DISSERTATION

THE GENETICS AND GENOMICS OF HERBICIDE RESISTANT *KOCHIA SCOPARIA* L.

Submitted by

Eric L. Patterson

Department of Bioagricultural Sciences and Pest Management

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2018

Doctoral Committee:

Advisor: Todd Gaines

Chris Saski
Daniel Sloan
Stephen Pearce

ABSTRACT

THE GENETICS AND GENOMICS OF HERBICIDE RESISTANT *KOCHIA SCOPARIA* L

Weed genomics resources lag behind other plant biology disciplines despite larger annual crop losses occurring due to weeds than to plant pathogens or invertebrate pests. To date only a handful of weed genomes are assembled, and what is available is generally incomplete, poorly annotated, or only useful to a small group of researchers. Recent advancements in sequencing and an increased interest in the genetic foundations of weedy traits have contributed to driving *de novo* genome assemblies for key weed species. The introduced weed species *Kochia scoparia* (kochia) is the most important weed species in Colorado and severely impacts yield in various crop systems including sugar beet, wheat, and corn. Additionally, kochia rapidly invades disturbed land including roadsides, drainage areas, rangelands, and pastures. Kochia spans a massive geographic distribution, from as far south as Mexico, as far north as Saskatoon, Canada, as far east as the Mississippi river, and as far west as Oregon. Locally, kochia populations are well adapted to various abiotic stresses including drought, cold, high salinity, and high wind.

Recently, and most importantly, kochia has evolved resistance to several modes of herbicide action. Currently kochia populations exist that are resistant to acetolactate synthase (ALS) inhibitors, photosystem II (PSII) inhibitors, several synthetic auxin compounds, and the 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) inhibitor, glyphosate. Individuals have even been identified that are resistant to all four modes of action (MOA) simultaneously. Each herbicide mode of action (MOA) resistance case is caused by different mutations or even different mutation types (target site SNPs, copy number variation, translocation changes, etc.).

Selection pressure from herbicides is intense as not having the proper allele is lethal; therefore, resistance alleles are selected and go to fixation quickly. Kochia populations may be especially prone to herbicide resistance for a variety of physiological reasons, as kochia plants can produce thousands of seeds, are wind pollinated, are primarily outcrossing, and have tumbleweed seed dispersal in the windier environments like eastern Colorado and Kansas. Additionally, there may be genetic and genomic explanations for rapid herbicide resistance evolution such as rapid mutation rates or dynamic responses to environmental stress.

Glyphosate resistance, in particular, has driven a significant amount of herbicide resistance research in this species. In this case, resistance is caused by copy number variation of the target gene, *EPSPS*. Over production of the EPSPS enzyme makes normally lethal doses of glyphosate inadequate for control. Many of the details underlying gene amplification are missing, such as what are its origins and what genes are included in the duplication event. Understanding mechanisms of gene duplication is fundamental to understanding the evolution of resistance, predicting future gene duplication events, and understanding the significance of fitness and inheritance studies.

ACKNOWLEDGEMENTS

I here acknowledge all the wonderful individuals who have supported my research and education; it’s a long time coming!

I especially want to thank my committee: Dr. Todd Gaines, who gave me the opportunity to do the work I love and for always being there as a friend, mentor and colleague, Dr. Dan Sloan who shaped my understanding of plant genomes, Dr. Chris Saski for forcing me to always improve my science and adopt new techniques, and Dr. Stephen Pearce for asking the tough questions.

Thank you to the weed lab for being excellent scientists and friends. I would have killed every plant if it wasn’t for Dr. Eric Westra and the undergraduates, Crystal Sparks, Rachel Chayer, and Abigail Barker. Thank you to Dr. Phil Westra, my grand-advisor, for always believing in me and delivering sage advice. Dr. Scott Nissen for all the fun times and life discussions.

And finally, thank you to Dr. Margaret Fleming, you already know all you have done for me.

DEDICATION

For my little scientists, thanks for all the "help" in the lab.

For my parents who never saw this coming.

For my grandma and her love.

And finally, for my wife who keeps me sharp and never stops making me better...

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGEMENTS iv

DEDICATION v

TABLE OF CONTENTS vi

CHAPTER 1: AN INTRODUCTION TO EPSPS GENE DUPLICATION 1

Summary 1
Introduction 2
Conclusion 18

REFERENCES 20

CHAPTER 2: THE GENOME OF KOCHIA SCOPARIA 35

Summary 35
Introduction 36
Methods 37
Results 44
Discussion 47
Conclusion 52
Tables 54
Figures 60

REFERENCES 65

CHAPTER 3: THE EPSPS LOCUS IN KOCHIA SCOPARIA 73

Summary 73
Introduction 74
Methods 76
Results 79
Discussion 83
Conclusion 88
Tables 89
Figures 94

REFERENCES 103

SUMMARY OF DISSERTATION 109

CHAPTER 1: AN INTRODUCTION TO EPSPS GENE DUPLICATION

Glyphosate resistance and *EPSPS* gene duplication: Convergent evolution in multiple plant species[1]

Summary

One of the increasingly widespread mechanisms of resistance to the herbicide glyphosate is copy number variation (CNV) of the 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) gene. *EPSPS* gene duplication has been reported in eight weed species, ranging from 3-5 extra copies to more than 150 extra copies. In the case of Palmer amaranth (*Amaranthus palmeri*), a section of >300 kb containing *EPSPS* and other genes has been replicated and inserted at new loci throughout the genome, resulting in significant increase in total genome size. The replicated sequence contains several classes of mobile genetic elements including helitrons, raising the intriguing possibility of extra-chromosomal replication of the *EPSPS*-containing sequence. In kochia (*Kochia scoparia*), from three to more than 10 extra *EPSPS* copies are arranged as a tandem gene duplication at one locus. In the remaining six weed species that exhibit *EPSPS* gene duplication, little is known about the underlying mechanisms of gene duplication or their entire sequence. There is mounting evidence that adaptive gene amplification is an important mode of evolution in the face of intense human-mediated selection pressure. The convergent evolution of CNVs for glyphosate resistance in weeds, through at least two different mechanisms, may be indicative of a more general importance for this mechanism of adaptation in plants. CNVs

[1] Patterson, E. L., Pettinga, D. J., Ravet, K., Neve, P., & Gaines, T. A. (2017). Glyphosate resistance and EPSPS gene duplication: Convergent evolution in multiple plant species. *Journal of Heredity*, *1*, 9.

warrant further investigation across plant functional genomics for adaptation to biotic and abiotic stresses, particularly for adaptive evolution on rapid time scales.

Introduction

The herbicide glyphosate has been described as a “once-in-a century-herbicide” due to its unique broad spectrum of weed control efficacy (Duke and Powles 2008). It inhibits the enzyme 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) which is found in both monocotyledon and dicotyledon plants (Steinrücken and Amrhein 1980). EPSPS catalyzes the reaction that metabolizes 3-phosphoshikimate into 5-enolpyruvylshikimate-3-phosphate, an essential step in the synthesis of aromatic amino acids. It is thought that glyphosate causes plant death by starving the plant of aromatic amino acids (Schönbrunn et al. 2001). The ecological toxicity profile of glyphosate has been shown to be extremely low due to rapid metabolism by soil microbes and tight binding of the chemical to soil (Giesy et al. 2000; Rueppel et al. 1977; Williams et al. 2000). Additionally, *EPSPS* is found only in plants and microorganisms with no homolog in animals (Herrmann and Weaver 1999). Glyphosate was introduced as a herbicide in the early 1970s (Baird et al. 1971) and has been used in non-selective applications (e.g., orchards, vineyards, fallow, prior to planting broadacre crops, postharvest) since its introduction. Beginning in 1996, the introduction of transgenic glyphosate-resistant crops including cotton, soybean, sugar beet, and corn extended glyphosate use to selective in-crop application (Duke and Powles 2008; Padgette et al. 1996).

The commercially successful transgenic glyphosate-resistant crops contain a gene of bacterial origin (*CP4 EPSPS*) that is glyphosate-insensitive and therefore confers a high level of resistance in plants (Padgette et al. 1996). However, attempts to discover genetic variation for glyphosate resistance in crops provide insights into the natural selection of glyphosate resistance in weeds. Several molecular and genetic approaches were utilized to develop glyphosate-resistant crops, although most of these were not commercialized. A perennial ryegrass variety was recurrently selected with increasing doses of glyphosate over 11 generations, but this selection experiment resulted in only moderate resistance (Johnston and Faulkner 1991). Chemical mutagenesis of over 1 million *Arabidopsis thaliana* seeds did not produce any resistant plants, leading to the conclusion, at the time, that a single point mutation in the target-site plant *EPSPS* may not be sufficient to confer resistance (Bradshaw et al. 1997; Haughn and Somerville 1987). Liquid plant cell cultures of chicory, petunia, tobacco, tomato, and carrot were exposed to increasing amounts of glyphosate and eventually some of the cells became resistant to the glyphosate in the media by over-expressing *EPSPS*, sometimes even by increases in gene copy number (Goldsbrough et al. 1990; Nafziger et al. 1984; Sellin et al. 1992; Shyr et al. 1993; Smith et al. 1986; Steinrücken et al. 1986; Wang et al. 1991). These resistant cell lines typically had issues that prevented their commercial release, such as instability of the increase in *EPSPS* gene copy number upon regeneration to a whole plant, loss of glyphosate resistance on regeneration, or infertility of the regenerated plant following glyphosate application. Experiments in alfalfa, soybean, and tobacco further demonstrated that *EPSPS* gene amplification can confer glyphosate resistance in plants (Widholm et al. 2001). Ultimately, the recurrent selection, mutagenesis, and cell culture methods suggested that there is limited standing genetic variation for glyphosate resistance in plants.

The first case of a naturally evolved glyphosate-resistant (GR) weed was annual ryegrass (*Lolium rigidum*), discovered in Australia in an orchard (Powles et al. 1998). To date, 37 species have been reported as GR (Heap 2017). These 37 species include both monocotyledon and dicotyledon weeds. Glyphosate resistance has evolved in a variety of situations including orchards, cereals, fence lines, and transgenic GR crops. We know now that glyphosate resistance in weeds can be conferred by several genetic mechanisms including point mutations in the active (target) site of *EPSPS*, reduced translocation of glyphosate to the meristems, and vacuole sequestration (reviewed by Sammons and Gaines 2014). One of the most interesting and increasingly widespread mechanisms of resistance to glyphosate is increased copy number of the *EPSPS* gene. In this review, we discuss the current information for each species that has evolved increased *EPSPS* gene copy number as a resistance mechanism and synthesize the current state of knowledge for this striking case of convergent evolution. We suggest that adaptive gene amplification can be an important mode of evolution on rapid time scales in the face of intense human-mediated selection pressure.

EPSPS Copy Number Variation

An increase in copy number of a gene produces copy number variation (CNV), referred to as gene amplification or gene duplication. *EPSPS* gene duplication is thought to confer resistance to glyphosate by over-production of the target protein, EPSPS. The increased protein pool of EPSPS requires an equivalent increase in applied glyphosate to inhibit sufficient amounts of EPSPS to cause lethality (Gaines et al. 2010; Sammons and Gaines 2014). Additionally, since glyphosate binding to the EPSPS protein is essentially irreversible, once glyphosate is bound it is effectively sequestered by the plant.

The first demonstration that *EPSPS* gene duplication conferring glyphosate resistance was in Palmer amaranth (*Amaranthus palmeri*) from Georgia, USA (Gaines et al. 2010). Six additional weedy species have independently evolved increased *EPSPS* copy number and one species has obtained high *EPSPS* copy number by hybridization with GR Palmer amaranth (Chen et al. 2015; Lorentz et al. 2014; Malone et al. 2016; Nandula et al. 2014; Ngo et al. 2017;

Salas et al. 2012; Wiersma et al. 2015). To date four of the resistant species are dicotyledons in the Chenopodiaceae/Amaranthaceae and four are monocotyledons in the Poaceae.

Palmer amaranth

GR Palmer amaranth was first reported in the US state of Georgia (Culpepper et al. 2006). Since that time, GR Palmer amaranth has become a substantial problem in several major crops in North and South America (Küpper et al. 2017; Norsworthy et al. 2014; Price et al. 2011; Sosnoskie and Culpepper 2014). Quantitative PCR using relative quantification with a single copy normalization gene has demonstrated that resistant Palmer amaranth contains from 50 to more than 150 copies of the *EPSPS* gene (Gaines et al. 2011; Küpper et al. 2017). In this species, increased *EPSPS* gene copy number is directly proportional to *EPSPS* mRNA and EPSPS protein abundance which is proportional to the quantity of glyphosate needed to control these plants (Gaines et al. 2010).

Cytogenetics approaches have proven highly useful in characterizing the molecular structure of gene duplications involved in herbicide resistance (Jugulam and Gill 2017). Cytogenetic studies using Fluorescence In Situ Hybridization (FISH) in GR Palmer amaranth showed that the *EPSPS* copies are dispersed across the genome on all chromosomes (Gaines et al. 2010). The duplicated *EPSPS* copies were shown to contain introns, indicating the duplication did not occur via an RNA-transposon, and multiple types of mobile genetic elements were found to be associated with the duplicated *EPSPS* genes (Gaines et al. 2013). More recently this has been confirmed using genomics techniques (Molin et al. 2017a). The amplified region that contains *EPSPS* was sequenced by generating a BAC library and probing for the *EPSPS* gene and then sequencing those clones with long read Pacific Biosciences sequencing technology. The amplified region was found to be ~300 kb, in high abundance (>100 copies), and dispersed across the genome (Molin et al. 2017a). Flow cytometry measurements for GR Palmer amaranth individuals show significantly larger genomes than glyphosate-susceptible (GS) Palmer amaranth due to the large size and high copy number of the *EPSPS* replicon. Calculations show the GR genome to be between 20-30 Mb (7-13%) larger than the GS genome (Molin et al. 2017a).

The amplified region contains 72 predicted genes, many of which were classified as transposable elements (TEs) based on a repetitive element database (Jurka et al. 2005), including LTR retrotransposons, non-LTR retrotransposons, class II transposons, and helitrons (Molin et al. 2017a). Several of the genes in this region show increased transcription but not always to the same magnitude as *EPSPS* suggesting that either 1) not all genes in the amplified region are always duplicated or 2) these other genes are regulated differently than *EPSPS* (Molin et al. 2017a). The potential that the >300 kb replicon may have a circular structure is especially intriguing, inviting speculation that the entire structure could replicate externally to the chromosome and insert and excise repeatedly throughout the genome. This is the first documented case of such a potentially mobile, large genetic structure associated with gene duplication and copy number variation in any species.

To understand inheritance of the resistance trait, several studies with GR Palmer amaranth crossed to susceptible plants measured *EPSPS* copy number in the F1 and F2 progeny (Chandi et al. 2012; Mohseni-Moghadam et al. 2013). As would be expected due to the large number of *EPSPS* gene copies and their distribution across multiple, unlinked locations on different chromosomes, inheritance of glyphosate resistance in these studies was non-Mendelian and segregated as a polygenic trait. There are also indications that Palmer amaranth can produce seeds asexually via facultative apomixis (Ribeiro et al. 2014), which may facilitate inheritance of the potentially meiotically-unstable *EPSPS* gene duplication when it occurs via transduplication throughout an individual plant genome. A segregating F_2 population contained individuals with complete loss of the *EPSPS* replicon (*EPSPS* copy number of one) as well as individuals with *EPSPS* gene copy number greater than the sum of both parents (Gaines et al. 2011). The apparent instability of the *EPSPS* CNV raises questions about the likelihood of multiple independent CNV events versus a single origin and spread, as spread via gene flow could be dependent on the stability of transmission of increased *EPSPS* gene copy number across multiple generations. Resequencing and alignment of the *EPSPS* replicon from multiple glyphosate-resistant populations across the USA showed high sequence homology, supporting a hypothesis of single origin of the *EPSPS* replicon in Palmer amaranth (Molin et al. 2017b). At this point in time, some combination of both multiple origins (convergent evolution) and spread via seed- and pollen-mediated gene flow seems most likely (Beard 2014).

Some mutations conferring herbicide resistance have associated fitness costs including reduced growth rate, fecundity, and/or competitiveness due to direct or pleiotropic effects of the mutation (reviewed by Vila-Aiub et al. 2009). The *EPSPS* gene duplication in Palmer amaranth could affect plant fitness (growth rate, fecundity, competitiveness) in several ways, including 1) the increased metabolic cost of *EPSPS* overproduction; 2) potential pleiotropic effects of over-expressing other genes in the replicon; and 3) genome instability and disruption of other genes due to *EPSPS* insertion events. Two separate studies found no observable fitness costs in physiological traits (Giacomini et al. 2014; Vila-Aiub et al. 2014). However, since Palmer amaranth is dioecious and therefore an obligate outcrossing species, no studies have used near isogenic lines for conclusive fitness studies. Indeed, due to the size, dispersion, and potential instability of the *EPSPS*-containing replicon, obtaining true-breeding lines may not be possible.

There may also be other fitness related traits that have not yet been measured that may demonstrate fitness costs of *EPSPS* gene amplification and genome expansion in Palmer amaranth.

Other Amaranthus Species

After the initial discovery of *EPSPS* gene amplification in Palmer amaranth, other GR *Amaranthus* weeds were evaluated for this mechanism. *EPSPS* copy number increase was described in waterhemp (*A. tuberculatus* syn. *rudis*) in several independent studies (Chatham et al. 2015a; Chatham et al. 2015b; Lorentz et al. 2014). *EPSPS* copy number in waterhemp was far fewer than in Palmer amaranth, with most resistant plants having between 4-8 copies up to a maximum of 16 copies (Chatham et al. 2015a; Chatham et al. 2015b). Dillon et al. (2017) grouped GR waterhemp into the following three categories of resistance magnitude: low glyphosate resistance (2-4 copies), moderate glyphosate resistance (4-7 copies), and high glyphosate resistance (7-16 copies). As shown in Palmer amaranth, genomic copy number was correlated with mRNA levels, shikimate accumulation (a biomarker for glyphosate inhibition of *EPSPS*), and glyphosate resistance level (Dillon et al. 2017). A fitness cost for increased *EPSPS* gene copy number in waterhemp was shown as a reduction in frequency of individuals carrying two or more *EPSPS* copies in a population grown for six generations without glyphosate selection (Wu et al. 2017).

Using FISH, it was discovered that the original copy of *EPSPS* in waterhemp is near the centromere in GS individuals (Dillon et al. 2017). There are several copies of *EPSPS* in tandem duplication at the same locus, near the centromere, in GR high copy number individuals. In the highest copy number individuals the *EPSPS* gene was also found on an extra chromosome, suggesting that tandem duplication may occur initially followed by transduplication and potentially replication of an extra chromosome (Dillon et al. 2017).

GR spiny amaranth (*Amaranthus spinosus*) exhibited up to a five-fold resistance to glyphosate in plants containing between 33-37 copies of *EPSPS* (Nandula et al. 2014). When the *EPSPS* gene was sequenced from GR individuals, the *EPSPS* gene was found to be identical to the gene from GR Palmer amaranth, having 29 single nucleotide polymorphisms when compared to the *EPSPS* gene from GS spiny amaranth. This evidence pointed to a hybridization event of spiny amaranth with high-copy number GR Palmer amaranth (Nandula et al. 2014). Inter-specific hybridization is known to occur within the *Amaranthus* genus (Trucco et al. 2005a; Trucco et al. 2005b; Trucco et al. 2009), including gene flow from Palmer amaranth to spiny amaranth (Gaines et al. 2012) and transfer of acetolactate synthase inhibitor resistance alleles between *Amaranthus* spp. (Franssen et al. 2001).

Kochia scoparia

Kochia scoparia (kochia) is a weed species in the Amaranthaceae common to the western Great Plains region of North America (Friesen et al. 2009) and GR kochia is a major agronomic challenge in this region (Kumar et al. 2014; Waite et al. 2013). The genus *Kochia* is related to the genus *Amaranthus* within the Amaranthaceae. Kochia has also evolved increased *EPSPS* copy number for glyphosate resistance (Godar et al. 2015; Wiersma et al. 2015), and currently is the only dicotyledon not in the *Amaranthus* genus with *EPSPS* CNV. Initially, GR kochia was shown to have *EPSPS* copy numbers between 3-9 (Kumar et al. 2015; Wiersma et al. 2015); however, in a survey from sugar beet fields, kochia plants were shown to occasionally have >10 copies of *EPSPS* (Gaines et al. 2016). Increased copy number has been correlated with increased mRNA and protein abundance as well as whole-plant resistance level in kochia (Gaines et al. 2016; Godar et al. 2015; Wiersma et al. 2015).

FISH in kochia has revealed that all copies of *EPSPS* occur at a single locus and Fiber-FISH suggests that all copies are located as a tandem duplication (Jugulam et al. 2014). Additionally, the Fiber-FISH results suggest several sizes for the tandem repeats, with the two most common being a repeat of ~45kb and a repeat of ~66kb. Additionally, some copies are slightly longer, >70kb, and one inversion was detected. The tandem duplication of *EPSPS* was proposed to be caused by an initial unequal crossing-over event that produced tandem *EPSPS* gene copies, followed by glyphosate selection pressure and further unequal crossing-over events during cell division that produced additional *EPSPS* copies in tandem duplication (Jugulam et al. 2014). Inheritance of the tandem *EPSPS* gene duplication was consistent with a single-gene pattern, as expected for a tandem duplication at a single locus (Jugulam et al. 2014).

An initial fitness study comparing high-copy number GR to GS kochia showed little to no fitness cost in most vegetative traits and little effect on reproductive traits (Kumar and Jha 2015). The two populations were collected from the same locality, but it is unknown how similar the genetic background is between the populations (Kumar and Jha 2015). More recently, researchers have made several crosses between GS and GR plants of varying copy number and measured several traits in the segregating F_2 population(s) (Martin et al. 2017). Some plants with elevated *EPSPS* copy number had delayed development, reduced fecundity, and reduced competitive ability. However, there was large variation among independent F_2 crosses in the magnitude of observed fitness costs, with fitness costs being either higher or absent depending on the specific cross (Martin et al. 2017). When comparing several GR and GS kochia populations in another study, it was observed that fitness costs were consistently found in germination characteristics but not necessarily in any vegetative characteristics (Osipitan and Dille 2017).

The Grasses

Several grass species in divergent genera of Poaceae appear to have independently evolved increased *EPSPS* copy number as a glyphosate resistance mechanism. Current information is limited to the occurrence of *EPSPS* gene duplication in the grasses, as no cytogenetic or sequencing studies have been completed. The species are Italian ryegrass (*Lolium perenne* ssp. *multiflorum*), ripgut brome (*Bromus diandrus*), goosegrass (*Eleusine indica*), and windmill grass (*Chloris truncata*), occurring in the USA, Australia, China, and Australia, respectively (Chen et al. 2015; Malone et al. 2016; Ngo et al. 2017; Salas et al. 2012). In all four grass species, increased copy number was associated with increased glyphosate resistance. In Italian ryegrass, *EPSPS* copy numbers were reported from 15 to 25 (Salas et al. 2012). In ripgut brome, *EPSPS* copy number ranged from 10 up to 36 copies (Malone et al. 2016). In goosegrass, *EPSPS* copy number was 89 in one population, 23-fold more copies than a susceptible population (Chen et al. 2015). Finally, in windmill grass, *EPSPS* copy number was reported from 32 up to 48 copies (Ngo et al. 2017). In these grass species, the inheritance, potential fitness costs, and cytogenetics of the *EPSPS* duplication events have not yet been reported.

Mechanisms of Copy Number Variation

Gene duplication is a relatively common process in evolutionary history and produces important raw material for adaptive evolution in mammalian cancer cells, bacteria, arthropods, and plants (Bass and Field 2011; Flagel and Wendel 2009; Gaines et al. 2010; Hastings et al. 2009; Schimke 1986; Wiersma et al. 2015). Plants can acquire additional gene copies in several ways. Mobile genetic elements such as transposable elements (TEs) are a well-studied mechanism of gene duplication. TE activity is usually suppressed because TE activity can have negative effects such as disrupting important genes or affecting their transcription, or causing genome instability (Jensen et al. 1999; Slotkin and Martienssen 2007). There is some evidence, however, that certain biotic and abiotic stresses can increase TE activity, resulting in genomic re-arrangements (Bennetzen 2005; Capy et al. 2000). These rearrangements can be the duplication of genes contained within the TE boundaries, the movement of regulatory elements, the disruption of genes near the TE insertion site, or changes in chromatin structure (Bennetzen 2005).

The type of mobile genetic element recently identified in Palmer amaranth shares similarities with helitron structures (Molin et al. 2017a). Helitrons are a type of transposable element that are hypothesized to use a “rolling circle” replication mechanism, mediated by a single stranded DNA intermediate (Kapitonov and Jurka 2001; Kapitonov and Jurka 2007; Thomas and Pritham 2015). Helitrons were first discovered in Arabidopsis and rice but have since been discovered in almost all eukaryotic lineages. Helitrons can be quite prevalent in some eukaryotic genomes, ranging from 0-5% of the total genetic content. The helitron-like sequence that is associated with *EPSPS* gene duplication in Palmer amaranth alone can cause a >5% increase in genome size (Molin et al. 2017a).

Another possibility for generating increased gene copy number is tandem duplication events. For tandem duplications to occur, unequal crossing-over must occur between homologous chromosomes. In humans, tandem duplication events are known to be generated by one of two mechanisms: non-allelic homologous recombination (NAHR) and microhomology-mediated events (Hastings et al. 2009). Anytime a double stranded break (DSB) occurs in a strand of DNA, the subsequent repair to the damaged location may introduce mistakes, such as if the repair proteins accidentally employ NAHR or microhomology-based unequal recombination while the damage is being repaired (Hastings et al. 2009). These events can happen in somatic or gametic cells, but only events in gametes or somatic cells that eventually differentiate into gametes are heritable and therefore relevant to evolution. Because plant somatic cells are totipotent and can differentiate into gametic cells at various stages, especially in long-lived plants, a mechanism exists by which somatic variation can eventually be incorporated into gametes. It is likely that a DSB or some other disruption near the *EPSPS* gene caused kochia to employ one of these unequal crossing-over mechanisms, inadvertently generating the tandem *EPSPS* duplications and copy number variation observed in this species (Jugulam et al. 2014).

Another way to generate additional copies of genes is via a polyploid event or gene flow from one organism to another. Polyploidy often shapes large-scale evolutionary events like speciation or genetic isolation and seems to be a relatively rare mechanism leading to single gene copy number changes, especially on short time scales (Adams and Wendel 2005; Ramsey and Schemske 1998). As previously mentioned, interspecific gene flow has occurred from Palmer amaranth to spiny amaranth, transferring duplicated copies of the *EPSPS* gene and glyphosate resistance (Gaines et al. 2012; Nandula et al. 2014).

In both animal and plant systems, it has been shown that environmental stress induces higher frequencies of CNVs (Hastings et al. 2000). The exact nature of the relationship between stress and CNVs is unclear. It could be that stress induces higher levels of DSB, resulting in more chances for gene duplications to occur and generate genetic diversity. Additionally, stress has been shown to change methylation patterns in several species which may be a way to regulate TE activity or the rate of DSB in certain genomic locations (Lämke and Bäurle 2017). There is evidence that unequal crossing-over events and TE insertions happen at hotspots mediated either by specific DNA sequences, epigenetics, or chromatin structure (Cai and Xu 2007; Drouaud et al. 2013; Gaut et al. 2007; Purandare and Patel 1997).

Copy Number Variation and Adaptation

Adaptation by gene duplication has been observed in bacteria, yeast, cancer cells, and plant cell cultures (Hyppa and Smith 2010; Slack et al. 2006; Suh et al. 1993; Watanabe et al. 2011). There are many reasons why gene duplications and CNV are a frequent mechanism underpinning adaptation. All genes contained within the region can have increased expression simultaneously, which may be adaptive, but not all genes necessarily have immediate changes in function. All genes within the region maintain their own promoters and all cis-regulatory elements used to modulate their expression. Due to redundancy in function, one or more of the gene copies is free from selection pressure to diverge through random mutations, assuming at least one copy maintains the original function. This divergence usually ends in pseudogenes but may also result in neo- or sub-functionalization, thereby generating novel genetic diversity which may be adaptive (Flagel and Wendel 2009; Lynch and Conery 2000).

Silent point mutations in the genome are a fairly consistent molecular clock and non-silent point mutations that change protein function are often subject to purifying selection (Drake et al. 1998). The rate of CNV generation, on the other hand, is variable and is subject to environmental factors. Under more intense selection pressures the number of CNV events in offspring increases, while under optimal conditions fewer genomic rearrangements are observed (DeBolt 2010). Species which have evolved higher rates of CNV, or more sensitivity to stress, may have increased genetic diversity, and therefore an increased chance of survival under strong selective pressures such as herbicide application (Kondrashov 2012; Żmieńko et al. 2014). This type of heritable, possibly adaptive, genetic variation due to CNV is especially important in plants that have short generational timescales and live in constantly changing environments with strong selective pressures, such as weeds in agricultural systems (DeBolt 2010; Hastings et al. 2009). The prevalence of CNV underlying glyphosate resistance provides further support for the importance of this mode of adaptation.

Gene amplification has been shown in arthropods to cause insecticide and miticide resistance for almost thirty years (Bass and Field 2011; Devonshire and Field 1991). A general expansion and functional diversification within gene families via gene duplication is evident in the genomes of pest species such as *Anopheles gambiae* when compared to *Drosophila melanogaster* (Ranson et al. 2002). In arthropods, gene amplification typically results in the overexpression of certain metabolic genes, including esterase (Hemingway 2000; Hemingway et al. 1998; Li et al. 2007; Ono et al. 1999; Raymond et al. 1989; Small and Hemingway 2000), glutathione-*S*-transferase (Vontas et al. 2001; Zhou and Syvanen 1997), and cytochrome P450 monooxygenase (Emerson et al. 2008; Schmidt et al. 2010). However, the target gene of insecticides and miticides can also be amplified and over-expressed to cause resistance, similar to the case of *EPSPS* gene duplication (Anthony et al. 1998; Kwon et al. 2010; Labbé et al. 2007b).

In the case of organophosphate resistance in *Culex pipiens*, the target gene acetylcholinesterase is duplicated and one of the copies carries a point mutation that generally confers a severe fitness cost. However, one copy maintains the wild-type sequence and continues to function normally, while the mutant copy confers a resistance benefit in the presence of the insecticide. In effect this series of genetic mutations (copy number variation followed by a single base pair mutation) has effectively resulted in a permanent heterozygous genotype with different alleles in duplicated genes (Bourguet et al. 1997; Labbé et al. 2007a; Labbé et al. 2007b). While this is an interesting example of how copy number variation can confer resistance, a more recent example in *Tetranychus urticae* links the number of copies of the target genes in a directly proportional relationship to the amount of target protein produced. Because the pool of target protein is larger, the amount of active ingredient needed to inhibit the protein pool also must increase, thereby conferring resistance to higher doses of organophosphate miticides (Kwon et al. 2010).

In animals (especially humans) copy number variation is often associated with genetic disorders, especially cancer; however, in plants there exist several examples of how copy number variations can generate genetic diversity useful for adaptation (Mishra and Whetstine 2016). In plants, resistance to the soybean root knot nematode in some soybean cultivars is due to duplication of three genes, resulting in over-expression of the three genes that is directly correlated with nematode resistance (Cook et al. 2012). Another example of the adaptive potential of CNVs is in clonally propagated potato which shows prolific and genome wide copy number variation. Clonally propagated varieties have upward of 30% of the genes in the genome duplicated or deleted. Additionally, there is a specific increase in the number of genes annotated as having roles in environmental stress tolerance. It is thought that clonally propagated plants tolerate a larger mutational load as they do not need to undergo meiosis and produce seed, both of which can be negatively affected by genomic rearrangements (Hardigan et al. 2016). Copy number variations may provide plants with novel genetic diversity, and their production may be stimulated by stress.

Recently resistance to Acetyl-CoA Carboxylase (ACCase)-inhibiting herbicides in hairy crabgrass (*Digitaria sanguinalis*) was reported to be due to 5 to 7-fold increase in *ACCase* gene copy number resulting in 3 to 9-fold increase in *ACCase* transcript abundance (Laforest et al.

2017). This provides the first example of CNV for resistance to a herbicide other than glyphosate, and further highlights the potential advantages of adaptive CNVs for rapidly generating increased gene expression phenotypes to confer herbicide resistance. Other than this recent example, to date gene duplication as a herbicide resistance mechanism has only been identified for *EPSPS* and glyphosate resistance, a target-site mechanism. This raises the question as to why there is a prevalence of the CNV-based mechanism for glyphosate. The *EPSPS* CNV may be an extremely rare event that is only revealed by intense selection over large geographical areas. Perhaps the genomic context of *EPSPS* happens to be more prone to duplication than other herbicide target-site genes, enabling tandem duplication and/or transduplication. The relatively low resistance level conferred by single nucleotide mutations in *EPSPS* (reviewed by Sammons and Gaines 2014) and the apparent high fitness cost of the highly-resistant double mutation T102I and P106S in *EPSPS* (TIPS) (Vila-Aiub et al. 2017; Yu et al. 2015) may indicate that *EPSPS* over-expression by gene duplication is a more efficient mechanism, in contrast to several other herbicide target genes for which target-site mutations are highly efficient and commonly selected (Powles and Yu 2010). However, the P106S mutation was recently shown to have a fitness advantage over *EPSPS* gene duplication in waterhemp, as the P106S mutation increased in frequency over six generations without glyphosate selection while the *EPSPS* CNV decreased in frequency (Wu et al. 2017). Additionally, previous research may have simply failed to consider gene duplication as a possible resistance mechanism, resulting in CNVs being overlooked in some cases of herbicide resistance evolution. Resistance to some herbicides is known to be caused by increased expression of non-target-site genes that metabolize the herbicide, including glutathione S-transferase (Cummins et al. 2013) and cytochrome P450 monooxygenase (Duhoux et al. 2015; Gaines et al. 2014; Gardin et al. 2015; Iwakami et al.

2014). In general the examples of increased non-target-site gene expression have not yet been evaluated for CNV.

Conclusion

To date, four dicotyledon species and four monocotyledon (grass) species have evolved *EPSPS* gene amplification resulting in glyphosate resistance. One of those species, spiny amaranth, obtained high copy numbers by interspecific gene flow while the other seven species seem to have evolved *EPSPS* gene amplification independently in a case of convergent evolution. In one species, Palmer amaranth, the mechanism of gene duplication is partially understood, involving transduplication of >300 kb of sequence containing EPSPS to multiple novel insertion sites, possibly through a helitron-like mechanism. Gene amplification in kochia is also well studied, occurring by a different mechanism with extra gene copies arranged as tandem duplications likely caused by unequal crossing over. In the remaining species, further investigation is required to elucidate the mechanisms that generated *EPSPS* gene amplification.

The convergent evolution of the same resistance mechanism, increased *EPSPS* gene copy number, via two different genomic mechanisms is quite striking and raises several questions. 1) Is *EPSPS* gene amplification present at initially low frequencies (i.e., rare standing genetic variation for *EPSPS* CNV) and how often does *EPSPS* gene amplification occur due to normal DNA repair processes or mobile genetic element activity (i.e., *de novo* genetic variation)? 2) Are potential fitness costs associated with *EPSPS* gene amplification, whether physiological (consequences of over-expressing *EPSPS* and/or other duplicated genes), genomic (disruption of other genes when the *EPSPS* replicon inserts at a novel locus), or energetic (increased ATP and amino acid usage to produce an over-abundance of EPSPS enzyme) likely to be balanced by ongoing selection for maximum resistance benefit with minimal fitness cost? 3) Given the previously observed instability of increased *EPSPS* gene copy number in plant cell culture and the instability of other gene duplications for xenobiotic resistance (e.g., in cancer cells), would *EPSPS* gene amplification be retained if glyphosate selection pressure were removed, and does the stability depend on the genomic mechanism (tandem duplication or dispersed transduplications)? 4) What genetic and genomic mechanisms underlie the production of high *EPSPS* copy numbers in these eight species? 5) Why has *EPSPS* gene duplication been observed to date only in the Amaranthaceae and Poaceae plant families? 6) Are CNVs more likely to arise independently in different populations of the same species, than to migrate via gene flow? The convergent evolution of CNVs for glyphosate resistance in weeds, through at least two mechanisms, may be indicative of a more general importance for this mechanism of adaptation in plants. CNVs warrant further investigation across plant functional genomics for adaptation to biotic and abiotic stresses, particularly for adaptive evolution on rapid time scales.

REFERENCES

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135-141.

Anthony N, Unruh T, Ganser D (1998) Duplication of the Rdl GABA receptor subunit gene in an insecticide-resistant aphid, *Myzus persicae*. Mol Gen Genet 260:165-175.

Baird DD, Upchurch RP, Homesley WB, Franz JE (1971) Introduction of a new broadspectrum postemergence herbicide class with utility for herbaceous perennial weed control. Proc North Central Weed Control Conf 26:64-68.

Bass C, Field LM (2011) Gene amplification and insecticide resistance. Pest Manag Sci 67:886-890.

Beard KE. 2014. Can very rapid adaptation arise without ancestral variation? Insight from the molecular evolution of herbicide resistance in genus *Amaranthus*. Clemson University, *All Dissertations*, 1797.

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Devel 15:621-627.

Bourguet D, Lenormand T, Guillemaud T, Marcel V, Fournier D, Raymond M (1997) Variation of dominance of newly arisen adaptive genes. Genetics 147:1225-1234.

Bradshaw LD, Padgette SR, Kimball SL, Wells BH (1997) Perspectives on glyphosate resistance. Weed Technol 11:189-198.

Cai X, Xu SS (2007) Meiosis-driven genome variation in plants. Curr Genomics 8:151-161.

Capy P, Gasperi G, Biémont C, Bazin C (2000) Stress and transposable elements: co-evolution or useful parasites? Heredity 85:101-106.

Chandi A, Milla-Lewis SR, Giacomini D, Westra P, Preston C, Jordan DL, York AC, Burton JD, Whitaker JR (2012) Inheritance of evolved glyphosate resistance in a North Carolina Palmer amaranth (*Amaranthus palmeri*) biotype. Int J Agron doi:10.1155/2012/176108.

Chatham LA, Wu C, Riggins CW, Hager AG, Young BG, Roskamp GK, Tranel PJ (2015a) *EPSPS* gene amplification is present in the majority of glyphosate-resistant Illinois waterhemp (*Amaranthus tuberculatus*) populations. Weed Technol 29:48-55.

Chatham LA, Bradley KW, Kruger GR, Martin JR, Owen MD, Peterson DE, Mithila J, Tranel PJ (2015b) A multistate study of the association between glyphosate resistance and *EPSPS* gene amplification in waterhemp (*Amaranthus tuberculatus*). Weed Sci 63:569-577.

Chen J, Huang H, Zhang C, Wei S, Huang Z, Chen J, Wang X (2015) Mutations and amplification of *EPSPS* gene confer resistance to glyphosate in goosegrass (*Eleusine indica*). Planta 242:859-868.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE (2012) Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. Science 338:1206-1209.

Culpepper AS, Grey TL, Vencill WK, Kichler JM, Webster TM, Brown SM, York AC, Davis JW, Hanna WW (2006) Glyphosate-resistant Palmer amaranth (*Amaranthus palmeri*) confirmed in Georgia. Weed Sci 54:620-626.

Cummins I, Wortley DJ, Sabbadin F, He Z, Coxon CR, Straker HE, Sellars JD, Knight K, Edwards L, Hughes D, Kaundun SS, Hutchings SJ, Steel PG, Edwards R (2013) Key role for a glutathione transferase in multiple-herbicide resistance in grass weeds. Proc Natl Acad Sci USA 110:5812-5817.

DeBolt S (2010) Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. Genome Biol Evol 2:441-453.

Devonshire AL, Field LM (1991) Gene amplification and insecticide resistance. Ann Rev Entomol 36:1-21.

Dillon AJ, Varanasi VK, Danilova T, Koo D-H, Nakka S, Peterson D, Tranel P, Friebe B, Gill BS, Jugulam M (2017) Physical mapping of amplified copies of the 5-enolpyruvylshikimate-3-phosphate synthase gene in glyphosate-resistant *Amaranthus tuberculatus*. Plant Physiol 173:1226-1234.

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148:1667-1686.

Drouaud J, Khademian H, Giraut L, Zanni V, Bellalou S, Henderson IR, Falque M, Mézard C (2013) Contrasted patterns of crossover and non-crossover at *Arabidopsis thaliana* meiotic recombination hotspots. PLoS Genet 9:e1003922.

Duhoux A, Carrère S, Gouzy J, Bonin L, Délye C (2015) RNA-Seq analysis of rye-grass transcriptomic response to an herbicide inhibiting acetolactate-synthase identifies transcripts linked to non-target-site-based resistance. Plant Mol Biol 87:473-487.

Duke SO, Powles SB (2008) Glyphosate: a once-in-a-century herbicide. Pest Manag Sci 64:319-325.

Emerson J, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science 320:1629-1631.

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. New Phytol 183:557-564.

Franssen AS, Skinner DZ, Al-Khatib K, Horak MJ, Kulakow PA (2001) Interspecific hybridization and gene flow of ALS resistance in *Amaranthus* species. Weed Sci 49:598-606.

Friesen LF, Beckie HJ, Warwick SI, Van Acker RC (2009) The biology of Canadian weeds. 138. *Kochia scoparia* (L.) Schrad. Can J Plant Sci 89:141-167.

Gaines TA, Shaner DL, Ward SM, Leach JE, Preston C, Westra P (2011) Mechanism of resistance of evolved glyphosate-resistant Palmer amaranth (*Amaranthus palmeri*). J Agric Food Chem 59:5886-5889.

Gaines TA, Ward SM, Bukun B, Preston C, Leach JE, Westra P (2012) Interspecific hybridization transfers a previously unknown glyphosate resistance mechanism in *Amaranthus* species. Evol Appl 5:29-38.

Gaines TA, Wright AA, Molin WM, Lorentz L, Riggins CW, Tranel PJ, Beffa R, Westra P, Powles SB (2013) Identification of genetic elements associated with *EPSPS* gene amplification. PLOS One 8:e65819.

Gaines TA, Barker AL, Patterson EL, Westra P, Westra EP, Wilson RG, Jha P, Kumar V, Kniss AR (2016) *EPSPS* gene copy number and whole-plant glyphosate resistance level in *Kochia scoparia*. PLOS ONE 11:e0168295.

Gaines TA, Lorentz L, Figge A, Herrmann J, Maiwald F, Ott MC, Han H, Busi R, Yu Q, Powles SB, Beffa R (2014) RNA-Seq transcriptome analysis to identify genes involved in metabolism-based diclofop resistance in *Lolium rigidum*. Plant J 78:865-876.

Gaines TA, Zhang W, Wang D, Bukun B, Chisholm ST, Shaner DL, Nissen SJ, Patzoldt WL, Tranel PJ, Culpepper AS, Grey TL, Webster TM, Vencill WK, Sammons RD, Jiang JM, Preston C, Leach JE, Westra P (2010) Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. Proc Natl Acad Sci USA 107:1029-1034.

Gardin JAC, Gouzy J, Carrere S, Delye C (2015) ALOMYbase, a resource to investigate non-target-site-based resistance to herbicides inhibiting acetolactate-synthase (ALS) in the major grass weed *Alopecurus myosuroides* (black-grass). BMC Gen 16:590.

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

Giacomini D, Westra P, Ward SM (2014) Impact of genetic background in fitness cost studies: An example from glyphosate-resistant Palmer amaranth. Weed Sci 62:29-37.

Giesy JP, Dobson S, Solomon KR (2000) Ecotoxicological risk assessment for Roundup® herbicide. *In* G. W. Ware, ed. Reviews of Environmental Contamination and Toxicology: Continuation of Residue Reviews. Springer New York, New York, NY. pp. 35-120.

Godar AS, Stahlman PW, Jugulam M, Dille JA (2015) Glyphosate-resistant kochia (*Kochia scoparia*) in Kansas: EPSPS gene copy number in relation to resistance levels. Weed Sci 63:587-595.

Goldsbrough PB, Hatch EM, Huang B, Kosinski WG, Dyer WE, Herrmann KM, Weller SC (1990) Gene amplification in glyphosate tolerant tobacco cells. Plant Sci 72:53-62.

Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. Plant Cell 28:388-405.

Hastings P, Bull HJ, Klump JR, Rosenberg SM (2000) Adaptive amplification: an inducible chromosomal instability mechanism. Cell 103:723-731.

Hastings P, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10:551.

Haughn G, Somerville C (1987) Selection for herbicide resistance at the whole-plant level. *In* H. Lebaron, et al., eds. Applications of Biotechnology to Agricultural Chemistry. ACS Publications, Washington, D.C. pp. 98-108.

Heap I The international survey of herbicide resistant weeds. Accessed May 23, 2017. Available on-line: www.weedscience.com. (2017).

Hemingway J (2000) The molecular basis of two contrasting metabolic mechanisms of insecticide resistance. Insect Biochem Mol Biol 30:1009-1015.

Hemingway J, Hawkes N, Prapanthadara L-a, Jayawardenal KI, Ranson H (1998) The role of gene splicing, gene amplification and regulation in mosquito insecticide resistance. Phil Trans Royal Soc London B: Biol Sci 353:1695-1699.

Herrmann KM, Weaver LM (1999) The shikimate pathway. Ann Rev Plant Phys Plant Mol Biol 50:473-503.

Hyppa RW, Smith GR (2010) Crossover Invariance Determined by Partner Choice for Meiotic DNA Break Repair. Cell 142:243-255.

Iwakami S, Endo M, Saika H, Okuno J, Nakamura N, Yokoyama M, Watanabe H, Toki S, Uchino A, Inamura T (2014) Cytochrome P450 CYP81A12 and CYP81A21 are associated with resistance to two acetolactate synthase inhibitors in *Echinochloa phyllopogon*. Plant Physiol 165:618-629.

Jensen S, Gassama M-P, Heidmann T (1999) Taming of transposable elements by homology-dependent gene silencing. Nat Genet 21:209-212.

Johnston D, Faulkner J (1991) Herbicide resistance in the Graminaceae—a plant breeder's view Herbicide Resistance in Weeds and Crops Butterworth-Heinemann, Oxford, UK. pp. 319-330.

Jugulam M, Gill BS (2017) Molecular cytogenetics to characterize mechanisms of gene duplication in pesticide resistance. Pest Manag Sci:In press.

Jugulam M, Niehues K, Godar AS, Koo D-H, Danilova T, Friebe B, Sehgal S, Varanasi VK, Wiersma A, Westra P, Stahlman PW, Gill BS (2014) Tandem amplification of a chromosomal segment harboring EPSPS locus confers glyphosate resistance in *Kochia scoparia*. Plant Physiol 166:1200-1207.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462-467.

Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714-8719.

Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521-529.

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Royal Soc B: Biol Sci 279:5048-5057.

Kumar V, Jha P (2015) Growth and reproduction of glyphosate-resistant and susceptible populations of *Kochia scoparia*. PloS One 10:e0142675.

Kumar V, Jha P, Reichard N (2014) Occurrence and characterization of kochia (*Kochia scoparia*) accessions with resistance to glyphosate in Montana. Weed Technol 28:122-130.

Kumar V, Jha P, Giacomini D, Westra EP, Westra P (2015) Molecular basis of evolved resistance to glyphosate and acetolactate synthase-inhibitor herbicides in kochia (*Kochia scoparia*) accessions from Montana. Weed Sci 63:758-769.

Küpper A, Borgato EA, Patterson EL, Netto AG, Nicolai M, Carvalho SJd, Nissen SJ, Gaines TA, Christoffoleti PJ (2017) Multiple resistance to glyphosate and acetolactate synthase inhibitors in Palmer amaranth (*Amaranthus palmeri*) identified in Brazil. Weed Sci 65:317-326.

Kwon D, Clark J, Lee S (2010) Extensive gene duplication of acetylcholinesterase associated with organophosphate resistance in the two-spotted spider mite. Insect Mol Biol 19:195-204.

Labbé P, Berticat C, Berthomieu A, Unal S, Bernard C, Weill M, Lenormand T (2007a) Forty years of erratic insecticide resistance evolution in the mosquito *Culex pipiens*. PLoS Genet 3:e205.

Labbé P, Berthomieu A, Berticat C, Alout H, Raymond M, Lenormand T, Weill M (2007b) Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the mosquito *Culex pipiens*. Mol Biol Evol 24:1056-1067.

Laforest M, Soufiane B, Simard M-J, Obeid K, Page E, Nurse RE (2017) Acetyl-CoA carboxylase overexpression in herbicide resistant large crabgrass (*Digitaria sanguinalis*). Pest Manag Sci:Early View.

Lämke J, Bäurle I (2017) Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. Genome Biol 18:124.

Li X, Schuler MA, Berenbaum MR (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu Rev Entomol 52:231-253.

Lorentz L, Gaines TA, Nissen SJ, Westra P, Strek H, Dehne HW, Ruiz-Santaella JP, Beffa R (2014) Characterization of glyphosate resistance in *Amaranthus tuberculatus* populations. J Agric Food Chem 62:8134-8142.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Malone JM, Morran S, Shirley N, Boutsalis P, Preston C (2016) EPSPS gene amplification in glyphosate-resistant *Bromus diandrus*. Pest Manag Sci 72:81-88.

Martin SL, Benedict L, Sauder CA, Wei W, da Costa LO, Hall LM, Beckie HJ (2017) Glyphosate resistance reduces kochia fitness: Comparison of segregating resistant and susceptible F2 populations. Plant Sci 261:69-79.

Mishra S, Whetstine JR (2016) Different facets of copy number changes: permanent, transient, and adaptive. Mol Cellular Biol 36:1050-1063.

Mohseni-Moghadam M, Schroeder J, Ashigh J (2013) Mechanism of resistance and inheritance in glyphosate resistant Palmer amaranth (*Amaranthus palmeri*) populations from New Mexico, USA. Weed Sci 61:517-525.

Molin WT, Wright AA, Lawton-Rauh A, Saski CA (2017a) The unique genomic landscape surrounding the *EPSPS* gene in glyphosate resistant *Amaranthus palmeri*: a repetitive path to resistance. BMC Gen 18:91.

Molin WT, Wright AA, VanGessel MJ, McCloskey WB, Jugulam M, Hoagland RE (2017b) Survey of the genomic landscape surrounding the EPSPS gene in glyphosate resistant Amaranthus palmeri from geographically distant populations in the United States. Pest Manag Sci: In press.

Nafziger ED, Widholm JM, Steinrucken HC, Killmer JL (1984) Selection and characterization of a carrot cell-line tolerant to glyphosate. Plant Physiol 76:571-574.

Nandula VK, Wright AA, Bond JA, Ray JD, Eubank TW, Molin WT (2014) *EPSPS* amplification in glyphosate-resistant spiny amaranth (*Amaranthus spinosus*): a case of gene transfer via interspecific hybridization from glyphosate-resistant Palmer amaranth (*Amaranthus palmeri*). Pest Manag Sci 70:1902-1909.

Ngo TD, Malone JM, Boutsalis P, Gill G, Preston C (2017) EPSPS gene amplification conferring resistance to glyphosate in windmill grass (*Chloris truncata*) in Australia. Pest Manag Sci: Early View.

Norsworthy JK, Griffith G, Griffin T, Bagavathiannan M, Gbur EE (2014) In-field movement of glyphosate-resistant Palmer amaranth (*Amaranthus palmeri*) and its impact on cotton lint yield: Evidence supporting a zero-threshold strategy. Weed Sci 62:237-249.

Ono M, Swanson JJ, Field LM, Devonshire AL, Siegfried BD (1999) Amplification and methylation of an esterase gene associated with insecticide-resistance in greenbugs, *Schizaphis graminum* (Rondani)(Homoptera: Aphididae). Insect Biochem Mol Biol 29:1065-1073.

Osipitan OA, Dille JA (2017) Fitness outcomes related to glyphosate resistance in kochia (*Kochia scoparia*): What life history stage to examine? Front Plant Sci 8:1090.

Padgette SR, Re DB, Barry GF, Eichholtz DE, Delannay X, Fuchs RL, Kishore GM, Fraley RT (1996) New weed control opportunities: Development of soybeans with a Roundup Ready gene. *In* S. O. Duke, ed. Herbicide-resistant crops: Agricultural, environmental, economic, regulatory, and technical aspects. CRC Press, Inc., Boca Raton, FL.

Powles SB, Yu Q (2010) Evolution in action: Plants resistant to herbicides. Annu Rev Plant Biol 61:317-347.

Powles SB, Lorraine-Colwill DF, Dellow JJ, Preston C (1998) Evolved resistance to glyphosate in rigid ryegrass (*Lolium rigidum*) in Australia. Weed Sci 46:604-607.

Price AJ, Balkcom K, Culpepper S, Kelton J, Nichols R, Schomberg H (2011) Glyphosate-resistant Palmer amaranth: a threat to conservation tillage. J Soil Water Cons 66:265-275.

Purandare SM, Patel PI (1997) Recombination hot spots and human disease. Genome Res 7:773-786.

Ramsey J, Schemske DW (1998) Pathways, mechanisms, and rates of polyploid formation in flowering plants. Ann Rev Ecol System 29:467-501.

Ranson H, Claudianos C, Ortelli F, Abgrall C, Hemingway J, Sharakhova MV, Unger MF, Collins FH, Feyereisen R (2002) Evolution of supergene families associated with insecticide resistance. Science 298:179-181.

Raymond M, Beyssat-Arnaouty V, Sivasubramanian N, Mouches C, Georghiou G, Pasteur N (1989) Amplification of various esterase B's responsible for organophosphate resistance in *Culex* mosquitoes. Biochem Genet 27:417-423.

Ribeiro DN, Pan Z, Duke SO, Nandula VK, Baldwin BS, Shaw DR, Dayan FE (2014) Involvement of facultative apomixis in inheritance of *EPSPS* gene amplification in glyphosate-resistant *Amaranthus palmeri*. Planta 239:199-212.

Rueppel ML, Brightwell BB, Schaefer J, Marvel JT (1977) Metabolism and degradation of glyphosate in soil and water. J Agric Food Chem 25:517-528.

Salas RA, Dayan FE, Pan Z, Watson SB, Dickson JW, Scott RC, Burgos NR (2012) *EPSPS* gene amplification in glyphosate-resistant Italian ryegrass (*Lolium perenne* ssp. *multiflorum*) from Arkansas. Pest Manag Sci 68:1223-1230.

Sammons DR, Gaines TA (2014) Glyphosate resistance: State of knowledge. Pest Manag Sci 70:1367-1377.

Schimke RT (1986) Methotrexate resistance and gene amplification: Mechanisms and implications. Cancer 57:1912-1917.

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, Bogwitz MR, Martin J, Daborn PJ, Goddard ME, Batterham P (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genet 6:e1000998.

Schönbrunn E, Eschenburg S, Shuttleworth WA, Schloss JV, Amrhein N, Evans JNS, Kabsch W (2001) Interaction of the herbicide glyphosate with its target enzyme 5-enolpyvuvylshikimate 3-phosphate synthase in atomic detail. Proc Natl Acad Sci USA 98:1376-1380.

Sellin C, Forlani G, Dubois J, Nielsen E, Vasseur J (1992) Glyphosate tolerance in *Cichorium intybus* L. var. Magdebourg. Plant Sci 85:223-231.

Shyr Y-YJ, Caretto S, Widholm JM (1993) Characterization of the glyphosate selection of carrot suspension cultures resulting in gene amplification. Plant Sci 88:219-228.

Slack A, Thornton PC, Magner DB, Rosenberg SM, Hastings PJ (2006) On the mechanism of gene amplification induced under stress in *Escherichia coli*. PLoS Genet 2:385-398.

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272.

Small GJ, Hemingway J (2000) Molecular characterization of the amplified carboxylesterase gene associated with organophosphorus insecticide resistance in the brown planthopper, *Nilaparvata lugens*. Insect Mol Biol 9:647-653.

Smith CM, Pratt D, Thompson GA (1986) Increased 5-enolpyruvylshikimic acid 3-phosphate synthase activity in a glyphosate-tolerant variant strain of tomato cells. Plant Cell Rep 5:298-301.

Sosnoskie LM, Culpepper AS (2014) Glyphosate-resistant Palmer amaranth (*Amaranthus palmeri*) increases herbicide use, tillage, and hand-weeding in Georgia cotton. Weed Sci 62:393-402.

Steinrücken HC, Amrhein N (1980) The herbicide glyphosate is a potent inhibitor of 5-enolpyruvylshikimic acid-3-phosphate synthase. Biochem Bioph Res Co 94:1207-1212.

Steinrücken HC, Schulz A, Amrhein N, Porter CA, Fraley RT (1986) Overproduction of 5-enolpyruvylshikimate-3-phosphate synthase in a glyphosate-tolerant *Petunia hybrida* cell line. Arch Biochem Biophys 244:169-178.

Suh H, Hepburn AG, Kriz AL, Widholm JM (1993) Structure of the amplified 5-enolpyruvylshikimate-3-phosphate synthase gene in glyphosate-resistant carrot cells. Plant Mol Biol 22:195-205.

Thomas J, Pritham EJ (2015) *Helitrons*, the eukaryotic rolling-circle transposable elements. *In* N. Craig, et al., eds. Mobile DNA, 3rd edition. American Society of Microbiology, Washington, DC. pp. 893-926.

Trucco F, Jeschke MR, Rayburn AL, Tranel PJ (2005a) *Amaranthus hybridus* can be pollinated frequently by *A. tuberculatus* under field conditions. Heredity 94:64-70.

Trucco F, Jeschke MR, Rayburn AL, Tranel PJ (2005b) Promiscuity in weedy amaranths: High frequency of female tall waterhemp (*Amaranthus tuberculatus*) x smooth pigweed (*A. hybridus*) hybridization under field conditions. Weed Sci 53:46-54.

Trucco F, Tatum T, Rayburn AL, Tranel PJ (2009) Out of the swamp: unidirectional hybridization with weedy species may explain the prevalence of *Amaranthus tuberculatus* as a weed. New Phytol 184:819-827.

Varanasi, V. K., Godar, A. S., Currie, R. S., Dille, A. J., Thompson, C. R., Stahlman, P. W., and Jugulam, M. (2015). Field-evolved resistance to four modes of action of herbicides in a single kochia (Kochia scoparia L. Schrad.) population. Pest management science, 71(9), 1207-1212.

Vila-Aiub MM, Han H, Jalaludin A, Yu Q, Powles SB (2017) Adaptive value of single and double spontaneous EPSPS mutations imparting glyphosate resistance in the agroecosystem. Global Herbicide Resistance Challenge Proceedings 2:78.

Vila-Aiub MM, Goh SS, Gaines TA, Han H, Busi R, Yu Q, Powles SB (2014) No fitness cost of glyphosate resistance endowed by massive *EPSPS* gene amplification in *Amaranthus palmeri*. Planta 239:793-801.

Vila-Aiub MM, Neve P, Powles SB (2009) Fitness costs associated with evolved herbicide resistance alleles in plants. New Phytol 184:751-767.

Vontas JG, Graham J, Hemingway J (2001) Glutathione S-transferases as antioxidant defence agents confer pyrethroid resistance in *Nilaparvata lugens*. Biochem J 357:65-72.

Waite J, Thompson CR, Peterson DE, Currie RS, Olson BLS, Stahlman PW, Al-Khatib K (2013) Differential kochia (*Kochia scoparia*) populations response to glyphosate. Weed Sci 61:193-200.

Wang Y, Jones JD, Weller SC, Goldsbrough PB (1991) Expression and stability of amplified genes encoding 5-enolpyruvylshikimate-3-phosphate synthase in glyphosate-tolerant tobacco cells. Plant Mol Biol 17:1127-1138.

Watanabe T, Tanabe H, Horiuchi T (2011) Gene amplification system based on double rolling-circle replication as a model for oncogene-type amplification. Nuc Acids Res 39:e106.

Widholm JM, Chinnala AR, Ryu JH, Song HS, Eggett T, Brotherton JE (2001) Glyphosate selection of gene amplification in suspension cultures of 3 plant species. Physiol Plantarum 112:540-545.

Wiersma AT, Gaines TA, Preston C, Hamilton JP, Giacomini D, Buell CR, Leach JE, Westra P (2015) Gene amplification of 5-enol-pyruvylshikimate-3-phosphate synthase in glyphosate-resistant *Kochia scoparia*. Planta 241:463-474.

Williams GM, Kroes R, Munro IC (2000) Safety evaluation and risk assessment of the herbicide Roundup and its active ingredient, glyphosate, for humans. Reg Toxic Pharma 31:117-165.

Wu C, Davis AS, Tranel PJ (2017) Limited fitness costs of herbicide-resistance traits in *Amaranthus tuberculatus* facilitate resistance evolution. Pest Manag Sci:In press.

Yu Q, Jalaludin A, Han H, Chen M, Sammons RD, Powles SB (2015) Evolution of a double amino acid substitution in the 5-enolpyruvylshikimate-3-phosphate synthase in *Eleusine indica* conferring high-level glyphosate resistance. Plant Physiol 167:1440-1447.

Zhou Z-H, Syvanen M (1997) A complex glutathione transferase gene family in the housefly *Musca domestica*. Mol Gen Genet 256:187-194.

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant genomes. Theor Appl Genet 127:1-18.

CHAPTER 2: THE GENOME OF KOCHIA SCOPARIA

Exploring Copy Number Variation in the Kochia Genome[2]

Summary

Kochia scoparia (kochia) is an important weed that has evolved resistance to several herbicides, chief among them is glyphosate. Resistance to glyphosate is conferred by gene copy duplication of the target gene 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*). We set out to understand the extent to which copy number variation (CNV) may exist at additional loci within the kochia genome. In this work, we generated the first assembly of the kochia genome from a combination of Illumina and PacBio data and then resequenced a glyphosate resistant line. We discovered hundreds of putative CNV events, but copy number exhibited little correlation with gene expression levels as measured by RNA-seq, indicating that transcriptional regulation may often supersede any expressional differences that could be produced by CNV. We also discovered that the only family of genes enriched in the glyphosate resistant line is a class of transposons, known as Fhy/FAR1 mutator-like transposases. These genes, thought to be "domesticated transposons" seem to still be actively duplicating and may be co-selected with *EPSPS* gene duplication or increasing activity in response to glyphosate pressure.

Introduction

Copy number variation (CNV) is known to be an important source of novel genetic variation (Flagel and Wendel 2009). In plants, CNVs have been found that faciltate resistance or

[2] Eric L. Patterson, Christopher A. Saski, Daniel B. Sloan, Karl Ravet, Pat J. Tranel, Philip Westra, Todd A. Gaines

tolerence to various abiotic and biotic stresses including heat, pathogenic nematodes, and continuous asexual reproduction (Debolt 2010; Cook et al. 2012; Hardigan et al. 2016). Stress, and in particular abiotic stress, has been shown to induce CNV events (Slack et al. 2006; Hull et al. 2017). There is also evidence that CNV events are not random; either they 1) can occur selectively to amplify particular genes or gene families more often than others or 2) occur at random intially but then are selected quickly so that some genes are more likely to remain duplicated then others. (Debolt 2010; Hull et al. 2017).

The plant species *Kochia scoparia* (kochia) is one of the most troublesome weeds in western United States (Casey 2009). Since its introduction from Eurasia, kochia has rapidly adapted to the high plains, developing tolerance to several abiotic stresses such as high salt, cold, and drought. Additionally, kochia has evolved resistance to several herbicide modes of action including acetolactate inhibitors, photosystem II inhibitors, synthetic auxins, and the 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) inhibitor, glyphosate (Foes et al. 1999; Cranston et al. 2001; Preston et al. 2009; Waite et al. 2013). Glyphosate resistance was first reported in kochia within a decade of Roundup™ ready technology introduction (Waite et al. 2013, Varanasi et al. 2015). This widespread herbicide resistance is a perfect example of evolution in action and underscores kochia's ability to rapidly adapt to new abiotic stresses (Beckie et al. 2012).

Recently, it was discovered that kochia is glyphosate resistant by way of *EPSPS* CNV (Wiersma et al. 2014; Gaines et al. 2016; Jugulam et al. 2014). Increased copy number of *EPSPS* results in the over-production of the EPSPS protein and therefore more glyphosate needs to be applied for the same lethal effect. Kochia is not the only plant to use *EPSPS* copy number variation to become glyphosate resistant. At least seven other, divergent species have evolved

EPSPS copy number increases to become resistant to glyphosate (Patterson et al. 2018). The mechanism of *EPSPS* duplication is only understood in three species, *Amaranthus palmeri*, *Amaranthus tuberculatus* and kochia, and it is clear that each has generated copy number increases by different mechanisms (Koo et al. 2018a; Koo et al. 2018b; Patterson et al. 2018).

In this paper, we sequenced the genome of a glyphosate susceptible *Kochia scoparia* line and compared it to whole genome resequencing data from a glyphosate resistant line. Using *EPSPS* as a positive control for novel CNV discovery, we identify other, novel genomic rearrangements between these lines and correlate genome resequencing data to changes in the gene expression of these new CNVs. This study provides a genomics platform for investigations into kochia's unique biology and explores CNV between a glyphosate resistant and susceptible line.

Methods

Tissue Collection and DNA Extraction

The herbicide-susceptible *K. scoparia* 7710 line (Preston et al. 2009; Pettinga et al. 2017) was used for genomic sequencing. Plants in this line were killed by glyphosate treatments at field rates of 860 g a.e. ha^{-1}. Plants were grown in a greenhouse at Colorado State University. After seeds germinated, they were transferred into 1-gallon pots filled with Fanfard 4P Mix supplemented with Osmocote fertilizer (Scotts Co. LLC), regularly watered, and grown under a 16-hour photoperiod. Temperatures in the greenhouse cycled between 25 ℃ day and 20 ℃ nights. A single, healthy individual was selected for bulk tissue collection. Several grams of leaf tissue were homogenized in liquid nitrogen using a pestle and mortar.

A glyphosate resistant line (M32) was developed from a field population in Akron, Colorado (40.162382, -103.172849) in the Fall of 2012. After glyphosate failed to control these plants in the field, seed was collected and brought to the lab. Seeds were germinated and treated with 860 g a.e. ha^{-1} of glyphosate mixed with ammonium sulfate (2% w/v). Survivors were then collected, crossed and seed was collected. This process was repeated for three generations until no susceptible individuals were observed. All plants were confirmed to have elevated *EPSPS* copy number using genomic qPCR (Gaines et al. 2016).

For Illumina sequencing of the two lines, DNA was extracted from samples using a modified CTAB extraction protocol that is described in Doyle 1991. First, 500 µl of extraction buffer (100 mM tris, 1.4 M NaCl, 20 mM EDTA, pH 8.0, 2% CTAB, 0.3% mercaptoethanol) with 5mg polyvinylpyrrolidone (PVP) was mixed with the tissue aliquots. The suspension was homogenized and incubated at 60 ℃ for 15 min. Next, 500 µl of chloroform:isoamyl alcohol (24:1) was added and the tubes were gently agitated on an orbital mixer for 15 min. The tubes were then centrifuged at 8000 rcf for 15 min and the top, aqueous phase was moved to a new tube. One µl of RNase A was added and incubated at 37 ℃ for 1 hour. The chloroform:isoamyl alcohol separation was performed again and the aqueous phase retained again. DNA was then precipitated by adding 1/10 volume 5M sodium acetate, pH 8 and three volumes of 100% ethanol. The samples were then centrifuged at 10,000 rcf for 10 min. The supernatant was poured off and the resulting pellet was rinsed with 70% ethanol and then allowed to dry. The pellet was re-suspended in 100 µl of water, checked for concentration and purity on a Nanodrop T1000, and sent for Illumina sequencing at the Roy J. Carver Biotechnology Center at The University of Illinois at Urbana-Champaign.

For large-fragment, PacBio sequencing of the glyphosate susceptible line, the CTAB protocol described above was modified to obtain more DNA of sufficiently large size (>10kb). Approximately one gram of finely chopped kochia young leaf tissue was added to 50ml conical tubes. To this tissue 15ml of CTAB extraction buffer and 60μg of PVP was added, mixed, and allowed to incubate for 30 min at 50 °C. The tubes were then centrifuged at 3600 rcf for 10 min. The liquid phase was separated into a new tube and 15 mL of chloroform:isoamyl alcohol (24:1) was added and mixed by inversion. They were then centrifuged at 3600 rcf for 10 min more and the upper phase transferred to a new tube. To this 4 µl of RNase A was added and incubated for 30 min at 37 °C. The chloroform:isoamyl alcohol separation was repeated and the final aqueous phase collected. The DNA was precipitated by adding 3 volumes of EtOH and 0.5 volume of NaCl 5M. The tubes were then incubated at -20 °C for 30 min, centrifuged at 3600 rcf for 10 min, and the pellet washed with 70% ethanol. The final pellets were dried and re-suspended in 2 ml of Tris-EDTA buffer. The DNA was further purified using the Genomic DNA Clean & Concentrator™-10 kit by Zymo, following the recommended protocol. The final DNA was then pooled, checked for purity, concentration, and size and sent to UC Davis Genome Center, DNA Technologies & Expression Analysis Cores at The University of California, Davis for PacBio sequencing.

Illumina and PacBio Sequence Data for Susceptible Genome Assembly

Three libraries of glyphosate susceptible kochia DNA were prepared for Illumina sequencing on a HiSeq 2500 at the University of Illinois, Biotechnology Center : 1) A high coverage 150bp, paired-end library on one full flow cell (8 lanes), 2) a 150bp, 5kb mate pair library (1 lane), and 3) a 150bp, 10kb mate pair library (1 lane). Quality of the raw sequence reads were assessed using FASTQC v0.10.1. Adapters were removed using Trimmomatic version 0.60 with the parameters "ILLUMINACLIP: tranel_adaptors.fa:2:30:10 TRAILING:30 LEADING:30 MINLEN:45" using these adaptors: "AGATCGGAAGAGCAC" and

“AGATCGGAAGAGCGT” to identify and remove adaptors as well as accepting trimmed sequences with a minimum length of 45.

DNA sent for PacBio sequencing was checked for quality using a NanoDrop 2000c and quantified using Qubit. Large insert DNA library was generated using the PacBio SMRT Library Prep for RSII followed by BluePippin size selection for fragments >10kb. The library was equally loaded across 12 Pac-Bio SMRT cells using the RSII chemistry after a titration cell to determine optimal loading. In total, 2,760,348 PacBio reads were generated with a read N50 of 6,576 bp with the largest read being 41,738 bp.

One hundred gigabases of Illumina data from each of the high-coverage 240bp kochia library, the *Arabidopsis thaliana* genome project, *Beta vulgaris* genome project, and *Amaranthus hypochodriachus* genome project were analyzed using Khmer to generate k-mer frequency distributions of 24-mers (Crusoe et al. 2015).

Genome Assembly

Two different assemblies were generated that integrated the PacBio and Illumina data. These two assemblies were then compared and merged by consensus for a single final assembly. For the first assembly, raw PacBio reads were error corrected using the high coverage 240 bp, paired-end Illumina library with the error correcting software Proovread 2.13.11 (Hackl et al. 2014). Proovread was run with standard parameters, using the high coverage 150 bp, paired-end Illumina library on each SMRT cell individually. Error corrected reads were then assembled using the Celera Assembler fork for long reads, Canu 1.0 (Koren et al. 2017). Canu was run with a predicted genome size of 1 Gb, and the PacBio-corrected settings. For the second assembly, an initial ALLPATHS-LG assembly was made with all three Illumina libraries (Butler et al. 2008). ALLPATHS was run assuming a haploid genome of 1 Gb. The resulting contigs were then scaffolded using the uncorrected PacBio reads using the software PBJelly 15.8.24 (English et al. 2012). PBJelly was run with the following blasr settings: -"minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 19 –noSplitSubreads". The two assemblies were then merged with GARM Meta assembler 0.7.3 to get a final genome assembly (Soto-Jimenez, Estrada, and Sanchez-Flores 2014). The final assembly from ALLPATHS was set to assembly "A" and the final Assembly from Canu was set as genome "B." All other parameters were kept standard.

Genome Annotation and the Arrangement of Contigs into Pseudomolecules

The merged assembly was annotated with the WQ-Maker 2.31.8 pipeline in conjunction with CyVerse (Cantarel et al. 2008; Thrasher et al. 2014). WQ-Maker was informed with the *Kochia scoparia* transcriptome developed by Wiersma et al. 2014, all expressed sequence tags (ESTs) from the Chenopodiaceae downloaded from NCBI, all protein sequence from the Chenopodiaceae family downloaded from NCBI, and Augustus using *Arabidopsis thaliana* gene models. The resulting predictions were then used to train SNAP (2013-02-16) through two rounds for final gene model predictions. Gene space completeness was assessed using BUSCO v3 using standard parameters (Simão et al. 2015).

The predicted gene transcripts (mRNA) and predicted translated protein sequence was then annotated using Basic Local Alignment Search Tool (BLAST) Nucleotide (BlastN) and Protein (BlastP) 2.2.18+ for similarity to known transcripts and proteins, respectively. Alignments were made to the entire NCBI nucleotide and protein databases. For all Blast homology searches the e-value was set at 1e^{-25} and only the best match was considered. Additionally, the predicted proteins were further annotated using InterProScan 5.28-67.0 for protein domain predictions (Camacho et al. 2009; Mi et al. 2005; Jones et al. 2014). InterProScan was run using standard settings. The complete assembly was analyzed using RepeatMasker 4.0.6 to search for small interspersed repeats, DNA transposon elements, and other known repetitive elements using the "viridiplantae" repeat database and standard search parameters (Tarailo-Graovac and Chen 2009).

The contigs in the final kochia genome assembly were aligned against the 9 chromosomes of the *Beta vulgaris* genome (accessed from NCBI on 11-20-17) using NUCmer from the Mummer 3.0 package (Kurtz et al. 2004) using standard parameters. Kochia scaffolds were then arranged in the order that maintained maximum synteny with the *Beta vulgaris* pseudomolecules using the maximal unique matches (Mums) from NUCmer. Mums were arranged by start/stop basepair from the *Beta vulgaris* assembly and the corresponding scaffold in kochia was moved into order.

Illumina Sequence Data for Resistant Line Resequencing and CNV discovery

DNA from the glyphosate resistant line was prepared for Illumina sequencing using Genomic DNA Sample Prep Kit from Illumina following the manufacturer's protocols. The library was sequenced on one entire lane of a HiSeq 2500 flow cell. Reads were aligned to the final genome assembly using the BWA-backtrack alignment program using standard parameters (Li and Durbin 2009). The resulting alignment was then analyzed using the software CNVnator v0.3.2 with a 1000bp sliding window to screen for large CNVs that have the potential for harboring genes (Abyzov et al. 2011). The output was then subjected to two filtering criteria: 1) a normalized read depth (nrd) of >2 or <0.5 above/below background, and 2) the presence of at least one entire gene model within the boundaries of the putative CNV.

To correct for the fact that our assembly of the reference genome is not complete and potentially contains collapsed repeats, the Illumina data from the initial assembly of the susceptible line were aligned back to the assembly. Read depth was then calculated for all genes. Genes that had read depths of >2 or <0.5 above background from this control alignment were removed from further analysis, as they were most likely not truly different between the resistant and susceptible line, but merely an artifact of having an incomplete reference.

Measuring Differential Gene Expression

RNA was extracted from young leaf tissue from four plants from each of the glyphosate susceptible and resistant lines using the Qiagen RNeasy Plus Mini Kit. Each replicate sample was normalized to a total mass of 200ng total RNA. Strand-specific RNA-seq libraries were prepared robotically on a Hamilton Star Microlab at the Clemson University Genomics and Computational Facility following in-house automation procedures and generally the TruSeq Stranded mRNAseq preparation guide. The prepared libraries were pooled and 100 bp paired end reads were sequenced using a NextSeq 500/550. Reads were aligned to the gene models from the genome assembly using the mem algorithm from the BWA alignment program version 0.7.15 under standard parameters. Read counts for each gene were extracted from this alignment using the software featureCounts in the Subread 1.6.0 package and the gene annotation generated by WQ-Maker (Liao, Smyth, and Shi 2014). Expression level and differential expression between the glyphosate susceptible and glyphosate resistant plants for all genes was calculated using the EdgeR package using the quasi-likelihood approach in the generalized linear model (glm) framework as described in the user manual (Robinson, McCarthy, and Smyth 2010). These expression data were then correlated with the read depth from the genome resequencing and the list of putative CNVs.

Results

K-*mer Analysis and Assembly Statistics*

The *k*-mer distribution graphs from unassembled Illumina data of the susceptible line from kochia showed a tri-modal distribution rather than the uni-modal distribution observed in *Arabidopsis thaliana*, *Amaranthus hypochondriachus*, and *Beta vulgaris* (Figure 2.1). The *Beta vulgaris k*-mer distribution exhibited a small, yet noticeable, second mode in its distribution. The second and third modes are comprised of *k*-mers that appear at approximately two or three times the abundance of the *k*-mers in the first mode. This indicates a high abundance of duplicated and triplicated sequence in the Illumina dataset.

Two approaches were used to integrate Illumina and PacBio data, and these two approaches were then consolidated into a single final assembly. This final assembly consisted of 19,671 scaffolds, spanning ~711Mb. The longest scaffold was 770kb and the N50 was ~62kb for this final assembly. Approximately 9.43% of the base pairs were unknown "N" bases that serve only as scaffolding and distance information (Table 2.1).

After annotation, 47,414 genes were predicted with an average transcript length of 943bp (Table 2.2) in kochia, compared to the 27,429 in *Beta vulgaris*. These genes were analyzed using BUSCO for completeness, which found 1,103 out of 1,440 (76.6%) ultra-conserved genes represented in the dataset (Table 2.3). Genes were then annotated by homology using BLASTN and BLASTP against the NCBI nucleotide and protein databases respectively and the predicted proteins were analyzed using InterProScan to classify functional protein domains. Approximately 62% of predicted kochia genes found one or more matches in the NCBI database(s) using a e value < 1 e^{-25} and almost 82% of predicted proteins were prescribed one or more functional InterPro domain(s) (Table 2.2). RepeatMasker uncovered 6.25% of the genome assembly consisting of interspersed repeats with the largest proportion being LTR elements of either the Ty1/Copia or Gypsy/DIRS1 variety. Simple repeats made up approximately 2.5% of the assembly (Table 2.4). For comparison, in the assembly of *Beta vulgaris* 252 Mb (42.3%) of the genome assembly consisted of repetitive DNA, with gypsy-like LTR retrotransposons making up 57.34 Mb (22%) of that repetitive content (Dohm et al. 2014)

Conservation of Synteny with Beta Vulgaris

Mummer was used to align the *Beta vulgaris* and kochia assemblies; finding regions that were >80% similarity for >500 bp ("links"). Mummer calculated 13,573 links between the kochia and *Beta vulgaris* assemblies spanning 364.5Mb in 5,451 contigs from the kochia assembly. These links were used as anchors for our kochia contigs that were then merged into pseudomolecules in the order of maximum synteny. Of the 13,573 links, 3,212 links connected to chromosomes outside of the pseudomolecule in which the kochia contig was placed. These breaks from synteny are non-resolvable without breaking the overall synteny between the kochia pseudomolecules and *Beta vulgaris* chromosomes (Figure 2.2).

Discovering novel CNVs between glyphosate resistant and susceptible lines

Shotgun Illumina sequence from the glyphosate-resistant kochia population was used to discover novel CNVs. This glyphosate resistant line was used for several reasons. First, the glyphosate resistant line is well characterized and has been inbred in the greenhouse for three generations and is no longer segregating for glyphosate resistance. This helps reduce individual variation in our analysis. Second, the well characterized *EPSPS* CNV served as a positive control for the discovery of novel CNVs.

CNVnator was used to identify regions with deviations in normalized read depth (nrd) of 2× or 0.5×. CNVnator initially predicted 3,522 CNV events with a >2× nrd and 11,012 CNV events with <0.5× nrd. Next, Illumina reads from the susceptible line were aligned to the reference and CNV events were called, as these could account for many false positives. After these were removed from the analysis, 2,802 CNV events had a >2× nrd and these regions contained 3,918 genes while 7,147 CNV events had a <0.5× nrd and containing 9,235 genes. The average length of all CNV events was ~13.5 kb (Table 2.5). CNVnator predicts the *EPSPS* duplication with high confidence (p-value <0.0001). The EPSPS CNV was approximately 62kb in length and consisted of 7 genes.

The InterPro IDs assigned to all the genes in this filtered list of putative CNVs were summed for events with >2× and <0.5× nrd. The most common term associated with genes within events with >2× nrd was IPR005162: Retrotransposon gag domain, while for genes within events with <0.5× nrd it was IPR012337: Ribonuclease H-like domain. Most of the top terms associated with but events with >2× and <0.5× nrd are also the top terms for the genome as a whole. Five terms have a higher proportion in events with >2× nrd then in the background genome annotation. The terms IPR005162 (Retrotransposon gag domain), IPR021109 (Aspartic peptidase domain), IPR031052 (FHY3/FAR1 family), IPR007527 (Zinc finger, SWIM-type), and IPR004330 (FAR1 DNA binding domain) are over-represented only for the genes within events with >2× nrd (Table 2.6).

We looked specifically at the loci annotated as Fhy/FAR-like genes. In the genome annotation, 578 loci are described with the InterPro ID IPR031052: FHY3/FAR1 family. Of those, 89 were indicated to be potential CNVs, with either increased or decreased nrd (Table 2.6). Of the 89 loci that were significant as potential CNVs, only 5 loci had <0.5× nrd while the remaining 84 showed >2× nrd (Figure 2.5). The resequencing read depth of these genes did not correlate with the expression of the Fhy/FAR-like genes (r = 0.079, p = 0.45).

Potential impact of novel CNVs on the transcriptome

We wanted to test the extent to which these putative CNVs influenced the expression of the genes contained within them. We measured the differential expression (DE) between our glyphosate resistant and susceptible lines by performing an RNA-seq with the gene models from the assembly. We then correlated the DE with the predicted CNV read depth and the CNVnator output. We used the refined list of putative CNVs (as defined above) and applied an additional cutoff of P-value <0.01 for both read depth from CNVnator and for differential expression from EdgeR. After filtering for P-Value, 489 genes within events with >2× nrd and 1,189 genes within events with <0.5× nrd remained. We saw little to no correlation between nrd and gene expression (r = 0.406, p = 0.096). One of the genes in the *EPSPS* CNV had low expression in both resistant and susceptible plants and was removed due to a DE p-value <0.01. Another showed over-expression but not to the extent predicted based on its genomic read depth. The final gene, despite being clearly co-duplicated with EPSPS, showed significantly decreased expression (Figure 2.3,2.4).

Discussion

K-mer Analysis and Assembly Statistics

Initial Illumina data and its corresponding *k*-mer curve show the potential for an extensive amount of sequence duplication and triplication in kochia. This *k*-mer distribution predicted that the genome of kochia would be 2.8 Gb; however, flow cytometry confirmed that the genome is ~1Gb. After assembly, we saw little evidence of extensive sequence duplication or repetitive regions. It could be that the repetitive elements are not resolved in the assembly or repetitive elements are large and during assembly these regions collapsed and appear as a single element when they are duplicated in the susceptible line. To test this, the Illumina data generated for the genome assembly from the susceptible line was realigned back to the assembly. This analysis revealed many such regions that were then removed from analysis as they were invariant regions between the two lines and most likely due to missing/collapsed regions in the assembly.

The final assembly accounted for only ~75% of the expected gene space as predicted by BUSCO and 83% of the predicted total genome size. Annotation of this assembly using WQ-Maker predicted 47,414 gene models, which is ~13,000 more than its relative, *Beta vulgaris*. Only ~80% of the genes were prescribed some sort of meaningful annotation by homology with proteins from the NCBI database or protein domain prediction and InterPro. We see great room for improvement of this initial assembly. The limited amount of PacBio data available means there are still regions with potentially high error rates and the more complex repetitive regions we are interested may still be missing from the assembly. In future drafts of the kochia genome, we hope to improve annotation by having higher accuracy sequence, with better homology to known genes in other species and with known protein domains. Additionally, kochia annotation will improve as related genomes such as *Beta vulgaris*, *Spinacia oleracea,* and *Chenopodium quinoa* become better annotated.

Beta vulgaris was used to order the contigs from our kochia assembly as kochia's nearest relative with a complete genome. We expected that there are large differences in the overall structure and order of the genomes as the genera of these two species are quite divergent (Muller and Borsch 2005); however, gene order is strongly conserved when the largest contigs from kochia are aligned against *Beta vulgaris* .

Discovering Novel CNVs in a Glyphosate Resistant Line

Having a kochia genome assembly allows us to not only understand the CNV event that led to glyphosate resistance, but also the effect that glyphosate selection has had on the genome at loci distal to the *EPSPS* locus. If generating novel CNVs provides an evolutionary advantage for glyphosate tolerance and resistance, then the plants may be generating other, novel CNVs inadvertently and these rearrangements may be co-selected with the *EPSPS* gene duplication. By resequencing a glyphosate resistant line, we could detect regions with high or low nrd. Thousands of genes were discovered in these regions with >2× nrd and <0.5× nrd. In future work, molecular validation of these CNVs will be critical for calculating the number of false positives as well as for determining the possibility for physiological impacts of these CNVs.

To understand the types of genes that were within these variable regions, we classified all genes using their corresponding InterPro IDs. It became apparent from this analysis that genes with some InterPro IDs appear more frequently in both high and low nrd areas; however, genes annotated with these IDs are also usually more abundant in the overall annotation. Genes annotated with IPR012337: ribonuclease H-Like domain, IPR005135: endonuclease, IPR026960: reverse transcriptase domain, and IPR025558: DUF 4283 were common in both high and low nrd events, but were also common at the same proportion in the genome annotation as a whole. Several of these InterPro IDs are associated with mobile elements, which is not surprising considering the amount of variation retroelements often show between individuals. It is interesting, however, that some ID terms were more common in events with >2× nrd. This includes genes annotated with IPR001878: Zinc finger/CCHC, IPR031052: FHY3/FAR1, IPR007527: Zinc Finger-Swim type elements, and IPR004332: transposase, MuDR (Mutator transposable elements).

The FAR1 family of proteins have very similar structure to mutator-like transposases, including an N-terminal zinc finger domain, a central transposase domain, and a C-terminal SWIM domain (Wand and Xing 2002). Often, a single gene is annotated with all four of these

InterPro IDs, therefore the over-representation of these four domains is the over-representation of a single family of mobile elements; the Fhy/FAR1 mutator-like transposases. Why these elements occur more consistently in areas with >2× nrd in the glyphosate resistant line is unclear. Generally, this class of proteins are thought of as transcriptional regulators that change gene expression in response to light (Wang and Xing 2002; Hudson, Lisch, and Quail 2003; Allen et al. 2006; Rongcheng Lin et al. 2007; R. Lin et al. 2008; W Tang et al. 2012). Evolutionarily, the Fhy/FAR genes are MULE transposases that have been "domesticated" to have a functional role in gene regulation. In fact, they are the only transposon-like gene with known host function (Alzohairy et al. 2013).

Potential impact of novel CNVs on the transcriptome

The power of CNV events to provide potential phenotypic advantages lies in the ability to over- and under-express genes within the boundaries of the event. Additionally, newly generated gene copies can sub- and neofunctionalize as they accumulate mutations (Flagel and Wendel 2009; Lynch and Conery 2000). Recent or young CNVs can be an effective way of changing expression because they keep the promoter of the duplicated genes intact and, therefore, the new gene copies maintain their regulatory network. Theoretically, doubling the number of copies of a gene should double the transcriptional output; however, there are many post-transcriptional activates that modulate gene expression or even repress it entirely. Additionally, different epigenetic signals on each copy may differentially regulate transcriptional output of each gene copy. With EPSPS, transcriptional and protein output is correlated with gene copy number (Gaines et al. 2016). However, eventually a physiological max is achieved and additional EPSPS protein no longer has a physiological benefit and EPSPS protein production in regulated (Gaines et al. 2016).

We performed an RNA-Seq experiment using young leaf tissue from four daughters of the glyphosate resistant plant used for Illumina resequencing versus four plants from the line used for genome assembly to test the expression of all genes contained within predicted CNV events. As expected, EPSPS and three of the other genes contained within that CNV event all showed positive correlation between over-expression and enhanced nrd; however, at a genome-wide level, nrd was not correlated with over-expression of genes. In fact, it was often the case that a gene would have >2x nrd but was under-expressed or *vice versa*.

We believe several things may account for this phenomenon. First, we may be incorrectly identifying CNV events or we are not applying strong enough criterion for determining a true CNV. In previous experiments, CNVnator results were verified empirically using comparative genomic hybridization and it was found that it can have a false discovery rate between 3-20% (Abyzov et al. 2011). We tried to reduce the number of false positives by only looking at CNV events with P-values less than that of the EPSPS CNV event; however, even these events showed no correlation between expression and nrd (Figure 2.4). As *in silico* predictions can vary greatly from reality, especially for CNV prediction, empirical molecular validation by quantitative PCR is needed so that a true false discovery rate can be calculated in future research. Second, overexpression of genes leads to gene silencing. For instance, it has been shown that inserting many transgenes under constitutive promoters into a single individual can lead to suppression of transgene expression, most likely due to RNA silencing (Finnegan and McElroy 1994; Wei Tang, Newton, and Weidner 2007; Vaucheret et al. 1998). Third, other regulatory machinery may override expression differences from changes in gene copy number. Since all regulatory machinery is still intact after a CNV event, genes are still subject to promoter based modifications to expression. If there are line specific differences in expression, a CNV event may not be enough to overcome the regulatory network in place. It may also be that the novel CNV events in the glyphosate resistant line initially led to correlated changes in expression but these plants quickly develop transcriptional regulatory machinery to compensate for what might be harmful changes in expression. Fourth, we only observed a single time point so gene expression that is regulated at different life stages or in different tissues may be masked by the tissue and time of sample collection. Finally, the individuals sampled for DE may have copy number variation among sibling plants rather than strictly between the two lines and/or they may not be representative of the re-sequenced plant. If a CNV event is different between individuals within each line (i.e., between siblings) then differential expression between lines becomes difficult to assess.

Conclusion

There exists a growing body of evidence that CNVs can be very important in adaptive evolution. Much work has been done in animal systems especially in human genetics, where somatic variation of CNVs have been repeatedly found to cause cancer (Schimke, Hill, and Johnston 1985; Xi et al. 2011). In insect systems however, it is known that CNVs have great potential for resistance to insecticides (Bass and Field 2011). It is clear that in some systems CNVs can be harmful and potentially lethal but in other systems, CNVs can offer an adaptive advantage. Many weed species, including kochia are r-selected species and produce thousands of offspring (Sakai et al. 2001; Pianka 1970); therefore, rearrangements that cause severe defects or that are lethal can be tolerated in the population if a few offspring get a sufficient evolutionary advantage, such as the case of glyphosate resistance.

By using both Illumina and PacBio data we assembled a draft of the kochia genome to serve as a platform to begin understanding how CNVs may be shaping kochia’s evolution and physiology. Even though the draft remains fragmented, we discovered novel CNVs by genomic resequencing of a glyphosate resistant line. As expected, the EPSPS CNV was obvious and genes within that region were overexpressed; however, thousands of other regions across the genome varied between the assembled glyphosate susceptible line and the re-sequenced glyphosate resistant line. Several of these regions strongly correlated with changes in gene expression and may have consequences for the plant’s physiology. Most importantly, the Fhy/FAR1 mutator-like transposases have increases in nrd and therefore may be selectively duplicated in the glyphosate resistant line, and it may be that they are still highly active Mule transposons. Future work, including an improved kochia genome with higher coverage PacBio and Hi-C guided scaffolding as well as expanding this work into new, locally adaptive populations, may reveal CNVs of great import, especially in local adaptation to abiotic stresses.

Tables

Table 2.1: A statistical summary of the kochia genome assembly.

Metric	Count	Percentage
Number of scaffolds	19,671	
Total size of scaffolds (bp)	711,356,803	
Longest scaffold (bp)	770,912	
Shortest scaffold (bp)	897	
Scaffold length/genome size		83.70%
Number of scaffolds > 1K nt	19,594	99.6%
Number of scaffolds > 10K nt	14,701	74.7%
Number of scaffolds > 100K nt	1,286	6.5%
Mean scaffold size (bp)	36,163	
N50 scaffold length (bp)	61,675	
%A		28.8%
%C		16.4%
%G		16.4%
%T		28.5%
%N		9.5%
Num. of contigs	61,353	
Num. of contigs in scaffolds	54,776	
Total size of contigs	643,547,114	

Table 2.2: A statistical summary of predicted genes in the kochia genome.

Metric	Count	Percentage
Proteome		
Total Length of Proteome aa	14,859,659	
Longest Protein	5,817	
Number of Transcripts > 500 aa	8,158	
Number of Transcripts > 1,000 aa	1204	
Mean Protein Size	313	
Median Protein Size	234	
Transcriptome		
Number of Coding Gene Models (Maker)	47,414	
Total Length of Transcripts	44,695,962	
Longest Transcript	17,454	
Number of Transcripts > 500 nt	30,953	65.3%
Number of Transcripts > 1K nt	16,209	34.2%
Number of Transcripts > 10K nt	12	0.0%
Mean Transcript size	943	
Median Transcript size	702	
N50 transcript length	1,311	
L50 transcript count	10,590	
scaffold %A		27.9%
scaffold %C		22.1%
scaffold %G		22.1%
scaffold %T		27.8%
scaffold %N		0.1%
Annotation		
Number of Proteins with Blast Hit (DataBase)	29,730	62.70%
Number of Proteins with InterPro Domain	38,779	81.79%

Table 2.3: Assessing the kochia genome assembly and annotation completeness with BUSCO

Metric	Count	Percentage
# of Ultra-conserved Genes Searched For	1440	
# Ultra-conserved Single Genes Found	987	68.5%
# Ultra-conserved Duplicated Genes Found	33	2.3%
# Ultra-conserved Partial Genes Found	83	5.7%
Total Ultra-conserved Genes Found	1103	76.6%
# Ultra-conserved Genes Missing	337	23.4%

Table 2.4: Analyses of repetitive content in the kochia genome using RepeatMasker

Interspersed repeat elements	Number	Length (BP)	% of Assembly
Retroelements	66,766	38,463,923	5.41%
SINEs:	178	26,154	0%
Penelope	8	787	0%
LINEs:	12,579	4,566,194	0.64%
CRE/SLACS	199	137,148	0.02%
L2/CR1/Rex	0	0	0%
R1/LOA/Jockey	0	0	0%
R2/R4/NeSL	0	0	0%
RTE/Bov-B	3,377	1,311,037	0.18%
L1/CIN4	9,011	3,123,321	0.44%
LTR elements:	54,009	33,871,575	4.76%
BEL/Pao	0	0	0%
Ty1/Copia	22,611	15,381,646	2.16%
Gypsy/DIRS1	30,306	18,264,655	2.57%
Retroviral	0	0	0%
DNA transposons	27,584	5,607,206	0.79%
hobo-Activator	10,360	1,763,567	0.25%
Tc1-IS630-Pogo	3,368	819,160	0.12%
En-Spm	0	0	0%
MuDR-IS905	0	0	0%
PiggyBac	0	0	0%
Tourist/Harbinger	1,508	538,707	0.08%
Other (Mirage, P-element, Transib)	2	74	0%
Rolling-circles	0	0	0%
Unclassified:	1,535	392,232	0.06%
Total interspersed repeats:		44,463,361	6.25%

Other Repeats	Number	Length (BP)	% of Assembly
Small RNA:	948	223,307	0.03%
Satellites:	256	22,750	0%
Simple repeats:	261,069	14,664,544	2.06%
Low complexity:	58,540	3,248,675	0.46%
Total Other repeats:		18,159,276	2.55%

Table 2.5: Summary comparing the resequencing data from the glyphosate resistant kochia line when it is aligned to the susceptible genome assembly.

Events with >2× nrd	Number
Number of CNVs	2,802
Number of Genes	3,918
Average number of genes per CNV	1.40
Average Length (bp)	13,987
Average Read Depth	0.253

Events with <0.5× nrd	Number
Number of CNVs	7,147
Number of Genes	9,235
Average number of genes per CNV	1.29
Average Length (bp)	13,504
Average Read Depth	7.359

Table 2.6: The most abundant InterPro IDs in the genome annotation and in lists of events with >2x nrd and with <0.5 x nrd. The proportion of those events within each list are provided. Terms with higher then expected abundance in either events with >2x nrd or with <0.5 x nrd are highlighted in grey.

	Total Genome Annotation (55,615 total)		Events with >2x nrd (5,659 total)		Events with <0.5x nrd (16,550 total)	
Top 15 InterPro Ids for Genome Annotation	Number of Genes	Proportion	Number of Genes	Proportion	Number of Genes	Proportion
IPR005162 - Retrotransposon gag domain	1712	3.1%	433	7.7%	460	2.8%
IPR012337 - Ribonuclease H-like domain	1691	3.0%	194	3.4%	535	3.2%
IPR021109 - Aspartic peptidase domain	1507	2.7%	293	5.2%	479	2.9%
IPR026960 - Reverse transcriptase zinc-binding domain	1279	2.3%	174	3.1%	408	2.5%
IPR005135 - Endonuclease/exonuclease/phosphatase	1215	2.2%	174	3.1%	438	2.6%
IPR027417 - P-loop nucleoside triphosphate hydrolase	898	1.6%	69	1.2%	298	1.8%
IPR025558 - Domain of unknown function DUF4283	894	1.6%	101	1.8%	294	1.8%
IPR011009 - Protein kinase-like domain	893	1.6%	59	1.0%	316	1.9%
IPR000719 - Protein kinase domain	772	1.4%	45	0.8%	250	1.5%
IPR001878 - Zinc finger, CCHC-type	767	1.4%	191	3.4%	212	1.3%
IPR032675 - Leucine-rich repeat domain, L domain-like	718	1.3%	57	1.0%	238	1.4%
IPR031052 - FHY3/FAR1 family	578	1.0%	128	2.3%	198	1.2%
IPR011990 - Tetratricopeptide-like helical domain	524	0.9%	38	0.7%	120	0.7%
IPR008271 - Serine/threonine-protein kinase, active site	491	0.9%	33	0.6%	152	0.9%
IPR007527 - Zinc finger, SWIM-type	446	0.8%	119	2.1%	145	0.9%

Figures

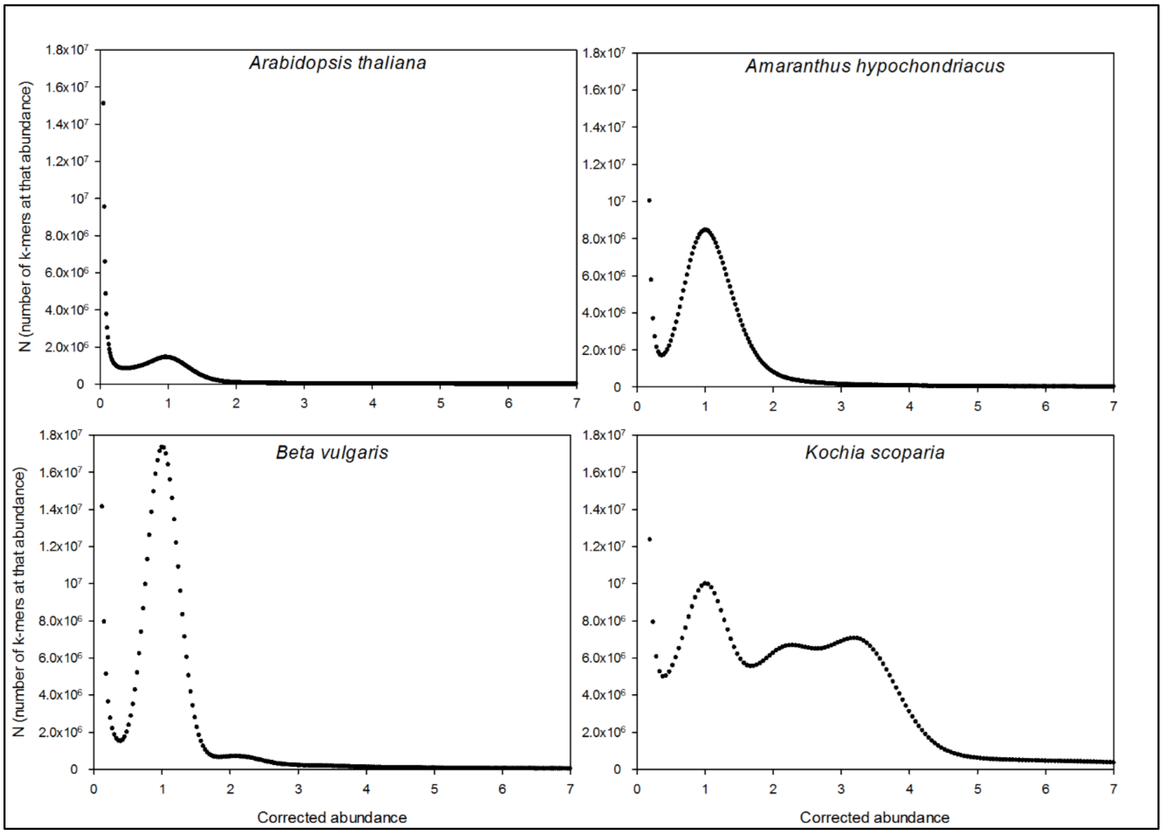

Figure 2.1: A K-mer (24-mer) distribution graph for unassembled Illumina data from four species: *Arabidopsis thaliana*, *Amaranthus hypochondriacus*, *Beta vulgaris*, and *Kochia scoparia*. Axes have been adjusted so that the first mode of each distribution is 1.

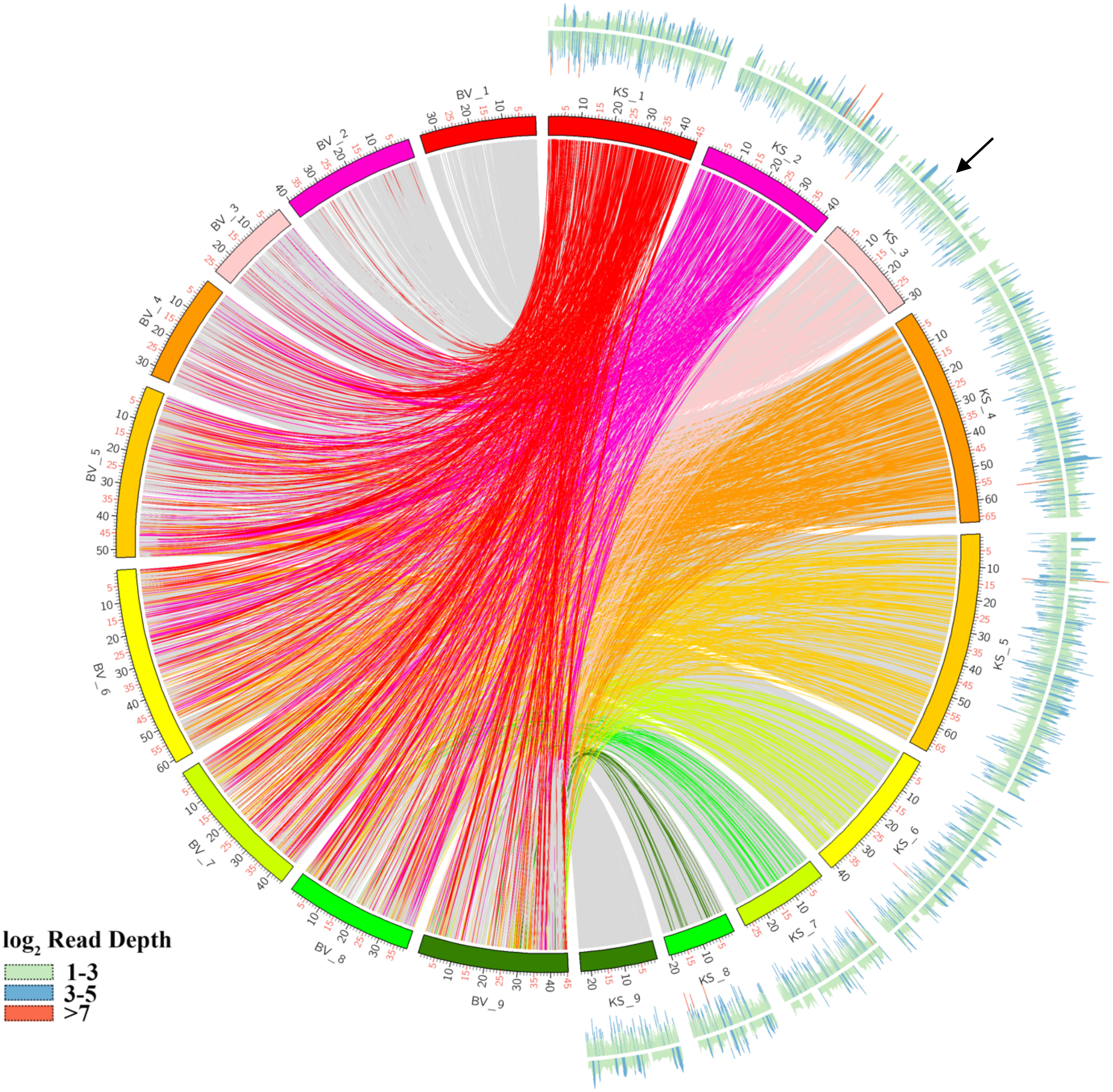

Figure 2.2: Kochia assembly contigs arranged into pseudomolecules based on synteny with *Beta vulgaris*. Grey/faded lines represent matches used to order the scaffolds while colored lines represent 500 bp alignments that are non-syntenic (align to other chromosomes) based on this arrangement of the contigs in these pseudomolecules. On the outer most track, peaks pointing inward represent dips in coverage (cutoff of 0.5× coverage) while peaks pointing outward represent increases in coverage (cutoff of 2× coverage) in the glyphosate resistant line. The black arrow shows the location of the *EPSPS* CNV event.

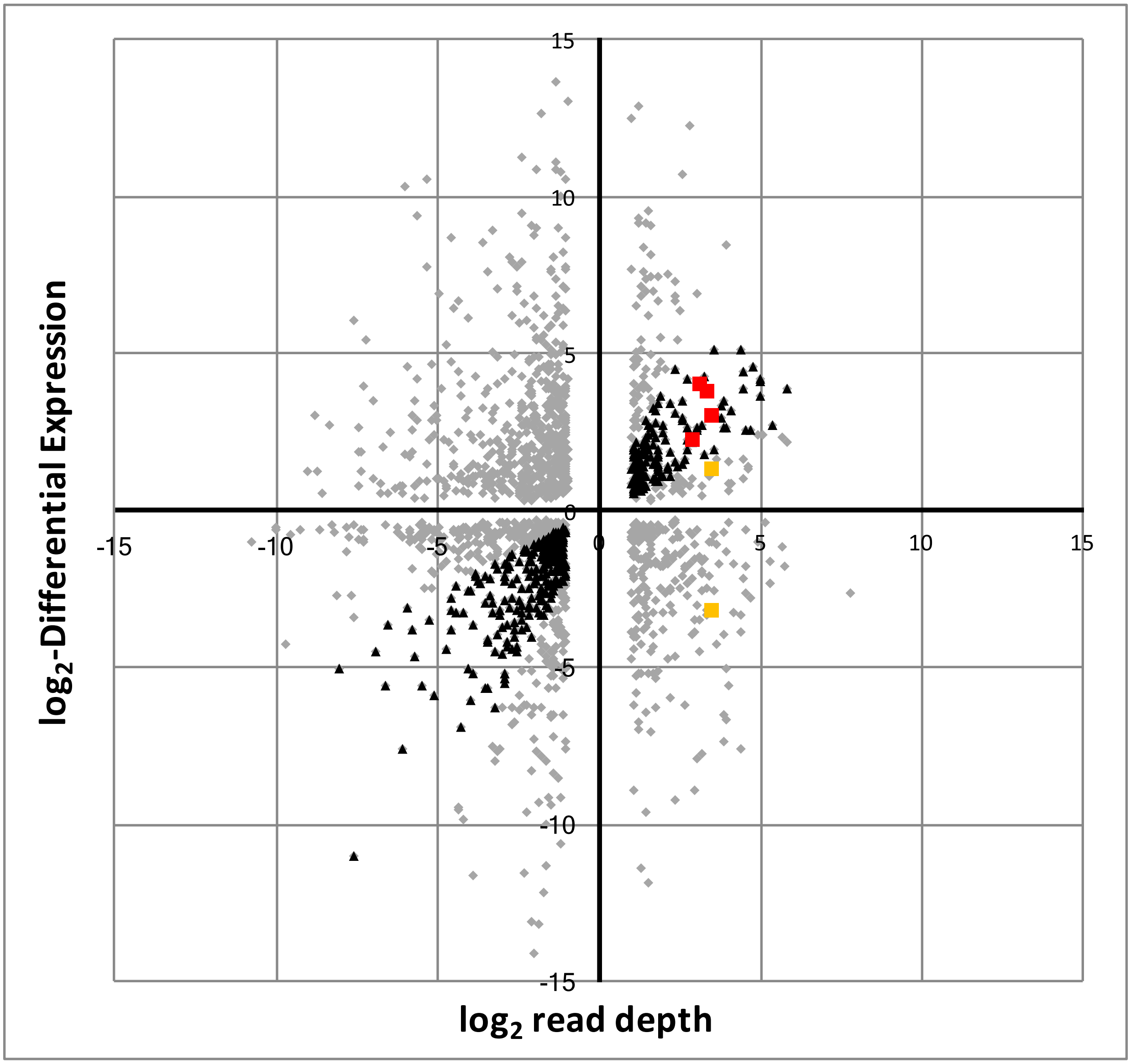

Figure 2.3: A plot of the $\log_2$ read depth for all genes with a p-value <0.01 from CNVnator versus the $\log_2$-fold change in expression for each gene with significance P-value <0.01 from EdgeR. Grey diamonds are all genes for which read depth and copy number are not correlated. Black triangles are genes for which read depth and copy number are correlated. Red squares are the four genes within the EPSPS CNV event that have significant differential expression and for which expression is correlated to read depth. Orange squares are genes within the EPSPS CNV event that have significant differential expression but whose expression is not correlated to read depth.

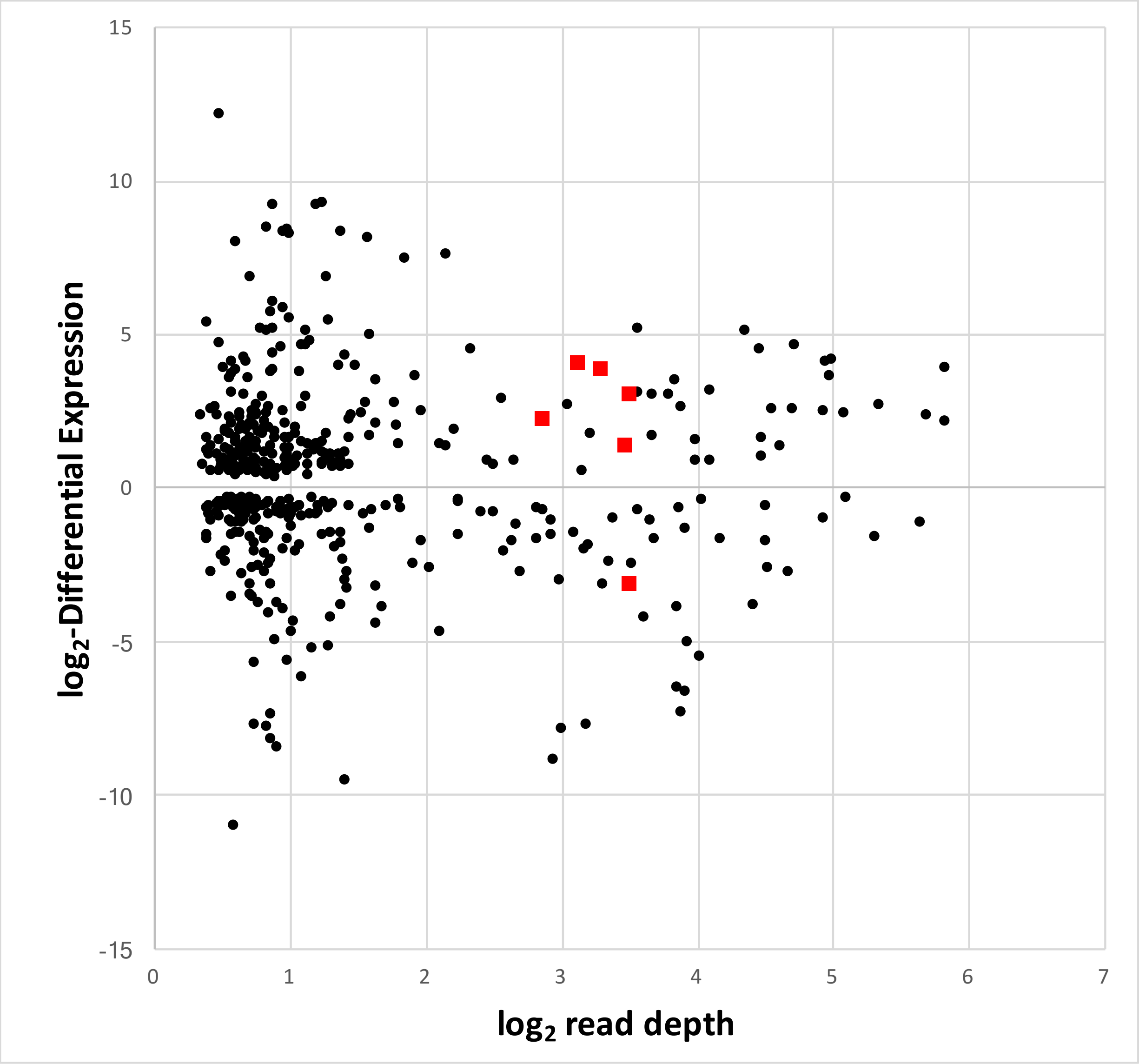

Figure 2.4: A plot of the $\log_2$ read depth for all genes with a p-value less than that of the genes in the EPSPS CNV from CNVnator versus the $\log_2$-fold change in expression for each gene with significance P-value <0.01 from EdgeR. Red squares are the genes within the EPSPS CNV event.

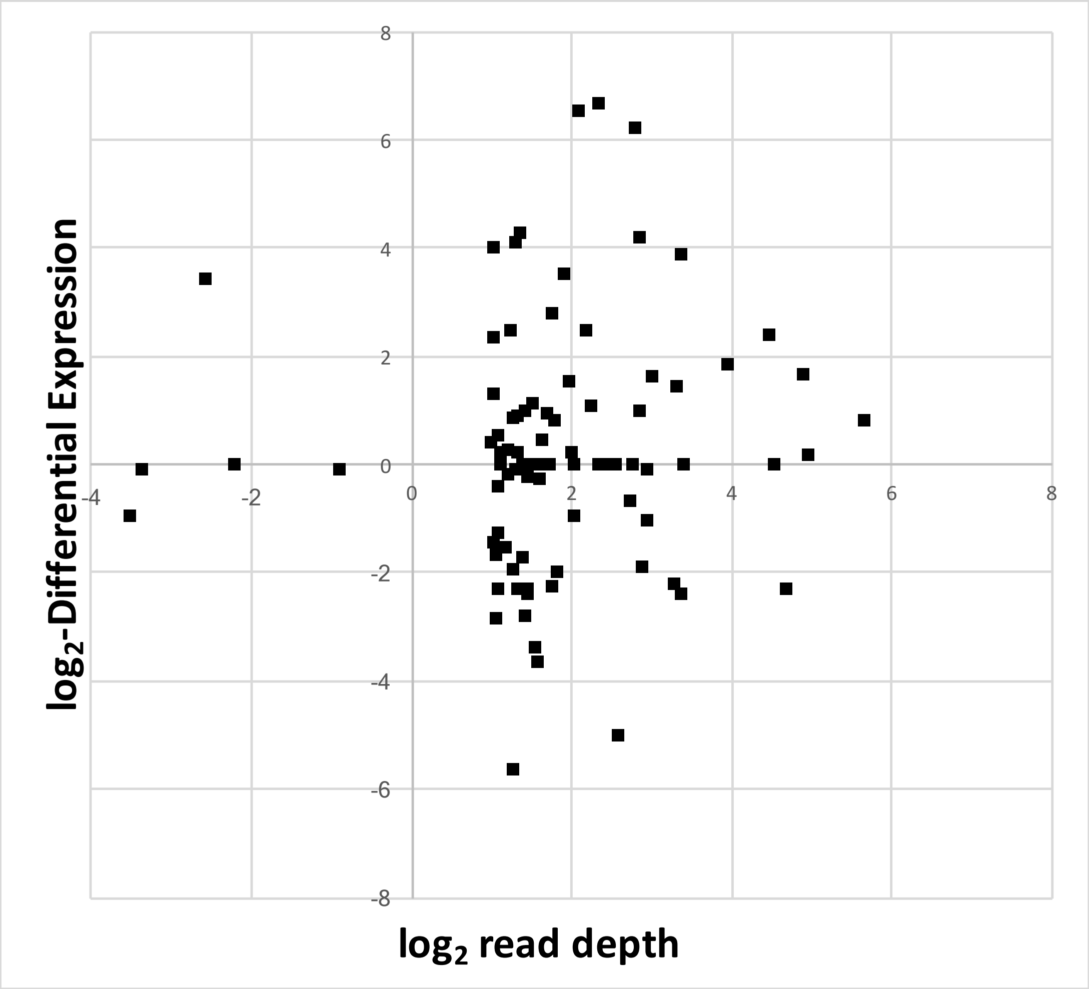

Figure 2.5: A plot of the $\log_2$ read depth for all genes annotated as Fhy/FAR related with a p-value <0.01 from CNVnator versus the $\log_2$-fold change in expression, regardless of expression.

REFERENCES

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein. 2011. "CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing." *Genome Research* 21 (6): 974–84. doi:10.1101/gr.114876.110.

Allen, T., A. Koustenis, G. Theodorou, D. E. Somers, S. A. Kay, G. C. Whitelam, and P. F. Devlin. 2006. "Arabidopsis FHY3 Specifically Gates Phytochrome Signaling to the Circadian Clock." *The Plant Cell Online* 18 (10): 2506–16. doi:10.1105/tpc.105.037358.

Alzohairy, Ahmed M., Gábor Gyulai, Robert K. Jansen, and Ahmed Bahieldin. 2013. "Transposable Elements Domesticated and Neofunctionalized by Eukaryotic Genomes." *Plasmid* 69 (1): 1–15. doi:10.1016/j.plasmid.2012.08.001.

Bass, C, and L M Field. 2011. "Gene Amplification and Insecticide Resistance." *Pest Management Science* 67 (8): 886–90. doi:10.1002/ps.2189.

Beckie, Hugh J., Ross M. Weiss, Julia Y. Leeson, and Owen O. Olfert. 2012. "Range Expansion of kochia (*Kochia scoparia*) in North America under a Changing Climate." Edited by Jerry A. Ivany and Robert E. Blackshaw. *Topics in Canadian Weed Science* 8 (February). Canadian Weed Science Society: 31–45.

Butler, Jonathan, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte, Eric S. Lander, Chad Nusbaum, and David B. Jaffe. 2008. "ALLPATHS: De Novo Assembly of Whole-Genome Shotgun Microreads." *Genome Research* 18 (5): 810–20. doi:10.1101/gr.7337908.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10. doi:10.1186/1471-2105-10-421.

Cantarel, Brandi L., Ian Korf, Sofia M.C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. “MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes.” *Genome Research* 18 (1): 188–96. doi:10.1101/gr.6743907.

Casey, C.A. 2009. “Plant Fact Sheet for Kochia (*Kochia scoparia*).” *USDA-Natural Resources Conservation Service, Kansas Plant Materials Center*. http://greatbasinseeds.com/wordpress/wp-content/uploads/2013/09/hycrest.crested.pdf.

Cook, David E, Tong Geon Lee, Xiaoli Guo, Sara Melito, Kai Wang, Adam M Bayless, Jianping Wang, et al. 2012. “Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean.” *Science (New York, N.Y.)* 338 (6111): 1206–9. doi:10.1126/science.1228746.

Cranston, Harwood J, Anthony J Kern, Josette L Hackett, Erica K Miller, Bruce D Maxwell, and William E Dyer. 2001. “Dicamba Resistance in Kochia.” *Weed Research* 49 (2): 164–70.

Crusoe, Michael R., Hussien F. Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, et al. 2015. “The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis.” *F1000Research*. doi:10.12688/f1000research.6924.1.

Debolt, Seth. 2010. “Copy Number Variation Shapes Genome Diversity in Arabidopsis over Immediate Family Generational Scales.” *Genome Biology and Evolution* 2 (1): 441–53. doi:10.1093/gbe/evq033.

Dohm, Juliane C., André E. Minoche, Daniela Holtgräwe, Salvador Capella-Gutiérrez, Falk

Zakrzewski, Hakim Tafer, Oliver Rupp, et al. 2014. "The Genome of the Recently Domesticated Crop Plant Sugar Beet (Beta Vulgaris)." *Nature* 505 (7484): 546–49. doi:10.1038/nature12817.

Doyle, Jeffrey. 1991. "DNA Protocols for Plants." In *Molecular Techniques in Taxonomy*, 283–93. doi:10.1007/978-3-642-83962-7_18.

English, Adam C., Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, et al. 2012. "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology." *PLoS ONE* 7 (11). doi:10.1371/journal.pone.0047768.

Finnegan, Jean, and David McElroy. 1994. "Transgene Inactivation: Plants Fight Back!" *Bio/Technology* 12 (9): 883–88. doi:10.1038/nbt0994-883.

Flagel, Lex E., and Jonathan F. Wendel. 2009. "Gene Duplication and Evolutionary Novelty in Plants." *New Phytologist* 183 (3): 557–64. doi:10.1111/j.1469-8137.2009.02923.x.

Foes, Matthew J, Lixin Liu, Gerald Vigue, Edward W Stoller, Loyd M Wax, and Patrick J Tranel. 1999. "A Kochia (*Kochia scoparia*) Biotype Resistant to Triazine and ALS-Inhibiting Herbicides." *Weed Science* 47 (1): 20–27.

Gaines, Todd A., Abigail L. Barker, Eric L. Patterson, Eric P. Westra, and Andrew R. Kniss. 2016. "EPSPS Gene Copy Number and Whole-Plant Glyphosate Resistance Level in *Kochia Scoparia*." *PLoS ONE* 11 (12). doi:10.1371/journal.pone.0168295.

Hackl, T., R. Hedrich, J. Schultz, and F. Forster. 2014. "Proovread: Large-Scale High-Accuracy PacBio Correction through Iterative Short Read Consensus." *Bioinformatics* 30 (21): 3004–11. doi:10.1093/bioinformatics/btu392.

Hardigan, Michael A., Emily Crisovan, John P. Hamilton, Jeongwoon Kim, Parker Laimbeer, Courtney P. Leisner, Norma C. Manrique-Carpintero, et al. 2016. "Genome Reduction

Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated *Solanum Tuberosum*." *The Plant Cell* 28 (2): 388–405. doi:10.1105/tpc.15.00538.

Hudson, Matthew E., Damon R. Lisch, and Peter H. Quail. 2003. "The *FHY3* and *FAR1* Genes Encode Transposase-Related Proteins Involved in Regulation of Gene Expression by the Phytochrome A-Signaling Pathway." *Plant Journal* 34 (4): 453–71. doi:10.1046/j.1365-313X.2003.01741.x.

Hull, Ryan M., Cristina Cruz, Carmen V. Jack, and Jonathan Houseley. 2017. "Environmental Change Drives Accelerated Adaptation through Stimulated Copy Number Variation." *PLoS Biology* 15 (6). doi:10.1371/journal.pbio.2001333.

Jones, Philip, David Binns, Hsin Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–40. doi:10.1093/bioinformatics/btu031.

Jugulam, Mithila, Kindsey Niehues, Amar S Godar, D.-H. Koo, Tatiana Danilova, Bernd Friebe, Sunish Sehgal, et al. 2014. "Tandem Amplification of a Chromosomal Segment Harboring 5-Enolpyruvylshikimate-3-Phosphate Synthase Locus Confers Glyphosate Resistance in *Kochia scoparia*." *Plant Physiology* 166 (3): 1200–1207. doi:10.1104/pp.114.242826.

Koo, D. H., Jugulam, M., Putta, K., Cuvaca, I. B., Peterson, D. E., Currie, R. S., M., Friebe, and Gill, B. S. (2018). Gene duplication and aneuploidy trigger rapid evolution of herbicide resistance in common waterhemp. Plant physiology, 176(3), 1932-1938.

Koo, D. H., Molin, W. T., Saski, C. A., Jiang, J., Putta, K., Jugulam, M., Friebe, B., and Gill, B. S. (2018). Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri. Proceedings of the National

Academy of Sciences, 201719354.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive κ-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36. doi:10.1101/gr.215087.116.

Kurtz, Stefan, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. 2004. “Versatile and Open Software for Comparing Large Genomes.” *Genome Biology* 5 (2): R12. doi:10.1186/gb-2004-5-2-r12.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. doi:10.1093/bioinformatics/btt656.

Lin, R., Y. Teng, H.-J. Park, L. Ding, C. Black, P. Fang, and H. Wang. 2008. “Discrete and Essential Roles of the Multiple Domains of Arabidopsis FHY3 in Mediating Phytochrome A Signal Transduction.” *Plant Physiology* 148 (2): 981–92. doi:10.1104/pp.108.120436.

Lin, Rongcheng, Lei Ding, Claudio Casola, Daniel R Ripoll, Cédric Feschotte, and Haiyang Wang. 2007. “Transposase-Derived Transcription Factors Regulate Light Signaling in Arabidopsis.” *Science* 318 (5854): 1302–5. doi:10.1126/science.1146281.

Lynch, M., and J. S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” *Science* 290 (5494): 1151–55. doi:10.1126/science.290.5494.1151.

Müller, K. and Borsch, T., 2005. Phylogenetics of Amaranthaceae based on matK/trnK sequence data: evidence from parsimony, likelihood, and Bayesian analyses. *Annals of the Missouri*

Botanical Garden, 66-102.

Mi, Huaiyu, Betty Lazareva-Ulitsky, Rozina Loo, Anish Kejariwal, Jody Vandergriff, Steven Rabkin, Nan Guo, et al. 2005. “The PANTHER Database of Protein Families, Subfamilies, Functions and Pathways.” *Nucleic Acids Research* 33 (DATABASE ISS.). doi:10.1093/nar/gki078.

Patterson, Eric L, Dean J Pettinga, Karl Ravet, Paul Neve, and Todd A Gaines. 2018. “Glyphosate Resistance and EPSPS Gene Duplication: Convergent Evolution in Multiple Plant Species.” *Journal of Heredity* 109 (2): 117–25. doi:10.1093/jhered/esx087.

Pettinga, Dean J, Junjun Ou, Eric L Patterson, Mithila Jugulam, Philip Westra, and Todd A Gaines. 2017. “Increased Chalcone Synthase (CHS) Expression Is Associated with Dicamba Resistance in *Kochia scoparia*.” *Pest Management Science*, December. doi:10.1002/ps.4778.

Pianka, Eric R. 1970. “On R- and K-Selection.” *The American Naturalist* 104 (940). University of Chicago Press : 592–97. doi:10.1086/282697.

Preston, Christopher, David S. Belles, Philip H. Westra, Scott J. Nissen, and Sarah M. Ward. 2009. “Inheritance of Resistance to The Auxinic Herbicide Dicamba in Kochia (*Kochia scoparia*).” *Weed Science* 57 (1). Weed Science Society of America: 43–47. doi:10.1614/WS-08-098.1.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. doi:10.1093/bioinformatics/btp616.

Sakai, Ann K., Fred W. Allendorf, Jodie S. Holt, David M. Lodge, Jane Molofsky, Kimberly A. With, Syndallas Baughman, et al. 2001. “The Population Biology of Invasive Species.”

Annual Review of Ecology and Systematics 32 (1). Annual Reviews: 305–32. doi:10.1146/annurev.ecolsys.32.081501.114037.

Schimke, R T, A Hill, and R N Johnston. 1985. "Methotrexate Resistance and Gene Amplification: An Experimental Model for the Generation of Cellular Heterogeneity." *British Journal of Cancer* 51 (4): 459–65.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12. doi:10.1093/bioinformatics/btv351.

Slack, Andrew, P. C. Thornton, Daniel B. Magner, Susan M. Rosenberg, and P. J. Hastings. 2006. "On the Mechanism of Gene Amplification Induced under Stress in Escherichia Coli." *PLoS Genetics* 2 (4): 385–98. doi:10.1371/journal.pgen.0020048.

Soto-Jimenez, Luz, Karel Estrada, and Alejandro Sanchez-Flores. 2014. "GARM: Genome Assembly, Reconciliation and Merging Pipeline." *Current Topics in Medicinal Chemistry* 14 (3): 418–24. doi:10.2174/1568026613666131204110628.

Tang, W, W Wang, D Chen, Q Ji, Y Jing, H. Wang, and R. Lin. 2012. "Transposase-Derived Proteins FHY3/FAR1 Interact with PHYTOCHROME-INTERACTING FACTOR1 to Regulate Chlorophyll Biosynthesis by Modulating HEMB1 during Deetiolation in Arabidopsis." *The Plant Cell* 24 (5): 1984–2000. doi:10.1105/tpc.112.097022.

Tang, Wei, Ronald J. Newton, and Douglas A. Weidner. 2007. "Genetic Transformation and Gene Silencing Mediated by Multiple Copies of a Transgene in Eastern White Pine." *Journal of Experimental Botany* 58 (3): 545–54. doi:10.1093/jxb/erl228.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. "Using RepeatMasker to Identify Repetitive

Elements in Genomic Sequences." *Current Protocols in Bioinformatics*. doi:10.1002/0471250953.bi0410s25.

Thrasher, Andrew, Zachary Musgrave, Brian Kachmarck, Douglas Thain, and Scott Emrich. 2014. "Scaling Up Genome Annotation Using MAKER and Work Queue." *Int. J. Bioinformatics Res. Appl.* 10 (4/5): 447–60. doi:10.1504/IJBRA.2014.062994.

Vaucheret, Hervé, Christophe Béclin, Taline Elmayan, Frank Feuerbach, Christian Godon, Jean Benoit Morel, Philippe Mourrain, Jean Christophe Palauqui, and Samantha Vernhettes. 1998. "Transgene-Induced Gene Silencing in Plants." *Plant Journal*. doi:10.1046/j.1365-313X.1998.00337.x.

Waite, Jason, Curtis R Thompson, Dallas E Peterson, Randall S Currie, Brian L S Olson, Phillip W Stahlman, and Kassim Al-Khatib. 2013. "Differential Kochia (*Kochia Scoparia*) Populations Response to Glyphosate." *Weed Science* 61 (2): 193–200. doi:10.1614/WS-D-12-00101.1.

Wang, Haiyang, and Wang Deng Xing. 2002. "Arabidopsis FHY3 Defines a Key Phytochrome A Signaling Component Directly Interacting with Its Homologous Partner FAR1." *EMBO Journal* 21 (6): 1339–49. doi:10.1093/emboj/21.6.1339.

Wiersma, Andrew T, Todd A Gaines, Christopher Preston, John P. Hamilton, Darci Giacomini, C. Robin Buell, Jan E Leach, and Philip Westra. 2014. "Gene Amplification of 5-Enol-Pyruvylshikimate-3-Phosphate Synthase in Glyphosate-Resistant *Kochia scoparia*." *Planta* 241: 463–74. doi:10.1007/s00425-014-2197-9.

Xi, Ruibin, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua Zhang, Mark D Johnson, et al. 2011. "Copy Number Variation Detection in Whole-Genome Sequencing Data Using the Bayesian Information Criterion." *PNAS* 108 (46): E1128-36.

CHAPTER 3: THE EPSPS LOCUS IN KOCHIA SCOPARIA

The structure of the *EPSPS* locus in glyphosate resistant and susceptible *Kochia scoparia*[3]

Summary

In the weedy plant species, *Kochia scoparia*, 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) copy number variation (CNV) confers glyphosate resistance. Kochia is not the only species to undergo EPSPS CNV; however, unlike the other well studied species, *Amaranthus palmeri*, kochia's copies of *EPSPS* are arranged in tandem and copy numbers have not been reported above 11 copies. In this study, we use a combination of genomics techniques to assess the size of the duplicated locus, discover the genes surrounding *EPSPS* that are co-duplicated, and identify a possible cause for the initial duplication event. First, we use information from the genome assembly and resequencing data of a glyphosate resistant kochia line to predict the size of the amplified region. From this we develop a bacterial artificial chromosome (BAC) genomic library for kochia and a set of three probes that allow us to isolate BACs upstream, downstream, and in the middle of the duplicated region. These BACs, when sequenced and assembled indicate that the *EPSPS* duplication appears in two forms, a larger 72kb repeat and a smaller 48.5kb repeat. Both contain the *EPSPS* gene, but different numbers of co-duplicated genes. Additionally, a large transposable element known as a Fhy/FAR1 mutator-like transposase has inserted both downstream and upstream of the *EPSPS* gene, but only in the glyphosate resistant line. We developed a series of qPCR markers for copy number assays that

[3] Eric L. Patterson, Christopher A. Saski, Daniel B. Sloan, Karl Ravet, Phil Westra, Todd A. Gaines

validate our BAC assemblies and the presence of the Fhy/FAR1 transposase insertion. Understanding the genomic differences between the resistant and susceptible *EPSPS* loci is the first step in understanding the origin of *EPSPS* gene duplication, and possibly other CNVs in *Kochia scoparia*.

Introduction

Gene copy number variation can be a double-edged sword when it comes to evolution and adaptation. While it can have serious consequences in some systems, i.e. causing cancer in humans, it can also increase genetic variation and provide an evolutionary advantage, especially in the more plastic genomes of plants (Schimke, Hill, and Johnston 1985; Xi et al. 2011; Debolt 2010; Lynch and Conery 2000; Hull et al. 2017).

Copy number variation of 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) is known to confer resistance to glyphosate, the world's most-used herbicide (Duke and Powles 2008; Sammons and Gaines 2014). Increased gene copy number of *EPSPS* causes the over-production of the EPSPS protein, glyphosate's target (Gaines et al. 2010; Wiersma et al. 2014). This overproduction of target protein makes it necessary for the application of more glyphosate to have the same lethal effect (Gaines et al. 2016). This phenomenon has been observed in eight weed species to date; however, the molecular and genomic mechanisms underlying *EPSPS* gene duplication are only known in one species, *Amaranthus palmeri* (Patterson et al. 2018; Molin et al. 2017). In the case of *A. palmeri*, *EPSPS* gene duplication is caused by a large, circular, extra-chromosomal DNA element that disperses copies across the genome (Koo et al. 2018, Molin et al. 2017). This mechanism sometimes results in *A. palmer* plants containing *EPSPS* copies estimated in the hundreds (Gaines et al. 2010).

Recently *EPSPS* gene duplication has been described in the weed species *Kochia scoparia* (kochia), one of the most important weeds in the Central Great Plains of the United States and Canada (Jugulam et al. 2014; Wiersma et al. 2014; Gaines et al. 2016). In kochia, *EPSPS* copy numbers typically range from 3 to 8 with the highest reports at 11 copies (Gaines et al. 2016). Additionally, fluorescence *in situ* hybridization (FISH) shows that the *EPSPS* copies in kochia are arranged in tandem and are most likely caused by unequal crossing over (Jugulam et al. 2014). More detailed cytogenetics studies using fiber-FISH show that the majority of repeats of the *EPSPS* loci are either 45 kb or 66 kb in length. Occasionally, inverted repeats or repeats of 70 kb in length have been observed (Jugulam et al. 2014). The initial causes of the *EPSPS* gene duplication event remain unresolved. One possibility is that low-level *EPSPS* gene amplification exists within natural standing variation or genomic rearrangements are being generated each generation at a low frequency, then these rearrangements are being selected by glyphosate because they confer survival in the glyphosate-treated environment.

In this paper, we explore the *EPSPS* locus from the recently assembled genome sequence of kochia and uncover the genes that are co-duplicated with *EPSPS*. Additionally, we sequenced and assembled the entire *EPSPS* locus by sequencing bacterial artificial chromosomes (BACs) generated from a glyphosate resistant kochia plant to look at differences between the structures of the *EPSPS* locus in resistant and susceptible individuals. The comparison between the susceptible and resistant assemblies allows us to define the various repeat types and the genetic content therein. Most importantly, we discovered a mobile element that is associated with the gene duplication event and that we hypothesize may be responsible for the origin of the *EPSPS* gene duplication event.

Methods

Analyzing the EPSPS Contig from the Glyphosate Susceptible Genome

The contig containing the *EPSPS* locus (Contig_00009) was found in the first draft of the kochia genome assembly from a glyphosate susceptible line; it happened to be the 10th largest in the assembly. This line was called "7710" and its origins and breeding are described in Preston et al. 2009 and Pettinga et al. 2017. Contig_00009 was aligned and compared to the scaffold containing *EPSPS* from *Beta vulgaris* and *Amaranthus palmeri* using the alignment software Mummer (Kurtz et al. 2004).

A glyphosate resistant line, so-called "M32", was developed from a field population in eastern Colorado. This population was initially identified after glyphosate application failed to control the plants in a wheat fallow system in 2012. Seed was collected in the Fall after the plants had fully matured and brought back to the greenhouse for screening, verification and purification. Seeds were grown in 10x10cm pots with one plant per pot. Once plants reached a height of 8-10 cm 870 g ae ha^{-1} of glyphosate mixed with ammonium sulfate (2% w/v) was applied. After three weeks, dead and highly injured plants were removed, and the remaining plants were allowed to grow, cross pollinate with other survivors, and set seed. This process was repeated for three generations. At this point, there were no more susceptible individuals in the offspring.

High quality DNA was extracted from a single glyphosate resistant individual using a modified CTAB DNA extraction protocol (Doyle 1991) (See Chapter 2). This DNA was then used to generate, whole genome, 100bp paired reads generated on an Illumina 2500 HiSeq sequencer. In total 142,961,780 read pairs were generated for a total of ~285Gb of sequence data. These reads were aligned to Contig_00009 using BWA v0.7.15 backtrack alignment program using standard parameters (See Chapter 2) (Li and Durbin 2009). Next, RNA was extracted from young leaf tissue of four, 10 cm tall, glyphosate-resistant and susceptible plants using a Qiagen RNeasy kit. Two hundred nanograms of this RNA was used to generate cDNA utilizing the TruSeq Stranded mRNAseq preparation guidelines. This cDNA was multiplexed and a single lane from an Illumina 2500 HiSeq was used to generate RNA-Seq data for all eight individuals (4 susceptible and 4 resistant) (See Chapter 2). For each sample, between 15,000,000 and 21,000,000, 150bp paired Illumina reads were obtained after the reads were trimmed and analyzed using FastQC v0.10.1. Reads were then aligned to the gene models in contig_00009 using Bowtie 2 and the differential expression of each gene within the contig was analyzed using the quasi-likelihood approach in the generalized linear model (glm) framework as described in the user manual of EdgeR (Robinson, McCarthy, and Smyth 2010; Langmead and Salzberg 2012). Bowtie was run using standard parameters; therefore, for reads that mapped to multiple locations, only the highest scoring match was reported. Contig_00009 was aligned to itself and a dotplot was generated using YASS (Noé and Kucherov 2005).

Sequencing BACs from a glyphosate resistant plant

A library of bacteria artificial chromosomes (BACs) was generated from a single glyphosate resistant kochia plant selected from the glyphosate resistant population following the protocol described in Luo and Wing 2003 with modifications as described in Molin et al. 2017. High molecular weight (HMW) DNA was extracted from young leaf tissue from a single glyphosate resistant plant using a modified CTAB DNA extraction protocol (See Chapter 2). This HMW DNA was ligated to a linearized vector and transformed into *E. coli* using electroporation. Recombinant colonies were then grown on LB plates. Radiolabeled probes were designed for the *EPSPS* gene itself, a sequence upstream, and a sequence downstream of the

EPSPS CNV. Predicted locations for the probes were determined by looking at the alignment of shot gun Illumina data from the glyphosate resistant line against the contig_00009. Several colonies containing the appropriate sequences were identified for each probe. These identified BACs were end sequenced to determine their approximate location and run on pulse-field gel electrophoresis to determine their approximate size. Colonies containing positive BACs of the correct position and size were isolated and cultured. HMW DNA was extracted from these colonies and prepared using a SMRTbell Template Prep Kit, 1.0 using the manufacture-recommended protocols. Finally, the HMW DNA was sent for RSII PacBio sequencing on two SMRT cells.

PacBio reads were assembled using the software Canu (Koren et al. 2017). The BAC vector sequence was then removed from the assembled contigs. These resistant contigs were then self-aligned and aligned to the susceptible contig using YASS. Additionally, the BACs insert sequences were run through the MAKER pipeline, informed with cDNA and protein annotations from the Chenopodiaceae and the gene models from the kochia genome (Cantarel et al. 2008) for gene annotation.

Markers for Confirming the Structure of the EPSPS CNV

Primers were designed that were spaced at regular intervals (~5kb-15kb) along this contig that spanned the putative CNV area for genomic qPCR analysis (Table 3.1). Additionally, qPCR primers were designed that spanned the junctions of the two dominant repeat types as well as for the large insert (Table 3.2). Primers were designed to closely mimic the primers already published for the *EPSPS* gene (Wiersma et al, 2016), including a melting temperature between 51 and 56 ℃, a GC content between 40 and 50%, and a length of between 20 and 24 base pairs. Furthermore, the resulting amplicon had to be between 100 and 200 base pairs long. All genomic

PCR was performed using the same protocol established for *EPSPS* copy number assay (Gaines et al, 2016).

Susceptible and resistant plants were grown in the greenhouse until they were ~10 cm tall and 100 mg of young expanding leaf tissue was taken from each plant. DNA was extracted from this tissue using the recommend protocol from the DNeasy Plant Mini Kit. The DNA quality and abundance was checked using a NanoDrop 1000 and diluted to 5 ng/ µl. For qPCR two genes were used as single-copy controls: acetolactate synthase (*ALS*) and copalyl di-phosphate synthetase 1 (*CPS*). Each qPCR reaction consisted of 12.5 µL PerfeCTa SYBR® green Super Mix (Quanta Biosciences), 1 µL of the forward and reverse primers at 10 µM final concentration, 10 ng gDNA (2 µL), and 9.5 µL of sterile water for a total volume of 25 µL.

A BioRad CFX Connect Real-Time System was used for qPCR. The temperature cycle for all reactions was as follows: an initial 3 min at 95 ℃ followed by 35 rounds of 95 ℃ for 30 sec and 53 ℃ for 30 secs with a fluorescence reading at 497 nm after each round. A melt curve was performed from 65–95 ℃ in 0.5 ℃ increments for each reaction to verify the production of a single PCR product. Additionally, all products from a susceptible line were run on a 1.5% agarose gel to verify a single product with low primer dimerization. Relative quantification was calculated using the comparative C_t method: $2^{\Delta C}$ ($\Delta C_t = (C_t^{(ALS)} + C_t^{(CPS)})/2 - C_t^{EPSPS}$) (Schmittgen and Livak 2008).

Results

Analyzing the EPSPS Contig from the Glyphosate Susceptible Genome

The susceptible contig containing the *EPSPS* locus from the genome assembly was 399,779 bp long. The *EPSPS* gene model was 5,551 bp long (UTR, Exons and Introns included) and located between base pairs 91,663-97,214 of the contig. When this contig was aligned to

Beta vulgaris near perfect synteny was observed; however, when compared to the sequence responsible for duplicating *EPSPS* from *Amaranthus palmeri*, little similarity existed outside of the *EPSPS* gene itself (Figure 3.1).

When shotgun Illumina data from a glyphosate resistant line was aligned to the contig, the read depth of *EPSPS* and its surrounding area was much greater (> 7.26 times) than the background read depth. Using this alignment, it was possible to predict the exact boundaries of the *EPSPS* CNV starting at base pair 41,684 and continuing to base pair 101,128 with the total length of the CNV being 59,444 bp (a "Type I" repeat). This region contains seven coding genes of various functions including *EPSPS* itself (Figure 3.2, Table 3.3). When differential expression of these genes was calculated using RNA-Seq data, five of the genes showed over expression in the glyphosate resistant line, one gene showed under-expression in the glyphosate resistant line and one showed no significant difference (FDR adjusted p-value < 0.05) (Figure 3.2, Table 3.3). Since gene expression is dynamic, depending on both environmental conditions and developmental stage, the genes not showing DE may be overexpressed in glyphosate resistant plants under different experimental conditions. When the *EPSPS* contig was aligned to itself, there was no evidence for sequence complexity (simple sequence repeats, inverted repeats, self-homology, etc.) at the predicted boundaries of the CNV (Figure 3.3).

Sequencing BACs from a glyphosate resistant plant

A BAC library was generated and probed using the *EPSPS* gene sequence and sequence upstream and downstream of the predicted CNV boundaries. These BACs were sent for PacBio sequencing and assembled. From this PacBio data we assembled three contigs that were 139,476 bp, 110,757 bp, and 43,607 bp long for the upstream, *EPSPS*, and downstream regions, respectively. These assemblies encompassed at least two repeats of the CNV and a significant portion of the surrounding sequence. The first repeat was a Type I repeat as defined above and contained the entire predicted duplicated region; however, the second repeat was smaller and contained only four of the seven co-duplicated genes (a Type II repeat). Both repeats end at the same base pair, directly after *EPSPS*; however, the beginning of the Type II repeat is 23,390 bp further downstream (Figure 3.5, 3.7). When all 3 BAC contigs were self-aligned, a large repeat structure appeared just downstream of every assembled *EPSPS* gene and at the upstream boundary of the *EPSPS* CNV (Figures 3.4, 3.5, 3.6). This repeat structure consisted of twelve, 135 bp sequences that were identical.

Enough overlap existed among the three BAC contigs to composite all three of our BAC assemblies together to make a representative sequence (a meta-assembly) that contained one type I repeat and one type II repeat as well as the flanking upstream and downstream sequence. When this BAC meta-assembly from glyphosate resistant kochia was aligned to the susceptible contig from the genome assembly, we observed near perfect agreement for the repeats; however, a large disparity was evident at the end of each copy and at the beginning of the *EPSPS* CNV event (Figure 3.7). A 16,037 bp sequence was inserted just downstream and upstream of both copies of *EPSPS* in the glyphosate resistant BAC assemblies. This insert shows no homology with any part of the susceptible contig; furthermore, when this insertion was aligned against the entire susceptible genome, this region was not found in its entirety.

We ran annotation using Maker on this insertion to predict gene models and identified four regions with putative coding genes. The first predicted gene belonged to the family of genes known as FHY3/FAR1 (IPR031052) and contained the domains: “AR1 DNA binding” and “zinc finger, SWIM-type” (IPR004330F, IPR007527 respectively). The second gene’s function was less clear but was identified to be part of the Ubiquitin-like domain superfamily (IPR029071).

The third gene's function was also unclear and was generally identified as belonging to the Endonuclease/exonuclease/phosphatase superfamily (IPR036691). The fourth and final gene had no identifiable InterPro domains, and BLASTed to uncharacterized proteins in NCBI. Additionally, this insertion was responsible for the large repetitive domain observed in the self-alignment. We refer to this insertion as the Fhy/FAR1-like insertion due to the annotation of one of the genes predicted in its borders.

When the full type I repeat from the glyphosate resistant BAC was aligned to the contig from the susceptible genome, two deletions >1,000bp were detectable in the resistant BAC. These could be real disparities between the lines or an error in the assembly of the susceptible contig. In total, these deletions account for 3,450 bp. If the Fhy/FAR1-like insertion and these deletions are accounted for, and assuming they are the same in every copy, then type I repeats are 72,022 bp long and type 2 repeats are 48,641 bp long.

Markers for Confirming the Structure of the EPSPS CNV

Quantitative PCR markers were developed dispersed across the entire CNV, including markers on both sides in regions that show no evidence of CNV (Table 3.1). Markers 1 and 2 showed low copy number (near 1) as they both sit upstream of the beginning of the CNV start site. Marker 3 only amplified in the resistant line and showed increased copy number ranging between 8 and 14 copies (depending on the individual). Marker 4 had fewer copies, between 3 and 10. Markers 5, 6, 7, and 8 were very tightly associated and co-varied for each individual ranging from 10-20 copies. Markers 9, 10, and 11 had one copy as they lie downstream of *EPSPS* and outside the borders of the CNV (Table 3.4). Additional qPCR markers were developed that only amplified when the Fhy/FAR1-like insertion was flanked by either the type I or type II repeat. Using these markers, we quantified the number of type I and type II repeats in several individuals. In our line, type II repeats were less frequent then type I repeats. The tested individuals each had approximately 2 type II repeats and between 5-7 type I repeats (Table 3.5). These markers did not amplify in any susceptible plants, indicating the Fhy/FAR1-like insertion is not present at the beginning of the susceptible *EPSPS* locus.

Additionally, we developed a marker internal to the Fhy/FAR1-like insertion. All susceptible individuals had approximately 4-5 copies of this marker; however, none of these regions were assembled in the kochia genome assembly. In resistant individuals, we detected 14-18 copies of the Fhy/FAR1-like insertion. If we account for the 4-5 copies that are in the susceptible individuals and if we consider that a Fhy/FAR1-like insertion exists at both the upstream and downstream boundary then we would predict 9-13 copies, which almost perfectly correlates with the copy number observed for qPCR markers 5, 6, 7, and 8. This would indicate that one copy of the Fhy/FAR1-like insertion is associated with each repeat, regardless of whether it is type I or type II (Table 3.5). With this information in conjunction with previously published cytogenetic work from Jugulam et al. 2014, we propose a model for the structure of the *EPSPS* CNV (Figure 3.8).

Discussion

Analyzing the EPSPS Contig from the Glyphosate Susceptible Genome

The *EPSPS* contig from kochia has near perfect synteny with *Beta vulgaris* along its entire length but little homology with a similar region from *Amaranthus palmeri*, another plant that undergoes *EPSPS* gene duplication but through a seemingly different mechanism (Figure 3.1) (Patterson et al. 2018; Molin et al. 2017; Jugulam et al. 2014). The length of the *EPSPS* contig and the location of *EPSPS* within that contig means that the boundaries of the CNV event were within the assembled contig. When whole genome resequencing of a glyphosate resistant line was performed, increased read depth was observed for an ~60 kb region (Figure 3.2).

RNA-Seq expression data shows that four of the six genes within the conserved region of the repeat are over-expressed at a rate commensurate with genomic resequencing read depth: RAD51, transketolase, tRNA N6-adenosine threonylcarbamoyltransferase, and *EPSPS* (FDR adjusted p-value <0.05). Interestingly, one of the genes within this region, golgin subfamily A member 6-like protein 6, shows decreased expression in the high duplication plant. The gene RAD51 is significantly overexpressed; however, it is not commensurate with its read depth; read depth is greater than the corresponding over expression. The gene NRT1/ PTR FAMILY 7.2-like gene had no difference in expression. We believe that this reduction and maintenance of expression may be due to gene silencing, similar to what happens when multiple copies of transgenes are inserted in the same plant (Finnegan and McElroy 1994; Wei Tang, Newton, and Weidner 2007). The obvious benefit of *EPSPS* over-expression is glyphosate resistance but the effects of these other genes remain unclear. We hypothesize that the co-amplification of these other genes is not adaptive but is being co-selected with *EPSPS* and repeated glyphosate application. Most interesting of these genes is the RAD51 homolog. Mis-expression or knockouts of RAD51 have been shown to cause cancer in animal tissues as RAD51 regulates crossing-over events during meiosis (Maacke et al. 2000) (Figure 3.2). In the future, it would be interesting to work in a model system to overexpress these other genes and observe the impacts they have on plant physiology and fitness.

When contig_00009 is aligned to itself, no complexities, such as SSRs or large homodimers of nucleotides, exist at the beginnings of either type I or type II repeats (Figure 3.3). This would indicate that the sequence in the susceptible locus alone is insufficient for explaining why this region has become a site copy number variation. Most likely homology exists at the upstream and downstream boundaries where an initial misalignment followed by crossing over occurred (Graur and Li 2000; Russell 2002).

Sequencing BACs from a glyphosate resistant plant

BACs generated from a glyphosate resistant line were sequenced using Pac-Bio to elucidate any differences between individuals with and without *EPSPS* duplicated. We assembled 3 contigs of 139,476 bp, 110,757 bp, and 43,607 bp that, when meta-assembled, encompassed one whole type 1 repeat, one whole type II repeat, and flanking sequence on either side of the *EPSPS* CNV event. When the meta-assembled BAC contig was aligned to contig_00009 and a large insertion was observed that contains several putative genes including a Fhy/FAR-1 transposon-like gene. Additionally, every instance of this insertion had a large repeat structure consisting of twelve 135bp repeats (Figure 3.4, 3.5, 3.6) that were not present in the susceptible contig. This insertion could not be found in the kochia genome assembly. Evolutionarily speaking, members of the Fhy/FAR gene family are derived from MULE transposons and have been "domesticated" to have a role in the regulation of genes involved in circadian rhythm and light sensing (Hudson, Lisch, and Quail 2003; W Tang et al. 2012; Wang and Xing 2002). We believe this is evidence that these elements may still be mobile and that they are not fully "domesticated." Because the insert appears to be both at the upstream and downstream borders of the CNV we hypothesize that the insertion of this Fhy/FAR-1 transposon-like element happened in two locations, flanking the *EPSPS* gene. These two insertions then led to misalignment as both sequences were identical and a crossing-over event happened somewhere along the length of the misaligned region generating two alleles – one with two, Type I repeats and the other with no EPSPS locus, the latter of which would presumably be lethal in the homozygous state. Interestingly, the insertion of the upstream Fhy/FAR element shares microhomology with the beginning of the Type II repeat. We propose that a subsequent double stranded break at the Fhy/FAR-1 downstream boundary incorrect implementation of microhomology mediated double-stranded break repair could have caused the formation of Type II repeats (Figure 3.9) (Ottaviani et al. 2014, Sfeir and Symington 2015).

In total, the presence of the Fhy/FAR1 insertion in conjunction with each *EPSPS* copy and a few minor differences between the susceptible and resistant contigs brings the size of type I and type II repeats to 72,022 bp and 48,641 bp long, respectively. These sizes are larger than the previously fiber-FISH predicted sizes of 66kb and 45kb respectively (Jugulam et al. 2014). What accounts for the differences between our assemblies and the previously reported fiber-FISH studies remains unclear, as Fiber-FISH generally has a resolution of ~1kb (Ersfeld 1994). It may be that different populations of kochia have different repeat sizes. Further testing and validation on the type and size of the *EPSPS* duplications in various, divergent populations is needed to confirm this. Additionally, a 38 kb inversion of the *EPSPS* CNV has been previously reported (Jugulam et al. 2014); however, in our work we did not detect any BACs with the inverted regions. The inversions may be absent from the glyphosate-resistant line we used, we may have been unable to computationally resolve an inverted copy, or we failed to select a colony that contained a BAC with an inversion.

Markers for Confirming the Structure of the EPSPS CNV

Quantitative PCR markers designed along the length of the CNV confirmed that there were two types of repeats, the longer type I repeat and shorter type II repeat. Four markers were highly duplicated and therefore present in both type I and type II repeats and two markers were duplicated to a lesser extent indicating they were only in the Type I repeats (Table 3.4). The results from the pair of primers that detected the presence and number of the Fhy/FAR transposable element was surprising. In the susceptible plant, approximately 4-6 copies were observed despite not appearing in the susceptible genome assembly; therefore, this specific Fhy/FAR transposable element was not assembled. It may be that these background copies lie in repetitive or difficult to assemble regions. In the resistant plants, the number of Fhy/FAR insert copies was always approximately equal to the *EPSPS* copy number plus 4-6 copies, indicating that the original copies found elsewhere in the genome are still present and the insert is being co-duplicated with every repeat of the *EPSPS* CNV. In *Amaranthus palmeri,* it has been shown that miniature inverted-repeat transposable elements (MITEs) as well as putative helitrons are closely associated with *EPSPS* gene duplication in resistant individuals (Gaines et al. 2013; Molin et al. 2017). It seems that mobile elements are a key factor in determining when and how the *EPSPS* locus becomes duplicated.

The development of our evolutionary history model allows us to test whether EPSPS duplication in this species happened once or multiple times. If all glyphosate resistant populations have the same genomic elements (Far-1 insertions, Type I and Type II repeats, upstream and downstream boundaries, etc) it would imply that duplication occurred once and is spreading via pollen or seed mediated gene flow. If; however, there are other types of rearrangements or mobile elements in divergent populations, it implies multiple evolutionary events of EPSPS gene duplication. Additionally, the insertion of two Far-1 elements near each other resulting in unequal crossing over may be testable in a model system. If we are able to transform a model plant so that two identical elements were near each other, we could try and induce unequal crossing over and CNVs.

Conclusion

By understanding the sequence and structure of the *EPSPS* locus in both resistant and susceptible kochia individuals it is possible to construct a testable hypothesis as to the history of molecular and genomic events that gave rise to glyphosate resistance in this species. We hypothesize that the insertion of two Fhy/FAR like transposons near the *EPSPS* gene has caused a genomic disruption that has led to subsequent unequal crossing-over and copy number variation of the *EPSPS* gene and the surrounding region. Several genes in this region surrounding *EPSPS* are co-duplicated and the duplication has impacts on their expression; however, the fitness penalties, if any exist, for the over-expression of these other genes is not yet investigated and therefore the full impact of gene duplication is still unknown. *EPSPS* gene duplication in kochia is an amazing case of genome plasticity and the adaptive potential of copy number variation. This study highlights the importance of the interactions between transposable elements, copy number variation, and adaptive evolution.

Tables

Table 3.1: Primers for qPCR markers for determining copy number at multiple locations near the *EPSPS* gene.

Primer name	Primer sequence	Melting Temp (℃)	GC Content (%)
1	5’-CATAGGTTGAGGGTGGACTTTC-3’	55.2	50
1	5’-GGTGTTTGTTTGACCACCTTTC-3’	54.8	45.5
2	5’-TTCTGCCTCAGCAAACATACT-3’	54.3	42.9
2	5’-CATGGTCACTTTGTGTGTCATTAG-3’	54.2	41.7
3	5’-CTCGGAAAGGATGGAAGAATG-3’	53.2	47.6
3	5’-GTTATGTCCTGTCTTCTGTGTG-3’	53.2	45.5
4	5’-TTTCGCTTTCCGAGGTAATAG-3’	52.4	42.9
4	5’-CAACTAACACGAACATTGTGTC-3’	52.2	40.9
5	5’-TCGAAGCCTGACATTAGATTAG-3’	51.9	40.9
5	5’-CTCTTTGTACCTGATCCCATC-3’	52.5	47.6
6	5’-CTCCTCCTCCCTCCTAATATC-3’	53	52.4
6	5’-CTTGTTTCCTCCTCTCGTTC-3’	52.9	50
7	5’-TCATCCCTTTCTCTCTCCTC-3’	52.9	50
7	5’-GATAAGTCCGTCAACACGATC-3’	53.1	47.6
8	5’-GACATCCTGTCATGGAGTAAG-3’	52.4	47.6
8	5’-CCTAAATAAACCGGAAGCAATC-3’	51.8	40.9
9	5’-TCAACACCCAACTCACATCTC-3’	54.7	47.6
9	5’-TAGAAGCACAGGAGAGAGAGAA-3’	54.5	45.5
10	5’-GGCATGTGGAGAAGATGTATAG-3’	52.7	45.5
10	5’-CTTTGTTGGTTCAATTGGAGG-3’	52.2	42.9
11	5’-TCGGATCCCTTAGATACACTAC-3’	52.8	45.5
11	5’-GTTACCTGTCTTGAGCAGTG-3’	53.1	50

Table 3.2: Primers for qPCR markers for determining copy number of Type I repeats, Type II repeats, and the Fhy/FAR Insertion.

Primer name	Primer sequence	Melting Temp (℃)	GC Content (%)	Length (bp)
Type I/II FP	5’-GACGGAAATACCCTCAATATAGACA-3’	54.0	40.0%	25
Type I RP	5’-ACGCCCAAGATGTACATTGATA-3’	54.0	40.9%	22
Type II RP	5’-CATGCCTTTGATGTCCAAGTTT-3’	54.1	40.9%	22
Fhy/FAR FP	5’-GAAGATAGCGAGACGTTTGAG-3’	53.0	47.6%	21
Fhy/FAR RP	5’-CGGCTTGATCGGTTAAGATAC-3’	53.2	47.6%	21

Table 3.3: List of genes near *EPSPS* that are in or flanking the *EPSPS* CNV event. Read depth is the $\log_2$ of the difference between the background read depth and the read depth of each gene. DE is the differential expression between four resistant and four susceptible individuals from RNA-Seq. P-value is the significance of DE and is adjusted for false discovery rate.

Gene	Beginning	Ending	Length	Orientation	Description	Part of the CNV?	Read Depth	DE	P-value
KS_00451	27,406	28,674	1,268	Reverse	GRAVITROPIC IN THE LIGHT 1-like	No	0	-0.43	0.00
KS_00452	35,728	36,696	968	Reverse	IRK-Interacting Protein	No	0	-2.62	0.05
KS_00453	37,839	41,640	3,801	Reverse	Nitroreductase family	No	0	0.74	0.00
KS_00454	43,124	47,121	3,997	Forward	arginase 1, mitochondrial	Only Type 1	2.86	2.23	0.00
KS_00455	47,240	52,651	5,411	Reverse	protein NRT1/ PTR FAMILY 7.2-like	Only Type 1	2.86	0.72	0.58
KS_00456	63,014	72,467	9,453	Forward	tRNA N6-adenosine threonylcarbamoyltransferase	Type 1 & 2	3.49	3.03	0.00
KS_00457	72,617	73,531	914	Reverse	golgin subfamily A member 6-like protein 6	Type 1 & 2	3.49	-3.18	0.00
KS_00458	76,342	81,181	4,839	Forward	DNA repair protein RAD51	Type 1 & 2	3.46	1.33	0.00
KS_00459	82,421	84,836	2,415	Forward	transketolase, chloroplastic-like	Type 1 & 2	3.29	3.83	0.00
KS_00460	91,663	97,214	5,551	Forward	3-phosphoshikimate 1-carboxyvinyltransferase 2 (EPSPS)	Type 1 & 2	3.12	4.01	0.00
KS_00461	106,901	109,241	2,340	Forward	NAD dependent epimerase	No	0	2.52	0.00
KS_00462	106,975	110,332	3,357	Reverse	uncharacterized protein	No	0	2.54	0.06
KS_00463	113,504	114,006	502	Reverse	DUF861	No	0	0.05	0.85

Table 3.4: Copy number data from all qPCR markers on three susceptible and five resistant individuals. Copy number is calculated as $\Delta C_t = (C_t^{(ALS)} + C_t^{(CPS)})/2 - C_t^{Marker}$. “N/A” stands for “No Amplification”.

Line	Biological Replicate	1	2	3	4	5	6	7	8	9	10	11
7710	1	0.9	0.7	N/A	1.1	1.6	1.1	1.3	1.2	0.7	1.9	0.8
	2	0.7	0.7	N/A	1.0	1.5	1.2	1.4	1.4	0.9	1.7	1.2
	3	0.7	0.6	N/A	0.9	1.0	1.2	0.7	1.3	1.0	1.6	1.1
M32	1	0.9	0.7	9.5	6.1	11.3	11.2	11.3	11.5	1.0	N/A	1.0
	2	0.8	0.7	9.5	6.0	12.6	12.1	12.4	13.3	1.0	N/A	1.1
	3	0.7	0.6	7.6	3.2	10.9	11.1	11.0	11.7	1.0	N/A	1.0
	4	0.7	0.7	8.1	5.1	10.8	9.9	10.4	9.9	0.9	N/A	0.9
	5	1.2	1.0	14.2	10.0	20.3	19.0	19.6	20.0	1.3	N/A	1.4

Table 3.5: Copy number data from Type I repeats, Type II repeats, and the Fhy/FAR Insertion on three susceptible and five resistant individuals. Copy number is calculated as $\Delta C_t = (C_t^{(ALS)} + C_t^{(CPS)})/2 - C_t^{Marker}$. “N/A” stands for “No Amplification”

Line	Replicate	Type 1	Type 2	FAR-1 TE
7710	1	N/A	N/A	3.9
	2	N/A	N/A	5.5
	3	N/A	N/A	4.7
M32	1	5.4	1.8	16.2
	2	5.1	1.9	17.4
	3	5.1	1.7	18.2
	4	5.3	1.7	14.1
	5	6.9	2.1	17.7

Figures

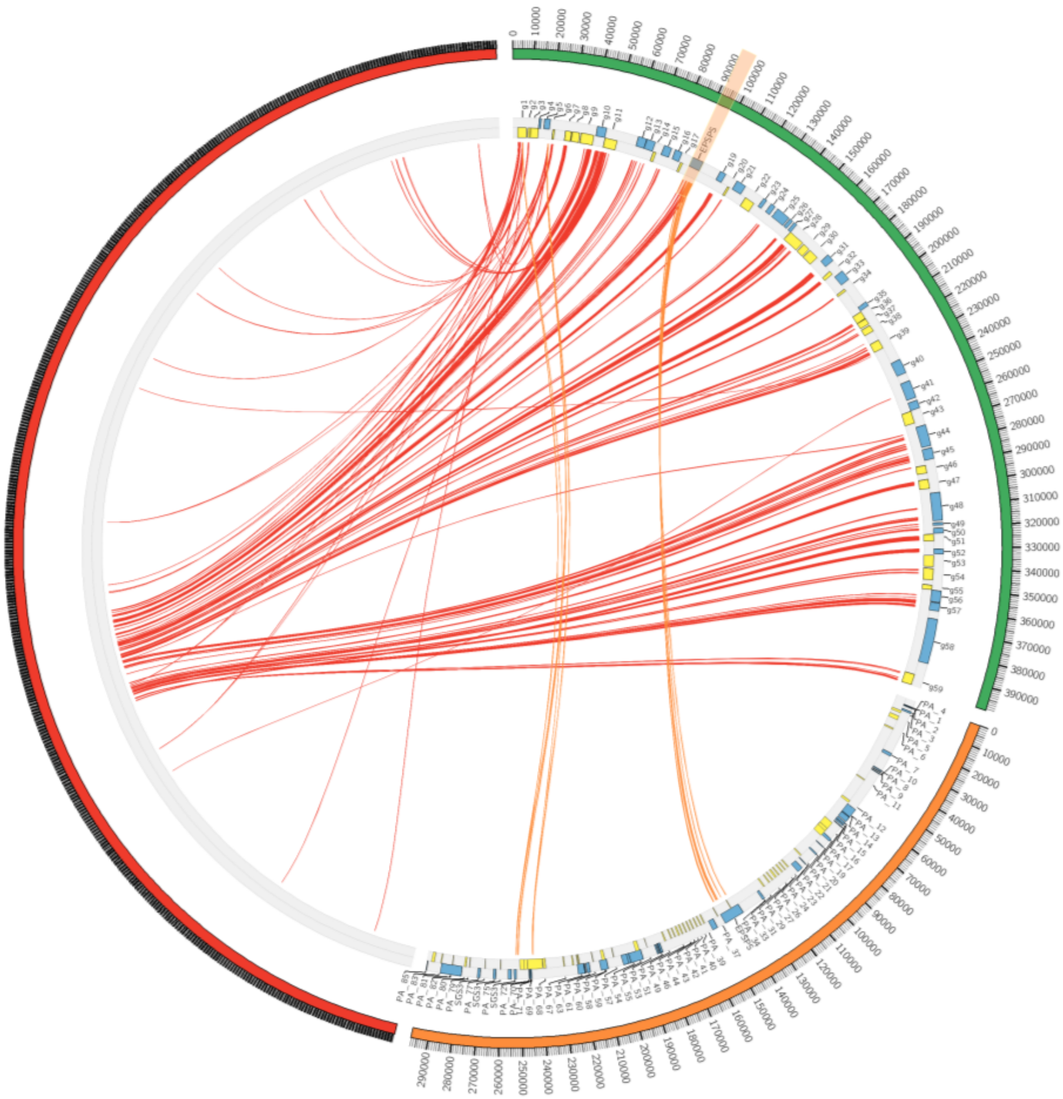

Figure 3.1: A comparison of the *EPSPS* contig from kochia (Green), a large segment from the *Beta vulgaris* genome (Red), and the *EPSPS* replicon from *Amaranthus palmeri* (Orange). Blue and yellow blocks indicate genes in the forward and reverse orientation, respectively. The *EPSPS* gene is highlighted in orange. Red, connecting lines indicate areas of high similarity between *Beta vulgaris* and kochia. Orange, connecting lines indicate areas of high similarity between *Amaranthus palmeri* and kochia. Number of base pairs in the alignment are listed on the outside track.

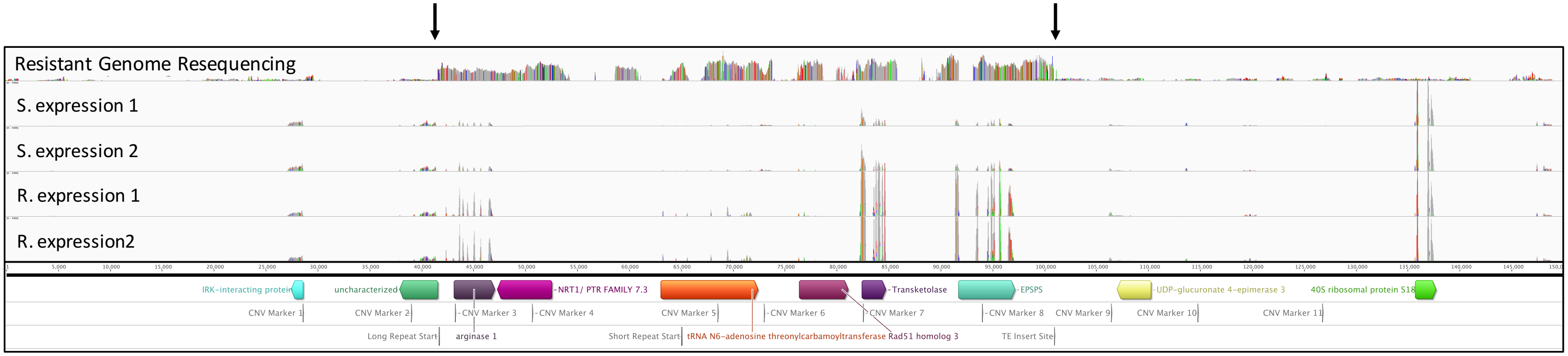

Figure 3.2: The first 150,000 bp from the *EPSPS* contig from the kochia genome assembly. Predicted genes are represented by the multicolored blocks and labeled with text of the corresponding color. The locations of copy number qPCR markers are indicated, as well as the beginning of the Type I, Type II, and Fhy/FAR insert site. The beginning and end of the duplication are indicated with black arrows. Alignments of RNA-Seq Illumina data from two resistant and two susceptible individuals are indicated as well as whole genome resequencing data from the resistant line.

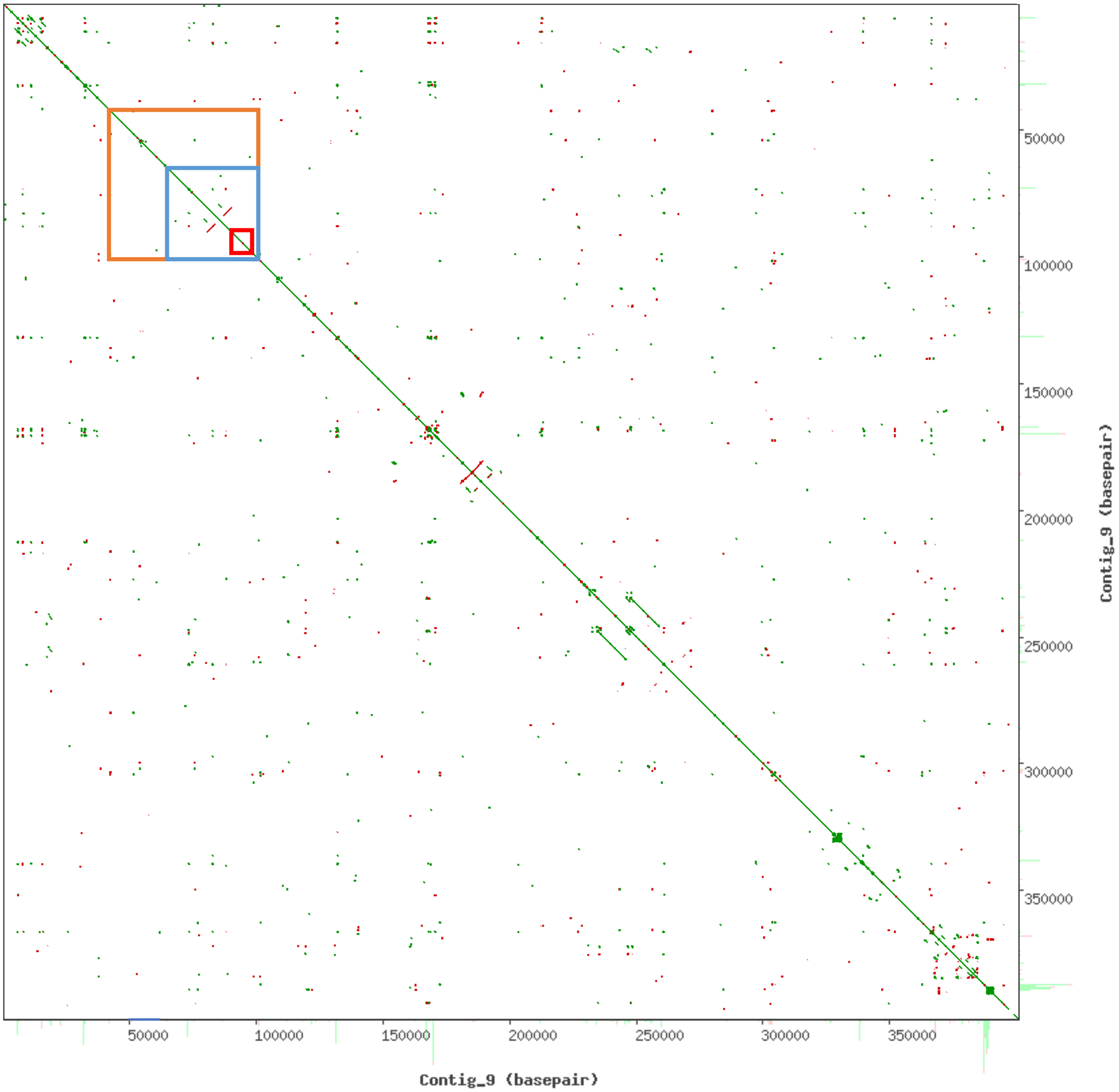

Figure 3: A self-alignment of the *EPSPS* contig from the kochia genome assembly. The location of the *EPSPS* gene is indicated with a red box. Type I repeats are indicated with an orange box, Type II repeats are indicated using a blue box.

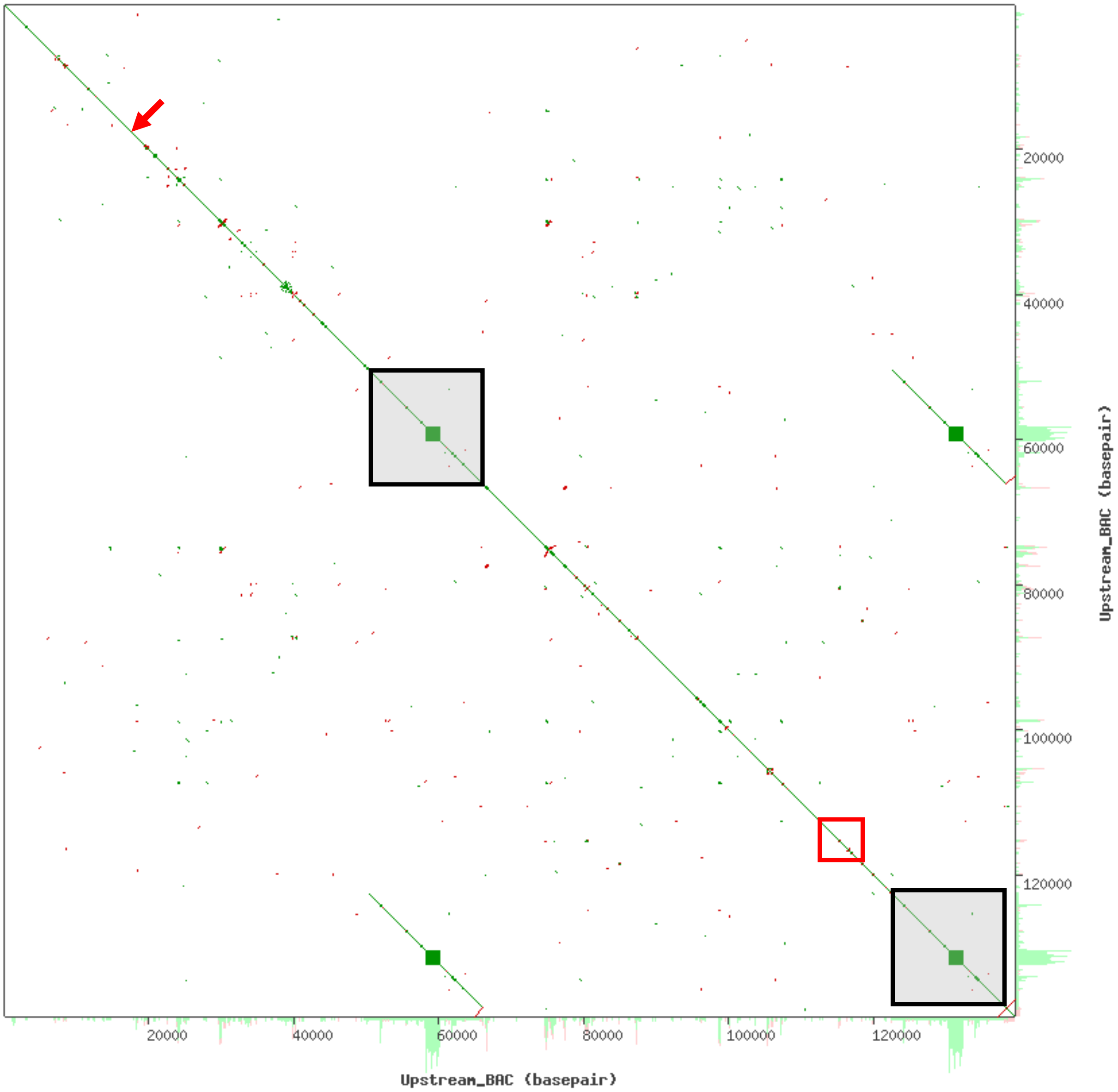

Figure 4: A self-alignment of the upstream contig from the resistant BAC assembly. The location of the *EPSPS* gene is indicated with a red box. The Fhy/FAR insertion is denoted with black boxes. The red arrow indicates the beginning of the *EPSPS* contig from the susceptible genome assembly.

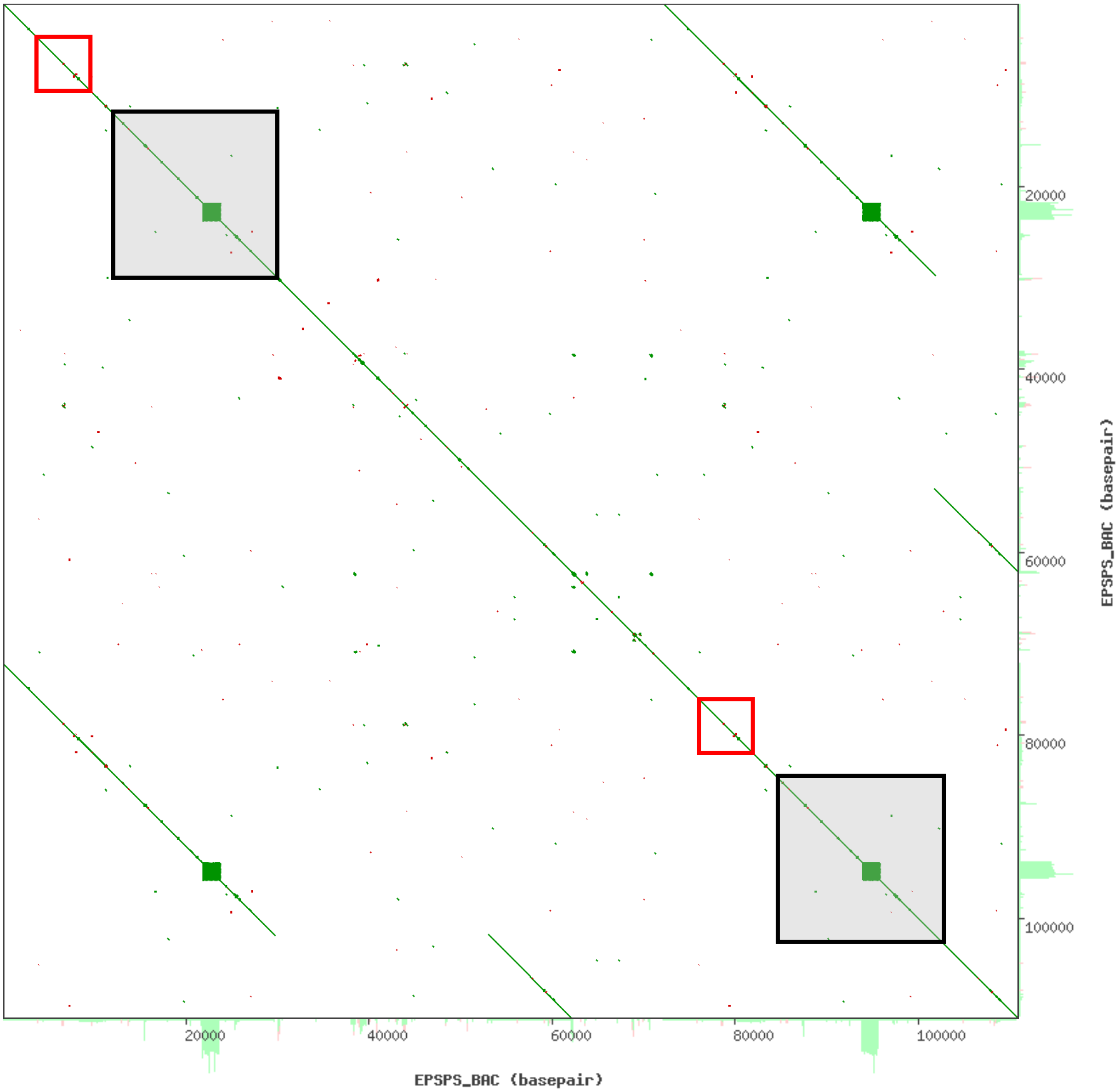

Figure 3.5: A self-alignment of the *EPSPS* contig from the resistant BAC assembly. The locations of the 2 *EPSPS* genes are indicated with red boxes. The Fhy/FAR insertion is denoted with black boxes

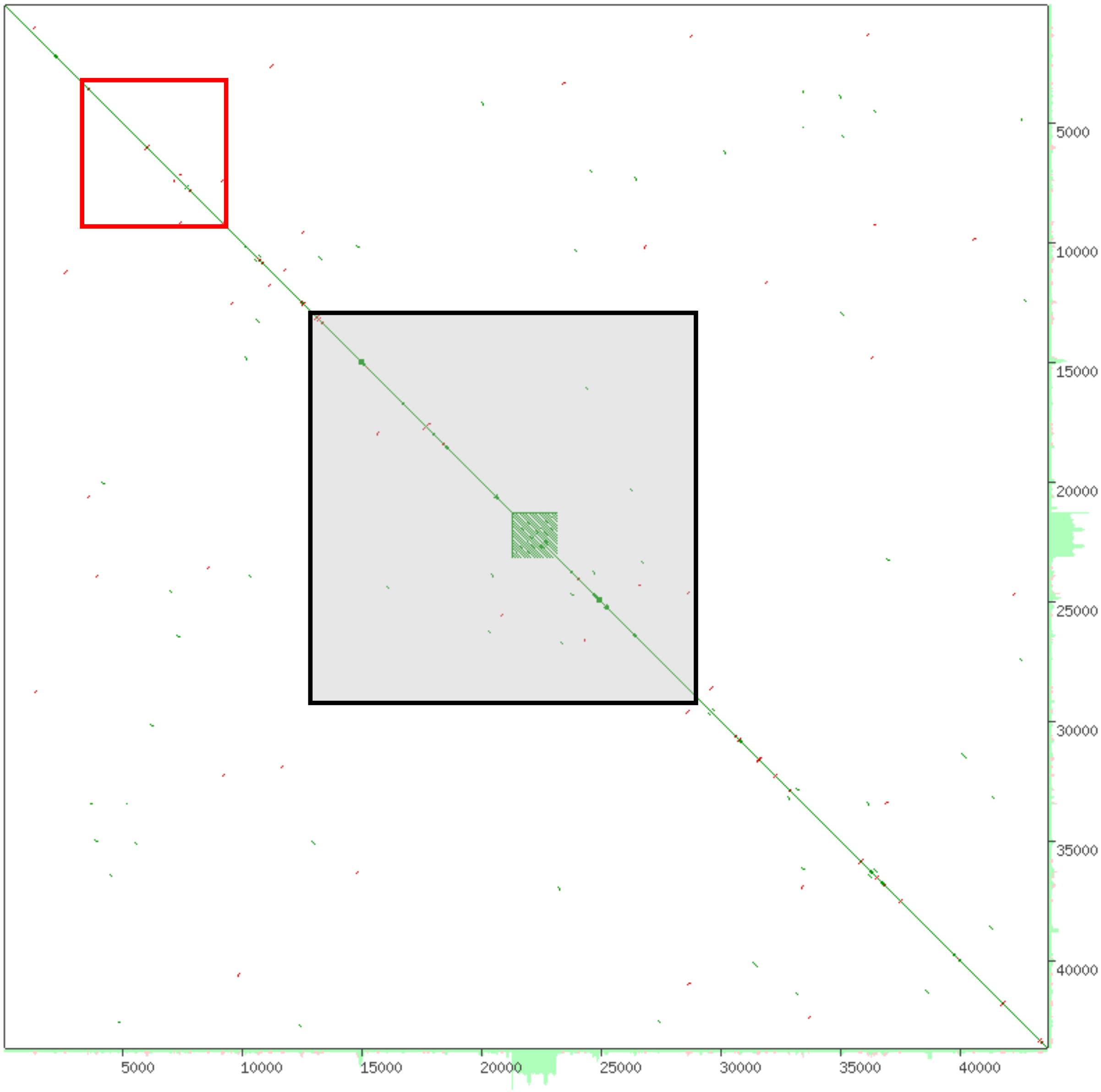

Figure 3.6: A self-alignment of the downstream contig from the resistant BAC assembly. The location of the *EPSPS* gene is indicated with a red box. The Fhy/FAR insertion is denoted with black boxes.

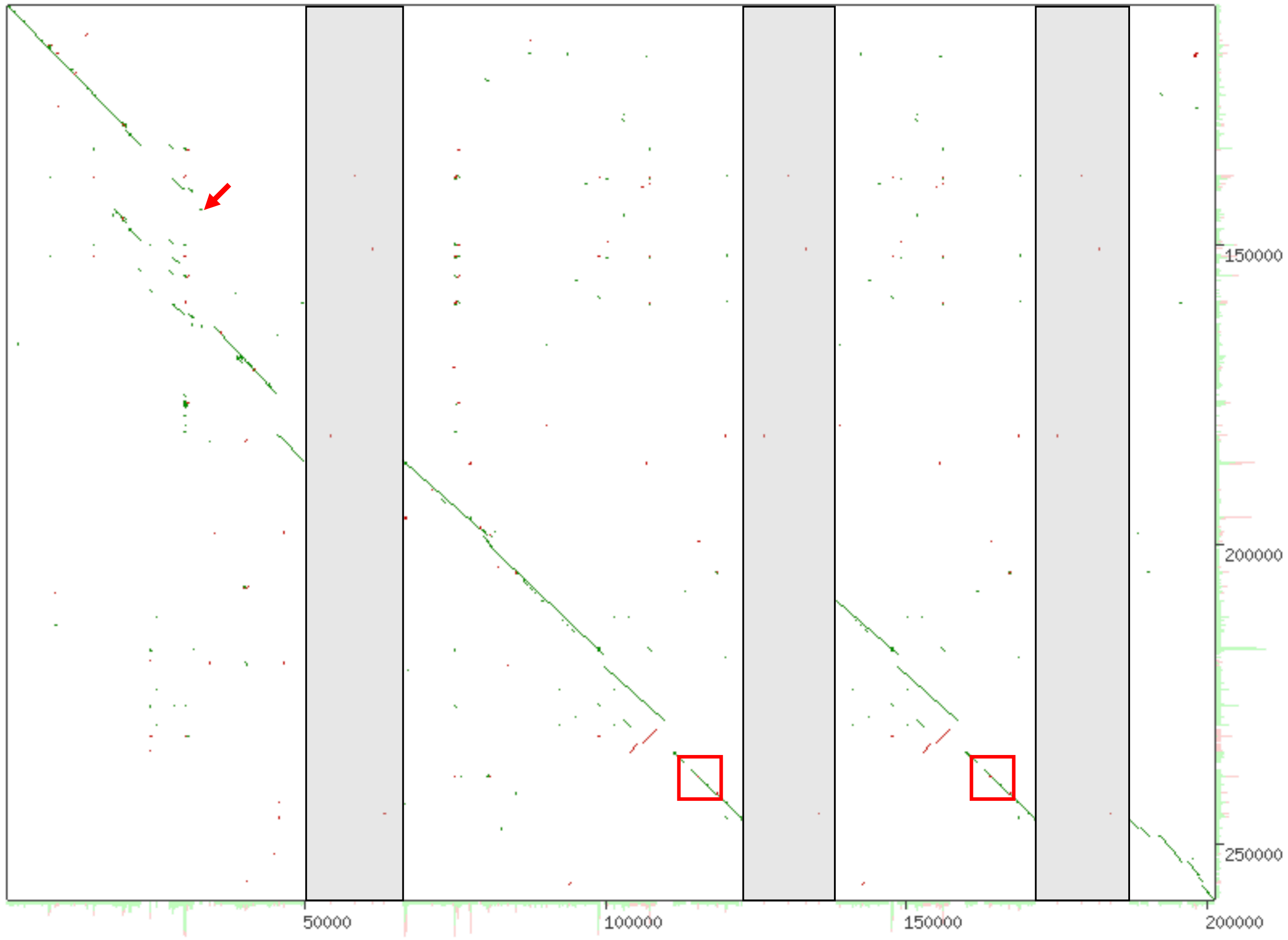

Figure 3.7: An alignment of a composite of all BAC contigs versus the *EPSPS* contig from the kochia genome assembly. The locations of the two *EPSPS* genes are indicated with red boxes. The Fhy/FAR insertions are denoted with the black boxes (there are no dots as this sequence is missing from the susceptible contig). The red arrow indicates the beginning of the *EPSPS* contig from the susceptible genome assembly

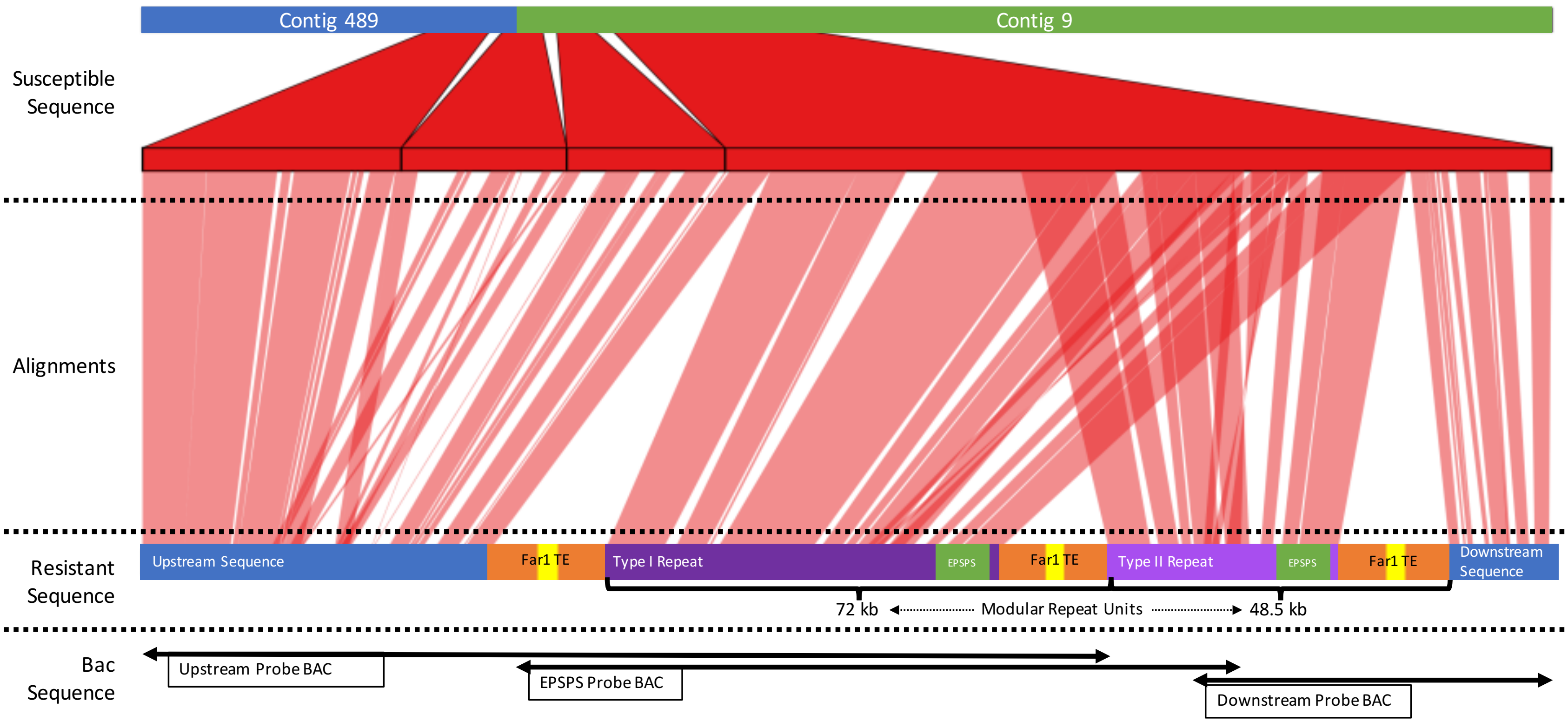

Figure 3.8: A Schematic of the *EPSPS* locus, insertion site of the Fhy/FAR insert, and the two Types of repeats.

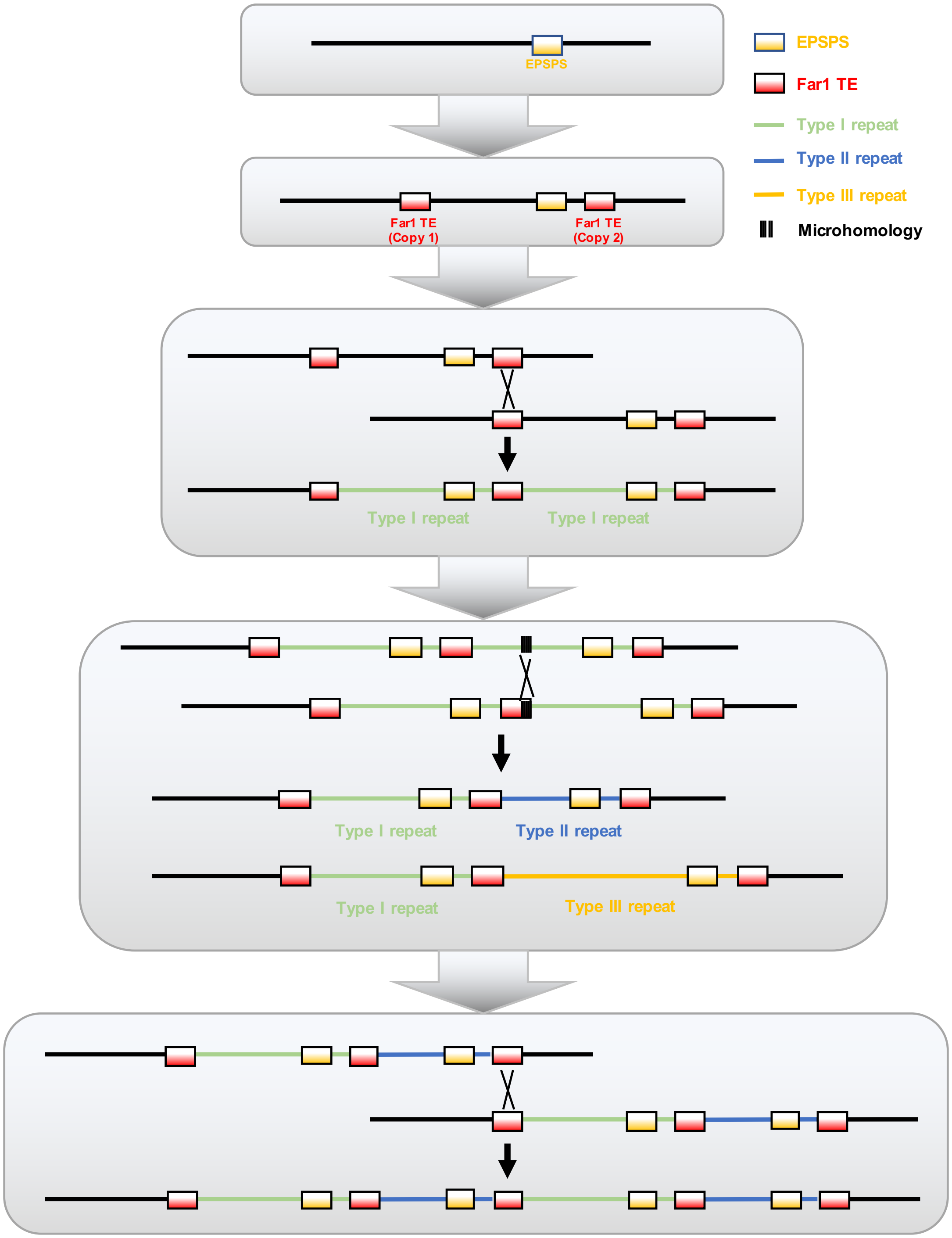

Figure 3.9: A hypothetical model for the generation and continued increase of *EPSPS* copy number. The initial event that led to *EPSPS* gene duplication was the insertion of two mobile elements both upstream and downstream of the *EPSPS* gene (Far1 TE). After unequal crossing over, gametes were produced with >1 *EPSPS* gene copy. Subsequently, a double stranded break occurred within the middle of the repeat region which was incorrectly repaired using microhomology mediated repair, instead using the end of the Far1 TE as the repair template, generating a shorter *EPSPS* copy (Type II).

REFERENCES

Cantarel, Brandi L., Ian Korf, Sofia M.C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. "MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes." *Genome Research* 18 (1): 188–96. doi:10.1101/gr.6743907.

Debolt, Seth. 2010. "Copy Number Variation Shapes Genome Diversity in Arabidopsis over Immediate Family Generational Scales." *Genome Biology and Evolution* 2 (1): 441–53. doi:10.1093/gbe/evq033.

Doyle, Jeffrey. 1991. "DNA Protocols for Plants." In *Molecular Techniques in Taxonomy*, 283–93. doi:10.1007/978-3-642-83962-7_18.

Duke, Stephen O, and Stephen B Powles. 2008. "Glyphosate: A Once-in-a-Century Herbicide." *Pest Management Science* 64 (4). John Wiley & Sons, Ltd.: 319–25. doi:10.1002/ps.1518.

Ersfeld, Klaus. 1994. "Fiber-FISH: Fluorescence In Situ Hybridization on Stretched DNA." In *Parasite Genomics Protocols*, 395–402. New Jersey: Humana Press. doi:10.1385/1-59259-793-9:395.

Finnegan, Jean, and David McElroy. 1994. "Transgene Inactivation: Plants Fight Back!" *Bio/Technology* 12 (9): 883–88. doi:10.1038/nbt0994-883.

Gaines, Todd A., Abigail L. Barker, Eric L. Patterson, Eric P. Westra, and Andrew R. Kniss. 2016. "EPSPS Gene Copy Number and Whole-Plant Glyphosate Resistance Level in *Kochia scoparia*." *PLoS ONE* 11 (12). doi:10.1371/journal.pone.0168295.

Gaines, Todd A., Alice A. Wright, William T. Molin, Lothar Lorentz, Chance W. Riggins, Patrick J. Tranel, Roland Beffa, Philip Westra, and Stephen B. Powles. 2013. "Identification of Genetic Elements Associated with EPSPS Gene Amplification." Edited by Jianwei Zhang. *PLoS ONE* 8 (6). Public Library of Science: e65819. doi:10.1371/journal.pone.0065819.

Gaines, Todd A., Wenli Zhang, Dafu Wang, Bekir Bukun, Stephen T Chisholm, Dale L Shaner, Scott J Nissen, et al. 2010. "Gene Amplification Confers Glyphosate Resistance in Amaranthus Palmeri." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 1029–34. doi:10.1073/pnas.0906649107.

Graur, Dan, and Wen-Hsiung. Li. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates.

Hudson, Matthew E., Damon R. Lisch, and Peter H. Quail. 2003. "The *FHY3* and *FAR1* Genes Encode Transposase-Related Proteins Involved in Regulation of Gene Expression by the Phytochrome A-Signaling Pathway." *Plant Journal* 34 (4): 453–71. doi:10.1046/j.1365-313X.2003.01741.x.

Hull, Ryan M., Cristina Cruz, Carmen V. Jack, and Jonathan Houseley. 2017. "Environmental Change Drives Accelerated Adaptation through Stimulated Copy Number Variation." *PLoS Biology* 15 (6). doi:10.1371/journal.pbio.2001333.

Jugulam, M., K. Niehues, A. S. Godar, D.-H. Koo, T. Danilova, B. Friebe, S. Sehgal, et al. 2014. "Tandem Amplification of a Chromosomal Segment Harboring 5-Enolpyruvylshikimate-3-Phosphate Synthase Locus Confers Glyphosate Resistance in *Kochia scoparia.*" *Plant Physiology* 166 (3): 1200–1207. doi:10.1104/pp.114.242826.

Koo, D. H., Molin, W. T., Saski, C. A., Jiang, J., Putta, K., Jugulam, M., Friebe, B., and Gill, B. S. (2018). Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri. *Proceedings of the National*

Academy of Sciences, 201719354.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive κ-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36. doi:10.1101/gr.215087.116.

Kurtz, Stefan, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. 2004. “Versatile and Open Software for Comparing Large Genomes.” *Genome Biology* 5 (2): R12. doi:10.1186/gb-2004-5-2-r12.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

Luo, Meizhong, and Rod A. Wing. 2003. “An Improved Method for Plant BAC Library Construction.” In *Plant Functional Genomics*, 3–20. New Jersey: Humana Press. doi:10.1385/1-59259-413-1:3.

Lynch, M., and J. S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” *Science* 290 (5494): 1151–55. doi:10.1126/science.290.5494.1151.

Maacke, Heiko, Sven Opitz, Kirsten Jost, Willem Hamdorf, Wilhelm Henning, Stefan Kr�ger, Alfred Ch. Feller, et al. 2000. “Over-Expression of Wild-Type Rad51 Correlates with Histological Grading of Invasive Ductal Breast Cancer.” *International Journal of Cancer* 88 (6). Wiley-Blackwell: 907–13. doi:10.1002/1097-0215.

Molin, William T., Alice A. Wright, Amy Lawton-Rauh, and Christopher A. Saski. 2017. “The Unique Genomic Landscape Surrounding the EPSPS Gene in Glyphosate Resistant

Amaranthus Palmeri: A Repetitive Path to Resistance." *BMC Genomics* 18 (1). BioMed Central: 91. doi:10.1186/s12864-016-3336-4.

Noé, Laurent, and Gregory Kucherov. 2005. "YASS: Enhancing the Sensitivity of DNA Similarity Search." *Nucleic Acids Research* 33 (SUPPL. 2). doi:10.1093/nar/gki478.

Ottaviani, D., LeCain, M. and Sheer, D., 2014. The role of microhomology in genomic structural variation. Trends in Genetics, 30(3), pp.85-94.

Patterson, Eric L, Dean J Pettinga, Karl Ravet, Paul Neve, and Todd A Gaines. 2018. "Glyphosate Resistance and EPSPS Gene Duplication: Convergent Evolution in Multiple Plant Species." *Journal of Heredity* 109 (2). Oxford University Press: 117–25. doi:10.1093/jhered/esx087.

Pettinga, Dean J, Junjun Ou, Eric L Patterson, Mithila Jugulam, Philip Westra, and Todd A Gaines. 2017. "Increased Chalcone Synthase (CHS) Expression Is Associated with Dicamba Resistance in *Kochia scoparia*." *Pest Management Science*, December. doi:10.1002/ps.4778.

Preston, Christopher, David S. Belles, Philip H. Westra, Scott J. Nissen, and Sarah M. Ward. 2009. "Inheritance of Resistance to The Auxinic Herbicide Dicamba in Kochia (*Kochia scoparia*)." *Weed Science* 57 (1). Weed Science Society of America: 43–47. doi:10.1614/WS-08-098.1.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. doi:10.1093/bioinformatics/btp616.

Russell, Peter J. 2002. *IGenetics*. Benjamin Cummings.

Sammons, Robert Douglas, and Todd A. Gaines. 2014. "Glyphosate Resistance: State of

Knowledge.” *Pest Management Science*. doi:10.1002/ps.3743.

Schimke, R T, A Hill, and R N Johnston. 1985. “Methotrexate Resistance and Gene Amplification: An Experimental Model for the Generation of Cellular Heterogeneity.” *British Journal of Cancer* 51 (4): 459–65.

Schmittgen, Thomas D, and Kenneth J Livak. 2008. “Analyzing Real-Time PCR Data by the Comparative C T Method.” *Nature* 3 (6): 1101–8. doi:10.1038/nprot.2008.73.

Sfeir, A. and Symington, L.S., 2015. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway?. Trends in biochemical sciences, 40(11), pp.701-714.

Tang, W, W Wang, D Chen, Q Ji, Y Jing, H. Wang, and R. Lin. 2012. “Transposase-Derived Proteins FHY3/FAR1 Interact with PHYTOCHROME-INTERACTING FACTOR1 to Regulate Chlorophyll Biosynthesis by Modulating HEMB1 during Deetiolation in Arabidopsis.” *The Plant Cell* 24 (5): 1984–2000. doi:10.1105/tpc.112.097022.

Tang, Wei, Ronald J. Newton, and Douglas A. Weidner. 2007. “Genetic Transformation and Gene Silencing Mediated by Multiple Copies of a Transgene in Eastern White Pine.” *Journal of Experimental Botany* 58 (3): 545–54. doi:10.1093/jxb/erl228.

Wang, Haiyang, and Wang Deng Xing. 2002. “Arabidopsis FHY3 Defines a Key Phytochrome A Signaling Component Directly Interacting with Its Homologous Partner FAR1.” *EMBO Journal* 21 (6): 1339–49. doi:10.1093/emboj/21.6.1339.

Wiersma, Andrew T, Todd A Gaines, Christopher Preston, John P. Hamilton, Darci Giacomini, C. Robin Buell, Jan E Leach, and Philip Westra. 2014. “Gene Amplification of 5-Enol-Pyruvylshikimate-3-Phosphate Synthase in Glyphosate-Resistant *Kochia scoparia*.” *Planta* 241: 463–74. doi:10.1007/s00425-014-2197-9.

Xi, Ruibin, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua

Zhang, Mark D Johnson, et al. 2011. “Copy Number Variation Detection in Whole-Genome Sequencing Data Using the Bayesian Information Criterion.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (46): E1128-36. doi:10.1073/pnas.1110574108.

SUMMARY OF DISSERTATION

The success of weedy plant species depends on their ability to rapidly adapt to new environments, tolerate novel stresses, and to compete with crops and desirable flora. In turn, these traits are determined by the genes in their genome and how those genes interact with environmental factors. Weed scientists that want to understand weedy traits at the molecular level depend on access to high quality genomic information. In Colorado, *Kochia scoparia* is the most important weed species in terms of economic impact. In the last decade, the ability to successfully control kochia has become more difficult as populations have evolved resistance to chemical control methods (herbicides), which have traditionally been the most effective and economic option.

Kochia has limited genomic information publically available. The nearest sequenced species is *Beta vulgaris*, which is quite diverged and has limited usefulness in investigating the genetics of the weedy traits found in kochia. To address this, we developed the first reference draft genome of kochia and used the genome as a platform to explore the hypothesis that genome plasticity in the form of gene copy number variation is an important weedy trait in kochia and that it might partially explain its success as a rapidly evolving weed. The reference draft genome was not complete (~80%) and remained highly fragmented (>19,000 contigs); however, the average contig length was much longer then a gene (~2,500 bp) and we were able to annotate >45,000 genes. Additionally, the contigs were long enough to perform a genome wide resequencing experiment to discover novel CNV events. We performed resequencing in a glyphosate resistant line to discover what regions, besides *EPSPS*, were being duplicated. We discovered thousands of potential novel CNV regions varying between these two lines. Most interestingly, the Fhy/FAR1 mutator-like transposases seem to be much more abundant in the glyphosate resistant line.

This work expands on what is known about genome plasticity and serves as a starting point for discovering novel genome rearrangements in this species. Furthermore, it gives the first description at the kinds of rearrangements that are associated with glyphosate resistance. With this tool and analysis in place, we can begin experiments to understand if these rearrangements are caused by the applied stress (i.e. glyphosate), whether they are co-selected with *EPSPS* CNV, and begin to understand how important CNVs are for generating genetic variation.

In this dissertation we also sequenced the *EPSPS* loci from a glyphosate susceptible (from the genome assembly) and from a resistant population using a BAC library. Several genetic elements were identified that, we believe, contributed to the evolution of *EPSPS* copy number variation in the resistant line. With the differences between the two lines, we constructed a model consisting of a series of events that explain one path to the initial duplication event and subsequent EPSPS copy number increases. The existence of two Fhy/FAR like transposons inserted flanking the *EPSPS* loci may have been the initial event that has led to subsequent unequal crossing-over and copy number variation of the *EPSPS* gene and the surrounding region.

We also discovered the genes that flank *EPSPS* and seem to be co-duplicated. The impact these co-duplicated genes have on normal plant physiology and possible fitness penalties remains unclear; however, the genomic tools we have developed will help tremendously in answering these questions in future work. This aspect of the dissertation highlights the amazing interplay between different genomic rearrangements; giving a concrete example of how transposable elements, like the Fhy/FAR transposon, can impact genome arrangement and structure beyond simply transposition. We now have mobile elements to investigate and look for in future studies.

I hope that the work done in this dissertation contributes to the communities' understanding of herbicide resistance, genome plasticity, and ultimately plant adaptation and evolution. The genome assembly of kochia allows us to explore new traits, new genes, and new ways the environment is shaping weed genome evolution. With this resource, we can now perform stronger scientific experiments including bulk segregate analysis (BSA), genome wide association mapping (GWAS), genotype by sequencing (GBS), and, once a transformation system is developed, directed transgenics for gene function discovery. In years to come, as genomics tools become more readily available and interest in invasive and weedy species increases, I believe weeds will be a source of new and amazing discoveries.