The evolution, diversity, and genetic architecture of sex in waterhemp (*Amaranthus tuberculatus*)

Julia M. Kreiner[1], Marco Todesco[2,3,4], Natalia Bercovich[2], Yunchen Gong[5], Cassandra Elphinstone[2], Patrick J. Tranel[6], Loren Rieseberg[2], Stephen Wright[5]

[1]Department of Ecology & Evolution, University of Chicago
[2]Department of Botany, University of British Columbia
[3]Michael Smith Laboratory, University of British Columbia, Vancouver
[4]Department of Biology, University of British Columbia, Kelowna
[5]Department of Ecology & Evolutionary Biology, University of Toronto
[6]Department of Crop Sciences, University of Illinois Urbana-Champaign

Abstract

The evolution of separate sexes is hypothesized to occur through distinct pathways involving few large-effect or many small-effect alleles. However, the genetic architecture of sex may itself evolve through selection to suppress recombination in order to maintain beneficial combinations of alleles, potentially at the cost of losing functional genetic variation. To explore these processes, we leveraged the recent transition of *Amaranthus tuberculatus* to dioecy within a predominantly monoecious genus, along with a sex-phenotyped population genomic dataset, and six newly generated chromosome-level haplotype phased assemblies. We identify a ~4 Mb region strongly associated with sex through genotype and sequence-depth based analyses. Comparative genomics of these proto-sex chromosomes within the species and across the Amaranthus genus demonstrates remarkable variability in their structure and genic content, including numerous polymorphic inversions. No such inversion underlies the extended linkage we observe associated with sex determination. Instead, we identify a complex copy number polymorphism that is differentiated across sexes, is variable across ancestral lineages, geographical scales, and habitats, but incompletely explains sex as phenotyped—over 10% of individuals show phenotype-genotype mismatch in the sex-linked region. Together with our novel observation of leakiness in expression of sex within the species, these findings imply the presence of multiple interacting determinants of sex. The evolution of separate sexes and its genetic architecture in *A. tuberculatus* therefore appears to be ongoing, with modifiers of sex that permit gene exchange and the maintenance of diversity in proto-sex chromosomes.

Introduction

Dioecy is thought to primarily evolve from cosexuality through two distinct pathways (gynodioecy and monoecy-paradioecy) that result in contrasting predictions about the importance of small versus large effect mutations [1]. In the gynodioecy pathway, the intermediate stage can be characterized by a single large-effect male sterility mutation that invades a cosexual population and transitions it to polymorphic for females and hermaphrodites, eventually resolved to dioecy by the invasion of a female sterility mutation [2]. Recombination suppression is thought to evolve to prevent fully sterile genotypes, and may be further extended due to sexually antagonistic selection that assembles male and/or female beneficial alleles on alternate haplotypes [3]. The lack of recombination in the sex-linked region is thought to drive the degeneration of the non-recombining Y/W chromosomes, although contemporary work shows extensive variation in size and the degree of degeneration, which is not a simple function of age [4] [5–9]. Nonetheless, young sex determining systems in the early stages of this process more often tend to demonstrate incomplete recombination suppression and leaky sex determination, despite multiple sexually antagonistic polymorphisms having evolved [10,11].

In dioecious species that evolve through the monoecy-paradioecy pathway, selection is thought to act on quantitative variation for sex allocation. Populations transition from individuals which are equally invested in male and female function to alternately fixed for one or the other through gradual disruptive selection [12–14]. Tests of this route to separate sexes have been limited by the relative infrequency of dioecious species in predominantly monoecious plant genera and families (but see *Sagittaria* [1], *Urtrica [15]*, and *Poplar* [16]). Dioecious species that have recently evolved from monoecy are particularly prone to leakiness in gender expression [13,17–19], implicating the possible role of gradual evolution of sex allocation, multiple loci and/or expression modifiers governing this transition. Recent models suggest that while polygenic variation for sex allocation is initially selected upon, this pathway may also resolve in the subsequent emergence of single-locus sex determination[14]. Therefore, recent transitions to dioecy from monoecy provide the opportunity to investigate the timescale of recombination suppression and the evolution of the genetic architecture of sex allocation and determination.

The genus *Amaranthus* (Amaranthaceae) contains about 70 species, of which the vast majority are monoecious. There have likely been multiple relatively recent transitions to dioecy, as two to three dioecious clades are spread across two out of the three subgenera (*Amaranthus* and *Acnida*, but not *Albersia*) [20–23]. All dioecious species in this genus are thought to result from a male heterogametic (XY) system, with homomorphic (lack of apparent morphological differences) sex chromosomes [24,25]. The genus is economically important, as it contains both domesticated pseudograin crops and invasive agricultural weeds, including the focal species for this study, the wind pollinated and annual *A. tuberculatus* (common waterhemp). *Amaranthus tuberculatus* can hybridize with monoecious relatives at relatively high rates under field conditions, highlighting the recent evolution of dioecy in the clade [24,26,27]. Understanding the genetic basis of sex in the genus is therefore important for improving breeding efforts and the efficiency of artificial selection (e.g. [28]), as well as for informing evolutionary management approaches such as gene drive to suppress females and accelerate local extirpation of invasive weed populations [29,30].

Work on the causes and consequences of sex in *A. tuberculatus* illustrates sexual dimorphism, apparent at both the phenotypic and genomic level, despite the likely recent origin of separate sexes. Phenotyped males and females are differentiated in key life history traits, with males being faster growing and earlier flowering than females, and with sex differences in genome size and content of particular repeat classes [31,32]. Furthermore, levels of sexual dimorphism between males and females vary by environment (whether plants were collected from natural or agricultural habitats) [32], suggesting that sexually antagonistic selection is heterogeneous across environments. Recent genomic work confirmed a male heterogametic system, with a cumulative 4.6 Mb of male-specific sequence identified through sex-specific RAD-tags [30], with loci spread across multiple scaffolds, likely due to reference quality. Following this work, Raiyemo et al. [33] produced a phased reference genome identifying chromosome 1 as enriched for sex linked SNPs, although their assemblies lacked a contiguous sex-linked region based on genome wide association (GWA) and F_{ST} analyses.

Here, we generate several new haplotype-phased assemblies and analyze them along with considerable resequencing resources to resolve the diversity and evolution of proto-sex chromosomes in *A. tuberculatus*. We first implement population genomic approaches to resolve the location of the SDR along the chromosome. Leveraging our multiple assemblies and comparative genomic approaches, we show sex determination in *A. tuberculatus* to result from a polymorphic region ~4.5 Mb long, not found within an inversion but rather differentiated in copy number across the sexes. The lack of 1:1 mapping of phenotype onto any genotype across the genome suggests that additional genetic or environmental modifiers may be contributing to sex expression in the system, and/or that unnoticed quantitative variation for sex may exist. This unquantified variation in the determination and expression of sex has allowed for on-going gene exchange between the proto-X and -Y, such that sex-linked haplotypes retain remarkable levels of variability structured by ancestry, geography, and habitat. Taken together, our results reveal the genomic heterogeneity in sex determination and expression that can result during the transition from monoecy to dioecy.

Results

Characterization of the copy number polymorphic sex-linked region

To identify the location of the sex-linked region in our newly produced reference genomes, we performed a GWA using 186 sex-phenotyped individuals (96 female, 92 male) from paired environmental collections [31] that we mapped to one randomly chosen haplotype (haplotype 2) of the highest coverage (70X) assembly. This genome came from a male plant from Walpole, Ontario, Canada, a population admixed for ancestry between the two varieties, var. *rudis* and var. *tuberculatus*. A GWA identified the largest scaffold, Scaffold_1 as highly enriched for sex-linked alleles, consistent with recent work ([33]; **Sup Figure 1)**. To attempt to resolve the mapping of reads with orthology to both the putative X and Y, we used the enrichment of sex-linked SNPs on Scaffold_1 to inform the production of a modified phased-haplotype level reference: containing only haplotype 1 for all scaffolds except Scaffold_1, which was represented by both haplotype 1 and 2. We remapped, recalled, and refiltered SNPs (see methods) before reanalyzing genome-wide associations (and all other population genomic analyses) for sex using this competitive mapping approach.

Genotype-phenotype associations resolved a large sex-linked region, ~4.0 Mb long, with a near continuous butte of extreme genotypic correlations with sex on Scaffold 1, haplotype 2, but not haplotype 1 (**Figure 1A**; **Sup Fig 2;** but note associations found across other scaffolds, explored further in *Complex copy number variation***)**. This sex-linked region contained 103 annotated genes, including sex-determining (SD) candidates such as TOE2 (*Target of Early Activation) [34]*, RAP2 [35], MADS8 [36], and FT/HD3 [37]. Notably, the sex-linked region is about 10 Mb downstream of the putative centromeric region according to RepeatOBserver (which uses a Fourier transform of DNA walks to identify putative centromere locations; [38]). Even relative to the nearby centromere, this region shows the highest patterns of 0-fold diversity and lowest gene density across the chromosome based on local regression fits (**Sup Fig 3, Sup Figure 4)**. This GWA also shows a large number of significant associations (beyond Bonferroni correction) off of Scaffold 1. We tested the redundancy/information value of these off-scaffold hits using a lasso regression approach. For the 595/2356 loci with no-missing data out of those with p values $< 10^{-15}$, allowing for the optimal shrinkage of predictors results in 26 loci with non-zero effects on sex, for which 22 are found on Scaffold 1 (**Sup Figure 5)**.

To identify whether the sex-linked region was male or female specific, we estimated copy number in these same samples. By scaling each individual's median read-mapped coverage in 100bp windows within the sex-linked region, by the median read-mapped coverage on Scaffold_1 outside of the sex-linked region, we attained estimates of fine-scale copy number. We see that while median male and female copy number show complete overlap for most of Scaffold_1, we see clear divergence among sexes in sequence presence/absence between 40.5-43.5 [and a secondary peak around 44.5] Mb. Most males show a median copy number on average around 0.75 across this region (light yellow in **Figure 1B**), while most females (black in **Figure 1B)** show a median copy number closer to 0.25. While this is distinct from the expectation of a Y-specific region, for which the median copy number should hover around 1 and 0 in the Y and X respectively, low divergence between the proto-sex chromosomes means that even with competitive mapping, Y reads may map to the X and vise-versa. These coverage

based analyses suggest that the sex-linked region identified in the GWA (and supported by further Fst and heterozygosity based analyses; **Sup Figure 6**) maps to the Y, which has either recently expanded in gene content relative to the X, or these genes have have been recently lost on the X. Comparison of the synteny of Scaffold_1 haplotype 1 and haplotype 2 assemblies using genespace[39] supports these population-level copy number polymorphism inferences, as we find our haplotype 2 (proto-Y) reference lacks syntenic orthologous genes on haplotype 1 (proto-X) in the SLR.

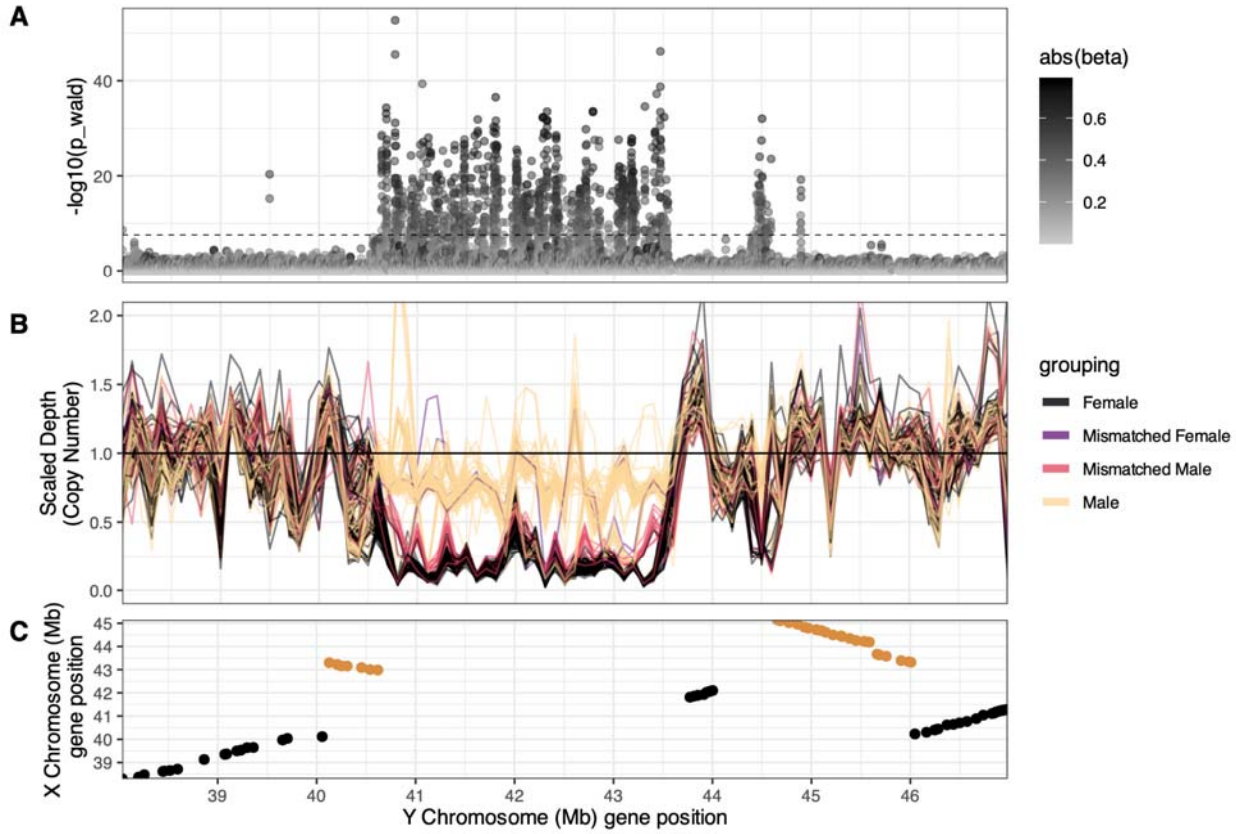


Figure 1. The sex-linked region (SLR) in *A. tuberculatus* on Scaffold 1**. A**) A GWA resolved strongly sex-linked alleles along a 4 Mb stretch on haplotype 2 of Scaffold 1, between 40.65-43.5[44.5] Mb. Color of points depicts the absolute allelic effect size from mixed model associations. The horizontal dashed bar represents the Bonferroni p=0.05 threshold. **B***)* Sex-based differentiation in copy number in the SLR illustrates fine-scale variation within sexes, and genotype-phenotype mismatch. Copy number was calculated through scaling depth in the SLR, by coverage on Scaffold_1 outside of the SLR, in 50 kb genomic windows. **C**) Genes falling within the SLR on the Y (haplotype 2 assembly) lack orthologous copies on the X (haplotype 1 assembly).

Given the apparent role of copy number variation in sex-determination, we tested whether genome-wide sex-associated loci also showed such a signal. To do so, we extracted all sex-associated alleles passing multiple test correction off of Scaffold_1 (n_{genome}=864, $n_{scaffold\_1}$=2011) and calculated the SNP-level ratio of male:female coverage. Coverage at these sites are elevated to an average ratio of 1.2 (95% CI = [1.01,1.52]). The mean of the distribution of male:female coverage at these genome-wide loci is similar to as is seen in the SLR (mean = 1.49; 95% CI = [0.99, 2.59]; **Sup Figure 7**). Since it is unlikely these represent independent sex-linked CNV distributed all across the genome, and as these reads/SNPs do not drop out after stringent

filtering based on mapping quality, this suggests that about 20% of males have sex-linked paralogues on their Y haplotype that are absent in our reference Y (as in [40]). Therefore, while Y-linked regions are typically thought of as degenerate and lacking variation, the Y haplotypes studied here are surprisingly diverse.

Structural variation across male waterhemp genomes and the sex-linked region

The six haplotype-phased chromosome-level assemblies we generated with PacBio HiFi and Hi-C data provided the opportunity to further resolve the extent of structural heterogeneity and evolution in *Amaranthus* (**Sup Table 1)**. Consistent with recent reports, our analyses show that the two largest scaffolds/chromosomes in *A. tuberculatus* (chromosomes 1 and 2) were formed through a chromosomal fusion (i.e. Robertsonian translocation) since the split of *A. tuberculatus* to *A. tricolor, palmeri, and crutensus* (**Figure 2A)**.

We find widespread signals of intraspecific structural variation in *A. tuberculatus*. Each genome shows widespread signals of an ancient whole genome duplication as suggested in [41], with ohnologues present on alternate scaffolds (e.g. **Sup Figure 8, 9)**. When comparing the 6 *A. tuberculatus* haplotype assemblies to each other, we find numerous inversions, duplications, and translocations (**Figure 2A**). To quantify the extent of structural rearrangement, we calculated the number of orthologous groups (contiguous tracks of syntenic sequence) within each scaffold for each pair of haplotype assemblies. We found on average ~5-8 orthologous groups per scaffold depending on the paired comparison (**Figure 2B**). Notably, the SDR containing Scaffold_1 is significantly enriched over all other scaffolds for the number of orthologous groups (with the upper and lower quartile ranging from 13-20; chromosome effect: $F = 2.9352$, $P = 0.00157$, all pairwise comparisons $p < 0.05$ except for with Scaffold_9), as is particularly apparent by the large number of inversions segregating across these haplotypes (n = 11; **Figure 2C, D**). Scaffold_1 is the largest scaffold in the assembly, and if we control for scaffold length, it is no longer an outlier for the number of orthologous groups. This suggests that while the amount of structural variation scales with chromosome size, sex-determining and sex-linked regions (SLRs) may be more likely to arise on or result in regions of the genome enriched for structural complexity.

Given the number of structural rearrangements present on the chromosome containing the SLR, and the common role of structural variation in limiting recombination along SLRs [42–44], we tested the extent to which they may be sex-linked. To do so, we interrogated both pairwise comparisons of the X-containing and Y-containing haplotypes of Scaffold_1 within an individual, and all-vs-all assembly comparisons (i.e. Genespace). While a number of private and low frequency inversions are present, segregating between just X-containing assemblies (**Sup Fig 10**) and between just Y containing assemblies (**Sup Fig 11**), one inversion appears to segregate consistently between the three X and Y containing haplotypes (**Figure 2C, 2E, Sup Table 2**). This inversion is directly upstream and neighboring the sex-linked region, but remarkably lacks any substantial signal of sex-linkage itself based on population genomic inference (i.e. GWA in **Figure 1**; F_{ST} and Heterozygosity in **Sup Fig 6**). We turned back to our population genomic data to further test whether there was any cryptic population structure by sex or other variables within the bounds of these inversions based on PCA derived from SNP genotypes. Sex does not predict PC1, PC2, or PC3 for any of the inversions contained on Scaffold_1 **(Sup Figure 12)**. However, other variables do: ancestry (proportion of var. *rudis*

versus var. *tuberculatus* based on K=2 grouping of a structure plot) loads on PC2 of Inversion3, PC1 on Inversion11, and PC1 of Inversion12; longitude on PC1 and PC3 of Inversion12; environment (Natural or Agricultural) marginally on PC3 of Inversion8; and latitude marginally on PC2 of Inversion8. These results imply inversions are not facilitating population-level recombination suppression in or around the sex-linked region.

Gene-presence absence variation in the sex-linked region is not specific to just X-Y haplotypic comparisons. When we look directly at the 4Mb sex-linked region and compare gene presence absence across other Y assemblies, we find a stark contrast in the relative abundance of core genes (present in all assemblies) versus pan genes, with core genes being at a frequency of 62% (15,002/24173) genome wide, but only 10% among Y haplotypes within the SLR. Indeed, this is also visually apparent in dot plot based pairwise comparisons of the Y haplotypes, for which orthologous synteny is variable within the sex-linked region (**Sup Fig 13).** Since this is a particularly copy number variable and relatively low gene-density region, it is also plausible that technical issues with assembly of this region has led to gene dropout.

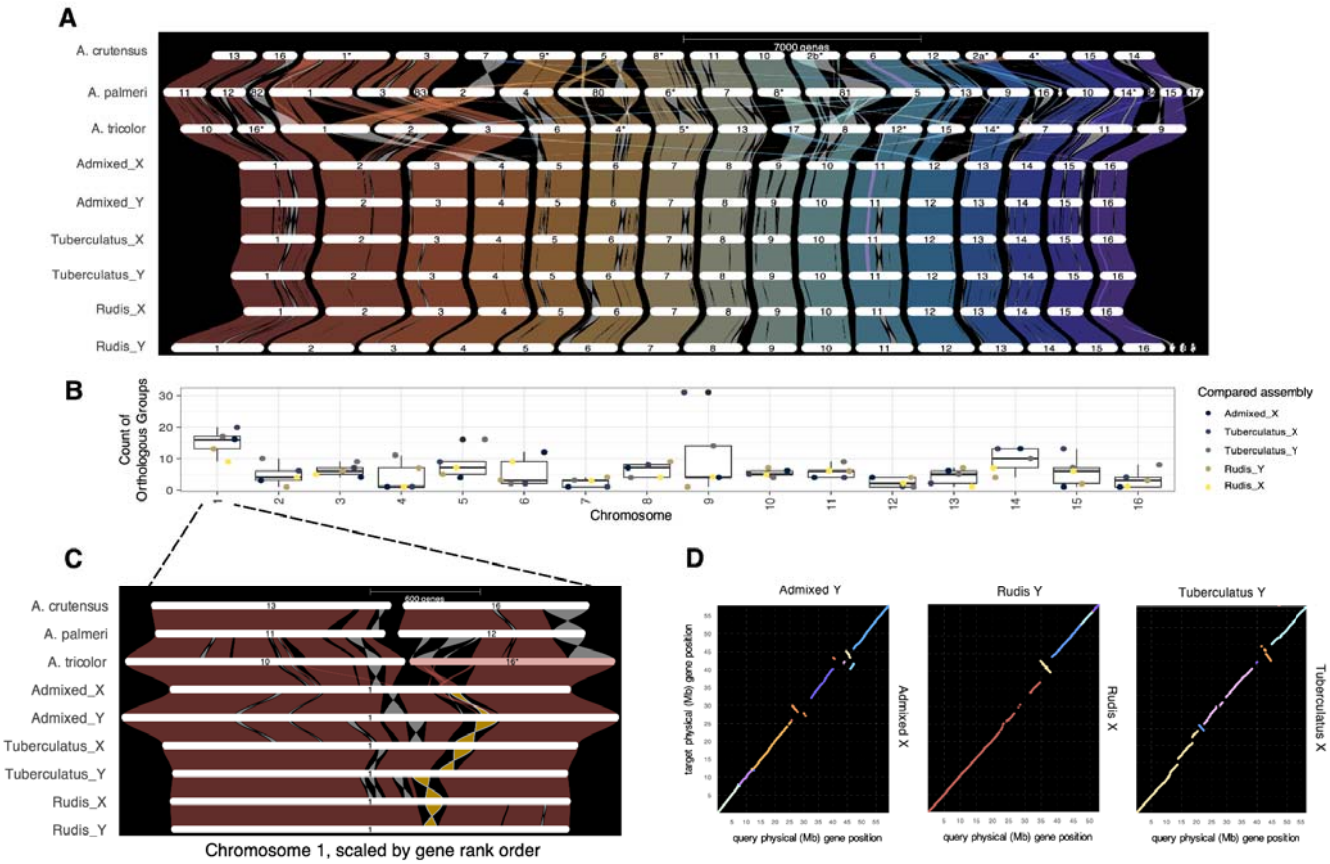


Figure 2. Structural variation within *A. tuberculatus* is enriched on the chromosome containing the sex-linked region. ***A***) A synteny plot from Genespace highlighting the syntenic gene orthologues within *A. tuberculatus* among haplotype-phased assemblies and between *A. tuberculatus* and three congeners (*A. tricolor, A. palmeri*, and *A. crutensus*) with chromosome level assemblies. Inversions are highlighted in light grey, as are translocations which can be distinguished from inversions by their connections between distinct chromosomes. ***B***) The count of orthologous groups between the Y-containing admixed, Walpole genome and each haplotype-level assembly, along with box plot summaries. ***C***) A zoomed in synteny plot of Chromosome 1, which contains the SLR, illustrating the widespread occurrence of inversions and a inversion (highlighted in yellow) that is heterozygous in all three of our males. ***D)*** Pairwise dot plots of syntenic genes between the proto-X and proto-Y chromosome-level assemblies for each male.

Genotype-phenotype mismatch and population structure in the sex-linked region

While in many systems, a single locus is sufficient to fully explain phenotypic variation in sex, this seems not to be the case in *A. tuberculatus.* We see a strong notable exception to the excess of male:female coverage in the SLR for ~10% of individuals in this dataset, which show a mismatch between sex and their average coverage in the SLR (**Figure 1B; Figure 3**). This genotype-phenotype mismatch is apparent not only based on copy number profiles (**Figure 1B, Figure 3C)** but also through PCA and phylogenetic reconstruction of unphased but competitively mapped genotypes in this region for 5/94 (5.3%) and 17/89 (19.1%) of individuals phenotyped as females and males, respectively (**Figure 3A, B)**.

One might expect that individuals that are proximal to the opposite sex in the genotype of their SLR would show phenotypic evidence of this intermediacy. Alternatively, one might consider the possibility of whether sex was mis-phenotyped. To consider these two possibilities, we first independently validated our findings of genotype-phenotype mismatch with a dataset of sexed herbarium sequenced samples for which phenotypes can be reexamined, finding that again, nearly 10% of samples show a genotype-phenotype mismatch **(Sup Fig 14)**. While neither the herbarium or common garden population genomic datasets provided evidence of phenotypic variation in the expression of sex, careful observation of waterhemp reproductive structures in other contexts revealed leakiness in its expression, which we revisit and report for the first time here. As pictured in **Figure 3D**, we discovered an individual from a separate accession and a separate grow out that is made up of entirely female flowers except for one male flower atop an inflorescence. At other times throughout the numerous years of growing waterhemp, we have observed other forms of leaky sex-expression in *A. tuberculatus*, including entire branches that have oppositely sexed flowers relative to the rest of the plant, and even a predominantly female plant with several perfect (hermaphroditic) flowers. This quantitative variation for sex-expression supports the role of multiple SD loci or modifiers acting in the system.

Mean copy number in the SLR (**Figure 3C)** is structured not just by their sex grouping (i.e., genotype-phenotype status; $F_{3,173} = 1867$, p-value $< 2x10^{-16}$), but also by var. *rudis* ancestry ($F_{1,173}=7.38$, p-value = 0.0073), the interaction between ancestry and the ancestry and sex grouping ($F_{3,173}=4.38$, p-value = 0.0054), and marginally habitat type (agricultural or natural; $F_{1,173} = 3.3849$, p-value = 0.067) (**Figure 3**). Genotypic structure in this region depicts an even stronger role of ancestry in shaping sex-linked haplotypes (ancestry: F = 59.6, p-value = 8×10^{-13}; ancestry x sex grouping: ($F_{3,173}=3.5$, p-value = 0.016) in addition to sex grouping ($F_{3,176} = 33$, p-value $< 2x10^{-16}$), latitude ($F_{1,173} = 15.75$, p-value = 0.0001), and habitat type (agricultural or natural; $F_{1,173} = 4.17$, p-value = 0.043). Clearly, the sex-linked region in *A. tuberculatus* retains plenty of diversity that reflects both deep and recent evolutionary history.

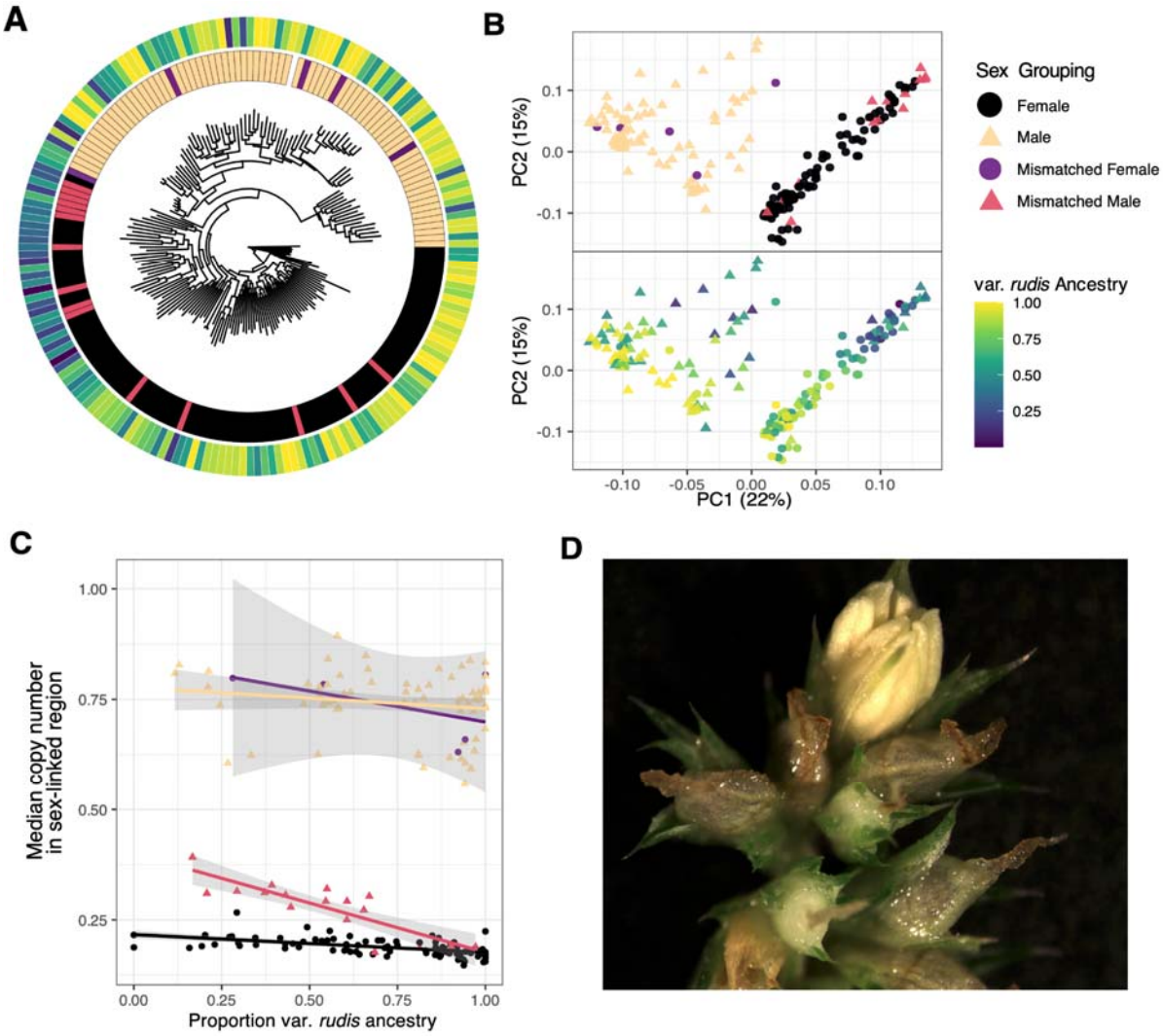


Figure 3. The sex-linked region shows SNP and copy number diversity that reflects strong population structure and leaky sex expression. A) A phylogeny inferred from IQtree of genotypes in the sex-linked region, with tips colored by sex grouping and var. *rudis* ancestry proportion (legend in B). **B**) PCA of genotypes the sex-linked region, colored by sex grouping and var. *rudis* ancestry composition. **C**) The median copy number in the SLR is predicted by the proportion of var. *rudis* ancestry, and by the interaction of sex grouping x ancestry. **D)** A male flower (emerging anthers in light yellow) atop a predominantly female inflorescence (stigmas in brown) in a waterhemp individual.

While multiple loci or modifiers may interact to determine sex, this architecture might also facilitate the maintenance of diversity in the sex-linked regions by allowing for gene-exchange. To test this, we performed LD based inference of recombination rate with LDhat [45] with the competitively mapped reads along Scaffold_1 (**Sup Fig 15**). While the mean effective recombination rate ($4N_eR$) shows a clear drop-off at the border of the sex-linked region, the smoothed moving average does not go to zero, and further indicates peak in recombination in the center of the SLR. Rare recombination events in the SLR might leave a geographic signal. We investigated this by calculating the difference between an individual's copy number in the SLR and the mean value for its respective sex-matched copy number, visualizing its distribution on a map (**Sup Fig 16**), and modeling its predictors in a multiple regression. The degree of copy

number mismatch shows a widespread geographic distribution, with mismatched individuals present in most populations. While latitude, longitude, environment, and ancestry do not predict the degree of mismatch, state is a significant predictor ($F_{4,178} = 2.4607$, $p = 0.047$). While this LD based inference of recombination is not free from technical artifacts resulting from mismapping of repetitive sequence, together these results suggest that gene exchange between the sexes is likely on-going, and contributes to variation at the SLR we describe here.

Despite the average genotype in the SLR being commonly mismatched to phenotypic sex, we sought out fine-scale locus-specific copy number variation that remained unique to one sex—a strong candidate for a SDR. To do so, we performed a GWA for sex based on copy number in 100bp windows, which further resolved heterogeneity in sex-linkage even within this 4Mb region **(Figure 4A)**. While copy number variation at the edges of the SLR tend to show the strongest fidelity between sexes, a CNV within a narrow 100kb tract around 42.32 to 42.33 Mb shows the strongest association with sex despite being surrounded by a valley of significance, and is more closely associated with sex than any SNP-based genotype genome wide (r^2, p-value for top SNP = 0.73, 2×10^{-53}; r^2, p-value for top CNV locus =0.73, 1.2×10^{-54}; **Figure 4B)**. Just downstream of the most significant locus lies the MADS-box transcription factor, SEP1, necessary for petal, stamen and carpel development in Arabidopsis. Copy number at this upstream 100bp window is on average 4 in males and 0 in females, although males show substantial variation (between 0 - 8). The genotype to sex association at this locus is still imperfect however, with a fewer number of mismatched individuals, only 4 males and 7 females resembling the copy number of the opposite sex (~5% percentage mismatch) (**Figure 4C)**.

Previous common garden phenotyping of these samples, not just for sex, but for key life history traits identified a continuum of sexual-dimorphism [31,32], allowing us to test hypotheses about the timing of the evolution of sexual dimorphism versus separate sexes[46,47]. When we perform GWA of traits with significant sexual dimorphism (i.e., biomass, flowering time), only the most sexually dimorphic trait – biomass – produces a GWA profile loaded onto the sex-linked alleles. Genotype-phenotype associations for dry biomass showed a 99.99% overlap with SNPs associated with sex, whereas flowering time, the second most dimorphic, showed 0 overlap with sex-associated SNPs (**Sup Fig 17**). This suggests that shifts in size may have occurred through primary genetic changes during the transition to separate sexes, whereas flowering time may have been a secondary genetic change [46]. Furthermore, we tested whether genotype-phenotype mismatched individuals were more likely to have trait values of the opposite sex, as might be expected if these individuals represent sex intermediates. However, we see no such pattern (**Sup Fig 18**).

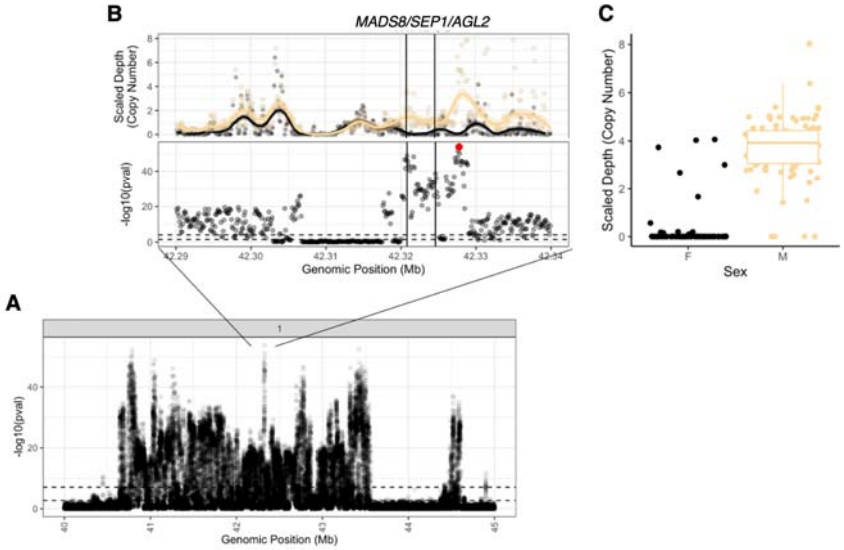


Figure 4. Copy number GWA resolves fine-scale variation in the association with sex within the sex-linked region. A) The Manhattan plot of significance of the association of sex with copy number in 100bp windows across the sex-linked region. **B**) Zooming in on the region of the genome with strongest copy number association with sex, [*upper*] the distribution of mean copy number in 100bp windows for individuals phenotyped as male (yellow) and female (black) [*lower*] and the significance profile of this copy number variation with sex. **C**) Copy number variation at the 100bp window with the strongest association with sex, highlighted in red in B).

Discussion

Our results demonstrate a proto-XY system in *A. tuberculatus* that retains considerable complexity and diversity despite evidence of extended linkage of loci associated with sex. We find that copy number polymorphism between the proto-X and proto-Y are involved in sex determination. However, we find no locus for which copy number variation (or SNP variation) fully explains sex as phenotyped. Nearly 10% of individuals phenotyped as male or female display the genotype of the alternate sex in the SLR. Together with observations of leaky sex expression in the species, our results implicate additional SD factors that facilitate on-going gene exchange at the primary sex-linked locus and therefore the maintenance of diversity in the architecture and expression of sex.

The evolution of dioecy from monoecy has classically contrasted pathways that rely on large or small effect mutations [1]. Recent models on the gradual evolution of separate sexes through selection on sex allocation however outline how polygenic variation for sex allocation can be concentrated to a single locus through ecological selection for sex specialization [14]. Under such a model, XY systems are more likely to evolve when fitness exponentially relates to female allocation, which could result from the coupling of survival and seed production in highly fecund females. Such a scenario is plausible in *Amaranthus tuberculatus*, as it is remarkably difficult to manage in field settings due to the large crop of seed produced by females, estimated to range up to 35,000-1,200,000 seeds per plant [48]. While we identify a large-effect locus capable of explaining 76% of the variation in sex, no SNP or copy number variant within the chromosome 1 SLR shows a perfect correlation with sex. Thus, despite *A. tuberculatus* showing a significant

concentration of the genetic architecture of sex within this region, the persistence of variation in sex determinants suggests that separate sexes may still be actively evolving. The integration of gene expression data with expanded GWA studies holds significant promise for uncovering additional modifiers of sex expression, while estimates of sex allocation-fitness curves in closely related monoecious species will be instrumental in elucidating the importance of a quantitative pathway towards separate sexes.

Species that have evolved dioecy under the quantitative monodioecy-paradioecy pathway are hypothesized to be more susceptible to modifiers of sex-expression [49], whether genetic or environmental. Extensive phenotypic observation of field and herbarium samples in the *Siparuna* genus, where multiple transitions to dioecy has occurred from monoecy, revealed inconsistent or "leaky" sexes as a relatively common phenomenon [50–52], as is also observed in *Mercurialus [53]*. The waterhemp congener, *A. palmeri* which likely represents an independent transition to dioecy [23,29], shows transient hermaphroditism with initial staminate (male) flowers producing both male and female reproductive organs, but the with development of the gynoecium primordia halted early on [54]. At the scale of phenotyping we performed in our initial common garden experiment, nearly four thousand individuals for numerous phenotypes [31], it is possible that we missed cryptic variation in sex expression amongst the millions of flowers present. In the meantime, evidence of leaky sex expression from waterhemp individuals reared in a separate experiment has come to light, with a predominantly female individual displaying a single male flower atop its inflorescence. That variation for sex-determination and expression still exists supports genomic evidence that the transition to dioecy is not complete, and that the evolution of separate sexes in *Amaranthus* proceeded through the quantitative monoecy-paradioecy pathway. Crossing of individuals with alternate matching and mismatching status at the primary SLR will help to map the number and effect size of secondary SD loci, if segregating.

Multiple mechanisms of sex determination might be selectively favored to prevent the degeneration of sex-linked regions [55]. The sex-linked locus shows evidence of substantial copy number heterogeneity both within and among sexes, with copy number in 1kb windows (even within genes) varying between 0 - 39.3X for males and 0 - 13.5X for females. When this fine-scale variability in copy number is associated with sex, it displays a characteristic "suspension bridge" pattern typical of recombination suppressed regions, with divergence concentrated near the breakpoints at the center of the region where selection counters recombination [43,56,57]. However, the peak of *4NeR* in the center of the sex-linked region, along with a 10% rate of phenotype-genotype mismatch, suggests that on-going gene exchange (i.e., "flux" through gene conversion or double recombination events) may be a considerable force countering sex-specific selection [58,59]. The distribution of genotype-phenotype mismatch does not show a particularly strong geographic signal, with mismatched individuals being present in nearly all populations, although home state is a significant predictor. This suggests that either recombination between sex chromosomes is not particularly rare, or that rare recombination is complemented by environmental modifiers of sex. In terms of a mechanism of gene exchange, gene conversion among sex-chromosomes has been observed in a number of species, from fruit flies to humans [60,61], and when combined with high levels of copy number variation as is seen here, may counteract the accumulation of deleterious mutations in recombination suppressed regions [62].

On-going gene flux has likely bolstered the maintenance of variation in sex-linked haplotypes across the native range of *A. tuberculatus.* Similar results have recently been observed for recombining inversions encoding *Heliconius* wing coloration[63], while remarkable diversity among Y haplotypes has been recently observed in guppies and fruit flies [40,64]. We see structure in sex-linked haplotypes shaped by ancestry and latitude. On one hand, this may be consistent with neutral divergence and a history of diverging sex-chromosomes on either side of the Mississippi river before secondary contact of var. *rudis* and var. *tuberculatus [48]*, but may also result from clinal selection from multivariate climatic conditions. Previous analyses have suggested an interaction between sexual and natural selection for flowering time, with males showing stronger divergence from females at lower latitudes [31,32], potentially supporting a role of divergent sex-linked haplotypes in climate adaptation.

Despite the potentially homogenizing force of gene exchange between the X and Y, we observe a narrow peak of association in the middle of the SLR at MADS8 and a relative dearth of associations around it suggests particularly strong sex-specific effects [65–67]. MADS box transcription factors play an exceptional role in developmental novelties, with their duplication and neofunctionalization in plants resulting in sex-specific development, flowers, fruits, and seeds ([68–71]), and in Populus, potentially even the transition from male to female [72]. MADS8 in particular has been characterized as a gene indispensable for female development in Rice (Oryza sativa) under high temperatures [36]. Such an environmental-dependent pathway could explain the lack of complete phenotype-to-genotype mapping we observe for MADS8. This might also explain the same finding for the gene *FloweringTime*, which previous work identified as a candidate for SD gene having been differentially expressed across male and female flowers and showing sex-specific markers [23,33], given that our assemblies show presence/absence variation of this gene within the Y itself. Quantifying differential expression of sex-linked genes in response to varying exogenous cues will help resolve the potential role and nature of environmental modifiers of sex in this system.

These results provide support for the concentration of the genetic architecture of sex during the transition from dioecy to monoecy. While large-effect alleles have accrued on the extended sex-linked region on chromosome 1, it seems likely the evolution of separate sexes is still incomplete, with multiple SD factors and variation in sex-expression still segregating, along with on-going gene exchange between the proto-X and Y. The practical implications of this complexity in genetic architecture of sex are that propositions of using gene drive on the SD locus as a mechanism of weed control may be practically challenging to implement. Further work to decipher the genetic architecture and fitness curve of sex-allocation in close monoecious relatives will help resolve the ecological mechanisms and the variation that has been lost during the evolution of separate sexes in Amaranthus.

Acknowledgements

Thank you to Wouter van der Bijl, Tyler Kent, Sally Otto, and the Rieseberg and Mank Labs for feedback on the paper. Thanks to Hayley McKay for generating the gene expression data used for gene annotation. Thanks to Federico Trucco for observing the leaky sexed waterhemp individual and permission to use the photo here. Julia Kreiner was funded by the University of

513 Chicago and the UBC Bioinformatic Biodiversity Research Centre Fellowship. Cassandra
514 Elphinstone was funded by a Weston Award in Northern Research. Stephen Wright, Loren
515 Rieseberg, and Marco Todesco were funded by NSERC Discovery Grants.
516
517

Methods

Data Production

Based on previous analyses of population structure across populations of *Amaranthus* spanning Ontario and the Midwestern United States [31,73], we selected three populations that spanned predominantly var. *Rudis* (Nune), admixed (Walpole), and predominantly var. t*uberculatus* (Nat) ancestry. Further, we selected an individual from a maternal line in an admixed population where its sibling was previously genotyped as segregating for high EPSPS copy number (~28 copies [73]). For each population, seed from 2 maternal lines were sown into 5 pots each in growth chambers in the Biodiversity Research Centre with day and night temperatures set to (15 degrees C/12hr, 25 degrees C/8hr). After germination, pots were thinned to a single plant, and relocated to the greenhouse in early May where they would grow under lengthening days to maximize tissue production and delay the time to flowering. From each line, we collected tissue from one male numerous times over the summer to maximize tissue for high molecular weight DNA extractions and Hi-C.

Hi-C libraries were prepared using a protocol based on Padmarasu *et al.*, 2019 [74] and Dong and Zhong, 2020 [75], with modifications. In brief, young leaves were collected and flash-frozen in liquid nitrogen. Approximately 0.75-1 g of tissue were used as starting material. Leaves were pulverized using a mortar and pestle, and nuclei were cross-linked in 1X PBS buffer + 1.5% formaldehyde for 15 minutes at room temperature in a rotisserie oven. Cross-linking was halted by adding glycine at a final concentration of 250 mM, and incubating the tissue for 5 minutes at room temperature, with rotation. Tissue was washed once in ice-cold 1x PBS buffer and resuspended in 10 ml of cold Nuclei Isolation Buffer (20 mM HEPES pH 8.0, 250 mM sucrose, 1 mM $MgCl_2$, 5 mM KCl, 40% glycerol v/v). It was then filtered, in successions, through two layers of cheesecloth and one layer of miracloth (MilliporeSigma, Burlington, MA, USA). Nuclei were then washed in Nuclei Isolation Buffer and fractionated using a 95% Percoll (MilliporeSigma) gradient buffer. DNA in purified nuclei was then treated with DpnII restriction enzyme, biotinylated, and proximity-ligated. DNA was extracted from the nuclei using phenol:chloroform:isoamyl alcohol (Invitrogen, Waltham, MA, USA) and quantified using a Broad Range kit on a Qubit fluorometer (Invitrogen). Biotinylated DNA fragments were isolated from 1-4 µg of the resulting DNA, and Illumina libraries were prepared following Todesco *et al.* 2020 [76]. Preliminary experiments performed in sunflower and cannabis showed that, due to the abundance of repetitive sequences in plant genomes, large fractions of read pairs in Hi-C libraries are not informative. Both proximally-ligated fragments in a read pair need to be mapped uniquely to the genome to provide information about long-range interactions; if one of the fragments derives from a repetitive region, and can therefore not be uniquely mapped, the read pair becomes uninformative. To reduce the abundance of repetitive regions in the Hi-C libraries, and therefore increase the proportion of informative reads, we applied an enzymatic repeat depletion treatment [76]. Hi-C libraries were sequenced by Novogene Corporation Inc. (Sacramento, CA, USA).

HMW DNA for PacBio sequencing was extracted using a CTAB-based protocol adapted from Stoffel, K. *et al.*, 2012 [77] in the Rieseberg Lab at UBC. Briefly, 0.8 g of flash frozen young leaves of *Amaranthus tuberculatus* were ground to thin powder using mortar and pestle in liquid

nitrogen. The powder was thoroughly resuspended in Extraction Buffer (100mM Tris-HCl pH 8.0, 1.4M NaCl, 20 mM EDTA pH 8.0, 2% w/v CTAB and 1% v/v b-mercaptoethanol) and incubated 1 hour at 55°C. Eventually the tube was cooled down to RT and extracted with one volume of chloroform:isoamyl alcohol (24:1). The aqueous phase was supplemented with NaCl up to a final concentration of 2.6 M and re-extracted with chloroform:isoamyl alcohol. The new aqueous phase was transferred to two new tubes that were topped up with at least 6 volumes of Precipitation Buffer (50 mM Tris-HCl pH8.0, 10 mM EDTA pH 8.0 and 1% w/v CTAB). Tubes were centrifuged for 30 minutes at RT and the combined pellets were washed with milliQ water. Eventually the pellet was gently and fully resuspended in NaCl 1.5 M and incubated with RNAseA for two hours at 37 degrees. After the RNAse treatment, a third chloroform extraction was followed by two washes with 75% ethanol. Finally the fully dried pellet was allowed to resuspend overnight in 10 mM Tris-HCl pH 8.0 at 4°C. Pipetting was reduced to the minimum along the procedure and if done was exclusively handled with wide-bore tips to avoid damaging DNA. HiFi sequencing was performed at the Yale Genomics Centre. We decided to sequence one individual from Walpole, Ontario, Canada, to high coverage (70X), and the other two individuals to the minimum coverage recommended by HiFiASM for haplotype-phased assembly + HiC to assess assembly quality at varying coverages (25X) [78].

HiFiASM was used to produce haplotype-phased assembly, informed by Hi-C data for each sample. Default settings were used, except for one sample (which happened to be lower coverage; var. *tuberculatus*) which showed asymmetric assembly sizes between haplotypes. As recommended, the setting for heterozygosity was adjusted to multiple values, however with little effect. We found that removing the Hi-C data from the initial HiFiASM assembly for this sample led to a much more symmetric haplotype size, and thus we proceeded with this approach. After the assembly of each haplotype's contigs, contigs were further scaffolded using the Hi-C data with the program YaHS [79], and the resulting scaffolds were then manually curated with juicebox based on the contact map. Briefly, scaffolds were joined together that had overrepresentation of long-distance contacts, and contigs were flipped in orientation that showed an excess of long range compared to short-range contacts within the scaffold. Finally, we performed pairwise mapping of each haplotype to each other, and to one haplotype of each other individual, using the SynMap2 tool in Coge [80] to confirm the consistency of scaffold naming among assemblies (such that Scaffold_1 is syntenic to Scaffold_1, and Scaffold_2 to Scaffold_2, etc., across all haplotypes).

All six haplotype assemblies were then annotated with Maker [81], using *ab-initio* prediction, along with protein sequence from a previous *Amaranthus tuberculatus* assembly [73], its congener *Amaranthus hypochondriacus [82]*, and using transcriptomic data we produced from 24 individuals spanning two populations from southwestern Ontario found in a Natural and Agricultural setting, with 12 males and 12 females sequenced. For these transcriptomes, RNA was extracted using RNAeasy plant kits in the Wright Lab, and libraries were prepared and sequenced at the Sickkids, Toronto sequencing facility on Illumina single-end 100bp technology.

Comparative Genomics

To identify orthologous groupings among assemblies and visualize their synteny, we ran Genespace [39] on all 6 *A. tuberculatus* assemblies along with three other congeners with

recently produced chromosome-level reference genomes: *Amaranthus crutensus* [83], *Amaranthus tricolor [84]*, and *Amaranthus palmeri [85]*. We used the gghits and riparian_plot functions to visualize pairwise and all vs all synteny. We also investigated the number of orthogroups (defined by the presence of X-1 breaks in synteny) between each genome and across each scaffold. To identify core genes and pan genes, we queried pangenes for their presence/absence across assemblies at the genome-wide level, and for particular chromosomes of interest. We refined these comparisons to assemblies we hypothesized to contain the male SDR.

Centromere identification and Repeat Calling

Previous work generated de-novo TE libraries [32] for a previous draft *A. tuberculatus* reference genome [73]. We used repeat masker based on these libraries to identify previously characterized repeats in all 6 of the new haplotype-phased chromosome level assemblies produced here. We also ran RepeatOBserver to identify the putative centromeres in our assembly, as inferred from the Shannon diversity index of Fourier transforms of DNA walks, with low diversity indicating highly repetitive low complexity regions as is typically found in centromeric regions [38].

Population genomics

We leveraged 187 resequenced individuals from 17 pairs of neighboring natural-agricultural populations collected from Michigan to Kansas (from [31]) to understand population genomic patterns of separate sexes. The samples were grown as a part of a much larger common garden experiment, in which 6000 individuals were grown, and ~4000 sexed and phenotyped for numerous traits [31]. Important to this study, plants were sexed at the early to mid-stages of flowering and so based on only a small proportion of flowers. Furthermore, due to the large-scale nature of this study, and the sheer number of flowers produced by an individual, we did not make an effort to look for variation in sex expression (nor did we know it existed). As a result, we could not comment on whether any individuals showed leakiness in their sex expresion. In the meantime, the Tranel lab at University of Illinois Urbana-Champaign was performing a routing grow out of waterhemp for herbicide phenotyping, and noticed the individual pictured (**Fig 3D).** In addition to the pictured female individual that harbors a male flower, the Tranel lab has observed, rarely, other instances of leaky sex expression including: a predominantly female individual with an entire branch of male flowers, and an individual that contained a perfect (hermaphroditic) flower.

We re-aligned the raw reads from this resequencing dataset using BWA-mem [86] to a randomly chosen haplotype of the highest-coverage individual (70X in total, 35X per haplotype; an individual from an admixed population in Ontario segregating for high EPSPS copy number). After calling SNPs using freebayes (options --use-best-n-alleles 2 --report-monomorphic) and filtering SNPs (using the following expressions in BCFtools [87]: QUAL >= 30, 'AB >= 0.25 & AB <= 0.75 | AB <= 0.01' , SAF > 0 & SAR > 0, MQM >=30 & MQMR >= 30', ((PAIRED > 0.05) & (PAIREDR > 0.05) & (PAIREDR / PAIRED < 1.75) & (PAIREDR / PAIRED > 0.25)) | ((PAIRED < 0.05) & (PAIREDR < 0.05))') we performed a genome-wide association study of the 11,040,109 SNPs genome-wide with sex using gemma and a minor allele threshold of 5% [88]. This GWAS identified Scaffold_1 as containing the major sex-linked region. Having applied a Bonferroni multiple test correction, we extract the genotypes of all remaining

significant SNPs and performed a lasso regression analysis on them to determine their co-linearity in explaining sex, after first performing cross-validation to identify the value of the shrinkage penalty (lambda) that minimizes the mean-square error. On this same genome, we also re-mapped two traits, biomass and flowering time, that were measured for these samples in [31].

To further distinguish population-level patterns between haplotypes (i.e. whether it was a ZW or XY system) and to distinguish between reads with the ability to map to both haplotypes, we produced a modified assembly that contained both haplotype 1 and haplotype 2 of Scaffold_1, but only haplotype 2 of all other scaffolds. All raw resequenced reads were then remapped to this modified assembly, but SNPs were only re-called on haplotype 1 and 2 of Scaffold_1. 1,881,197 SNPs were associated sex across both haplotype of Scaffold_1 in the subsequent association study.

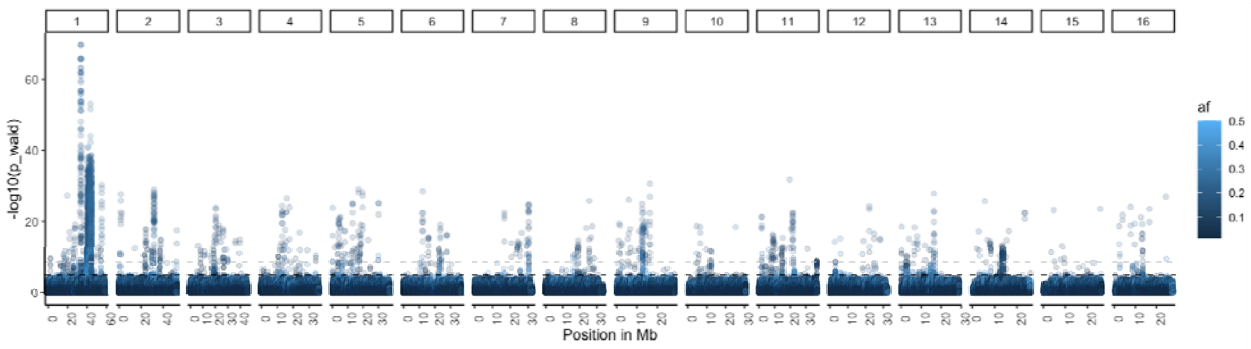
Using competitively mapped SNPs, we then implemented a number of population genetic investigations. PCAs were performed on particular regions of interest, including each of the inversions identified in Genespace and the major sex-linked region on Scaffold_1 using plink2 (--pca). Previous work inferred the ancestry of each sample with FastStructure at K=2 [31], and we used these ancestry proportions that relate to the proportion of var. *rudis* or var. *tuberculatus* ancestry in a multiple regression along with sex-grouping (with female, male, mismatched female, mismatched male as factor levels), the ancestry x sex-grouping interaction, habitat (natural or agricultural), latitude, and longitude of a sample to predict PC values of each inversion. To compare the proportion of heterozygous genotypes between males and females for each locus, we used the --geno-count option in plink for separate runs that included either only males or only females. For each sex, we then calculated the heterozygous proportion as the number of heterozygous genotypes divided by the total number of genotypes called at that locus (i.e. accounting for missing data at each locus). We also inferred the phylogenetic tree of the sex-linked region using IQtree2 [89], with the algorithm that accounts for invariant sites (options -st DNA -nt 80 -m TESTONLY+ASC). Finally, we used $Pi_{XY[90]}$ to calculate weir and cockerham's F_{ST} and diversity between and within sexes, while controlling for missing data. For both F_{ST} and the difference in the proportion of heterozygous genotypes between males and females, we used R to calculate the 95% confidence interval of all autosomal scaffolds using quantile(probs = c(0.025,0.975) to provide context for their distributions along the sex-linked region on Scaffold_1. All plotting was done with ggplot2 [91] and the cowplot [92] packages in R.

We calculated fine-scale copy number in 100bp windows along the sex-linked region. Specifically, we used mosdepth [93] to estimate coverage in 100bp windows for each individual directly from mapped reads. We calculated copy number as the depth in a focal 100bp window along Scaffold_1 (the SLR contained scaffold) divided by the mean coverage outside of the SLR on Scaffold_1. To visualize fine-scale copy number by sample in the sex-linked region and reduce noise in these estimates, we plotted the smoothed means of copy number in 100 kb windows. Mean individual copy number was also calculated across the entire sex-linked region. Using this estimate of an individual's mean copy number in the SLR, we performed multiple regression analyses with var. *rudis* ancestry proportion, sex grouping (defined based on male and female, and whether or not they were genotype-phenotype matched [n=4 groups]), latitude, longitude, and habitat (natural or agricultural) as predictors. We also performed a copy number-phenotype association analysis for each 100bp window on Scaffold_1. This was done using a

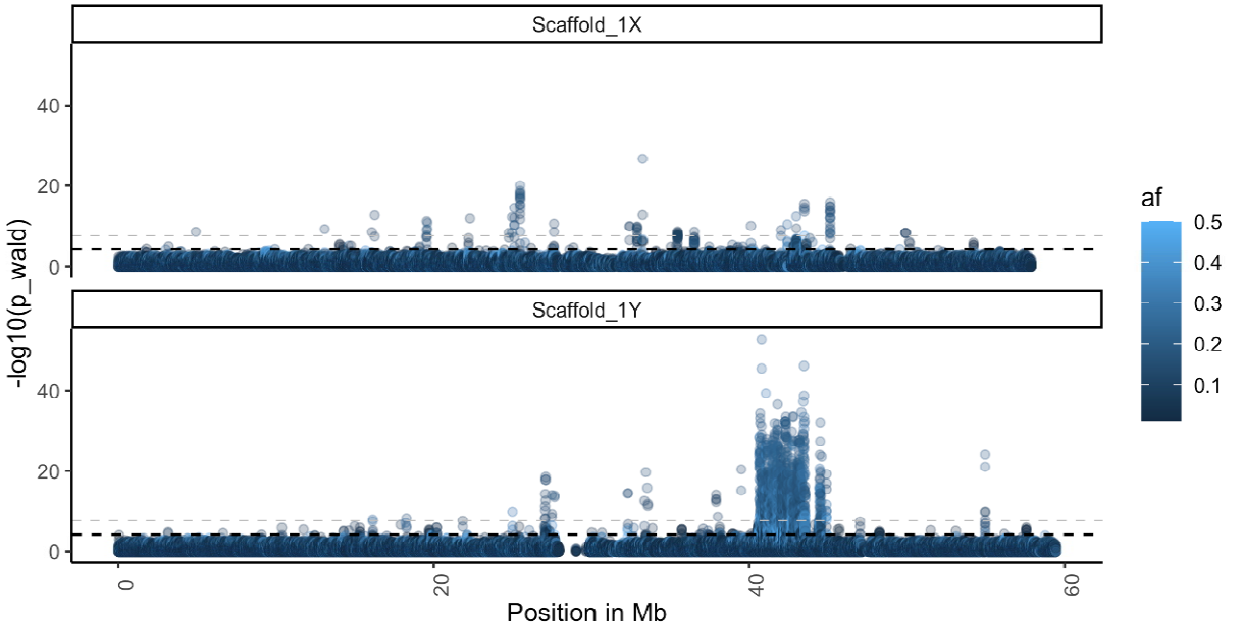
custom script, for which a linear regression of sex against copy number was run for each window across Scaffold_1. The -log10(p-value) of these tests were plotted, aftering inferring the raw p-value cutoff using a Bonferroni multiple test correction threshold of 0.05.

To validate the likelihood of sex mis-phenotyping, we investigated patterns of genotype-phenotype mismatch in an independent dataset, inferred local recombination rate, and the distribution of genotype-phenotype mismatch across the range. We mapped resequenced herbarium samples [94] for which the phenotype value for sex was verified and validated based on deposited images, to our focal reference genome. Mapping was performed after merging and de-duplicated reads with DeDup [95], and rescaling base-quality scores to account for DNA damage [96]. As for our contemporary samples, we then inferred copy number based on scaled depth in the sex-linked region. On contemporary samples, we inferred the effective population recombination rate (rho) using LDhat [45]. To do so, we used a precomputed look up table (theta = 0.01) and a subsampled VCF with 100 genotypes (25 males and 25 females). Rho was computed between each SNP, each point estimate was plotted along with the loess curve along Scaffold_1. Since we also had phenotype data for these samples, we were able to perform a multiple linear regression analysis for two key traits (biomass and flowering time), to ask whether genotype-phenotype match status predicted trait values, while controlling for experimental treatment, environment, geography, and sex of samples. Finally, we calculated the difference in an individual's copy number relative to the respective sex-matched mean value. We tested whether there was a strong geographic signal in the distribution in the degree of this mismatch in a multiple regression framework, with latitude, longitude, environment, and state as predictors.

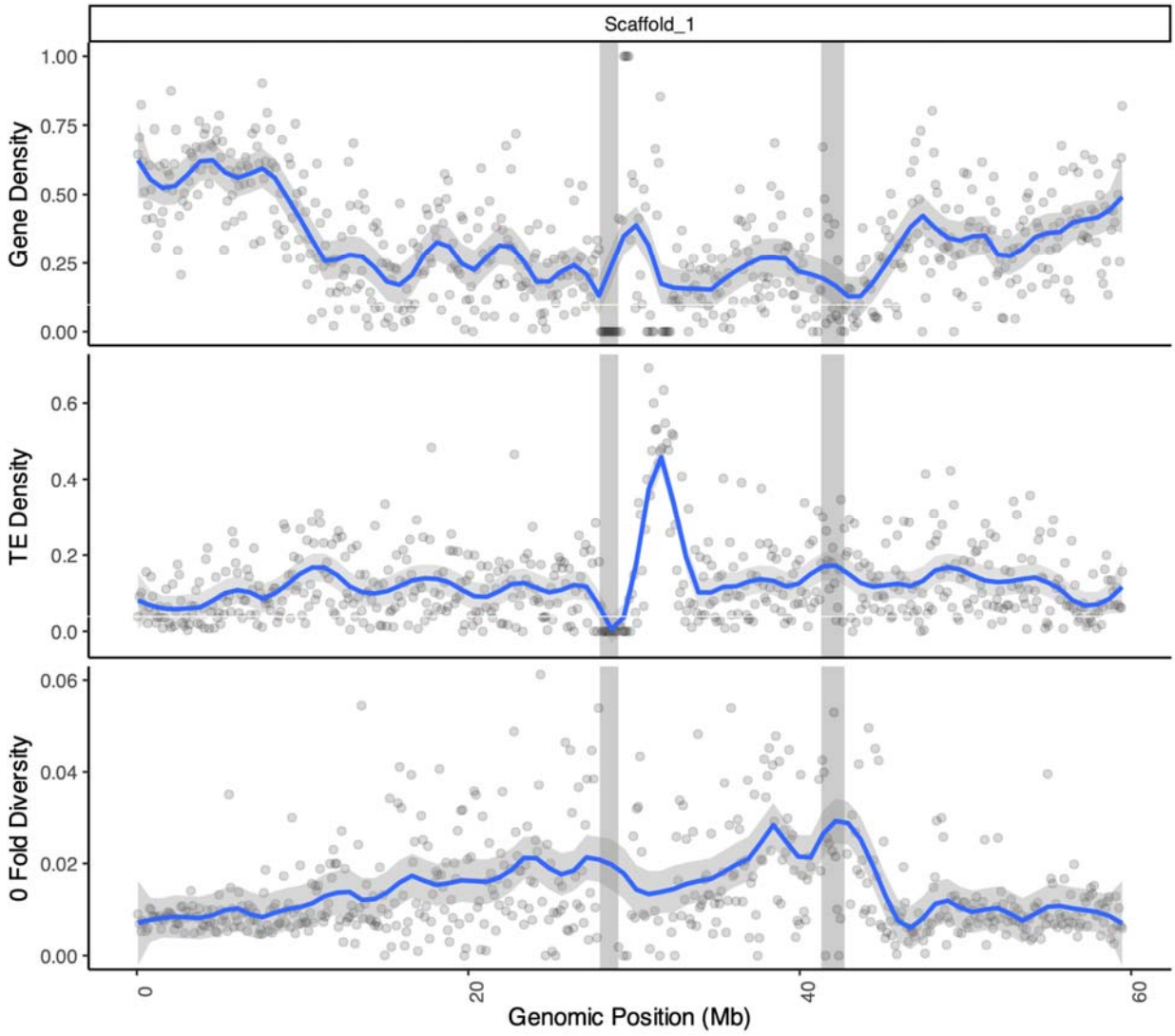
Supplementary Figures



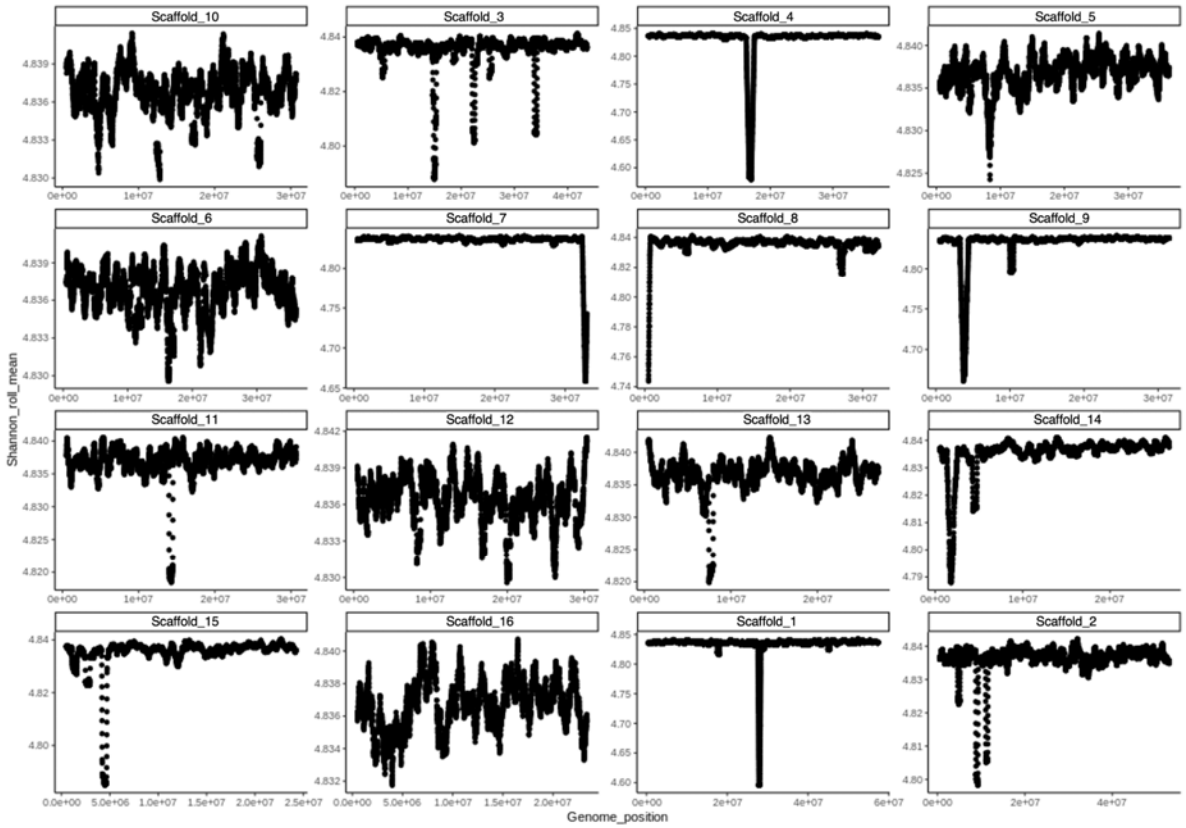
Sup Figure 1. A Manhattan plot of the strength of the -log10(p-value) of the association of SNPs across the genome with sex phenotypes. Analysis was done on SNPs called from reads mapped only to haplotype_2 (i.e., not competitively mapped).



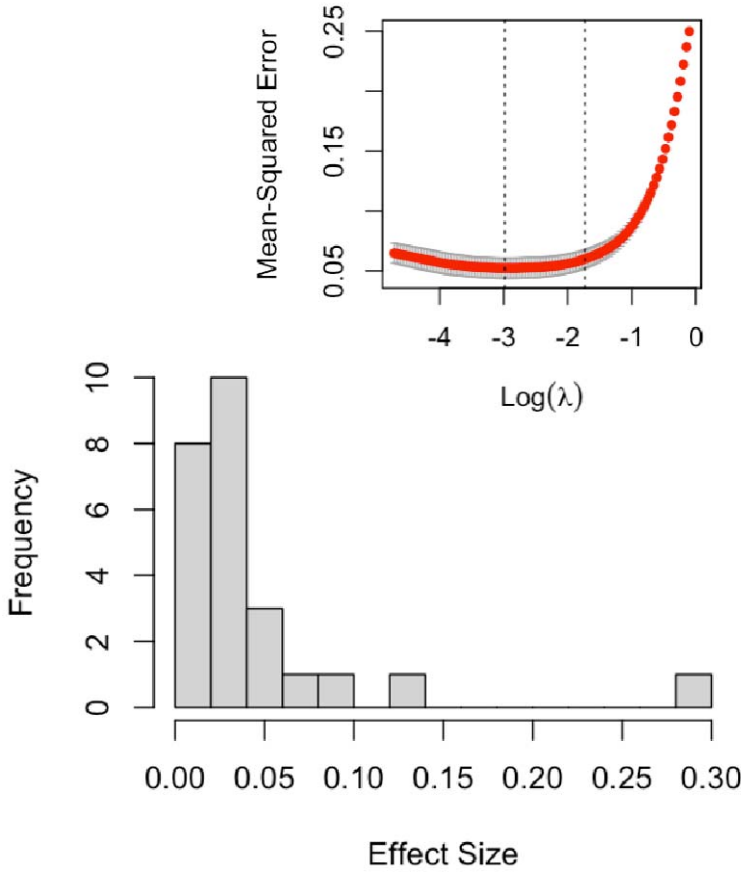
Sup Figure 2. A Manhattan plot of the strength of the -log10(p-value) of the association of SNPs along haplotype_1 (the proto-X) and haplotype_2 (the proto-Y) of Scaffold 1. Analysis was done on SNPs called from reads competitively mapped to haplotype 1 and 2.



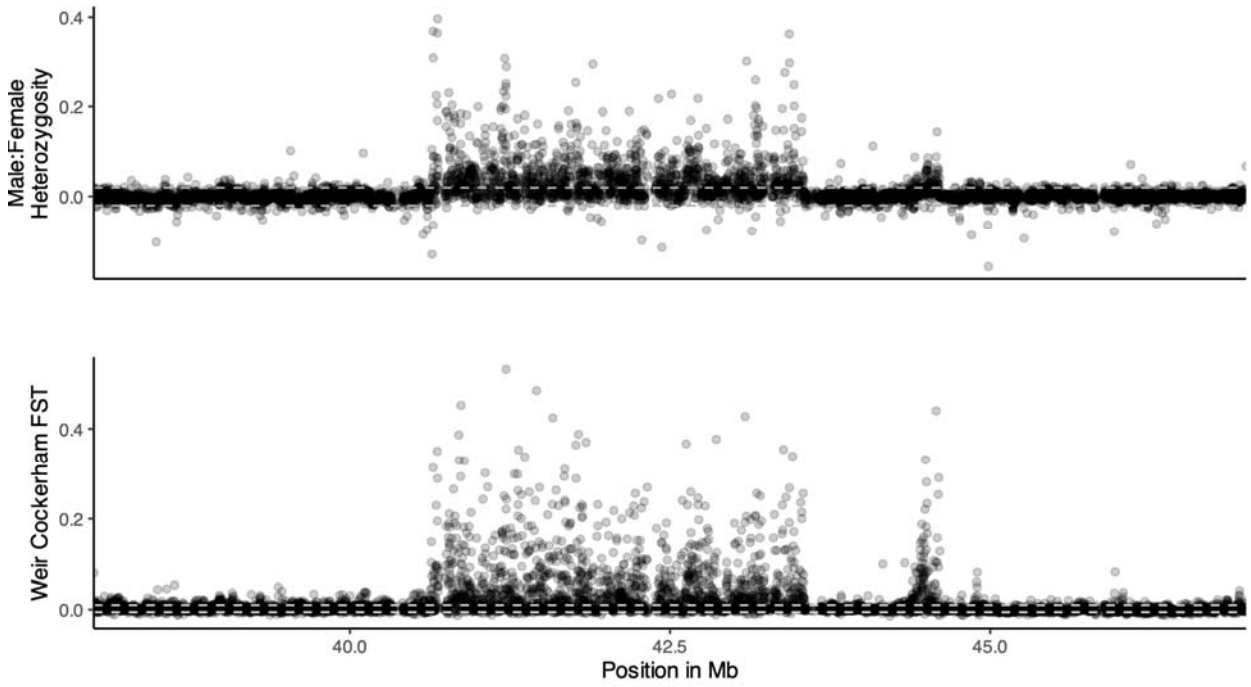
Sup Figure 3. Structure across chromosome 1, containing the sex determining region between 40-43.5 Mb (right vertical bar), a region which shows the highest 0 fold diversity and among the lowest gene density across chromosome 1. Analysis was done by mapping all reads to only haplotype 2 (i.e. no competitive mapping). TE density (based on a previously curated TE library) is greatly enriched at ~32 Mb, just neighboring the inferred location of the centromeric region (left vertical bar; as inferred from RepeatOBserver).



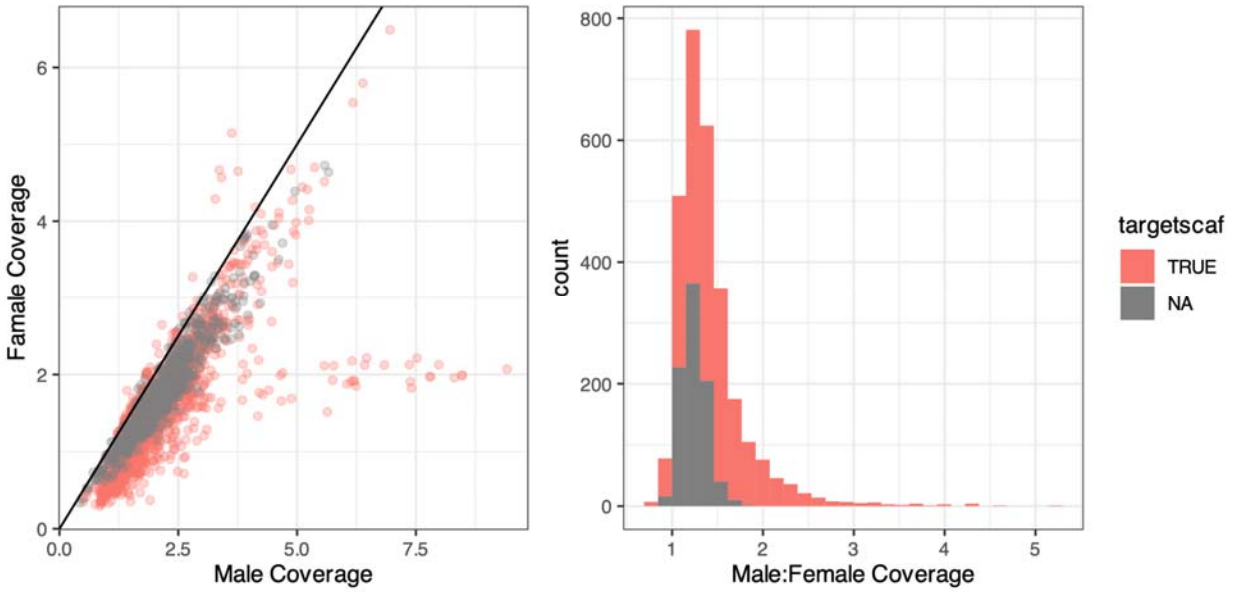
Sup Figure 4. RepeatOBserver inferred shannon-diversity index of sequence complexity across each scaffold for our focal reference, where minimum values are predictive of centromere location. The sex-linked region region is present on Scaffold_1 around ~41Mb, and shows a tertiary minimization of Shannon diversity.



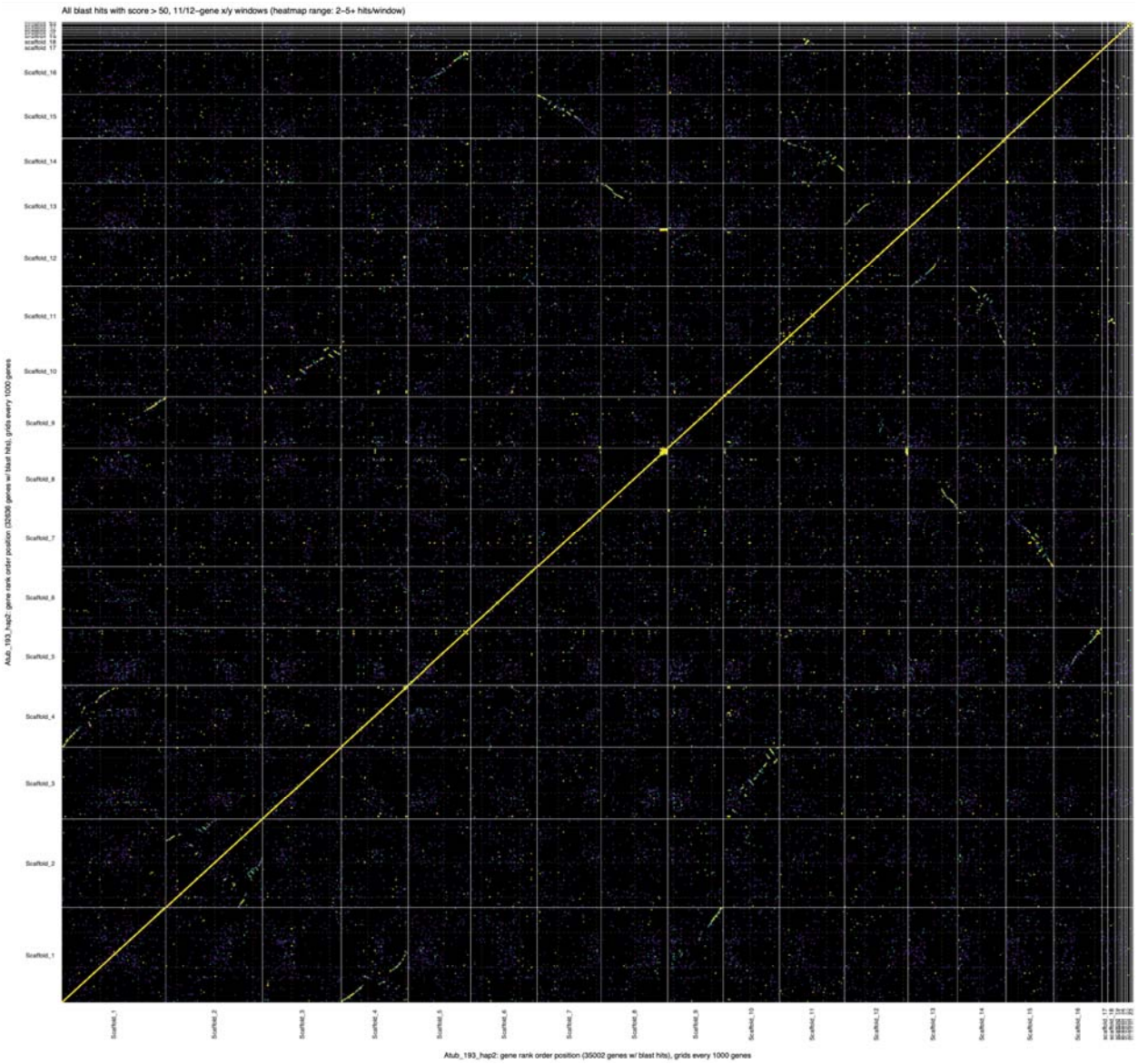
Sup Figure 5. Results for the lasso regression analyses of the predictive independence of loci with significant associations with sex, after Bonferroni correction. Upper) The minimization of MSE occurs when lambda approaches zero (~= 0.01), such that no one predictor is eliminated from the model, their coefficients are just shrunken towards 0. B) The distribution of effect sizes for the 26 loci with remaining non-zero effects on sex.



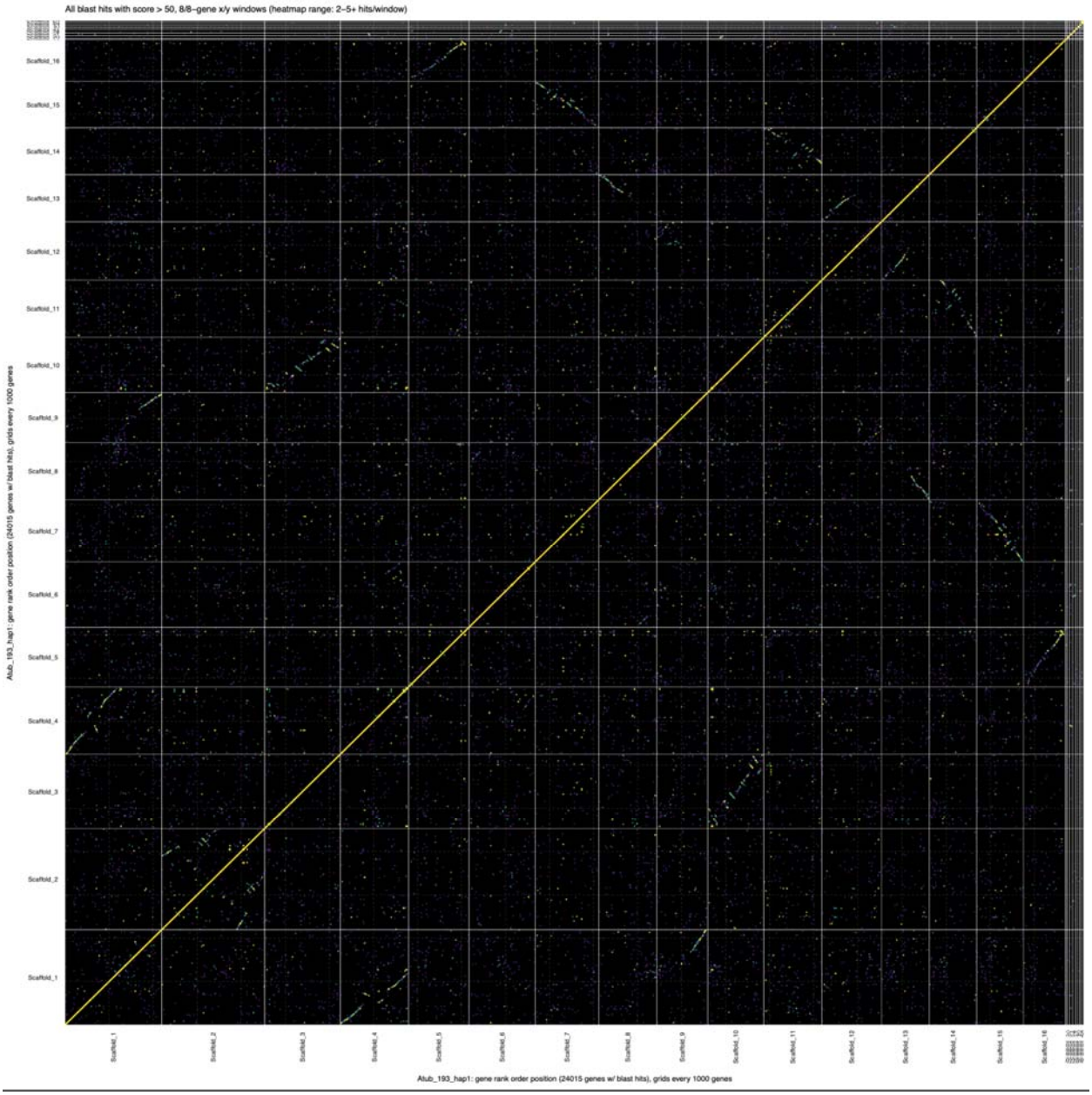
Sup Figure 6. Upper) The difference in heterozygosity between males and females is heightened in the sex-linked region, along with allelic differentiation (Fst: Middle Lower) between males and females, where points represent the mean in 1 kb genomic windows. In both plots, the 95% CI of these summary statistics across autosomes is delimited by the horizontal dashed white lines.



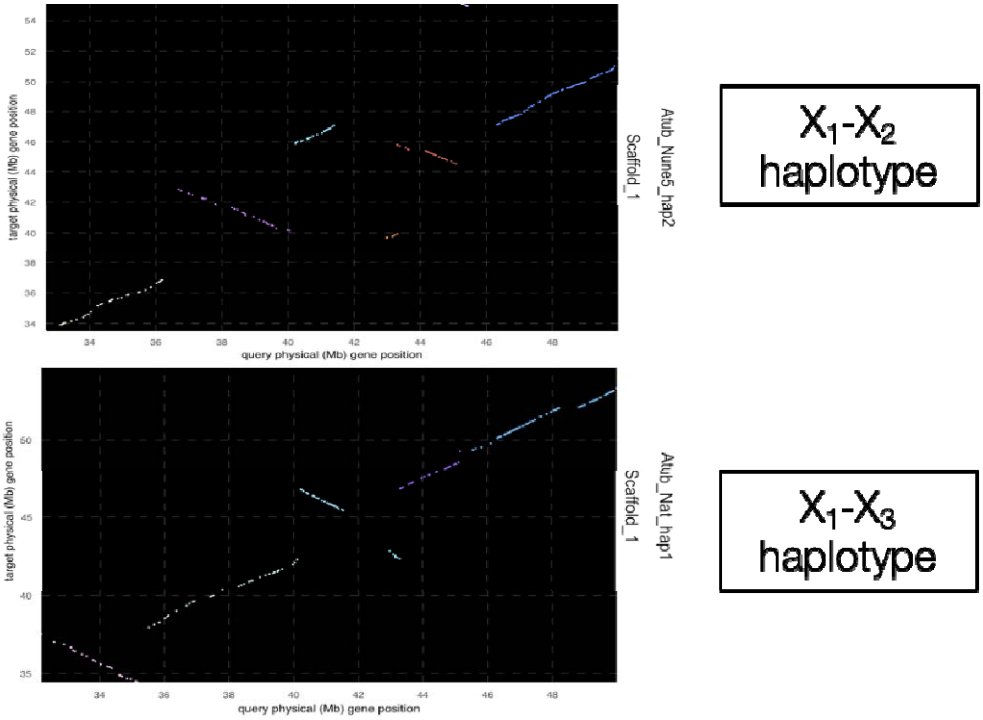
Sup Figure 7. Male versus Female coverage for loci significantly associated with sex (FDR < 0.05). Color indicates whether loci are located on the sex-linked, "target" scaffold_1, or on other scaffolds across the genome.



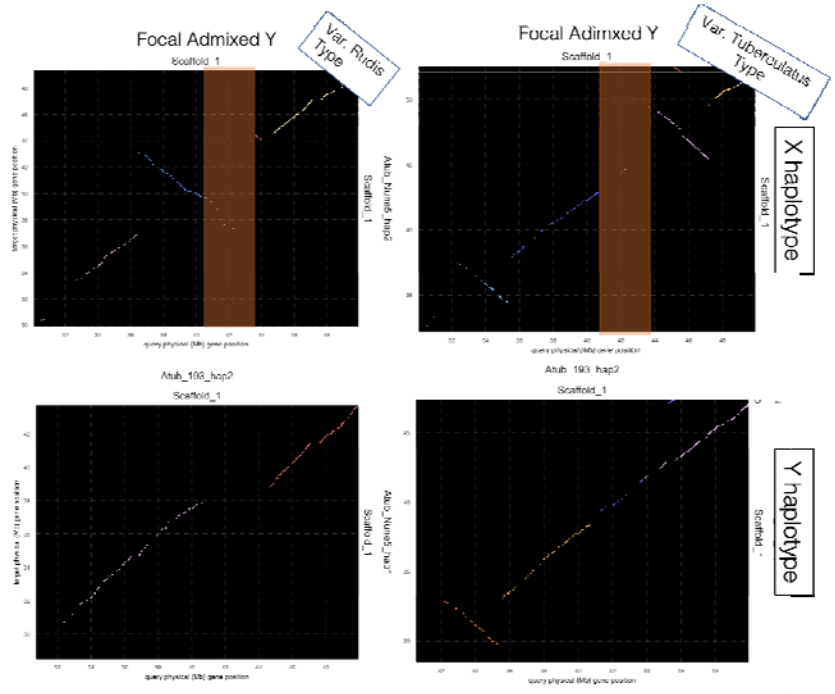
Sup Figure 8. Dotplot of synteny between haplotype 2 and itself (our focal assembly for which population genomic was aligned to, the individual from Walpole]), where dots are gene blast hits with score > 50 inferred from the genespace pipeline. Evidence of a past whole-genome duplication event is evident, as every scaffold shows contiguous stretches of synteny with one or more scaffolds.



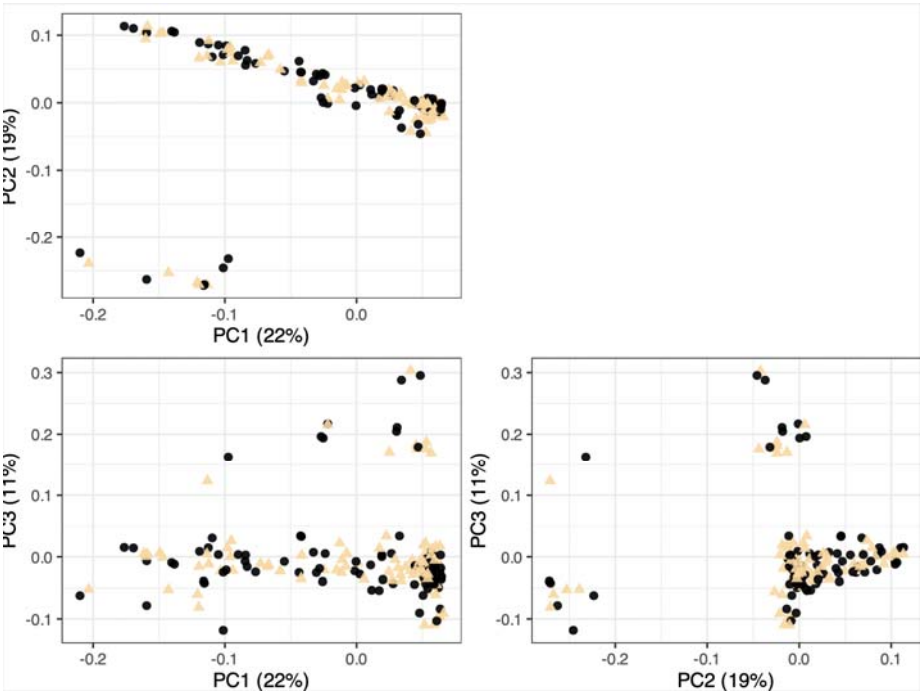
Sup Figure 9. Dotplot of synteny between haplotype 1 against itself [from the Walpole male], where dots are gene blast hits with score > 50 inferred from the genespace pipeline. Evidence of a past whole-genome duplication event is evident, as every scaffold shows contiguous stretches of synteny with one or more scaffolds.



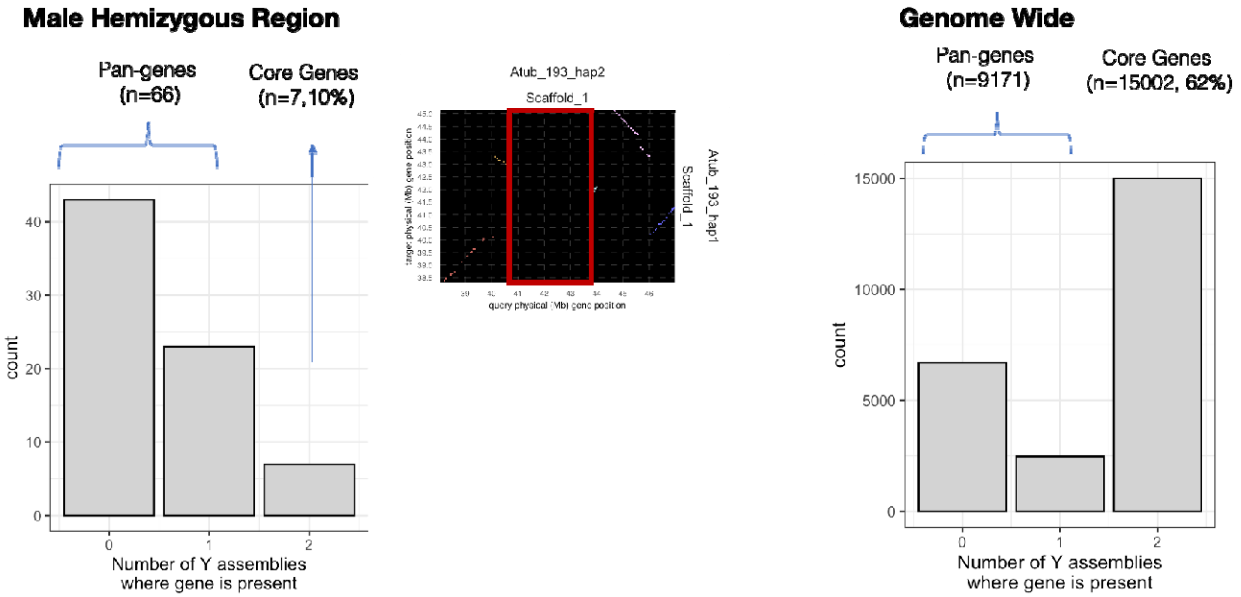
Sup Figure 10. Dotplots of syntenic genes between X haplotype assemblies of the sex-linked region. Colors represent orthologous grouping (i.e. tracts of contiguous syntenic sequence).



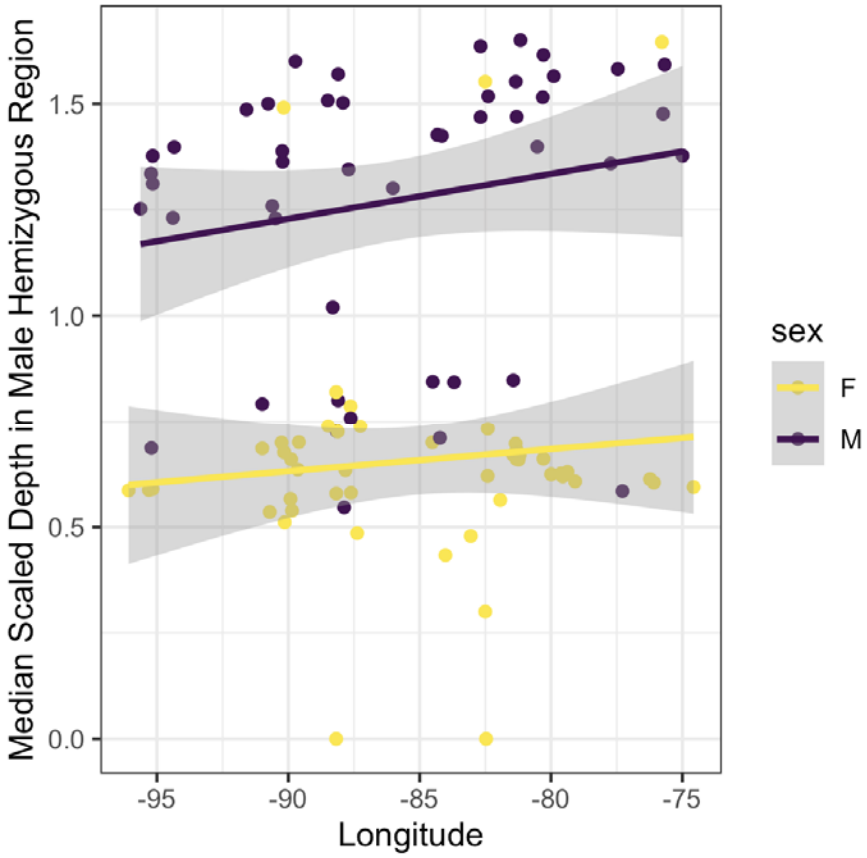
Sup Figure 11. Pairwise comparisons of synteny for our focal-Y (admixed type) and the X and Y haplotypes of the var. *rudis* type individual (left) var. *tuberculatus* type (right).



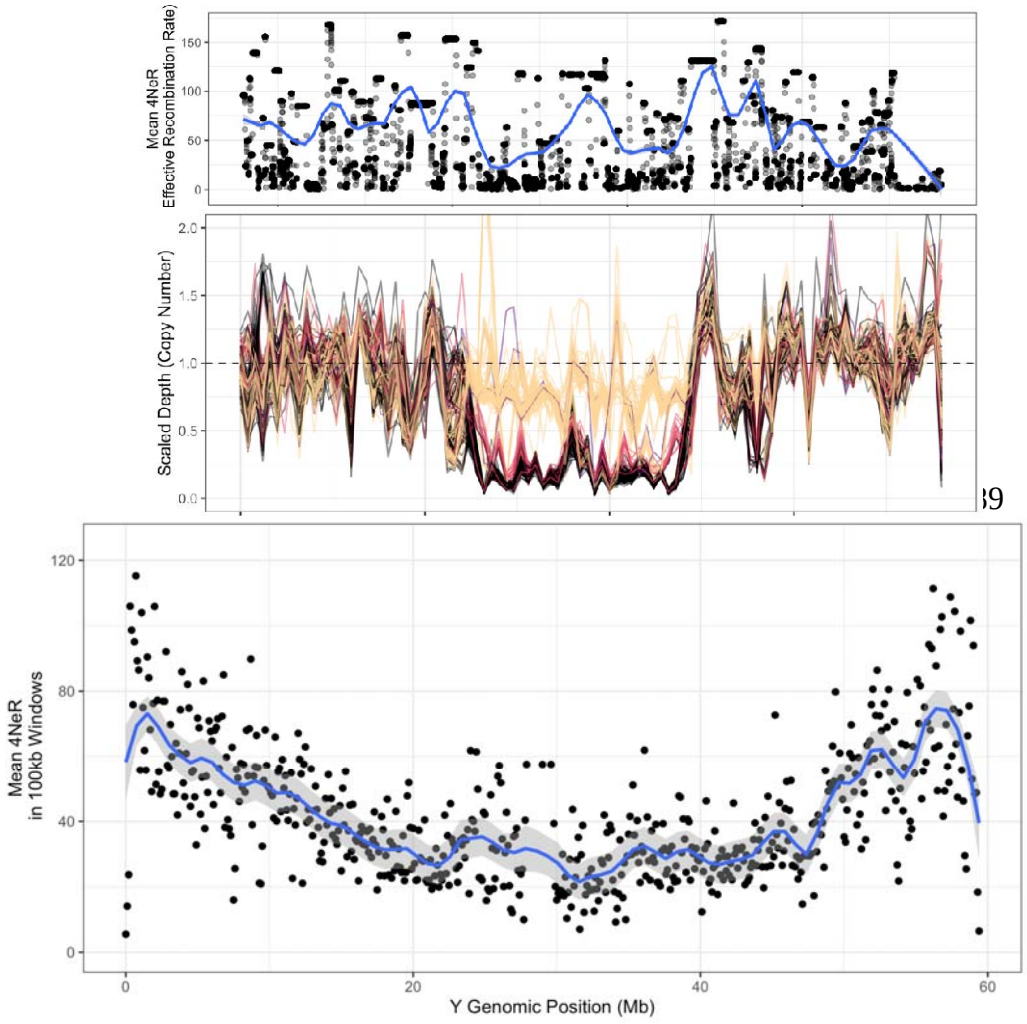
Sup Figure 12. Structure in the inversion (Sup Table 1 number 5* and 14*) just upstream of the sex-linked region. Points (individual genomes) coloured by phenotypic sex assignment (cream = male, black = female), illustrating the lack of sex-based structure in this inversion.



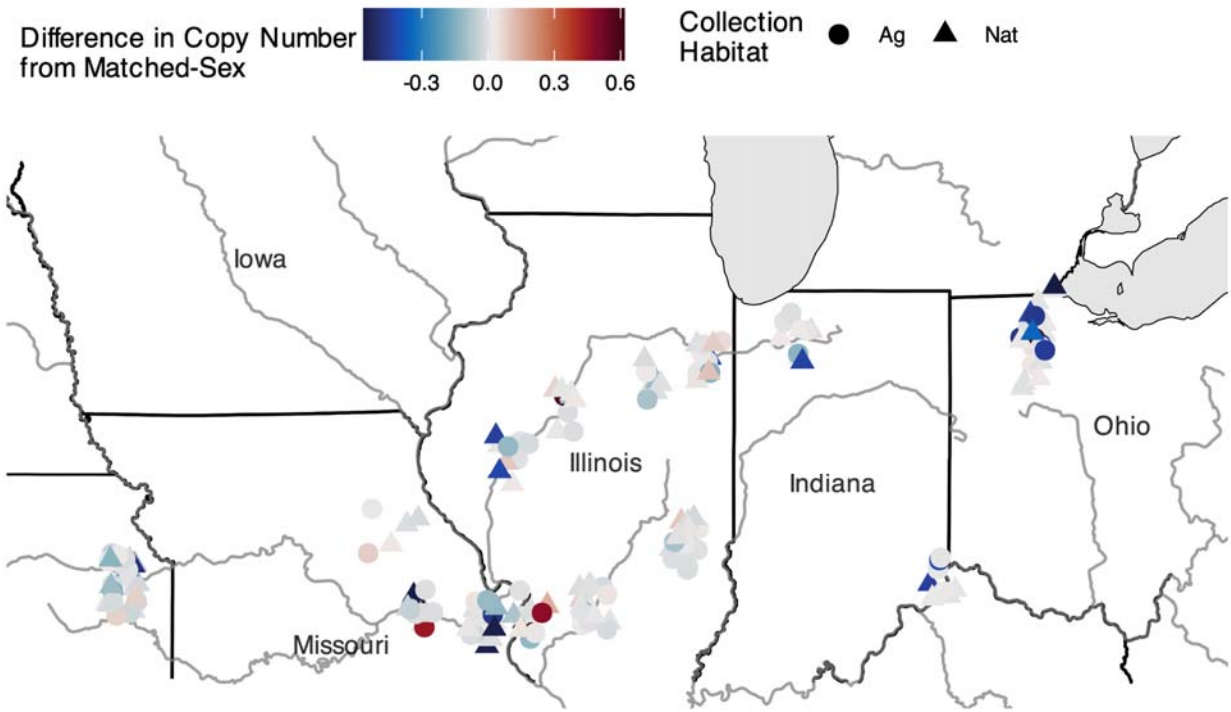
Sup Figure 13. Enriched gene presence-absence variation in the sex-linked region, compared to genome wide.



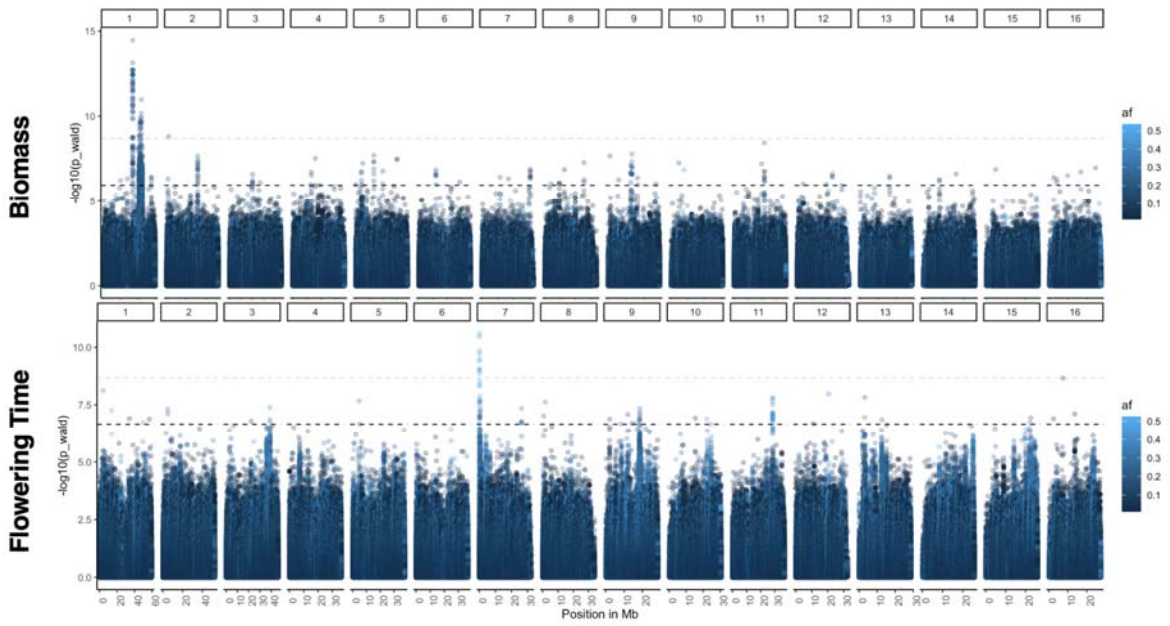
Sup Figure 14. Median scaled depth in sex linked region for an independent dataset [94], illustrating genotype-phenotype mismatch. 10 males show depth profiles more similar to females than expected, and 3 females show depth profiles more similar to females.



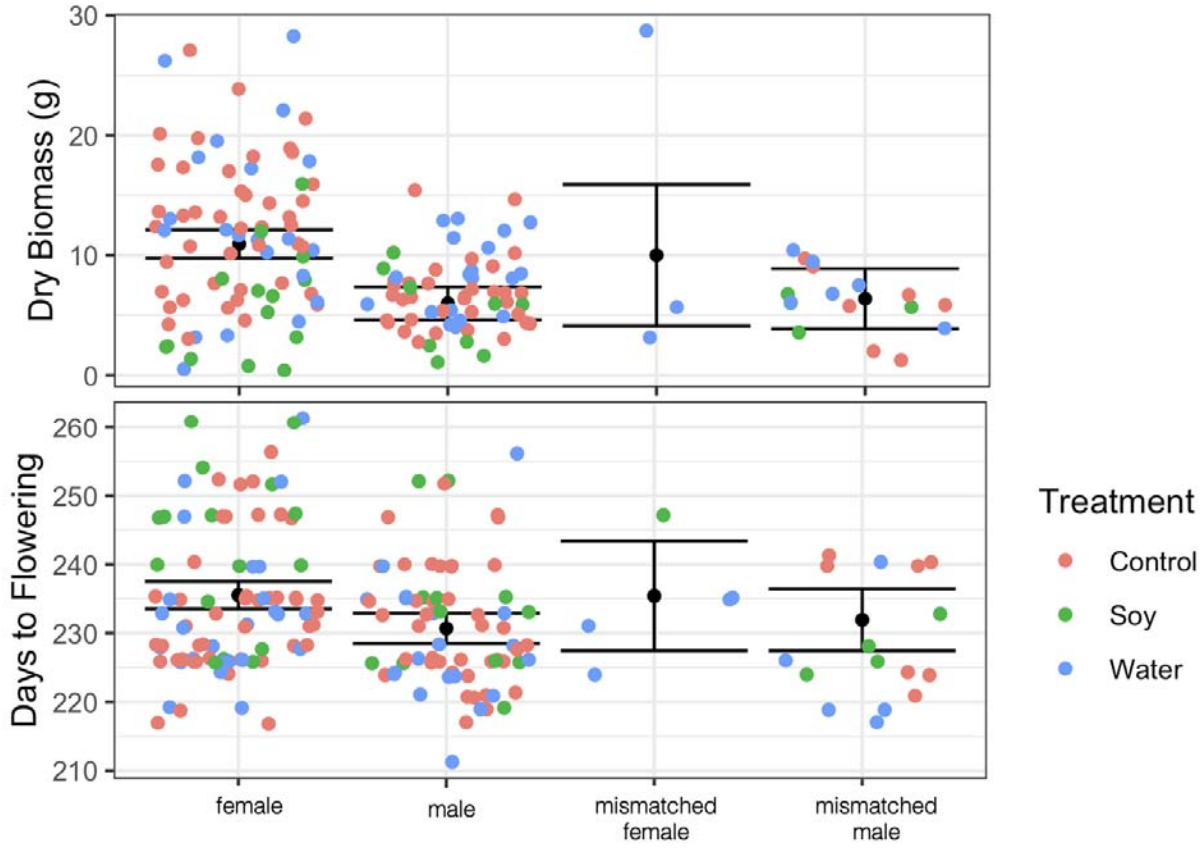
Sup Figure 15. LDhat inferred effective recombination rate. *Top*) Effective recombination rate along the sex linked region on haplotype 2 (the Y containing haplotype) of Scaffold 1, along with copy number variation for context (*Middle),* and the effective recombination rate landscape across the entire haplotype 2 of Scaffold 1 (*Bottom).*



Sup Figure 16. The geographic distribution of the difference in copy number between an individual and mean copy number of their respective sex-matched genotype. Color coding represents the degree of mismatch based on an individual's mean copy number across the region, and the mean copy number of the average sex-matched individual. Negative numbers indicate males who genotypically resemble most females, and who therefore have less sequence content in the SLR than most males, whereas positive values indicate females who genotypically resemble most males (have an excess of sequence content in the SLR).



Sup Figure 17. Manhattan plots of GWA of family-mean biomass (top) and family-mean flowering time, based on mapping of reads to haplotype 2 of the Walpole admixed reference. The lower black horizontal dashed line represents the 5% FDR p-value threshold, while the upper dashed horizontal line represents the Bonferroni corrected 5% p-value threshold.



Sup Figure 18. The distribution of biomass and days to flowering for individuals with different phenotype-genotype groupings for sex. We see no significant difference in phenotypes between mismatched individuals and their respective matched sex.

Supplementary Tables

Sup Table 1. Statistics for 6 haplotype-phased chromosome-level assemblies, from 3 male *A. tuberculatus*. *indicates that the initial HiFiasm assembly step was performed without Hi-C data.

	Walpole_hap1	**Walpole_hap2**	**Nune_hap1**	**Nune5_hap2**	**Nat_hap1***	**Nat_hap2***
Main genome scaffold total	1984	1060	1570	777	1652	574
Main genome contig total	2464	1497	2167	1311	2153	1001
Main genome scaffold seq total	678.737 MB	654.704 MB	648.107 MB	620.003 MB	646.622 MB	647.614 MB
Main genome contig seq total	678.641 MB	654.616 MB	647.988 MB	619.896 MB	646.522 MB	647.528 MB
Gap	0.014%	0.013%	0.018%	0.017%	0.015%	0.013%
Main genome scaffold N/L50	9/32 MB	8/34.048 MB	8/33.576 MB	8/33.357 MB	8/33.422 MB	8/33.162 MB
Main genome contig N/L50	48/2.327 MB	41/2.803 MB	51/2.194 MB	48/2.368 MB	53/2.428 MB	45/3.333 MB
Main genome scaffold N/L90	122/136.416 KB	25/910.168 KB	42/371.629 KB	16/24.153 MB	47/312.591 KB	17/3.405 MB
Main genome contig N/L90	498/93.793 KB	339/247.302 KB	467/152.867 KB	348/275.056 KB	421/155.35 KB	277/375.396 KB
Max scaffold length	57.876 MB	59.453 MB	54.404 MB	59.099 MB	61.224 MB	56.815 MB
Max contig length	19.248 MB	20.689 MB	16.635 MB	17.086 MB	19.341 MB	21.347 MB
Number of scaffolds > 50 KB	426	353	317	287	301	280
% main genome in scaffolds > 50 KB	93.24%	96.46%	94.88%	97.78%	93.96%	98.63%

Sup Table 2. List of Inversions relative to Atub_Walpole_hap2. Note that letter code demarcates the same inversion present across multiple pairwise comparisons (each referenced by the inversion number and alt_genome). Inversion C (5 and 14) represent the variant directly neighboring the SLR in Atub_Walpole_hap2.

Inversion Number	**Ref_Scaf**	**Start**	**End**	**Alt_Genome**
1	Scaffold_1	25970434	30554490	Atub_Walpole_hap1
2A	Scaffold_1	7153812	7609088	Atub_Walpole_hap1
3	Scaffold_1	40126975	40143010	Atub_Walpole_hap1
4B	Scaffold_1	11910786	11950873	Atub_Walpole_hap1
5C	Scaffold_1	44656825	46017919	Atub_Walpole_hap1
6A	Scaffold_1	7153812	7609088	Atub_Nune_hap2
7	Scaffold_1	36429003	43364478	Atub_Nune_hap2
8	Scaffold_1	43375038	44017851	Atub_Nune_hap2
9A	Scaffold_1	7153812	7609088	Atub_Nune_hap1
10A	Scaffold_1	7153812	7609088	Atub_Nat_hap1
11B	Scaffold_1	11910786	12359692	Atub_Nat_hap1
12D	Scaffold_1	20165508	21398392	Atub_Nat_hap1
13	Scaffold_1	32372322	35290168	Atub_Nat_hap1
14C	Scaffold_1	43649837	47205224	Atub_Nat_hap1
15A	Scaffold_1	7153812	7609088	Atub_Nat_hap2
16D	Scaffold_1	20165508	21398392	Atub_Nat_hap2
17	Scaffold_1	21953637	23506917	Atub_Nat_hap2
18	Scaffold_1	32218868	35290168	Atub_Nat_hap2

975 **References**

976 1. Dorken ME, Barrett SCH. Sex determination and the evolution of dioecy from monoecy in Sagittaria
977 latifolia (Alismataceae). Proc Biol Sci. 2004;271: 213–219.

978 2. Lloyd DG. The transmission of genes via pollen and ovules in gynodioecious angiosperms. Theor
979 Popul Biol. 1976;9: 299–316.

980 3. Charlesworth D. Evolution of recombination rates between sex chromosomes. Philos Trans R Soc
981 Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0456

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

984 5. Ohyama K, Takemura M, Oda K, Fukuzawa H, Kohchi T, Nakayama S, et al. Gene content,
985 organization and molecular evolution of plant organellar genomes and sex chromosomes — Insights
986 from the case of the liverwort Marchantia polymorpha. Proc Jpn Acad Ser B Phys Biol Sci. 2009;85:
987 108–124.

988 6. Akagi T, Fujita N, Masuda K, Shirasawa K, Nagaki K, Horiuchi A, et al. Rapid and dynamic
989 evolution of a giant Y chromosome in Silene latifolia. bioRxiv. 2023. p. 2023.09.21.558759.
990 doi:10.1101/2023.09.21.558759

991 7. Filatov DA. Heterochiasmy and Sex Chromosome Evolution in Silene. Genes . 2023;14.
992 doi:10.3390/genes14030543

993 8. Hough J, Hollister JD, Wang W, Barrett SCH, Wright SI. Genetic degeneration of old and young Y
994 chromosomes in the flowering plant Rumex hastatulus. Proc Natl Acad Sci U S A. 2014;111: 7713–
995 7718.

996 9. Sacchi B, Humphries Z, Kružlicová J, Bodláková M, Pyne C, Choudhury BI, et al. Phased Assembly
997 of Neo-Sex Chromosomes Reveals Extensive Y Degeneration and Rapid Genome Evolution in
998 Rumex hastatulus. Mol Biol Evol. 2024;41. doi:10.1093/molbev/msae074

999 10. Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, Ayyampalayam S, et al. The asparagus
1000 genome sheds light on the origin and evolution of a young Y chromosome. Nat Commun. 2017;8:
1001 1279.

1002 11. Liu Z, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu Q, et al. A primitive Y chromosome in
1003 papaya marks incipient sex chromosome evolution. Nature. 2004;427: 348–352.

1004 12. Charlesworth D, Charlesworth B. Population genetics of partial male-sterility and the evolution of
1005 monoecy and dioecy. Heredity . 1978;41: 137–153.

1006 13. Lloyd DG. THE DISTRIBUTIONS OF GENDER IN FOUR ANGIOSPERM SPECIES
1007 ILLUSTRATING TWO EVOLUTIONARY PATHWAYS TO DIOECY. Evolution. 1980;34: 123–
1008 134.

1009 14. Lesaffre T, Pannell Charles Mullon J. A model for the gradual evolution of dioecy and
1010 heterogametic sex determination. doi:10.1101/2023.03.24.534076

1011 15. Grosse-Veldmann B, Weigend M. The geometry of gender: hyper-diversification of sexual systems

in Urtica L. (Urticaceae). Cladistics. 2018;34: 131–150.

16. Cronk Q. The distribution of sexual function in the flowering plant: from monoecy to dioecy. Philos Trans R Soc Lond B Biol Sci. 2022;377: 20210486.

17. Delph LF. Sexual dimorphism in gender plasticity and its consequences for breeding system evolution. Evol Dev. 2003;5: 34–39.

18. Ehlers BK, Bataillon T. “Inconstant males” and the maintenance of labile sex expression in subdioecious plants. New Phytol. 2007;174: 194–211.

19. Russell JRW, Pannell JR. Sex determination in dioecious Mercurialis annua and its close diploid and polyploid relatives. Heredity . 2015;114: 262–271.

20. Mosyakin SL, Robertson KR. Amaranthus. Flora of North America Editorial Committee, editor. New York: Oxford University Press; 2003.

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

22. Waselkov KE, Boleda AS, Olsen KM. A Phylogeny of the Genus Amaranthus (Amaranthaceae) Based on Several Low-Copy Nuclear Loci and Chloroplast Regions. Syst Bot. 2018;43: 439–458.

23. Raiyemo DA, Bobadilla LK, Tranel PJ. Genomic profiling of dioecious Amaranthus species provides novel insights into species relatedness and sex genes. BMC Biol. 2023;21: 37.

24. Murray MJ. The Genetics of Sex Determination in the Family Amaranthaceae. Genetics. 1940;25: 409–431.

25. Grant WF. Cytogenetic studies in Amaranthus: I. Cytological aspects of sex determination in dioecious species. Can J Bot. 1959;37: 413–417.

26. Sauer J. RECENT MIGRATION AND EVOLUTION OF THE DIOECIOUS AMARANTHS. Evolution. 1957;11: 11–31.

27. Trucco F, Jeschke MR, Rayburn AL, Tranel PJ. Amaranthus hybridus can be pollinated frequently by A. tuberculatus under field conditions. Heredity . 2005;94: 64–70.

28. Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S, et al. Distinct genetic architectures for male and female inflorescence traits of maize. PLoS Genet. 2011;7: e1002383.

29. Neves CJ, Matzrafi M, Thiele M, Lorant A, Mesgaran MB, Stetter MG. Male Linked Genomic Region Determines Sex in Dioecious Amaranthus palmeri. Journal of Heredity. 2020. doi:10.1093/jhered/esaa047

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

31. Kreiner JM, Caballero A, Wright SI, Stinchcombe JR. Selective ancestral sorting and de novo evolution in the agricultural invasion of Amaranthus tuberculatus. Evolution. 2021. doi:10.1111/evo.14404

32. Kreiner JM, Hnatovska S, Stinchcombe JR, Wright SI. Quantifying the role of genome size and repeat content in adaptive variation and the architecture of flowering time in Amaranthus tuberculatus. PLoS Genet. 2023;19: e1010865.

33. Raiyemo DA, Cutti L, Patterson EL, Llaca V, Fengler K, Montgomery JS, et al. A phased chromosome-level genome assembly provides insights into the evolution of sex chromosomes in Amaranthus tuberculatus. bioRxiv. 2024. p. 2024.05.30.596720. doi:10.1101/2024.05.30.596720

34. Nowak K, Morończyk J, Grzyb M, Szczygieł-Sommer A, Gaj MD. miR172 Regulates WUS during Somatic Embryogenesis in Arabidopsis via AP2. Cells. 2022;11. doi:10.3390/cells11040718

35. Lai Y-S, Zhang X, Zhang W, Shen D, Wang H, Xia Y, et al. The association of changes in DNA methylation with temperature-dependent sex determination in cucumber. J Exp Bot. 2017;68: 2899–2912.

36. Shen C, Zhang Y, Li G, Shi J, Wang D, Zhu W, et al. MADS8 is indispensable for female reproductive development at high ambient temperatures in cereal crops. Plant Cell. 2023;36: 65–84.

37. Kojima S, Takahashi Y, Kobayashi Y, Monna L, Sasaki T, Araki T, et al. Hd3a, a rice ortholog of the Arabidopsis FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions. Plant Cell Physiol. 2002;43: 1096–1105.

38. Elphinstone C, Elphinstone R, Todesco M, Rieseberg L. RepeatOBserver: tandem repeat visualization and centromere detection. bioRxiv. 2023. p. 2023.12.30.573697. doi:10.1101/2023.12.30.573697

39. Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Elife. 2022;11. doi:10.7554/eLife.78526

40. van der Bijl W, Shu J, Goberdhan VS, Sherin LM, Cortázar-Chinarro M, Corral□López A, et al. Deep learning reveals the role of copy number variation in the genetic architecture of a highly polymorphic sexual trait. bioRxiv. 2024. doi:10.1101/2023.09.29.560175

41. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. BMC Biol. 2017;15: 74.

42. Charlesworth B. The Evolution of Sex Chromosomes. Science. 1991;251: 1030–1033.

43. Kirkpatrick M. The Evolution of Genome Structure by Natural and Sexual Selection. J Hered. 2017;108: 3–11.

44. Furman BLS, Metzger DCH, Darolti I, Wright AE, Sandkam BA, Almeida P, et al. Sex Chromosome Evolution: So Many Exceptions to the Rules. Genome Biol Evol. 2020;12: 750–763.

45. McVean G, Auton A. LDhat 2.1: a package for the population genetic analysis of recombination. Department of Statistics, Oxford, OX1 3TG, UK. 2007. Available: https://pdfs.semanticscholar.org/37c0/71c2bcd115bcce4bb2e4eb89a7dbb861c136.pdf

46. Charlesworth D. Does sexual dimorphism in plants promote sex chromosome evolution? Environ Exp Bot. 2018;146: 5–12.

47. Lesaffre T, Pannell JR, Mullon C. The joint evolution of separate sexes and sexual dimorphism. bioRxiv. 2024. p. 2024.05.31.596835. doi:10.1101/2024.05.31.596835

48. Costea M, Weaver SE, Tardif FJ. The Biology of Invasive Alien Plants in Canada. 3. Amaranthus tuberculatus (Moq.) Sauer var. rudis (Sauer) Costea & Tardif. Can J Plant Sci. 2005;85: 507–522.

49. Freeman DC, Doust JL, El-Keblawy A, Miglia KJ, McArthur ED. Sexual specialization and inbreeding avoidance in the evolution of dioecy. Bot Rev. 1997;63: 65–92.

50. Feil J. Reproductive ecology of dioecious Siparuna (Monimiaceae) in Ecuador: a case of gall midge pollination. Bot J Linn Soc. 1992;110: 171–203.

51. RENNER, S. S., AND G. HAUSNER. Siparunaceae and Monimiacea. In: Andersson GHA, editor. Flora of Ecuador. Sweden: Berlings Arlov; 1997. pp. 1–125.

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

53. Cossard GG, Pannell JR. Enhanced leaky sex expression in response to pollen limitation in the dioecious plant Mercurialis annua. J Evol Biol. 2021;34: 416–422.

54. Wu W, Jernstedt J, Mesgaran MB. Comparative floral development in male and female plants of Palmer amaranth (Amaranthus palmeri). Am J Bot. 2023;110: e16212.

55. Blaser O, Grossen C, Neuenschwander S, Perrin N. Sex-chromosome turnovers induced by deleterious mutation load. Evolution. 2013;67: 635–645.

56. Guerrero RF, Rousset F, Kirkpatrick M. Coalescent patterns for chromosomal inversions in divergent populations. Philos Trans R Soc Lond B Biol Sci. 2012;367: 430–438.

57. Kapun M, Mitchell ED, Kawecki TJ, Schmidt P, Flatt T. An Ancestral Balanced Inversion Polymorphism Confers Global Adaptation. Mol Biol Evol. 2023;40. doi:10.1093/molbev/msad118

58. Navarro A, Betrán E, Barbadilla A, Ruiz A. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics. 1997;146: 695–709.

59. Schaeffer SW, Anderson WW. Mechanisms of genetic exchange within the chromosomal inversions of Drosophila pseudoobscura. Genetics. 2005;171: 1729–1739.

60. Chang C-H, Gregory LE, Gordon KE, Meiklejohn CD, Larracuente AM. Unique structure and positive selection promote the rapid divergence of Drosophila Y chromosomes. Elife. 2022;11. doi:10.7554/eLife.75795

61. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature. 2003;423: 873–876.

62. Connallon T, Clark AG. Gene duplication, gene conversion and the evolution of the Y chromosome. Genetics. 2010;186: 277–286.

63. De-Kayne R, Gordon IJ, Terblanche RF, Collins S, Omufwoko KS, Martins DJ, et al. Extensive haplotype diversity in a butterfly colour pattern supergene is fuelled by incomplete recombination suppression. bioRxiv. 2024. p. 2024.07.26.605145. doi:10.1101/2024.07.26.605145

64. Almeida P, Sandkam BA, Morris J, Darolti I, Breden F, Mank JE. Divergence and Remarkable Diversity of the Y Chromosome in Guppies. Mol Biol Evol. 2021;38: 619–633.

65. Charlesworth B, Barton NH. The Spread of an Inversion with Migration and Selection. Genetics. 2018;208: 377–382.

66. Kapun M, Flatt T. The adaptive significance of chromosomal inversion polymorphisms in Drosophila melanogaster. Mol Ecol. 2019;28: 1263–1282.

67. Charlesworth B, Flatt T. On the fixation or non-fixation of adaptive inversions. Authorea Preprints. Authorea, Inc.; 2021. doi:10.22541/au.161474567.72404976/v1

68. Shore P, Sharrocks AD. The MADS-box family of transcription factors. Eur J Biochem. 1995;229: 1–13.

69. Masiero S, Colombo L, Grini PE, Schnittger A, Kater MM. The emerging importance of type I MADS box transcription factors for plant reproduction. Plant Cell. 2011;23: 865–872.

70. Airoldi CA, Davies B. Gene duplication and the evolution of plant MADS-box transcription factors. J Genet Genomics. 2012;39: 157–165.

71. Qiu Y, Li Z, Köhler C. Ancestral duplication of MADS-box genes in land plants empowered the functional divergence between sporophytes and gametophytes. New Phytol. 2024. doi:10.1111/nph.20065

72. Cronk Q, Müller NA. Default Sex and Single Gene Sex Determination in Dioecious Plants. Front Plant Sci. 2020;11: 1162.

73. Kreiner JM, Giacomini DA, Bemm F, Waithaka B, Regalado J, Lanz C, et al. Multiple modes of convergent adaptation in the spread of glyphosate-resistant Amaranthus tuberculatus. Proc Natl Acad Sci U S A. 2019. doi:10.1073/pnas.1900870116

74. Padmarasu S, Himmelbach A, Mascher M, Stein N. In Situ Hi-C for Plants: An Improved Method to Detect Long-Range Chromatin Interactions. Methods Mol Biol. 2019;1933: 441–472.

75. Dong P, Zhong S. Characterization of Plant 3D Chromatin Architecture, In Situ Hi-C Library Preparation, and Data Analysis. Methods Mol Biol. 2020;2093: 147–157.

76. Todesco M, Owens GL, Bercovich N, Légaré J-S, Soudi S, Burge DO, et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. Nature. 2020;584: 602–607.

77. Stoffel K, van Leeuwen H, Kozik A, Caldwell D, Ashrafi H, Cui X, et al. Development and application of a 6.5 million feature Affymetrix Genechip® for massively parallel discovery of single position polymorphisms in lettuce (Lactuca spp.). BMC Genomics. 2012;13: 185.

78. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18: 170–175.

79. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btac808

80. Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics. 2017;33: 2197–2198.

81. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18: 188–196.

82. Adhikary D, Deyholos MK, Délano-Frier JP. The Amaranth Genome. Springer Nature; 2021.

83. Ma X, Vaistij FE, Li Y, Jansen van Rensburg WS, Harvey S, Bairu MW, et al. 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. 2021;107: 613–628.

84. Wang H, Xu D, Wang S, Wang A, Lei L, Jiang F, et al. Chromosome-scale Amaranthus tricolor genome provides insights into the evolution of the genus Amaranthus and the mechanism of betalain biosynthesis. DNA Res. 2023;30. doi:10.1093/dnares/dsac050

85. Montgomery JS, Giacomini D, Waithaka B, Lanz C, Murphy BP, Campe R, et al. Draft Genomes of Amaranthus tuberculatus, Amaranthus hybridus, and Amaranthus palmeri. Genome Biol Evol. 2020;12: 1988–1993.

86. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

87. Danecek P, Bonfield JK, Liddle J, Marshall J. Twelve years of SAMtools and BCFtools. 2021. Available: https://academic.oup.com/gigascience/article-abstract/10/2/giab008/6137722

88. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44: 821–824.

89. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530–1534.

90. Korunes KL, Samuk K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Mol Ecol Resour. 2021;21: 1359–1368.

91. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media; 2009.

92. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2 (2020). R package version 1.1. 1. 2021.

93. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34: 867–868.

94. Kreiner JM, Latorre SM, Burbano HA, Stinchcombe JR, Otto SP, Weigel D, et al. Rapid weed adaptation and range expansion in response to agriculture over the past two centuries. Science. 2022;378: 1079–1085.

95. Peltzer A. Computational methods for ancient genome reconstruction. Eberhard Karls Universität Tübingen. 2019. Available: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3022008

96. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29: 1682–1684.

1199