RESOLUTION OF SEX CHROMOSOMES IN WATERHEMP AND PALMER AMARANTH

BY

DAMILOLA ALEX RAIYEMO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Crop Sciences
in the Graduate College of the
University of Illinois Urbana-Champaign, 2024

Urbana, Illinois

Doctoral Committee:

Professor Patrick J. Tranel, Chair
Professor Aaron G. Hager
Associate Professor Anthony J. Studer
Assistant Professor Steven J. Burgess

ABSTRACT

The genus *Amaranthus* L. includes nine dioecious species: *Amaranthus acanthochiton* J.D. Sauer, *Amaranthus arenicola* I.M. Johnson, *Amaranthus australis* (A. Gray) J.D. Sauer, *Amaranthus cannabinus* (L.) J.D. Sauer, *Amaranthus floridanus* (S. Watson) J.D. Sauer, *Amaranthus tuberculatus* (Moq.) J.D. Sauer, *Amaranthus greggii* S. Watson, *Amaranthus watsonii* Standley, and *Amaranthus palmeri* S. Watson that are native to North America and grouped into the subgenus *Acnida* (L.) Aellen ex K.R. Robertson. Two of the dioecious species, *A. tuberculatus* (waterhemp) and *A. palmeri* (Palmer amaranth), are agronomically important weeds that have evolved resistance to herbicides from several modes of action and have spread beyond their native ranges. A strategy that has been proposed for the management of these weedy dioecious species to complement existing tools (such as herbicide technology, crop rotation, seedbank depletion etc.) is a genetic control, whereby sex ratios could be biased towards one gender (e.g., males) and the genetic factors involved are inherited in a non-Mendelian pattern via a gene drive system. The depletion of the other gender (i.e., females) required for outcrossing over multiple generations thus results in population collapse. Adopting such a genetic control strategy however requires a comprehensive understanding of the factors and mechanisms responsible for sex determination as well as evolutionary relationships among the dioecious species. While previous studies confirmed males were the heterogametic sex and identified several candidate genes within putative sex-determining regions, as well as differentially expressed genes between males and females with likely role in sex determination, the contiguity of these sex-determining regions and the genomic architecture of the chromosome (colloquially referred to as the "sex chromosome") that harbors the sex determinants remain poorly understood.

Chapter 1 includes a brief overview of genome sequencing and assembly of sex chromosomes; evolution and mechanisms of sex determination in plants; and the genetics of dioecy evolution within the *Amaranthus* genus. Chapter 2 explores the relationships among the dioecious amaranths using Mash distances, as well as the conservation of candidate genes previously identified within the *A. palmeri* and *A. tuberculatus* sex-determining regions in other dioecious amaranths. The results of this study confirmed Sauer's taxonomic ordering of the dioecious amaranths, which was based on comparative morphology, and indicated two possible independent origins of dioecy evolution within the *Amaranthus* genus. Chapter 3 further explores the relationships among the dioecious amaranths utilizing their complete chloroplast genomes. Although some species relationships were consistent with the previous Mash tree, the chloroplast phylogeny indicates that the relationships of the *A. australis + A. cannabinus* lineage to the other dioecious species remains unclear and showed that *A. palmeri* and *A. watsonii* are more genetically related than previously reported. We also provide a framework for investigating evolutionary relationships among the amaranths and demonstrate that the use of complementary phylogenetic approaches coupled with proper species identification could be very informative in examining the complex evolutionary history of the genus. Chapter 4 focuses on the sequencing, assembly, phasing, and annotation of the *A. tuberculatus* genome as well as identification of the sex chromosomes within the assembly. We present a chromosome-level haplotype-resolved genome of *A. tuberculatus* and report a contiguous ~32.8 Mb region near the middle of chromosome 1 as the sex-determining region. We show that the sex chromosome in *A. tuberculatus* likely originated from the fusion of two ancestral chromosomes. Chapter 5 describes the sequencing, assembly, and annotation of *A. palmeri*, *A. retroflexus*, and *A. hybridus* genomes, and the identification of the sex chromosomes in the *A. palmeri* assembly. Here, we

present a chromosome-level phased genome of three *Amaranthus* species (*A. palmeri*, *A. retroflexus*, and *A. hybridus*) and report a contiguous ~2.84 Mb region at the distal end of chromosome 3 of *A. palmeri* as the sex-determining region. Finally, Chapter 6 provides concluding remarks on sex determination within the *Amaranthus* genus and future directions on finding the candidates genes and mechanisms involved.

ACKNOWLEDGMENTS

I express my sincere gratitude to my advisor, Dr. Patrick Tranel, for providing me with the opportunity to obtain my Ph.D. degree through his lab. The depth of his intellect and critical evaluation of my research were valuable throughout my program. I also thank my committee members: Dr. Aaron Hager, Dr. Anthony Studer, and Dr. Steven Burgess for their guidance and expertise. I am grateful for the support and friendship from past and present members of Dr. Tranel's lab, including Dr. Yousoon Baek, Dr. Brent Murphy, Dr. Lucas Bobadilla, Jacob Montgomery, Isabel Werle, Alexander Lopez, Vanessa Soliz, and Filipi Machado. Thank you for the conversations and insights that have helped me become a better scientist.

I also thank my parents; Theophilus and Victoria Raiyemo, and my siblings; Funmilola and Eniola, for their love and support throughout my time at the University of Illinois. Finally, I dedicate this work to all members of the Raiyemo family around the world.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 GENOME SEQUENCING, ASSEMBLY, AND IDENTIFICATION OF SEX CHROMOSOMES

In the early 2000s, DNA sequencing adopted the Sanger dideoxy synthesis or the Maxam-Gilbert chemical sequencing methods that were highly laborious and expensive (Slatko et al. 2018; Hu et al. 2021). The sequencing of the model plant *Arabidopis thaliana* cost about $100 million despite the small genome size of ~140 Mb (Kersey 2019), and The Human Genome Project (HGP) cost about $3 billion (Shendure et al. 2017). The introduction of short-read or next-generation sequencing (NGS) brought a reduction in sequencing cost and offered more capabilities and higher throughput. However, *de novo* assembly of sequenced genomes and discovery of long structural variants were practically impossible due to the limitation imposed by the short read lengths of the NGS platforms, which could not span long repetitive or other complex regions of the genome (e.g., centromeric or AT-rich regions) (Goodwin et al. 2016).

To assemble a genome using only short reads, DNA would be extracted from the tissue of interest (e.g., leaf tissues), library preparation and size selection carried out, and sequenced on an Illumina platform. Adapters and low-quality bases are then removed from the single- or paired-end reads as part of quality control and the cleaned reads run through a *de bruijn* graph assembler [e.g., ABySS (Jackman et al. 2017), SOAPdenovo (Luo et al. 2012), Velvet (Zerbino 2010), or SPAdes (Bankevich et al. 2012)]. The *de novo* assembly was computationally complex, had a long run-time, and was characterized by errors or gaps around repeat regions and with heterozygous regions split into multiple contigs, yielding an assembly that was fragmented, incomplete, redundant, and erroneous (Fan and Li 2012; Amarasinghe et al. 2020). Further approaches to improve assemblies were developed, including reads with longer inserts (e.g.,

Illumina mate pairs ranging from 2 – 10 Kb) that could be combined with the short reads and assembled using a hybrid assembler [e.g., MaSuRCA (Zimin et al. 2013)].

The launching of Pacific Biosciences (PacBio) RS system in 2011 and subsequently the PacBio Sequel system and Oxford Nanopore (ONT) in 2014 heralded a new era of improved assembly contiguity whereby the single-molecule real-time (SMRT) sequencing technologies could produce long or ultra-long reads of 10 Kb – 2 Mb (Jain et al. 2018). A limitation of these technologies however were high error rates (>10%), and the need for high molecular weight DNA. When an organism is sequenced on the PacBio RS system, the early Sequel system, or Oxford NanoPore and assembled [e.g., with Canu (Koren et al. 2017) or Wtdbg2 (Ruan and Li 2020)], additional steps of polishing the assembly [e.g., with Arrow and Pilon (Walker et al. 2014)] for error correction, fixing mis-assemblies, and gap filling using the more accurate Illumina short reads are required. Alternatively, the contiguity of a genome that was previously assembled using only short-reads sequence could be improved by scaffolding with the PacBio long reads [e.g., with SSPACE (Boetzer et al. 2011)].

Despite the sequencing and computational advances, accuracy (as measured by the three Cs – correctness, completeness and contiguity) in genome assemblies still differed among organisms as some stretches of the genome or genomic regions are by nature more difficult to sequence (https://www.pacb.com/blog/understanding-accuracy-in-dna-sequencing/) e.g., sex chromosomes, centromeres, and telomeres with abundant repetitive sequences, AT/GC-rich regions, and palindromic sequences. In addition, how to accurately obtain haplotypes i.e., separating (phasing) the maternally and paternally inherited copies of each chromosome for diploid or polyploid organisms was a challenge. The variation between haplotypes was often collapsed into a single mixed sequence for unphased assemblies (Koren et al. 2018). In some

scenarios where the traditional PacBio long reads or the continuous long reads (CLR) are assembled and phased, the diploid-aware genome assembler (e.g., FALCON-Unzip or Canu) begins by first creating a single mixed or collapsed assembly, and then uses heterozygous single nucleotide polymorphisms (SNPs) and structural variants (SVs) to partition the reads by haplotype, and reassembles them into haplotigs, resulting in primary contigs and associate contigs (Koren et al. 2018; Kronenberg et al. 2019; Zhang et al. 2020).

In 2020, PacBio introduced the Sequel Ile platform, which was based on circular consensus sequencing (CCS) (Wenger et al. 2019), as opposed to the CLR sequencing that was used in its previous platforms (Eid et al. 2009). The high-fidelity (HiFi) reads from this new technology had read accuracy >99%, compared to the ~90% accuracy of the traditional long reads, and comparable to >99% accuracy of Illumina short reads. PacBio HiFi long reads circumvented the haplotype-phasing limitations of short reads sequencing and even the traditional error-prone long reads. In addition, typical PacBio HiFi libraries for human whole-genome sequencing (WGS) are 18 – 20 Kb, much longer than the ~300 bp Illumina short reads. This greater contiguity thus improves the results in the calling of variants, haplotype phasing, coverage uniformity, and *de novo* assemblies. Furthermore, phased genomes have been shown to provide allelic information than single mixed genomes, which could be valuable in understanding species evolution (Chin et al. 2016).

The integration of linked-read sequencing technology, also known as chromosome conformation capture mapping or Hi-C (Phase Genomics or Dovetail) and optical mapping (Bionano Genomics) with HiFi and ONT long reads has now become the state-of-the-art approach in assembling chromosome-scale genomes (Mascher et al. 2017; McCord et al. 2020; Kronenberg et al. 2021). This combination allows for the detection of large-scale structural

variation, validation of genome assembly completeness, and scaffolding of genomes as they span genomic regions that could be difficult to sequence (for reference, optical map lengths are ~225 Kb) (Yuan et al. 2020). Several recent genome sequencing projects, including phased chromosome-scale assemblies of human and other vertebrates (Garg et al. 2021; Cheng et al. 2022), apple (Khan et al. 2022), tetraploid potato (Sun et al. 2022), *Arabidopsis thaliana* (Wang et al. 2022), endemic species, *Bletilla striata* (Jiang et al. 2022) and telomere-to-telomere (T2T) gap-free maize genome assembly (Chen et al. 2023), have utilized this sequencing and assembly approach. Phased chromosome-scale assemblies now enable the accurate identification of sex chromosomes in dioecious species and the elucidation of their organization and evolution. Using a newly assembled chromosome-level genome of *Salix dunnii*, He et al. (2021) was able to identify a contiguous 3.21 Mb sex-determining region (SDR), and also confirmed male-heterogamety in the species despite the homomorphic (i.e., indistinguishable) nature of sex chromosomes in *Salix*. In fact, their finding of male-heterogamety is contrary to the well-known female-heterogamety in several other species of *Salix*, thus attesting to the importance of high-quality chromosome assembly in sex chromosome studies. Similarly, Darolti et al. (2022) revealed that the most substantial source of variation in the results of different studies on sex chromosome divergence in guppy (*Poecillia* spp.) could be attributed to the choice of reference genome used. With high-quality phased chromosome-level genomes, QTL mapping, genome-wide association studies (GWAS), and *k*-mer approaches can be used to identify which of the assembled chromosomes are the sex chromosomes (Akagi et al. 2014; Müller et al. 2020; Carey et al. 2022; Ma et al. 2022).

1.2 EVOLUTION AND MECHANISMS OF SEX DETERMINATION IN PLANTS

The question of how sedentary organisms have separate male and female reproductive systems while it appears they have evolved from an ancestral hermaphroditic (or bisexual) reproductive system has captivated many researchers (Renner and Müller 2022). This phenomenon of distinct male and female plants, termed 'dioecy' represents only 5-6% of the total flowering plant species (Renner 2014), and confers evolutionary advantages by enforcing outcrossing; thus enhancing high genetic diversity and adaptation rates (Käfer et al. 2017; Muyle et al. 2021).

One model of dioecy evolution (known as the two-factor model) proposed by Charlesworth and Charlesworth (1978) posits that a recessive male-sterility mutation (feminizing) occurs first in a population, followed by a dominant female-sterility mutation (masculinizing). The two mutations then become chromosomally linked within a region that undergoes recombination suppression. The two-factor model assumes that the development of unisexual flowers and dioecy evolved from hermaphroditism through a gynodioecious pathway, and also is more common in insect-pollinated species (Charlesworth and Charlesworth 1978; Standley 1985). It has been suggested that the model does not put into consideration the development of unisexual flowers in monoecious species or when dioecy evolves from monoecy (Renner and Ricklefs 1995; Renner and Won 2001), or the roles of environmental stress or hormones on sex determination, i.e., environmentally induced sexual plasticity (Golenberg and West 2013; Renner and Müller 2021).

Another model is the monoecy-paradioecy-dioecy pathway (Lloyd 1980; Standley 1985; Renner and Won 2001). A paradioecy pathway begins with monoecy and proceeds by a divergence of the sexes in the ratio of male to female flowers, which is made possible through

the regulation of a developmental switch determining the differentiation of female or male flowers (Golenberg and West 2013). This regulation of gene expression leading to sexual segregation could result from external environmental or internal physiological cues. The environmentally determined sexual development differs in expression patterns of regulatory regions or a network determining sex but it is still under genetic control (Golenberg and West 2013). Moreso, the regulatory states suppressing male and female functions already exist, and potentially are connected to plant hormones that could act as a toggle switch between female and male development. If a change in the regulatory system causes a disequilibrium in the male-to-female ratio, a genetic mutation promoting a switch to the opposite direction could be positively selected and fixed, resulting in the establishment of a single-factor sex-determination system (Henry et al. 2018).

The two factors or a single factor responsible for sex determination are located within a sex-determining region (SDR), which is located on the Y chromosome (species with an XY system e.g., *Populus tremula*) or on the W chromosome (species with a ZW system e.g., *Populus alba*). The sex chromosome harboring these sex determinants could be homomorphic (i.e., cytogenetically indistinguishable chromosomes) or heteromorphic. Notable plant species for the two sex-determining genes are kiwifruit and garden asparagus. In kiwifruit (*Actinidia deliciosa*), a male-specific type-C cytokinin response regulator (*SHY GIRL*, *SyGl*) that suppresses feminization, and a male-promoting tapetum regulator (*FRIENDLY BOY*, *FrBy*) that aids in the development of androecia, are the two genes crucial for sex determination (Akagi et al. 2019). Similarly, two genes, a suppressor of female function (*SOFF*) and a MYB transcription factor expressed only in males (*DEFECTIVE IN TAPETUM DEVELOPMENT AND FUNCTION 1*,

TDF1), are involved in sex determination in garden asparagus (*Asparagus officinalis*) (Harkess et al. 2020).

Sex determination in diploid persimmon (*Diospyros lotus*) and poplar (*Populus* spp.) are however known to be controlled by a single gene. In *Diospyrus lotus*, the gene *OGI* present within the SDR represses the activity of an autosomal gene, *MeGI* (*MALE GROWTH INHIBITOR*), thereby resulting in the development of androecia (Akagi et al. 2014). Investigation of the regulatory networks involved in the *MEGI/OGI* mechanism suggests that a cytokinin signaling pathway could also be involved in female floral organ differentiation (Yang et al. 2019). In *Populus* spp., partial duplicate of *ARR17* (*ARABIDOPSIS RESPONSE REGULATOR 17*) within the SDR represses the activity of a Y-linked *ARR17* on chromosome 19 (although not within the SDR) via a small-RNA-mediated DNA methylation (Müller et al. 2020). Additional evidence indicated that while the *ARR17* had translocated or switched heterogamety in species of poplar, it is still being recruited in the sex determination pathway within the genus and thus has evolved independently a few times (Müller et al. 2020; Montalvão Leite et al. 2022). Spinach is another species in which sex determination was recently proposed to be controlled by a single gene, *NRT1/PTR6.4* (transporter of nitrate, peptide, or hormones). The *NRT1/PTR6.4* gene was hypothesized to utilize two pathways for carpel development suppression and stamen initiation. The regulation of sex determination by *NRT1/PTR6.4* involved jasmonic acid biosynthesis, gibberellic acid signaling, and B-class genes (Ma et al. 2022).

Experimental validation of single-gene sex determination has also been confirmed by engineering monoecious species. In melon (*Cucumis melo*), a network of three genes controls sex expression (Boualem et al. 2008, 2015). *CmACS11* controls the development of pistillate

flowers (like *SILKLESS* in maize), *CmWIP1* suppresses female flower and stamen development (like *TASSEL SEED* in maize), and *CmACS7* represses male flower development. By selecting on natural variation to synthesize a population that is fixed for a null form of feminizing gene *CmACS11* and segregating for a functional *WIP1*, Boualem et al. (2008) engineered the transition from monoecy to dioecy. The same gene(s) controlling sex expression in monoecious species could thus be involved in sex determination in related dioecious species, e.g., observed in diploid persimmon (*Diospyros lotus*) (Renner and Müller 2021).

Although information on sex determination for flowering plants is still scanty and many species studied so far adopt varying mechanisms for sex determination, evidence in the literature points to some similarities in pathways or mechanisms utilized in sex determination or differentiation (e.g., cytokinin, JA or other hormone biosynthesis pathway, tapetum development genes, pollen development, and fertility genes, or small-RNA-mediated DNA methylation) across different systems [reviewed in (Montalvão Leite et al. 2021)].

1.3 GENETICS OF DIOECY EVOLUTION WITHIN THE *AMARANTHUS* GENUS

The *Amaranthus* genus is made up of 70 - 80 species, nine of which are dioecious [*Amaranthus acanthochiton* J.D. Sauer, *Amaranthus arenicola* I.M. Johnson, *Amaranthus australis* (A. Gray) J.D. Sauer, *Amaranthus cannabinus* (L.) J.D. Sauer, *Amaranthus floridanus* (S. Watson) J.D. Sauer, *Amaranthus tuberculatus* (Moq.) J.D. Sauer, *Amaranthus greggii* S. Watson, *Amaranthus watsonii* Standley, and *Amaranthus palmeri* S. Watson] while the remaining species are monoecious (Sauer 1955, 1972; Bayón and Peláez 2012; Bayón 2015; Waselkov et al. 2018). Early works on dioecy in the amaranths began with Murray (1940), who observed via hybridization studies that males were the heterogametic sex (i.e., XY sex system). This finding was later confirmed using sex-specific markers developed through reduced

representation ddRAD sequencing (Montgomery et al. 2019). Murray (1940) hypothesized that the XY chromosome pair in dioecious *Amaranthus* species used in the studies carry differential sex factors and that the autosomes, if they carry sex factors, are homozygous. In addition, Murray observed a 1:1 sex ratio, and no abnormalities on pistillate plants and thus concluded that *A. tuberculatus* sexual states were extremely stable.

Further studies looking at sex chromosome structure in dioecious amaranths via cytology failed to identify heteromorphic sex chromosomes (Grant 1959), and thus it is believed that the dioecious amaranths have homomorphic sex chromosomes. The availability of draft genome assemblies for both *A. tuberculatus* and *A. palmeri* enabled the identification of male-specific region in both species (Montgomery et al. 2021). This region for *A. tuberculatus* spanned several contigs with a total length of 4.6 Mb and contained 147 predicted gene models while the region was ~1.3 Mb in *A. palmeri* and contained 121 gene models. The lack of synteny between the male-specific regions of both species and the clustering of *A. palmeri* with monoecious species from a previous nuclear phylogeny (Waselkov et al. 2018) led Montgomery et al. (2021) to infer that the two species likely evolved dioecy independently.

Although the genomic resources were instrumental in revealing the male-specific regions in both species, they could not provide any inference on sex chromosome structure, content, organization, or evolution. The draft *A. tuberculatus* genome was sequenced on the PacBio Sequel II system, assembled, and polished with Illumina short reads to arrive at 841 contigs while the *A. palmeri* genome was sequenced on PacBio Sequel I system and scaffolded with Dovetail Hi-C to arrive at 303 scaffolds (Montgomery et al. 2020). The differences in sequencing, assembly pipelines, and the unphased nature of the genomes further complicated downstream comparative analysis. Fully phased chromosome-level assembly of the dioecious

Amaranthus species will allow for robust syntenic analyses, thus yielding novel insights into their sex chromosome structure, organization, and evolution.

1.4 ATTRIBUTION

Chapter 2: I procured the dioecious *Amaranthus* seeds used in the study from the USDA Germplasm Resources Information Network (GRIN). I performed all greenhouse and lab experiments, carried out the bioinformatics analyses, and wrote the manuscript. Coauthors Dr. Lucas Bobadilla and Dr. Patrick Tranel conceived the original research study and revised the manuscript. This chapter was published in *BMC Biol* 21, 37 (2023). https://doi.org/10.1186/s12915-023-01539-9

Chapter 3: I conducted all computational analyses and wrote the manuscript. Coauthor Dr. Patrick Tranel conceived the study design and contributed to the revision of the manuscript. This chapter was published in *BMC Ecol Evo* 23, 15 (2023). https://doi.org/10.1186/s12862-023-02121-1

Chapter 4: Jacob Montgomery and Dr. Sarah Morran grew plants, harvested, and prepared samples for shipping to the Genome Center of Excellence at Corteva. Dr. Victor Llaca and Kevin Fengler performed DNA and RNA extraction, PacBio HiFi sequencing, Bionano DLS, Hi-C Seq, Iso-Seq sequencing, as well as genome data integration and assembly. Dr. Eric Patterson and Dr. Luan Cutti annotated the genome and provided genome assembly metrics. I carried out genome-wide association analysis, adaptive evolution analysis, transposable element analysis, comparative genomics analyses, statistical analyses, and wrote the manuscript. Dr. Patrick Tranel, Dr. Todd Gaines, and Dr. Eric Patterson conceived the research study, and contributed to the revision of the manuscript. This chapter has been submitted for journal publication.

Chapter 5: I carried out the genome-wide association analysis, transposable element analysis, comparative genomics analyses, and wrote the manuscript. The contribution of other coauthors to this chapter is the same as described in Chapter 4 above.

1.5 REFERENCES

Akagi T, Henry IM, Tao R, Comai L (2014) A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. Science (80) 346:646

Akagi T, Pilkington SM, Varkonyi-Gasic E, Henry IM, Sugano SS, Sonoda M, Firl A, McNeilage MA, Douglas MJ, Wang T, Rebstock R, Voogd C, Datson P, Allan AC, Beppu K, Kataoka I, Tao R (2019) Two Y-chromosome-encoded genes determine sex in kiwifruit. Nat Plants 5:801–809

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21:1–16

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Bayón ND (2015) Revisión taxonómica de las especies monoicas de *Amaranthus* (Amaranthaceae): *Amaranthus* subg. *Amaranthus* y *Amaranthus* subg. *Albersia*. Ann Missouri Bot Gard 101:261–383

Bayón ND, Peláez C (2012) A new species of *Amaranthus* (Amaranthaceae) from Salta, Argentina. Novon 22:133–136

Boetzer M, Henkel C V., Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579

Boualem A, Fergany M, Fernandez R, Troadec C, Martin A, Morin H, Sari MA, Collin F, Flowers JM, Pitrat M, Purugganan MD, Dogimont C, Bendahmane A (2008) A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons. Science (80) 321:836–838

Boualem A, Troadec C, Camps C, Lemhemdi A, Morin H, Sari MA, Fraenkel-Zagouri R, Kovalski I, Dogimont C, Perl-Treves R, Bendahmane A (2015) A cucurbit androecy gene reveals how unisexual flowers develop and dioecy emerges. Science (80) 350:688–691

Carey SB, Lovell JT, Jenkins J, Leebens-Mack J, Schmutz J, Wilson MA, Harkess A (2022) Representing sex chromosomes in genome assemblies. Cell Genomics 2:100132

Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. Am Nat 112:975–997

Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, Zhao H, Song W, Hufford MB, Schnable JC, Jin W, Lai J (2023) A complete telomere-to-telomere assembly of the maize genome. Nat Genet 55:1221–1231

Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, Li H (2022) Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol 40:1332–1335

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054

Darolti I, Almeida P, Wright AE, Mank JE (2022) Comparison of methodological approaches to the study of young sex chromosomes: A case study in *Poecilia*. J Evol Biol:1646–1658

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B,

Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, DeWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science (80) 323:133–138

Fan W, Li R (2012) Test driving genome assemblers. Nat Biotechnol 30:330–331

Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, Maguire J, Mahmoud M, Cheng H, Heller D, Zook JM, Moemke T, Marschall T, Sedlazeck FJ, Aach J, Chin CS, Church GM, Li H (2021) Chromosome-scale, haplotype-resolved assembly of human genomes. Nat Biotechnol 39:309–312

Golenberg EM, West NW (2013) Hormonal interactions and gene regulation can link monoecy and environmental plasticity to the evolution of dioecy in plants. Am J Bot 100:1022–1037

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351

Grant WF (1959) Cytogenetic studies in *Amaranthus*. Can J Bot 37:413–417

Harkess A, Huang K, van der Hulst R, Tissen B, Caplan JL, Koppula A, Batish M, Meyers BC, Leebens-Mack J (2020) Sex determination by two Y-linked genes in garden asparagus. Plant Cell 32:1790–1796

He L, Jia KH, Zhang RG, Wang Y, Shi T Le, Li ZC, Zeng SW, Cai XJ, Wagner ND, Hörandl E, Muyle A, Yang K, Charlesworth D, Mao JF (2021) Chromosome-scale assembly of the genome of *Salix dunnii* reveals a male-heterogametic sex determination system on

chromosome 7. Mol Ecol Resour 21:1966–1982

Henry IM, Akagi T, Tao R, Comai L (2018) One hundred ways to invent the sexes: Theoretical and observed paths to dioecy in plants. Annu Rev Plant Biol 69:553–575

Hu T, Chitnis N, Monos D, Dinh A (2021) Next-generation sequencing technologies: An overview. Hum Immunol 82:801–811

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I (2017) ABySS 2.0 : Resource-Efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–777

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338–345

Jiang L, Lin M, Wang H, Song H, Zhang L, Huang Q, Chen R, Song C, Li G, Cao Y (2022) Haplotype-resolved genome assembly of *Bletilla striata* (Thunb.) Reichb.f. to elucidate medicinal value. Plant J 111:1340–1353

Käfer J, Marais GAB, Pannell JR (2017) On the rarity of dioecy in flowering plants. Mol Ecol 26:1225–1241

Kersey PJ (2019) Plant genome sequences: past, present, future. Curr Opin Plant Biol 48:1–8

Khan A, Carey SB, Serrano A, Zhang H, Hargarten H, Hale H, Harkess A, Honaas L (2022) A phased, chromosome-scale genome of 'Honeycrisp' apple (*Malus domestica*). Gigabyte 2022:1–15

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL,

Smith TPL, Phillippy AM (2018) *De novo* assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol 36:1174–1182

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res 27:722–736

Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, Hiendleder S, Fedrigo O, Jarvis ED, Adam M, Eichler EE, Williams JL, Smith TPL, Hall RJ, Shawn T, Kingan SB (2019) Extended haplotype phasing of *de novo* genome assemblies with FALCON-Phase. bioRxiv:1–27

Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, Porubsky D, Kuhn K, Mueller KA, Low WY, Hiendleder S, Fedrigo O, Liachko I, Hall RJ, Phillippy AM, Eichler EE, Williams JL, Smith TPL, Jarvis ED, Sullivan ST, Kingan SB (2021) Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nat Commun 12:1–10

Lloyd DG (1980) The distributions of gender in four angiosperm species illustrating two evolutionary pathways to dioecy. Evolution 34:123–134

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J (2012) SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1:18

Ma X, Yu L, Fatima M, Wadlington WH, Hulse-Kemp AM, Zhang X, Zhang S, Xu X, Wang J, Huang H, Lin J, Deng B, Liao Z, Yang Z, Ma Y, Tang H, Van Deynze A, Ming R (2022)

The spinach YY genome reveals sex chromosome evolution, domestication, and introgression history of the species. Genome Biol 23:1–30

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Stanková H, Vrána J, Chan S, Munõz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doleael J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433

McCord RP, Kaplan N, Giorgetti L (2020) Chromosome conformation capture and beyond: Toward an integrative view of chromosome structure and function. Mol Cell 77:688–708

Montalvão Leite P ana, Kersten B, Kim G, Fladung M, Müller NA (2022) ARR17 controls dioecy in *Populus* by repressing B-class MADS-box gene expression. Phil Trans R Soc B 377:20210217

Montalvão Leite PA, Kersten B, Fladung M, Müller NA (2021) The diversity and dynamics of sex determination in dioecious plants. Front Plant Sci 11:1–12

Montgomery JS, Giacomini D, Waithaka B, Lanz C, Murphy BP, Campe R, Lerchl J, Landes A, Gatzmann F, Janssen A, Antonise R, Patterson E, Weigel D, Tranel PJ (2020) Draft genomes of *Amaranthus tuberculatus*, *Amaranthus hybridus*, and *Amaranthus palmeri*.

Genome Biol Evol 12:1988–1993

Montgomery JS, Giacomini DA, Weigel D, Tranel PJ (2021) Male-specific Y-chromosomal regions in waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). New Phytol 229:3522–3533

Montgomery JS, Sadeque A, Giacomini DA, Brown PJ, Tranel PJ (2019) Sex-specific markers for waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). Weed Sci 67:412–418

Müller NA, Kersten B, Leite Montalvão AP, Mähler N, Bernhardsson C, Bräutigam K, Carracedo Lorenzo Z, Hoenicka H, Kumar V, Mader M, Pakull B, Robinson KM, Sabatti M, Vettori C, Ingvarsson PK, Cronk Q, Street NR, Fladung M (2020) A single gene underlies the dynamic evolution of poplar sex determination. Nat Plants 6:630–637

Murray MJ (1940) The genetics of sex determination in the family Amaranthaceae. Genetics 25:409–431

Muyle A, Martin H, Zemp N, Mollion M, Gallina S, Tavares R, Silva A, Bataillon T, Widmer A, Glémin S, Touzet P, Marais GAB (2021) Dioecy is associated with high genetic diversity and adaptation rates in the plant genus *Silene*. Mol Biol Evol 38:805–818

Renner SS (2014) The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database. Am J Bot 101:1588–1596

Renner SS, Müller NA (2021) Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. Nat Plants 7:392–402

Renner SS, Müller NA (2022) Sex determination and sex chromosome evolution in land plants. Philos Trans R Soc B Biol Sci 377

Renner SS, Ricklefs RE (1995) Dioecy and its correlates in the flowering plants. Am J Bot

82:596

Renner SS, Won H (2001) Repeated evolution of dioecy from monoecy in Siparunaceae (Laurales). Syst Biol 50:700–712

Ruan J, Li H (2020) Fast and accurate long-read assembly with wtdbg2. Nat Methods 17:155–158

Sauer J (1955) Revision of the dioecious amaranths. Madroño 13:5–46

Sauer J (1972) The dioecious amaranths: A new species name and major range extensions. Madrono 21:426

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH (2017) DNA sequencing at 40: Past, present and future. Nature 550

Slatko BE, Gardner AF, Ausubel FM (2018) Overview of next generation sequencing technologies (and bioinformatics) in cancer. Mol Biol 122:1–15

Standley LA (1985) Paradioecy and gender ratios in *Carex macrocephala* (Cyperaceae). Am Midl Nat 113:283–286

Sun H, Jiao WB, Krause K, Campoy JA, Goel M, Folz-Donahue K, Kukat C, Huettel B, Schneeberger K (2022) Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet 54:342–348

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, Dong Q, Ye K (2022) High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. Genomics, Proteomics Bioinforma 20:4–13

Waselkov KE, Boleda AS, Olsen KM (2018) A phylogeny of the genus *Amaranthus* (Amaranthaceae) based on several low-copy nuclear loci and chloroplast regions. Syst Bot 43:439–458

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin CS, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37:1155–1162

Yang HW, Akagi T, Kawakatsu T, Tao R (2019) Gene networks orchestrated by *MeGI*: a single-factor mechanism underlying sex determination in persimmon. Plant J 98:97–111

Yuan Y, Chung CYL, Chan TF (2020) Advances in optical mapping for genomic research. Comput Struct Biotechnol J 18:2051–2062

Zerbino DR (2010) Using the Velvet *de novo* assembler for short-read sequencing technologies. Curr Protoc Bioinforma

Zhang X, Wu R, Wang Y, Yu J, Tang H (2020) Unzipping haplotypes in diploid and polyploid genomes. Comput Struct Biotechnol J 18:66–72

Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677

CHAPTER 2: GENOMIC PROFILING OF DIOECIOUS *AMARANTHUS* SPECIES PROVIDES NOVEL INSIGHTS INTO SPECIES RELATEDNESS AND SEX GENES

ABSTRACT

Amaranthus L. is a diverse genus consisting of domesticated, weedy, and non-invasive species distributed around the world. Nine species are dioecious, of which *Amaranthus palmeri* S. Watson and *Amaranthus tuberculatus* (Moq.) J.D. Sauer are troublesome weeds of agronomic crops in the United States and elsewhere. Shallow relationships among the dioecious *Amaranthus* species and the conservation of candidate genes within previously identified *A. palmeri* and *A. tuberculatus* male-specific regions of the Y (MSYs) in other dioecious species are poorly understood. In this study, seven genomes of dioecious amaranths were obtained by paired-end short-read sequencing and combined with short reads of seventeen species in the family Amaranthaceae from NCBI database. The species were phylogenomically analyzed to understand their relatedness. Genome characteristics for the dioecious species were evaluated and coverage analysis was used to investigate the conservation of sequences within the MSY regions.

We provide genome size, heterozygosity, and ploidy level inference for seven newly sequenced dioecious *Amaranthus* species and two additional dioecious species from the NCBI database. We report a pattern of transposable element proliferation in the species, in which seven species had more *Ty3* elements than *copia* elements while *A. palmeri* and *A. watsonii* had more *copia* elements than *Ty3* elements, similar to the TE pattern in some monoecious amaranths. Using a Mash-based phylogenomic analysis, we accurately recovered taxonomic relationships among the dioecious *Amaranthus* species that were previously identified based on comparative morphology. Coverage analysis revealed eleven candidate gene models within the *A. palmeri*

MSY region with male-enriched coverages, as well as regions on Scaffold 19 with female-enriched coverage, based on *A. watsonii* reads alignments. A previously reported *FLOWERING LOCUS T* (*FT*) within *A. tuberculatus* MSY contig was also found to exhibit male-enriched coverages for three species closely related to *A. tuberculatus* but not for *A. watsonii* reads. Additional characterization of the *A. palmeri* MSY region revealed that 78% of the region is made of repetitive elements, typical of a sex determination region with reduced recombination.

The results of this study further increase our understanding of the relationships among the dioecious species of the *Amaranthus* genus as well as revealed genes with potential roles in sex function in the species.

2.1 INTRODUCTION

The genus *Amaranthus* L. is a diverse plant group of 70 – 80 species distributed across the world's temperate and tropical regions (Sauer 1967). Nine of these species [*Amaranthus acanthochiton* J.D. Sauer, *Amaranthus arenicola* I.M. Johnson, *Amaranthus australis* (A. Gray) J.D. Sauer, *Amaranthus cannabinus* (L.) J.D. Sauer, *Amaranthus floridanus* (S. Watson) J.D. Sauer, *Amaranthus tuberculatus* (Moq.) J.D. Sauer, *Amaranthus greggii* S. Watson, *Amaranthus watsonii* Standley, and *Amaranthus palmeri* S. Watson] are dioecious (i.e., separate male and female individual plants), native to North America and grouped collectively into the subgenus *Acnida* (L.) Aellen ex K.R. Robertson (Sauer 1955, 1972; Mosyakin and Robertson 1996).

The *Amaranthus* genus has been described as taxonomically challenging due to morphological similarities among species (Costea and DeMason 2001). Relationships among species of the genus, including the dioecious ones, were previously investigated using several molecular markers and phylogenetic frameworks (Lanoue et al. 1996; Chan and Sun 1997; Xu and Sun 2001; Stetter and Schmid 2017; Waselkov et al. 2018). Stetter and Schmid (2017), with

an objective to elucidate the domestication history of cultivated amaranths, used genotyping-by-sequencing (GBS) for 35 species of the genus in neighbor joining and multispecies coalescent (MSC) frameworks to infer *A. hybridus* as likely ancestor of the cultivated amaranths, *A. caudatus*, *A. cruentus* and *A. hypochondriacus*. In the most recent attempt to reconstruct the evolutionary relationships among the species of the genus, Waselkov et al. (2018) sampled 58 species, including the nine dioecious species, and used six molecular markers (ITS, *A36*, *G3PDH*, *waxy*, *trnL5'-trnL3'* and *matk/trnK*) in a maximum parsimony and Bayesian inference phylogenetic framework. Trees from both studies were congruent with high support for deeper node relationships, such as species clustering or clades corresponding to previously defined three subgenera, *Acnida*, *Amaranthus* and *Albersia* (Mosyakin and Robertson 1996). Relationships among the dioecious species along "shallow" nodes however were poorly resolved with weak supports and, thus, some relationships remain unclear (e.g., is *A. tuberculatus* more closely related to *A. arenicola* than to *A. floridanus*?).

While advances in molecular phylogenetics have increased the level of inference we can draw on trait evolution or species relationships, poorly-resolved trees resulting from biological processes (e.g., ancient or recent hybridization, incomplete lineage sorting, introgression or rapid radiation) or systematic errors (e.g., low parsimony-informativeness of markers) still make inference on trait evolution intractable for some genera (Renner 2014). Several methods estimating phylogenetic relationships that put into consideration these biological processes have gained attention (Durand et al. 2011; Wen et al. 2018; Kubatko and Chifman 2019), however, few are able to explicitly estimate species trees from phylogenomic data taking into account several sources of conflict and heterogeneity in molecular substitution (Morales-Briones et al. 2021). Thus, complementary approaches are often required for robust relationship inference.

Phylogenetic approaches (e.g., *k*-mer-based method) that by-pass challenges inherent in alignment- or assembly-based methods have been proposed, offering flexibility to sequence analysis and better use of computing power compared to alignment-based methods (Leimeister et al. 2014; Sarmashghi et al. 2019). For instance, the MinHash algorithm (Broder 1997) was implemented in a sequence clustering tool, Mash (Ondov et al. 2016), and among 74 alignment-free (AF) methods, Mash was shown to have the highest performance for genome-based phylogeny of plants using unassembled reads (Zielezinski et al. 2019).

Aside from interests in the evolutionary relationship among *Amaranthus* species, there is also renewed interests in the dioecious species for their weedy trait characteristics (Ward et al. 2013; Tranel 2021) and their mechanisms of sex determination or dioecy evolution (Neves et al. 2020; Montgomery et al. 2021). Although, many of the dioecious species are restricted to their geographic range and currently of little economic importance with regards to food source relative to cultivated monoecious species (Sauer 1950, 1967; Aderibigbe et al. 2022), *Amaranthus tuberculatus* and *A. palmeri* are two agronomically important weeds in North America (Steckel 2007), and have been the focus of many research studies (Trucco et al. 2005; Tranel et al. 2011; Gaines et al. 2012; Kreiner et al. 2018; Tranel 2021). The dioecious nature of both species ensures obligate outcrossing, thus enhancing high genetic diversity, prolific seed production, rapid adaptation and spread of herbicide resistance (Ward et al. 2013; Shergill et al. 2018; Tranel 2021; Heap 2023). While dioecy confers evolutionary advantages (Käfer et al. 2017; Muyle et al. 2021), a disadvantage, however, believed to be taking place naturally, is that bottleneck events could result in populations that are depleted of one of the two sexes, and if not for sex reversion, the population would collapse and thus become locally extinct (Henry et al. 2018). Considering this disadvantage an advantage from a weed management standpoint, artificial gender

manipulation, whereby sex ratios could be biased towards one gender and the genetic factors involved are inherited in a non-mendelian pattern via a gene drive system, was proposed as a possible strategy for management of weedy dioecious *Amaranthus* species (Tranel and Trucco 2009; Neve 2018).

Only until recently have the genes and the mechanisms involved in sex determination been elucidated for a few plant species (Akagi et al. 2014, 2019; Harkess et al. 2020; Müller et al. 2020; Montalvão Leite et al. 2021, 2022; Ma et al. 2022). For the amaranths, previous work on dioecy confirmed males of *A. tuberculatus* and *A. palmeri* are heterogametic and, thus, have an XY sex chromosome system (Montgomery et al. 2019; Neves et al. 2020). The male-specific region of the Y (MSY) for both species were subsequently identified, spanning a ~1.3 Mb region with 121 gene models for *A. palmeri* while several contigs with a total length of 4.6 Mb and containing 147 gene models were identified for the *A. tuberculatus* MSY region (Montgomery et al. 2020, 2021; Neves et al. 2020). Lack of synteny between the MSY regions of both species (Neves et al. 2020; Montgomery et al. 2021), and the clustering of *A. palmeri* with monoecious species in the nuclear tree from Waselkov et al.'s phylogeny (Waselkov et al. 2018), led Montgomery et al. (Montgomery et al. 2021) to infer that the two species likely evolved dioecy independently. However, the chloroplast tree from the same study that generated the nuclear tree showed a single monophyletic clade for the dioecious *Amaranthus* species (Waselkov et al. 2018). Simultaneously, Neves et al. (23) also demonstrated that dioecy in both *A. palmeri* and *A. tuberculatus* could be under the control of separate genomic regions. Based on the above evidence, we hypothesize two origins of dioecy: one shared by *A. palmeri* and *A. watsonii* and another shared by the remaining dioecious amaranths (29). While male-specific regions in

closely related species could differ in size or content, there is evidence that the same gene(s) or dioecy mechanism could still be recruited across the species (42).

The objective of this research was to use comparative genomics to investigate dioecy within the *Amaranthus* genus. We obtained whole-genome sequence from seven dioecious amaranths, and report genome characteristics, transposable element (TE) proliferation patterns, and phylogenomic relationships among the species. We identified genomic regions including candidate genes within *A. palmeri* and *A. tuberculatus* MSYs region that exhibit male-enriched coverages across other dioecious *Amaranthus* species and could have roles in sex function. Finally, we elucidated repeat contents for the *A. palmeri* MSY region to test the hypothesis that typical sex determination regions have suppressed recombination and accumulate repetitive sequences (Charlesworth 2013, 2016; Hobza et al. 2015).

2.2 MATERIALS AND METHODS

2.2.1 Plant material, DNA extraction and Illumina sequencing

Accessions of seven dioecious amaranths were obtained from USDA Germplasm Resources Information Network (GRIN) (Appendix A Table A.1). Voucher specimens of all accessions sequenced in this study can be found at the Illinois Natural History Survey (ILLS) Herbarium at the University of Illinois Robert A. Evers Laboratory. Voucher barcodes are included in Appendix A Table A.1. Seeds were grown in containers filled with a growing media that included Sunshine LC1 (Sun Gro Horticulture, 770 Silver Street Agawam, MA) growing mix, soil, peat, and torpedo sand (3:1:1:1 by weight). Two or three young leaves were harvested from each species following flower formation and visual identification of gender. Leaf tissues collected were frozen in liquid nitrogen and stored in -80 C pending DNA extraction. Genomic DNA was extracted from one male of each species and from one female each of *A.*

acanthochiton, *A. cannabinus*, *A. greggii* and *A. watsonii* following standard CTAB protocol (Doyle and Doyle 1990). DNA integrity was determined using a spectrophotometer (Nanodrop1000 Spectrophotometer, Thermo Fisher Scientific, 81 Wyman Street, Waltham, MA 02451) and by resolving the DNA on 1% agarose gel by electrophoresis. The absence of band shearing or smearing indicated high molecular weight DNA with sufficient purity and quality required for sequencing. The eleven DNA samples were submitted to the Roy J. Carver Biotechnology Center at the University of Illinois, Urbana–Champaign for sequencing. Shotgun genomic libraries were prepared with Hyper Library construction kit from Kapa Biosystems (Roche, Basel, Switzerland), and the libraries were size selected, pooled, quantitated by qPCR and paired-end sequenced (2 x 150 bp) on one S4 lane for 151 cycles on Illumina NovaSeq6000. Sequences of seventeen other species belonging to either the *Amaranthus* genus or broadly a member of the family Amaranthaceae were downloaded from the NCBI database. Sequencing platforms for these genomes varied from Illumina Hiseq 2500 to Novaseq 6000 (Appendix A Table A.1).

2.2.2 Genome size, heterozygosity, and ploidy analysis

The genome sizes for the species sequenced were estimated with GenomeScope v2.0 [Ranallo-Benavidez et al. (38); https://github.com/tbenavi1/genomescope2.0]. A *k*-mer length, *k*, of 21 was chosen for genome size estimation based on the recommendations from the authors, which was seen as a balance between speed of computation and accuracy. *K*-mer frequencies were generated from the adapter trimmed Illumina sequences for each of the nine dioecious amaranth species with Jellyfish v2.3.0 (Marçais and Kingsford 2011) using parameters: count -C -m 21 -s 3G -t 6 /dev/fd/0 -o output_reads.jf, and histograms of *k*-mer frequencies were obtained using the 'histo' sub-command and --high=1000000 flag. Genome sizes were then estimated

from the histograms using GenomeScope v2.0 with parameters: -i reads.histo -o output_dir -k 21 -m 1000000. The *k*-mer histograms obtained from previous steps were further analyzed with two *k*-mer-based tools, CovEST v0.5.6 [(Hozza et al. 2015); https://github.com/Alexdami17/covest] and FindGSE [(Sun et al. 2018); https://github.com/schneebergerlab/findGSE]. We used both the "basic" and "repeats" model of CovEST with default parameters, except -r 150. The "basic" model is for simple genomes without repeats; however, species of the *Amaranthus* genus have been shown to be made of at least 50% repetitive elements (Lightfoot et al. 2017; Ma et al. 2021). Moreover, the "repeats" model is error-aware, accounts for repeat structures and performs well on data with low sequencing coverage (Hozza et al. 2015).

The ploidy levels for each of the genomes were also estimated using Smudgeplot v0.2.3 [(Ranallo-Benavidez et al. 2020); https://github.com/KamilSJaron/smudgeplot]. *K*-mer frequencies were first generated using KMC v3.1.1 [(Kokot et al. 2017); https://github.com/tbenavi1/KMC] with parameters: -k21 -t10 -m30 -ci1 -cs10000 @FILES kmer_counts tmp and then converted to *k*-mer frequency histogram using parameters: kmc_tools transform kmcdb histogram species_k21.hist -cx10000. 'FILES' contain the raw read names for forward and reverse reads. The 'smudgeplot.py cutoff species_k21.hist L/U' was then used to estimate *k*-mer coverage thresholds from the histogram file. *K*-mers in the coverage range from L to U were extracted with the command 'kmc_tools transform', and smudge_pairs command was used to reduce the file to compute set of *k*-mer pairs. The smudgeplots showing proposed ploidy for each of the genomes were then generated with coverages of identified *k*-mer pairs (i.e., species_coverages.tsv file) using 'smudgeplot.py plot' command. Haploid *k*-mer coverages were estimated directly from the histogram generated by KMC, rather than supplied from GenomeScope output.

2.2.3 Transposable element analysis of unassembled *Amaranthus* genomes

We analyzed repetitive elements in the unassembled Illumina raw reads from males of sequenced dioecious *Amaranthus* species using a similarity-based clustering tool, RepeatExplorer2 on a dedicated cloud galaxy instance (Novák et al. 2010, 2020) (https://repeatexplorer-elixir.cerit-sc.cz/galaxy). The sex of *A. palmeri* was unidentified by the authors in their study (Molin et al. 2017); however, we included the raw reads sequence for comparison. A recommendation from authors of RepeatExplorer2 is that coverage greater than 1x be avoided while coverage between 0.1 – 0.5X is optimal. We therefore subsampled all reads to 0.3x with rasusa v0.6.1 (Hall 2022) (https://github.com/mbhall88/rasusa) using parameters: -i r1.fq -i r2.fq --coverage 0.3 --genome-size estimated-genomesize-from-genomescope -o out.r1.fq -o out.r2.fq -s 15. For each species: 1,263,202 (*A. acanthochiton*), 1,142,178 (*A. arenicola*), 1,739,786 (*A. australis*), 1,424,616 (*A. cannabinus*), 1,212,116 (*A. floridanus*), 1,237,964 (*A. tuberculatus*), 1,394,368 (*A. greggii*), 806,748 (*A. watsonii*) and 910,966 (*A. palmeri*) read pairs were kept after subsampling. Reads of *A. hybridus* (894,080), *A. hypochondriacus(*1,279,884), and *A. cruentus* (979,906) subsampled to 0.3x were also included for comparisons. The FastQ read pairs for each species were quality filtered and interleaved with "Preprocessing of FASTQ paired-end reads" tool in RepeatExplorer Utilities on the galaxy instance. The interleaved reads were then analyzed for repeats with RepeatExplorer2 clustering tool using default parameters. The clusters of repeats within each supercluster were manually inspected to ensure accuracy of the automated repeats prediction. Repeat proportions from the curated cluster table were then estimated using the "Repeat proportions from CLUSTER_TABLE" tool also on the galaxy instance.

We complemented our repeat discovery approach using dnaPipeTE v1.3.1 (Goubert et al. 2015). First, we constructed a representative repeat library for the amaranths from the genome of *A. hypochondriacus* (Lightfoot et al. 2017). *De novo* identification of species-specific repeats in the genome was carried out with RepeatModeler v2.0.2 using default parameters (Flynn et al. 2020). A curated RepBase database (110; RepeatmaskerEdition-20181026) was combined with RepeatMasker default Dfam3.2 database, and 'famdby.py' utility was used to query the combined database to obtain a library of 'viridiplantae' repeats with parameters: -i RepeatMaskerLib.h5 families --format fasta_name --include-class-in-name --ancestors --descendants 'viridiplantae'. We performed additional LTR structural analysis using LTR_retriever pipeline (Ou and Jiang 2018), first by analyzing the genome with LTR_harvest (Ellinghaus et al. 2008) from genometools v1.6.0 using the parameters: -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes, and then through LTR_FINDER_parallel (Ou et al. 2018) using default parameters. Output from both LTR_harvest and LTR_FINDER_parallel were concatenated and analyzed with LTR_retriever v2.9.0 to obtain a non-redundant LTR library using default parameters (Ou and Jiang 2018). The non-redundant LTR library was then merged with 'viridiplantae' repeats and the species-specific consensus library of repeats. To reduce redundancy, the final repeat library was clustered using CD-HIT-EST v4.6 (Li and Godzik 2006) with parameters: -c 0.8 -G 1 -s 0.9 -aL 0.8 -aS 0.8 -M 5000 -T 6 -i. The repeat library was then used with dnaPipeTE for repeat discovery in each species using parameters: -RM_lib repeat library -genome_size estimated-genome-size-from-genomescope -genome_coverage 0.3 and other parameters default. Prior to repeat analysis with dnaPipeTE, we first mapped the raw reads of each species to *A. hypochondriacus* chloroplast genome (GenBank accession number KX279888) (Chaney et al. 2016), and subsequently to *Beta*

vulgaris mitochondrial genome (GenBank accession number BA000009) (Kubo et al. 2000) using Bowtie v2.4.4 (Langmead and Salzberg 2012) while keeping only non-aligned reads with parameters: -p 32 -X 1000 --un-conc. This step was taken to avoid the assembly of organellar DNA into contigs that could spuriously be annotated as repeats.

For TE quantification in available *Amaranthus* genome assemblies, we constructed species-specific libraries for each of the species following the method described previously. Final curated libraries were then used to analyze and annotate repeats in the genomes using RepeatMasker v4.1.2-p1 (http://www.repeatmasker.org/RepeatMasker/) with default parameters.

2.2.4 Mash-based whole-genome phylogenetic analysis

Quality of the Illumina raw reads obtained from 17 species in the NCBI database was accessed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and aggregated with MultiQC v1.5 (Ewels et al. 2016). Low quality bases and adapters were then removed with Trimmomatic (Bolger et al. 2014) using parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36. *Chenopodium quinoa* raw reads (Project number PRJNA821252) had Nextera adapter sequences and were thus removed using parameters: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36. To then determine relatedness among the seven sequenced dioecious *Amaranthus* genomes in this study and the 17 other species from the public repository, we used an assembly/alignment-free tool, Mashtree v1.2.0 with the following parameters: --mindepth 0 --numcpus 6 *FORWARD.fastq.gz > mashtree.dnd (Katz et al. 2019). Mashtree handles only single reads, therefore we used only forward reads from the paired read sequences. We included the female reads of four species (*A. acanthochiton*, *A. cannabinus*, *A. greggii*, and *A. watsonii*) to ascertain the robustness of the alignment-free approach in that males are expected to cluster with

the respective females of the species. Mashtree uses a *k*-mer strategy in a two-step approach, first adopting the MinHash algorithm of Mash to create genome sketches (Ondov et al. 2016), and second using the sketches to determine distances between genomes as a pairwise distance matrix, which is subsequently used to build a neighbor-joining tree in QuickTree (Howe et al. 2002). The output tree (.dnd) from Mashtree was visualized and annotated with FigTree v1.4.4 (https://github.com/rambaut/figtree).

2.2.5 Whole-sequence alignments and coverage analysis of *Amaranthus palmeri* and *Amaranthus tuberculatus* male-specific region of the Y.

Demultiplexing of Fastq raw reads was carried out with Illumina bcl2fastq v2.20 Conversion Software, and quality control, including adapter trimming from the reads, was carried out by the sequencing facility. A total of ~6.23 Gb of raw reads were obtained corresponding to 528,703,130 (*A. acanthochiton* female, 127x genome coverage), 604,304,170 (*A. acanthochiton* male, 145x), 533,480,886 (*A. arenicola* male, 142x), 642,410,494 (*A. australis* male, 121x), 676,006,832 (*A. cannabinus* female, 144x), 592,414,420 (*A. cannabinus* male, 126x), 572,691,874 (*A. floridanus* male, 143x), 540,070,720 (*A. greggii* female, 118x), 525,935,576 (*A. greggii* male, 115x), 489,955,162 (*A. watsonii* female, 183x) and 525,324,158 (*A. watsonii* male, 197x) read pairs. The quality of reads for *A. palmeri* and *A. tuberculatus* obtained from the NCBI database was accessed as previously described.

All reads (seven males and four females) were then mapped to the *A. palmeri* and *A. tuberculatus* draft genomes (Montgomery et al. 2020) with BWA-MEM v0.7.5 using default settings (Li 2013). The tool ‘fixmate’ within SAMtools v1.14 was used to fill mate coordinates and insert size fields (Li et al. 2009), and duplicates in the reads alignments were marked with Picard v2.26.9 (http://broadinstitute.github.io/picard/). SAMtools flagstat was then used to

compute overall summary statistics of reads alignment. Alignment files for each species were filtered with SAMtools to remove reads with mapping quality (MAPQ) < 5, alternative hits (tag XA:Z) and supplementary alignments (tag SA:Z). Coverage analysis was then carried out with the filtered alignments using DifCover (Smith et al. 2018; Timoshevskiy et al. 2019), which puts into consideration the modal coverage of male and female samples for depth normalization and also accounts for problematic region such as highly repetitive regions or gaps. DifCover was recently implemented in a computational workflow (SexFindR) that identifies sex chromosomal regions (Grayson et al. 2022). The estimated genome-wide coverages represented as the ratio of log2 male-to-female reads mapped to both *A. palmeri* and *A. tuberculatus* genome assemblies were then plotted with the R packages tidyverse (Wickham et al. 2019) and ggpubr (https://github.com/kassambara/ggpubr).

Additionally, read coverages for scaffold 20, the location of the *A. palmeri* MSY region, were calculated and normalized from the filtered alignments using bamCoverage v3.5.1 (Ramírez et al. 2014) with parameters: -b input.bam -o output_cov -of bigwig -bs 20 -r region-of-interest --effectiveGenomeSize estimated-genome-size-from-genomescope --normalizeUsing RPGC --smoothLength 60 --extendReads 150 --ignoreDuplicates --exactScaling -p 5. Coverages and gene annotations were then plotted and visualized using rtracklayer v1.54.0 (Lawrence et al. 2009), GenomicFeatures v1.46.5 (Lawrence et al. 2013), and Gviz v1.38.3 (Hahne and Ivanek 2016) in R v4.1.2 (R Core Team 2021).

We also accessed the presence of *NRT1/PTR6.4* recently proposed as a sex determinant in spinach (Ma et al. 2022) in the genomes of *A. hypochondriacus*, *A. palmeri* and *A. tuberculatus* genomes using a BLAST search on CoGE (Lyons and Freeling 2008). Although both spinach (subfamily Chenopodoideae) and amaranth (subfamily Amaranthoideae) lineages are

paraphyletic from previous phylogeny (Yao et al. 2019; Morales-Briones et al. 2021), our hypothesis that the gene could be present in the amaranths was informed by the two lineages belonging to the family Amaranthaceae and show some relationships in the previous trees.

2.2.6 Transcription factors and repetitive elements within *Amaranthus palmeri* male-specific region of the Y.

Identification of transcription factors among candidate gene models within the *A. palmeri* MSY region was carried out using PlantTFcat (Dai et al. 2013). A custom repeat library for *A. palmeri* genome (Montgomery et al. 2020) was also prepared as previously described for the *A. hypochondriacus* genome. The library of repeats was then used to analyze and annotate repetitive elements within the MSY region of *A. palmeri* using RepeatMasker v4.1.2-p1 with default parameters.

2.3 RESULTS

2.3.1 Genome size, heterozygosity, and ploidy estimation

We employed *k*-mer-based tools to estimate genome sizes, heterozygosity, and ploidy for dioecious amaranths (Figure 2.1). Estimates of genome sizes using GenomeScope (Ranallo-Benavidez et al. 2020) were 793.3 Mb (*A. australis*), 702.0 Mb (*A. cannabinus*), 684.6 Mb (*A. greggii*), 621.5 Mb (*A. acanthochiton*), 615.8 Mb (*A. tuberculatus*), 596.6 Mb (*A. floridanus*), 563.1 Mb (*A. arenicola*), 399.9 Mb (*A. watsonii*) and 374.4 Mb (*A. palmeri*) (Appendix A Figure A.1 – A.9). The genome size estimates fall within the confidence bounds of previously reported genome sizes for *A. australis* (95% CI 735.7 – 912.8), *A. floridanus* (95% CI 543.5 – 772.9) and *A. palmeri* (95% CI 307.1 – 536.5) based on flow cytometry while the estimate for *A. tuberculatus* was 5.6 Mb lower than the lower confidence limit from previous estimate (95% CI 621.4 – 729.8) (Stetter and Schmid 2017). Analysis of raw reads of monoecious species (*A.*

hybridus SRR12075659, *A. hypochondriacus* SRR2106212 and *A. cruentus* SRR13980261) also revealed genome size estimates consistent with previous flow cytometry results (Appendix A Figure A.10 – A.12). The estimate of 398 Mb for *A. cruentus* reported in Ma et al. (Ma et al. 2021) however appears to be underestimated based on our reanalysis. We report a genome size of 489 Mb for the species which is consistent with previous estimates from flow cytometry (Appendix A Figure A.12).

CovEST repeats model and FindGSE yielded apparent overestimations of genome sizes for all species while CovEST basic model gave both apparent over and underestimates (Appendix A Table A.2). CovEST and FindGSE, like GenomeScope, estimate genome characteristics from *k*-mer frequencies; however, they differ in the distribution or models adopted. GenomeScope fits a non-linear least square to a negative binomial distribution using Levenberg-Marquardt algorithm (Ranallo-Benavidez et al. 2020), CovEST use a Poisson distribution for *k*-mer abundance spectrum adopting a probabilistic framework (Hozza et al. 2015), and FindGSE fits *k*-mer frequencies with a skew normal distribution (Sun et al. 2018). It is possible that the distribution or model adopted in fitting *k*-mer frequencies by CovEST or FindGSE is less suitable considering our *k*-mer counts data, thereby resulting in the inflation of genome sizes. Similar observation where estimates from CovEST Repeat model was higher than estimates from GenomeScope were reported for species of beetles (Pflug et al. 2020).

Estimates of heterozygosity for *A. palmeri* (2.72%), *A. watsonii* (2.07%), and *A. arenicola* (2.06%) were higher than those of the other species, which ranged from 0.03% for *A. cruentus* to 1.97% for *A. acanthochiton* (Appendix A Figure A.1 – A.12), indicating that high allelic variation could introduce assembly difficulty for some of the species (Asalone et al. 2020).

We predicted the ploidy level for each of the genomes using Smudgeplot (Ranallo-Benavidez et al. 2020) in order to determine if species were polyploids, which may impact downstream analysis (e.g., reads mapping). All seven of the dioecious species sequenced from this study, including two other dioecious species and monoecious ones, were inferred as diploids. The *k*-mer coverage (kcov) in GenomeScope plots also corresponds to the haploid *k*-mer coverage (1n) in Smudgeplot, indicating the accuracy of ploidy prediction (Appendix A Figure A.1 – A.12). Smudgeplot initially inferred tetraploidy for *Amaranthus greggii* when it was allowed to automatically detect haploid *k*-mer coverage at 44, similar to when Smudgeplot originally predicted tetraploidy for the diploid *Fragaria iinumae* strawberry genome (Ranallo-Benavidez et al. 2020). However, rerunning Smudgeplot with the *k*-mer coverage from GenomeScope (kcov = 42) and increasing the lower *k*-mer coverage threshold value, L, to 20 caused it to infer diploidy (Additional file 2: Fig. S7). Nevertheless, the proportion of "AABB" smudge was as high as "AB" smudge for *A. greggii* relative to other species, indicating higher rates of duplications or paralogs (Appendix A Figure A.7).

2.3.2 Transposable element analysis of unassembled *Amaranthus* genomes

To gain insight into the impact of repetitive elements on genome structure of dioecious *Amaranthus* species, we subjected subsampled read pairs of the nine dioecious species to RepeatExplorer2 (Novák et al. 2020), a graph-based repetitive sequence clustering and characterization tool for Illumina raw reads. Subsampled reads correspond to 0.3X coverage for each genome (see methods). Results of the repeat analysis is presented in Appendix A Table A.3A.

The total TE content identified in the nine genomes of the dioecious amaranths in RepeatExplorer2 pipeline was less than the total TE content discovered in the genome

assemblies of the species, *A. hypochondriacus* at 51.76% (Lightfoot et al. 2017), *A. cruentus* at 57.7% (Ma et al. 2021), *A. hybridus* at 57.34%, *A. palmeri* at 56.03% or *A. tuberculatus* at 66.06% (Appendix A Table A.4 – A.5). The total composition of TE for *A. tuberculatus* male genome reported here is similar to the 66.28% reported for a previously assembled female genome of the same species (Hnatovska 2022). It is worth mentioning that 57.68% of *A. hypochondriacus* genome [9.49% *copia* and 7.88% *Ty3*] was made up of repetitive elements when the genome was reanalyzed using more recent TE discovery tools (Appendix A Table A.4). A similar observation was reported for the human genome, where RepeatMasker identified 48% of the genome as TEs, a proportion that further increased to 53% on re-analysis of the genome with the addition of Dfam2.0 database (Goerner-Potvin and Bourque 2018).

Reanalysis of the short reads with dnaPipeTE pipeline and using a species-specific library from *A. hypochondriacus* identified more proportion of total TEs in the genomes (Appendix A Table A.3B). Although both dnaPipeTE and RepeatExplorer2 operate on the same principle, dnaPipeTE could annotate a larger fraction of TEs (Goubert et al. 2015). Our analysis identified the abundance of low copy repeats as a major source of discrepancies between dnaPipeTE and RepeatExplorer2 repeat quantification for the amaranths (Appendix A Table A.3B, Appendix A A.13 – A.24). The total TE estimates for *A. tuberculatus* and *A. hybridus* using dnaPipeTE were 10% less than the total TE in their genome assemblies (Appendix A Table A.4). For *A. palmeri*, *A. hypochondriacus* and *A. cruentus*, differences in total TE between dnaPipeTE and the genome assembly were 19%, 18% and 22%, respectively.

Despite TEs being underestimated in our study, the dynamics of relative TE accumulation for species within the genus is still interesting. *Amaranthus acanthochiton, A. arenicola, A. australis, A. cannabinus, A. floridanus, A. tuberculatus* and *A. greggii* had more

Ty3 element than *copia* element (Appendix A Table A.4). This pattern of relative TE composition using raw reads of *Amaranthus tuberculatus* [6.62% *copia* and 8.29% *Ty3*] is similar to TE composition in its assembled genome, where *copia* elements made up 12.58% while *Ty3* elements made up 17.01% of the genome (Appendix A Table A.4). *Amaranthus watsonii*, however, had more *copia* (4.11%) than *Ty3* elements (2.71%), similar to *A. palmeri* (3.46% *copia* and 2.64% *Ty3*). The pattern of LTR composition in the unassembled raw reads of *A. palmeri* is also similar to its genome assembly (9.73% *copia* and 7.79% *Ty3*) (Appendix A Table A.5) and to assembly of other monoecious species, *A. hybridus* (9.32% *copia* and 8.66% *Ty3*; Appendix A Table A.4), *A. cruentus* [13.9% *copia* and 10.5% *Ty3*; Ma et al. (Ma et al. 2021)], or *A. hypochondriacus* [6.93% *copia* and 4.81% *Ty3*; Lightfoot et al. (Lightfoot et al. 2017)]. DnaPipeTE, like Repeatexplorer2, also estimated slightly more total repeats composition for *A. cannabinus* than *A. australis* despite our previous genome size estimation indicating *A. australis* genome is larger than that of *A. cannabinus*. Both species however had the highest genome sizes and highest total TE discovered relative to other dioecious species (Appendix A Table A.3).

2.3.3 Mash-based phylogenomic analysis

Considering the inconsistent tree topologies observed in previous phylogenetic studies of *Amaranthus* genus, and to avoid phylogenetic errors or noise that could result from assembling short reads, we investigated relatedness among the sequenced *Amaranthus* genomes and other members of the order Caryophyllales using an assembly- or alignment-free *k*-mer approach implemented in Mashtree (Katz et al. 2019). As expected, sequenced females from four species included in the tree construction grouped together with their respective males (Figure 2.2). Our analysis of genome relatedness showed species clustering corresponding to the three subgenera:

Acnida, *Amaranthus* and *Albersia* (Figure 2.2), previously recognized based on fruit, bract and tepal characteristics of pistillate flowers (Mosyakin and Robertson 1996). The *Acnida* subgenus, which corresponds to the dioecious species, is split into two separate clades in our Mash-based phylogeny (Figure 2.2), consistent with the split in previous studies (Stetter and Schmid 2017; Waselkov et al. 2018). All dioecious species were placed in one clade, excluding *A. palmeri* and *A. watsonii*, which were placed with monoecious species in the subgenus *Amaranthus*. Although the Dioecious/Pumilus clade in Waselkov et al.'s (Waselkov et al. 2018) nuclear phylogeny is congruent with our Mash-based phylogeny, only the sister-species relationships between *A. australis* and *A. cannabinus* and between *A. palmeri* and *A. watsonii* were supported in our analysis. *Amaranthus tuberculatus* was more closely related to *A. floridanus* than to other dioecious species in our study, similar to Stetter and Schmid (Stetter and Schmid 2017), while *A. arenicola* was more related to *A. greggii*.

The clustering of *A. caudatus*, *A. quitensis*, *A. hybridus*, *A. hypochondriacus* and *A. cruentus* was consistent with previous tree topologies based on chloroplast markers (Waselkov et al. 2018) or biallelic SNPs (Stetter and Schmid 2017). We recovered the same relationships among the five monoecious species reported in Xu and Sun's study (Xu and Sun 2001), which was based on combined AFLP and ISSR datasets. Moreover, the genetic similarity between *A. quitensis* and *A. caudatus* has been suggested to be due to gene flow because the former was often found in *A. caudatus* fields (Sauer 1967; Waselkov et al. 2018).

It is worth mentioning that organellar DNA has been demonstrated not to impact Mash-based phylogeny construction in previous studies, being that their high copy numbers are not represented among low-frequency *k*-mers used in Mash phylogeny (Wascher et al. 2022). Although assembly- or alignment-free *k*-mer-based methods are optimal in analysis of genome

relatedness, they are not without cons in that they are based on assumptions that do not model complex evolutionary processes (Ondov et al. 2016). A single value is computed as distance per pair of species, and therefore conclusions on the contribution of specific genomic regions to species divergence are difficult to obtain. Moreover, low-depth coverage, variation in library sizes or missing data could impact the accuracy of MinHash methods whereby distances deviate from true genetic distances (VanWallendael and Alvarez 2022). While we did not set out to evaluate these sources of bias, we note that *A. hypochondriacus*, *A. caudatus*, *A. quitensis* and *B. vulgaris* short reads from the NCBI database had 96, 98, 100 and 124 bp read lengths, respectively, compared to >130 bp read lengths for other species. Nevertheless, Mash accurately recovered Sauer's taxonomic ordering of the dioecious amaranths (Sauer 1957) as well as the relationships among monoecious species in the subgenus *Amaranthus*, demonstrating the robustness of Mash in our study.

Also intriguing is the relationship between species clustering from our Mash-based phylogeny and the total TE composition from our dnaPipeTE repeats analysis. *Amaranthus cannabinus* and *A. australis* (60.67% and 60.48%, respectively) had a higher total TE composition than *A. tuberculatus* and *A. floridanus* (56.4% and 54.2%, respectively), followed by *A. acanthochiton, A. arenicola* and *A. greggii,* which were all similar in their total TE composition (51.79%, 53.02% and 52.84%, respectively) and *A. watsonii* and *A. palmeri,* which had the least TE compositions (44.03% and 37.02%, respectively).

2.3.4 Whole-sequence mashs and coverage analysis of *Amaranthus palmeri* and *Amaranthus tuberculatus* male-specific regions of the Y

Mapping of Illumina paired-end short reads of sequenced dioecious *Amaranthus* species to draft genomes of both *A. tuberculatus* and *A. palmeri* showed differences in reads alignment

(Appendix A Table A.6 – A.7). As expected, *A. tuberculatus* reads mapped back to its genome assembly had >90% reads in proper pairs (Appendix A Table A.6). Although >90% of *A. palmeri* reads mapped to its genome assembly, only 77% reads were in proper pairs (Appendix A Table A.7). Five species, *A. acanthochiton*, *A. arenicola*, *A. australis*, *A. cannabinus*, and *A. floridanus*, had >70% of paired reads in proper pairs when mapped to *A. tuberculatus* genome while *A. watsonii* had <67% of paired reads in proper pairs (Appendix A Table A.6). However, when the short reads sequences were mapped to *A. palmeri* genome, the five species that mapped well to *A. tuberculatus* had <63% of paired reads in proper pairs, while *A. watsonii* had >75% of its paired reads in proper pairs (Appendix A Table A.7). *Amaranthus greggii*, however, had <66% of its paired reads in proper pairs when mapped to either *A. tuberculatus* or *A. palmeri* draft genomes, perhaps due to its high level of paralogy (discussed above). Structural differences or sequence divergence among the species could have resulted in non-proper pairing of reads for the six genomes when mapped to *A. palmeri* genome. *Amaranthus watsonii*, based on previous phylogenetic studies, including our Mash-based phylogeny, was closely related to *A. palmeri* (Waselkov et al. 2018), which is congruent with our mapping results.

Coverage analysis for sequenced reads mapped to the *A. palmeri* genome revealed male- or female-enriched regions across the genome (Figure 2.3A, Appendix A Table A.8 – A.11). Only *A. watsonii*-mapped reads showed regions with significant spans of male-enriched coverages (Figure 2.3A, Appendix A Table A.11). A total of 84 scaffolds had regions exhibiting male-enriched coverages for *A. watsonii* mapped reads, in which 29 were reported in Neves et al. (Neves et al. 2020) and 13 were reported in Montgomery et al. (Montgomery et al. 2021). It is worth mentioning that all the male-specific scaffolds reported by Montgomery et al. were among the 42 scaffolds reported by Neves et al. The MSY region of *A. palmeri* was previously

identified to span a region of ~1.3 Mb on scaffold 20 (503,282 – 1,770,936 bp), with 121 candidate gene models within the region (Neves et al. 2020; Montgomery et al. 2021). Consistent with the two prior studies, scaffold 20 (MSY region) had the highest window and largest bases spanned for male-enriched coverages in our analysis (Figure 2.3A, Appendix A Table A.11). A total of 101 scaffolds had regions with female-enriched coverages, however, several of the scaffolds that were female-enriched were among those exhibiting male-enriched coverages (Appendix A Table A.11, Appendix A Figure A.25). Interestingly, scaffold 19 exhibited significant spans of female-enrichment. Scaffold 19 is 2.23 Mb in length and contains 115 predicted gene models, including pentatricopeptide repeat-containing protein (PPR), serine/arginine-rich splicing factor, and several proteins of unknown function (Appendix A Figure A.26). It is worth noting that scaffolds with enrichment more than scaffolds 19 or 20 have shorter lengths (<200 kb) relative to both scaffolds. Mapped reads of the other three species from both male and female individuals showed no contiguous region was significantly enriched for male or female coverages (Figure 2.3A). The fact that some *A. watsonii* female reads also mapped within the MSY region on scaffold 20 suggest that the region is not entirely male-specific, and some portions could be part of the pseudo-autosomal region (PAR) that is still recombining with the X chromosome (Figures 2.3B and 2.3C).

We identified 11 sex-linked genes with a combined length of 21,680 bp (~22 Kb) exhibiting male-enriched coverage for *A. watsonii* reads that mapped to the *A. palmeri* MSY region (Appendix A Table A.18). Only three of these genes had informative annotations, one each as pentatricopeptide repeat-containing protein (PPR), serine/arginine-rich splicing factor, and magnesium protoporphyrin IX methyltransferase. A BLAST search of the remaining 8 genes to the non-redundant protein database on NCBI showed two genes, g4825 and g4829, matched to

Zinc finger CCHC-type (*Artemisia annua* L.) and serine/arginine-rich splicing factor (*Arachis hypogea* L.) homologs, respectively, while the remaining 6 genes matched to uncharacterized proteins or had no similarity matches. The *PPR* gene within the sex-determining region was particularly interesting in that six of its seven exons had male-enriched coverages for *A. watsonii* mapped reads, while the three other species had reads from both male and female individuals mapped to the gene (Figure 2.3B).

To identify regions within *A. tuberculatus* genome assembly with male or female-enriched coverages, we included the short reads of female individual of three species, *A. acanthochiton*, *A. cannabinus*, and *A. greggii*, in addition to *A. watsonii*, in that they were farther away from *A. tuberculatus* based on Waselkov et al.'s phylogeny (Waselkov et al. 2018). We reasoned that gene(s) crucial for sex functions should be conserved across species sharing a common dioecy evolutionary event and, therefore, including the most distally related species would identify the most crucial genes. We included previously sequenced short reads of two males and two females of *A. tube*rculatus from Kreiner et al. (Kreiner et al. 2019), which were sequenced to 10x depth.

Among the previously reported MSY contigs, a few were found to exhibit male-enriched coverages for only some species (Figure 2.4A). For example, contig 00001274 had male-enriched coverages for *A. cannabinus*, *A. greggii* and *A. tuberculatus*, contig 0000298 had male-enriched coverages for *A. greggii* and *A. tuberculatus*, and contig 00100752 had male-enriched coverage for *A. cannabinus*, *A. greggii* and *A. tuberculatus*, although variation existed in the length of bases spanned for the coverages (Appendix A Table A.12 – A.17). Only contig 00004323 had male-enriched coverages for all 5 species, while contigs 00000336, 00000340, 00003161, 00004353 and 00100771 were not enriched for either male- or female-specific

coverages for any species. As expected, *A. tuberculatus* had the most significantly enriched contigs (Figure 2.4B) and the highest number of contigs (~300) for both male- and female-enriched regions, while *A. watsonii* mapped reads had the least number of contigs for male- and female-enriched regions (Figure 2.4C). Interestingly, contigs 0000298, 00001274, 00001293 and 00001713, which were previously identified as male-specific, had no female-enriched coverages (Appendix A Table A.12 – A.17).

A 200-bp *FLOWERING LOCUS T* (*FT*) on contig 00000542 identified as one of the MSY genes in Montgomery et al. (Montgomery et al. 2021) was also found to exhibit male-enriched coverages across *A. acanthochiton*, *A. cannabinus*, *A. greggii* and *A. tuberculatus* while reads from *A. watsonii* did not map to the *FT* gene (Appendix A Figure A.27). The gene next to the 200 bp *FT*, although annotated as 'unknown,' also showed male-enriched coverage across the three species, including *A. tuberculatus* (Appendix A Figure A.27) and had its second and longest exon (pos:14302-14525) match to predicted *Beta vulgaris* subsp. *vulgaris Heading date 3a* (LOC104890180) with 84% homology. The 200 bp *FT* also matched to the same *Heading date 3a* locus, but at a different position, and thus we consider this second fragment part of the *FT* gene. In total, there are 4 exons of the *Heading date 3a* in *Beta vulgaris* compared to the two fragments, one with one exon and the other with two exons, in the *A. tuberculatus* contig assembly at this locus.

2.3.5 Transcription factors and repetitive elements within *Amaranthus palmeri* male-specific region of the Y

Transcription factors (TF) have been implicated in sex functions in flowering plants; however, only a few gene models out of the 121 gene models within the *A. palmeri* MSY region had informative annotation. To therefore identify any transcription factors with potential sex

functions among the gene models, a reference plant TF and transcriptional regulator categorization tool, PlantTFcat (Dai et al. 2013) was used for TF prediction. Seven transcription factors from three family types and four families, one of which was *LBD*, were identified (Appendix A Table A.19). The TF families with the highest number of genes predicted from the analysis was *CCHC(Zn)* with 4 genes, followed by one gene each for *ASL-LOB*, BED-type(Zn) and GRF (Appendix A Table A.19). These transcription factors are multifunctional or involved in several processes, including epithelial development, cell adhesion, leaf development or overall plant growth and development (Laity et al. 2001; Majer and Hochholdinger 2011; Kim and Tsukaya 2015).

Additional characterization of the ~1.3 Mb MSY region of *A. palmeri* for transposable elements revealed consistency with a typical sex determination region, with the accumulation of repetitive sequences and the presence of predominantly male-specific sequences (Charlesworth 2013, 2016; Henry et al. 2018). The MSY region was made up of 78.49% repetitive elements (Appendix A Table A.5). The long-interspersed nuclear elements (LINE/*L1*) made up the highest composition at 19.13%, followed by *copia* and *Ty3* at 15.64% and 12.91%, respectively. The proportion of repeats within the MSY region is higher relative to the entire *A. palmeri* genome (56.03%), indicating that this region has indeed accrued repetitive elements during its evolution. The composition of repeats within this region is also consistent with other studies, e.g., 76.9% of the 1.5 Mb *Mercurialis annua* SDR is made up of repeats, and LTRs were most abundant (Veltsos et al. 2018). Similarly, 77% of the 8.1 Mb *Carica papaya* hermaphroditic specific Y region (HSY) is made up of repeats with *Ty3* being most abundant (Na et al. 2014).

2.4 DISCUSSION

We inferred genome characteristics, shallow relationships and gained further understanding of conserved genomic regions with potential roles in sex function among dioecious *Amaranthus* species using comparative genomics. Genome size, repeats proportion, heterozygosity, polyploidy, and GC content are documented genome characteristics that could influence *de novo* assembly quality (Chen et al. 2013; Dominguez Del Angel et al. 2018; Asalone et al. 2020), and thus genome profiling provides valuable consideration towards a high quality assembly. *k*-mer analysis of genome sizes for the dioecious *Amaranthus* species were generally consistent with estimates from flow cytometry for previously reported species (Stetter and Schmid 2017). Heterozygosity estimates differed across species; although such differences might be species specific, they also could reflect differences due to accessions used, and the number of crosses made to propagate the accessions. Ploidy inference analysis also affirms the previously reported diploid state of the species sequenced in this study (Grant 1959). The entire *Amaranthus* genus has been hypothesized to be a paleoallotetraploid (Murray 1940; Grant 1959; Sauer 1967), however, *Amaranthus dubius* Mart. Ex Thell. is the only known extant allotetraploid (2n=64) species, with others being diploid (2n = 32 or 34). Although, diploidy was inferred for *A. greggii*, the higher number of duplicated sequences or paralogs suggests a possible pre- or post-speciation event could have led to the retention of the sequences.

Repeats analysis revealed transposable elements contributing to genome structure differences in dioecious amaranths. The long terminal repeats (LTRs) proliferation and their elimination is the primary mechanism contributing to genome size variation in dioecious *Amaranthus* species. There is a well-established correlation between genome size and LTR element abundance (Janicki et al. 2011; Bennetzen and Wang 2014), however, it is intriguing

that the LTR superfamily *copia* element was more abundant than the *Ty3* element for two dioecious species, *A. palmeri* and *A. watsonii*, similar to the pattern for some monoecious *Amaranthus* species. It is possible that the removal of the LTR elements via ectopic recombination differs between the dioecious and monoecious species (Bennetzen and Wang 2014; Bourque et al. 2018). The mechanistic process involved in such differential LTR removal however remains elusive. The similar TE pattern between *A. palmeri-A. watsonii* and monoecious species is congruent with other studies that have shown some relationships between the two species and the monoecious species (Wassom and Tranel 2005; Riggins et al. 2010; Waselkov et al. 2018). Franssen et al. (65) also suggested that the pollen of *A. palmeri* was less similar to that of the other dioecious *Amaranthus* species sampled (*A. tuberculatus* and *A. arenicola*), and more closely resembled pollen of the monoecious species.

Our complementary repeats discovery methods whereby we analyzed and compared TEs in genome assemblies to TEs from short reads allowed us to identify the abundance of low copy repeats for the amaranths. Various families of transposable elements are known to exist in high copy numbers in the plant genome (Feschotte et al. 2002) and repeat discovery tools could identify these high or medium copy repeats (Novák et al. 2013). However, it is nontrivial to estimate absolute repeat composition of plant genomes using short reads sequences, and methods relying only on raw reads for genomes with low copy repeats return lower TE contents (Treangen and Salzberg 2012; Goerner-Potvin and Bourque 2018). Other factors that could result in TE underestimation include short insert size library, novel or diverged repeats in species of interest relative to the annotation database from other species, and difficulty in detecting nested repeats with short reads. Given the analysis of TEs in genome assemblies from our study and the literature, we hypothesize that the composition of TEs in the amaranths range from 55 – 75% of

the genome. Overall, our findings are congruent with other studies demonstrating the contribution of specific TEs (e.g., LTRs) in genome size variation within a genus, such as in *Oryza* spp. (Zuccolo et al. 2007).

Interestingly, our phylogenomic analysis of genome relatedness appears to be highly consistent with the early taxonomic works of Jonathan D. Sauer on dioecious amaranths based on comparative morphology and the species' geographic distributions (2,3,33). For example, *A. arenicola* is closely related to *A. greggii* based on morphology and their proximity around the tropical Gulf coast (Sauer 1972) while *A. watsonii* and *A. palmeri* share an overlapping range, with the former sometimes confused for *A. palmeri* (Sauer 1955). The sympatry of *A. australis* (southern water hemp) and *A. cannabinus* (eastern water hemp) was also reported, with both species having similar habitat requirements (e.g., salty and fresh water tolerance, and both found in wet sand of coastal marshes) (Sauer 1957). Of keen interest is the relationship between *A. tuberculatus* and *A. floridanus* in our study (Figure 2.2), the former being noxious and expands rapidly while the latter is restricted to Florida. The close relationship between both species was also previously established using biallelic SNPs data in SNAPP (Stetter and Schmid 2017). *Amaranthus tuberculatus* however has been previously suggested to be more related to *A. arenicola* than many other *Amaranthus* species based on morphology (Sauer 1957; Waselkov et al. 2018). The higher number of hybrids between *A. tuberculatus and A. arenicola* and the limited habitat data for *A. floridanus*, as well as limited to no herbarium collections documenting hybrids between *A. tuberculatus* and *A. floridanus* could have led to the suggestion of their relationships (Sauer 1957). Mash-based phylogeny have been shown to be robust in species relationship inference with Wascher et al. (2022) using it to trace the domestication of cultivated sugar beet to wild relatives in Greece. Similarly, Mash recovered accurate cladograms for

polyploid species, where assembly- or alignment-based approaches would have been intractable (VanWallendael and Alvarez 2022).

Furthermore, our analysis identified regions within the *A. palmeri* genome assembly that are male-enriched, congruent with male-specific scaffolds that were previously reported (23,24). We also found scaffold 19 exhibited female-enriched coverages, thus indicating that the scaffold could be part of the X chromosome. Several candidate genes exhibiting male-enriched coverages, including pentatricopeptide repeat-containing protein (PPR) and serine/arginine-rich splicing factor (SC35) were identified within the *A. palmeri* MSY region. Although the genes have no known direct links to sex determination in flowering or dioecious plants, they have been reported to play some roles in sex functions. *PPR*s act as restorers of fertility (*Rf*) i.e., restore partial or normal pollen production to plants via suppression of the cytoplasmic male sterility (CMS) locus (Chen and Liu 2014; Gaborieau et al. 2016). In radish, the *PPR* gene, *Rfo,* was found to restore fertility by specifically downregulating the expression of the CMS locus, *orf138*, in the tapetum of anthers (Uyttewaal et al. 2008). Whether the *PPR* within *A. palmeri* MSY carries any restoration activity, i.e., has a post-transcriptional action on mitochondrial gene expression, is not known. Recently, a *PPR* was reported as one of the SDR genes in the gymnosperm plant, *Gingko biloba* (Gong and Filatov 2022). Sex-linked genes in other dioecious plant species have been shown to exhibit male-specific coverages within the sex-determining regions e.g., the sex-determinant factors, *SOFF* in garden asparagus (*Asparagus officinalis* L.) (Harkess et al. 2017), *SyGI* and *FrBy* in kiwifruit (*Actinidia* sp) (Akagi et al. 2019) and *NRT1/PTR6.4* in spinach (*Spinacea oleracea* L.) (Ma et al. 2022). For *A. tuberculatus* MSY contigs, we identified several contigs with male-enriched coverages, in which four had no female-enriched coverages for all species. However, only a previously identified flowering locus

T (*FT*) gene had male-specific coverage for mapped reads of three species, but not for *A. watsonii*, indicating the conservation of the *FT* gene and its possible role in conferring male fitness as previously hypothesized (Montgomery et al. 2021). The difference in mapping pattern between *A. watsonii* and the three other dioecious species is consistent with our hypothesis of a different dioecy evolutionary event in *A. watsonii* and *A. palmeri* relative to the other dioecious amaranths.

2.4.1 Implications for dioecy evolution within the *Amaranthus* genus

An open question with regards to dioecy within the *Amaranthus* genus has been the evolution of dioecy and the mechanisms involved, and if these could be explained with existing models (Charlesworth and Charlesworth 1978; Standley 1985; Renner and Ricklefs 1995; Renner and Won 2001; Henry et al. 2018). The phylogenetic study of *Amaranthus* from Waselkov et al. (Waselkov et al. 2018) included 58 out of 74 species with all nine dioecious species and is also rooted, providing directionality in ancestry relationship. The Dioecious/Pumilus clade and the Hybridus clade (monoecious) from the study shared a recent common ancestor (Bayesian posterior probability value of 1 and bootstrap support value of 99). Both clades then shared a recent common ancestor with the Galapagos clade, in which all species are monoecious. Although, the nuclear-based and the chloroplast-based trees from the study were discordant, characterized by occasional polytomies, less supported nodes, or poorly resolved clades, dioecy within the genus appeared to have originated from a monoecious ancestor.

The *Amaranthu*s genus has been shown to be closely related to *Chamissoa altissima* (Jacq.) Kunth (Kadereit et al. 2003), which is hermaphroditic (Bullock 1985). A maximum-likelihood phylogeny of Amaranthaceae family constructed from 936-nuclear gene supermatrix also showed monophyly of Amaranthoids and Celosioids (Morales-Briones et al. 2021). The

Amaranthoids are characterized by their unisexual flowers while *Celosia argentea*, a member of the Celosioids, has bisexual flowers, indicating the possibility of a hermaphroditic ancestor in the evolution of the *Amaranthus* genus.

It is unclear how dioecy evolved in the genus, whether via hermaphroditism-gynodioecy/androdioecy-dioecy pathway (Charlesworth and Charlesworth 1978), via monoecy-paradioecy-dioecy pathway (Lloyd 1980; Standley 1985; Renner and Ricklefs 1995; Renner and Won 2001), or via an environmentally/physiologically-induced mechanism (Golenberg and West 2013; Henry et al. 2018; Renner and Müller 2021). The origin of dioecy evolution has implications for what mechanisms could be involved in dioecy. Species evolving dioecy via a hermaphroditism-gynodioecy/androdioecy-dioecy pathway have two sex determinant factors or genes (female suppressor and male activator) that are linked within a region of suppressed recombination on the Y chromosome (MSY or SDR region), which has been observed in *Asparagus officinalis* (Harkess et al. 2017, 2020) and *Actinidia* spp (Akagi et al. 2018, 2019). However, species evolving dioecy from hermaphroditism through monoecious populations could utilize a single gene for sex determination, which has been observed in *Diospyros lotus* (Akagi et al. 2014). The *Amaranthus* genus is made up of 74 species, 9 of which are dioecious while others are monoecious (Sauer 1955, 1972; Bayón and Peláez 2012; Bayón 2015), primarily wind-pollinated (Murray 1940; Bram and Quinn 2000; Trucco et al. 2005), and no evidence of gynodioecy within the genus points to a likely evolution of dioecy from monoecy. The presence of species with bisexual flowers at the subfamily level (Amaranthoideae) suggests monoecy could have arisen from an ancestral hermaphroditic population, giving rise to a hermaphroditism-monoecy-dioecy pathway (Henry et al. 2018; Cronk 2022).

If this is the case, a single gene could thus be sufficient for sex determination in dioecious species of the *Amaranthus* genus. Sex determination in spinach was recently proposed to be controlled by a single gene, *NRT1/PTR6.4* (transporter of nitrate, peptide or hormones), utilizing two pathways for carpel development suppression and stamen initiation (Ma et al. 2022). Although the subfamilies Amaranthoideae (*Amaranthus* genus) and Chenopodoideae (*Spinacia* genus) are in the family Amaranthaceae, comprehensive phylogenetic studies have not shown a convincing support for their relationship (Yao et al. 2019; Morales-Briones et al. 2021). A BLAST search of the spinach *NRT1/PTR6.4* against *A. palmeri, A. tuberculatus* or *A. hypochondriacus* on CoGe (Lyons and Freeling 2008) revealed no orthologs in the amaranths, indicating that *Spinacea* and *Amaranthus* lineage evolved dioecy independently and utilize separate dioecy mechanisms.

Based on our whole-genome analysis of relatedness and other evidence from this study, *A. palmeri* and *A. watsonii* are closely related, and likely utilize a similar dioecy mechanism. The other dioecious species form subclades (e.g., close relationship between *Amaranthus tuberculatus* and *A. floridanus*, *A. cannabinus* and *A. australis* and *A. arenicola* and *A. greggii*) within a larger clade. Whether species within this clade and the *A. palmeri-A. watsonii* cluster evolved dioecy independently but still recruited the same gene(s) or pathways for such independent evolution is unclear (Montalvão Leite et al. 2021; Montgomery et al. 2021). The availability of chromosome-scale reference genome assemblies and genetic maps for the species will allow further characterization of their sex chromosomes.

2.5 CONCLUSION

We report genome characteristics, including size, heterozygosity, and ploidy for seven newly sequenced dioecious species within the *Amaranthus* genus. Although our transposable

element analysis does not capture the full suite of repetitive elements in the respective genomes, it offered a new view of TE dynamics among the dioecious *Amaranthus* species, especially for the species with no high-quality reference or even draft genomes. Furthermore, a pattern of TE proliferation is emerging in the genus, in which some dioecious species have a higher proportion of *Ty3* than *copia* elements, but the reverse is the case for *A. palmeri*, *A. watsonii* and some monoecious species. It is unclear what the 'correct' topology for dioecious species relationship is within the *Amaranthus* genus. Nevertheless, we provide additional evidence supporting early taxonomic relationships among the dioecious *Amaranthus* species based on comparative morphology, i.e., close relationship between *A. palmeri* and *A. watsonii*, *A. australis* and *A. cannabinus*, and *A. tuberculatus* and *A. floridanus*, as well as their relationship to the monoecious species in the subgenus *Amaranthus*. We report 11 gene models, including a pentatricopeptide repeat-containing protein and serine/arginine-rich splicing factor within the *A. palmeri* MSY region that also exhibit male-specific coverages for *A. watsonii*. In addition, a previously reported *FT* within an *A. tuberculatus* MSY contig was found to exhibit male-specific coverage for three species but not for *A. watsonii*. Overall, our findings support the previous hypothesis that dioecy evolved separately in *A. tuberculatus* and *A. palmeri*.

2.6 FIGURES

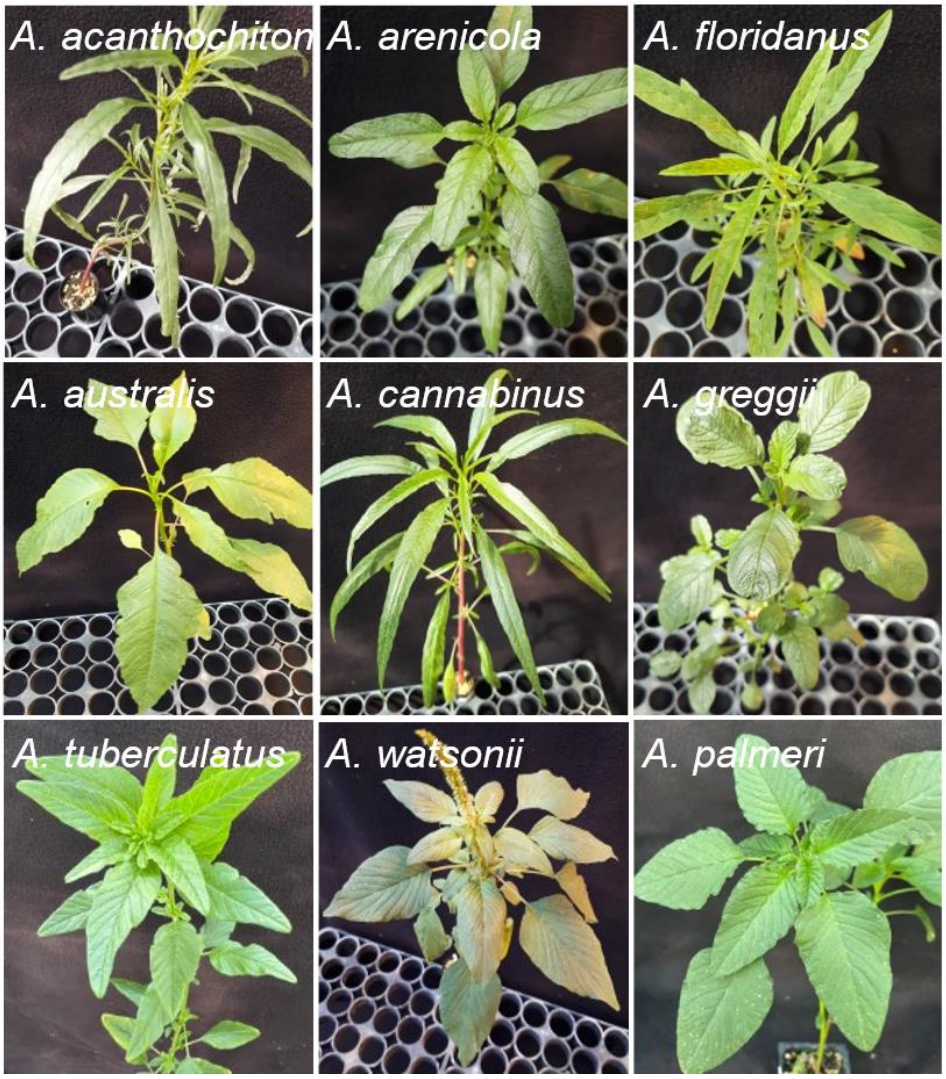

Figure 2.1 Representative individuals of the nine dioecious *Amaranthus* species.

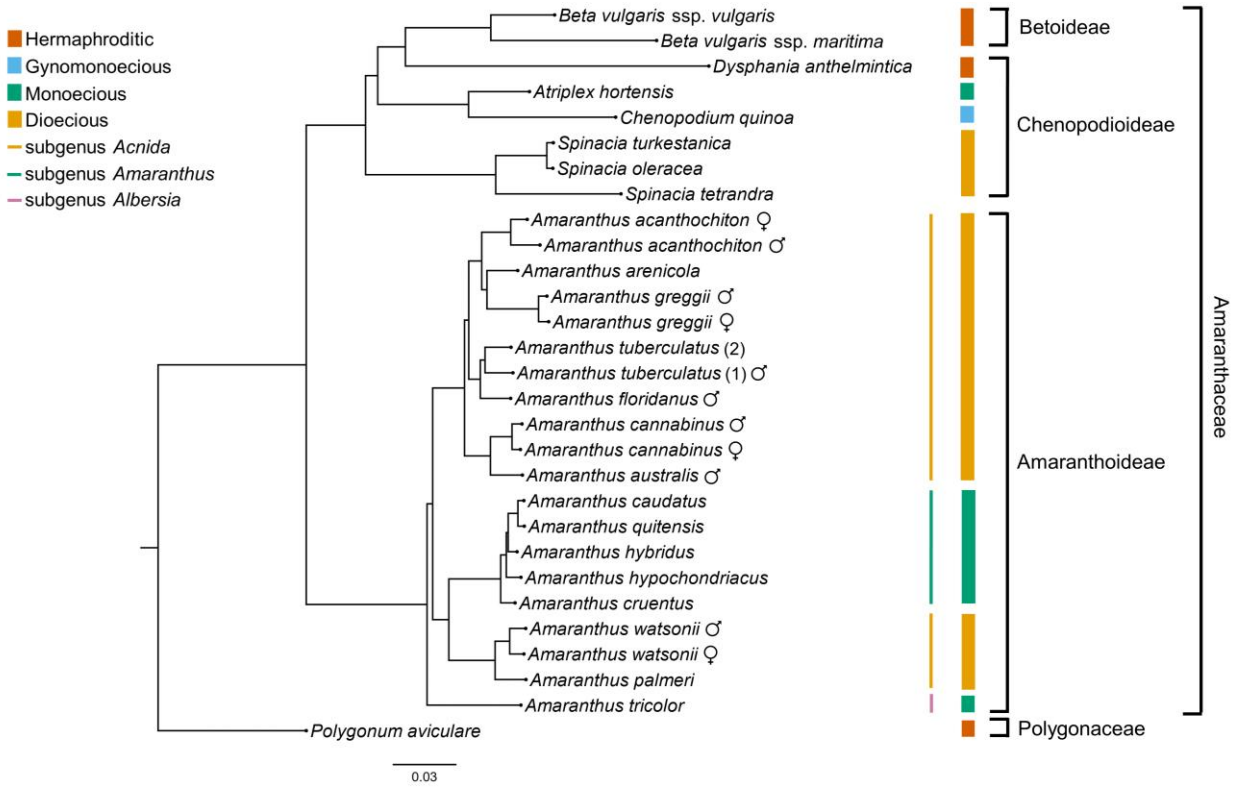

Figure 2.2 A Mash-based phylogeny using Illumina raw reads of dioecious *Amaranthus* species and other species of the Amaranthaceae family. *Polygonum aviculare* was used as outgroup.

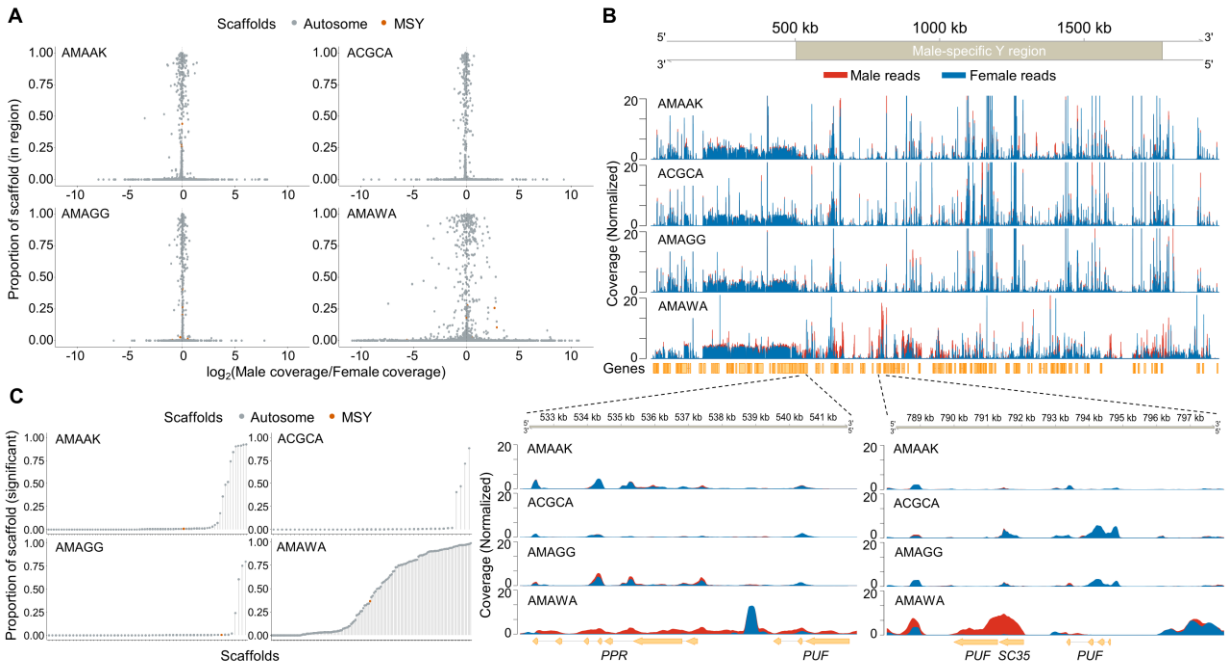

Figure 2.3 Coverage differences between male and female reads of four dioecious *Amaranthus* species mapped to *A. palmeri* scaffold assembly. **A** Analysis of scaffold regions with male- or female-enriched coverages with DifCover pipeline. The y-axis represents the proportion of scaffold the specific region occupies. Orange color is used to indicate regions on the previously identified male-specific region of the Y on scaffold 20. **B** Reads alignment coverage from bamCoverage analysis for scaffold 20. Genes exhibiting male-enriched coverages were visualized within a 10-kb window. **C** All significantly different regions for each scaffold plotted as total proportion of the scaffold length. Species name abbreviations represent the EPPO code for the five dioecious species: AMAAK (*Amaranthus acanthochiton*) ACGCA (*Amaranthus cannabinus*), AMAGG (*Amaranthus greggii*), AMAWA (*Amaranthus watsonii*) and AMATU (*Amaranthus tuberculatus*).

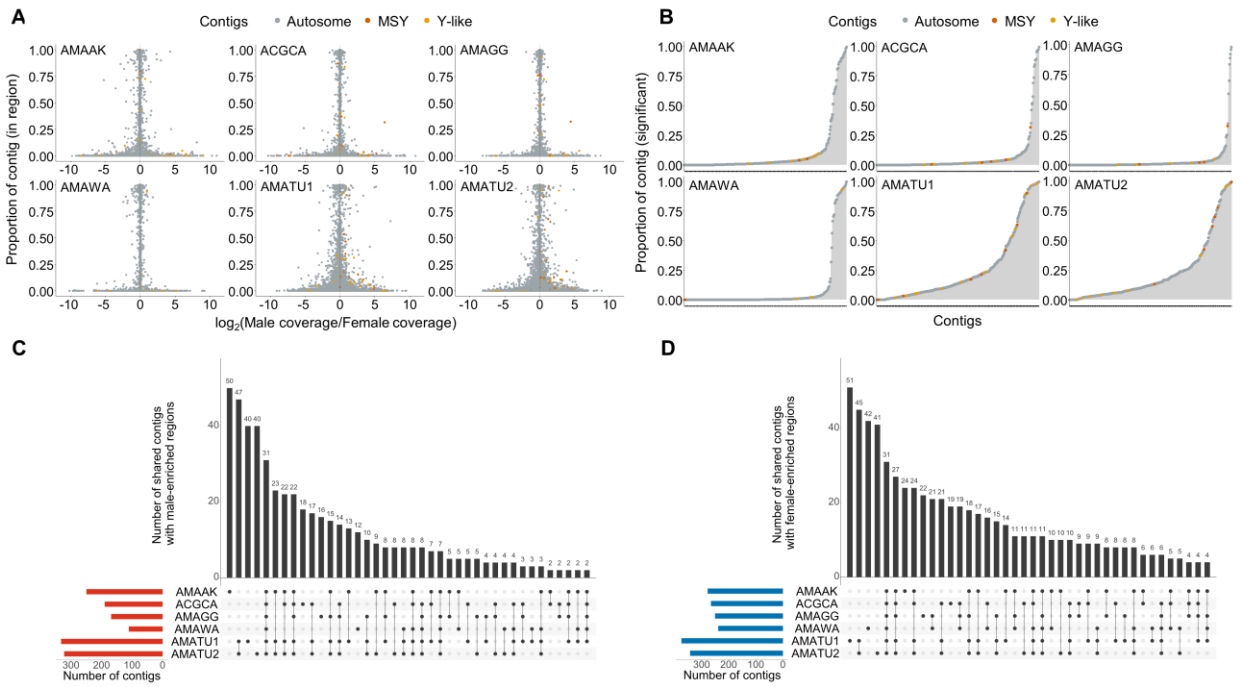

Figure 2.4 Coverage differences between male and female reads of five dioecious *Amaranthus* species mapped to *A. tuberculatus* contig assembly. **A** Analysis of contig regions with male or female-enriched coverages with DifCover pipeline. The y-axis represents the proportion of contig the specific region occupies. Orange color (designated as MSY) is used to indicate regions in the top 10 contigs with both male-specific 15-mer and RAD-tag alignments in Montgomery et al. (Montgomery et al. 2021) while yellow color (designated as Y-like) represents regions in 13 other contigs with either the 15-mer or RAD-tag alignments. **B** All significantly different regions for each contig plotted as total proportion of the contig length. **C,D** Upset plots delineating the number of shared contigs with male or female-enriched coverages. Species name abbreviations represent the EPPO code for the five dioecious species: AMAAK (*Amaranthus acanthochiton*) ACGCA (*Amaranthus cannabinus*), AMAGG (*Amaranthus greggii*), AMAWA (*Amaranthus watsonii*) and AMATU (*Amaranthus tuberculatus*).

2.7 REFERENCES

Aderibigbe OR, Ezekiel OO, Owolade SO, Korese JK, Sturm B, Hensel O (2022) Exploring the potentials of underutilized grain amaranth (*Amaranthus* spp.) along the value chain for food and nutrition security: A review. Crit Rev Food Sci Nutr 62:656–669

Akagi T, Henry IM, Ohtani H, Morimoto T, Beppu K, Kataoka I, Tao R (2018) A Y-encoded suppressor of feminization arose via lineage-specific duplication of a cytokinin response regulator in kiwifruit. Plant Cell 30:780–795

Akagi T, Henry IM, Tao R, Comai L (2014) A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. Science (80-) 346:646

Akagi T, Pilkington SM, Varkonyi-Gasic E, Henry IM, Sugano SS, Sonoda M, Firl A, McNeilage MA, Douglas MJ, Wang T, Rebstock R, Voogd C, Datson P, Allan AC, Beppu K, Kataoka I, Tao R (2019) Two Y-chromosome-encoded genes determine sex in kiwifruit. Nat Plants 5:801–809

Asalone KC, Ryan KM, Yamadi M, Cohen AL, Farmer WG, George DJ, Joppert C, Kim K, Mughal MF, Said R, Toksoz-Exley M, Bisk E, Bracht JR (2020) Regional sequence expansion or collapse in heterozygous genome assemblies. PLoS Comput Biol 16:1–22

Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:4–9

Bayón ND (2015) Revisión taxonómica de las especies monoicas de *Amaranthus* (Amaranthaceae): *Amaranthus* subg. *Amaranthus* y *Amaranthus* subg. Albersia. Ann Missouri Bot Gard 101:261–383

Bayón ND, Peláez C (2012) A new species of *Amaranthus* (Amaranthaceae) from Salta, Argentina. Novon 22:133–136

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65:505–530

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C (2018) Ten things you should know about transposable elements. Genome Biol 19:1–12

Bram MR, Quinn JA (2000) Sex expression, sex-specific traits, and the effects of salinity on growth and reproduction of *Amaranthus cannabinus* (Amaranthaceae), a dioecious annual. Am J Bot 87:1609–1618

Broder AZ (1997) On the resemblance and containment of documents. Proc Int Conf Compression Complex Seq:21–29

Bullock SH (1985) Breeding systems in the flora of a tropical deciduous forest in Mexico. Biotropica 17:287–301

Chan KF, Sun M (1997) Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of *Amaranthus*. Theor Appl Genet 95:865–873

Chaney L, Mangelson R, Ramaraj T, Jellen EN, Maughan PJ (2016) The complete chloroplast genome sequences for four *Amaranthus* species (Amaranthaceae). Appl Plant Sci 4:1600063

Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. Am Nat 112:975–997

Charlesworth D (2013) Plant sex chromosome evolution. J Exp Bot 64:405–420

Charlesworth D (2016) Plant sex chromosomes. Annu Rev Plant Biol 67:397–420

Chen L, Liu YG (2014) Male sterility and fertility restoration in crops. Annu Rev Plant Biol 65:579–606

Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC (2013) Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. PLoS One 8

Costea M, DeMason D (2001) Stem morphology and anatomy in *Amaranthus* L . (Amaranthaceae), taxonomic significance. J Torrey Bot Soc 128:254–281

Cronk Q (2022) The distribution of sexual function in the flowering plant: from monoecy to dioecy. Philos Trans R Soc B Biol Sci 377

Dai X, Sinharoy S, Udvardi M, Zhao PX (2013) PlantTFcat: An online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics 14

Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, Amselem J, Bouri L, Bocs S, Klopp C, Gibrat JF, Vlasova A, Leskosek BL, Soler L, Binzer-Panchal M, Lantz H (2018) Ten steps to get started in genome assembly and annotation. F1000Research 7

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Mol Biol Evol 28:2239–2252

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: Where genetics meets genomics. Nat Rev Genet 3:329–341

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451–9457

Franssen AS, Skinner DZ, Al-Khatib K, Horak MJ (2001) Pollen morphological differences in *Amaranthus* species and interspecific hybrids. Weed Sci 49:732–737

Gaborieau L, Brown GG, Mireau H (2016) The propensity of pentatricopeptide repeat genes to evolve into restorers of cytoplasmic male sterility. Front Plant Sci 7:1816

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

Goerner-Potvin P, Bourque G (2018) Computational tools to unmask transposable elements. Nat Rev Genet 19:688–704

Golenberg EM, West NW (2013) Hormonal interactions and gene regulation can link monoecy and environmental plasticity to the evolution of dioecy in plants. Am J Bot 100:1022–1037

Gong W, Filatov DA (2022) Evolution of the sex-determining region in *Ginkgo biloba*. Philos Trans R Soc B Biol Sci 377

Goubert C, Modolo L, Vieira C, Moro CV, Mavingui P, Boulesteix M (2015) *De novo* assembly and annotation of the Asian tiger mosquito (*Aedesal bopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol 7:1192–1205

Grant WF (1959) Cytogenetic studies in *Amaranthus*. Can J Bot 37:413–417

Grayson P, Wright A, Garroway CJ, Docker MF (2022) SexFindR: A computational workflow to identify young and old sex chromosomes. bioRxiv:1–31

Hahne F, Ivanek R (2016) Visualizing genomic data using Gviz and bioconductor. Methods Mol Biol 1418:335–351

Hall M (2022) Rasusa: Randomly subsample sequencing reads to a specified coverage. J Open Source Softw 7:3941

Harkess A, Huang K, van der Hulst R, Tissen B, Caplan JL, Koppula A, Batish M, Meyers BC, Leebens-Mack J (2020) Sex determination by two Y-linked genes in garden asparagus. Plant Cell 32:1790–1796

Harkess A, Zhou J, Xu C, Bowers JE, Van Der Hulst R, Ayyampalayam S, Mercati F, Riccardi P, McKain MR, Kakrana A, Tang H, Ray J, Groenendijk J, Arikit S, Mathioni SM, Nakano M, Shan H, Telgmann-Rauber A, Kanno A, Yue Z, Chen H, Li W, Chen Y, Xu X, Zhang Y, Luo S, Chen H, Gao J, Mao Z, Pires JC, Luo M, Kudrna D, Wing RA, Meyers BC, Yi K, Kong H, Lavrijsen P, Sunseri F, Falavigna A, Ye Y, Leebens-Mack JH, Chen G (2017) The asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nat Commun 8

Heap IM (2023) The International Survey of Herbicide Resistant Weeds. http://www.weedscience.org/. Accessed December 27, 2022

Henry IM, Akagi T, Tao R, Comai L (2018) One hundred ways to invent the sexes: Theoretical and observed paths to dioecy in plants. Annu Rev Plant Biol 69:553–575

Hnatovska S (2022) Genome size and repeat abundance variation in *Amaranthus tuberculatus*. University of Toronto

Hobza R, Kubat Z, Cegan R, Jesionek W, Vyskot B, Kejnovsky E (2015) Impact of repetitive DNA on sex chromosome evolution in plants. Chromosom Res 23:561–570

Howe K, Bateman A, Durbin R (2002) QuickTree: Building huge neighbour-joining trees of

protein sequences. Bioinformatics 18:1546–1547

Hozza M, Vinař T, Brejová B (2015) How big is that genome? Estimating genome size and coverage from *k*-mer abundance spectra. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 9309:199–209

Janicki M, Rooke R, Yang G (2011) Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosom Res 19:787–808

Kadereit G, Borsch T, Weising K, Freitag H (2003) Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C4 photosynthesis. Int J Plant Sci 164:959–986

Käfer J, Marais GAB, Pannell JR (2017) On the rarity of dioecy in flowering plants. Mol Ecol 26:1225–1241

Katz L, Griswold T, Morrison S, Caravas J, Zhang S, Bakker H, Deng X, Carleton H (2019) Mashtree: a rapid comparison of whole genome sequence files. J Open Source Softw 4:1762

Kim JH, Tsukaya H (2015) Regulation of plant growth and development by the GROWTH-REGULATING FACTOR and GRF-INTERACTING FACTOR duo. J Exp Bot 66:6093–6107

Kokot M, Dlugosz M, Deorowicz S (2017) KMC 3: counting and manipulating *k*-mer statistics. Bioinformatics 33:2759–2761

Kreiner JM, Giacomini DA, Bemm F, Waithaka B, Regalado J, Lanz C, Hildebrandt J, Sikkema PH, Tranel PJ, Weigel D, Stinchcombe JR, Wright SI (2019) Multiple modes of convergent adaptation in the spread of glyphosate-resistant *Amaranthus tuberculatus*. Proc Natl Acad Sci U S A 116:23363

Kreiner JM, Stinchcombe JR, Wright SI (2018) Population genomics of herbicide resistance: adaptation via evolutionary rescue. Annu Rev Plant Biol 69:611–635

Kubatko LS, Chifman J (2019) An invariants-based method for efficient identification of hybrid species from large-scale genomic data. BMC Evol Biol 19:1–13

Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T (2000) The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* L.) reveals a novel gene for tRNA(Cys)(GCA). Nucleic Acids Res 28:2571–2576

Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: New insights into structural and functional diversity. Curr Opin Struct Biol 11:39–46

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Lanoue KZ, Wolf PG, Browning S, Hood EE (1996) Phylogenetic analysis of restriction-site variation in wild and cultivated *Amaranthus* species (Amaranthaceae). Theor Appl Genet 93:722–732

Lawrence M, Gentleman R, Carey V (2009) rtracklayer: An R package for interfacing with genome browsers. Bioinformatics 25:1841–1842

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ (2013) Software for computing and annotating genomic ranges. PLoS Comput Biol 9:1–10

Leimeister CA, Boden M, Horwege S, Lindner S, Morgenstern B (2014) Fast alignment-free sequence comparison using spaced-word frequencies. Bioinformatics 30:1991–1999

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v2 00:1–3

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079

Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein

or nucleotide sequences. Bioinformatics 22:1658–1659

Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. BMC Biol 15:1–15

Lloyd DG (1980) The distributions of gender in four angiosperm species illustrating two evolutionary pathways to dioecy. Evolution 34:123–134

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53:661–673

Ma X, Vaistij FE, Li Y, Jansen van Rensburg WS, Harvey S, Bairu MW, Venter SL, Mavengahama S, Ning Z, Graham IA, Van Deynze A, Van de Peer Y, Denby KJ (2021) A chromosome-level *Amaranthus cruentus* genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop. Plant J 107:613–628

Ma X, Yu L, Fatima M, Wadlington WH, Hulse-Kemp AM, Zhang X, Zhang S, Xu X, Wang J, Huang H, Lin J, Deng B, Liao Z, Yang Z, Ma Y, Tang H, Van Deynze A, Ming R (2022) The spinach YY genome reveals sex chromosome evolution, domestication, and introgression history of the species. Genome Biol 23:1–30

Majer C, Hochholdinger F (2011) Defining the boundaries: Structure and function of LOB domain proteins. Trends Plant Sci 16:47–52

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27:764–770

Molin WT, Wright AA, Lawton-Rauh A, Saski CA (2017) The unique genomic landscape

surrounding the EPSPS gene in glyphosate resistant *Amaranthus palmeri*: A repetitive path to resistance. BMC Genomics 18:1–16

Montalvão Leite P ana, Kersten B, Kim G, Fladung M, Müller NA (2022) ARR17 controls dioecy in *Populus* by repressing B-class MADS-box gene expression. Phil Trans R Soc B 377:20210217

Montalvão Leite PA, Kersten B, Fladung M, Müller NA (2021) The diversity and dynamics of sex determination in dioecious plants. Front Plant Sci 11:1–12

Montgomery JS, Giacomini D, Waithaka B, Lanz C, Murphy BP, Campe R, Lerchl J, Landes A, Gatzmann F, Janssen A, Antonise R, Patterson E, Weigel D, Tranel PJ (2020) Draft genomes of *Amaranthus tuberculatus, Amaranthus hybridus,* and *Amaranthus palmeri*. Genome Biol Evol 12:1988–1993

Montgomery JS, Giacomini DA, Weigel D, Tranel PJ (2021) Male-specific Y-chromosomal regions in waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). New Phytol 229:3522–3533

Montgomery JS, Sadeque A, Giacomini DA, Brown PJ, Tranel PJ (2019) Sex-specific markers for waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). Weed Sci 67:412–418

Morales-Briones DF, Kadereit G, Tefarikis DT, Moore MJ, Smith SA, Brockington SF, Timoneda A, Yim WC, Cushman JC, Yang Y (2021) Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.l. Syst Biol 70:219–235

Mosyakin SL, Robertson KR (1996) New infrageneric taxa and combinations in *Amaranthus* (Amaranthaceae). Ann Bot Fenn 33:275–281

Müller NA, Kersten B, Leite Montalvão AP, Mähler N, Bernhardsson C, Bräutigam K, Carracedo Lorenzo Z, Hoenicka H, Kumar V, Mader M, Pakull B, Robinson KM, Sabatti M, Vettori C, Ingvarsson PK, Cronk Q, Street NR, Fladung M (2020) A single gene underlies the dynamic evolution of poplar sex determination. Nat Plants 6:630–637

Murray MJ (1940) The genetics of sex determination in the family Amaranthaceae. Genetics 25:409–431

Muyle A, Martin H, Zemp N, Mollion M, Gallina S, Tavares R, Silva A, Bataillon T, Widmer A, Glémin S, Touzet P, Marais GAB (2021) Dioecy is associated with high genetic diversity and adaptation rates in the plant genus *Silene*. Mol Biol Evol 38:805–818

Na JK, Wang J, Ming R (2014) Accumulation of interspersed and sex-specific repeats in the non-recombining region of papaya sex chromosomes. BMC Genomics 15:1–12

Neve P (2018) Gene drive systems: do they have a place in agricultural weed management? Pest Manag Sci 74:2671–2679

Neves CJ, Matzrafi M, Thiele M, Lorant A, Mesgaran MB, Stetter MG (2020) Male linked genomic region determines sex in dioecious *Amaranthus palmeri*. J Hered 111:606–612

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:1–12

Novák P, Neumann P, Macas J (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc 15:3745–3776

Novák P, Neumann P, Pech J, Steinhaisl J, MacAs J (2013) RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792–793

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016)

Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biol 17:1–14

Ou S, Chen J, Jiang N (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 46:e126

Ou S, Jiang N (2018) LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol 176:1410–1422

Pflug JM, Holmes VR, Burrus C, Spencer Johnston J, Maddison DR (2020) Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). G3 Genes, Genomes, Genet 10:3047–3060

R Core Team (2021) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T (2014) DeepTools: A flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42:187–191

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11

Renner SS (2014) The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database. Am J Bot 101:1588–1596

Renner SS, Müller NA (2021) Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. Nat Plants 7:392–402

Renner SS, Ricklefs RE (1995) Dioecy and its correlates in the flowering plants. Am J Bot 82:596

Renner SS, Won H (2001) Repeated evolution of dioecy from monoecy in Siparunaceae (Laurales). Syst Biol 50:700–712

Riggins CW, Peng Y, Stewart CN, Tranel PJ (2010) Characterization of *de novo* transcriptome for waterhemp (*Amaranthus tuberculatus*) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes. Pest Manag Sci 66:1042–1052

Sarmashghi S, Bohmann K, Gilbert MTP, Bafna V, Mirarab S (2019) Assembly-free and alignment-free sample identification using genome skims. Genome Biol 20:34

Sauer J (1955) Revision of the dioecious amaranths. Madroño 13:5–46

Sauer J (1957) Recent migration and evolution of the dioecious amaranths. Evolution 11:11–31

Sauer J (1972) The dioecious amaranths: A new species name and major range extensions. Madrono 21:426

Sauer JD (1950) The grain amaranths: A survey of their history and classification. Ann Missouri Bot Gard 37:561–632

Sauer JD (1967) The grain amaranths and their relatives: A revised taxonomic and geographic survey. Ann Missouri Bot Gard 54:103–137

Shergill LS, Barlow BR, Bish MD, Bradley KW (2018) Investigations of 2,4-D and multiple herbicide resistance in a Missouri waterhemp (*Amaranthus tuberculatus*) population. Weed Sci 66:386–394

Smith JJ, Timoshevskaya N, Ye C, Holt C, Keinath MC, Parker HJ, Cook ME, Hess JE, Narum SR, Lamanna F, Kaessmann H, Timoshevskiy VA, Waterbury CKM, Saraceno C, Wiedemann LM, Robb SMC, Baker C, Eichler EE, Hockman D, Sauka-Spengler T, Yandell M, Krumlauf R, Elgar G, Amemiya CT (2018) The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat Genet 50:270–277

Standley LA (1985) Paradioecy and gender ratios in *Carex macrocephala* (Cyperaceae). Am

Midl Nat 113:283–286

Steckel LE (2007) The dioecious *Amaranthus* spp.: Here to stay. Weed Technol 21:567–570

Stetter MG, Schmid KJ (2017) Analysis of phylogenetic relationships and genome size evolution of the *Amaranthus* genus using GBS indicates the ancestors of an ancient crop. Mol Phylogenet Evol 109:80–92

Sun H, DIng J, Piednoël M, Schneeberger K (2018) FindGSE: Estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. Bioinformatics 34:550–557

Timoshevskiy VA, Timoshevskaya NY, Smith JJ (2019) Germline-specific repetitive elements in programmatically eliminated chromosomes of the sea lamprey (*Petromyzon marinus*). Genes 10:832

Tranel PJ (2021) Herbicide resistance in *Amaranthus tuberculatus*†. Pest Manag Sci 77:43–54

Tranel PJ, Riggins CW, Bell MS, Hager AG (2011) Herbicide resistances in *Amaranthus tuberculatus*: A call for new options. J Agric Food Chem 59:5808–5812

Tranel PJ, Trucco F (2009) 21st-century weed science: a call for *Amaranthus* genomics. In: Stewart CN, editor. Weedy and invasive plant genomics. Oxford: Wiley-Blackwell; 2009. p. 53–81

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nat Rev Genet 13:36–46

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

Uyttewaal M, Arnal N, Quadrado M, Martin-Canadell A, Vrielynck N, Hiard S, Gherbi H, Bendahmane A, Budar F, Mireau H (2008) Characterization of *Raphanus sativus*

pentatricopeptide repeat proteins encoded by the fertility restorer locus for Ogura cytoplasmic male sterility. Plant Cell 20:3331–3345

VanWallendael A, Alvarez M (2022) Alignment-free methods for polyploid genomes: Quick and reliable genetic distance estimation. Mol Ecol Resour 22:612–622

Veltsos P, Cossard G, Beaudoing E, Beydon G, Bianchi DS, Roux C, González-Martínez SC, Pannell JR (2018) Size and content of the sex-determining region of the Y chromosome in dioecious *Mercurialis annua*, a plant with homomorphic sex chromosomes. Genes 9

Ward SM, Webster TM, Steckel LE (2013) Palmer amaranth (*Amaranthus palmeri*): A review. Weed Technol 27:12–27

Wascher FL, Stralis-Pavese N, McGrath JM, Schulz B, Himmelbauer H, Dohm JC (2022) Genomic distances reveal relationships of wild and cultivated beets. Nat Commun 13:1–13

Waselkov KE, Boleda AS, Olsen KM (2018) A phylogeny of the genus *Amaranthus* (Amaranthaceae) based on several low-copy nuclear loci and chloroplast regions. Syst Bot 43:439–458

Wassom JJ, Tranel PJ (2005) Amplified fragment length polymorphism-based genetic relationships among weedy *Amaranthus* species. J Hered 96:410–416

Wen D, Yu Y, Zhu J, Nakhleh L (2018) Inferring phylogenetic networks using PhyloNet. Syst Biol 67:735–740

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019) Welcome to the Tidyverse. J Open Source Softw 4:1686

Xu F, Sun M (2001) Comparative analysis of phylogenetic relationships of grain amaranths and

their wild relatives (*Amaranthus*; Amaranthaceae) using internal transcribed spacer, amplified fragment length polymorphism, and double-primer fluorescent intersimple sequence repeat. Mol Phylogenet Evol 21:372–387

Yao G, Jin JJ, Li HT, Yang JB, Mandala VS, Croley M, Mostow R, Douglas NA, Chase MW, Christenhusz MJM, Soltis DE, Soltis PS, Smith SA, Brockington SF, Moore MJ, Yi TS, Li DZ (2019) Plastid phylogenomic insights into the evolution of Caryophyllales. Mol Phylogenet Evol 134:74–86

Zielezinski A, Girgis HZ, Bernard G, Leimeister CA, Tang K, Dencker T, Lau AK, Röhling S, Choi JJ, Waterman MS, Comin M, Kim SH, Vinga S, Almeida JS, Chan CX, James BT, Sun F, Morgenstern B, Karlowski WM (2019) Benchmarking of alignment-free sequence comparison methods. Genome Biol 20:1–18

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim HR, Collura K, Kudrna D, Wing RA (2007) Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. BMC Evol Biol 7:1–15

CHAPTER 3: COMPARATIVE ANALYSIS OF DIOECIOUS *AMARANTHUS* PLASTOMES AND PHYLOGENOMIC IMPLICATIONS WITHIN AMARANTHACEAE S.S.

ABSTRACT

The genus *Amaranthus* L. consists of 70 – 80 species distributed across temperate and tropical regions of the world. Nine species are dioecious and native to North America; two of which are agronomically important weeds of row crops. The genus has been described as taxonomically challenging and relationships among species including the dioecious ones are poorly understood. In this study, we investigated the phylogenetic relationships among the dioecious amaranths and sought to gain insights into plastid tree incongruence. A total of 19 *Amaranthus* species' complete plastomes were analyzed. Among these, seven dioecious *Amaranthus* plastomes were newly sequenced and assembled, an additional two were assembled from previously published short reads sequences and 10 other plastomes were obtained from a public repository (GenBank).

Comparative analysis of the dioecious *Amaranthus* species' plastomes revealed sizes ranged from 150,011 – 150,735 bp and consisted of 112 unique genes (78 protein-coding genes, 30 transfer RNAs and 4 ribosomal RNAs). Maximum likelihood trees, Bayesian inference trees and splits graphs support the monophyly of subgenera *Acnida* (7 dioecious species) and *Amaranthus*; however, the relationship of *A. australis* and *A. cannabinus* to the other dioecious species in *Acnida* could not be established, as it appears a chloroplast capture occurred from the lineage leading to the *Acnida* + *Amaranthus* clades. Our results also revealed intraplastome conflict at some tree branches that were in some cases alleviated with the use of whole chloroplast genome alignment, indicating non-coding regions contribute valuable phylogenetic

signals toward shallow relationship resolution. Furthermore, we report a very low evolutionary distance between *A. palmeri* and *A. watsonii*, indicating that these two species are more genetically related than previously reported.

Our study provides valuable plastome resources as well as a framework for further evolutionary analyses of the entire *Amaranthus* genus as more species are sequenced.

3.1 INTRODUCTION

The genus *Amaranthus* L. consists of 70 – 80 species dispersed across the temperate and tropical regions of the world (Sauer 1967). The genus has been described as taxonomically challenging and species identification could be difficult due to small or inconspicuous reproductive organs (Costea and DeMason 2001; Iamonico 2020; Bayón 2022). Accurate identification of species in the genus thus requires the use of habit, leaf size and shape, fruit type, bracts, bracteoles, and sepals of pistillate flowers. Species in the genus are characterized by their alternate distal leaves and unisexual flowers, which is distinct from closely related genera in the Amaranthaceae family with distal opposite leaves and bisexual flowers (Bayón 2022). The genus is divided into three subgenera, *Amaranthus* subgenus *Amaranthus*, *Amaranthus* subgenus *Albersia* (Kunth) Gren. & Godr. and *Amaranthus* subgenus *Acnida* (L.) Aellen ex K.R. Robertson (Mosyakin and Robertson 1996).

The subgenus *Acnida* is made up of nine dioecious species that are native to North America and is further classified into three sections, *Acnida* sect. *Acnida* (L.) Mosyakin & K.R. Robertson [comprise of *A. australis* (A. Gray) J.D. Sauer, *A. cannabinus* (L.) J.D. Sauer, *A. floridanus* (S. Watson) J.D. Sauer, *A. tuberculatus* (Moq.) J.D. Sauer], *Acnida* sect. *Acanthochiton* (Torr.) Mosyakin & K.R. Robertson [comprises *A. acanthochiton* J.D. Sauer] and *Acnida* sect. *Saueranthus* Mosyakin & K.R. Robertson [comprise of *A. arenicola* I.M. Johnson,

A. greggii S. Watson, *A. watsonii* Standley, and *A. palmeri* S. Watson] (Sauer 1955, 1957, 1972; Mosyakin and Robertson 1996; Steckel 2007). The infrageneric classification above was based on combinations of morphological characteristics: dehiscent or indehiscent fruits, presence/absence of foliaceous bracts, presence/absence of tepals of pistillate flowers, shape of the tepals and whether they are well developed or not (Sauer 1955, 1957; Mosyakin and Robertson 1996).

Several species within the *Amaranthus* genus are economically important in that they offer nutritional benefits and are either grown for their grains (e.g., *A. hypochondriacus* L., *A. cruentus* L. and *A. caudatus* L.) or as leafy vegetables in parts of Asia and Africa (e.g., *A. tricolor* L., *A. blitum* L. and *A. dubius* L.) (Sauer 1950; Riggins and Mumm 2021; Aderibigbe et al. 2022; Sarker et al. 2022). However, twenty species are widespread as weeds of crop lands and non-agrarian areas around the world, with *A. tuberculatus* and *A. palmeri* being particularly troublesome due to their rapid adaptability to changing climatic conditions, management strategies and herbicide management (Ward et al. 2013; Riggins and Mumm 2021; Tranel 2021). Investigation of species' relationships within the genus could enable better comprehension of trait evolution (e.g., weediness).

Previous studies investigating the relationships among the amaranths have utilized either plastid DNA markers (e.g., *matK*, *trnL*), nuclear ribosomal internal transcribed spacer (ITS), low-copy nuclear genes (e.g., *Waxy*, *A36*), nuclear markers (e.g., *ALS*, AFLP), biallelic single nucleotide polymorphisms or chloroplast genomes (Xu and Sun 2001; Wassom and Tranel 2005; Riggins et al. 2010; Stetter and Schmid 2017; Waselkov et al. 2018; Xu et al. 2022). Waselkov et al. (2018) in their phylogenetic studies reported partial support for the infrageneric classification of Mosyakin and Robertson (1996), with grouping of some species corresponding to the three

subgenera. It was however noted that the infrageneric taxa may not reflect the evolutionary history of species in the genus (Mosyakin and Robertson 2003; Waselkov et al. 2018). Moreover, many of the previous phylogenetic studies have either sequenced and assembled chloroplast genomes as genomic resource and sampled very few dioecious species or used few markers for tree construction. Neither strategy has offered convincing support for the relationship between the dioecious *Amaranthus* species.

Chloroplast genomes provide an advantage in inferring evolutionary relationships among species because they are highly conserved with stable gene content, gene order and overall lower substitution rates relative to nuclear genomes (Duchene and Bromham 2013; Smith 2015). They have a typical quadripartite structure consisting of a large single copy region (LSC), a small single copy region (SSC) and a pair of inverted repeats (IRs) with small sizes ranging from 115 – 165 Kb for most photosynthetic organisms (Howe et al. 2003; Jansen et al. 2005; Dobrogojski et al. 2020). Although methods including plastid DNA enrichments and bacterial artificial chromosome (BAC) were earlier proposed to obtain chloroplast genomes from plants (Jansen et al. 2005), advances in genome sequencing, bioinformatics and phylogenomic methods have simplified the acquisition of chloroplast genomes using next-generation sequencing as well as their subsequent analysis (McPherson et al. 2013; Twyford and Ness 2017; Wang et al. 2018). Complete chloroplast genomes thus possess more parsimony-informative sites and, in many cases, provide better resolution in deciphering species relationships than do a few molecular markers (Wambugu et al. 2015; Song et al. 2020; Zhao et al. 2021).

There are about 23 *Amaranthus* species' plastomes available in public repositories; some with incomplete annotations and others remain unverified after author's submission [NCBI GenBank database (Sayers et al. 2020), accessed on July 7, 2022]. The low number of available

chloroplast sequences for species in the *Amaranthus* genus is thus insufficient. In this study, we report the complete chloroplast sequence data for the nine dioecious species of the *Amaranthus* genus. The objectives of this study are to 1) investigate the structural organization of plastomes of dioecious *Amaranthus* species, 2) identify divergence hotspots that could be useful in species delimitation or development of barcoding markers and 3) provide a comprehensive plastid-based phylogenetic resource for comparison with tree topologies that are derived from nuclear genomes. In addition to seven newly sequenced and assembled plastomes of dioecious *Amaranthus* species, we further assembled plastomes from previously reported short reads of species in the family Amaranthaceae s.s. for comparative analyses.

3.2 MATERIALS AND METHODS

3.2.1 Plant material, DNA extraction and Illumina sequencing

Seeds of seven dioecious species of the *Amaranthus* genus (*A. acanthochiton*, *A. arenicola*, *A. australis*, *A. cannabinus*, *A. floridanus*, *A. greggii* and *A. watsonii*) were obtained from USDA Germplasm Resources Information Network (GRIN). Voucher specimens of the accessions grown and sequenced have been deposited at the Illinois Natural History Survey (ILLS) Herbarium at the University of Illinois Robert A. Evers Laboratory (Appendix B Table B.1). The DNA extraction and sequencing procedure have been described previously (Raiyemo et al. 2023). Briefly, seeds were grown in containers with a mixture of Sunshine LC1 (Sun Gro Horticulture, 770 Silver Street Agawam, MA) growing mix, soil, peat, and torpedo sand (3:1:1:1 by weight). Two or three young fresh leaves were harvested from each species, flash frozen in liquid nitrogen and stored at -80^{o}C. Genomic DNA was extracted following standard CTAB protocol (Doyle and Doyle 1990), and DNA integrity was determined using a spectrophotometer (Nanodrop1000 Spectrophotometer, Thermo Fisher Scientific, 81 Wyman Street, Waltham, MA

02451). The DNA samples were submitted to the Roy J. Carver Biotechnology Center at the University of Illinois, Urbana–Champaign for paired-end sequencing (2 x 150 bp) on Illumina NovaSeq6000. Other chloroplast genome assemblies or raw reads of species belonging to the family Amaranthaceae s.s. used in this study were downloaded from the NCBI database and are described further in supplementary file (Appendix B Table B.2).

3.2.2 Genome assembly and annotation

Quality of the sequenced raw reads and those from the NCBI database was evaluated with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and aggregated with MultiQC v1.5 (Ewels et al. 2016). Low quality bases and adapters were removed with Trimmomatic (Bolger et al. 2014) using parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36. The complete chloroplast genomes for the dioecious *Amaranthus* species including other species from the NCBI database were *de novo* assembled with GetOrganelle v1.7.6.1 (Jin et al. 2020) using the default parameters, except -R 45. All *Amaranthus* species' assemblies were seeded with *A. hypochondriacus* reference cp genome (GenBank accession number KX279888). Assembly graphs were visualized with Bandage (Wick et al. 2015), and synteny plots generated with MUMmer (Marçais et al. 2018) were used to confirm that each assembly had the same SSC orientation as the reference chloroplast genome used to seed the assembly. All assembled chloroplast genomes were then annotated with GeSeq (Tillich et al. 2017). Annotation steps included the use of the following: BLAT search, ARAGORN v1.2.38, and MPI-MP chloroplast reference set along with the default settings (Tillich et al. 2017). The annotations were further verified with additional tools, tRNAscan-SE v2.0.7 within GeSeq and a standalone plastid annotation pipeline, Chloe v0.1.0 (https://chloe.plastid.org/annotate.html). Visualization of the

chloroplast genome annotation was carried out with the program OGDRAW (Greiner et al. 2019).

3.2.3 Analysis of simple sequence repeats (SSRs), repetitive sequences and codon usage bias

Microsatellites or simple sequence repeats from the chloroplast genomes were identified with MISA v2.1 (https://webblast.ipk-gatersleben.de/misa/) using the following search parameters: 12, 6, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats, respectively (Beier et al. 2017). Repetitive sequences, including forward, palindromic, reverse, or complementary repeats in the genomes were detected with REPuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer) using a minimal repeat size set to 30 bp and a hamming distance of 3 (Kurtz et al. 2001). Codon usage and relative synonymous codon usage (RSCU) were evaluated with CodonW v1.4.4 (Peden 1999).

3.2.4 Comparison of dioecious *Amaranthus* chloroplast genomes

The assembled chloroplast genomes of the nine dioecious *Amaranthus* species were compared to the reference chloroplast genome of *A. hypochondriacus* with mVISTA (https://genome.lbl.gov/vista/mvista/submit.shtml) using the shuffle-LAGAN mode (Brudno et al. 2003). Comparison of boundaries between the LSC, IR and SSC regions (i.e., LSC/IRb/SSC/IRa) among the chloroplast genomes were carried out with IRSCOPE (https://irscope.shinyapps.io/irapp/) (Amiryousefi et al. 2018). To avoid data duplication, the IRa region was removed from each of the plastomes prior to alignment. The plastome sequences were then aligned using the FFT-NS-2 method in MAFFT v7.5 (Katoh et al. 2002; Katoh and Standley 2013). The alignment of the nine dioecious *Amaranthus* species was then used to determine the values of nucleotide variability (π) (Nei and Li 1979). Nucleotide variability values were also calculated separately for the alignment of four weedy species (*A. tuberculatus*,

A. palmeri, *A. hybridus* and *A. retroflexus*). Sliding window analyses were carried out with DnaSp v6.12 (Rozas et al. 2017) using a window length of 800 bp and a step size of 200 bp.

3.2.5 Phylogenetic analysis

Thirty plastomes belonging to Amaranthaceae s.s, including the newly assembled nine of the dioecious *Amaranthus* species, were used for phylogenetic analyses (Appendix B Table B.2 – B.3). Three species in the family Achatocarpaceae were included as outgroups. Our phylogenetic analyses were focused on understanding the relationship between the dioecious *Amaranthus* species, and therefore did not include other members of the Amaranthaceae s.l. Phylogenetic analyses were carried out using two datasets: 1) seventy-eight protein-coding sequences (CDS) extracted from the cp assemblies and 2) whole chloroplast genomes with IRa removed. All datasets were aligned with MAFFT v7.5 (Katoh et al. 2002; Katoh and Standley 2013) using the FFT-NS-2 method. The alignments were visually inspected and columns with less than 50% occupancy were removed in Jalview v2.11.2.4 (Waterhouse et al. 2009). Alignment statistics were then assessed with MEGA11 (Tamura et al. 2021).

For the concatenated 78 protein-coding sequences, the analyses were carried out with a partitioning scheme – allowing substitution patterns to vary across genes. A Maximum Likelihood (ML) tree implemented in RAxML v8.2.12 (Stamatakis 2014) was carried out with the alignment using the GTRGAMMA substitution model and 1,000 rapid bootstrap replicates. The degree of conflict on each node given the individual gene trees was assessed via the internode certainty all (ICA) which was calculated in RAxML using the extended majority rule consensus tree (Salichos et al. 2014). In addition, Quartet Sampling (Pease et al. 2018) with 1,000 replicates was carried out to differentiate between strong conflict and weak branch

support. The ML bootstrap trees from RAxML were also used to estimate species tree in ASTRAL-III (Zhang et al. 2018).

We complemented our analysis in RAxML by further implementing another ML tree in IQ-TREE v2.1.2 (Minh et al. 2020b), first without partitioning and second with the previous partitioning scheme used, but allowing an optimal model to be determined by ModelFinder (Kalyaanamoorthy et al. 2017). Topology tests between the partitioned and unpartitioned tree was assessed with the approximately unbiased (AU) test (Shimodaira 2002). Concordance factors between gene trees and species trees were calculated in IQ-TREE (Minh et al. 2020). Additionally, conflicting and concordant bipartitions among gene trees were calculated in Phyparts (Smith et al. 2015).

Bayesian inference (BI) analyses was carried out with MrBayes v3.2.7 (Ronquist et al. 2012) following the partitioning scheme adopted for RAxML. The Markov chain Monte Carlo (MCMC) analyses consisted of two independent runs and four heated chains of 20 million generations each, sampling every 1,000 generations using a GTR + G model and a 25% burn-in. The parameters for each partition were unlinked. Convergence of parameter estimates was first assessed by inspecting the average standard deviation of split frequencies in MrBayes, followed by further assessment using Tracer v1.7.2 (Rambaut et al. 2018).

For the plastome alignment, ambiguously aligned regions with < 50% occupancy were also inspected and removed from the sequence alignment in Jalview. A ML tree with the optimal model, TVM + F + R2, suggested by ModelFinder was then implemented in IQ-TREE 2 on the alignment without data partitioning. For Bayesian inference phylogeny, the GTR + I + G substitution model was used on the unpartitioned dataset. The Markov chain Monte Carlo (MCMC) analyses consisted of two independent runs and four heated chains of 6 million

generations each, sampling every 1,000 generations and a 25% burn-in. Parameter convergence was evaluated as previously described. All tree files were visualized and edited in FigTree v1.4.4 (https://github.com/rambaut/figtree) and Dendroscope v3.8.3 (Huson and Scornavacca 2012).

Since bifurcating trees may sometimes be inadequate in depicting the relationships between taxa with reticulation events (Bapteste et al. 2013; Schliep et al. 2017), we further evaluated the relationship among the dioecious *Amaranthus* species with a tree-based bootstrap consensus network that maps bipartition frequencies (e.g., from RAxML bootstrap trees) onto network edges and a distance-based Neighbor-Net algorithm (Bryant and Moulton 2004) that uses uncorrected *p*-distances in SplitsTree v4.18.3 (Huson 1998; Kloepper and Huson 2008).

We assessed the monophyly of dioecious *Amaranthus* species by constraining all dioecious species to be in one clade following our previous analysis and model in RAxML. Testing the monophyletic dioecious amaranths hypothesis was informed by the observed paraphyly between *A. australis*, *A. cannabinus* and the other seven dioecious species. The per site log-likelihoods of both the unconstrained and constrained trees were computed in RAxML, and used for an approximately unbiased (AU) test in CONSEL v1.20 (Shimodaira and Hasegawa 2002).

3.2.6 Evolutionary distance between the two dioecious species, *A. palmeri* and *A. watsonii*

Amaranthus palmeri and *A. watsonii* are two dioecious species with very similar morphological characteristics and exhibited sister relationships in previous phylogenies (Waselkov et al. 2018). To understand the relationship between both species, we used the whole plastome alignment (minus IRa) as input for MEGA11 (Tamura et al. 2021) to calculate evolutionary distances (uncorrected *p*-distances). Additionally, we assembled the nuclear ribosomal DNA (rDNA) genes, 18S (small subunit, SSU), 5.8S, 26S (large subunit, LSU) and

their internal transcribed spacers, ITS1 and ITS2 from short reads sequences of the dioecious species with GetOrganelle v1.7.6.1 (Jin et al. 2020). Each of the rDNA genes were identified from the assembly using Rfam 14.8 [(Wheeler and Eddy 2013; Kalvari et al. 2021); http://rfam.xfam.org/] and the ITS regions were further verified with the tool, ITSx (Bengtsson-Palme et al. 2013). Both the complete ITS region (18S-ITS1-5.8S-ITS2-26S) and the full rDNA were then aligned using MAFFT. To reduce assembly artifacts due to the difficulty in assembling externally transcribed spacer (ETS) and intergenic spacer (IGS) from short reads, we removed columns with <50% occupancy from the full rDNA alignment. Evolutionary distances were then calculated as previously described.

3.3 RESULTS

3.3.1 Characteristics of the dioecious *Amaranthus* chloroplast

Raw reads data from which seven dioecious *Amaranthus* chloroplast genomes were assembled are available under the NCBI Sequence Read Archive (SRA) project number PRJNA836903 while information on the other two dioecious species is provided in the supplementary file (Appendix B Table B.2). The assembled chloroplast genomes of the nine dioecious *Amaranthus* species ranged from 150,011 bp (*A. australis*) to 150,735 bp (*A. greggii*). The genomes have a typical quadripartite structure consisting of a large single copy (LSC) region (83,244 – 83,986 bp), and a small single copy (SSC) region (18,026 – 18,088 bp), separated by two inverted repeat (IR) regions (24,346 – 24,352 bp) (Figure 3.1, Table 3.1). The average GC content for the nine genomes ranged from 36.56 (*A. cannabinus*) to 36.62 (*A. australis*) (Table 3.1). The genomes contained 133 genes including 88 protein-coding genes, 37 tRNA genes and 8 rRNA genes. The LSC region contained 83 genes out of which 61 were protein-coding and 22 were tRNAs, while the SSC region contained 11 protein-coding genes and 1 tRNA. The IR

region (IRb) contained 17 genes (6 protein-coding, 7 tRNAs and 4 rRNAs) and a *ycf1* fragment while IRa also had the 17 genes present in IRb and an *rps19* fragment. The partial fragments of both *ycf1* and *rps19* in the *Amaranthus* chloroplast genomes are consistent with previous reports for chloroplast genomes that have suggested the pseudogenization of both genes (Huang et al. 2013; Hu et al. 2015; Gonçalves et al. 2019). There were 17 distinct genes (*ndhB*, *petB*, *petD*, *atpF*, *clpP1*, *ndhA*, *rpl16*, *rpoC1*, *rps12*, *rps16*, *pafI*, $trnG^{UCC}$, $trnI^{GAU}$, $trnL^{UAA}$, $trnA^{UGC}$, $trnK^{UUU}$, $trnV^{UAC}$) with introns, in which 3 (*rps12*, *clpP1* and *ycf3*) had two introns. The gene $trnK^{UUU}$ had the longest intron at 2,586 bp. Overall, 78 protein-coding genes, 30 tRNA genes and 4 rRNA genes, making a total of 112 genes, represent the unique genes found in the chloroplast genomes of dioecious *Amaranthus* species (Table 3.1). Although Geseq annotated the gene *rpl23* in the genomes, Chloe did not annotate this gene. Previous studies have reported the pseudogenization of *rpl23* in the order Caryophyllales and several angiosperm taxa (Wicke et al. 2011; Yao et al. 2019). We therefore did not consider it further in subsequent analysis.

3.3.2 Simple sequence repeats (SSRs), repetitive sequences and codon usage bias patterns

Simple sequence repeats in the chloroplast genomes of the nine dioecious *Amaranthus* species ranged from 31 (*A. acanthochiton*) to 37 (*A. cannabinus*), of which the mononucleotides (12 – 17) and tetranucleotides (10 – 14) repeats were most abundant. All nine species had one hexanucleotide SSR while only *A. cannabinus* had one pentanucleotide repeat (Table 3.2). Composition of repetitive sequence types across the species ranged from 36 in four species (*A. acanthochiton*, *A. cannabinus*, *A. watsonii* and *A. palmeri*) to 39 in *A. greggii*. Forward and palindromic repeats across the species ranged from 14 – 16 and 21 – 23, respectively. One reverse repeat was identified in all species except *A. acanthochiton*, *A. australis* and *A.*

cannabinus, which had none. No complementary repeat was detected in any of the nine species at the threshold used to find the repeats (Table 3.3).

Codon usage frequency is believed to differ across genomes or among genes, and codons that are optimal are important for efficient and accurate translation (Akashi and Eyre-Walker 1998; Hershberg and Petrov 2008; Frumkin et al. 2018). The codon usage and relative synonymous codon usage (RSCU) for the *A. tuberculatus* chloroplast genome was calculated based on 78 protein-coding sequences in the genome (61 within the LSC, 6 within IR and 11 within the SSC regions). The 78 protein-coding genes were encoded by 21,260 codons, excluding stop codons (Appendix B Table B.5). Codons with the third position nucleotide of A or T were used more often than codons ending with G or C. The most common amino acid codon in the *A. tuberculatus* cp genome was leucine at 2,233 codons (10.5%), while the least frequent was cysteine at 665 codons (3.12%) (Appendix B Table B.5).

3.3.3 Comparative analysis of dioecious *Amaranthus* chloroplast genome structure

Pairwise comparison of sequence divergence across the nine dioecious *Amaranthus* species and the reference *A. hypochondriacus* chloroplast genome using mVISTA revealed highly conserved coding regions while the non-coding regions were more divergent (Figure 3.2). Although, the intergenic region, *psaA-ycf3* appears to be more conserved across six species, it appears to be less conserved across *A. arenicola*, *A. floridanus* and *A. tuberculatus*. The intergenic region, *psbM-trnD*^{GUC} also showed a high divergence in *A. australis*. Other intergenic regions such as *rpl32-trnL*^{UAG}, *trnK*^{UUU}*-rps16*, *trnS*^{GCU}*-trnG*^{UCC}, and *ndhE-ndhG* also exhibited variations relative to the reference. These intergenic spacer regions have been reported to be variable in other plant species and hold valuable phylogenetic signals for resolving species relationship (Lee and Wen 2004; Yamane et al. 2006; Spalik et al. 2009; Dong et al. 2012; Liu et

al. 2017). Analysis of the LSC/IRb/SSC/IRa boundaries showed that *rps19* is located at the boundary of LSC/IRb with 119 bp of its length within the LSC region and 160 bp of its length within IRb region, while *ycf1* is located at the SSC/IRa boundary with 4,008 bp of its length within the SSC region and 1,387 bp of its length within the IRa region (Figure 3.3). Contraction and expansion of IR regions contributes to size variation and rearrangement of the LSC/IRb/SSC/IRa boundaries in angiosperms (Wang et al. 2008). However, there were no differences between the LSC/IRb, IRb/SSC, and SSC/IRa boundaries across the nine dioecious *Amaranthus* species in our study (Figure 3.3). Thirteen mutational hotspots (9 in LSC, 3 in SSC and 1 in IR regions) exhibited nucleotide diversity, π, greater than 0.006 when comparing the nine dioecious species (Figure 3.4A) while ten hotspots (7 in LSC and 3 in IR regions) exhibited π greater than 0.008 when comparing four weedy *Amaranthus* species (Figure 3.4B). Across the 19 *Amaranthus* species with available plastome sequences, twelve hotspots exhibited π greater than 0.008 (Appendix B Figure B.1). The overall low nucleotide variability among the *Amaranthus* species indicates high level of sequence conservation.

3.3.4 Phylogenetic analysis

There were 58,259 conserved sites, 9,073 variable sites and 7,203 parsimony-informative sites in a total of 67,333 alignments for the concatenated 78 protein-coding genes. Maximum likelihood and Bayesian inference phylogeny revealed high support for many branches on the tree, including the additional taxa belonging to 8 other genera in Amaranthaceae s.s., with bootstrap support values close to 100 and posterior probabilities close to 1. We recovered the monophyly of the subgenera *Acnida* (dioecious species) and *Amaranthus* (monoecious species), which corresponds to previously reported classification based on morphology (Figure 3.5) (Mosyakin and Robertson 1996; Costea and DeMason 2001; Waselkov et al. 2018). Seven

dioecious species (*A. tuberculatus*, *A. floridanus*, *A. arenicola*, *A. watsonii*, *A. palmeri*, *A. acanthochiton*, and *A. greggii*) within the subgenus *Acnida* formed a monophyletic group with full support (BS = 100, PP = 1, ICA = 1.00). Within this clade, the relationship of *A. tuberculatus* to *A. floridanus* was less supported (BS = 54, ICA = 0.11) although both species were sister to *A. arenicola*. Two other dioecious species, *A. australi*s and *A. cannabinus* formed a monophyletic clade but were less supported in their relationship with the *Acnida* + *Amaranthus* clades (BS=56, PP=0.77).

The low ICA scores, 0.01 and 0.09, for the branch leading to a common ancestor between *A. australis*, *A. cannabinus*, and *Acnida* + *Amaranthus* clades, and the branch leading to *A. quitensis*, *A. dubius*, *A. hypochondriacus* and *A. caudatus*, respectively, indicates that the two most prevalent conflicting bipartitions have almost similar or at least close frequency of support (Figure 3.5). Bootstrap consensus network also revealed that while 55.8% support the first bipartition leading to a common ancestor between *A. australis*, *A. cannabinus* and *Acnida* + *Amaranthus* clades, 43.5% support the second bipartition or branch leading to *A. australis*, *A. cannabinus* and species in the *Albersia* subgenus (Figure 3.6). Similarly, 54.4% support the first bipartition or branch leading to *A. floridanus* and *A. tuberculatus* while 30% support the second bipartition or branch leading to *A. arenicola* and *A. tuberculatus* (Figure 3.6). Although, NeighborNet fit for the 78 CDS was 99.185%, indicating that the data is tree-like or bifurcating, the incongruence among the tree described above was further confirmed in the splits graph, thus corroborating the bootstrap consensus network (Figure 3.7).

Quartet Concordance (QC), Quartet Differential (QD) and Quartet Informativeness (QI) (collectively referred to as Quartet internodal score) indicate strong or perfect support for many of the tree branches i.e., 1/-/1 (Appendix B Figure B.2); however, the branch leading to *A.*

floridanus and *A. tuberculatus* had a low QI score (0.067), similar to the branch leading to the common ancestor between *A. floridanus, A. tuberculatus, A. arenicola, A. watsonii* and *A. palmeri* (QI = 0.18), an indication of low information for the branches. The relationship between some species in the subgenus *Amaranthus* also appears to be weak with QC scores ranging from 0.068 – 0.51, QD scores from 0 – 0.52, and QI scores from 0.36 – 0.97. A low score for the three measures reflects a weak consensus relationship among species, possibility of competing alternative history or presence of a supported secondary evolutionary history, perhaps due to introgressive gene flow, and in some cases low information for branches. The relationship between *A. australis, A. cannabinus* and other dioecious *Amaranthus* spp. based on ICA was not clear as evidenced in the counter-support for the branch leading to a common ancestor between the two species and the *Acnida* + *Amaranthus* clades (QC = -0.43, QD = 0.045). Overall, there was full support along the backbone relating the *Acnida* clade (seven dioecious species) and the *Amaranthus* clade (Appendix B Figure B.2). Quartet Fidelity (QF) scores for the 33 taxa ranged from 0.6 – 0.94, indicating that many of the taxa sampled in this study were not misplaced (a term sometimes referred to as "rogue" taxa) (Appendix B Figure B.2).

Approximately unbiased (AU) test to determine if there is significant difference between trees with or without partitioning revealed both approaches were not significantly different ($p > 0.5$), therefore, results of the partitioned tree in IQTREE are presented and discussed. The topology and support for the tree generated in IQTREE adopting an optimal model was similar to the tree from RAxML (Appendix B Figure B.3). Although many branches had high support, the gene concordance factor (gCF) and site concordance factor (sCF) values corroborate the discordance or conflicts among branches earlier reported (Appendix B Figure B.3). For instance, the branch leading to *A. floridanus, A. tuberculatus* and *A. arenicola* had a 100% BS; however,

only 19% of the genes and 98% of the sites are concordant with the focal branch. Also, the gCF calculated in IQTREE corresponds to the conflicting/concordant bipartitions among gene trees obtained in Phyparts (i.e., for a gCF value of 15.4% for the branch leading to *A. floridanus* and *A. tuberculatus*, only 12 genes out of 78 support that branch) (Appendix B Figure B.4). Interestingly, the level of discordance in gene trees is less pronounced for the other species of Amaranthaceae s.s. included in the tree as could be observed in the proportion of gene trees that supports their branches, further indicating that complex conflicts exist within the *Amaranthus* genus. Considering the "backbone" of *Amaranthus* using the 19 species, 71 genes support the backbone phylogeny or species tree while only 7 genes were discordant (Appendix B Figure B.4), similar to Morales-Briones et al. (Morales-Briones et al. 2021) where 62 genes were in concordance with the species tree for the *Amaranthus* genus while only 6 were discordant (see Supplementary Figure S5 in Morales-Briones et al.).

The test of topology based on approximately unbiased (AU) test to determine if an *a priori* constraint tree where all dioecious species are placed together would be better than an unconstraint tree revealed that the constraint tree is significantly different from the unconstraint one (p = 6e-07). The result of the AU test is also congruent with an initial log-likelihood test (Shimodaira-Hasegawa test) reported in RAxML, with the constraint tree indicted as significantly worse than the unconstraint tree (RAxML does not output p-values for log-likelihood tests). The topology test thus suggests that the two species *A. australis* and *A. cannabinus* are less closely related to the other dioecious amaranths based on their chloroplast genomes.

For the plastome alignment excluding IRa, there were 103,019 conserved sites, 23,246 variable sites and 18,803 parsimony-informative sites in a total of 126,265 columns. The

topology of the tree using 78 plastid protein-coding genes and whole plastome sequences were very similar, except the sister relationship between *A. arenicola* and *A. tuberculatus* was now established and had full support (BS = 100, PP=1, ICA=1.00). *Amaranthus australi*s and *A. cannabinus* once again did not cluster with the other dioecious species; however, the support for their relationship with the *Acnida* + *Amaranthus* clades increased (BS = 98, PP = 1, ICA = 0.89). Support values for other nodes also increased (Figure 3.8). There was also no difference in topology and bootstrap support between IQTREE (TVM + F + R2 model) and RAxML (GTRGAMMA model) trees, except the node that had 60% bootstrap support in IQTREE had 49% bootstrap support in RAxML, therefore results from IQTREE are presented (see Appendix B Figure B.5 Bootstrap consensus network for RAxMLbootstrap support values). Bootstrap values measure the standard error of the inferred tree mean from a full dataset in which the standard error decreases with more samples or loci (Minh et al. 2020a); therefore, bootstrap support values are expected to be higher for the whole plastome alignment as opposed to the set of 78 protein-coding genes. Bootstrap consensus network and NeighborNet splits graph (fit = 99.661%) also showed a highly supported bipartition for *A. arenicola* + *A. tuberculatus*, and *A. australi*s + *A. cannabinus* lineages. However, 48.8% support the first bipartition or branch leading to *A. polygonoides* and the other species in Amaranthaceae s.s. while 32.6% support the second bipartition or branch leading to *A. viridis*, *A. tricolor* and other species in Amaranthaceae s.s. (Appendix B Figure B.5 – B.6). The Quartet internodal scores (QC/QD/QI) for the cp genome alignment for most branches, including the other species of Amaranthaceae s.s., was 0/0/1, respectively while taxon QF score ranged from 0.03 – 0.3 (data not shown). These scores differ considerably from the Quartet internodal scores obtained with the 78 protein-coding

sequences, thus reflecting a very complex conflict that could not be resolved from modeling the evolution of the species while assuming the concatenated plastid supermatrix as a "single-gene".

3.3.5 Evolutionary distance between *A. palmeri* and *A. watsonii*

Adjusting the method for distance calculation by using *p*-distance, Maximum Composite Likelihood, LogDet or changing rates to Gamma or Gamma and a proportion of invariable sites, or changing the Gamma rate parameter to 8 had no noticeable effects on the distances calculated. Therefore, we report the uncorrected *p*-distances. The evolutionary distance between *A. palmeri* and *A. watsonii* based on cp genome (minus IRa) was 0.0000476, which is considerably low compared to the distances between *A. tuberculatus* and *A. arenicola* (0.000143), *A. tuberculatus* and *A. floridanus* (0.000254) and *A. arenicola* and *A. floridanus* (0.000254). *Amaranthus australis* and *A. cannabinus* have also been shown to be sister taxa, however, the distance between both species was higher (0.0021688). The internal transcribed spacer (ITS) and full nuclear ribosomal cistron (rDNA) regions were 5,819 and 10,674 bp, respectively. Assembly size for the full rDNA ranged from 9,894 – 11,582 bp (Appendix A Table B.4). A BLAST search of 722 bp *A. tuberculatus* ITS (GenBank accession number MG685285) from Waselkov et al. (Waselkov et al. 2018) against our assembled *A. tuberculatus* nuclear rDNA revealed 96.8% similarity to a region in the assembly, suggesting that the assembly contained the total ITS sequence region used in their study. Evolutionary distance between *A. palmeri* and *A. watsonii* and between *A. caudatus*, *A. cruentus* and *A. quitensis* based on the ITS region was 0.000000 (Appendix B Table B.6). The very low distance (0) between these species indicates the low informativeness of the ITS region in distinguishing between the species. Only 38 parsimony-informative sites were found in the ITS region across the 14 *Amaranthus* species with short reads available for rDNA assembly. When the full rDNA assembly (containing sequences from ETS

and possibly IGS) was used for distance calculation, the distance between *A. palmeri* and *A. watsonii* was still low (0.000453) relative to the distances between *A. tuberculatus* and *A. arenicola* (0.003036), *A. tuberculatus* and *A. floridanus* (0.006462), and *A. arenicola* and *A. floridanus* (0.003645). The evolutionary distance between *A. hybridus* and *A. quitensis* was 0.016139, similar to the distance between *A. cruentus* and *A. quitensis* (0.016233) (Appendix B Table B.7).

3.4 DISCUSSION

3.4.1 Dioecious *Amaranthus* species' plastome features

We report the complete chloroplast genomes of nine dioecious *Amaranthus* species and their composition. The size of the cp genomes is consistent with the size of 150 – 151 kb reported for other *Amaranthus* species (Chaney et al. 2016; Xu et al. 2022). Similarly, GC content, number of protein-coding genes, transfer RNAs, ribosomal RNAs and overall structure are highly conserved across the dioecious *Amaranthus* species. Our comparative analysis revealed regions that differed across the species e.g., *trnL*^{UAG}*-ccsA-ndhD*, were highly divergent across the nineteen *Amaranthus* species and could be valuable in marker development or DNA barcoding. This region among others has been reported to be very variable across flowering plants (Shaw et al. 2014; Shahzadi et al. 2020). Moreover, the low nucleotide diversity (see Appendix B Figure B.1 for highest π value at 0.016) among *Amaranthus* species also suggests a high genetic similarity, which may impact phylogenetic signals. A similar pattern of low nucleotide variability was observed among species of *Aldama* (Asteraceae), where the most variable region had a π value between 0.02936 and 0.0305 (Loeuille et al. 2021). Although chloroplast size variation in several species could be attributed to expansion and contraction of IR regions (Palmer et al. 1987; Dugas et al. 2015; Mower and Vickrey 2018), the

LSC/IRb/SSC/IRa boundaries, including their positions, were very conserved across the dioecious amaranths. Our analysis of microsatellites and repeats also revealed patterns consistent with previous studies of SSRs and repetitive sequences in the amaranths (Chaney et al. 2016; Xu et al. 2022). The relative synonymous codon usage for dioecious amaranths is also similar to *A. hypochondriacus* and other plant cp genomes (Chaney et al. 2016; Wen et al. 2021).

3.4.2 Phylogenetic incongruence within the dioecious amaranths

Of particular interest to us is the relationships among the dioecious amaranths, which have been elusive. Waselkov et al. (2018) studied the phylogeny of the amaranths using six molecular markers and attributed observed cytonuclear tree discordance to incomplete lineage sorting (ILS) and chloroplast capture. Xu et al. (2022), although they did not sample all dioecious amaranths, produced trees using complete chloroplast sequences but did not detect tree topology incongruence. Nontree-like signals in a phylogenetic tree could be due to either statistical reasons (incorrect model specification, sequence errors or short alignments) or biological factors such as hybridization, incomplete lineage sorting, ancestral gene flow or low mutation rate (Degnan 2018). We therefore evaluated if factors including poor loci resolution contributes to gene tree incongruence and if the use of more markers could provide better phylogenetic resolution.

Using a series of complementary approaches, we identified internodes or branches with low degrees of certainty. A combination of strong conflicts in phylogenetic signal and sometimes absence or low informative signals contributed to the conflict in reconstructing the true relationship between the amaranths. We found strong support along the “backbone” relating species in the *Acnida* clade (all nine of the dioecious species except *A. australis* and *A. cannabinus*) and species in the *Amaranthus* clade, and strong support for the sister relationship

between both clades, consistent with the nuclear phylogeny in Waselkov et al. (2018). The relationship of *A. australis + A. cannabinus* lineage to the other dioecious species however remains obscure, the topology test of monophyly did not support the placement of both species in the same clade as the other seven dioecious species. Chloroplast genomes are non-recombining and uniparentally inherited, and it is possible that the chloroplast in *A. australis + A. cannabinus* lineage was inherited after a hybridization event or chloroplast capture from an ancestor leading to the *Acnida + Amaranthus* clades.

Summary coalescent methods are known to be more robust than concatenation methods in the presence of high levels of ILS (Yu et al. 2011; Mirarab et al. 2021), and we have inferred species tree from the plastid protein-coding genes using a summary coalescent analysis. Genes with short lengths and uninformative loci that is typical of chloroplast genomes may however contribute to gene trees with topology inconsistencies at some branches and a subsequent species tree that is less accurate (Mirarab et al. 2014; Xi et al. 2015). Nevertheless, the higher proportion of gene trees (> 50%) concordant with the species tree for Amaranthaceae s.s. (tribes Celosieae, Aerveae, Achyrantheae and Gomphreneae) but not for *Amaranthus* species (see Appendix B Figure B.4), indicates inherent processes within the *Amaranthus* genus that contribute to conflicting phylogenetic signals. The inclusion of species belonging to these four tribes in our phylogenetic analysis therefore proved informative as it allowed us to validate the relationship of the tribes to Amarantheae. We recovered clades corresponding to relationships between the five tribes previously described in the Angiosperm Phylogeny Group (APG) IV system of classification (Chase et al. 2016) and previous studies (Kadereit et al. 2003; Müller and Borsch 2005; Morales-Briones et al. 2021).

It is expected that all genes in the plastomes would share the same evolutionary history based on their inheritance patterns. However, recent findings for angiosperms reveal chloroplast genes exhibit well-supported conflict and do not appear to share the same evolutionary history (Gonçalves et al. 2019; Walker et al. 2019). Plastid gene tree incongruence among five major clades of Amaranthaceae s.l. was recently hypothesized to be likely due to heteroplasmy (Morales-Briones et al. 2021). It is difficult to determine the exact causes of conflict in plastid gene trees within the *Amaranthus* genus in our study, whether it is a result of varying evolutionary histories of the genes or a result of systematic or other analytical methods e.g., lack of information or misalignment. There is also a debate over the impact of taxon sampling on the accuracy of phylogenetic analysis, with some authors reporting the contribution of low taxon sampling to tree conflicts (Heath et al. 2008) while others note no impact on tree inference (Rosenberg and Kumar 2001) [see Nabhan and Sarkar (2012) for a review on taxon sampling controversy]. Nevertheless, we sampled all the species in the dioecious clade (subgenus *Acnida*) as well as several species in the Hybridus clade (subgenus *Amaranthus*) and therefore tree conflicts in our study are not due to low taxon sampling.

Contrary to studies where data partitioning has improved phylogenetic inference (Xi et al. 2012), topology tests between partitioned and unpartitioned data sets for the 78 CDS revealed no differences between both approaches (Xiao et al. 2020). However, we recommend data partitioning, as the analysis of the whole plastome data sets yielded branches with high support but also highly complex conflicts that could not be easily interpreted. While we did not specifically investigate the contribution of tRNA, rRNA and introns by including partitions for them in the phylogenetic tree, the full support for the sister relationship between *A. arenicola* and *A. tuberculatus* using whole plastome alignment, which was not clear from using 78 protein-

coding regions, indicates that more signals favoring this relationship is coming from non-coding regions. Non-coding regions also hold phylogenetic information that could be useful in resolving shallow evolutionary relationships (Shaw et al. 2014; Walker et al. 2019). Their impact on tree inference would need to be further evaluated for the amaranths.

Additional studies into the relationship between the amaranths is required to understand their evolutionary history. Using a *k*-mer-based phylogenomic analysis, Raiyemo et al. (2023) reported the relatedness between the dioecious *Amaranthus* species. Although, the *k*-mer method was alignment-free and do not model complex evolutionary processes, sister-species relationships (e.g., between *A. australis* and *A. cannabinus*, *A. arenicola* and *A. greggii*, or *A. tuberculatus* and *A. floridanus*) were obtained, which is congruent with the previous infrageneric classification that was based on morphological characteristics. Nonetheless, phylogenetic studies incorporating morphological data, nuclear genes (perhaps obtained via a hybrid capture-based target enrichment) and mitochondrial data would still be required to enhance our understanding of the evolution of the *Amaranthus* genus and to provide additional insights into tree discordance in the genus (Koenen et al. 2020). Our work provides a framework for further investigation of the relationship between the amaranths as more species within the genus are sequenced.

3.4.3 Are *A. palmeri* and *A. watsonii* two species or a single polymorphic species?

Although both *A. palmeri* and *A. watsonii* had long been considered separate species by various authorities (Sauer 1955, 1957; Waselkov et al. 2018), the similarity in morphological characteristics, high degree of species range overlap and a low evolutionary distance between both species could indicate a single polymorphic species. Based on Sauer's (Sauer 1955) reported morphological characteristics, both species are very similar (1m tall; 5 stamens, 5 tepals, and inner tepal length of 2.5 – 3 mm for male flowers; 5 tepals with 2 – 2.5 mm length for

female flowers; utricle length of 1.5 mm; 2 or sometimes 3 style branches; and seed with obovate shape and dark reddish brown color), but differ in length of thyrses and shape of leaf blade. Historically, both species were considered important food plant; as a potherb and source of grain for various Indian tribes (Sauer 1955). Furthermore, Sauer (1955) hypothesized that the Colorado River and the associated irrigation projects provided the opportunity for which *A. watsonii* mixes with *A. palmeri*, and have moved into Southern California as a weed of irrigated fields. Both species are native to California and Arizona and are sympatric in San Bernadino and Imperial counties of California, and Yuma and Maricopa counties of Arizona (https://plants.usda.gov/home) (USDA 2022).

Stelkens and Seehausen (Stelkens and Seehausen 2009) in a study of evolutionary distances for hybridizing species using ITS1 and ITS2 reported a distance of 0.0155 between *A. retroflexus* and *A. cruentus*, which is congruent with the distance values between some closely related monoecious species in our study. The lowest distance in their study was between *Mimulus lewisii* and *M. cardinalis* (0.002), which was much higher than the distance between *A. palmeri* and *A. watsonii* (0.000453) in our study. Although, *A. palmeri* is now widespread and has become a troublesome weed of different agricultural systems (Ward et al. 2013), little is known about *A. watsonii*, or interspecific hybridization between both species that may have resulted in novel hybrid traits. Nevertheless, the very low distance between both species in our study based on complete chloroplast genomes and rDNA in addition to previously reported morphological similarities indicate that the two species are more genetically related than previously reported. Our study reinforces the taxonomic reconsideration of *A. palmeri* and *A. watsonii* as a single polymorphic species, or the latter be considered a variety of *A. palmeri*.

3.5 CONCLUSION

Although, the *Amaranthus* genus has been described as taxonomically challenging to work with due to similarities in species morphology and difficulty in accurate identification, we demonstrate that the use of complementary phylogenetic approaches coupled with proper species identification could be very informative in examining the genus' complex evolutionary history. We provide additional clarification on the relationships among the dioecious species of the *Amaranthus* genus, which have been conflicting based on previous studies where few molecular markers were used. Important open questions remain for the amaranths: 1) When in the evolutionary and biogeographic time scale did speciation events occurred? 2) When did chloroplast capture events take place? 3) Was there rapid radiation or ancient hybridization in the genus and at what time could this have taken place?

3.6 TABLES AND FIGURES

Table 3.1 Chloroplast genome features of nine dioecious *Amaranthus* species (LSC, large single copy; SSC, small single copy; IR, inverted repeat).

Species	**Length (bp)**	**Coverage depth (x)**	**LSC (bp)**	**SSC (bp)**	**IR (bp)**	**GC (%)**	**Number of unique genes**			
							Protein-coding	**tRNA**	**rRNA**	**Total**
A. acanthochiton	150,653	522.6	83,927	18,034	24,346	36.59	78	30	4	112
A. arenicola	150,655	511.4	83,926	18,037	24,346	36.61	78	30	4	112
A. australis	150,011	615.2	83,244	18,065	24,351	36.62	78	30	4	112
A. cannabinus	150,677	774.8	83,888	18,085	24,352	36.56	78	30	4	112
A. floridanus	150,670	514.0	83,935	18,043	24,346	36.60	78	30	4	112
A. tuberculatus	150,632	740.1	83,901	18,039	24,346	36.61	78	30	4	112
A. greggii	150,735	519.1	83,955	18,088	24,346	36.58	78	30	4	112
A. watsonii	150,706	520.6	83,986	18,026	24,347	36.61	78	30	4	112
A. palmeri	150,708	484.5	83,988	18,026	24,347	36.60	78	30	4	112

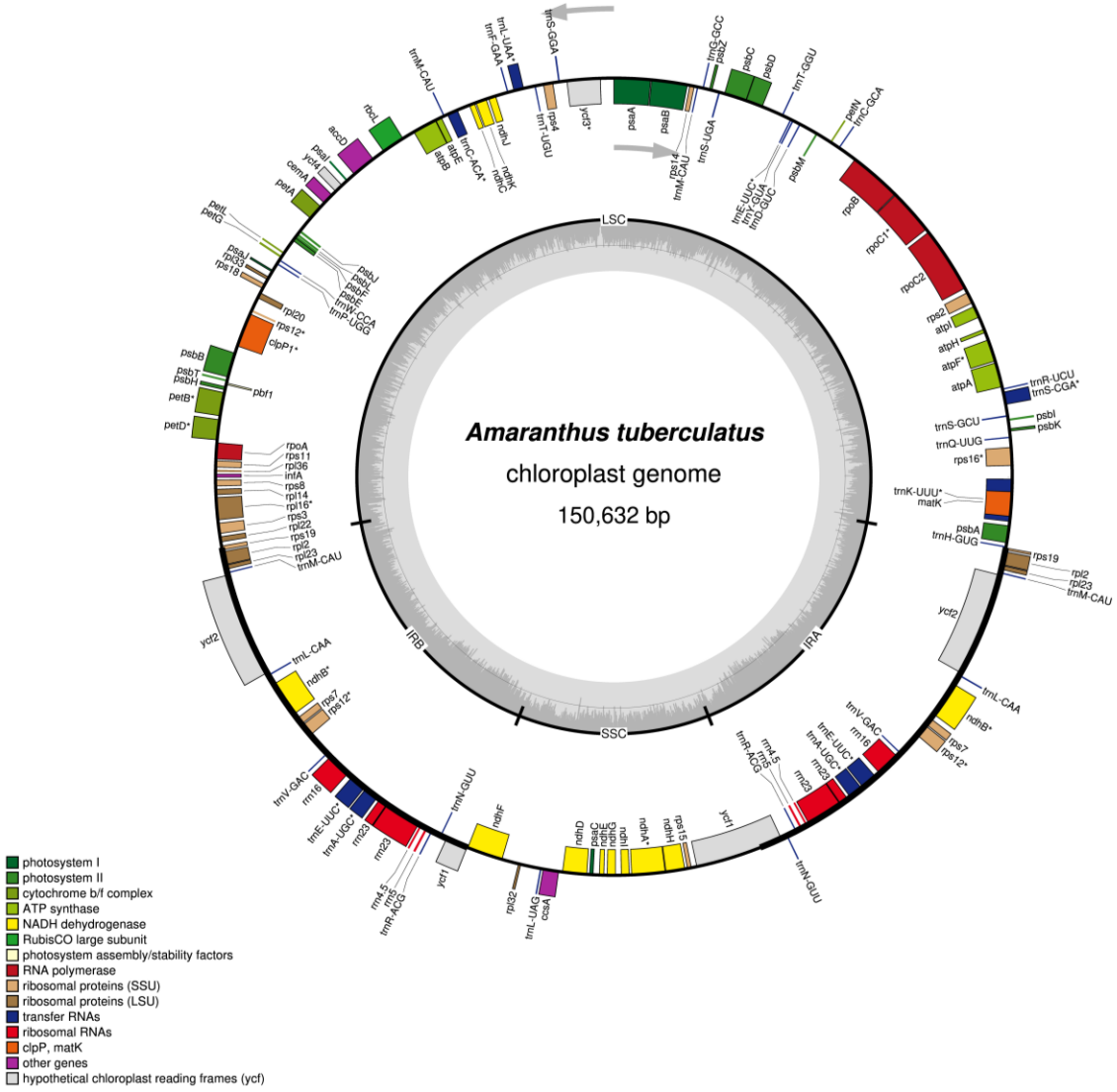

Figure 3.1 Annotated chloroplast gene map of *Amaranthus tuberculatus*. Genes depicted on the inside of the circle are transcribed clockwise while genes shown on the outside of the circle are transcribed counterclockwise. Genes with asterisk have introns. The grey area within the circle represents the GC content across the chloroplast genome.

Table 3.2 Simple sequence repeats (SSRs) in the nine dioecious *Amaranthus* chloroplast genomes

Species	**Mono**	**Di**	**Tri**	**Tetra**	**Penta**	**Hexa**	**Compound**	**Total**
A. acanthochiton	12	2	4	11	0	1	1	31
A. arenicola	13	2	4	10	0	1	2	32
A. australis	10	2	5	12	0	1	3	33
A. cannabinus	14	2	4	14	1	1	1	37
A. floridanus	17	2	4	10	0	1	2	36
A. tuberculatus	15	2	4	10	0	1	2	34
A. greggii	14	1	4	11	0	1	1	32
A. watsonii	12	3	4	11	0	1	1	32
A. palmeri	12	3	4	11	0	1	1	32

Table 3.3 Number of repetitive sequence types in nine dioecious *Amaranthus* chloroplast genomes.

Species	Forward	Palindrome	Reverse	Complement	Total
A. acanthochiton	14	22	0	0	36
A. arenicola	14	22	1	0	37
A. australis	15	23	0	0	38
A. cannabinus	15	21	0	0	36
A. floridanus	14	22	1	0	37
A. tuberculatus	14	23	1	0	38
A. greggii	16	22	1	0	39
A. watsonii	14	21	1	0	36
A. palmeri	14	21	1	0	36

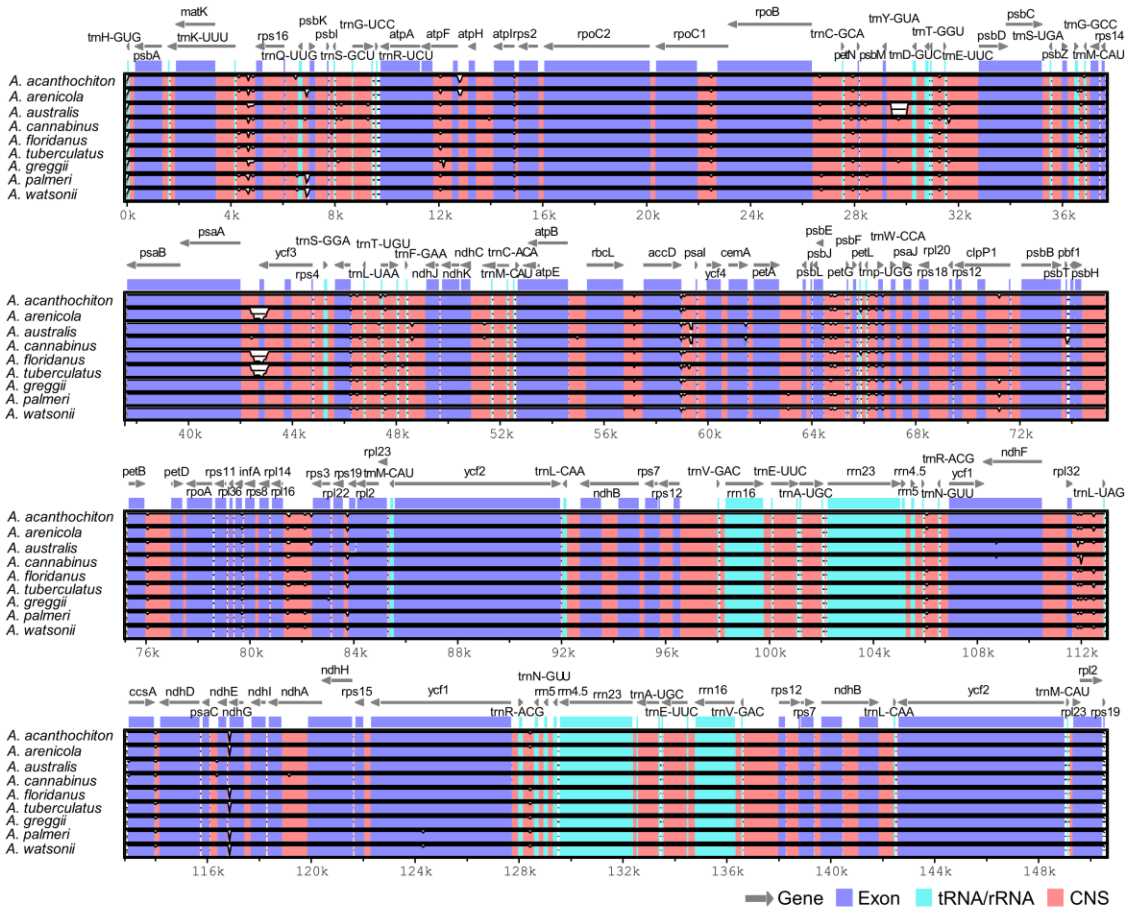

Figure 3.2 Sequence alignment of complete chloroplast genomes of nine dioecious *Amaranthus* species to the *A. hypochondriacu*s chloroplast genome (KX279888) using mVISTA. The y-axis within each species bar corresponds to percentage sequence identity (50-100%). The grey arrows indicate annotated genes within the genomes and their transcriptional direction. Genomic regions are color-coded as protein-coding (exon), transfer or ribosomal RNA (tRNA/rRNA), and conserved non-coding sequences (CNS).

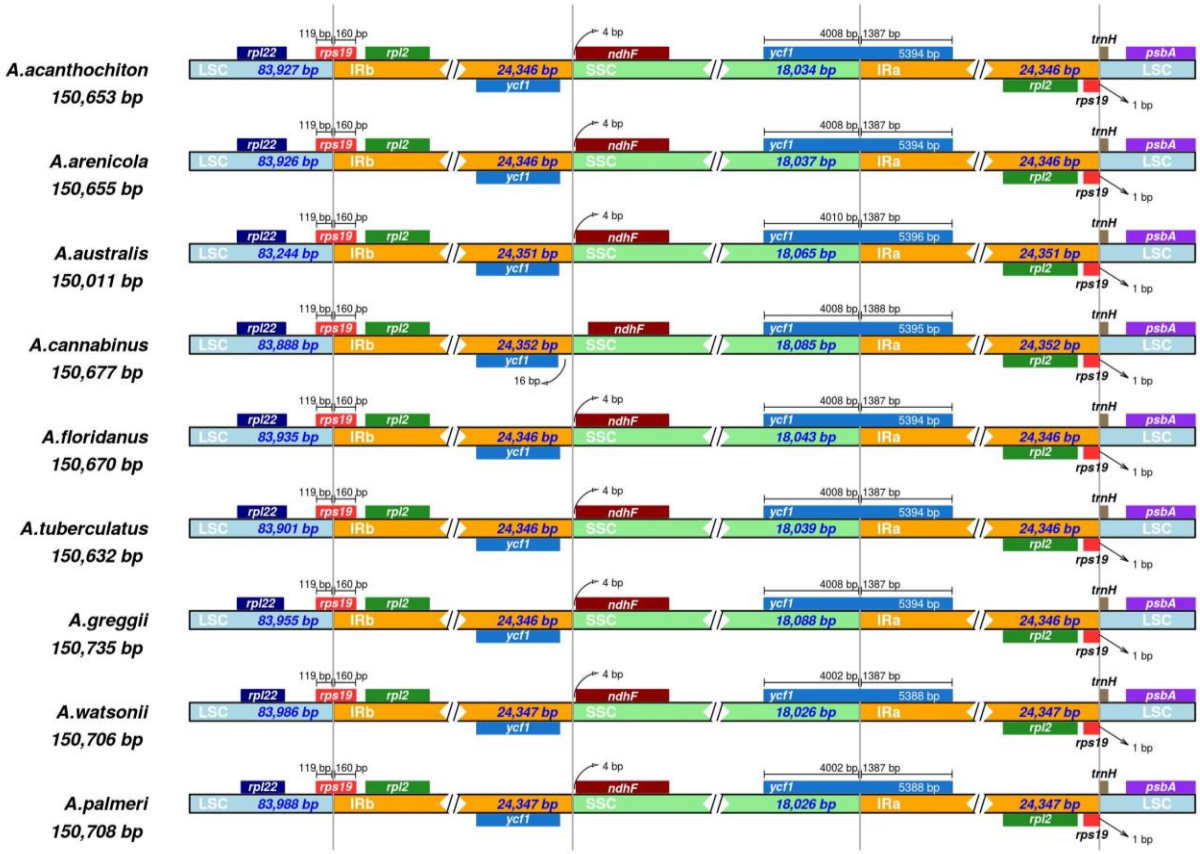

Figure 3.3 Comparison of large single copy (LSC), small single copy (SSC) and inverted repeats (IR) border regions among the nine dioecious *Amaranthus* chloroplast genomes. Genes preceded by the Greek letter psi (ψ) represent possible pseudogenes.

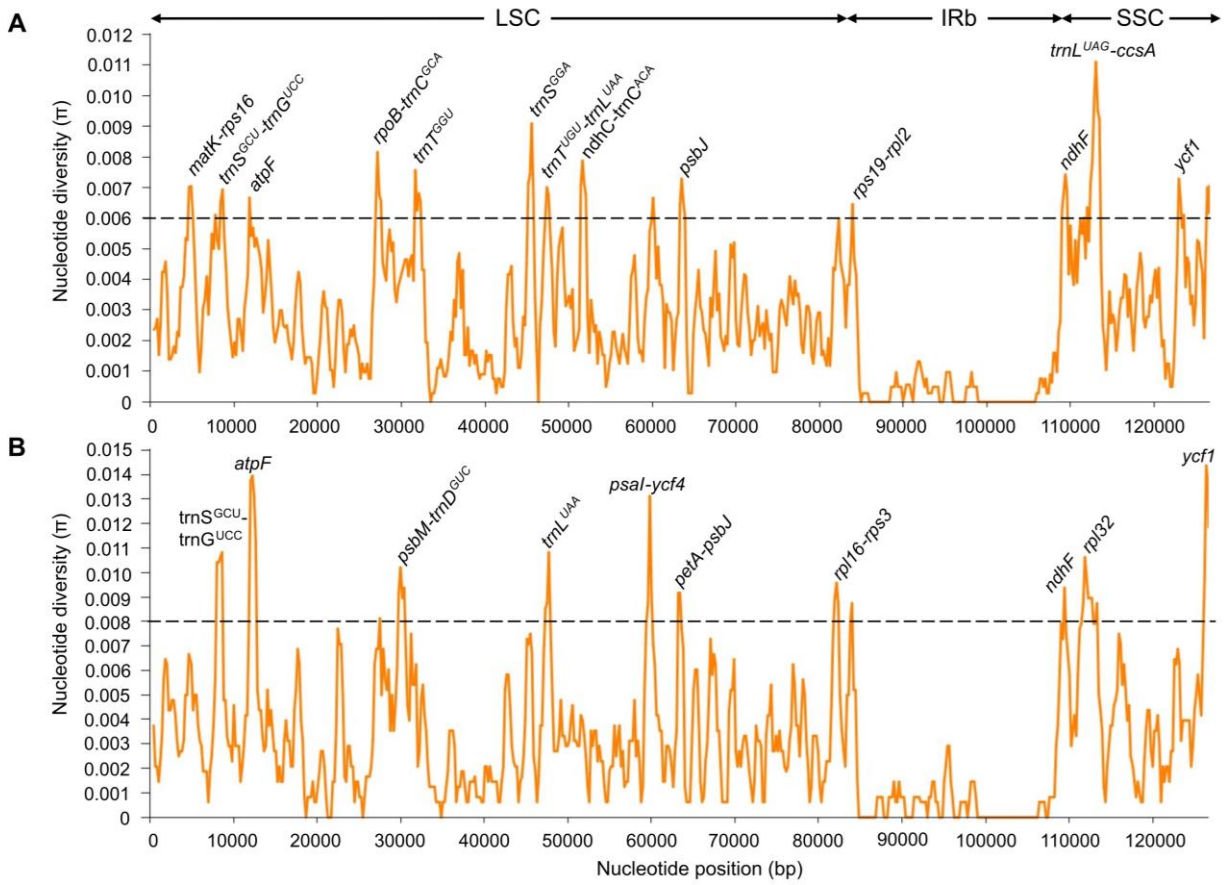

Figure 3.4 Sliding window analysis of nucleotide diversity within *Amaranthus* plastomes. A) comparison among nine dioecious species and B) comparison among four weedy species: *A. tuberculatus*, *A. palmeri*, *A. hybridus* and *A. retroflexus* (GenBank Accession number MW646089). Window length: 800 bp; step size: 200 bp.

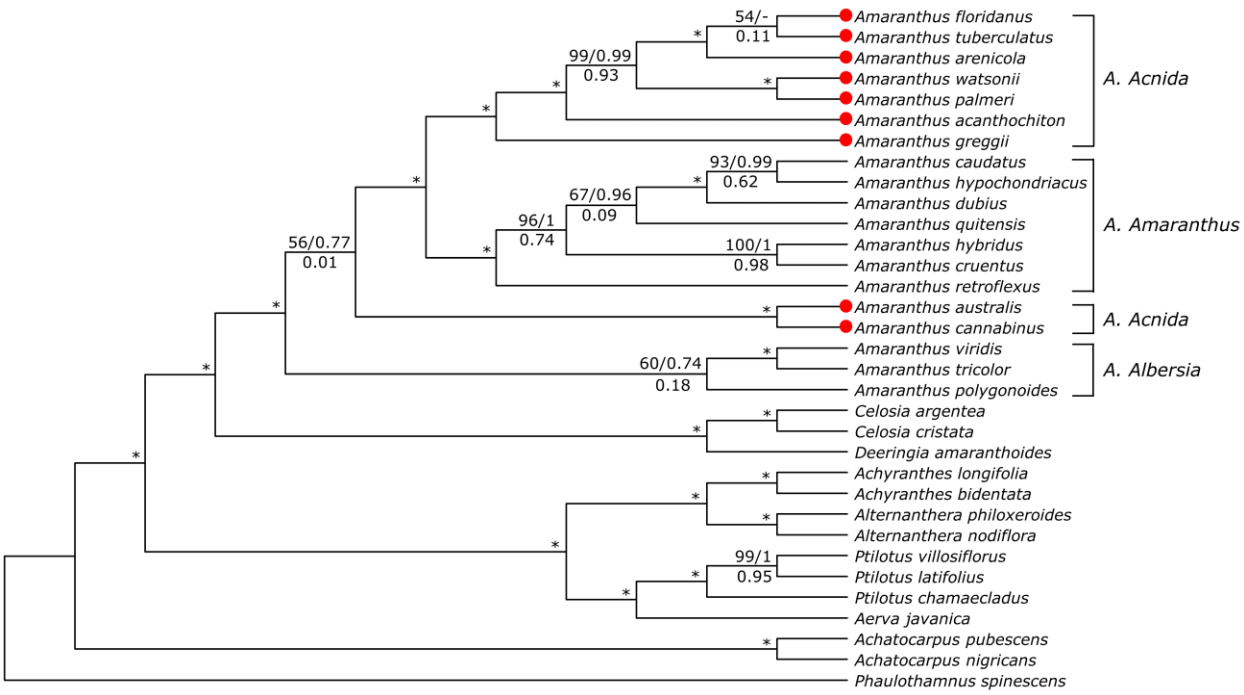

Figure 3.5 Cladogram of *Amaranthus* species and other species in Amaranthaceae s.s. based on 78 plastid protein-coding genes. Numbers above branches represent RAxML maximum likelihood bootstrap support (BS) and Bayesian posterior probability (PP) values, while values below branches represent Internode Certainty All (ICA) values. Asterisks indicate full support (BS = 100, PP = 1, ICA = 1.00). Terminal tips in red represent newly assembled plastid genomes in this study.

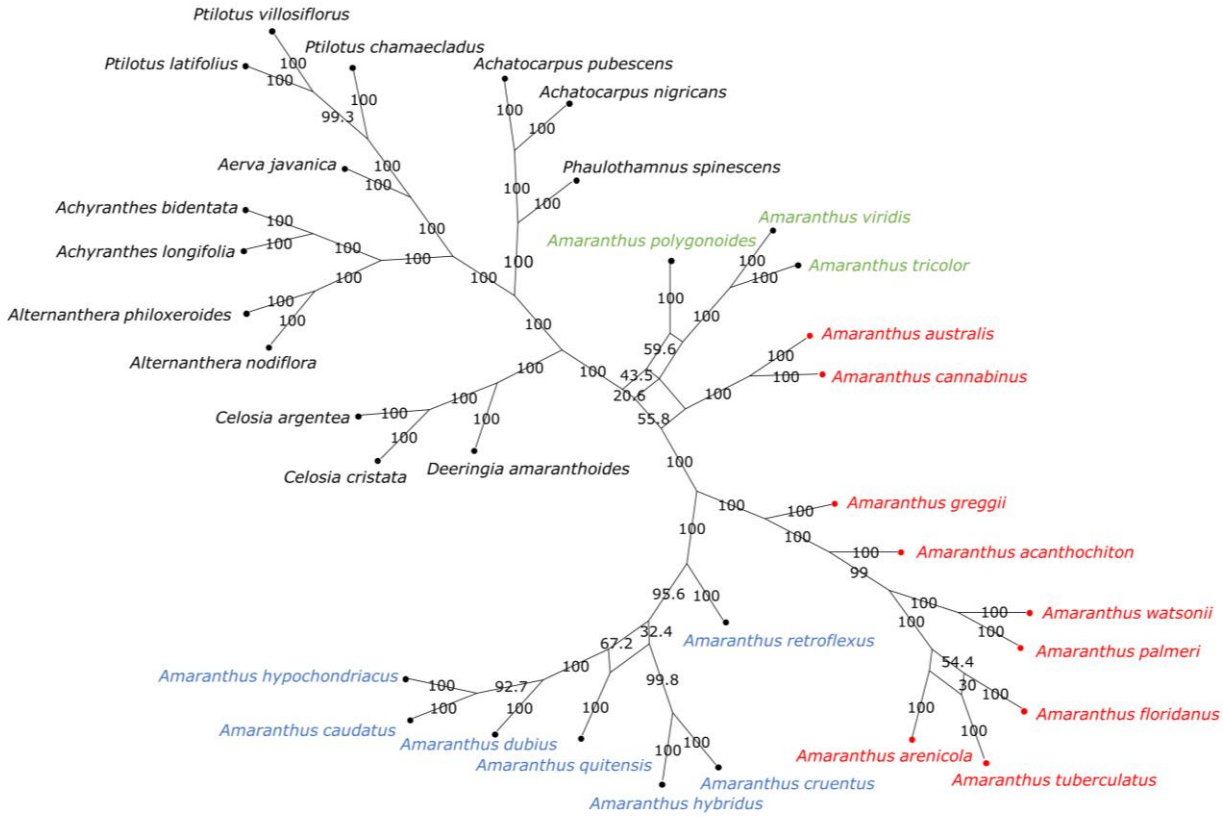

Figure 3.6 Bootstrap consensus network inferred from the maximum likelihood tree analysis for *Amaranthus* species and other species in Amaranthaceae s.s. based on 78 plastid protein-coding genes. Filtering threshold was 0.2, i.e., display splits or taxon bipartitions that occurred in at least 20% of the bootstrap replicates. Numbers on edges of the splits network are bootstrap support values. Species in red denotes subgenus *Acnida* while terminal tips in red are species with chloroplast genomes assembled in this study. Species in blue represents the subgenus *Amaranthus* while species in green represent subgenus *Albersia*.

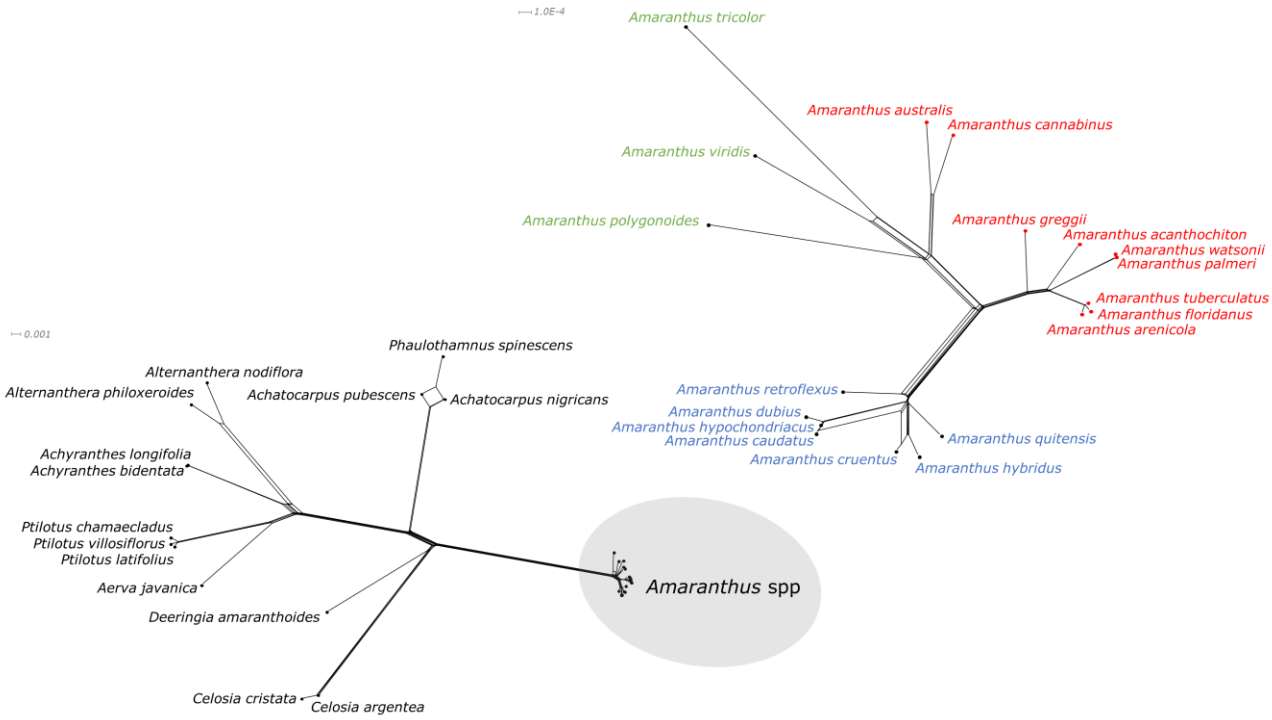

Figure 3.7 NeighborNet splits graph of *Amaranthus* species and other species in Amaranthaceae s.s. based on 78 plastid protein-coding genes. Split graph of *Amaranthus* spp. in the gray circle is enlarged in the top right. Species in red denotes subgenus *Acnida* while terminal tips in red are species with chloroplast genomes assembled in this study. Species in blue represents the subgenus *Amaranthus* while species in green represent subgenus *Albersia*. Scale bars (substitutions per site) are presented at the top-left corner of the graphs.

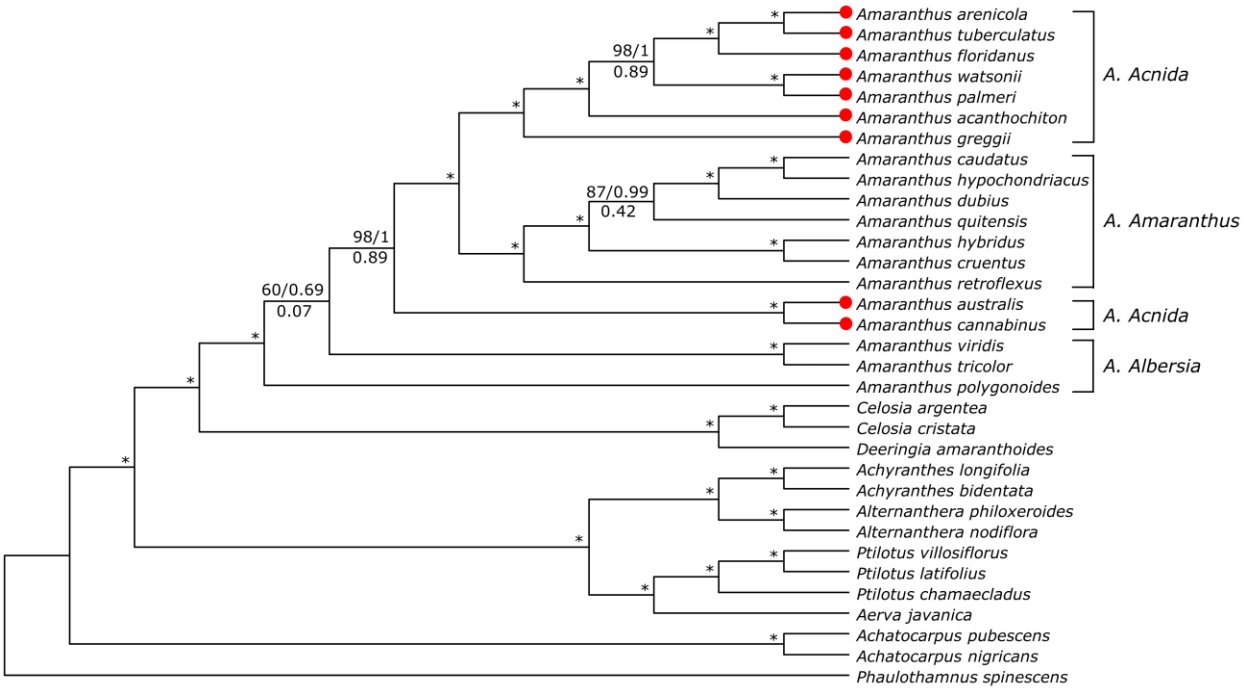

Figure 3.8 Cladogram of *Amaranthus* species and other species in Amaranthaceae s.s. based on whole chloroplast genomes. Numbers above branches represent IQ-TREE maximum likelihood ultrafast bootstrap support (UFBoot) and Bayesian posterior probability (PP) values, while values below branches represent RAxML Internode Certainty All (ICA) values. Asterisks indicate full support (BS = 100, PP = 1, ICA = 1.00). Terminal tips in red represent newly assembled plastid genomes in this study.

3.7 REFERENCES

Aderibigbe OR, Ezekiel OO, Owolade SO, Korese JK, Sturm B, Hensel O (2022) Exploring the potentials of underutilized grain amaranth (*Amaranthus* spp.) along the value chain for food and nutrition security: A review. Crit Rev Food Sci Nutr 62:656–669

Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–693

Amiryousefi A, Hyvönen J, Poczai P (2018) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34:3030–3031

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, Whitfield J (2013) Networks: Expanding evolutionary thinking. Trends Genet 29:439–441

Bayón ND (2022) Identifying the weedy amaranths (*Amaranthus*, Amaranthaceae) of South America. Adv Weed Sci 40:1–9

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: A web server for microsatellite prediction. Bioinformatics 33:2583–2585

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger I, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson KM, Vik U, Veldre V, Nilsson RH (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol Evol 4:914–919

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S (2003) Glocal alignment: Finding rearrangements during alignment. Bioinformatics 19

Bryant D, Moulton V (2004) Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265

Chaney L, Mangelson R, Ramaraj T, Jellen EN, Maughan PJ (2016) The complete chloroplast genome sequences for four *Amaranthus* species (Amaranthaceae). Appl Plant Sci 4:1600063

Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF, Briggs B, Brockington S, Chautems A, Clark JC, Conran J, Haston E, Möller M, Moore M, Olmstead R, Perret M, Skog L, Smith J, Tank D, Vorontsova M, Weber A (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20

Costea M, DeMason D (2001) Stem morphology and anatomy in *Amaranthus* L. (Amaranthaceae). J Torrey Bot Soc 128:254–281

Degnan JH (2018) Modeling hybridization under the network multispecies coalescent. Syst Biol 67:786–799

Dobrogojski J, Adamiec M, Luciński R (2020) The chloroplast genome: a review. Acta Physiol Plant 42:1–13

Dong W, Liu J, Yu J, Wang L, Zhou S (2012) Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS One 7:1–9

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue

Duchene D, Bromham L (2013) Rates of molecular evolution and diversification in plants:

Chloroplast substitution rates correlate with species-richness in the Proteaceae. BMC Evol Biol 13

Dugas D V., Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT, Hajrah NH, Alharbi NS, Al-Malki AL, Sabir JSM, Bailey CD (2015) Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. Sci Rep 5:1–13

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048

Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y (2018) Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci USA 115:E4940–E4949

Gonçalves DJP, Simpson BB, Ortiz EM, Shimizu GH, Jansen RK (2019) Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. Mol Phylogenet Evol 138:219–232

Greiner S, Lehwark P, Bock R (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res 47:W59–W64

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257

Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287–299

Howe CJ, Barbrook AC, Koumandou VL, Nisbet RER, Symington HA, Wightman TF, Fray R, Leaver CJ, Walker JE, Gray JC, Douglas AE, Cavalier-Smith T, Allen JF, Hermann RG, Blankenship RE (2003) Evolution of the chloroplast genome. Philos Trans R Soc B Biol Sci

358:99–107

Hu S, Sablok G, Wang B, Qu D, Barbaro E, Viola R, Li M, Varotto C (2015) Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats. BMC Genomics 16:1–14

Huang YY, Matzke AJM, Matzke M (2013) Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). PLoS One 8:1–12

Huson DH (1998) SplitsTree: Analyzing and visualizing evolutionary data. Bioinformatics 14:68–73

Huson DH, Scornavacca C (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067

Iamonico D (2020) Nomenclatural survey of the genus *Amaranthus* (Amaranthaceae). 11. dioecious *Amaranthus* species belonging to the sect. *Saueranthus*. Darwiniana 8:567–575

Jansen RK, Raubeson LA, Boore JL, DePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl J V., McNeal JR, Leebens-Mack J, Cui L (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol 395:348–384

Jin JJ, Yu W Bin, Yang JB, Song Y, Depamphilis CW, Yi TS, Li DZ (2020) GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol 21:1–31

Kadereit G, Borsch T, Weising K, Freitag H (2003) Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C4 photosynthesis. Int J Plant Sci 164:959–986

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E, Eddy SR, Finn RD,

Bateman A, Petrov AI (2021) Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res 49:D192–D200

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589

Katoh K, Misawa K, Kuma KI, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780

Kloepper TH, Huson DH (2008) Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol 8:1–7

Koenen EJM, Ojeda DI, Steeves R, Migliore J, Bakker FT, Wieringa JJ, Kidner C, Hardy OJ, Pennington RT, Bruneau A, Hughes CE (2020) Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. New Phytol 225:1355–1369

Kurtz S, Choudhuri J V., Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642

Lee C, Wen J (2004) Phylogeny of *Panax* using chloroplast *trnC-trnD* intergenic region and the utility of *trnC-trnD* in interspecific studies of plants. Mol Phylogenet Evol 31:894–903

Liu LX, Li R, Worth JRP, Li X, Li P, Cameron KM, Fu CX (2017) The complete chloroplast genome of chinese bayberry (*Morella rubra*, myricaceae): Implications for understanding the evolution of Fagales. Front Plant Sci 8:1–15

Loeuille B, Thode V, Siniscalchi C, Andrade S, Rossi M, Pirani JR (2021) Extremely low

nucleotide diversity among thirty-six new chloroplast genome sequences from *Aldama* (Heliantheae, Asteraceae) and comparative chloroplast genomics analyses with closely related genera. PeerJ 9

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol 14:1–14

McPherson H, van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M (2013) Capturing chloroplast variation for molecular ecology studies: A simple next generation sequencing approach applied to a rainforest tree. BMC Ecol 13

Minh BQ, Hahn MW, Lanfear R (2020a) New methods to calculate concordance factors for phylogenomic datasets. Mol Biol Evol 37:2727–2733

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R, Teeling E (2020b) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534

Mirarab S, Bayzid MS, Warnow T (2014) Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst Biol 65:366–380

Mirarab S, Nakhleh L, Warnow T (2021) Multispecies coalescent: Theory and applications in phylogenetics. Annu Rev Ecol Evol Syst 52:247–268

Morales-Briones DF, Kadereit G, Tefarikis DT, Moore MJ, Smith SA, Brockington SF, Timoneda A, Yim WC, Cushman JC, Yang Y (2021) Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.l. Syst Biol 70:219–235

Mosyakin SL, Robertson KR (1996) New infrageneric taxa and combinations in *Amaranthus*

(Amaranthaceae). Ann Bot Fenn 33:275–281

Mosyakin SL, Robertson KR (2003) *Amaranthus*. *In* Flora of North America Editorial Committee, ed. Flora of North America North of Mexico, vol 4. 4th ed. Oxford University Press, Oxford. 410–435 p

Mower JP, Vickrey TL (2018) Structural diversity among plastid genomes of land plants. Advances in Botanical Research. 1st ed. Elsevier Ltd. 263–292 p

Müller K, Borsch T (2005) Phylogenetics of Amaranthaceae based on matK / trnK sequence data: Evidence from Parsimony, Likelihood, and Bayesian analyses. Ann Missouri Bot Gard 92:66–102

Nabhan AR, Sarkar IN (2012) The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy. Brief Bioinform 13:122–134

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Palmer JD, Nugent JM, Herbon LA (1987) Unusual structure of geranium chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proc Natl Acad Sci 84:769–773

Pease JB, Brown JW, Walker JF, Hinchliff CE, Smith SA (2018) Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am J Bot 105:385–403

Peden JF (1999) Analysis of codon usage. University of Nottingham, UK

Raiyemo DA, Bobadilla LK, Tranel PJ (2023) Genomic profiling of dioecious *Amaranthus* species provides novel insights into species relatedness and sex genes. BMC Biol 21:1–18

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in

Bayesian phylogenetics using Tracer 1.7. Syst Biol 67:901–904

Riggins CW, Mumm RH (2021) Amaranths. Curr Biol 31:R834–R835

Riggins CW, Peng Y, Stewart CN, Tranel PJ (2010) Characterization of *de novo* transcriptome for waterhemp (*Amaranthus tuberculatus*) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes. Pest Manag Sci 66:1042–1052

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) Mrbayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542

Rosenberg MS, Kumar S (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci USA 98:10751–10756

Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34:3299–3302

Salichos L, Stamatakis A, Rokas A (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol 31:1261–1271

Sarker U, Lin YP, Oba S, Yoshioka Y, Hoshikawa K (2022) Prospects and potentials of underutilized leafy amaranths as vegetable use for health-promotion. Plant Physiol Biochem 182:104–123

Sauer J (1955) Revision of the dioecious amaranths. Madroño 13:5–46

Sauer J (1957) Recent migration and evolution of the dioecious amaranths. Evolution (N Y) 11:11–31

Sauer J (1972) The dioecious amaranths: A new species name and major range extensions. Madrono 21:426

Sauer JD (1950) The grain amaranths: A survey of their history and classification. Ann Missouri Bot Gard 37:561–632

Sauer JD (1967) The grain amaranths and their relatives: A revised taxonomic and geographic survey. Ann Missouri Bot Gard 54:103–137

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I (2020) GenBank. Nucleic Acids Res 48:D84–D86

Schliep K, Potts AJ, Morrison DA, Grimm GW (2017) Intertwining phylogenetic trees and networks. Methods Ecol Evol 8:1212–1220

Shahzadi I, Abdullah, Mehmood F, Ali Z, Ahmed I, Mirza B (2020) Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses, mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. Genomics 112:1454–1463

Shaw J, Shafer HL, Rayne Leonard O, Kovach MJ, Schorr M, Morris AB (2014) Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. Am J Bot 101:1987–2004

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

Shimodaira H, Hasegawa M (2002) CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247

Smith DR (2015) Mutation rates in plastid genomes: They are lower than you might think. Genome Biol Evol 7:1227–1234

Smith SA, Moore MJ, Brown JW, Yang Y (2015) Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC

Evol Biol 15:1–15

Song Y, Yu W Bin, Tan YH, Jin JJ, Wang B, Yang JB, Liu B, Corlett RT (2020) Plastid phylogenomics improve phylogenetic resolution in the Lauraceae. J Syst Evol 58:423–439

Spalik K, Downie SR, Watson MF (2009) Generic delimitations within the *Sium* alliance (Apiaceae tribe Oenantheae) inferred from cpDNA *rps16-5'trnK* (UUU) and nrDNA ITS sequences. Taxon 58:735–748

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Steckel LE (2007) The dioecious *Amaranthus* spp.: Here to stay. Weed Technol 21:567–570

Stelkens R, Seehausen O (2009) Genetic distance between species predicts novel trait expression in their hybrids. Evolution 63:884–897

Stetter MG, Schmid KJ (2017) Analysis of phylogenetic relationships and genome size evolution of the *Amaranthus* genus using GBS indicates the ancestors of an ancient crop. Mol Phylogenet Evol 109:80–92

Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38:3022–3027

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6–W11

Tranel PJ (2021) Herbicide resistance in *Amaranthus tuberculatus*†. Pest Manag Sci 77:43–54

Twyford AD, Ness RW (2017) Strategies for complete plastid genome sequencing. Mol Ecol Resour 17:858–868

USDA N (2022) The PLANTS Database (http://plants.usda.gov, 08/27/2022).

Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW (2019) Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ 2019:1–31

Wambugu PW, Brozynska M, Furtado A, Waters DL, Henry RJ (2015) Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. Sci Rep 5:1–9

Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol 8:1–14

Wang W, Schalamun M, Morales-Suarez A, Kainer D, Schwessinger B, Lanfear R (2018) Assembly of chloroplast genomes with long- and short-read data: A comparison of approaches using *Eucalyptus pauciflora* as a test case. BMC Genomics 19:1–15

Ward SM, Webster TM, Steckel LE (2013) Palmer amaranth (*Amaranthus palmeri*): A review. Weed Technol 27:12–27

Waselkov KE, Boleda AS, Olsen KM (2018) A phylogeny of the genus *Amaranthus* (Amaranthaceae) based on several low-copy nuclear loci and chloroplast regions. Syst Bot 43:439–458

Wassom JJ, Tranel PJ (2005) Amplified fragment length polymorphism-based genetic relationships among weedy *Amaranthus* species. J Hered 96:410–416

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Wen F, Wu X, Li T, Jia M, Liu X, Liao L (2021) The complete chloroplast genome of *Stauntonia chinensis* and compared analysis revealed adaptive evolution of subfamily Lardizabaloideae species in China. BMC Genomics 22:1–18

Wheeler TJ, Eddy SR (2013) Nhmmer: DNA homology search with profile HMMs. Bioinformatics 29:2487–2489

Wick RR, Schultz MB, Zobel J, Holt KE (2015) Bandage: Interactive visualization of *de novo* genome assemblies. Bioinformatics 31:3350–3352

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

Xi Z, Liu L, Davis CC (2015) Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol Phylogenet Evol 92:63–71

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, Davis CC (2012) Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci USA 109:17519–17524

Xiao TW, Xu Y, Jin L, Liu TJ, Yan HF, Ge XJ (2020) Conflicting phylogenetic signals in plastomes of the tribe Laureae (Lauraceae). PeerJ 8:1–23

Xu F, Sun M (2001) Comparative analysis of phylogenetic relationships of grain amaranths and their wild relatives (*Amaranthus*; Amaranthaceae) using internal transcribed spacer, amplified fragment length polymorphism, and double-primer fluorescent intersimple sequence repeat. Mol Phylogenet Evol 21:372–387

Xu H, Xiang N, Du W, Zhang J, Zhang Y (2022) Genetic variation and structure of complete chloroplast genome in alien monoecious and dioecious *Amaranthus* weeds. Sci Rep 12:1–9

Yamane K, Yano K, Kawahara T (2006) Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. DNA Res 13:197–204

Yao G, Jin JJ, Li HT, Yang JB, Mandala VS, Croley M, Mostow R, Douglas NA, Chase MW, Christenhusz MJM, Soltis DE, Soltis PS, Smith SA, Brockington SF, Moore MJ, Yi TS, Li DZ (2019) Plastid phylogenomic insights into the evolution of Caryophyllales. Mol Phylogenet Evol 134:74–86

Yu Y, Than C, Degnan JH, Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol 60:138–149

Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:15–30

Zhao F, Chen YP, Salmaki Y, Drew BT, Wilson TC, Scheen AC, Celep F, Bräuchler C, Bendiksby M, Wang Q, Min DZ, Peng H, Olmstead RG, Li B, Xiang CL (2021) An updated tribal classification of Lamiaceae based on plastome phylogenomics. BMC Biol 19:1–27

CHAPTER 4: A PHASED CHROMOSOME-LEVEL GENOME ASSEMBLY PROVIDES INIGHTS INTO THE EVOLUTION OF SEX CHROMOSOMES IN *AMARANTHUS TUBERCULATUS*

ABSTRACT

Amaranthus tuberculatus (waterhemp) is a troublesome weed species of agronomic importance that is dioecious with an XY sex-determination system. The evolution of sex chromosomes, the contiguity of sex-determining region (SDR) and the expression pattern of genes within the SDR remain poorly understood. We assembled the genome of a male *A. tuberculatus*, phased the genome into two chromosome-level haplotypes, and performed restriction site-associated DNA genome-wide association (RAD-GWA) analysis, comparative genomics, adaptive evolution analysis, and, with existing data, transcriptomic profiling to characterize the species' sex chromosomes. Comparative analysis enabled the identification of a ~32.8 Mb SDR on chromosome 1 that is gene-poor, abundant in long terminal repeat (LTR) retrotransposons, and harbors two inversions. Synteny analysis revealed that chromosome 1 likely originated from the fusion of two ancestral chromosomes, and mRNA data indicated 93 genes out of the 531 protein-coding genes within the SDR of haplome 2 were differentially expressed between mature male and female flowers, with several of the genes enriched for Gene Ontology (GO) terms involved in floral development. Beyond adding to our understanding of sex chromosome evolution, the genomic resource provided here will be valuable for addressing further questions on adaptive trait evolution in the *Amaranthus* genus.

4.1 INTRODUCTION

Dioecy, the separation of male and female reproductive systems on different plants, has evolved multiple times independently across many lineages, occurring in as many as 6% of

flowering plant species (Renner and Ricklefs 1995; Ming et al. 2011; Renner 2014), and via several mechanisms (Charlesworth and Charlesworth 1978; Bawa 1980; Lloyd 1980; Henry et al. 2018). One model (two-gene model) postulates that dioecy could evolve via a gynodioecy pathway requiring two mutations (Charlesworth and Charlesworth 1978; Akagi and Charlesworth 2019; Cronk and Müller 2020) while the second model (one-gene model) postulates that dioecy could evolve via a single regulatory factor (Henry et al. 2018). Both models have been supported by the discovery of either two [e.g., in *Actinidia* spp. (Akagi et al. 2019) and *Asparagus officinalis* (Harkess et al. 2020)] or one sex-determining gene(s) [e.g., in *Diospyros lotus* (Akagi et al. 2014) and several *Populus* species (Müller et al. 2020)].

The consequence of these models in the evolution of sex chromosome is that a sex-determining region (SDR) could evolve into a small region (e.g., ~150 kb in *Vitis* spp.) (Massonnet et al. 2020) or a larger non-recombining region that may have different evolutionary histories (i.e., strata) (e.g., 17.42 Mb in *Spinacia oleracea*) (Ma et al. 2022) and have accumulated repetitive sequences due to less effective selection in low recombination regions (Na et al. 2014; Hobza et al. 2015; Charlesworth 2016). The study of sex chromosome evolution has been complicated due to assembly challenges posed by structural variations and repetitive sequences common within SDRs (Charlesworth 2019), and also the possibility of numerous floral development genes acting as the sex-determining gene(s) (Ming et al. 2011). Nevertheless, the availability of long-read sequencing technologies, the ability to now detect genome-wide chromatin interactions (e.g., Hi-C), optical mapping, and overall improvements in computational approaches have made the sequencing and assembly of whole genomes, and more importantly, entire sex chromosomes, possible, thereby facilitating our understanding of sex chromosome

evolution in several species (Akagi et al. 2023; Du et al. 2023; Healey et al. 2023; Kafkas et al. 2023).

Amaranthus tuberculatus (Moq.) J.D. Sauer (waterhemp) is a troublesome dioecious weed of agronomic crops, native to the Midwest of the United States, with a range that has expanded globally (Sauer 1957; Steckel 2007). It is one of nine dioecious amaranths in the subgenus *Acnida* (L.) (Sauer 1957, 1972; Mosyakin and Robertson 1996). Due to the failure of several important herbicide chemistries in managing *A. tuberculatus*, novel approaches focusing on seedbank depletion are being explored, including gene drive to reduce or eliminate populations (Liu et al. 2020; Schleich et al. 2023; Soltani et al. 2023). In general, these genetic control strategies apply more broadly to all dioecious *Amaranthus* species (Tranel and Trucco 2009; Neve 2018); however, utilizing such strategies requires a deeper understanding of the basis of sex determination in the species (Montgomery et al. 2023).

Previous studies investigating sex determination in amaranths confirmed males were the heterogametic sex in *A. tuberculatus* (Murray 1940; Montgomery et al. 2019; Neves et al. 2020), and with indistinguishable chromosomes (Grant 1959). More recently, a draft genome of *A. tuberculatus* assembled into 841 contigs was used to develop genetic markers for genotyping sex (Montgomery et al. 2019). The genome combined with *k*-mer counting was further used to determine male-specific contigs (~4.6 Mb containing 147 gene models) that likely contain the sex-determining region (Neves et al. 2020; Montgomery et al. 2021). This draft assembly together with short-read sequencing from four dioecious amaranths was also used to identify the conservation of male-specificity of a copy of *FLOWERING LOCUS T* (Raiyemo et al. 2023). Furthermore, transcriptomic analysis revealed differentially expressed genes between male and female flowers, including *MADS18*, *MADS2*, *CMB2*, *CYP710A1*, *bHLH60*, and *bHLH91*

(Bobadilla et al. 2023). Due to the fragmentation of the draft assembly, the orientation and order of the male-specific contigs could not be determined. Moreso, none of the genes present on the male-specific contigs were part of the genes that were differentially expressed in the transcriptomic study.

With the rapid advances in sequencing technology, chromosome-level genomic resources have been developed for three monoecious amaranths, *A. hypochondriacus* L. (Lightfoot et al. 2017), *A. cruentus* L. (Ma et al. 2021) and *A. tricolor* L. (Wang et al. 2023), while the dioecious species have draft genomes assembled and scaffolded to pseudochromosome contiguity (Montgomery et al. 2023). Using a combination of PacBio long-read sequencing, Hi-C scaffolding, Bionano optical mapping, and haplotype phasing of a male individual, we generated two high-quality chromosome-level assemblies of *A. tuberculatus*, both of which significantly improve on the previous draft assembly. Comparison of the two haplotype assemblies (haplomes) allowed us to identify a contiguous sex-determining region in the genome, compare structural rearrangements between the two haplomes and to chromosome-level assemblies of three monoecious amaranths, and clarify the positions of previously reported differentially expressed genes between male and female plants.

4.2 MATERIALS AND METHODS

4.2.1 Plant material and growth conditions

Seeds from a herbicide susceptible accession designated “WUS” (GRIN accession number PI 698378) were sown in presoaked potting soil (Lambert LM-GPS) and watered through subirrigation. Upon reaching ~5 cm in height, seedlings were transplanted to 16-cm pots (America Clay Works I-A650MP) filled with the same soil and grown under greenhouse conditions (25/20C and 16/8-h day/night cycles). A single flowering male plant was placed in the

dark for 72 h, after which 4 g of fresh leaf tissue was sampled for PacBio HiFi library preparation and 2 g of fresh leaf tissue for Hi-C library preparation. In each case, tissue was flash frozen in liquid nitrogen and stored at -80 C until use. The same individual plant was used to harvest 4 g of fresh tissue that was immediately shipped on damp paper towels at 4 C for Bionano library preparation. Finally, a mix of four male and four female plants, assigned genotypically as described by (Montgomery et al. 2021), were grown as previously described. Root, stem, leaf, and meristem tissue was sampled from young (~6-10 leaf) and old (flowering) plants and combined with floral tissue from all stages of development for RNA extraction (Zymo Direct-zol RNA Miniprep). Following quantification of concentration and quality, three µg of RNA was used for PacBio Iso-Seq library preparation. All samples (~4 g) were shipped on dry ice (except for Bionano tissue, which was shipped at 4 C) to the Genome Center of Excellence at Corteva Agriscience for DNA extraction, library preparation, and sequencing.

4.2.2 Genome sequencing, assembly, and annotation

The protocols for library preparation, genome sequencing, assembly, and annotation are described in detail in Appendix C Note 1. The methods describing the analyses of genome characteristics including BUSCOs, transposable elements, LTR assembly index, telomeric, and centromeric repeats are provided in Appendix C Note 2.

4.2.3 Genome-wide association analysis for sex in waterhemp populations

A total of 353 individuals (175 females and 178 males) derived from an artificially generated mapping population (Wu et al. 2018) and previously genotyped using RAD-seq (Montgomery et al. 2019) were used for GWA analysis. The single-end raw reads were demultiplexed and cleaned using the "process_radtags" command in Stacks version 2.64 (Rochette et al. 2019). Each sample obtained was aligned to Hap2 of the *A. tuberculatus* genome

assembly using BWA-MEM v0.7.17 (Li 2013). Individuals were grouped by sex, and the "gstacks" command of Stacks was used to build loci from the aligned single-end reads. We first ran the “populations” command employing a minimal filtering (--min-maf 0.05), and then assessed the level of missingness in the two data groups using VCFtools (--missing-indv) following a previously described approach (Cerca et al. 2021). We filtered out samples with greater than 60% missing data from the list of individuals (popmap), and then ran the “populations” command once again on the streamlined samples utilizing additional filtering criteria (--min-maf 0.05, -p 2, -r 0.4). The set of variants obtained were then used for downstream analyses. Principal component analysis was performed using PLINK v1.90b7 (Chang et al. 2015), and GWA analysis was carried out with BLINK-C (Huang et al. 2019) using the first five principal components to account for population stratification and familial relatedness. Sex of the individuals were used as binary input (phenotypes) in the analysis. Calculation of f-statistics (F_{ST}) was performed using VCFtools v0.1.16 (Danecek et al. 2011), and linkage disequilibrium (LD) analysis was carried out using LDBlockShow v1.40 (Dong et al. 2021).

4.2.4 Location of previous sex markers and genes in the assemblies

A search of the previously reported primer sets (WHMS, MU-976, MU-657.2, and MU-533) that amplified a male-specific region of *A. tuberculatus* (Montgomery et al. 2019, 2021) was carried out against both haplotype assemblies with BLASTN (Camacho et al. 2009) using parameters: -task blastn-short -db haplotypes -query primers -outfmt 6 -out output. Reciprocal best hit (RBH) search between genes on the previously identified male-specific contigs and genes from the haplotype assemblies was carried out with MMseqs 2 (Steinegger and Söding 2017).

4.2.5 Synteny and intragenomic analysis

MCScan (Tang et al. 2008) from JCVI utility libraries v1.1.22 was utilized in defining collinear gene blocks between haplotype 1 and 2 assemblies using a C-score cutoff of 0.99. MCScanX (Wang et al. 2012b) was used to investigate duplicated genes in the haplotypes, and the output collinearity files were converted to ".anchors" format using "jcvi.compara.synteny" from JCVI for circos plotting. Sequence alignment between both haplotypes, and also between the haplotypes and a previous *A. tuberculatus* male draft assembly, were carried out with Minimap2 v2.24-r1122 (Li 2021). Alignments were visualized as dotplot using D-Genies (Cabanettes and Klopp 2018). Structural rearrangements between the two haplotypes from the previous alignment were further refined using SyRI (Goel et al. 2019). Syntenic orthologues among the haplotypes and three chromosome-level monoecious *Amaranthus* species (*A. hypochondriacus*, *A. cruentus* and *A. tricolor*) were evaluated and visualized using GENESPACE v1.2.3 (Lovell et al. 2022).

4.2.6 Sequence divergence and detection of positive selection

We followed previously described protocols to test for the evidence of adaptive evolution for 496 single-copy orthologous genes on chromosome 1 (Jeffares et al. 2014; Álvarez-Carretero et al. 2023). Briefly, each orthogroup on chromosome 1 from the previous orthofinder run were aligned separately using custom perl scripts, "multiple_sequence_spliiter.pl" and "align_orthologs.pl" from Jeffares *et al.*, (2014). Synonymous (d_S) substitution rates between the Hap1- and Hap2-linked genes (i.e., pairwise comparison using the aligned single-copy orthologous genes) were estimated with the maximum-likelihood (ML) method of Goldman & Yang (1994) implemented in the CODEML program (runmode = -2, CodonFreq = 2) in PAML package v4.10.7 (Yang 2007). We included chromosomes 4, 7, and 16 in the analysis to evaluate synonymous divergence for autosomes. Eight genes representing outliers were each removed on

chromosomes 1 and 4. We then proceeded to test for evidence of adaptive evolution for genes on chromosome 1 using a series of models. We compared the simplest or null model against a nested alternative site model (M0 vs M1a) that allows neutral sites ($\omega = 1$) to be evaluated. We then asked if adding a third class with $\omega > 1$ fits the data better than a model with only two classes, $\omega < 1$ or $\omega = 1$ (M1a vs M2a), as certain sites could be under positive selection.

Given the alignment of the single-copy genes on chromosome 1 of chromosome-level assembly of three monoecious amaranths and the two haplotypes, we used the branch models to determine if ω differs among lineages (M0 vs free-ratio), and if specific foreground branches have different ω from background branches (M0 vs two-ratio). Using the branch-site model (MA_{null}, $\omega = 1$ vs MA, $\omega > 1$), we determined if the previously defined foreground branches are more likely to contain sites under positive selection (see Fig. S14 of schematic representation of trees with defined foreground and background branches). For all hypotheses tested, we implemented the mutation-selection model (FMutSel) with observed codon frequencies used as estimates (CodonFreq = 7, estFreq = 0). The model has been indicated to account for mutation bias and selection affecting codon usage, and is preferable over other models (Yang and Nielsen 2008; Álvarez-Carretero et al. 2023). *P*-values were adjusted using the Benjamini–Hochberg method (Benjamini and Hochberg 1995) to account for multiple comparisons.

4.2.7 Expression profiling and gene ontology (GO) enrichment analysis

mRNA-sequencing data for three tissue types (mature flowers, shoot apical meristem, and floral meristems) from a previous study (Bobadilla et al. 2023) were mapped to Hap2 of the *A. tuberculatus* genome using STAR aligner v2.7.10b (Dobin et al. 2013). Two replicates from mature flower category were removed due to low mapping quality (26.47% and 28.44% uniquely mapped reads for replicate 2 and 4, respectively). Both samples were also removed from

downstream analyses in the previous study. Gene counting was carried out using featureCounts v2.0.6 from the subread package (Liao et al. 2014), and the differential expression (DE) analyses between sex for each tissue types were carried out with edgeR (Robinson et al. 2009). Prior to the DE analysis, genes with zero counts were filtered (i.e., CPM values less than 1), and counts were normalized using the TMM normalization in edgeR. A negative binomial generalized log-linear model was then fitted to the normalized read counts. Genes were assessed as differentially expressed based on FDR < 0.05 and FC > 1.2 thresholds using the '*glmTreat*' function within edgeR. Translated protein-coding sequences (CDS) of Hap2 were also assigned GO annotations using eggNOG-mapper v2.1.12 (Cantalapiedra et al. 2021). GO term enrichment analysis was then carried out using topGO with nodeSize = 10, which is the minimal number of genes to keep per term. The enrichment test was performed using Fisher's exact test and the "elim" algorithm. GO terms were assessed as significantly enriched at the default $p < 0.01$ threshold.

4.2.8 Phylogenetic analysis of *FLOWERING LOCUS T*

A 200 bp *FT* sequence, previously reported as male-specific and conserved among three other dioecious amaranths closely related to waterhemp, was searched against the haplotype assemblies using BLAST (Camacho et al. 2009). Failure to obtain perfect match to the candidate sex chromosome (Chr1) prompted further investigation. We queried the haplotype assemblies and three monoecious amaranth (*A. hypochondriacus*, *A. cruentus* and *A. tricolor*) assemblies for all homologs of *FLOWERING LOCUS T*, and also searched the 200 bp *FT* sequence against raw reads of *A. tuberculatus* from a previous study (Kreiner et al. 2019) using SRA-BLAST (Leinonen et al. 2011). The top hit (ranked by evalue and bitscore) and another hit were selected from each male while only the top hit for females were selected. All the *FT* sequences were aligned using MAFFT version 7 online (Katoh and Standley 2013), and columns with less than

15% occupancy in the alignment were removed with Jalview v2.11.3.2 (Waterhouse et al. 2009). A maximum likelihood tree was then constructed with the alignment in RAxML v8.2.12 (Stamatakis 2014) using a GTRGAMMA substitution model and 1000 rapid bootstrap replicates. The resulting tree was visualized with Dendroscope v3.8.10 (Huson and Scornavacca 2012).

4.2.9 Statistical analyses

The sample sizes used for analyses are indicated in the figures. Gene count per 500 kb, TE proportion per 500 kb, and LTR insertion times data were analyzed with the non-parametric Kruskal-Wallis test. Post-hoc test of multiple comparisons was carried out with Conover-Iman test, following the rejection of the null hypothesis from the Kruskal-Wallis test. *P*-values were adjusted for multiple comparisons with Benjamini–Hochberg correction (Benjamini and Hochberg 1995). The pairwise synonymous divergence (d_S) between species (i.e., the two haplotypes and three monoecious amaranths) previously estimated with CODEML were also analyzed following the method described above, except Dunn's test was used for multiple comparisons. All analyses were carried out with the PMCMRplus package in R v4.1.2 (R Core Team 2021).

4.3 RESULTS

4.3.1 Assembly metrics and genome repetitive landscape

Analysis of both haplome 1 (Hap1) and haplome 2 (Hap2) assembly completeness using BUSCO "embryophyta_odb10" database revealed 96.9% complete BUSCOs. LTR assembly index (LAI) also revealed high quality assemblies with average LAI scores for Hap1 and Hap2 at 19.19 and 20.49, respectively (Table 4.1). Repeat analysis using RepeatMasker revealed that 68.6% of Hap1 and 66.3% of Hap2 were made up of repetitive elements. The LTR/*Ty3* elements were the most abundant retrotransposons, representing 18.1% and 16.3% of the genome in Hap1

and Hap2, respectively (Table S1). The total repeat content and the abundance of *Ty3* elements are consistent with the 66.0% total repeats and 17.0% *Ty3* elements reported for *A. tuberculatus* draft genome assembly (Raiyemo et al. 2023).

Analysis of high-copy tandem repeats using StainedGlass heatmaps revealed that chromosomes 1, 2, 3 4, 6, and 8 appear to be submetacentric while chromosomes 9, 10, 11, 12, 14, 15, and 16 appear to be telocentric in both haplomes (Appendix C Figure C.1 and Appendix C Figure C.2). Some regions identified as centromeric by StainedGlass were also identified as centromeric from CentroMiner (Appendix C Table C.2). BLAST search of the simple telomeric repeat, TTTAGGG against both haplotypes revealed telomeric repeat sequences at 29 and 28 out of the possible 32 telomeric ends for Hap1 and Hap2, respectively (Appendix C Table C.3). The number of *A. tuberculatus* telomeres assembled is comparable to 30 out of 34 telomeric ends reported for *A. tricolor* (Wang et al. 2023).

Further annotation of the genome revealed Hap1 and Hap2 had 1,373 and 1,344 genes annotated as transcription factors (TF), respectively (Appendix C Table C.4 and Appendix C Table C.5). In addition, Hap1 had 1,346 genes annotated as disease resistance genes while Hap2 had 1,348 genes annotated as such (Appendix C Table C.6 and Appendix C Table C.7). Annotation of transfer RNAs revealed 2,096 tRNA genes in Hap1 and 2,026 tRNA genes in Hap2 (Appendix C Table C.8). Other non-coding RNAs in Hap1 (90 miRNAs, 890 rRNAs, 176 snRNA, 900 snoRNA) and Hap2 (93 miRNAs, 579 rRNAs, 194 snRNA, 979 snoRNA) were also annotated (Appendix C Table C.9).

Visualization of the genomic features analyzed above indicates an inverse relationship between gene density and LTR proportions, whereby gene-rich regions are LTR poor and vice versa (Figure 4.1).

4.3.2 Identification of the candidate sex-determining region

Preliminary analysis to filter out individuals with more than 60% missing data retained 44 female and 48 male samples, and the number of variants also reduced from 558,762 to 204,641. Genome-wide association (GWA) analysis with the 92 individuals revealed that the most significant single-nucleotide polymorphisms (SNPs) associated with sex were located on chromosome 1 and spanned from 13.95 – 46.60 Mb (Figure 4.2a). All identified significant SNPs resided within intergenic regions (Table C.10). A QQ plot revealed no evidence of systematic bias (e.g., from analytical method, model choice, genotyping error, or population structure) in the GWA analysis (Figure 4.2b). Genetic differentiation (F_{ST}) along chromosome 1 between females and males above the top 5% threshold spanned from 14.65 – 42.6 Mb, with a second peak at 51.6 – 51.7 Mb (Figure 4.2c and Figure 4.2d). Although no clearly defined linkage blocks were observed, the sex-determining region identified above corresponds to the region indicated as containing SNPs in possible linkage disequilibrium (Figure 4.2e). The region between 0 – 13.7 Mb did not show much differentiation, and thus could be considered a pseudoautosomal region (PAR) that is still actively recombining between the haplomes (Figure 4.2c and Figure 4.2d). Considering the lines of evidence above, we defined an approximate boundary of the sex-determining region (SDR) as a region of chromosome 1 on Hap2 between 13.8 – 46.6 Mb (~ 32.8 Mb), and the Hap1 equivalent is between 14.88 – 46.0 Mb (~ 31.12 Mb).

A BLAST query of previously reported primer sets, used to amplify male-specific regions (Montgomery et al. 2021), against both haplotypes also revealed perfect matches to chromosome 1 on Hap2. One set of primers (WHMS) matched to a 572 bp region (31,810,427 – 31,810,999 bp) (Figure 4.3a) while primer sets MU-976, MU-657.2, and MU-533 also matched to regions within the SDR on Hap2 (Appendix C Table C.11). The 572 bp male-specific marker

had homology to an LTR/*Ty3* retrotransposon (Appendix C Figure C.3). Similarly, MU-657.2 was previously reported to have homology to a LTR/*Ty3* element from sugar beet (Montgomery et al. 2019). Sequence alignment of the previous draft *A. tuberculatus* male assembly to both haplotypes and visualization of the identified male-specific Y contigs indicated several of the contigs aligned to chromosome 1 in both haplotypes (Appendix C Figure C.4 – C.10). Although tig00000542 was previously proposed as one end of the male-specific Y region while tig00100752 was the other end (Montgomery et al. 2021), our alignment revealed that tig00000298 and tig00000455 precedes tig00000542 while tig00100752 precedes tig00000336 and tig00000340 (Appendix C Figure C.5 – C.10). Structural variation between waterhemp populations could have caused these differences in the contig positions, as the previous draft genome was assembled from an individual selected from a different population than the new assembly.

There are 2,140 and 2,077 protein-coding genes on chromosome 1 of Hap1 and Hap2, of which 528 and 532 are within the SDR, respectively (Table 4.2). Although 5 microRNAs were identified on chromosome 1 in which three were within the putative SDR, none were haplotype-specific (Table C.12). Reciprocal best hit search with the 147 genes on the previously identified male-specific contigs and the haplotype assemblies revealed 13 and 18 genes had best matches on chromosome 1 of Hap1 and Hap2 assemblies, respectively (Table C.13). The genes are thus conserved between the draft assembly and the new assembly. However, 14 of the genes were syntenic between Hap1 and Hap2 based on our analysis from GENESPACE. Only 4 genes appear to be Hap2-specific with no syntenic hits to Hap1; however, these genes were annotated as proteins of unknown functions (Table C.13). Taken together, the line of evidence above supports chromosome 1 as the likely sex chromosome.

4.3.3 Comparative analysis between the SDR of the two haplomes

Synteny patterns between the two haplomes indicate a 1:1 relationship in gene content (Appendix C Figure C.11). Out of the 20,027 gene pairs representing the reciprocal best matches between the two haplotypes, there were 1,752 pairs for chromosome 1, which were used to define the boundaries of regions within the SDR. The synteny analysis revealed two inversions on either side of a collinear region within the SDR (Fig. 3a). One inversion (INV 1) is from 15,684,914 – 31,008,021 bp on Hap1, and from 17,851,790 – 31,641,701 bp on Hap2, based on the first and last gene pairs for the inversion. The other inversion (INV 2) ranges from 33,192,221 – 39,549,589 bp on Hap1, and from 34,354,662 – 39,995,341 bp on Hap2. Further analysis of structural rearrangements using SyRI revealed 262 inversions spread across the 16 chromosomes between Hap1 and Hap2 (Table C.14). INV 1 and INV 2 within the SDR were the two largest inversions out of the 262 inversions identified using SyRI (Figure 4.3a).

We observed a region that was not syntenic between Hap1 and Hap2 upstream of INV 1 on Hap2 and sought to determine if the region was a translocation or duplication from elsewhere in the genome. The boundary of this region within the SDR was defined as the difference between the end coordinate of the last gene in collinear block 1 (14,660,708 bp) and the start coordinate of the first gene for INV 1 on Hap2 (17,851,789 bp), and the presence of the difference above within regions identified as “not aligned” by SyRI. Out of 14,792 regions designated as “not aligned” between the two haplomes, chromosome 1 had the highest span of a region not aligned (~1.87 Mb; 15,186,930 – 17,059,186) (Table C.17). We therefore took the “non-syntenic” region as spanning a length of 3,191,081 bp (~3.19 Mb) on Hap2. Further analysis revealed that 1,797,397 bp (56.33%) of the region is made up of Ns and we designated the region as a “gap” (Figure 4.3a).

Analysis of the genomic architecture of chromosome 1 revealed gene density for the regions (collinear 1, INV 1, collinear 2, INV 2, and collinear 3) within both Hap1 and Hap2 were statistically different with $p < 3.35e\text{-}14$ and $p < 1.81e\text{-}10$, respectively. Among pairwise comparisons of regions, only collinear 1 or 3 region had significantly higher gene densities relative to either INV 1, collinear 2, or INV 2 regions for both haplotypes (Figure 4.3b, Table C.16). LTR proportions for the regions were also statistically different ($p < 4.37e\text{-}11$ and $p < 1.60e\text{-}09$, respectively). Pairwise comparison also indicated that the LTR proportions for INV 1, collinear 2, or INV 2 were significantly higher than that of collinear 1 or 3 (Figure 4.3c, Table C.17). Therefore, collinear 1 and 3, outside of the SDR, are gene-rich but LTR-poor while the SDR region (INV 1, collinear 2, and INV 2) is LTR-rich but gene-poor (Figure 4.3b and Figure 4.3c). Although analysis of insertion times revealed slight variation in the time of insertion of intact LTR retrotransposons (LTR-RTs) for collinear 2 and INV 2 relative to the other regions (Figure 4.3d, Table C.18), both haplotypes appear to have had several of their intact LTR-RTs inserted into the regions less than 0.5 Mya, as seen in the peaks around this time (Figure 4.3e).

Analysis of structural rearrangements between the haplotype assemblies and three monoecious amaranths indicates a highly conserved gene order, except for a few chromosomes (Figure 4.4a). Chromosome 1 in *A. tuberculatus* appears to have originated from the fusion of two chromosomes that are ancestral to chromosomes 13 and 16 from *A. hypochondriacus* and *A. cruentus*, and chromosomes 10 and 16 in *A. tricolor* (Figure 4.4b and Figure 4.4c). While the order of genes appears to have been maintained between monoecious species and Hap1 (Figure 4.4b), the inversions discussed above appear to have occurred on Hap2 (Figure 4.4c).

4.3.4 Sequence divergence and detection of genes under positive selection

Pairwise synonymous divergence (d_s) was estimated between single-copy genes on chromosome 1 for the five regions (collinear 1, INV 1, collinear 2, INV 2, and collinear 3) of the haplomes and also between their orthologs in three monoecious *Amaranthus* species. The inclusion of randomly selected chromosomes 4, 7, and 16 in the analysis allowed us to assess synonymous divergence in autosomes. The mean d_S among the chromosomes including chromosome 1 were similar, with consistency across species comparisons (Appendix C Figure C.12). On chromosome 1, the mean d_S for collinear 1, INV 1, collinear 2, INV 2, and collinear 3 for Hap1 vs. Hap2 comparison were 0.0426, 0.0283, 0.0399, 0.0308, and 0.0451, respectively. Comparisons between the five regions showed INV 1 d_S was significantly lower than that of collinear 1 (p = 0.0372) and collinear 3 (p = 0.0142) (Table C.21), indicating that recombination suppression within INV 1 occurred more recently. The higher mean d_S for collinear 1 and 3, and the lack of statistical evidence that the two regions differ between haplotypes, indicates that the two regions are still recombining. Both regions had a comparable mean d_S to Chr4 (0.0425), Chr7 (0.0471) and Chr16 (0.0508) (Appendix C Figure C.12). For all pairwise species comparisons, collinear 1 and 3 were not statistically different in mean d_S; however, collinear 2 and inversion 2 tend to exhibit more variation, indicating a higher sequence divergence for both regions relative to the other regions (Appendix C Figure C.13). The mean d_S of Hap1 or Hap2 to *A. tricolor* comparison was lower across the five regions, relative to the mean d_S of Hap1 or Hap2 to *A. cruentus* or to *A. hypochondriacus* comparison, supporting previous phylogenetic evidence that indicated *A. tricolor* (subgenus *Albersia*) is more related to *A. tuberculatus* (subgenus *Acnida*) than to the other monoecious species in the subgenus *Amaranthus* (Waselkov et al. 2018; Raiyemo et al. 2023; Wang et al. 2023).

Analysis of adaptive evolution using the M1a vs. M2a model revealed 17 genes out of 504 single-copy genes on chromosome 1 were significant ($\alpha = 0.05$) for sites under positive selection (Table C.20). Using the branch model, M0 vs. free-ratio, where omega (ω) was allowed to vary, there were 20 significant genes with different ω among the lineages. When Hap1 was used as the foreground branch to determine if ω differs for this branch relative to the background branches (i.e., the M0 vs. two-ratio model), there were no genes with significantly different ω. However, when Hap2 was used as the foreground branch, only one gene encoding a serine/arginine-rich SC35-like splicing factor *SCL33* had significantly different ω, indicating it confers some fitness benefit. When both haplomes were specified as the foreground branch [i.e., (Hap1 #1, Hap2 #1)], only two genes, encoding bark storage protein B and thaumatin-like protein, had ω that differed for the specified branch. However, when the branch leading to the common ancestor of the two haplomes were included in the foreground branch (i.e., (Hap1 #1, Hap2 #1) #1)], seven genes encoding uncharacterized protein LOC110699187, bark storage protein B, LOB domain-containing protein 19-like, LRR receptor-like serine/threonine-protein kinase, thaumatin-like protein, and two copies of polygalacturonase had ω that differed for the foreground branch compared to the background branches.

Adopting the foreground and background branches used for the branch model above but utilizing a branch-site model (i.e., MA_{null}, $\omega = 1$ vs. MA, $\omega > 1$), 10 genes were more likely to contain sites with $\omega > 1$ when Hap1 was used as the foreground branch. When Hap2 was used as the foreground branch, 16 genes were more likely to contain sites with $\omega > 1$. When both haplomes were used as the foreground branch, 17 genes were more likely to contain sites with $\omega > 1$. When the branch leading to the common ancestor of the two haplomes was included, 22 genes were more likely to contain sites with $\omega > 1$ (Appendix C Figure C.14, Appendix C Table

C.20, and Appendix C Table C.21). Within the SDR, genes for LOB domain-containing protein 19-like, vacuolar ion transporter homolog 4-like, DEAD-box ATP-dependent RNA-helicase 39, zinc finger CCCH domain-containing protein 18-like, and prefoldin subunit 3 were more likely to be under positive selection based on the branch-site model. Taken together, our analysis above reveals genes that are potentially important in sex-specific adaptation.

4.3.5 Expression analysis identifies genes involved in floral development

Clean mRNA reads (i.e., adapter trimmed and low-quality bases removed) from three tissue types (shoot apical meristem, floral meristem, and mature flower) reported by Bobadilla *et al.*, (2023) that were mapped to the Hap2 assembly had uniquely mapped reads for males ranged from 82.42 – 88.22%, and uniquely mapped reads for females ranged from 74.37 – 89.05%. Out of the 23,160 annotated Hap2 protein-coding genes, 20,259 gene were retained for DE analysis after filtering and TMM normalization (Figure 4.5). Among the 1,794 genes retained on chromosome 1, two at the shoot apical meristem stage, four at the floral meristem stage, and 435 (231 upregulated and 204 downregulated) at the mature flower stage were differentially expressed between males and females (Table C.22 – C.24). Within the SDR on chromosome 1, two genes at the shoot apical meristem stage, three genes at the floral meristem stage, and 93 genes (55 upregulated and 38 downregulated) at the mature flower stage were differentially expressed (Table C.22 – C.24). Two genes within the SDR (specifically within INV 1) encoding MADS-box protein FLOWERING LOCUS C-like and LOB domain-containing protein 19-like were consistently downregulated across the three tissue types for male plants (Table C.23 and Table C.24).

Gene ontology (GO) term enrichment analysis was performed to gain insight into biological processes that could be involved in sex determination. DEGs were selected based on

an FDR threshold of $p < 0.05$ and FC > 1.2. Biological processes including pollen tube growth, regulation of cell development, cellular component morphogenesis, anther wall tapetum development, pollen germination, and brassinosteroid-mediated signaling pathway were identified among the top 20 enriched GO terms (Figure 4.5, Table C.25). *FLOWERING LOCUS C*-like and other genes including *agamous*-like *MADS-box AGL15*, *transcription factor CYCLOIDEA*-like, *transcription factor MYB44*-like, *two-component response regulator ORR24*, transcription factor *bHLH118*, probable *WRKY transcription factor 23*, and *zinc finger protein WIP2*-like were part of the "regulation of transcription factor, DNA-templated" enriched GO term (Table C.26). The top five terms enriched for molecular function included transmembrane receptor protein serine/threonine kinase activity, mannan synthase activity, calmodulin binding, GTPase activator activity, and phosphatidylinositol binding (Table C.27), while the top five terms enriched for cellular function included plasma membrane, apical plasma membrane, pollen tube, actin filament, and endomembrane system (Table C.28).

4.3.6 Male-specificity of *FLOWERING LOCUS T*

BLAST search of the 200 bp *FT* sequence, previously reported as male-specific (Raiyemo et al. 2023), to the new assembly revealed no perfect matches, but 89.55% homology to a gene annotated as *HEADING DATE 3A* on chromosome 15 of both haplotypes (Hap1: 16,630,366 – 16,630,566, and Hap2: 14,349,263 – 14,349,463). Phylogenetic analysis using all homologs of *FLOWERING LOCUS T* or *HEADING DATE 3A* from the haplotype assemblies, three monoecious amaranths, and fragmented copies (< 150 bp) obtained from resequenced *A. tuberculatus* individuals via SRA-BLAST (Table C.29) revealed that the copy we previously identified to be male-specific and conserved in three species closely related to *A. tuberculatus* was present in the draft assembly but absent in the new assembly (Appendix C Figure C.15). A

region of tig00000542 (1 – 8222 bp) where the *FT* is located within the draft assembly did not map to any region of either Hap1 or Hap2 assemblies, further indicating that the copy of the *FT* was not assembled in the new genome and could be specific to some populations.

4.4 DISCUSSION

We present a high-quality haplotype-resolved assembly of *A. tuberculatus*, representing the first chromosome-level assembly in the subgenus *Acnidia,* which consists of all the dioecious species in the *Amaranthus* genus. We provide evidence that chromosome 1 of the assembled genome is the sex chromosome; harboring the two largest inversions in the genome within a ~32.8 Mb region that is gene-poor but abundant in LTR retrotransposons. Since the regions flanking the SDR are gene-rich and LTR-poor, it is possible that they are still actively recombining between the haplomes. There has been speculation over the correlation between chromosome sizes and peripheral recombination (Brazier and Glémin 2022), with evidence in *S. latifolia* indicating that large chromosomes tend to have peripheral recombination (Yue et al. 2023). Although chromosome 1 of the *A. tuberculatus* assembly is the largest chromosome (57.6 Mb), our conjecture on peripheral recombination within the flanking pseudoautosomal region is based on evidence from synteny, and on the negative relationship between gene and LTR density.

Our analysis of high-order repeats typical of centromeric regions (Melters et al. 2013) indicates that chromosome 1 is either submetacentric or metacentric with the first inversion (INV 1) located in the pericentromeric region of the chromosome. Inversions have been observed within sex chromosomes of different species, including plants and animals (Wang et al. 2012a; Natri et al. 2019; Hearn et al. 2022; Ma et al. 2022), and are known evolutionary drivers of sex determination systems (Natri et al. 2019). Chromosome 1 also appears to have originated from

the fusion of two ancestral chromosomes following the divergence of the *Acnidia* subgenus from *Albersia*. The fusion appears to have also occurred near the centromere, perhaps as an additional mechanism to suppress recombination. Whether the chromosomal fusion event is specific to the *Acnidia* subgenus or arose independently in *A. tuberculatus* remains unclear without more chromosome-level assemblies within the *Amaranthus* genus.

The similar synonymous substitution (d_S) rates between the haplomes for flanking regions, collinear 1 and 3 of the SDR and to autosomes, suggest that the two regions are still recombining, like the autosomal regions. Much of the variation in d_S across species comparisons from collinear 2 and inversion 2 thus reflects the expansion of the region and accumulation of non-coding sequences due to suppressed recombination. Considering that inversion 1 has the lowest d_S among the regions between the two haplotypes, this inversion likely occurred more recently. With no clear differences between the two haplotypes and given the similar numbers of genes present on both haplomes, *A. tuberculatus* may not have an extensive completely Y-linked region that has undergone genetic degeneration leading to loss of gene functions and deletions of genes. A similar scenario of no completely Y-linked region has been reported for spinach (Ma et al. 2022). Alternatively, the Y-haplotype (i.e., Hap2) may be missing Y-specific sequences, considering the ~1.8 Mb gap region in the assembly.

Comparison between this work and previous studies points to the influence of assembly choice on inferences drawn in sex chromosome studies. The absence of the *FT* gene fragment in this assembly that we previously reported as male-specific could either mean that it was not assembled or was specific to the population that was sequenced for the draft assembly. A search of the *FT* gene fragment in the Iso-seq data using a relaxed BLAST search parameter resulted in <90% homology to another copy. Although several authors now consider waterhemp as a single

species (Waselkov and Olsen 2014; Iamonico 2020), it is worth noting that prior to Pratt & Clark (2001), it was considered two species, *A. tuberculatus* (primarily east of the Mississippi River) and *A. rudis* (primarily west of the Mississippi River) (Sauer 1955, 1972) and some authors still consider them varieties, with var. *rudis* being the more weedy variety (Costea and Tardif 2003). The population used for the draft assembly, designated ACR, was a weedy population from Illinois, while the population used for the new assembly, designated WUS, was from a riparian population in Ohio.

Reanalysis of the mRNA data from Bobadilla *et al.*, (2023) revealed consistent expression patterns between this work and the previous study. The two genes encoding a MADS-box protein FLOWERING LOCUS C-like and LOB domain-containing protein 19-like we reported as downregulated in males across three tissue types were also reported as downregulated in the previous study. We, however, clarify that the two downregulated genes were annotated as MADS-box transcription factor 18 (*MADS18*) and *LOB domain-containing protein 31* in that study (Fig. S16). Six genes (encoding ethylene-responsive transcription factor RAP2-7-like, CBP-diacylglycerol-glycerol-3-phosphate, protein CLT2 chloroplastic, TBC1 domain family member 8B-like, and two uncharacterized proteins) present in both the previously identified male-specific contigs of the draft assembly and within the SDR in the new assembly were also differentially expressed. Although the genes had a small difference in expression given the fold-change threshold (FC > 1.2) used in our analysis, we note that further increasing the fold-change threshold would have resulted in the same conclusion as Bobadilla et al. where none of the genes on the male-specific contigs were differentially expressed.

In sum, our study provides valuable insights into the evolution of sex chromosomes in a dioecious weedy species and adds genomic resources to the genus *Amaranthus*. We report sex-

linked as well as sex-biased genes with potential roles in sex determination and flowering. Functional validation of several of the genes elicited in this study could pave the way for potential candidates for a genetic control strategy. Also, intraspecies variation in sex chromosomes or sex-determining region are being observed in some species, such as in spinach (Ma et al. 2022; She et al. 2023) and the human Y chromosomes (Hallast et al. 2023). Future research could utilize a comparative genomics approach with other dioecious species within the genus, as well as the assembly of multiple populations of *A. tuberculatus* to further understand the variation and evolution of the sex-determining region.

4.5 TABLES AND FIGURES

Table 4.1 Comparison of assembly statistics between the two haplomes of *A. tuberculatus* genome assemblies.

Genome characteristics	**Hap1**	**Hap2**
Assembly size (Mbp)	613.8	588.8
Scaffold N50 (Mbp)	38.92	36.16
Scaffold L50	7	7
GC content (%)	35.0	34.7
Complete BUSCO (%)	96.9	96.9
Size of Ns (Mbp)	8.10	9.03
LTR assembly index (LAI)	19.19	20.49
Protein-coding genes	23,253	23,160
Mean gene length (bp)	5,371	5,347
Mean CDS length (bp)	1,195	1,198
Mean exon length (bp)	284	285
Mean exon per gene	5.7	5.7
Number of tRNA	2,096	2,026
Number of genes in orthogroups	22,942	22,905
Percentage (%) of genes in orthogroups	98.7	98.9

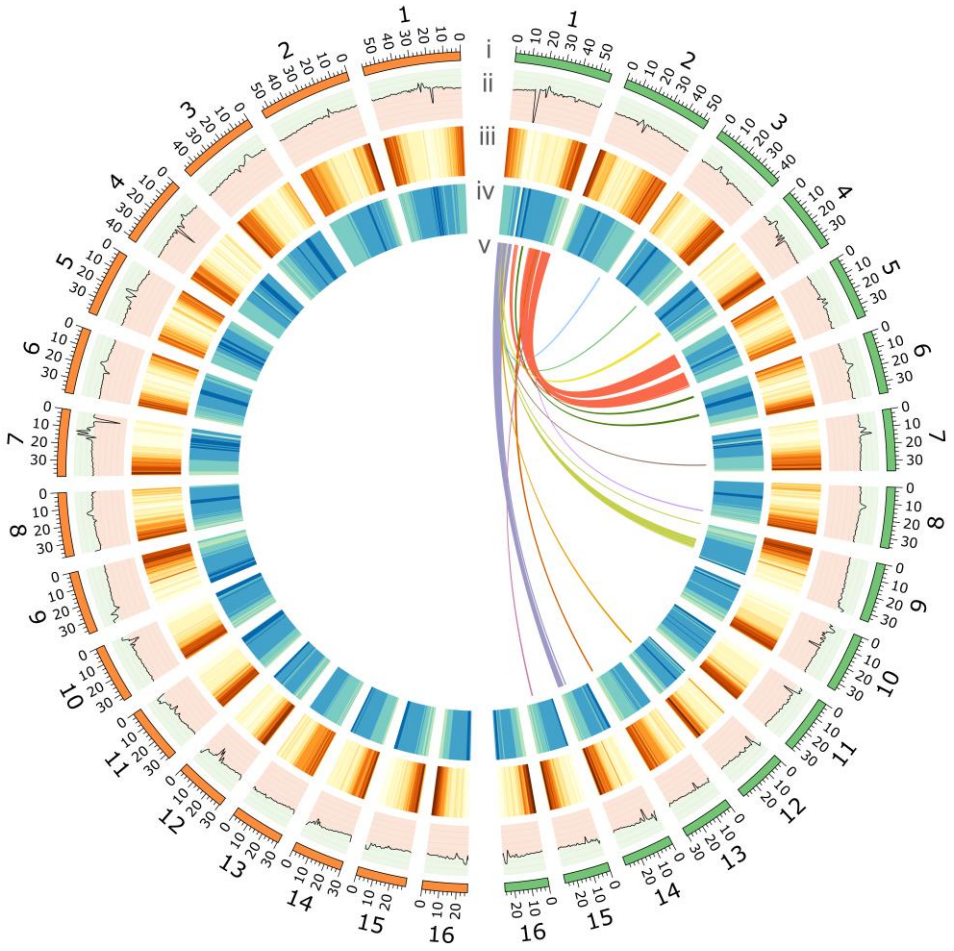

Figure 4.1 Genomic features of *A. tuberculatus* Hap1 (left) and Hap2 (right) assemblies. Circos plot depicts i) number and length (Mb) of chromosomes, ii) GC content along the chromosomes, with peaks in light green area representing GC content greater than the median (> 0.339317) and peaks in light red area representing GC content less than the median (< 0.339317), iii) gene density across the chromosomes, with brown representing gene-rich regions and yellow representing gene-poor regions, iv) LTR (long terminal repeats) density along chromosomes, with blue representing LTR-rich regions and green representing LTR-poor regions, v) inner ribbons represent duplicated genes on chromosome 1 of Hap2. Duplicated genes on chromosome 1 of Hap1 are not shown in the figure to avoid redundancy. Window size of 1 Mb and step size of 500 kb for ii-iv.

Table 4.2 Summary statistics for chromosome 1, sex-determining region (SDR), collinear regions and inversions (INV 1 and INV 2) for the two haplotype assemblies.

		Hap1	**Hap2**
Length (bp)	Chromosome 1	57,604,171	57,035,016
	SDR	31,116,274	32,798,384
	INV 1	15,323,107	13,789,911
	INV 2	6,357,368	5,640,679
Number of protein-coding genes	Chromosome 1	2,140	2,077
	SDR	528	531
	INV 1	188	186
	INV 2	114	111

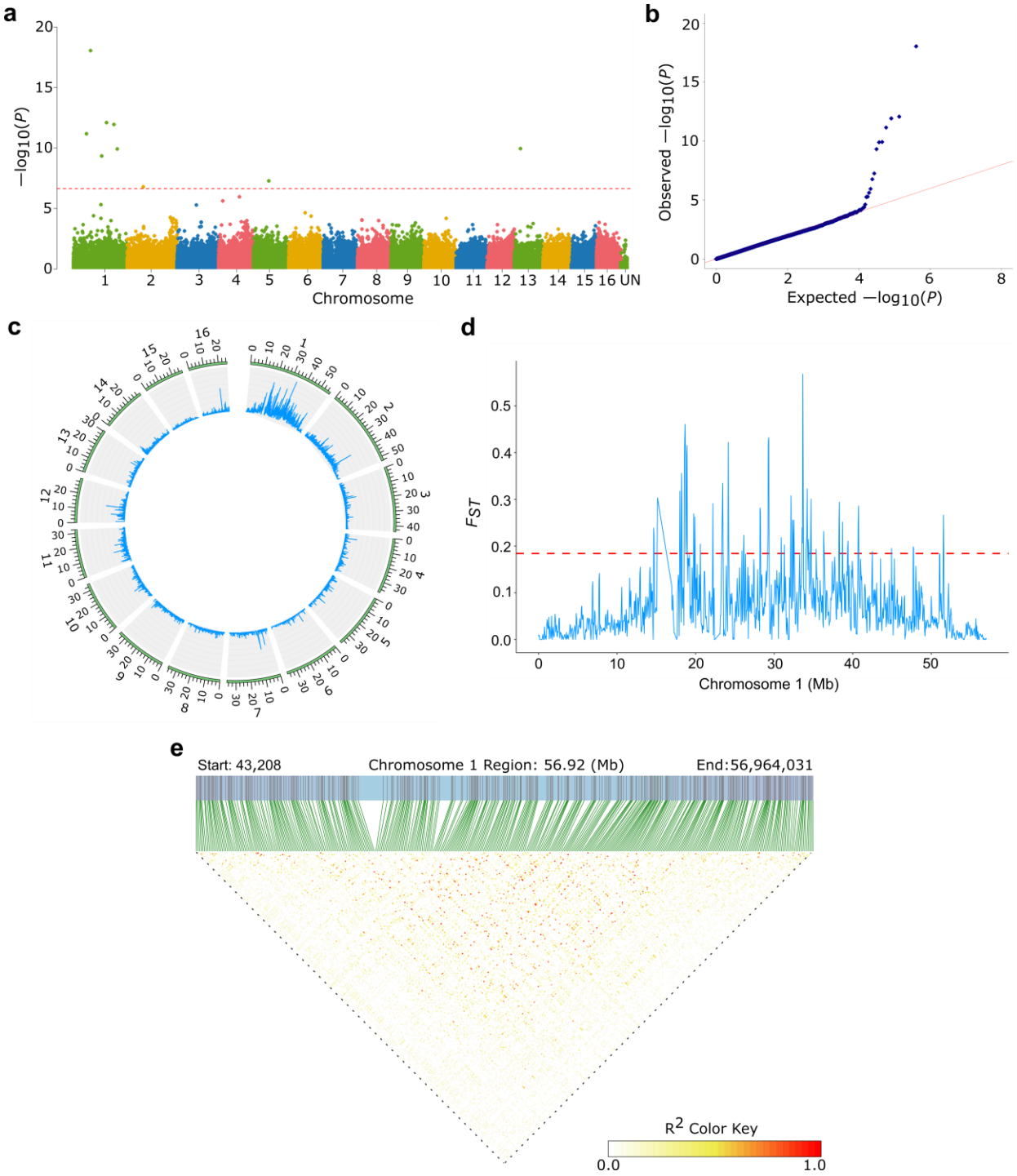

Figure 4.2 Identification of sex-determining region on chromosome 1. **a** Manhattan plot of GWA analysis using RAD-seq data from 44 females and 48 males. The dashed red line indicates Bonferroni threshold of $-\log_{10}(P) = 6.6120$. **b** Quantile-quantile (QQ) plot of the GWA analysis. **c** Fixation index (F_{ST}) between females and males across all 16 chromosomes (window size 100

Figure 4.2 (Cont.)

kb; step size 50 kb). **d** F_{ST} between females and males for chromosome 1 with dashed red line representing the top 5% threshold at 0.1841 (window size 100 kb; step size 50 kb). **e** plot of linkage disequilibrium analysis using 375 pruned SNPs at 1 SNP/50 kb across chromosome 1.

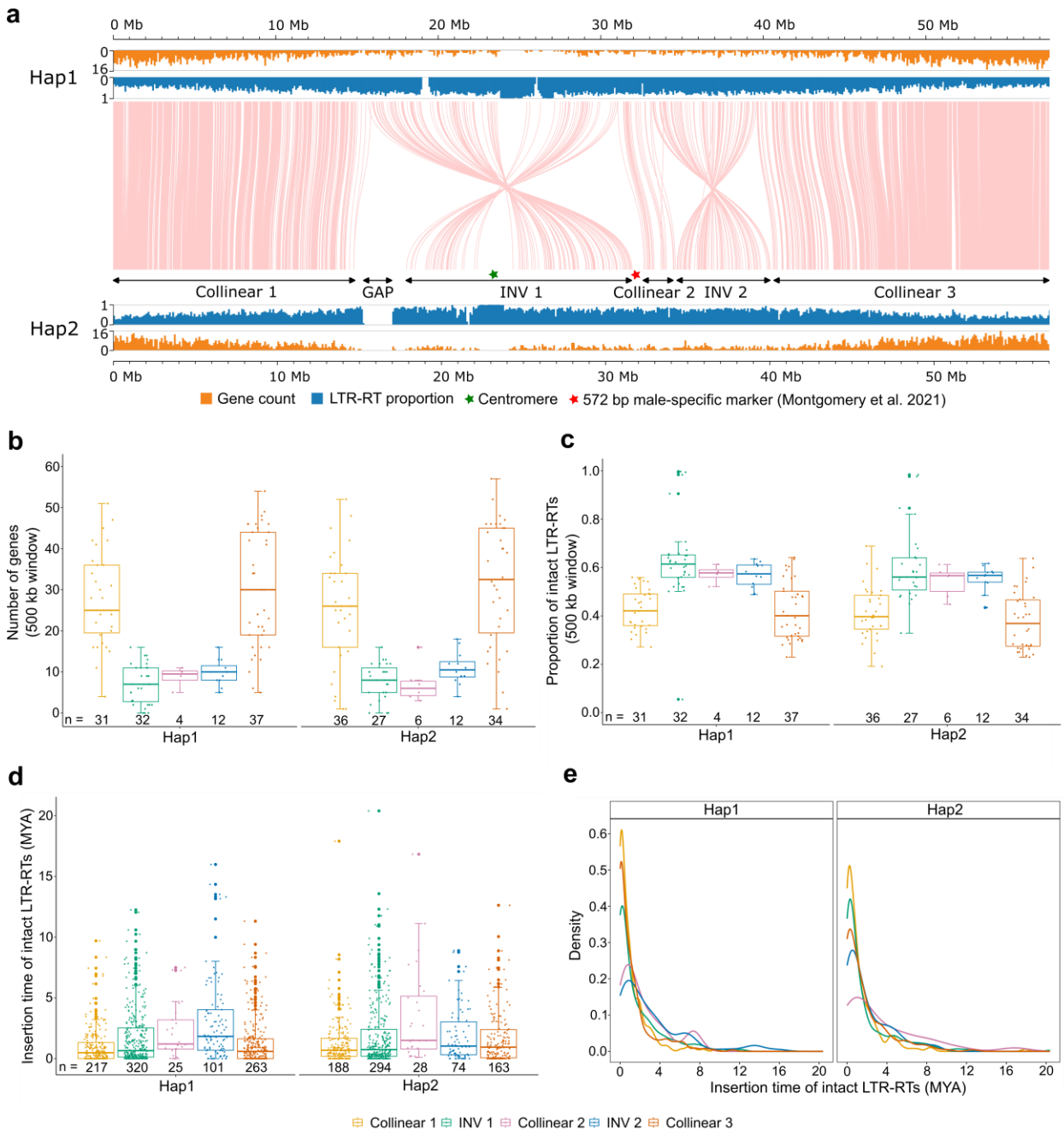

Figure 4.3 Comparative analysis between chromosome 1 of the two haplomes. **a** Synteny plot based on gene order, showing collinear regions and inversions (INV) on chromosome 1. The red asterisk represents a previously reported 572 bp male-specific marker(Montgomery et al. 2021) that matched to a region on Hap2. Gene count and LTR-RT proportion were calculated based on

Figure 4.3 (Cont.)

100 kb non-overlapping windows. **b** Number of genes across five regions (three collinear regions and two inversions) on the chromosome. Gene densities are calculated per 500 kb non-overlapping windows. **c** Proportion of intact LTR-RTs across the five regions on the chromosome. LTR-RT proportions are calculated per 500 kb non-overlapping windows. **d** Insertion time of intact LTR-RTs across the five regions. **e** Density distribution of intact LTR-RTs insertion times across the five regions on the chromosome.

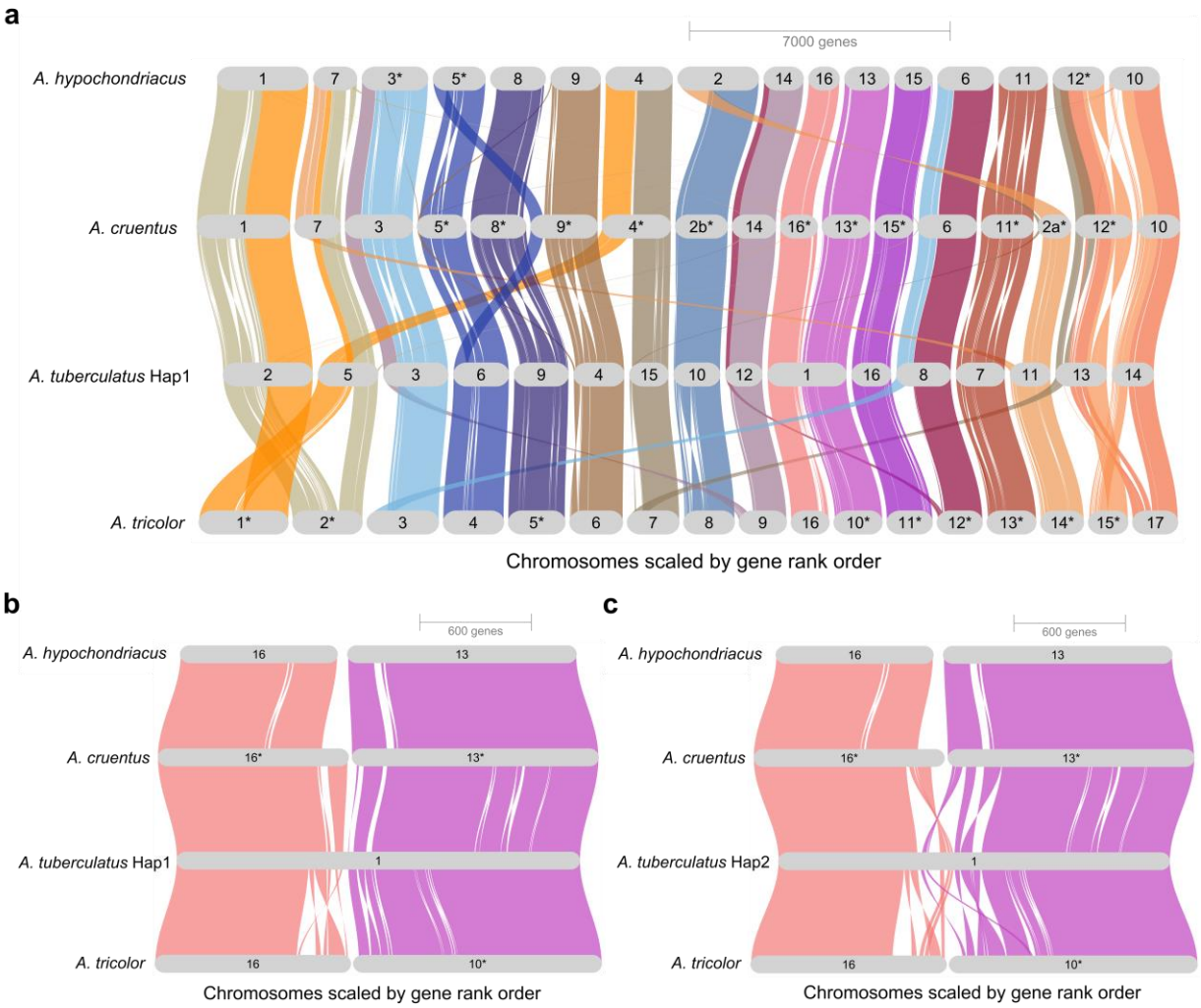

Figure 4.4 Synteny plot. **a** Synteny between the haplotype assemblies of *A. tuberculatus* and chromosome-level assemblies of three monoecious *Amaranthus* species**. b, c** Highlight of possible fusion of two separate chromosomes in *A. tuberculatus*. Asterisks indicate chromosomes that were manually inverted to keep the gene order consistent with *A. tuberculatus*.

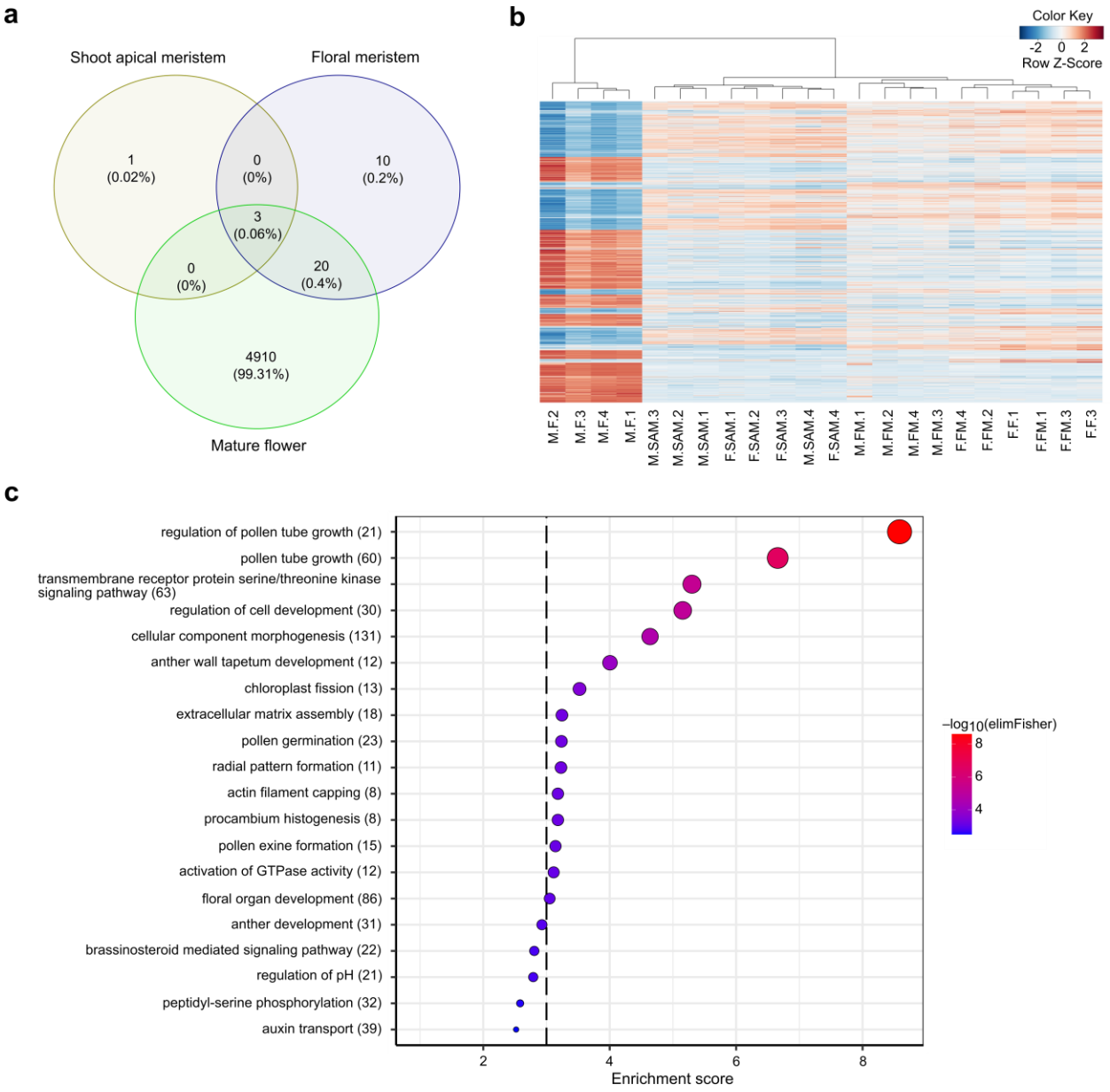

Figure 4.5 Differential gene expression analysis between male and female individuals across three tissue types. **a** Numbers of differentially expressed genes for male versus female comparison for shoot apical meristem, floral meristem, and mature flower. **b** Heatmap of log-CPM values for all 4,933 differentially expressed genes in male versus female comparison for mature flower. Samples are ordered using the hierarchical clustering method. Red depicts genes with relatively high expression, white depicts intermediate expression levels and blue represents

Figure 4.5 (Cont.)

genes with relatively low expression levels. FM: floral meristem, SAM: shoot apical meristem, M: mature flower. M or F as prefix indicate male or female, **c** Gene ontology enrichment analysis showing significantly overrepresented terms for male versus female differentially expressed genes. Values in parentheses represent the number of genes within the term. Dashed black line indicates significance level at p=0.001.

4.6 REFERENCES

Akagi T, Charlesworth D (2019) Pleiotropic effects of sex-determining genes in the evolution of dioecy in two plant species. Proc R Soc B Biol Sci 286: 20191805

Akagi T, Henry IM, Tao R, Comai L (2014) A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. Science (80) 346:646

Akagi T, Pilkington SM, Varkonyi-Gasic E, Henry IM, Sugano SS, Sonoda M, Firl A, McNeilage MA, Douglas MJ, Wang T, Rebstock R, Voogd C, Datson P, Allan AC, Beppu K, Kataoka I, Tao R (2019) Two Y-chromosome-encoded genes determine sex in kiwifruit. Nat Plants 5:801–809

Akagi T, Varkonyi-Gasic E, Shirasawa K, Catanach A, Henry IM, Mertten D, Datson P, Masuda K, Fujita N, Kuwada E, Ushijima K, Beppu K, Allan AC, Charlesworth D, Kataoka I (2023) Recurrent neo-sex chromosome evolution in kiwifruit. Nat Plants 9:393–402

Álvarez-Carretero S, Kapli P, Yang Z (2023) Beginner's guide on the use of PAML to detect positive selection. Mol Biol Evol 40:1–18

Bawa KS (1980) Evolution of dioecy in flowering plants. Annu Rev Ecol Syst 11:15–39

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc 57:289–300

Bobadilla LK, Baek Y, Tranel PJ (2023) Comparative transcriptomic analysis of male and females in the dioecious weeds *Amaranthus palmeri* and *Amaranthus tuberculatus*. BMC Plant Biol 23:1–26

Brazier T, Glémin S (2022) Diversity and determinants of recombination landscapes in flowering plants. PLoS Genet 18:1–29

Cabanettes F, Klopp C (2018) D-GENIES: Dot plot large genomes in an interactive, efficient and

simple way. PeerJ 2018

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10:1–9

Cantalapiedra CP, Herņandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829

Cerca J, Maurstad MF, Rochette NC, Rivera-Colón AG, Rayamajhi N, Catchen JM, Struck TH (2021) Removing the bad apples: A simple bioinformatic method to improve loci-recovery in *de novo* RADseq data for non-model organisms. Methods Ecol Evol 12:805–817

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4:1–16

Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. Am Nat 112:975–997

Charlesworth D (2016) Plant sex chromosomes. Annu Rev Plant Biol 67:397–420

Charlesworth D (2019) Young sex chromosomes in plants and animals. New Phytol 224:1095–1107

Costea M, Tardif FJ (2003) Conspectus and notes on the genus *Amaranthus* in Canada. Rhodora 105:260–281

Cronk Q, Müller NA (2020) Default sex and single gene sex determination in dioecious plants. Front Plant Sci 11:1–5

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Dong SS, He WM, Ji JJ, Zhang C, Guo Y, Yang TL (2021) LDBlockShow: A fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief Bioinform 22:1–6

Du Q, Wu Z, Liu P, Qing J, He F, Du L, Sun Z, Zhu L, Zheng H, Sun Z, Yang L, Wang L, Du H (2023) The chromosome-level genome of *Eucommia ulmoides* provides insights into sex differentiation and α-linolenic acid biosynthesis. Front Plant Sci 14:1–12

Goel M, Sun H, Jiao WB, Schneeberger K (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 20:1–13

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Grant WF (1959) Cytogenetic studies in *Amaranthus*. Can J Bot 37:413–417

Hallast P, Ebert P, Loftus M, Yilmaz F, Audano PA, Logsdon GA, Bonder MJ, Zhou W, Höps W, Kim K, Li C, Hoyt SJ, Dishuck PC, Porubsky D, Tsetsos F, Kwon JY, Zhu Q, Munson KM, Hasenfeld P, Harvey WT, Lewis AP, Kordosky J, Hoekzema K, O'Neill RJ, Korbel JO, Tyler-Smith C, Eichler EE, Shi X, Beck CR, Marschall T, Konkel MK, Lee C (2023) Assembly of 43 human Y chromosomes reveals extensive complexity and variation. Nature 621:355–364

Harkess A, Huang K, van der Hulst R, Tissen B, Caplan JL, Koppula A, Batish M, Meyers BC, Leebens-Mack J (2020) Sex determination by two Y-linked genes in garden asparagus. Plant Cell 32:1790–1796

Healey AL, Piatkowski B, Lovell JT, Sreedasyam A, Carey SB, Mamidi S, Shu S, Plott C,

Jenkins J, Lawrence T, Aguero B, Carrell AA, Nieto-Lugilde M, Talag J, Duffy A, Jawdy S, Carter KR, Boston LB, Jones T, Jaramillo-Chico J, Harkess A, Barry K, Keymanesh K, Bauer D, Grimwood J, Gunter L, Schmutz J, Weston DJ, Shaw AJ (2023) Newly identified sex chromosomes in the *Sphagnum* (peat moss) genome alter carbon sequestration and ecosystem dynamics. Nat Plants 9:238–254

Hearn KE, Koch EL, Stankowski S, Butlin RK, Faria R, Johannesson K, Westram AM (2022) Differing associations between sex determination and sex-linked inversions in two ecotypes of *Littorina saxatilis*. Evol Lett 6:358–374

Henry IM, Akagi T, Tao R, Comai L (2018) One hundred ways to invent the sexes: Theoretical and observed paths to dioecy in plants. Annu Rev Plant Biol 69:553–575

Hobza R, Kubat Z, Cegan R, Jesionek W, Vyskot B, Kejnovsky E (2015) Impact of repetitive DNA on sex chromosome evolution in plants. Chromosom Res 23:561–570

Huang M, Liu X, Zhou Y, Summers RM, Zhang Z (2019) BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8:1–12

Huson DH, Scornavacca C (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067

Iamonico D (2020) Nomenclatural survey of the genus *Amaranthus* (Amaranthaceae). 11. dioecious *Amaranthus* species belonging to the sect. *Saueranthus*. Darwiniana 8:567–575

Jeffares DC, Bartlomiej T, Sojo V, dos Reis M (2014) A Beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *In* C Peacock, ed. Parasite Genomics Protocols. Methods in Molecular Biology. New York, NY: Humana Press. 1–365 p

Kafkas S, Ma X, Zhang X, Topçu H, Navajas-Pérez R, Wai CM, Tang H, Xu X, Khodaeiaminjan M, Güney M, Paizila A, Karcı H, Zhang X, Lin J, Lin H, Herrán R de la, Rejón CR, García-Zea JA, Robles F, Muñoz C del V, Hotz-Wagenblatt A, Min XJ, Özkan H, Motalebipour EZ, Gozel H, Çoban N, Kafkas NE, Kilian A, Huang HX, Lv X, Liu K, Hu Q, Jacygrad E, Palmer W, Michelmore R, Ming R (2023) Pistachio genomes provide insights into nut tree domestication and ZW sex chromosome evolution. Plant Commun 4: 100497

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780

Kreiner JM, Giacomini DA, Bemm F, Waithaka B, Regalado J, Lanz C, Hildebrandt J, Sikkema PH, Tranel PJ, Weigel D, Stinchcombe JR, Wright SI (2019) Multiple modes of convergent adaptation in the spread of glyphosate-resistant *Amaranthus tuberculatus*. Proc Natl Acad Sci USA 116:23363

Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. Nucleic Acids Res 39:2010–2012

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v2 00:1–3

Li H (2021) New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572–4574

Liao Y, Smyth GK, Shi W (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930

Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. BMC Biol 15:1–

15

Liu C, Neve P, Glasgow L, Wuerffel RJ, Owen MDK, Kaundun SS (2020) Modeling the sustainability and economics of stacked herbicide-tolerant traits and early weed management strategy for waterhemp (*Amaranthus tuberculatus*) control. Weed Sci 68:179–185

Lloyd DG (1980) The distributions of gender in four angiosperm species illustrating two evolutionary pathways to dioecy. Evolution 34:123–134

Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J (2022) GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Elife 11:1–20

Ma X, Vaistij FE, Li Y, Jansen van Rensburg WS, Harvey S, Bairu MW, Venter SL, Mavengahama S, Ning Z, Graham IA, Van Deynze A, Van de Peer Y, Denby KJ (2021) A chromosome-level *Amaranthus cruentus* genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop. Plant J 107:613–628

Ma X, Yu L, Fatima M, Wadlington WH, Hulse-Kemp AM, Zhang X, Zhang S, Xu X, Wang J, Huang H, Lin J, Deng B, Liao Z, Yang Z, Ma Y, Tang H, Van Deynze A, Ming R (2022) The spinach YY genome reveals sex chromosome evolution, domestication, and introgression history of the species. Genome Biol 23:1–30

Massonnet M, Cochetel N, Minio A, Vondras AM, Lin J, Muyle A, Garcia JF, Zhou Y, Delledonne M, Riaz S, Figueroa-Balderas R, Gaut BS, Cantu D (2020) The genetic basis of sex determination in grapes. Nat Commun 11:1–12

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J,

Rank D, Garcia JF, DeRisi JL, Smith T, Tobias C, Ross-Ibarra J, Korf I, Chan SWL (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:1–20

Ming R, Bendahmane A, Renner SS (2011) Sex chromosomes in land plants. Annu Rev Plant Biol 62:485–514

Montgomery JS, Giacomini DA, Weigel D, Tranel PJ (2021) Male-specific Y-chromosomal regions in waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). New Phytol 229:3522–3533

Montgomery JS, Morran S, MacGregor DR, Scott McElroy J, Neve P, Neto C, Vila-Aiub MM, Victoria Sandoval M, Menéndez AI, Fan L, Caicedo AL, Maughan PJ, Assis Barbosa Martins B, Mika J, Collavo A, Merotto Jr A, Subramanian NK, Bagavathiannan V, Cutti L, Mazharul Islam M, Gill BS, Cicchillo R, Gast R, Soni N, Wright TR, Zastrow-Hayes G, May G, Sehgal D, Shankhar Kaundun S, Dale RP, Juan B, Peters B, Lerchl J, Tranel PJ, Beffa R, Jugulam M, Fengler K, Llaca V, Patterson EL, Gaines T (2023) The International Weed Genomics Consortium: Community resources for weed genomics research. Preprint:1–53

Montgomery JS, Sadeque A, Giacomini DA, Brown PJ, Tranel PJ (2019) Sex-specific markers for waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). Weed Sci 67:412–418

Mosyakin SL, Robertson KR (1996) New infrageneric taxa and combinations in *Amaranthus* (Amaranthaceae). Ann Bot Fenn 33:275–281

Müller NA, Kersten B, Leite Montalvão AP, Mähler N, Bernhardsson C, Bräutigam K, Carracedo Lorenzo Z, Hoenicka H, Kumar V, Mader M, Pakull B, Robinson KM, Sabatti

M, Vettori C, Ingvarsson PK, Cronk Q, Street NR, Fladung M (2020) A single gene underlies the dynamic evolution of poplar sex determination. Nat Plants 6:630–637

Murray MJ (1940) The genetics of sex determination in the family Amaranthaceae. Genetics 25:409–431

Na JK, Wang J, Ming R (2014) Accumulation of interspersed and sex-specific repeats in the non-recombining region of papaya sex chromosomes. BMC Genomics 15:1–12

Natri HM, Merilä J, Shikano T (2019) The evolution of sex determination associated with a chromosomal inversion. Nat Commun 10:145

Neve P (2018) Gene drive systems: do they have a place in agricultural weed management? Pest Manag Sci 74:2671–2679

Neves CJ, Matzrafi M, Thiele M, Lorant A, Mesgaran MB, Stetter MG (2020) Male linked genomic region determines sex in dioecious *Amaranthus palmeri*. J Hered 111:606–612

Pratt DB, Clark LG (2001) *Amaranthus rudis* and *A. tuberculatus*, one species or two? J Torrey Bot Soc 128:282–296

R Core Team (2021) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing

Raiyemo DA, Bobadilla LK, Tranel PJ (2023) Genomic profiling of dioecious *Amaranthus* species provides novel insights into species relatedness and sex genes. BMC Biol 21:1–18

Renner SS (2014) The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database. Am J Bot 101:1588–1596

Renner SS, Ricklefs RE (1995) Dioecy and its correlates in the flowering plants. Am J Bot 82:596

Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: A Bioconductor package for differential

expression analysis of digital gene expression data. Bioinformatics 26:139–140

Rochette NC, Rivera-Colón AG, Catchen JM (2019) Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol 28:4737–4754

Sauer J (1955) Revision of the dioecious amaranths. Madroño 13:5–46

Sauer J (1957) Recent migration and evolution of the dioecious amaranths. Evolution 11:11–31

Sauer J (1972) The dioecious amaranths: A new species name and major range extensions. Madrono 21:426

Schleich AH, Licht MA, Owen MDK, Yadav R (2023) Managing herbicide-resistant waterhemp (*Amaranthus tuberculatus* [Moq.] J.D. Sauer) seedbanks by integrating several management tactics. Agrosystems, Geosci Environ 6:1–8

She H, Liu Z, Li S, Xu Z, Zhang H, Cheng F, Wu J, Wang X, Deng C, Charlesworth D, Gao W, Qian W (2023) Evolution of the spinach sex-linked region within a rarely recombining pericentromeric region. Plant Physiol 00:1–18

Soltani N, Shropshire C, Sikkema PH (2023) Integrated weed management strategies for the depletion of multiple herbicide-resistant waterhemp (*Amaranthus tuberculatus*) seed in the soil seedbank. Weed Technol 37:108–112

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Steckel LE (2007) The dioecious *Amaranthus* spp.: Here to stay. Weed Technol 21:567–570

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

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res 18:1944–1954

Tranel PJ, Trucco F (2009) 21st-century weed science: a call for *Amaranthus* genomics. Weedy Invasive Plant Genomics:53–81

Wang H, Xu D, Wang S, Wang A, Lei L, Jiang F, Yang B, Yuan L, Chen R, Zhang Y, Fan W (2023) Chromosome-scale *Amaranthus tricolor* genome provides insights into the evolution of the genus *Amaranthus* and the mechanism of betalain biosynthesis. DNA Res 30:1–15

Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, Aryal R, VanBuren R, Murray JE, Zhang W, Navajas-Pérez R, Feltus FA, Lemke C, Tong EJ, Chen C, Wai CM, Singh R, Wang ML, Min XJ, Alam M, Charlesworth D, Moore PH, Jiang J, Paterson AH, Ming R (2012a) Sequencing papaya X and Y^h chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci USA 109:13710–13715

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH (2012b) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:1–14

Waselkov KE, Boleda AS, Olsen KM (2018) A phylogeny of the genus *Amaranthus* (Amaranthaceae) based on several low-copy nuclear loci and chloroplast regions. Syst Bot 43:439–458

Waselkov KE, Olsen KM (2014) Population genetics and origin of the native North American agricultural weed waterhemp (*Amaranthus tuberculatus*; Amaranthaceae). Am J Bot 101:1726–1736

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Wu C, Davis AS, Tranel PJ (2018) Limited fitness costs of herbicide-resistance traits in *Amaranthus tuberculatus* facilitate resistance evolution. Pest Manag Sci 74:293–301

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568–579

Yue J, Krasovec M, Kazama Y, Zhang X, Xie W, Zhang S, Xu X, Kan B, Ming R, Filatov DA (2023) The origin and evolution of sex chromosomes, revealed by sequencing of the *Silene latifolia* female genome. Curr Biol 33:2504-2514.e3

CHAPTER 5: THE GENOMES OF *AMARANTHUS PALMERI* (PALMER AMARANTH), *AMARANTHUS RETROFLEXUS* (REDROOT PIGWEED), AND *AMARANTHUS HYBRIDUS* (SMOOTH PIGWEED) SHED LIGHT INTO SEX CHROMOSOME EVOLUTION AND STRUCTURAL REARRANGEMENTS

ABSTRACT

Amaranthus palmeri (Palmer amaranth), *A. retroflexus* (redroot pigweed), and *A. hybridus* (smooth pigweed) are troublesome weeds that are economically and agriculturally damaging to several cropping systems. Collectively referred to as 'pigweeds', many biotypes have evolved resistance to various herbicide modes of action and continue to expand beyond their native range. To deepen our understanding of the biology of the weed species, including dioecy evolution, and provide genomic resources for rapid elucidation of mechanistic processes involved in herbicide resistance, we sequenced and assembled chromosome-level genomes of the three weed species. Combining the haplotype-resolved assembly of *A. palmeri* with existing restriction site-associated DNA (RAD-seq) and transcriptomic data for genome-wide association (GWA) analysis and differential gene expression, we identified a ~2.84 Mb region on chromosome 3 of Hap1 that is male-specific and contains 37 genes. Two genes within the male-specific region, encoding an Rf1 protein (restorer-of-fertility) and TLC domain-containing protein, were upregulated in male individuals across shoot apical meristem, floral meristem, and mature flower, and thus could be involved in sex determination in *A. palmeri*. This work advances our understanding of sex determination in a dioecious weed species that is of agronomic importance and provides genomic resources to further investigate adaptive trait evolution within the *Amaranthus* genus.

5.1 INTRODUCTION

Amaranthus palmeri S. Watson (Palmer amaranth), *A. retroflexus* L. (redroot pigweed), and *A. hybridus* L. (smooth pigweed) are troublesome weeds of several cropping systems (Sellers et al. 2003). They possess characteristics that make their control in agronomic crops challenging, including prolific seed production (e.g., *A. palmeri* can produce up to 1,000,000 seeds per plant), fast growth rate, long seed viability, and extended period of germination (Bensch et al. 2003; Steckel et al. 2004; Ward et al. 2013). The similarities in plant morphology among the weedy amaranths (collectively referred to as 'pigweeds') also complicate proper identification of individual species, which is necessary for timely and effective management decisions (Thapa and Blair 2018; Aderibigbe et al. 2022). Evolution and spread of herbicide resistance in pigweeds is now a major concern for crop production, and numerous studies have documented significant yield losses that could occur from interference of pigweeds with crops such as corn, soybean, or cotton (Massinga et al. 2001; Morgan et al. 2001; Hager et al. 2002; Bensch et al. 2003; Manalil et al. 2017).

While *A. retroflexus* and *A. hybridus* are monoecious and primarily self-pollinated, *A. palmeri* is a dioecious (i.e., with separate male and female plants), obligate outcrossing species that has now spread beyond its native origins of northwestern Mexico, southern California, New Mexico, and Texas to several parts of the US (Steckel 2007; Trucco et al. 2007). Sauer (1957) referred to *A. palmeri* as the most successful weedy invader of artificial habitats among the dioecious amaranths. It is now one of the most troublesome agronomic weeds in southeastern US, and reported in 45 countries where it is problematic in several agricultural systems (Ward et al. 2013; Roberts and Florentine 2022). Climate models indicate that *A. palmeri* will further expand beyond its present distribution, and by 2050 it is likely to have expanded to most of

Africa, Australia, Eastern Asia, Europe and the USA (Kistner and Hatfield 2018). The notoriety of the pigweeds as damaging to agricultural crops has spurred the exploration of several options for their control (e.g., cover crops, herbicide rotations, mechanical seed destruction, flood irrigation management, tillage, grazing, and other prevention strategies). A strategy that has been proposed for the management of dioecious weeds (e.g., *A. palmeri*) is a genetic control strategy whereby certain genetic factors that could bias sex ratio towards a particular gender are caused to be inherited across multiple generations in a non-mendelian fashion (e.g., via a meiotic drive), thereby leading to a population collapse at some point when the population lacks the other sex required for outcrossing (Neve 2018; Barrett et al. 2019; Rode et al. 2019; Legros et al. 2021). To utilize such a weed control strategy, however, requires understanding the mechanism of sex determination in the dioecious weed species.

Previous studies on sex determination in amaranths identified 34 diploid chromosomes in *A. palmeri*, although differences in chromosome pairs could not be ascertained (Murray 1940; Grant 1959). It was also proposed that males were the heterogametic sex with an XY system, which was subsequently confirmed using reduced representation sequencing (RAD-seq) data (Murray 1940; Montgomery et al. 2019; Neves et al. 2020). As part of an effort to make genomic resources available for *A. palmeri*, a draft genome for a male individual was assembled into 303 scaffolds, and a *k*-mer analysis was used to identify a ~1.3 Mb male-specific Y region (MSY) on scaffold 20 (Montgomery et al. 2020, 2021). This MSY region contained 121 gene models (Montgomery et al. 2021). The draft assembly was also combined with short reads of *A. watsonii* to detect the male-specificity of a copy of pentatricopeptide repeat-containing protein (PPR) within the MSY region on scaffold 20 (Raiyemo et al. 2023). Similarly, transcriptomic analyses between male and female individuals for shoot apical meristem, floral meristem, and mature

flower revealed the upregulation of *PPR247* in males, which mapped to a *PPR* gene copy within the MSY region on scaffold 20 (Bobadilla et al. 2023). The expression of this gene in males across the three tissue types led Bobadilla et al. to infer a single-gene model of sex determination for *A. palmeri*.

To gain insight into sex determination, and characterize sex chromosomes in *A. palmeri*, we assembled a chromosome-level genome of a male individual utilizing PacBio long-read sequencing, Hi-C scaffolding, and Bionano optical mapping. Phasing the assembly into two haplotypes and comparing the haplotype assemblies enabled us to identify contiguous male-specific region, and the genes present within the region that are likely candidates for sex determination. Sequencing and assembly of chromosome-level genomes of two additional amaranths (*A. retroflexus* and *A. hyridus*) further enabled the characterization of structural rearrangements between the species and other *Amaranthus* genome assemblies in the literature.

5.2 MATERIALS AND METHODS

5.2.1 Plant material and growth conditions

The plant materials sequenced in this study are publicly available with USDA Germplasm Resources Information Network (GRIN). *Amaranthus palmeri* and *A. retroflexus* seeds are available under the accession numbers PI 632235 and PI 572263, while the seeds of *A. hybridus* are from a population (FT-21605-14) that was sequenced and assembled in Montgomery et al. (2020). Seeds from the three species were sown separately in pots filled with premoistened soil (Lambert LM-GPS), and bottom irrigated. Seedlings were transplanted at ~5 cm height into 16-cm pots (America Clay Works I-A650MP) filled with the same soil. Plants were grown at a temperature of 25/20 C and a photoperiod of 16 h/8 h (light/dark) regimes. Tissue collection, DNA and RNA extractions, and library preparation have been previously

described (Chapter 4). All tissue samples were shipped either on dry ice or at 4 C to the Genome Center of Excellence at Corteva Agriscience for DNA extraction, library preparation, and sequencing.

5.2.2 Genome sequencing, assembly, and annotation

The sequencing, assembly, and annotation of *A. palmeri*, *A. retroflexus*, and *A. hybridus* genomes follow the previously described protocols used for *A. tuberculatus*. The methods were fully described in Chapter 4.

5.2.3 Sex phenotyping, SNP genotyping, and analysis of RAD-seq data

A RAD-seq dataset that was previously used to develop sex specific markers in Montgomery et al. (2019) was used for genome-wide association analysis. The sex of 384 individuals (192 females and 192 males) were phenotyped in the study. The single-end raw reads data were demultiplexed, and cleaned using the "process_radtags" command in Stacks version 2.65 (Rochette et al. 2019), followed by mapping each sample to the Hap1 assembly of *A. palmeri* using BWA-MEM v0.7.17 (Li 2013). The "gstacks" command from Stacks was then used to build loci from the aligned reads. To determine the level of missingness in the data, we grouped the individuals by sex, and then ran the “populations” command utilizing a minimal filtering criterion (--min-maf 0.05). We then accessed the level of missingness in the two data groups using VCFtools v0.1.16 (--missing-indv) following a previous approach (Cerca et al. 2021). We removed individuals that had greater than 60% missing data, and then ran the “populations” command once again on the streamlined samples utilizing additional filtering criteria (--min-maf 0.05, -p 2, -r 0.4). Using the set of variants obtained, we performed a principal component analysis with PLINK v1.90b7 (Chang et al. 2015), and then carried out a genome-wide association analysis with BLINK-C (Huang et al. 2019) using the first five

principal components from PLINK to account for population stratification and familial relatedness. Sex of the individuals were used as binary input (phenotypes) in the analysis. Calculation of f-statistics (F_{ST}) was carried out using VCFtools. Manhattan and F_{ST} plots were generated using CMplot package (Yin et al. 2021) in R.

5.2.4 Identification of previous sex marker and genes in the *A. palmeri* assembly

A search of the previously reported primer sets, PAMS-940 (Montgomery et al. 2019) and JM940 (Montgomery et al. 2021), that amplified a male-specific region of *A. palmeri* was carried out against both haplotype assemblies with BLASTN (Camacho et al. 2009) adopting the following parameters: -task blastn-short -db haplotypes -query primers -outfmt 6 -out output. Reciprocal best hit (RBH) searches between genes on the previously identified scaffold 20 and the haplotype assemblies, as well as between the two haplotype assemblies, were carried out using MMseqs 2 (Steinegger and Söding 2017).

5.2.5 Synteny and intragenomic analyses

MCscan-python version (Tang et al. 2008) from JCVI utility libraries v1.1.22 was used to identify reciprocal best hit (RBH), and collinear gene blocks between haplotype 1 and 2 of *A. palmeri* genome assemblies. MCScanX (Wang et al. 2012) was also used to investigate duplicated genes within each of the *A. palmeri* haplotypes and in *A. retroflexus*, and *A. hybridus* genome assemblies. To plot the duplicated genes with circos (Krzywinski et al. 2009), the output collinearity files from MCScanX were converted to ".anchors" format using "jcvi.compara.synteny" from JCVI. Sequence alignment between the two haplotypes of *A. palmeri* assemblies was carried out using Minimap2 v2.24-r1122 (Li 2021), and structural rearrangements between the two haplotypes were further refined using SyRI (Goel et al. 2019). Alignment between Hap1 and a previous draft assembly of a male *A. palmeri*, which contained

male-specific regions on scaffold 20, was carried out using Minimap2. The alignments were visualized as dotplot using D-Genies (Cabanettes and Klopp 2018). Syntenic orthologues among available chromosome-level *Amaranthus* species; three from this study and four (*A. hypochondriacus*, *A. cruentus*, *A. tricolor*, and *A. tuberculatus*) from previous studies were evaluated and visualized using GENESPACE v1.2.3 (Lovell et al. 2022).

5.2.6 Transcriptome profiling and gene ontology (GO) enrichment analysis

Quality control (QC) accessed and adapter trimmed mRNA reads for three tissue types (mature flowers, shoot apical meristem and floral meristems) from Bobadilla et al. (2023) were mapped to Hap1 of the *A. palmeri* genome assembly using STAR v2.7.10b (Dobin et al. 2013). Two replicates from mature flower category were removed from downstream analyses as it was done in the previous study due to low mapping quality i.e., 61.89% and 31.45% uniquely mapped reads for replicate 2 and 4, respectively. Gene counting was then carried out using featureCounts v2.0.6 from the subread package (Liao et al. 2014). Counts with zero were removed i.e., with CPM values less than 1, and the counts were normalized using the TMM normalization in edgeR (Robinson et al. 2009). The count normalization was then followed by the analyses of differential expression (DE) between male and female individuals for each tissue type in edgeR. Genes were assessed as differentially expressed based on FDR < 0.05 and $Log_2FC > 1.2$ thresholds using the '*glmTreat*' function within edgeR. Heatmaps were constructed with pheatmap package after a log2-transformed normalization of read counts in DESeq2 (Love et al. 2014). The counts were filtered to retain only genes that were differentially expressed from the previous analysis prior to the heatmap construction.

The translated protein-coding sequences (CDS) of Hap1 were first assigned GO annotations using eggNOG-mapper v2.1.12 (Cantalapiedra et al. 2021). GO term enrichment

analysis was then carried out using topGO with nodeSize = 10. The enrichment test was performed using Fisher's exact test and the "elim" algorithm. GO terms were assessed as significantly enriched at the default $p < 0.01$ threshold. The enrichment plot was generated using ggplot2 (Wickham et al. 2019) in R.

5.3 RESULTS

5.3.1 Genome assembly, annotation and repeat analyses

Evaluation of assembly completeness for *A. palmeri* haplome 1 (Hap1), *A. palmeri* haplome 2 (Hap2), *A. retroflexus* and *A. hybridus* genomes using BUSCO "embryophyta_odb10" database revealed greater than 97% complete BUSCOs for all the species (Table 5.1). Additional evaluation of assembly completeness using LTR assembly index (LAI) revealed average LAI scores for *A. palmeri*, Hap1, *A. palmeri* Hap2, *A. retroflexus*, and *A. hybridus* were 18.25, 21.85, 10.78 and 15.29, respectively (Table 5.1). Analysis of repetitive elements using RepeatMasker revealed that 54.27% of *A. palmeri* Hap1, 55.46% of *A. palmeri* Hap2, 57.84% of *A. retroflexus*, and 53.25% of *A. hybridus* were made up of transposable elements. The LTR/*Copia* retrotransposon was the most abundant, representing 9.34%, 10.33%, 9.78%, and 9.75% in *A. palmeri* Hap1, *A. palmeri* Hap2, *A. retroflexus*, and *A. hybridus* respectively (Appendix D Table D.1). Analysis of centromeric repeats using CentroMiner indicates chromosomes had varying types of centromeres e.g., chromosomes 1 and 5 appear telomeric while chromosomes 2 and 3 appear metacentric or submetacentric for *A. palmeri* (Appendix D Table D.2).

A search of the simple telomeric repeat, TTTAGGG, against the assemblies using BLASTN revealed telomeric repeat sequences at 27 out of 34 telomeric ends for *A. palmeri* Hap1, 26 out of 34 telomeric ends for *A. palmeri* Hap2, 26 out of 34 telomeric ends for *A. retroflexus*, and 8 out of 32 telomeric ends for *A. hybridus* (Appendix D Table D.3). It is well

known that highly repetitive regions including centromeres or AT-rich regions are known to pose limitations for assembly pipelines, which could further be exacerbated by the degree of heterozygosity of the species (Miga 2020; Sun et al. 2022). The total number of predicted protein-coding genes for the assemblies ranged from 22,771 genes for *A. hybridus* to 27,377 genes for *A. retroflexus* (Table 5.1, Appendix D Table D.4, and Appendix D Table D.5).

Transcription factors among the protein-coding genes for *A. palmeri* Hap1 (1,351), *A. palmeri* Hap2 (1,369), *A. retroflexus* (1,427), and *A. hybridus* (1,271) were annotated. The number of transcription factor families identified for both haplotypes of *A. palmeri*, *A. retroflexus*, and *A. hybridus* were 58, 57, and 56 TF families, respectively (Appendix D Table D.6 – D.9). In addition, the number of disease-resistance genes annotated for *A. palmeri* Hap1, *A. palmeri* Hap2, *A. retroflexus*, and *A. hybridus* were 1,356 (22 classes), 1,300 (24 classes), 1,358 (25 classes), and 1,362 (25 classes), respectively (Appendix D Table D.10 – D.13). The genomic features presented above were further visualized using a circos plot, which indicates regions abundant in gene content are poor in LTR retrotransposons (LTR-RTs) while regions that are poor in gene content are abundant in LTR-RTs (Figure 5.1).

5.3.2 Identification of the sex-determining region (SDR) in *A. palmeri*

Assessment of the level of missingness in the RAD-seq data revealed that both female and male samples had 40 – 100% missing data. Individuals with more than 60% missing data were then removed using the "populations" command in Stacks, which retained 52 female samples and 53 male samples. The filtering also improved the quality of the data (i.e., 18 – 31% missing data for females, 15 – 35% missing data for males and, 24.57% missing data across the 105 individuals). Overall, 234,066 variants were retained across the 105 individuals for genome-wide association (GWA) analysis. The analysis revealed four significant SNPs above the

Bonferroni threshold that are associated with sex; two on chromosome 3 (24,609,596 bp, *P* = 1.7481e-20; 24,595,358 bp, *P* = 6.8453e-17), and one each on chromosome 4 (13,423,105 bp, *P* = 1.0254e-08), and chromosome 16 (7,114,501 bp, *P* = 1.0704e-17). The most significant SNPs were located on chromosome 3 (Figure 5.2a). All significant SNPs identified resided within intergenic regions. The QQ plot did not reveal any evidence of systematic bias in the GWA analysis (Figure 5.2b). Analysis of genetic differentiation (F_{ST}) between females and males also points to chromosome 3 as one of the highly differentiated chromosomes between female and male individuals (Figures 5.2c and 5.2d).

5.3.3 Synteny and intergenomic analyses between the two haplotype assemblies

Analysis of synteny patterns between the haplotype assemblies indicated a one-to-one relationship in gene content between the haplotypes (Appendix D Figure D.1a and D.1b). MCscan analysis of syntenic orthologues revealed 21,421 gene pairs represent the reciprocal best matches between the two haplotypes, in which 1,357 gene pairs are on chromosome 3. A region on chromosome 3 between 22,994,134 – 25,836,006 (~ 2.84 Mb) is present at the distal end of Hap1 but absent from Hap2, and thus was designated as a male-specific Y region (MSY) within the SDR (Figure 5.2e). In the absence of this ~2.84 Mb region, Hap1 and Hap2 were highly syntenic (Figure 5.2e). Further analysis revealed 1,857,183 bp (65.35%) of the MSY region is made up of gaps. There are 37 genes within the MSY region out of 1,620 protein-coding genes on chromosome 3 of Hap1 (Appendix D Table D.14). Investigation of the presence of inversions, typical of some SDRs in plant species, using SyRI revealed a total of 93 inversions spread across the 17 chromosomes between Hap1 and Hap2 (Appendix D Table D.15). Although inversions were present on chromosome 3, they were however smaller (< 260 kb) when compared to

inversions on chromosomes 4, 7, 14, or 17, which were greater than 1 Mb (Appendix D Table D.15).

Query of a primer set (PAMS-940), used in Montgomery et al. (2019) to amplify MSY regions, against both haplotype assemblies of *A. palmeri* revealed a perfect match to chromosome 3 of Hap1 with a size of 51 bp (24,464,780 – 24,464,831 bp) (Fig. 5.2e). This size is consistent with the amplified product size reported in Montgomery et al. (2019). The forward and reverse primer sequence for JM940 from Montgomery et al. (2021) also matched to chromosome 3, however, both sequences were not within the expected distance (i.e., ~100 bp) from each other, possibly due to variation in non-coding sequences or transposable elements across populations. Sequence alignment of the previous draft *A. palmeri* male assembly to both haplotype assemblies indicated the draft assembly was highly fragmented, and the scaffolds were not in order (Appendix D Figure D.2a and D.2b). Scaffolds 5 and 20 both aligned to chromosome 3 in the two haplotype assemblies; however, the region identified as male-specific on scaffold 20 only aligned to a region on chromosome 3 of Hap1 (Appendix D Figure D.2a,b,c,d,f). Reciprocal best hit search between genes in the draft assembly and the haplotype assemblies, and filtering for match to scaffold 20, revealed 41 and 37 genes had matches on chromosome 3 of Hap1 and Hap2 assemblies, respectively (Appendix D Table D.16). However, only four genes (three unknown proteins and protein HEADING DATE 3A) that matched between the draft assembly and Hap1 were within the MSY region in Hap1 (Appendix D Table D.16). Taken together, we consider chromosome 3 as the likely sex chromosome candidate.

Analysis of structural rearrangements between the three *Amaranthus* genome assemblies from this study, *A. tuberculatus*, and three monoecious amaranths indicated a highly conserved gene order, except for a few chromosomes (Figure 5.3). For instance, chromosome 4 in *A.*

hypochondriacus, *A. hybridus*, and *A. cruentus* appears to have been derived from the fusion of two ancestral chromosomes (Figure 5.3). It is worth noting that some genes on chromosome 1 of the three genome assemblies appear to have paralogs (i.e., duplications) on chromosome 1 (Figure 5.1). This pattern of duplication for chromosome 1 was earlier observed in *A. hypochondriacus* (Lightfoot et al. 2017) and *A. cruentus* (Ma et al. 2021), and thus seem to be conserved across species in the subgenera *Acnida* and *Amaranthus*. Chromosome 1 (or chromosome 2 in *A. tuberculatus*) in species within the two subgenera however appears to have originated from the fusion of parts of chromosomes 1 and 2 in *A. tricolor*, which belongs to the subgenus *Albersia* (Figure 5.3). Similarly, chromosome 2 in *A. palmeri* and *A. retroflexus* (or chromosome 3 in *A. hybridus*) have duplicated copies on chromosome 10 (or chromosome 2 in *A. hybridus*) (Figure 5.1).

5.3.4 Transcriptome profiling between male and female flowers, and enrichment analysis

A previous mRNA dataset that had passed quality control (i.e., adapter trimmed reads and low-quality bases removed) reported in Bobadilla et al. (2023) were mapped to the Hap1 assembly of *A. palmeri* using STAR. Uniquely mapped reads for males ranged from 81.79 – 90.88% while uniquely mapped reads for females ranged from 76.40 – 91.61% across shoot apical meristem, floral meristem, and mature flower. Out of the 24,873 annotated Hap1 protein-coding genes, 20,046 gene were retained for DE analysis after filtering and TMM normalization (Figure 5.4a and 5.4b). Eight genes were differentially expressed between male and female comparison for shoot apical meristem, while 29 genes were differentially expressed for floral meristem, and 2,595 genes for mature flower (Figure 5.4a and 5.4b). Among the 1,283 genes retained on chromosome 3 after filtering and normalization, 6 at the shoot apical meristem stage (5 upregulated and 1 downregulated), 11 at the floral meristem stage (9 upregulated and 2

downregulated), and 183 (131 upregulated and 52 downregulated) at the mature flower stage were differentially expressed (Appendix D Table D.17 – D.19). Five genes (encoding a serine/threonine-protein kinase EDR1-like, two Rf1 proteins, an unknown protein, and a TLC domain-containing protein) were differentially expressed between males and female individuals across the three tissue types. A search of the unknown protein to NCBI nonredundant protein database using BLASTP revealed 71% homology to a wall-associated receptor kinase-like protein in several species. The five genes were upregulated in males and are present on chromosome 3 (Appendix D Table D.17 – D.19). Out of the five genes, two genes encoding a protein Rf1 (a restorer of male fertility gene that matched to pentatricopeptide repeat-containing protein in NCBI non-redundant protein sequence database from several species) and a TLC domain-containing protein were upregulated in males across the three tissue types, and both genes are within the MSY region on chromosome 3 (Appendix D Table D.17 – D.19).

Gene ontology (GO) term enrichment analysis to gain insight into biological processes that could be involved in sex determination was performed. DEGs were selected based on an FDR threshold of $p < 0.05$ and $Log_2FC > 1.2$. Biological processes including pollen tube growth, regulation of pollen tube growth, pollen germination, regulation of cell development, pollen sperm cell differentiation, cellular component morphogenesis, anther wall tapetum development, and pollen wall assembly were among the top 20 enriched GO terms identified (Figure 5.4c, Appendix D Table D.20). An unknown protein, although upstream of the MSY region on chromosome 3, was upregulated in males and part of the GO term enriched for pollen tube growth, regulation of pollen tube growth, regulation of cell development, and abscisic acid-activated signaling pathway (Appendix D Table D.21). A search of the unknown protein sequence against NCBI non-redundant protein sequence database using BLASTP revealed 98%

homology to calcium-dependent protein kinase 17. The top three terms enriched for molecular function included pectinesterase activity, pectinesterase inhibitor activity, and phosphatidylinositol phosphate kinase activity (Appendix D Table D.22), while the top three terms enriched for cellular function include apical plasma membrane, pollen tube, and pollen tube tip (Appendix D Table D.23).

5.3.5 Comparison of this work to previous studies

Bobadilla et al. (2023) reported three genes, *PPR247*, *ACD6*, and *WEX*, as likely candidates involved in sex determination in *A. palmeri*. They proposed that the presence of *PPR247* within the MSY region of *A. palmeri* results in the post-transcriptional silencing of *ACD6* and *WEX*, thus enabling the formation or development of male reproductive organs. Our findings revealed that the *WEX* gene on chromosome 3 (AmaPaChr03Ag064000) has 100% homology (at protein level) to a copy on scaffold 259 (8601 – 9125) of the draft assembly, and not the copy on scaffold 20 as earlier suggested in the study. The copy on scaffold 20 (g4679) had no match in the Hap1 assembly. *ACD6* from the scaffold assembly (g4650) had 96.6% homology to AmaPaChr03Ag064850. Although the gene was annotated as ankyrin-3-like, a search of the protein sequence to NCBI non-redundant protein sequence database using BLASTP indicated the gene was incorrectly annotated, as it matched to *ACD6* in *Amaranthus tricolor*, *Spinacea oleracea*, and *Beta vulgaris*. *ACD6* in Hap1 is downstream of the MSY region, unlike in the scaffold assembly where it was upstream. Further, the *PPR* on scaffold 20 (g4709) had 90.8% homology to AmaPaChr03Ag063890 (annotated as protein Rf1). Additional search of the nucleotide sequence of g4709 *PPR* to Hap1 revealed 99.2% homology to AmaPaChr03Ag063890 while g4710 *PPR* matched to AmaPaChr03Ag064390 (annotated as *PPR* At3g22470) with 99.9% homology. AmaPaChr03Ag064390 was also upregulated in males

but only in mature flowers. To further understand the relationship among the *PPRs* on chromosome 3, we aligned the coding sequences (CDS) of all annotated *PPR* and *Rf1* genes on the chromosome using MUSCLE and constructed a Neighbor-Joining tree using the resulting alignment in MEGA11. While g4710, g4709, AmaPaChr03Ag063890, and AmaPaChr03Ag064390 clustered together as predicted from the BLAST analysis, the transcript copies (TRINITY_GG_4349_c0_g1 and TRINITY_GG_4346_c2_g1) were grouped into a separate cluster, even though one of them mapped to a copy on scaffold 20 in the previous study. Interestingly, the *Rf1* gene (AmaPaChr03Ag063970) that was upregulated in males across the three tissue types and present within the MSY region on Hap1 grouped together with two other *PPRs* (AmaPaChr03Ag064300 and AmaPaChr03Ag064440), and in a separate cluster.

Comparison of this work to the study from Bobadilla et al. (2023) thus points to the influence of assembly choice on sex chromosome inferences. A substantial source of variation in the results of several studies on sex chromosome divergence in *Poecillia* spp. was attributed to the differences in reference genomes, software and parameters used in the studies (Darolti et al. 2022). The draft genome of *A. palmeri* that was previously used to identify the MSY region, and utilized in the transcriptomic analysis of Bobadilla et al., was sequenced on PacBio Sequel I system and scaffolded with Dovetail Hi-C into 303 scaffolds with no haplotype phasing (Montgomery et al. 2020). While the MSY region was contiguous in the draft assembly, we note that the X and Y chromosomes are a chimeric mix that have been collapsed into a single chromosome representation, further complicating downstream sex chromosome inferences.

5.4 DISCUSSION

We present the first haplotype-resolved chromosome-level assembly of *A. palmeri*, *A. retroflexus*, and *A. hybridus*; three weed species that have become troublesome weeds in

numerous agricultural systems around the world. Our analysis identified an approximately 2.84 Mb region at the distal end of chromosome 3 on Hap1 that is male-specific, which is longer than the 1.3 Mb region that was previously identified as male-specific for *A. palmeri* (Montgomery et al. 2021). With 65.35% of the MSY region made up of gaps, it is likely that our assembly is still missing male-specific sequences within the region. Nevertheless, we identified key genes (encoding Rf1 and TLC domain-containing protein) that are likely candidates involved in sex determination. While the function of TLC domain-containing protein is unclear in plants, Rf1 protein is well-documented as restorer of fertility in some crops e.g., sorghum and wheat (Klein et al. 2005; Melonek et al. 2021). More so, over half of cloned restorer (*Rf*) genes encode pentatricopeptide repeat-containing (PPR) proteins (Uyttewaal et al. 2008; Chen and Liu 2014; Gaborieau et al. 2016).

Recently, Wu et al. (2023) showed that while staminate flowers in *A. palmeri* initiate both carpel and stamen primordia at an early stage, the carpel primordium remains undeveloped, and individuals become functionally male whereas pistillate flowers initiate and develop only carpel primordium. Given that both androecium and gynoecium are initiated in staminate flowers, but only functional androecium develops, it is plausible that the expression of a dominant activator of maleness (*M*) or the gain-of-function allele at stage 4/5 of floral organogenesis reported by Wu et al. could result in the initiation and establishment of stamen primordia while the expression of a dominant suppressor of female organs (*supF*) could result in the arrest of the initial gynoecium primordia at stage 7. It is therefore possible that the *Rf1* gene within the MSY region of *A. palmeri* is a male-promoting factor, and its presence and expression results in the formation of male reproductive organs. However, whether the *Rf1* gene also acts as the suppressor of female function is not clear. In theory, a CRISPR knockout of the *Rf1* gene

would result in male-to-female conversion if the *Rf1* gene acts as both the male activator and female suppressor while the knockout of the *Rf1* gene would result in male-to sterile conversion if a separate gene acts as the suppressor of female function. Development of a working transformation protocol for *A. palmeri* would facilitate the functional validation of the *Rf1* gene and other candidates within the MSY region.

The similarity between chromosome 3 of Hap1 and Hap2 assemblies in the absence of the MSY region in Hap1 suggests that hemizygosity (i.e., presence of the MSY region at the distal end of chromosome 3 in Hap1 but absence in Hap2) is responsible for the recombination suppression between the two haplotypes rather than the presence of large inversions that were previously observed within the sex-determining region on chromosome 1 of *A. tuberculatus* (see Chapter 4). Similar findings on reduced recombination due to hemizygosity has been reported in garden asparagus (Harkess et al. 2020). The lack of one-to-one orthology or synteny between genes on chromosome 3 of *A. palmeri* and chromosome 1 of *A. tuberculatus* further supports our previous conclusions on independent sex evolution in both species.

Comparison of repetitive elements from assemblies in this study to a previous study on repeat composition in unassembled genomes or raw reads of *Amaranthus* spp. reveals a consistent pattern of transposable element proliferation that was previously reported (Raiyemo et al. 2023), whereby monoecious species (including the dioecious *A. palmeri*) were abundant in LTR/*Copia* elements than LTR/*Ty3* elements while the dioecious species including *A. tuberculatus* had more *Ty3* elements than *Copia* elements. Specific reasons for such preferential accumulation of transposable elements between monoecious and dioecious species is not clear.

The similarity between the gene ontology terms enriched for *A. palmeri* and *A. tuberculatus* suggests that the downstream genes or pathway recruited in flowering are similar

even though previous evidence and findings from this work supports that both species have evolved dioecy independently, and perhaps utilize different mechanisms of sex determination. For instance, the most significantly overrepresented terms for male versus female differentially expressed genes in both species were pollen tube growth and regulation of pollen tube growth.

In sum, our study highlights candidate genes that could be involved in sex determination in a dioecious weedy species. While gene order was fairly conserved across the assemblies of *Amaranthus* species, chromosomal fusions were part of structural rearrangement events that contributed to the evolution of chromosomes in the genus. Also, the overall pattern of gene duplication among the amaranths appears quite similar, and thus likely preceded speciation. The genomic resources provided here will also be valuable for furthering the study of adaptive trait evolution within the genus.

5.5 TABLE AND FIGURES

Table 5.1 Comparison of assembly statistics between the two haplomes of *A. palmeri*, *A. retroflexus*, and *A. hybridus* genome assemblies.

Genome characteristics	***A. palmeri***		***A. retroflexus***	***A. hybridus***
	Haplome 1	**Haplome 2**		
Assembly				
Assembly size (Mbp)	384.1	364.5	446.79	437.99
Scaffold N50 (Mbp)	23.59	22.01	24.31	27.31
Scaffold L50	8	8	8	7
GC content (%)	33.4	33.4	33.0	33.0
Complete BUSCO (%)	97.7	97.3	97.3	97.3
Size of Ns (Mbp)	19.14	7.02	21.64	40.79
LTR assembly index (LAI)	18.25	21.85	10.78	15.29
Annotation				
Protein-coding genes	24,873	24,791	27,377	22,771
Mean gene length (bp)	4,637	4,633	4,623	4,874
Mean CDS length (bp)	1,178	1,174	1,092	1,184
Mean exon length (bp)	273	272	298	270
Mean exon per gene	5.5	5.5	5.2	5.5
Number of tRNA	1,480	1,508	1,499	1,535
Total annotated genes	26,353	26,299	28,876	24,306

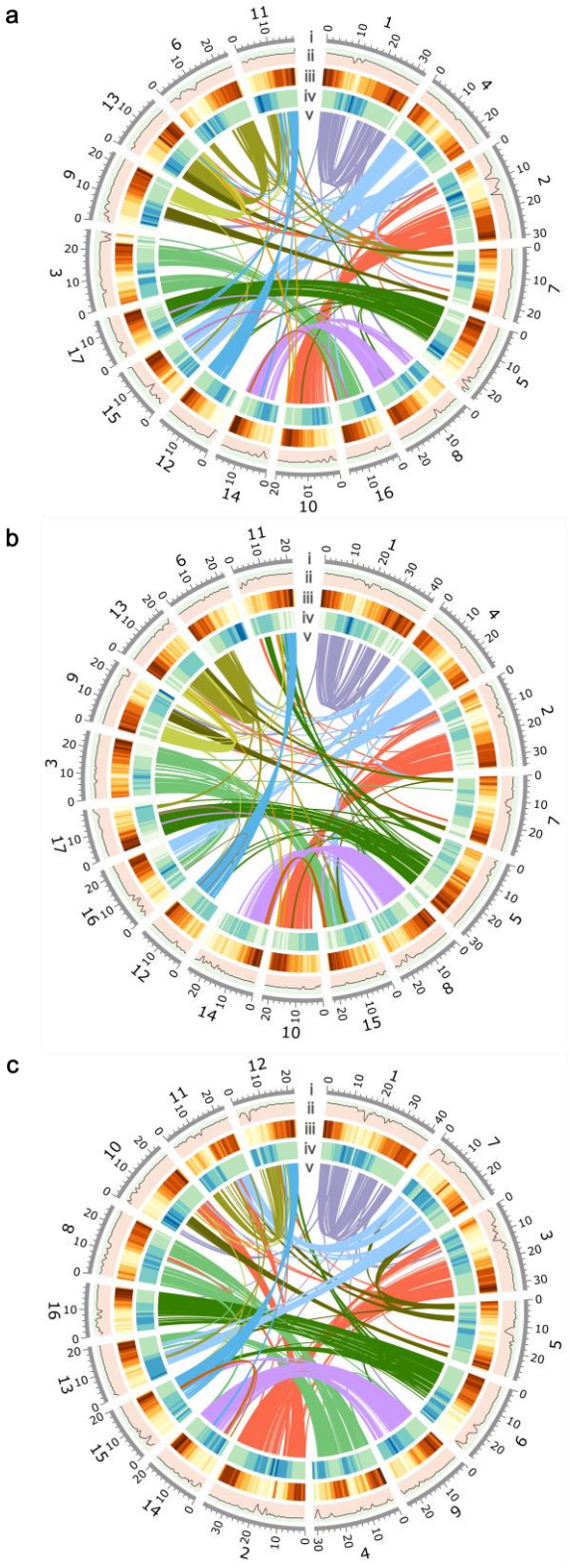

Figure 5.1 Genomic features of **a.** *A. palmeri* Hap1 **b.** *A. retroflexus* and **c.** *A. hybridus* assemblies. Circos plot depicts i) chromosome number and length (Mb), ii) GC content along the

Figure 5.1 (Cont.)

chromosomes, with peaks in light green area representing GC content greater than the median and peaks in light red area representing GC content less than the median (median GC contents: *A. palmeri* Hap1, 0.3245; *A. retroflexus*, 0.3235; and *A. hybridus*, 0.3188), iii) gene density across the chromosomes, with brown representing gene-rich regions and yellow representing gene-poor regions, iv) LTR (long terminal repeats) density along chromosomes, with blue representing LTR-rich regions and green representing LTR-poor regions, v) inner ribbons represent duplicated genes. Chromosomes are ordered based on synteny from GENESPACE in Figure 5.3. Window size of 1 Mb and step size of 500 kb for ii-iv.

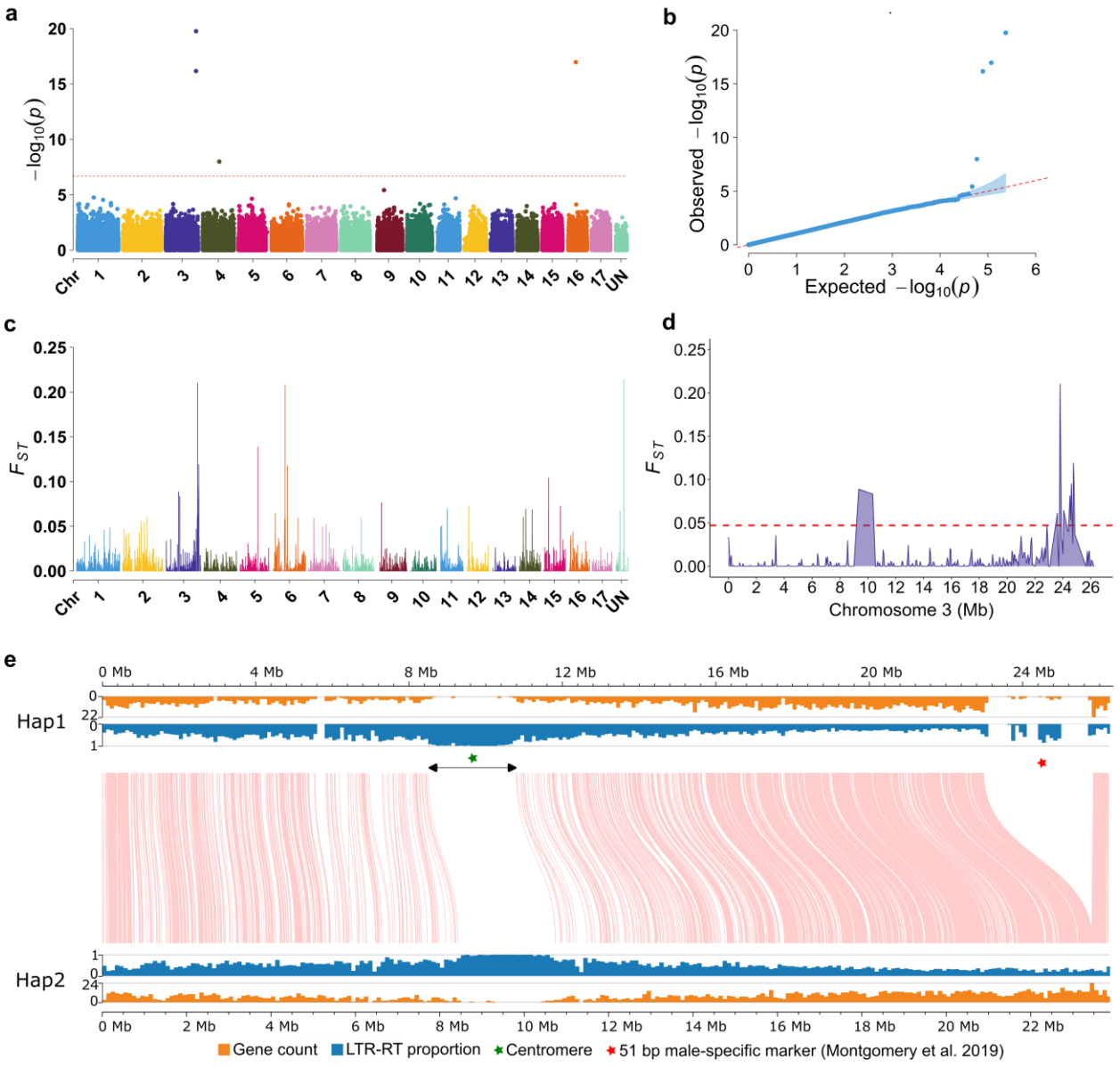

Figure 5.2 Identification of the sex-determining region on chromosome 3. **a** Manhattan plot of GWA analysis using RAD-seq data from 52 females and 53 males. The dashed red line indicates Bonferroni threshold of $-\log_{10}(P) = 6.6704$. **b** Quantile-quantile (QQ) plot of the GWA analysis. **c** Fixation index (F_{ST}) between females and males across all 17 chromosomes (window size 100 kb; step size 50 kb). **d** F_{ST} between females and males for chromosome 3 with dashed red line representing the top 1% threshold at 0.0472 (window size 100 kb; step size 50 kb). **e** Synteny plot showing collinear regions on chromosome 3. The green asterisk represents the centromere

Figure 5.2 (Cont.)

while the red asterisk represents a previously reported 51 bp male-specific marker (Montgomery et al. 2019) that matched to a region on hap 1. Gene count and LTR-RT proportion were calculated based on 100 kb non-overlapping windows.

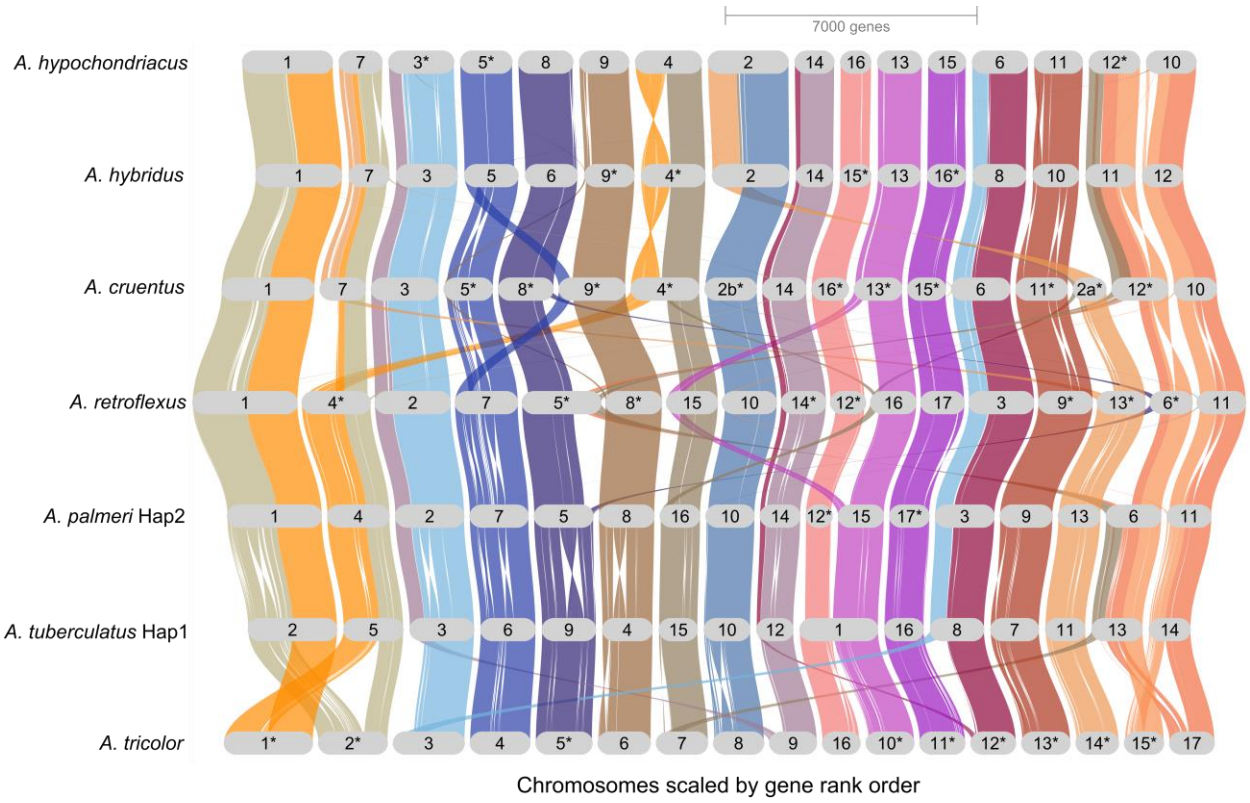

Figure 5.3 Synteny plot between the haplotype assemblies of four previously reported chromosome-level *Amaranthus* spp. assemblies and three newly assembled ones (*A. palmeri*, *A. retroflexus*, and *A. hybridus*). Asterisks indicate chromosomes that were manually inverted to keep the gene order consistent with *A. tuberculatus*. Species are ordered to reflect phylogenetic relationships from STAG in OrthoFinder, which also corresponds to the species tree in Wang et al. (2023).

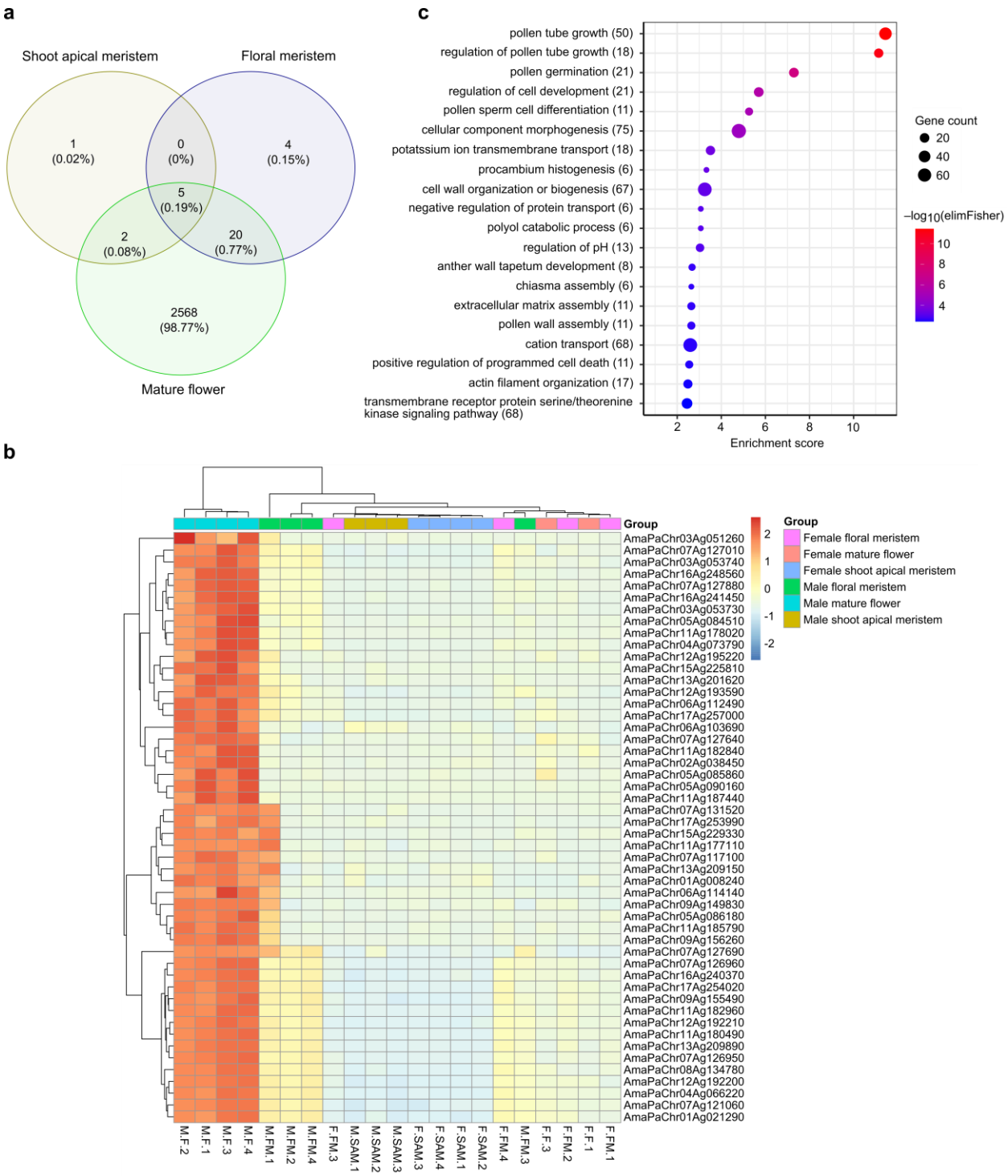

Figure 5.4 Differential gene expression analysis between male and female individuals across three tissue types. **a** Number of genes differentially expressed between male and female comparison for shoot apical meristem, floral meristem, and mature flower. **b** heatmap of top 50

Figure 5.4 (Cont.)

variable genes from the list of differentially expressed genes for male versus female comparison of mature flower. Samples are ordered using the hierarchical clustering method. **c** Gene ontology enrichment analysis showing significantly overrepresented terms for male versus female differentially expressed genes (values in parentheses represent the number of genes within the term).

5.6 REFERENCES

Aderibigbe OR, Ezekiel OO, Owolade SO, Korese JK, Sturm B, Hensel O (2022) Exploring the potentials of underutilized grain amaranth (*Amaranthus* spp.) along the value chain for food and nutrition security: A review. Crit Rev Food Sci Nutr 62:656–669

Barrett LG, Legros M, Kumaran N, Glassop D, Raghu S, Gardiner DM (2019) Gene drives in plants: Opportunities and challenges for weed control and engineered resilience. Proc R Soc B Biol Sci 286:1–9

Bensch CN, Horak MJ, Peterson D (2003) Interference of redroot pigweed (*Amaranthus retroflexus*), Palmer amaranth (*A. palmeri*), and common waterhemp (*A. rudis*) in soybean. Weed Sci 51:37–43

Bobadilla LK, Baek Y, Tranel PJ (2023) Comparative transcriptomic analysis of male and females in the dioecious weeds *Amaranthus palmeri* and *Amaranthus tuberculatus*. BMC Plant Biol 23:1–26

Cabanettes F, Klopp C (2018) D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ 2018

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10:1–9

Cantalapiedra CP, Herņandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829

Cerca J, Maurstad MF, Rochette NC, Rivera-Colón AG, Rayamajhi N, Catchen JM, Struck TH (2021) Removing the bad apples: A simple bioinformatic method to improve loci-recovery in *de novo* RADseq data for non-model organisms. Methods Ecol Evol 12:805–817

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4:1–16

Chen L, Liu YG (2014) Male sterility and fertility restoration in crops. Annu Rev Plant Biol 65:579–606

Darolti I, Almeida P, Wright AE, Mank JE (2022) Comparison of methodological approaches to the study of young sex chromosomes: A case study in *Poecilia*. J Evol Biol:1646–1658

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Gaborieau L, Brown GG, Mireau H (2016) The propensity of pentatricopeptide repeat genes to evolve into restorers of cytoplasmic male sterility. Front Plant Sci 7:1816

Goel M, Sun H, Jiao WB, Schneeberger K (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 20:1–13

Grant WF (1959) Cytogenetic studies in *Amaranthus*. Can J Bot 37:413–417

Hager AG, Wax LM, Stoller EW, Bollero GA, Hager AG, Bollero GA (2002) Common waterhemp (*Amaranthus rudis*) interference in soybean. Weed Sci 50:607–610

Harkess A, Huang K, van der Hulst R, Tissen B, Caplan JL, Koppula A, Batish M, Meyers BC, Leebens-Mack J (2020) Sex determination by two Y-linked genes in garden asparagus. Plant Cell 32:1790–1796

Huang M, Liu X, Zhou Y, Summers RM, Zhang Z (2019) BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8:1–12

Kistner EJ, Hatfield JL (2018) Potential geographic distribution of Palmer amaranth under current and future climates. Agric Environ Lett 3:1–5

Klein RR, Klein PE, Mullet JE, Minx P, Rooney WL, Schertz KF (2005) Fertility restorer locus Rf1 of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12. Theor Appl Genet 111:994–1012

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: An information aesthetic for comparative genomics. Genome Res 19:1639–1645

Legros M, Marshall JM, Macfadyen S, Hayes KR, Sheppard A, Barrett LG (2021) Gene drive strategies of pest control in agricultural systems: Challenges and opportunities. Evol Appl 14:2162–2178

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v2 00:1–3

Li H (2021) New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572–4574

Liao Y, Smyth GK, Shi W (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930

Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. BMC Biol 15:1–15

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:1–21

Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J (2022) GENESPACE tracks regions of interest and gene copy number

variation across multiple genomes. Elife 11:1–20

Ma X, Vaistij FE, Li Y, Jansen van Rensburg WS, Harvey S, Bairu MW, Venter SL, Mavengahama S, Ning Z, Graham IA, Van Deynze A, Van de Peer Y, Denby KJ (2021) A chromosome-level *Amaranthus cruentus* genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop. Plant J 107:613–628

Manalil S, Coast O, Werth J, Chauhan BS (2017) Weed management in cotton (*Gossypium hirsutum* L.) through weed-crop competition: A review. Crop Prot 95:53–59

Massinga RA, Currie RS, Horak MJ, Boyer J, Massinga RA, Currie RS (2001) Interference of Palmer amaranth in corn. Weed Sci 49:202–208

Melonek J, Duarte J, Martin J, Beuf L, Murigneux A, Varenne P, Comadran J, Specel S, Levadoux S, Bernath-Levin K, Torney F, Pichon JP, Perez P, Small I (2021) The genetic basis of cytoplasmic male sterility and fertility restoration in wheat. Nat Commun 12

Miga KH (2020) Centromere studies in the era of 'telomere-to-telomere' genomics. Exp Cell Res 394:112127

Montgomery JS, Giacomini D, Waithaka B, Lanz C, Murphy BP, Campe R, Lerchl J, Landes A, Gatzmann F, Janssen A, Antonise R, Patterson E, Weigel D, Tranel PJ (2020) Draft genomes of *Amaranthus tuberculatus*, *Amaranthus hybridus*, and *Amaranthus palmeri*. Genome Biol Evol 12:1988–1993

Montgomery JS, Giacomini DA, Weigel D, Tranel PJ (2021) Male-specific Y-chromosomal regions in waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). New Phytol 229:3522–3533

Montgomery JS, Sadeque A, Giacomini DA, Brown PJ, Tranel PJ (2019) Sex-specific markers

for waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). Weed Sci 67:412–418

Morgan G, Baumann P, Chandler J (2001) Competitive impact of Palmer amaranth (*Amaranthus palmeri*) on cotton (*Gossypium hirsutum*) development and yield. Weed Technol 15:408–412

Murray MJ (1940) The genetics of sex determination in the family Amaranthaceae. Genetics 25:409–431

Neve P (2018) Gene drive systems: do they have a place in agricultural weed management? Pest Manag Sci 74:2671–2679

Neves CJ, Matzrafi M, Thiele M, Lorant A, Mesgaran MB, Stetter MG (2020) Male linked genomic region determines sex in dioecious *Amaranthus palmeri*. J Hered 111:606–612

Raiyemo DA, Bobadilla LK, Tranel PJ (2023) Genomic profiling of dioecious *Amaranthus* species provides novel insights into species relatedness and sex genes. BMC Biol 21:1–18

Roberts J, Florentine S (2022) A review of the biology, distribution patterns and management of the invasive species *Amaranthus palmeri* S. Watson (Palmer amaranth): Current and future management challenges. Weed Res 62:113–122

Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

Rochette NC, Rivera-Colón AG, Catchen JM (2019) Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol 28:4737–4754

Rode NO, Estoup A, Bourguet D, Courtier-Orgogozo V, Débarre F (2019) Population management using gene drive: molecular design, models of spread dynamics and assessment of ecological risks. Conserv Genet 20:671–690

Sauer J (1957) Recent migration and evolution of the dioecious amaranths. Evolution 11:11–31

Sellers BA, Smeda RJ, Johnson WG, Kendig JA, Ellersieck MR, Science SW, Jun M, Jun NM, Sellers BA, Smeda RJ, Johnson WG, Kendig JA (2003) Comparative growth of six *Amaranthus* species in Missouri. Weed Sci 51:329–333

Steckel LE (2007) The dioecious *Amaranthus* spp.: Here to stay. Weed Technol 21:567–570

Steckel LE, Sprague CL, Stoller EW, Wax LM, Science W, Sprague CL (2004) Temperature effects on germination of nine *Amaranthus* species. Weed Sci 52:217–221

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

Sun Y, Shang L, Zhu QH, Fan L, Guo L (2022) Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci 27:391–401

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res 18:1944–1954

Thapa R, Blair MW (2018) Morphological assessment of cultivated and wild amaranth species diversity. Agronomy 8

Trucco F, Zheng D, Woodyard AJ, Walter JR, Tatum TC, Lane Rayburn A, Tranel PJ (2007) Nonhybrid progeny from crosses of dioecious amaranths: Implications for gene-flow research. Weed Sci 55:119–122

Uyttewaal M, Arnal N, Quadrado M, Martin-Canadell A, Vrielynck N, Hiard S, Gherbi H, Bendahmane A, Budar F, Mireau H (2008) Characterization of *Raphanus sativus* pentatricopeptide repeat proteins encoded by the fertility restorer locus for Ogura cytoplasmic male sterility. Plant Cell 20:3331–3345

Wang H, Xu D, Wang S, Wang A, Lei L, Jiang F, Yang B, Yuan L, Chen R, Zhang Y, Fan W

(2023) Chromosome-scale *Amaranthus tricolor* genome provides insights into the evolution of the genus *Amaranthus* and the mechanism of betalain biosynthesis. DNA Res 30:1–15

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH (2012) MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:1–14

Ward SM, Webster TM, Steckel LE (2013) Palmer amaranth (*Amaranthus palmeri*): A review. Weed Technol 27:12–27

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019) Welcome to the Tidyverse. J Open Source Softw 4:1686

Wu W, Jernstedt J, Mesgaran MB (2023) Comparative floral development in male and female plants of Palmer amaranth (*Amaranthus palmeri*). Am J Bot 110:1–10

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X (2021) rMVP: A Memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics, Proteomics Bioinforma 19:619–628

CHAPTER 6: CONCLUDING REMARKS

Pigweeds, including *Amaranthus tuberculatus* and *A. palmeri*, are emerging as among the most invasive and economically and agriculturally damaging weed species in the world. For example, *A. palmeri* has been reported in 45 countries and has been predicted to likely further expand to additional regions due to changing climatic conditions. This rapid expansion of pigweeds is not just a threat to crops but also to native species. While chemical control has historically been used and is still used for managing pigweeds, the evolution of resistant biotypes to herbicides from different modes of action have spurred scientists and farmers to begin to explore more options for pigweed control. The similarity in morphology among pigweeds also adds to the challenge of proper identification that hampers timely and effective management decisions. With recent developments in the field of sequencing and computational biology, we can further deepen our understanding of the biology of pigweeds, including the evolution of dioecy, adaptive traits, and herbicide resistance.

In Chapter 2, I began by exploring the relationships among the dioecious amaranths utilizing a Mash distance method that bypasses limitations associated with alignment-based methods. Result of this analysis revealed clustering patterns that were consistent with earlier taxonomic ordering of the dioecious amaranths that were based on comparative morphology (see Figure 3 in Sauer 1957) (Raiyemo et al. 2023). Although inference on species evolution is not recommended with Mash distances, the grouping of waterhemp and Palmer amaranth into separate clusters further supports previous findings on at least two origins of dioecy evolution within the *Amaranthus* genus (Raiyemo et al. 2023). In addition, the robustness of Mash distance in this study indicates its usefulness in species diagnostics, where unknown individuals could be sequenced at low coverage and clustered with known species using Mash.

The short-read sequence data from Chapter 2 also enabled a closer look into the relationships among the dioecious amaranths using their complete chloroplast genomes, which could provide better resolution than previous studies where few chloroplast markers were used (Waselkov et al. 2018). In Chapter 3, I assembled the chloroplast genomes of the dioecious amaranths and explored phylogenetic relationships among the species. Similar to the clustering from the previous Mash tree that utilized nuclear genome sequences, some relationships among the dioecious amaranths were consistent (e.g., relationships between *A. australis* and *A. cannabinus*, *A. arenicola* and *A. greggii*, and *A. tuberculatus* and *A. floridanus*). The relationship of *A. australis* + *A. cannabinus* lineage to the rest of the dioecious amaranths however remains unclear (Raiyemo and Tranel 2023). Rapid radiation was suggested as a reason why the relationships among five major clades of Amaranthaceae s.l. may remain unknown (Morales-Briones et al. 2021). It is possible that such a phenomenon is playing out at the species level, and thus the true relationships among some dioecious amaranths will remain difficult to ascertain. *Amaranthus palmeri* and *A. watsonii* have long been thought to be closely related, occupying the same range (Sauer 1957). It now appears both species might be difficult to distinguish based on chloroplast sequence markers (Raiyemo and Tranel 2023). Murphy et al. (2023) recently developed a protein biotyping assay that uses Matrix-Assisted Laser Desorption Ionization Time of Flight Mass Spectrometry (MALDI-TOF-MS) to identify *Amaranthus* species. While the assay had a high accuracy in distinguishing several *Amaranthus* species, it could not differentiate between *A. palmeri* and *A. watsonii*.

In Chapter 4, I characterized the sex chromosomes in *A. tuberculatus* using comparative genomics approaches. The result of the analyses revealed a contiguous region of ~32.8 Mb in the middle of chromosome 1 that is gene-poor but LTR-rich as the sex-determining region. While

there were no obvious candidate genes involved in sex determination, two genes within the sex-determining region, encoding a FLOWERING LOCUS C-like and a LOB domain-containing protein 19-like were downregulated in males across shoot apical meristem, floral meristem, and mature flower. Alternative splicing of *FLOWERING LOCUS C* intron1 and the transcription of *FLC* have been shown to be regulated by serine/arginine-rich proteins which in turn affects flowering time in *Arabidopsis* (Yan et al. 2017; Wang et al. 2023) while LOB domain-containing proteins are multifunctional and have been implicated in floral organ initiation or development in several species (Majer and Hochholdinger 2011; Zhang et al. 2020; Ma et al. 2022). In addition, chromosome 1 (the sex chromosome) of *A. tuberculatus* likely evolved from the fusion of two chromosomes that are ancestral to chromosomes 16 and 10 in *A. tricolor* (or 12 and 15 in *A. palmeri*).

Characterization of the sex chromosome of *A. palmeri* in Chapter 5 revealed a contiguous 2.84 Mb male-specific region at the distal end of chromosome 3, that is syntenic to a region of scaffold 20 that was previously identified as male-specific in a draft assembly. Thirty-seven genes within the region were identified, in which two genes, encoding an Rf1 (restorer of fertility) protein and TLC domain-containing protein were upregulated in males across shoot apical meristem, floral meristem, and mature flower. We hypothesize that the *Rf1* gene could be functioning as the male-promoting factor, although it is unclear whether the gene also acts as the suppressor of female function. The interplay among male sterility (caused by cytoplasmic-nuclear interactions), sex determination, and pollen fitness in *A. palmeri* will however require further investigation. Contrary to *A. tuberculatus* sex-determining region, where the presence of inversions within the centromeric and pericentromeric region of chromosome 1 contributes to

recombination suppression, we hypothesize that hemizygosity or the absence of the male-specific region of the Y-haplotype on the X chromosome contributes to recombination suppression.

In sum, the genomic analyses in this dissertation provide information that will be valuable for future research. I have identified the sex-determining regions of two important dioecious agronomic weeds (waterhemp and Palmer amaranth), the genes within the region, and sex-biased expression of the genes. The mechanistic process for which the genes within the sex-determining region are regulated, or how they affect sex determination requires additional experimentation. While the long-term goal of studying dioecy in amaranths is to find sex determinants that could be utilized in a meiotic drive to suppress or eliminate populations that are difficult to control, it is noteworthy that any of the candidate genes within the sex-determining region or autosomes with role in floral organ development could in theory be utilized for a meiotic drive e.g., inheritance of a genetic factor that leads to pollen abortion. However, such a genetic factor might require constant or yearly introduction into the population for any long-term seed bank depletion compared to when a sex determinant is used for a meiotic drive to bias population towards a particular sex. Overcoming the challenges with functional validation will provide the opportunity to explore which of the candidate genes identified in this study could be utilized for a genetic control strategy.

6.1 REFERENCES

Ma X, Yu L, Fatima M, Wadlington WH, Hulse-Kemp AM, Zhang X, Zhang S, Xu X, Wang J, Huang H, Lin J, Deng B, Liao Z, Yang Z, Ma Y, Tang H, Van Deynze A, Ming R (2022) The spinach YY genome reveals sex chromosome evolution, domestication, and introgression history of the species. Genome Biol 23:1–30

Majer C, Hochholdinger F (2011) Defining the boundaries: Structure and function of LOB domain proteins. Trends Plant Sci 16:47–52

Morales-Briones DF, Kadereit G, Tefarikis DT, Moore MJ, Smith SA, Brockington SF, Timoneda A, Yim WC, Cushman JC, Yang Y (2021) Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.l. Syst Biol 70:219–235

Murphy M, Hubert J, Wang R, Galindo-González L (2023) Seed protein biotyping in *Amaranthus* species: a tool for rapid identification of weedy amaranths of concern. Plant Methods 19:1–14

Raiyemo DA, Bobadilla LK, Tranel PJ (2023) Genomic profiling of dioecious *Amaranthus* species provides novel insights into species relatedness and sex genes. BMC Biol 21:1–18

Raiyemo DA, Tranel PJ (2023) Comparative analysis of dioecious *Amaranthus* plastomes and phylogenomic implications within Amaranthaceae s.s. BMC Ecol Evol 23:15

Sauer J (1957) Recent migration and evolution of the dioecious amaranths. Evolution (N Y) 11:11–31

Wang T, Wang X, Wang H, Yu C, Xiao C, Zhao Y, Han H, Zhao S, Shao Q, Zhu J, Zhao Y, Wang P, Ma C (2023) *Arabidopsis* SRPKII family proteins regulate flowering via phosphorylation of SR proteins and effects on gene expression and alternative splicing.

New Phytol 238:1889–1907

Waselkov KE, Boleda AS, Olsen KM (2018) A phylogeny of the genus *Amaranthus* (Amaranthaceae) based on several low-copy nuclear loci and chloroplast regions. Syst Bot 43:439–458

Yan Q, Xia X, Sun Z, Fang Y (2017) Depletion of *Arabidopsis* SC35 and SC35-like serine/arginine-rich proteins affects the transcription and splicing of a subset of genes. PLoS Genet 13:1–29

Zhang Y, Li Z, Ma B, Hou Q, Wan X (2020) Phylogeny and functions of LOB domain proteins in plants. Int J Mol Sci 21

APPENDIX A: SUPPLEMENTARY INFORMATION TO CHAPTER 2

Appendix A Figure A.1 – A.12: GenomeScope plots and smudgeplots for the nine dioecious *Amaranthus* species. Available at IDEALS (https://)

Appendix A Figure A.13 – A.24: Proportion of repeats identified in the nine dioecious Amaranthus species. Available at IDEALS (https://)

Appendix A Figure A.25 – A.27: Upset plots delineating the number of shared scaffolds with male or female-enriched coverages and reads alignment coverage of male-to-female individuals for dioecious *Amaranthus* species across *A. palmeri* Scaffold 19 and *A. tuberculatus FT* on tig00000542. Available at IDEALS (https://)

Appendix A Table A.1: *Amaranthus* species and other members of Caryophyllales used in this study. Available at IDEALS (https://)

Appendix A Table A.2 – A.7: Genome size estimates, repeat composition, and statistics of short-reads alignment to *A. palmeri* or *A. tuberculatus* draft genome assemblies. Available at IDEALS (https://)

Appendix A Table A.8 – A.17: Coverage analysis of short reads of dioecious amaranths mapped to *A. palmeri* or *A. tuberculatus* draft genome assemblies. Available at IDEALS (https://)

Appendix A Table A.18 – A.19: Gene models within *A. palmeri* male-specific Y region exhibiting male-specific coverage for *A. watsonii* mapped reads, and transcription factors within *A. palmeri* male-specific Y region. Available at IDEALS (https://)

APPENDIX B: SUPPLEMENTARY INFORMATION TO CHAPTER 3

Appendix B Figure B.1: Sliding window analysis of nucleotide diversity among nineteen chloroplast genomes of *Amaranthus* species. Available at IDEALS (https://)

Appendix B Figure B.2 – B.6: Phylogenetic trees, bootstrap consensus network, and NeighborNet splits graph of *Amaranthus* species and other species in Amaranthaceae s.s. Available at IDEALS (https://)

Appendix B Table B.1 – B.4: Sequence information for species used in phylogenomic analysis, chloroplast features, and nuclear rDNA assembly size of species assembled in this study. Available at IDEALS (https://)

Appendix B Table B.5: Relative synonymous codon usage (RSCU) of 78 protein-coding genes in the chloroplast genome of *Amaranthus tuberculatus*. Available at IDEALS (https://)

Appendix B Table B.6: Estimates of evolutionary divergence between ITS sequences of 14 species. Available at IDEALS (https://)

Appendix B Table B.7: Estimates of evolutionary divergence between nuclear rDNA sequence assembly of 14 *Amaranthus* species. Available at IDEALS (https://)

APPENDIX C: SUPPLEMENTARY INFORMATION TO CHAPTER 4

Appendix C Figure C.1 – C.16: Heatmaps of tandem repeat structures, search of the 572 bp male-specific marker to a database of transposable element encoded proteins, dotplots of alignment male-specific contigs in Montgomery et al. (2021) and *A. tuberculatus* haplotype assemblies, genomic features of *A. tuberculatus* haplotype assemblies, pairwise synonymous divergence (d_s), schematic representation of trees used in CODEML analysis, phylogenetic tree of *FLOWERING LOCUS T* copies, sequence alignment of MADS-box transcription factor 18 (MADS18) and MADS-box protein FLOWERING LOCUS C-like (FLC-like), and phylogenetic tree of LOB domain-containing protein (LBD). Available at IDEALS (https://)

Appendix Note 1: Library preparation and sequencing, contig assembly, hybrid scaffolding and pseudomolecule construction, and genome annotation. Available at IDEALS (https://)

Appendix Note 2: Genome characteristics and repeat analysis. Available at IDEALS (https://)

Appendix C Table C.1 – C.29: Repeat composition, centromeric and telomeric repeats identification, statistics of protein-coding genes, non-coding RNAs and tRNA annotations, disease resistance gene annotation, transcription factor annotation, significant SNPs from GWA analysis, microRNAs on chromosome 1, reciprocal best hit (RBH) search of 147 genes on male-specific contigs in Montgomery et al. (2021) to the haplotype assemblies, inversions between haplomes 1 and 2 across the genome, gene density parameter estimates, Intact-LTR proportion and insertion times (mya) parameter estimates, pairwise species *ds* comparisons for regions on chromosome 1, test of positive selection under three models for single-copy orthologous genes on chromosome 1, differentially expressed genes in male vs female comparison across shoot

apical meristem, floral meristem and mature flower for *A. tuberculatus*, and GO enrichment analyses. Available at IDEALS (https://)

APPENDIX D: SUPPLEMENTARY INFORMATION TO CHAPTER 5

Appendix D Figure D.1 – D.16: Genomic features of the two haplotype assemblies of *A. palmeri*, dotplots of sequence alignment between the haplotype assemblies and previous draft assembly, and phylogenetic tree of coding sequences of *PPRs* and *Rf1* on chromosome 3 of both haplotype assemblies. Available at IDEALS (https://)

Appendix D Table D.1 – D.23: Repeat composition, centromeric and telomeric repeats identification, statistics of protein-coding genes and tRNA annotations, disease resistance gene annotation, transcription factor annotation, reciprocal best hit (RBH) search of 121 genes on male-specific contigs in Montgomery et al. (2021) to the haplotype assemblies, inversions between haplomes 1 and 2 across the genome, differentially expressed genes in male vs female comparison across shoot apical meristem, floral meristem and mature flower for *A. palmeri*, and GO enrichment analyses. Available at IDEALS (https://)