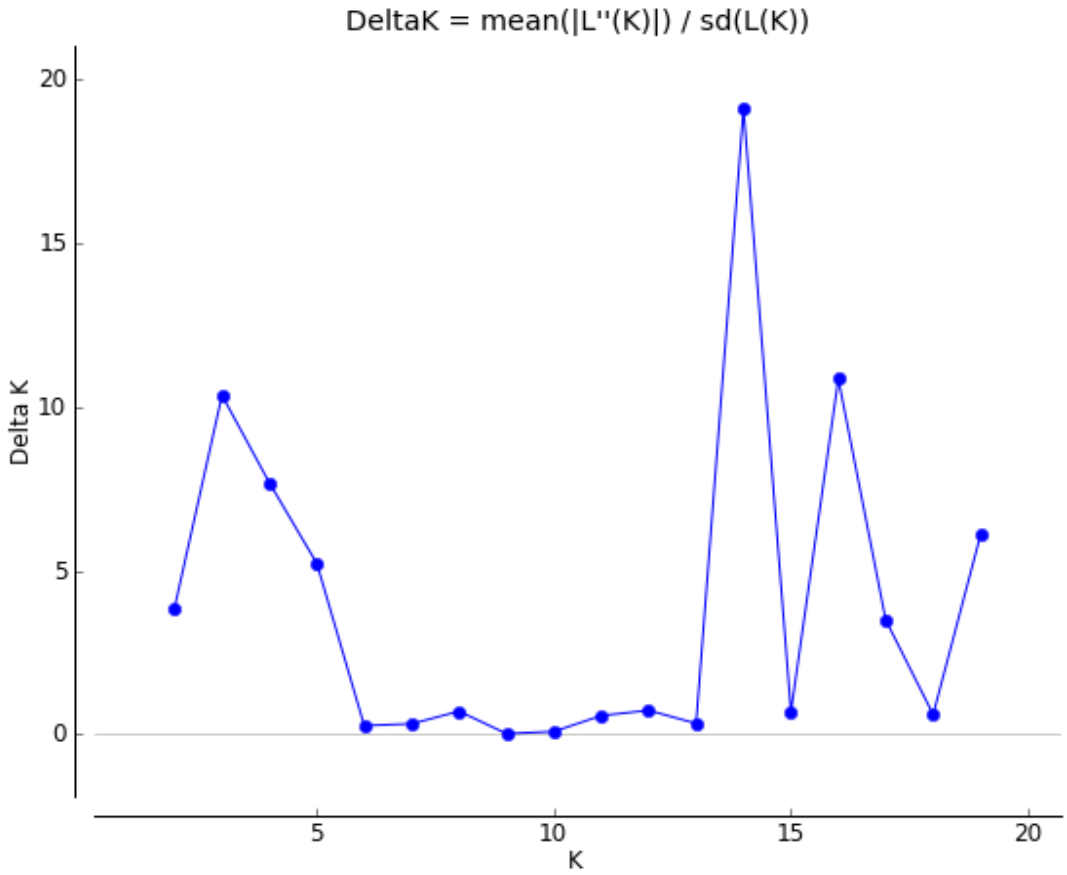
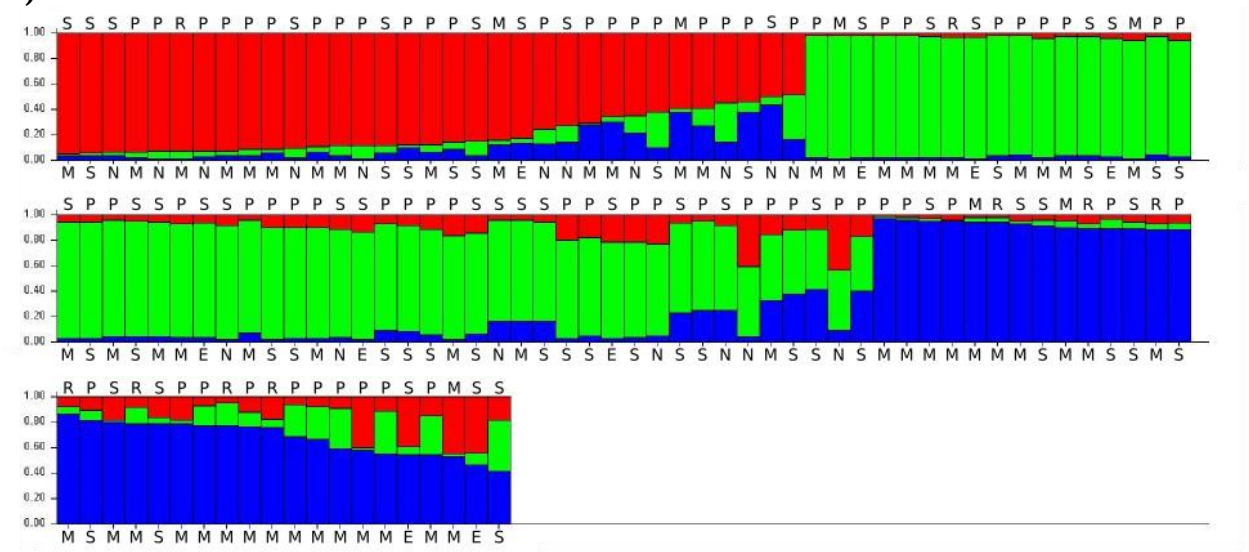


Figure A.5. STRUCTURE HARVESTER plot. Predicts the optimal K values that best fit the data.

A)



B)

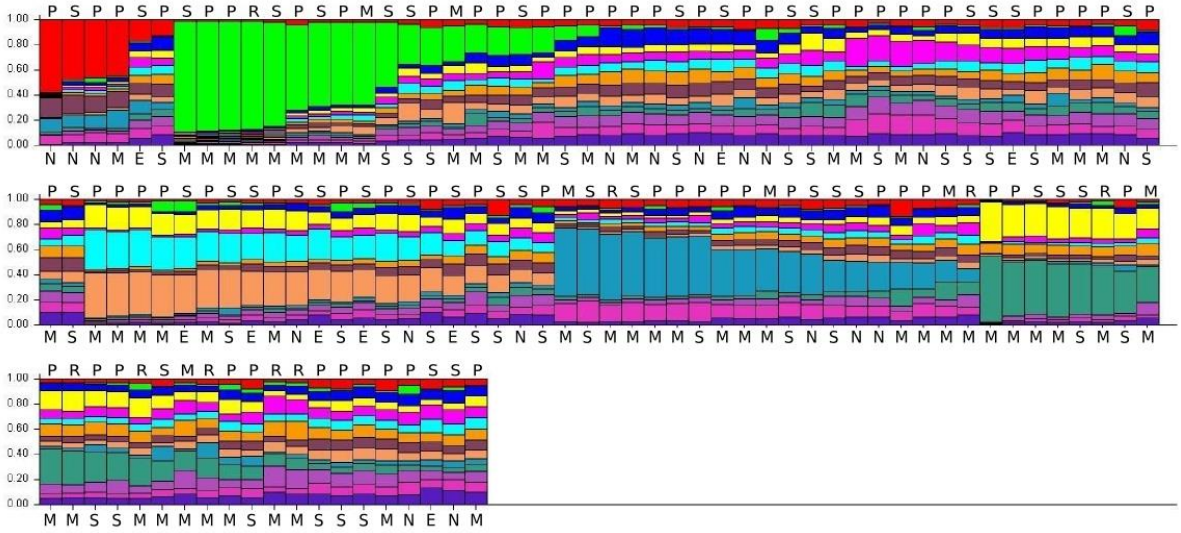


Figure A.6. STRUCTURE plots of *S. hermonthica* individuals. STRUCTURE plots at A) K=3 and at B) K=14. Letters below the columns indicate country where samples were collected: Mali (M), Senegal (S), Niger (N), and Eritrea (E). Letters above the columns indicate host plant from which samples were collected: sorghum (S), pearl millet (P), maize (M), and rice (R).