FHY3/FAR1 transposable elements generate adaptive genetic variation in the *Bassia scoparia* genome

Nathan Hall[1], Jinyi Chen [1,2], Christopher Saski[3], Phil Westra[4], Todd Gaines[4], Eric Patterson [1]

[1] Michigan State University, Department of Plant, Soil, and Microbial Sciences
[2] Nanjing Agricultural University, College of Plant Protection
[3] Clemson University, Department of Plant and Environmental Sciences
[4] Colorado State University, Department of Agricultural Biology

Abstract

Transposable element activity has greatly impacted the evolution of *Kochia scoparia* syn. *Bassia scoparia* (kochia), an important agricultural weed. In this study, an improved kochia genome assembly of the glyphosate-susceptible line 7710 was analyzed in comparison to a glyphosate-resistant line, M32. A near complete genome assembly consisting of 14 pseudochromosomes with a total length of 969.6 Mb, and N50 scaffold length of 99.88 Mb, was generated with hybrid short-read (Illumina), and single-molecule (PacBio/Bionano genomics) data, in addition to chromatin interaction mapping (Hi-C). The 9 largest scaffolds corresponded to the 9 chromosomes of a close relative, *Beta vulgaris*. From this assembly, 54,387 protein-coding gene loci were annotated. After gene annotation it was revealed that 58% of the whole-genome consisted of transposable elements. We identified that gene families containing *FHY3* (FAR-Red Elongated Hypocotyl 3) and *FAR1* (FAR-Red Impaired Response 1) functional domains have undergone a massive *Bassia*-lineage specific expansion. *FHY3* and *FAR1* domains were first described as a "domesticated" (non-active) sub-type of Mutator Don-Robertson transposase (MuDR) that function as transcription factors; however, the domains may still function as transposases in the correct genomic context, but this has yet to be observed. Mutator elements are associated with large DNA segmental transposition within the genome including gene duplication events, which can greatly influence plant phenotypes and evolution. We discovered in our work that putative MuDR elements with detectable *FHY3*/*FAR1* domains were tightly associated with segmental duplications of 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) that confer resistance to the herbicide glyphosate. Further, in a case study of MuDR activity, we characterized a new MuDR subtype, named here as "Muntjac", which contributes to the evolution of herbicide resistance in kochia through the process of transduplication. Collectively, our study provides insights into the de-domestication of a *FHY3/FAR1* transposon and provides new perspectives on the evolution of glyphosate resistance in kochia.

Keywords: Transposable Elements, *MuDR*, *FHY3/FAR1*, Glyphosate Resistance, Kochia, reference genome, genome resequencing, 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*), copy number variation

1. Introduction

Transposable elements (TEs) are engines of genetic change that can fundamentally alter an organisms genome and genetic diversity, even within relatively short evolutionary timescales[1]. TEs are active and mobile in the genome and propagate through replication cascades that are independent from the duplication of the background genome. While primarily thought of as detrimental as they can disrupt proper gene function and expression, there may be benefits from novel genetics that these TEs generate at the population and species level. The two major classes of TEs are defined by their propagation mechanism, known as Class I and II[2]. Class I TEs replicate via an RNA intermediate, a so-called "copy and paste" mechanism, while Class II TEs are without RNA intermediates and mobilize DNA through excision and re-insertion, via a "cut and paste" mechanism. When TEs insert into a new location in the genome, they may impart local structural changes to the DNA, such as methylation and gene expression[3,4]. At a chromosomal scale, TE insertions can facilitate larger-scale genomic rearrangements such as deletions, inversions, and segmental duplications[5,6]. At the gene scale, TEs can cause loss of gene function or gene shuffling through the process of transduplication[7–9]. Finally, and most rarely, TEs may contribute novel genes to the genome through a process of TE domestication, as in the case of the genes *FHY3* (Far-red elongated hypocotyl 3) and *FAR1* (far-red-impaired response). These

domesticated TEs were originally derived from a Class II TE known as a Mule (Mutator like) transposon in the MuDR (Mutator Don Robertson) TE clade but they have been hypothesized to have lost their ability to transpose and neofunctionalized as transcription factors that coordinate response to light signals[10].

All functioning, autonomous MuDR elements have a shared set of characteristics: 1) a DDE motif in the transposase domain, 2) terminal inverted repeats (TIR), and 3) target site duplications (TSD) that are created when the MuDR is inserted in a novel location[10]. When Mule elements move, including MuDR, their excision and reinsertion is known to pick up a 'payload' of genomic material composed of the genes found in the MuDR flanking sequence[11]. The transposase domain of MuDRs have been well characterized and is mechanistic in the complete excision of both the forward and reverse DNA strands from the genome, frequently leaving an excision footprint[12]. MuDR TIRs aid in the formation of their circularized double stranded DNA intermediate during transposition. TSDs form at the site of DNA insertion at a new location, which are typically short (8-11 bp) directed repeats adjacent to the TIRs[12]. *FHY3* and *FAR1* are notable in that as part of their domestication process they have lost their TIRs, resulting in the prediction that *FHY3* and *FAR1* cannot form the DNA intermediate needed for translocation of a Class II TE[13]. An undomesticated TE, named *Jittery,* is sister to the *FHY3* and *FAR1* genes and retains its TIRs and mobility[14].

Bassia scoparia (L.) Schrad. (kochia) is a common and economically important weed in crop and ruderal areas across the Canadian Prairies and the Great Plains of the USA[15]. Kochia is diploid with a somatic chromosome number of 2n=18 and an estimated haploid genome size of 0.7-1 Gb[16,17]. Glyphosate resistance in kochia is due to a tandemly arrayed copy number variation (CNV) of glyphosate's target, 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*)[18,19]. In previous work, a mobile genetic element (MGE) that contained a gene annotated as having a *FHY3* and a *FAR1* protein domain was identified flanking the *EPSPS* CNV suggested that the origin of the EPSPS CNV was closely related with *FHY3* and *FAR1* mobility[17]. In this research, we investigate a massive expansion of the *FHY3*/ *FAR1*-like gene family in the kochia lineage and provide evidence that some or even most of these *FHY3*/ *FAR1* like genes are not domesticated but are still actively mobile, at least in the kochia genome. Furthermore, we specifically describe the *FHY3/FAR1* MGE involved in glyphosate resistance and characterize it in the broader context of the MuDR transposon clade. We rename this specific transposing *FHY3/FAR1* gene "Muntjac", for the eponymous ship imagined by Lev Grossman in The Magicians Trilogy in order to differentiate it from domesticated *FHY3/FAR1* genes[20,21]. We propose that the transposable element, like the fabled ship, is highly mobile, at least in the kochia genome and that it has exhibiting degrees of contraction and expansion even at recent timescales. Furthermore, we propose that the activity of Muntjac is a signature of the evolutionary trajectory of the kochia genome in response to human selection from the herbicide glyphosate.

2. Material and methods

2.1. Germplasm

Several collections of glyphosate susceptible and resistant kochia plants were used for these experiments. The herbicide-susceptible reference genome assembly was generated from the highly inbred lab strain known as '7710', originally collected from western Nebraska[22]. This line has been used extensively as a susceptible control for many herbicide studies. The 7710 line has a consistent phenotype and it is well-studied compared to many kochia collections[23,24]. The primary resistant line used for comparison was the inbred lab line known as 'M32', collected from north-eastern Colorado[25], which has been confirmed to have highly amplified copy numbers of *EPSPS* and consistent (non-segregating) inheritance of the glyphosate resistant phenotype[17,26]. Furthermore, the *EPSPS* locus in M32 has been extensively studied and assembled using BAC sequencing[17].

To discover the original locus of Muntjac we collected two glyphosate-susceptible populations, one located from Israel in the native range of kochia (referred to as Israel), and a second from northern Colorado (referred to as CO-S). Additionally, we used glyphosate resistant plants with high *EPSPS* copy number but no amplification of the Muntjac gene (referred to as Oregon)[27]. For all studies, seeds were sowed in SureMix® and kept in growth chamber (25 °C in the day and 20 °C at night, 14/10 h light period) and watered every other day.

2.2. Kochia genome sequencing and assembly

DNA was extracted from leaf tissue of a single glyphosate susceptible kochia plant from the line 7710 using a modified CTAB protocol[17,28,29]. DNA was sent for PAC-BIO Sequel sequencing at The University of Delaware DNA Sequencing & Genotyping Center. Expanding leaf tissue was sent to Phase genomics for HiC library prep and sequencing. An optical map was constructed at the McDonnell Genome Institute using the BioNano Genomics Saphyr system. Short, high-coverage, paired end reads were generated at the University of Illinois on an Illumina HiSeq. These Illumina reads were used to assess the genome size of the kochia glyphosate-susceptible line 7710. Paired-end Illumina reads (144.34 Gb) (with an insert size of 350 bp) were selected to generate 17-mer frequency data for k-mer analysis. The k-mer depth distribution was plotted against the k-mer frequency, with the highest peak occurring at a depth of 97. Based on the total number of k-mers (71,593,662,168 bp), the kochia genome size was calculated to be 969.61 Mb, with genome heterozygosity of 0.06%. Additional flow cytometry data confirmed this size estimate with individuals from 900-1050 Mb. Like most chenopods, kochia has a chromosome count of n = 9[30].

Long read sequences were initially assembled using MeCat2[31]. Short reads were then aligned to this assembly using HiSat2 v2.1.0[32]. The assembly and alignment were used in conjunction to polish the genome using Pilon[33]. Phased haplotigs were merged using HaploMerger2[34]. Merged contigs were scaffolded using HiC data and the SALSA pipeline[35]. After initial scaffolding, the assembly was further scaffolded using the optical map at the McDonnell Genome Institute and the BioNano Solve software[36]. This assembly was then gap filled using PBJelly[37]. Newly filled gaps were polished once again using Pilon. Finally, super-scaffolds were manually scaffolded into pseudo-chromosomes using the HiC data and Juicer[38].

Gene models in the assembly were predicted using the GenSAS pipeline (https://www.gensas.org/gensas)[39]. The genome assembly was uploaded to the GENSAS webserver along with RNA-seq data for annotation[23]. The RepeatMasker and RepeatModeler tools were used to identify the repeats, and mask the genome for gene model calling[40,41]. Publicly available RNAseq data (NCBI SRA: SRX3276189- SRX3276192) were aligned with the masked genome, using BLAST+, BLAT, HISAT2, TopHat, and PASA [42–45]. The genome and these alignments were run through the Augustus, BRAKER, GeneMark-ES, Genscan, and SNAP tools for gene model prediction with default parameters[46–49]. Next, we made consensus gene models using EvidenceModeler. Each predictive tool was uniquely weighted. The following final weights were assigned to each prediction: 10 for transcript alignments and BRAKER; 5 for Augustus and protein alignments; 1 for GeneMark-ES and SNAP. PASA refinement was conducted using BRAKER-hisat2 as OGS. For the functional annotation, predicted genes were aligned to the NCBI plant protein database using both BLASTp and Diamond. Targeting and subcellular localization was predicted using Pfam, SignalP and TargetP, with default parameters for plants. Additional functional annotation was performed with Multiloc2, Iprscan, and an MMseqs2 search of Uniref50 database[50–52]. The final annotation was solidified after running BUSCO using the embryophyta_odb10 lineage database on the predicted proteins to assess the completeness of the annotation[53] .

2.3. Gene Family Expansions and the FHY3/ FAR1-like gene family in kochia

To determine the history of Muntjac and if *FHY3*/*FAR1* domains were prevalent in gene expansions, orthogroup clustering was performed using Orthofinder v2.4.0[54]. All annotated genes in the kochia genome were subjected to orthologous gene clustering with all annotated genes from eight closely related species from the Caryophyllales clade (Chenopodiaceae and Amaranthaceae). The species tree and related gene trees were inferred by Orthofinder using representative proteins produced using CGAT v0.06[54,55]. A CAFÉ analysis was conducted to detect significant expansions within orthogroups[56]. Distances within species tree produced by Orthofinder were scaled by a factor of 100 with ETE3 and CAFÉ was run with default settings[57].

The *FHY3/FAR1*-like susceptible protein described by Patterson et al[17] was used to validate the annotation of Muntjac. The gene sequence was identified by protein blast to the assembly and reads from three different SRA runs (ERR364385, SRR6165247, and SRR1198330) were mapped to the gene with HISAT2 to update the annotation. These results were compared to HISAT2 junction files, and matching junctions were searched for in 31 SRAs. The manually produced transcript was searched for ORFs using NCBI ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/ last accessed 3/8/2021), searched for conserved domains using NCBI Conserved Domain Search

(https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi last accessed 3/8/2021) and using InterPro online search portal (https://www.ebi.ac.uk/interpro/, last accessed 3/8/2021).

Past research has demonstrated that kochia genomes may exhibit appreciable structural variability. To classify types of gene duplication and detect perturbations to genome structure, MCSCANX was run according to its documentation. To confirm structural differences between the glyphosate susceptible line 7710 and glyphosate resistant line M32, micro-assemblies were created for the genomic regions surrounding an intact Muntjac (Bs.00g352450) and a fragmented Muntjac insertion (Bs.00g145150). Resistant reads were mapped against the kochia genome with HiSat2, and mapped reads from target genes and the surrounding sequence were extracted with SAMTools v1.10[58]. Extracted reads were assembled with SPAdes v3.14.0 (default settings)[59]. Assembled regions from resistant lines were compared back to each locus of origin to determine if an excision had occurred and search for target site duplications and terminal inverted repeats (TIRs).

To determine the relationship of Muntjac to *FHY3*, *FAR1* and *Jittery*, a gene tree was produced using protein sequences that represented several TE families including Jittery, Hop, and MuDRs. Briefly, full protein records were found on NCBI by searching reported transposase domains against NCBI using BLAST online search utility with searches including eight NCBI taxon IDs. Amino acid sequences were aligned and curated. Curated alignment was analyzed using IQTree v1.6.12 for Windows with flags (-alrt 1000, -bb 1000). For the purposes of visualization, the reduced set of proteins was searched for Pfam domains using NCBI Conserved Domain Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, last accessed 21/8/2021). Detected Pfam regions were compared with a down sampled gene tree produced with ETE3 v3.1.2 and Python v6.5.3 (https://github.com/NDHall/kochia_scoparia_analysis). Figures prepared with Inkscape v0.92.4 (https://inkscape.org/, last accessed 3/8/2021).

2.4. Assessing copy number differences and expression differences.

To determine genome wide copy number variation (CNV) in the previously characterized glyphosate resistant line M32, an individual was selected for genomic Illumina resequencing[17]. Paired end data were aligned to the susceptible reference genome assembly (i.e. 7710) using HiSat2 on standard parameters. Additionally, Illumina paired end data from the susceptible plant was also aligned to the reference to control for shared CNV events between the resistant and susceptible plants using HiSat2 on standard parameters. The resulting alignments were scanned using CNVnator v0.4.1 and a 1,000 bp sliding window to detect read coverage changes outside of the normal read depth distribution[60]. Duplicated and deleted regions were identified using a read depth cutoff of 2x and 0.5x coverage for duplications and deletions (respectively) and an adjusted p-value < 0.05. CNV regions shared between the susceptible and resistant lines were removed from consideration as they most likely exist in the background of the kochia genome and were not subject to glyphosate selection. The genes in the resulting glyphosate-resistant CNVs were collected and the gene ontology terms associated with those genes were analyzed for GO enrichment.

To measure the relationship of predicted gene copy number and expression, publicly available RNAseq data (4-7710 and 4-M32 individuals) was aligned to the genome-predicted transcriptome using HiSat2[17,32]. Counts were extracted from alignments using samtools idxstats and differential expression was calculated using recommend settings in EdgeR[58,61]. Relative expression was correlated to relative gene copy number differences between the S and R lines.

2.5. Muntjac and its MGE activity and transposition

PCR primers were designed to identify 1) the upstream and downstream boundaries of the MGE containing Muntjac's insertion on chromosome 9 (chr9), 2) the excision of the MGE from chr9 in a glyphosate resistant individual, 3) the insertion of the MGE at the *EPSPS* locus on chr1, and the 4) the presence of the flanking genomic sequence at the chr9 position. For this assay, we used several individuals from the five kochia collections described in section 2.1: 7710 (S), CO-S (S), Israel (S), M32 (R), and Oregon (R). DNA was extracted from individual plants of all populations. There were two replicates per population. Primers were designed based on kochia chr9 assembled sequences and were blasted to the whole kochia genome to confirm their specificity. The five pairs of primers were used for validation (Table 3). The PCR was conducted in a 25-µL volume, consisting of 1 µL of DNA, 1 µL of each primer (10 µM) and 12.5 µL of GoTaq Green Master mix (Promega®, USA). The PCR

was run in a thermocycler (BioRad, USA) with the following profile: 95 °C for 5 min, 35 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min, followed by a final extension step of 5 min at 72 °C.

3. Results

3.1. The Kochia Genome

The genome was assembled using a combination of high coverage RSII PacBio single-molecule real-time (SMRT) sequencing (130 Gb of raw data), deep-coverage Illumina shotgun sequencing data (144 Gb of raw data), a Hi-C linkage dataset (87 Gb of raw data), and a BioNano Optical map. The assembly consisted of 14 scaffolds with an N50 scaffold length of 99.88 Mb and an N50 contig length of 0.91 Mb. Total sequenced length of the assembly was 837.60 Mb, with 132.0 Mb in gaps. The scaffolds were ordered and numbered as pseudochromosomes against the *Beta vulgaris* genome, kochia's phylogenetically nearest sequenced relative. Due to *Beta vulgaris* and kochia being significantly phylogenetically divergent, several inversions, duplications and deletions are apparent in the whole genome alignment. The 5 remaining, much smaller, scaffolds from kochia remain unincorporated in the chromosome assembly due to a lack of information placing them in the larger pseudomolecules. For the most part, they are not included in further analyses. The average length of the 9 pseudomolecules is 100.6 Mb, among which the shortest was 84.85 Mb (chr2), and the longest was 117.96 Mb (chr5) (Figure 1).

Structural and functional gene annotation revealed that a majority (95.0%) of the 1,614 BUSCOs in the embryophyta_odb10 BUSCO dataset were either complete or fragmented, which indicated a high degree of assembly and annotation completeness (Table 1). In total, 54,387 gene models were predicted, 45,372 received functional annotation and 8,333 unique Interpro annotations were assigned to 30,046 genes and 2,234 unique Gene Ontologies were assigned to 19,357 genes.

3.2. Gene Family Expansions and the FHY3/ FAR1-like gene family in kochia

Orthologous gene clustering using CAFÉ and all annotated genes from eight closely related species from the Caryophyllales clade (Chenopodiaceae and Amaranthaceae) revealed 206 of gene families that are significantly expanded with 0 families contracting in kochia relative to all other chenopods (Figure 2, Table 2). Of the expanded gene families, 10 gene families contain proteins annotated as containing FHY3/FAR1 family and/or FAR1 DNA binding domain (IPR031052 and IPR004330 respectively) accounting for 267 genes total. Furthermore, using the results from Orthofinder and MCSCANX we see an especially high enrichment of *FHY3* and *FAR1* related genes annotated with IPR004330, IPR031052, or IPR037766 interpro identifiers within segmental duplications. A total 80 of the 282 whole genome/segmental duplications contained at least one *FAR1* related gene that had undergone duplication. *FAR1* related genes and their segmental duplications seem to be concentrated in gene rich areas of the chromosomes rather than the centromere (Figure 1).

3.3. Copy number variation and expression differences

Previous work identified a CNV event that increases the copy number of the gene *EPSPS*, the target of the herbicide glyphosate, which in turn increases mRNA and protein production. The increased EPSPS protein production due to the *EPSPS* CNV is not inhibited by typical glyphosate application rates and therefore confers high levels of glyphosate resistance[62–64]. This CNV is widespread in kochia populations throughout North America[27,65]. Several different arrangements of the *EPSPS* CNV have been suggested[27]; however, the *EPSPS* CNV event being described here is confirmed 1) to be in a tandem arrangement (i.e., not dispersed) with two distinct repeat units of different sizes, 2) to include 7 genes (one of which was *EPSPS*), and 3) to be associated with the insertion of one or more 15 kb mobile genetic element(s) (MGE) containing the complete *Muntjac* gene. This specific arrangement was previously confirmed in the resistant line M32 by BAC sequencing[17]. This MGE is flanked by TIRs suggesting that the MGE is composed of Muntjac plus additional genetic payload.

Shotgun Illumina resequencing from a glyphosate resistant line with an *EPSPS* CNV was aligned to the susceptible genome assembly. CNV detection was then performed based on changes in read depth using CNVnator. Dozens of small CNV events were predicted between the glyphosate susceptible and glyphosate resistant lines but none so large or nearly as high copy number as the *EPSPS* CNV event. The MGE is present as a single copy in the susceptible genome chr9, and this MGE sequence from the susceptible genome is nearly identical to the MGE located next to *EPSPS* on chr1 in the resistant line

(>99.9% sequence similarity). Breaks in read mapping coverage indicate the presence of a small structural variant between the MGE at chr9 in susceptible and the MGE on chr1 in resistant, which is also detected by CNVnator (Figure 3). RNA-Seq data from the two populations show that slight changes in predicted genomic copy number do not correlate well with expression changes for several the 7 genes located in the *EPSPS* CNV. The additional copies of *EPSPS* resulted in higher levels of *EPSPS* expression in a linear correlation.

3.4. Muntjac activity and transposition

Muntjac and the MGE contain three key traits of an active autonomous transposase: 1) an intact Mutator domain, 2) target site duplication, and 3) TIRs. The results of CNVnator and read mapping indicate that Muntjac is internal of the MGE that is present in the CNV related to herbicide resistance in kochia. Interruptions and variations in the aligned reads are due to the MGE containing Muntjac that occurs at an entirely different locus within the glyphosate resistant line consistent with transposase activity (Figure 3). Micro-assemblies of the susceptible Chr9 and resistant Chr1 insertion sites of the entire MGE provide further support for Muntjac as an autonomous Mutator Transposase (Figure 4). First, a target site duplication was detected at both the insertion next to *EPSPS* on Chr1 (~bp 6,110,000) in the resistant line as well as at its location on Chr9 (~bp 98,280,000) in the assembled susceptible line. Second, a TIR of the appropriate length (88 bp) was identified adjacent to each TSD. A survey of protein domains within MuDR type demonstrates that Muntjac retains key conserved domains for transposase activity including a Mutator domain (MULE) and a SWIM zinc-finger (Figure 5). The detection of a putative YCF1 (Yeast Cadmium Factor-1) domain within the Muntjac predicted protein is unique and does not appear to be in any other *FHY3* or *FAR1* genes in the susceptible genome. Finally, phylogenetic analysis places Muntjac within the MuDR TEs, sister to *FHY3* and *FAR1*. This placement of Muntjac within the MuDR clade indicates that Muntjac, *FHY3,* and *FAR1* are more closely related than they are to Jittery (Figure 6). It should be noted that we also identified, through functional annotation, what might be another copy of Muntjac at a novel position on Chr1 (~bp 22,725,000); however, it was missing TIRs and had fused with another unrelated gene suggesting that this copy is now non-functional (Figure 4).

3.5. Molecular verification of Muntjac activity and transposition

Active MuDR elements are mobile and therefore are at different loci in different individuals. To determine if Muntjac exhibited mobility with the kochia genome, we designed five pairs of PCR primers to detect the absence/presence of Muntjac at Chr9 (~bp 98,280,000) (Table 3) in a variety of kochia plants from diverse locales and glyphosate resistance. For these assays, if the MGE at its position in Chr9, the primer pair F1/R1 will exhibit no bands since the amplicon is too long to be amplified with the applied settings. Additionally, if the MGE is present at the Chr9 locus, then primer pairs F1/R4 and F3/R3 should amplify since R4 and F3 are both located within the conserved Muntjac sequence in the MGE and F1 and R3 are located in the flanking region, effectively detecting both the upstream and downstream boundaries of the MGE. These primers were tested and validated using DNA from populations M32 and 7710 since we have genomic information about these lines. As predicted, the F1/R1 primer pair did not amplify in 7710 since the MGE fragment (~16,000 bp) was too long to be amplified under the PCR setup (Table 4). In contrast, M32 plants had the predicted ~900 bp target band, indicating the MGE was not in its Chr9 position in this resistant line. Both M32 and 7710 showed target length single bands with these two primer pairs. Primer pairs F1/R4 and F3/R3 amplified the junction between MGE and the insertion point, which further demonstrated the presence of the full MGE in Chr9. As expected, 7710 plants showed target band with both primers while M32 plants did not. Primer pairs F1/R2 and F2/R3 were used to verify that the assembled DNA sequence flanking the Chr9 insertion site was present and correct (Figure 7, Table 4).

We also screened novel kochia populations that had no genomic data, both with and without the *EPSPS* CNV and glyphosate resistance, to determine if the MGE containing Muntjac on Chr9 is the native location and the position on Chr1 in the resistant line M32 is the novel insertion site. We selected a resistant line from Oregon and plants from two susceptible populations CO-S and Israel. In these populations, plants Oregon-1, CO-S-1, Israel-1, and Israel-2 showed target length bands with primer F1/R1, suggesting the MGE is not present at the Chr 9 locus in these resistant and susceptible populations. As anticipated from this result the primer used to test the boundaries of the MGE, F1/R4

and F3/R3, also did not amplify (Table 4). Primer pairs F1/R2 and F2/R3 confirmed the DNA around the Chr9 insertion line was present and normal (Table 4). Therefore, in these plants, MGE is not inserted in the chromosome 9 locus as was found in the 7710 population. For plants Oregon-1 and CO-S-1, there were target bands under F1/R2 and F2/R3, but three primer pairs F1/R1, F1/R4 and F3/R3, used to detect the MGE presence/absence did not amplify and confirm its presence or absence. In Oregon-2, there was no MGE insertion at Chr9, as F1/R1 worked, and F1/R4 and F3/R3 did not. However, in this individual, multiple bands were identified with primers F1/R2 and F2/R3. Results in this plant suggest the presence of nucleotide variations in this region among different kochia individuals. The PCR product of F1/R1, F1/R2 and F2/R3 from each population was sequenced. All sequences were highly similar with the kochia genome template, although some indels were present. Overall, the primer pairs used for validation here can be applied to most kochia populations and demonstrated that MGE position varies with kochia plants/populations and that the MGE at the Chr9 position is not shared among all susceptible populations.

4. Discussion

4.1. The Kochia Genome

The genome of Kochia scoparia comprises approximately 1 Gb of genetic material, which is characterized by highly repetitive and challenging to assemble centromeres and telomeres, analogous to other chenopod species. Through an integrative approach that employed hybrid short and long read sequencing techniques, nucleosome contact mapping, and optical mapping, we achieved an assembly of approximately 83.8% of the total sequence data, covering 97% of the predicted genome length. Although numerous gaps remain in these scaffolds, the gaps are notably gene-poor and consist mainly of highly repetitive sequences. The assessment of gene space completeness using the embryophyte odb10 dataset via BUSCO estimates was found to be 95.0%. Although the development of newer sequencing and assembly technologies may allow for a more contiguous Kochia genome, such refinement was not necessary for the present study, given the assembly's comprehensive coverage of gene space and relevant EPSPS and Muntjac loci in our susceptible reference.i

Beta vulgaris represents the nearest sequenced species to kochia despite being phylogenetically distant. Like most chenopods, both species have nine chromosomes. Upon alignment of the two genomes, a significant degree of macro-synteny and collinearity is observed. However, owing to their evolutionary divergence, numerous inversions, gaps, and transpositions have occurred. Notably, the regions adjacent to the centromere are particularly divergent. It may be that misassembles of the kochia genome also account for some of these differences. Furthermore, available data indicates that the Beta vulgaris genome is estimated to encompass approximately 758 Mb, approximately 75% of Kochia's genome size. The basis of this discrepancy is not yet apparent; however, preliminary analysis indicates the presence of a high number of transposable elements (TEs) in Kochia. Notably, there is currently no evidence of any polyploidy events or significant genomic duplication in either species since their divergence.

4.2. FHY3/ FAR1-like gene family in kochia

When the annotated genes of kochia are compared to other sequenced chenopods and amaranths, there are several gene families with conserved gene functions that are significantly expanded in kochia according to our Café Analysis (Figure 2). Many of these enriched families contain functions related to TEs (*e.g.* IPR005162: Retrotransposon gag domain, IPR001878: Zinc finger, CCHC-type superfamily, and IPR00752: Zinc finger, SWIM-type); however, of all the enriched genes, members of the *FHY3/FAR1* family are particularly interesting (Figure 1). Past research indicated that one of these *FHY3/FAR1* genes was in a MGE that caused a CNV of the gene *EPSPS*. This CNV, in turn, endows resistance to the herbicide glyphosate and is the major mechanism by which kochia has adapted to this herbicide[18,26]. *FHY3/FAR1* are generally considered one of the rare cases of 'domesticated' transposons and have been though to lose their ability to readily translocate and self-duplicate[13,66]. The expansion of the *FHY3/FAR1* may indicate that the *FHY3/FAR1* genes have either retained their transposable element capabilities or they have re-acquired transposable element capabilities in the *Bassia scoparia* lineage.

The interaction between genomic rearrangement(s) and adaptation underscores the profound impact that genomic plasticity can have on plant evolution, particularly under sustained abiotic stresses. While

386 not a novel concept, previous research has shown that TE activity and CNVs are essential pathways for
387 generating novel genetic diversity. Nevertheless, the potential contribution of structural variation as a
388 driver of adaptive genetic variation is often underestimated as supported by several studies[67,68].
389 Structural variation has the unique advantage of changing gene expression without changing the protein
390 structure or the regulation of the promotor region. This process is not stable and instead is dynamic and
391 responsive to abiotic stress with CNV numbers being driven higher by higher levels of selection[69].
392 This study utilized deep-coverage, short read genomic sequences of a glyphosate resistant individual to
393 examine the *EPSPS* locus and the *FHY3/FAR1* MGE described previously [17] for mapping the location
394 of the MGE. A genome-wide survey was conducted to explore potential loci, beyond the EPSPS locus,
395 that demonstrated high frequency within the resistant genome. As anticipated, given the sequence
396 divergence between the two Kochia individuals, the analysis revealed hundreds of small candidate copy
397 number variations (CNVs) at both lower and higher copy number than the EPSPS locus. Many of the
398 CNVs at very high read depths contained genes annotated as TEs. Most CNVs had little correlation to
399 gene expression, indicating that the plants were able to maintain some level of control over the
400 expression of duplicated genes from these other CNVs, the exception being *EPSPS* in the *EPSPS* CNV.
401
402 *4.3. Muntjac activity and transposition*
403 Muntjac's locus is variable within the genomes of several kochia samples. We designed PCR
404 markers that amplify if, and only if, the MGE is at its position on chr9 like the susceptible genome.
405 Additionally, we developed markers that amplify only if the MGE is absent (Figure 7b). The former
406 primers span the upstream and downstream boundaries of the MGE and the surrounding genome, while
407 the latter span the insertion site. If the entire MGE is present, the amplicon is too large to amplify (~16
408 Kb) but if it is absent, it is only 900 bp. We ran these molecular markers on the susceptible kochia line
409 from which the genome was developed as well as from the glyphosate resistant line from which we first
410 identified Muntjac. We confirm that Muntjac-MGE is indeed missing from its position on chr9 in the
411 glyphosate resistant line, M32, while it is present in the glyphosate susceptible line, 7710. From these
412 data we hypothesized that the Muntjac-MGE transposed from its position on chr9 to the EPSPS locus
413 on chr1 using a "cut-and paste" mechanism. To test whether chr9 is indeed the "native" location for
414 Muntjac we used our markers on two unique susceptible lines (divergent to the genome line) and a
415 glyphosate resistant line unrelated to the glyphosate resistant line we have been working with. Muntjac
416 was not at the position on chr9 or the position on chr1 (underlying resistance in concert with the *EPSPS*
417 CNV), in any of these populations or individuals, which indicates that a conserved, original, locus of
418 Muntjac is still unknown if it exists at all, and that the location on chr9 in the susceptible genome and
419 insertion in chr1 in the resistant line are due to independent and recent insertions at these loci. Recently,
420 *EPSPS* duplication without Muntjac as part of the CNV has been identified in some kochia populations,
421 indicating multiple unique paths to *EPSPS* copy number variation[27]. This suggests high levels of
422 structural variation and therefore genetic diversity exists in the kochia species as a whole.
423 The *FHY3/ FAR1* gene in the MGE that caused the *EPSPS* CNV is complete, containing key protein
424 domains for transposition. Close molecular examination of the MGE and Muntjac in R lines from
425 Oregon show evidence of differential transposition when compared to resistant and susceptible lines
426 from Colorado and Israel. This line of evidence demonstrates the variability of the insertion site
427 described from M32 on chr9 and suggests that the insertion that caused *EPSPS* duplication in M32
428 occurred recently. The detection of TSD, TIRs, and the apparent current mobility of Muntjac provides
429 strong lines of evidence supporting active transposition of Muntjac within kochia.
430
431 **Conclusion**
432 We assembled the genome of a glyphosate susceptible kochia plant to chromosome scale. By
433 performing orthologous gene clustering of related chenopods and amaranths, we observed an expansion
434 in genes containing *FHY3* and/or *FAR1* annotations in kochia. Resequencing of a glyphosate resistant
435 plant using high coverage Illumina data revealed the location of a MGE that is involved in the formation
436 of the *EPSPS* CNV. Within this element, a gene annotated as having *FHY3* and *FAR1* protein domains
437 was identified. Furthermore, this element had signature target-site duplications, terminal inverted
438 repeats, and a functional transposase domain in the *FHY3/FAR1* gene. We believe this element has all
439 the necessary features to be an autonomous Class II transposon and represents a "de-domestication" of
440 the *FHY3/FAR1* like genes. We therefore rename this element as "Muntjac" to distinguish it from other

441 Mules and *FHY3/FAR1* related genes. We verified using molecular markers that Muntjac exists in at
442 least two locations: Chr9 (~bp 98,280,000) in the glyphosate susceptible individual we assembled and
443 flanking the glyphosate target gene *EPSPS* on Chr1 (~bp 6,110,000) in a glyphosate resistant individual.
444 Furthermore, we used these markers to attempt to decipher which one of these locations is Muntjac's
445 "native" location; however, in other susceptible lines (namely plants from Israel, from kochia's native
446 range), Muntjac is at neither location. We hypothesize that Muntjac is still very active within kochia
447 and that its native location and recent activity may only be resolvable using a more in-depth pan-
448 genomic approach. Regardless, we believe Muntjac's involvement with generating glyphosate
449 resistance alleles is an strong demonstration of the critical role TEs play with the production of novel
450 variation and therefore the adaptive potential of weedy species.
451
452 **Data availability.**
453 The assembled genomes, associated GFF annotation files, and all functional annotation information
454 are publicly available through the International Weed Genomics Consortium online database,
455 'Weedpedia' (https://weedpedia.weedgenomics.org/).
456
457 **Acknowledgement**
458
459 This work was partially supported by USDA-NIFA Hatch project COL00783 at Colorado State
460 University.

461 **References**

462 1. Lisch, D. How important are transposons for plant evolution? *Nature Reviews Genetics* vol. 14
463 49–61 (2013).

464 2. Finnegan, D. J. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**,
465 103–107 (1989).

466 3. Lisch, D. Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.* **60**,
467 43–66 (2009).

468 4. Burgess, D., Li, H., Zhao, M., Kim, S. Y. & Lisch, D. Silencing of mutator elements in maize
469 involves distinct populations of small RNAs and distinct patterns of DNA methylation.
470 *Genetics* **215**, 379–391 (2020).

471 5. Bennetzen, J. L. Transposable element contributions to plant gene and genome evolution.
472 *Plant Mol. Biol.* **42**, 251–269 (2000).

473 6. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate
474 recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**, 1075–1079
475 (2002).

476 7. Hsing, Y. I. *et al.* A rice gene activation/knockout mutant resource for high throughput
477 functional genomics. *Plant Mol. Biol.* **63**, 351–364 (2007).

478 8. Candela, H. & Hake, S. The art and design of genetic screens: Maize. *Nat. Rev. Genet.* **9**, 192–
479 203 (2008).

480 9. Juretic, N., Hoen, D. R., Huynh, M. L., Harrison, P. M. & Bureau, T. E. The evolutionary fate
481 of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* **15**, 1292–1297
482 (2005).

483 10. Dupeyron, M., Singh, K. S., Bass, C. & Hayward, A. Evolution of Mutator transposable
484 elements across eukaryotic diversity. *Mob. DNA* **10**, (2019).

485 11. Lisch, D. Mutator and MULE Transposons . *Microbiol. Spectr.* **3**, (2015).

486 12. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev.*
487 *Genet.* **54**, 539–561 (2020).

488 13. Lisch, D. R., Freeling, M., Langham, R. J. & Choy, M. Y. Mutator transposase is widespread
489 in the grasses. *Plant Physiol.* **125**, 1293–1303 (2001).

490 14. Xu, Z. *et al.* Jittery, a Mutator distant relative with a paradoxical mobile behavior: Excision
491 without reinsertion. *Plant Cell* **16**, 1105–1114 (2004).

492 15. Kumar, V., Jha, P., Jugulam, M., Yadav, R. & Stahlman, P. W. Herbicide-resistant kochia
493 (Bassia scoparia) in North America: A review. *Weed Sci.* **67**, 4–15 (2019).

494 16. Friesen, L. F., Beckie, H. J., Warwick, S. I. & Van Acker, R. C. The biology of Canadian
495 weeds. 138. Kochia scoparia (l.) schrad. *Can. J. Plant Sci.* **89**, 141–167 (2009).

496 17. Patterson, E. L. *et al.* The Draft Genome of Kochia scoparia and the Mechanism of Glyphosate
497 Resistance via Transposon-Mediated EPSPS Tandem Gene Duplication. *Genome Biol. Evol.*
498 **11**, 2927–2940 (2019).

499 18. Wiersma, A. T. *et al.* Gene amplification of 5-enol-pyruvylshikimate-3-phosphate synthase in
500 glyphosate-resistant Kochia scoparia. *Planta* **241**, 463–474 (2015).

501 19. Jugulam, M. *et al.* Tandem amplification of a chromosomal segment harboring 5-
502 Enolpyruvylshikimate-3-Phosphate synthase locus confers glyphosate resistance in Kochia
503 scoparia. *Plant Physiol.* **166**, 1200–1207 (2014).

504 20. Grossman, L. *The magician king*. (2012).

505 21. Grossman, L. *The Magician's Land*. (2015).

506 22. Preston, C., Belles, D. S., Westra, P. H., Nissen, S. J. & Ward, S. M. Inheritance of Resistance
507 to The Auxinic Herbicide Dicamba in Kochia (Kochia scoparia). *Weed Sci.* **57**, 43–47 (2009).

508 23. Pettinga, D. J. *et al.* Increased chalcone synthase (CHS) expression is associated with dicamba
509 resistance in Kochia scoparia. *Pest Manag. Sci.* **74**, 2306–2315 (2018).

510 24. Busi, R. *et al.* Weed resistance to synthetic auxin herbicides. *Pest Manag. Sci.* **74**, 2265–2276
511 (2018).

512 25. Westra, E. P., Nissen, S. J., Getts, T. J., Westra, P. & Gaines, T. A. Survey reveals frequency
513 of multiple resistance to glyphosate and dicamba in kochia (Bassia scoparia). *Weed Technol.*
514 **33**, 664–672 (2019).

515 26. Gaines, T. A. *et al.* EPSPS Gene Copy Number and Whole-Plant Glyphosate Resistance Level
516 in Kochia scoparia. *PLoS One* **11**, 1–11 (2016).

517 27. Ravet, K. *et al.* Genomic-based epidemiology reveals independent origins and gene flow of
518 glyphosate resistance in Bassia scoparia populations across North America. *Mol. Ecol.* **30**,
519 5343–5359 (2021).

520 28. Jinlu, L., Shuo, W., Jing, Y., Ling, W. & Shiliang, Z. A Modified CTAB Protocol for Plant
521 DNA Extraction. *Chinese Bull. Bot.* **48**, 72–78 (2013).

522 29. Doyle J J & Doyle J L. A rapid DNA isolation procedure for small quantities of fresh leaf
523 tissue. *Phytochem. Bull.* **19**, 11–15 (1987).

524 30. Cheng, Y. *et al.* Studies on genome size estimation, chromosome number, gametophyte
525 development and plant morphology of salt-tolerant halophyte Suaeda salsa. *BMC Plant Biol.*
526 **19**, (2019).

527 31. Xiao, C. Le *et al.* MECAT: Fast mapping, error correction, and de novo assembly for single-
528 molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).

529 32. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment
530 and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

531 33. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and
532 genome assembly improvement. *PLoS One* **9**, (2014).

533 34. Huang, S., Kang, M. & Xu, A. HaploMerger2: Rebuilding both haploid sub-assemblies from
534 high-heterozygosity diploid genome assembly. *Bioinformatics* **33**, 2577–2579 (2017).

535 35. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C. S. Scaffolding of long read assemblies
536 using long range contact information. *BMC Genomics* **18**, (2017).

537 36. Staňková, H. *et al.* BioNano genome mapping of individual chromosomes supports physical
538 mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **14**, 1523–
539 1531 (2016).

540 37. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-
541 Read Sequencing Technology. *PLoS One* **7**, (2012).

542 38. Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nat. Methods*
543 **14**, 679–685 (2017).

544 39. Humann, J. L., Lee, T., Ficklin, S. & Main, D. Structural and functional annotation of
545 eukaryotic genomes with GenSAS. *Methods Mol. Biol.* **1962**, 29–51 (2019).

546 40. Tempel, S. Using and understanding repeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).

547 41. Smit, A. & Hubley, R. RepeatModeler open-1.0. *Available at http://www.repeatmasker.org*
548 2008–2015 (2015).

549 42. Cock, P. J. A., Chilton, J. M., Grüning, B., Johnson, J. E. & Soranzo, N. NCBI BLAST+
550 integrated into Galaxy. *Gigascience* **4**, (2015).

551 43. D, K. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions,
552 deletions, and gene fusions. *Genome Biol.* **14**, R36 (2013).

553 44. James, K. W. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

554 45. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and
555 the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, (2008).

556 46. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: A web server for gene
557 finding in eukaryotes. *Nucleic Acids Res.* **32**, (2004).

558 47. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1:
559 Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS.
560 *Bioinformatics* **32**, 767–769 (2016).

561 48. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification
562 in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506
563 (2005).

564 49. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in
565 novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.*
566 **18**, 1979–1990 (2008).

567 50. Cock, P. J. A., Grüning, B. A., Paszkiewicz, K. & Pritchard, L. Galaxy tools and workflows
568 for sequence analysis with applications in molecular plant pathology. *PeerJ* **2013**, (2013).

569 51. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
570 analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

571 52. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: Comprehensive
572 and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

573 53. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and
574 phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

575 54. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative
576 genomics. *Genome Biol.* **20**, (2019).

577 55. Sims, D. *et al.* CGAT: Computational genomics analysis toolkit. *Bioinformatics* **30**, 1290–
578 1291 (2014).

579 56. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: A computational tool for the
580 study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).

581 57. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of
582 Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

583 58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
584 2079 (2009).

585 59. Anton, B. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell
586 sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

587 60. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover,
588 genotype, and characterize typical and atypical CNVs from family and population genome
589 sequencing. *Genome Res.* **21**, 974–984 (2011).

590 61. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for
591 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140
592 (2009).

593 62. Patterson, E. L., Pettinga, D. J., Ravet, K., Neve, P. & Gaines, T. A. Glyphosate resistance and
594 EPSPS gene duplication: Convergent evolution in multiple plant species. *J. Hered.* **109**, 117–
595 125 (2018).

596 63. Gaines, T. A., Patterson, E. L. & Neve, P. Molecular mechanisms of adaptive evolution
597 revealed by global selection for glyphosate resistance. *New Phytol.* **223**, 1770–1775 (2019).

598 64. Gaines, T. A. *et al.* Gene amplification confers glyphosate resistance in Amaranthus palmeri.
599 *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1029–1034 (2010).

600 65. Martin, S. L. *et al.* High gene flow maintains genetic diversity following selection for high
601 EPSPS copy number in the weed kochia (Amaranthaceae). *Sci. Rep.* **10**, (2020).

602 66. Hudson, M. E., Lisch, D. R. & Quail, P. H. The FHY3 and FAR1 genes encode transposase-
603 related proteins involved in regulation of gene expression by the phytochrome A-signaling
604 pathway. *Plant J.* **34**, 453–471 (2003).

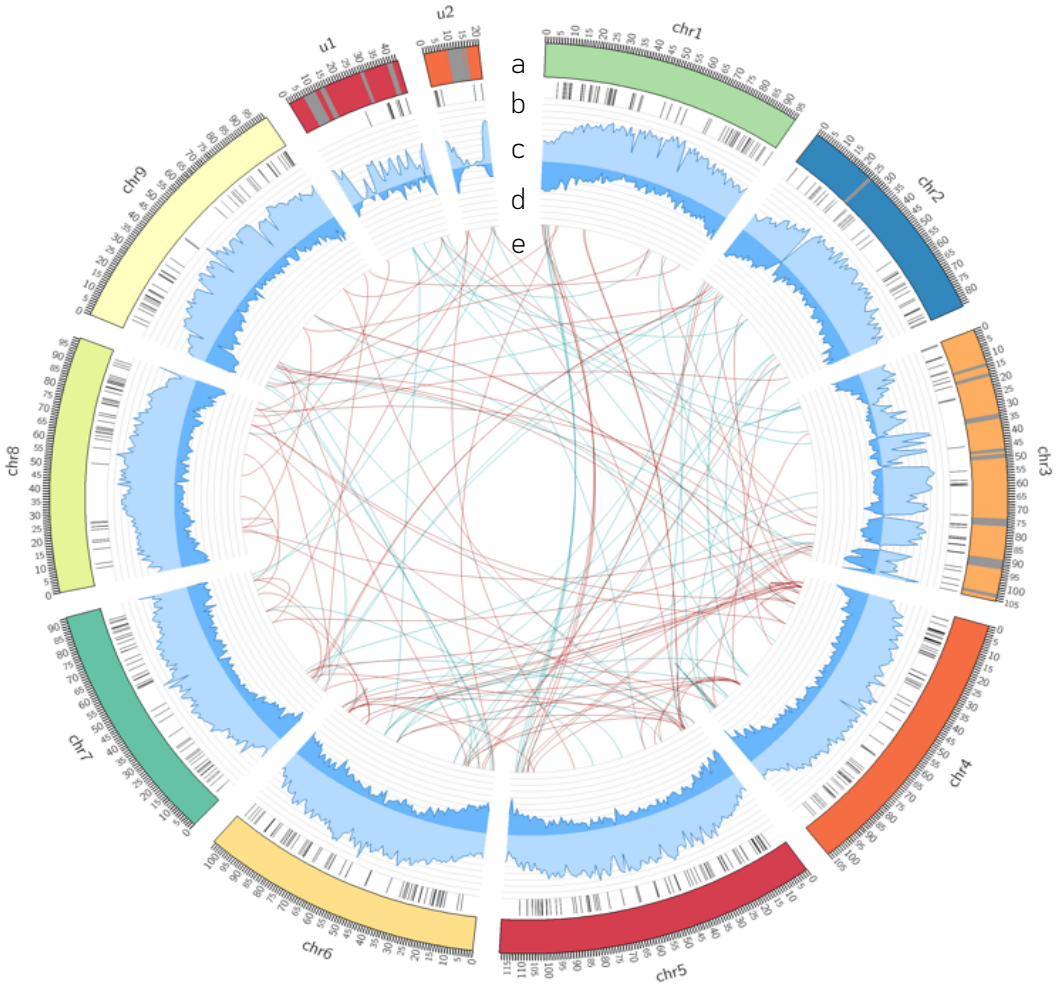
605 67. Hardigan, M., Crisovan, E., Hamilton, J. & Kim, J. Genome reduction uncovers a large
606 dispensable genome and adaptive role for copy number variation in asexually propagated
607 Solanum tuberosum. *The Plant* (2016).

608 68. Fuentes, R. R. *et al.* Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880
609 (2019).

610 69. Horváth, V., Merenciano, M. & González, J. Revisiting the Relationship between
611 Transposable Elements and the Eukaryotic Stress Response. *Trends Genet.* **33**, 832–841
612 (2017).

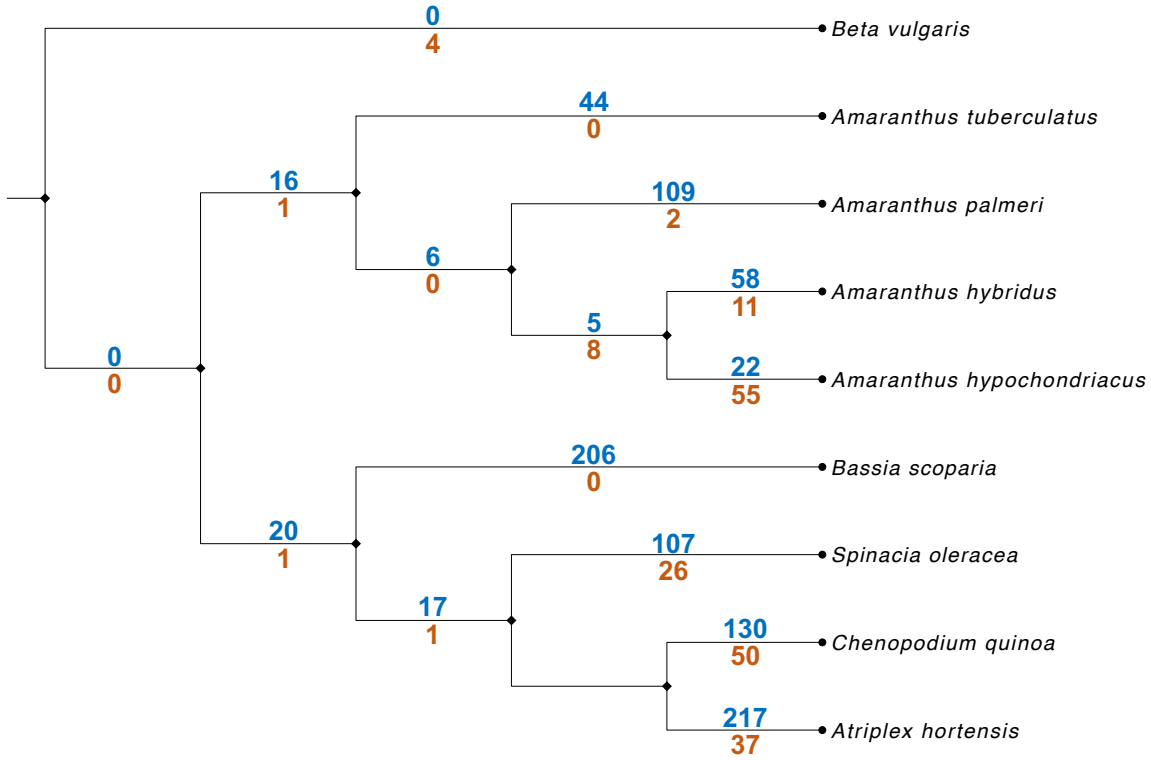
Figures

Figure 1: Circos plot of *Bassia scoparia* (kochia) genome assembly; a) chromosomes (chr1..chr9) and the two largest unplaced scaffolds (u1,u2) with gaps larger than 10 Mb indicated in gray; b) the placement of genes were part of a target set of genes containing some combination of the following Interpro designations *FHY1* (IPR037766), *FAR1* DNA binding domain (IPR004330), and *FHY3*/ *FAR1* family (IPR031052); c) repetitive sequence coverage per 1 Mb; d) gene coverage per 1 Mb; e) segmental duplications detected using MCScanX. Red lines indicate segmental duplications contains a target set gene. Green lines do not contain a target set gene.

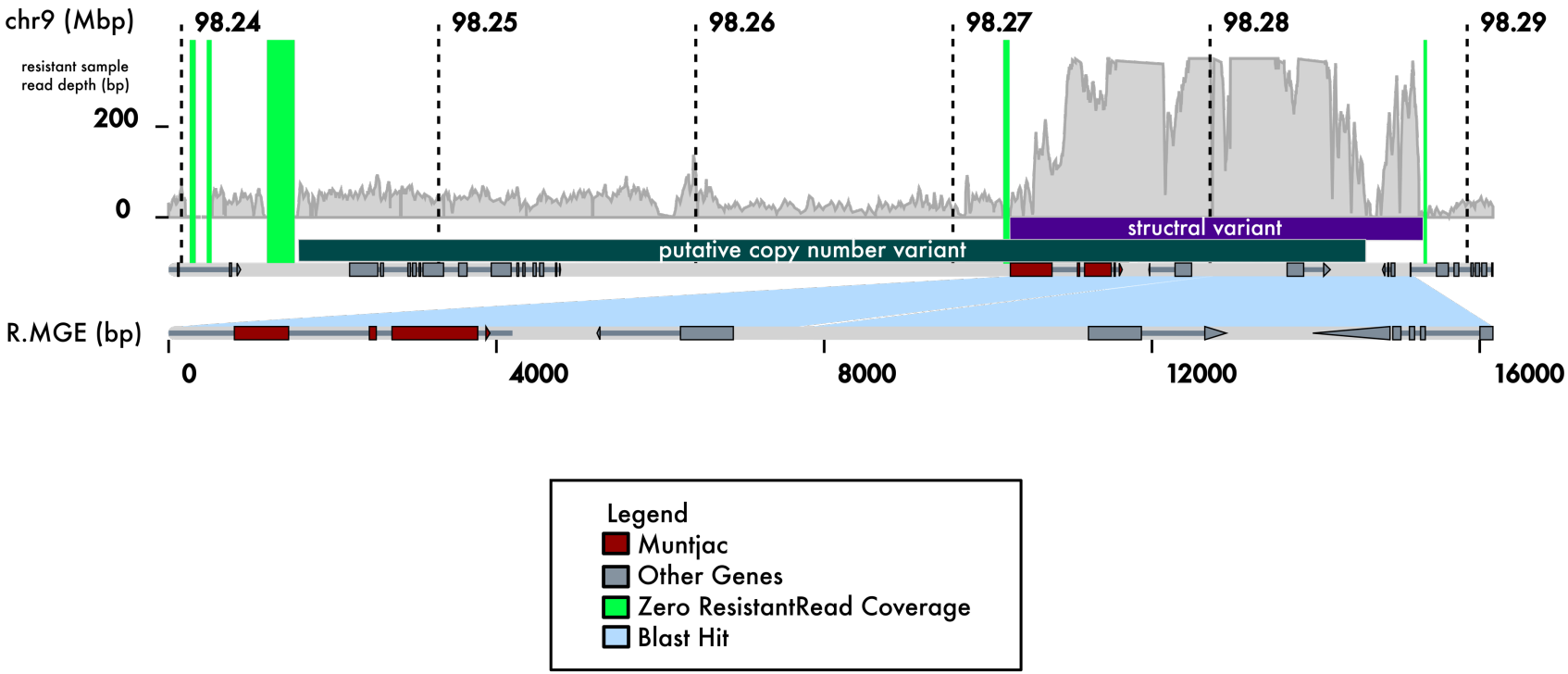


Legend
a: Karyogram
Gap 10Mb or larger
b: Target Gene
Gene containing FAR1, Fhy3, or FAR1 binding domain
c: Repetitive Sequence Coverage
Repetitive sequence coverage per 1Mb
d: Gene Coverage
Gene coverage per 1 Mb
e: Segmental Duplications
Contains FAR1 genes
Does not contain target genes

Figure 2: A cladogram of *Bassia scoparia* and several other *Amaranthaceae* species that have sequenced genomes. Significant gene family expansions for each clade are listed above the branch in blue while significant gene family contractions are reported below the branch in orange as calculated by Computational Analysis of gene Family Evolution (Cafe).



628 Figure 3: Genomic DNA Illumina reads from the resistant *Bassia scoparia* (kochia) line (M32) mapped to the susceptible genome. Putative copy number
629 variants predicted using CNVnator. Structural variants predicted using complete breaks in read mapping, and micro assembly.



630

Figure 4: Micro assemblies from the glyphosate resistant *Bassia scoparia* (kochia) line (M32) were assembled for two Muntjac insertions in the susceptible line (7710). a) A completely intact Muntjac insertion on Chr9, with detected target site duplication (TSD), and detectable terminal inverted repeats (TIR). b) A fragmented insertion on Chr1. The Muntjac gene has been pseudogenized or incorporated into another a gene. The TSD provides evidence of insertion, but the fragmentation of Muntjac and the absence of TIRs suggest that this insertion on Chr1 is no longer active.

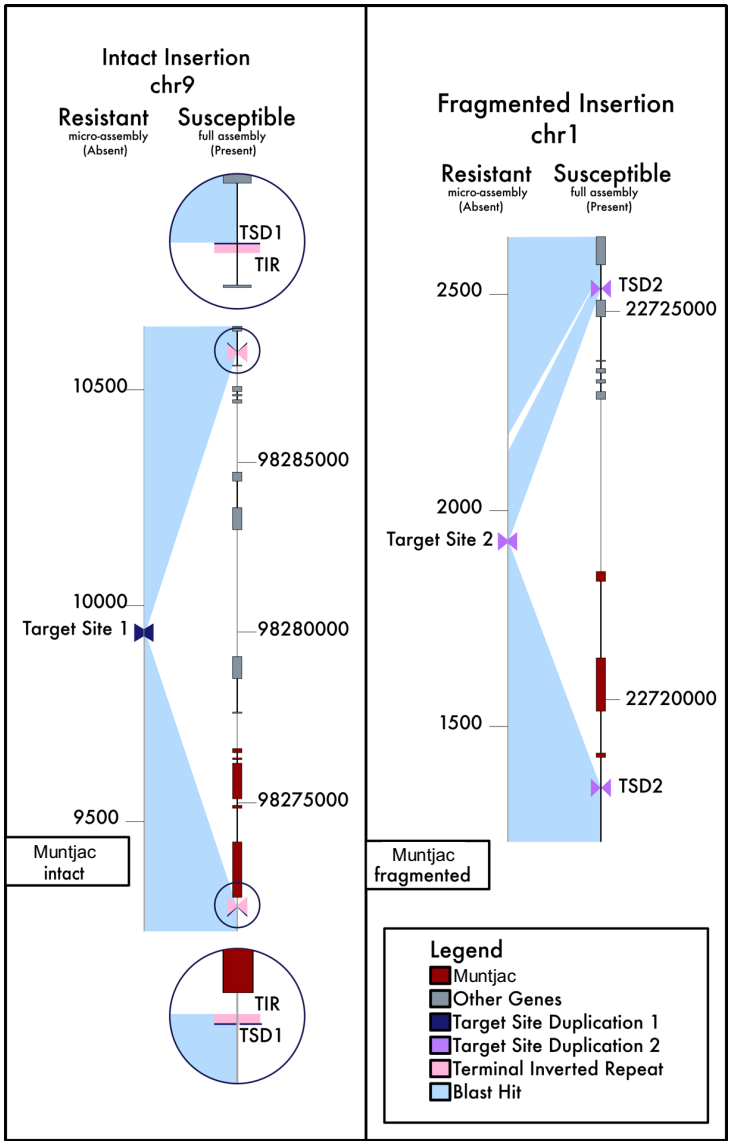


Figure 5: A reduced gene tree produced from an alignment of representative MuDR sequences searched against representative plant clades showing the representative domains within the MuDR clade. *FHY3*/ *FAR1* domain was detected in all sequences but is not shown for clarity.

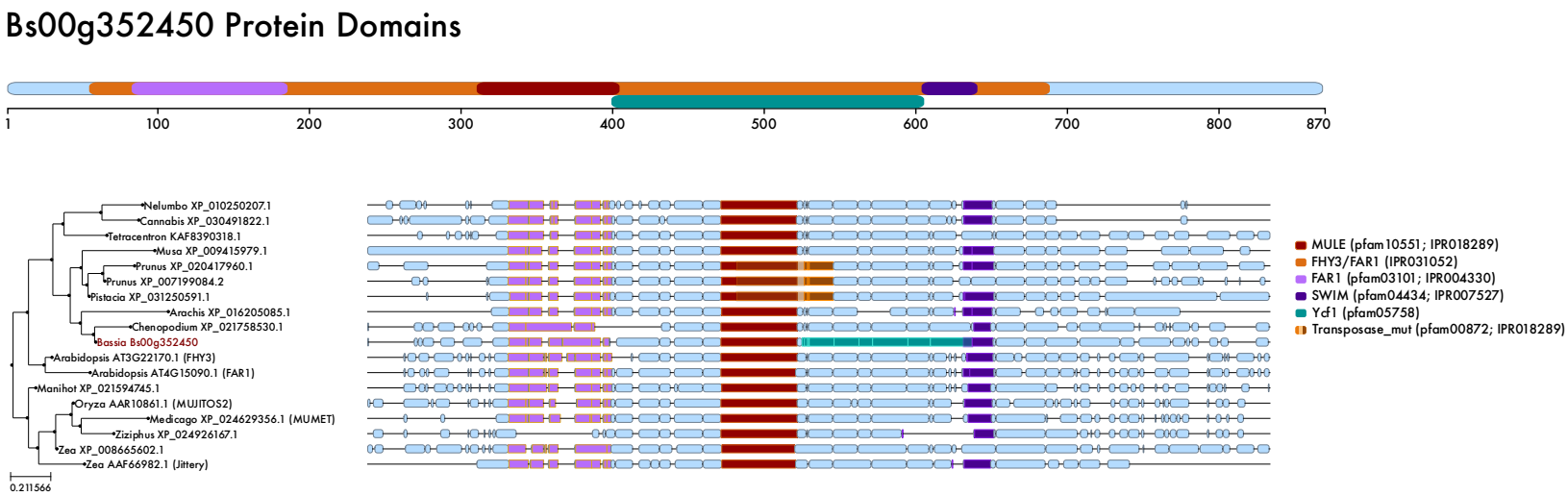


Figure 6: Simplified gene tree showing relationship of named MuDRs, *FHY3*, *FAR1*, and Muntjac. Muntjac clade is sister to *FHY3/FAR1* clade. The Muntjac clade diverged from *FHY3/FAR1* at the base of angiosperms, but it is more closely related to *FHY3/FAR1* than Jittery. Gene tree was produced taking representative MuDR sequences and searching against representative plant clades. Alignment was produced with COBALT, trimmed with GGBlocks and phylogeny estimated using IQTree.

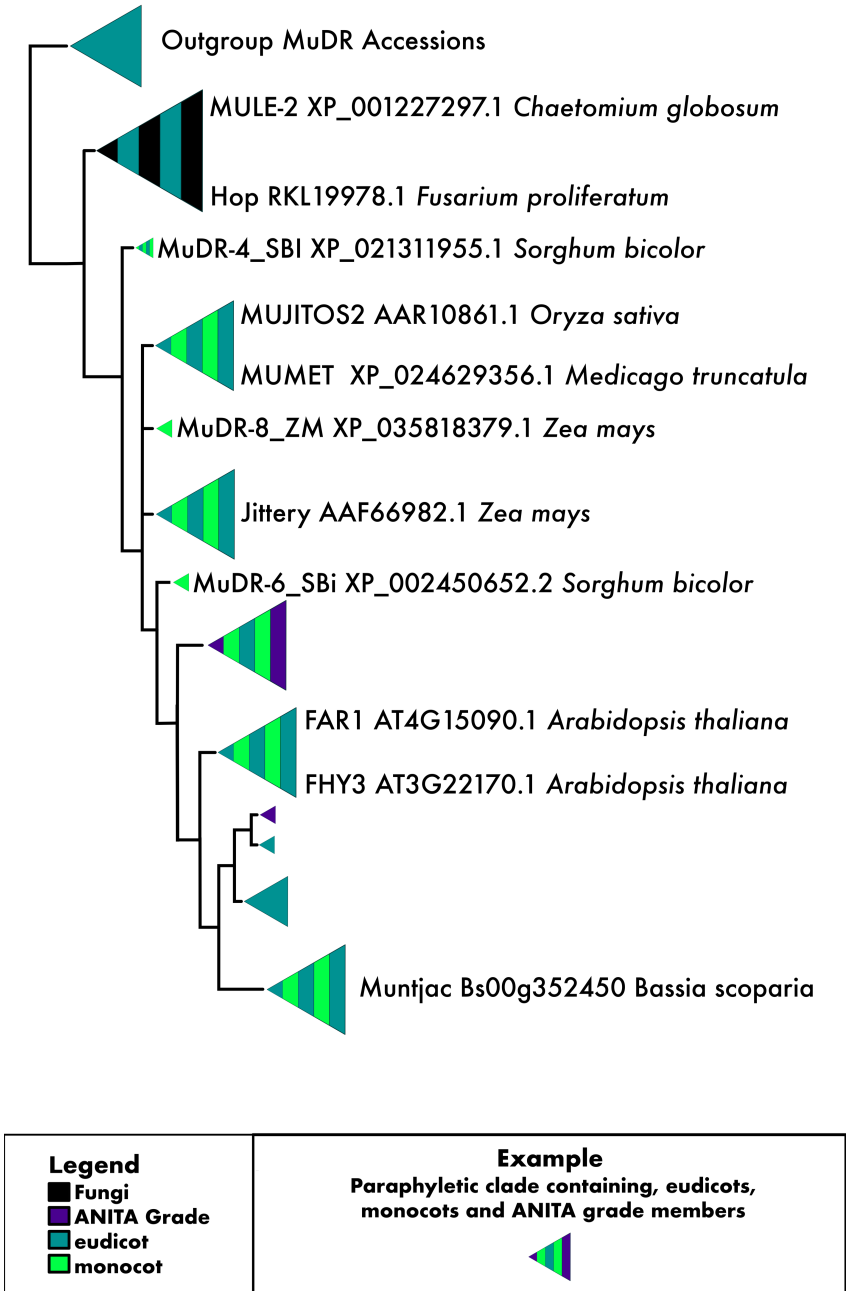
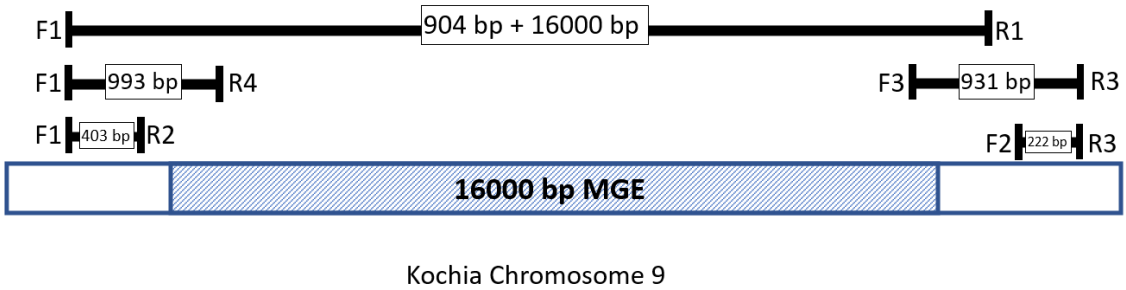


Figure 7: Diagram of the primer positions used to verify the presence/absence of MGE sequence on Chr9 in glyphosate resistant and susceptible *Bassia scoparia* (kochia) individuals.



655 **Tables**

656 Table 1: BUSCO scores for the kochia transcriptome using the Embryophyta_ODB 10 database.

657

	Number of BUSCOs	Percentage of BUSCOs
Total BUSCO groups searched	1,614	
Complete BUSCOs (C)	1,505	93.3%
Complete and single-copy BUSCOs (S)	1,468	91.0%
Complete and duplicated BUSCOs (D)	37	2.3%
Fragmented BUSCOs (F)	27	1.7%
Missing BUSCOs (M)	82	5.0%

658

Table 2: A summary of the output from Computational Analysis of gene Family Evolution (Cafe) indicating where in the evolution of the Bassia scoparia gene family expansions and contractions occurred. Terminal branches are labeled with the species while internal branches are bolded and species included in those branches are made explicit using a 6 character Genus:species ID.

Branch Label	Expansions	Genes Gained	Contractions	Genes Lost	Average Expansion	Significant Expansions	Significant Contractions
Beta vulgaris	16	19	25,664	27,667	-1.008	0	4
AmaHyp, AmaHyb, AmaPal, AmaTub, CheQui, AtrHor, SpiOle, BasSco	142	175	0	0	0.006	0	0
All Amaranthus Sp.	656	794	8,019	8,050	-0.265	16	1
Amaranthus tuberculatus	3,750	6,811	5,063	5,235	0.057	44	0
AmaHyp, AmaHyb, AmaPal	462	550	993	1,010	-0.017	6	0
Amaranthus palmeri	3,989	8,704	5,312	5,480	0.118	109	2
AmaHyp, AmaHyb	926	1,029	1,960	2,111	-0.039	5	8
Amaranthus hybridus	1,423	2,588	3,385	3,560	-0.035	58	11
Amaranthus hypochondriacus	2,059	2,638	2,082	2,484	0.006	22	55
CheQui, AtrHor, SpiOle, BasSco	928	1,231	4,161	4,181	-0.108	20	1
Bassia scoparia	5,331	22,938	6,405	6,541	0.598	206	0
CheQui, AtrHor, SpiOle	837	1,081	3,002	3,147	-0.075	17	1
Spinacia oleracea	3,334	10,766	4,605	5,272	0.200	107	26
CheQui, AtrHor	1,857	3,344	2,399	2,513	0.030	110	6
Chenopodium quinoa	11,787	18,049	2,284	2,907	0.552	130	50
Atriplex hortensis	2,170	8,704	4,121	4,713	0.146	217	37

664 Table 3 Information of kochia MGE validation primers

Assay	Primer name	Primer sequence (5' to 3')	Target length (bp)	Annealing temperature (°C)	Function
1	F1 R1	CATCTTCTCTTCTCAACCACTC ACCACACCCACTCATTCAC	904	55	Test for the presence of MGE
2	F1 R2	CATCTTCTCTTCTCAACCACTC CGGAAAACAACACTTAGAATGC	403		Amplify upstream of MGE insertion site
3	F2 R3	GTGAATGAGTGGGTGTGGT TGAACAACTTAGCAGCAAGG	222		Amplify downstream of MGE insertion site
4	F1 R4	CATCTTCTCTTCTCAACCACTC CTACAGTCAACAAAACCAAACC	993		Amplify the upstream junction of MGE insertion
5	F3 R3	ACCCAAATTCTTCAACAGCAC TGAACAACTTAGCAGCAAGG	931		Amplify the downstream junction of MGE insertion

665

666 Table 4 Results of MGE validation primer pairs.

		Populations									
Assay	Primer pair	Oregon		CO-S		Israel		M32		7710 (S mixture)	
		Oregon-1	Oregon-2	CO-S-1	CO-S-2	Israel-1	Israel-2	M32-1	M32-2	S-1	S-2
1	F1/R1	×	✓	×	✓	✓	✓	✓	✓	×	×
2	F1/R2	✓	× *	✓	✓	✓	✓	✓	✓	✓	✓
3	F2/R3	✓	× *	✓	✓	✓	✓	✓	✓	✓	✓
4	F1/R4	×	×	×	×	×	×	×	×	✓	✓
5	F3/R3	×	×	×	×	×	×	×	×	✓	✓

667 ✓: band of/close to target length amplified

668 ×: band of/close to target length not amplified

669 * Multiple bands amplified