The role of breakpoint mutations, supergene effects, and ancient nested rearrangements in the evolution of adaptive chromosome inversions in the yellow monkey flower, *Mimulus guttatus*

Authors: Leslie M. Kollar[1,2,3], Lauren E. Stanley[1], Sunil Kumar Kenchanmane Raju[4], David B. Lowry[1,2,3], Chad E. Niederhuth[1]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI, 48824

[2]Plant Resilience Institute, Michigan State University, East Lansing, MI, 48824

[3]Program in Ecology, Evolution, and Behavior, Michigan State University, East Lansing, MI, 48824

[4]Center for Genomics and Systems Biology, New York University, NY, 10003

First Author: Leslie M. Kollar

Corresponding Author: Chad E. Niederhuth and Leslie M. Kollar

Preprint servers: biorxiv

Classifications: Biological sciences

Keywords: chromosome inversion, supergene, breakpoint mutation

Abstract

Large chromosomal inversion polymorphisms are ubiquitous across the diversity of diploid organisms and play a large role in the evolution of adaptations in those species. Despite their importance, the underlying mechanisms by which inversions produce their adaptive phenotypic effects and become geographically widespread are still poorly understood. One way inversions could cause phenotypic effects is through meiotic recombination suppression, which can result in the formation of a supergene containing linked adaptive alleles. This supergene hypothesis has been promoted by theoreticians, but thus far, no studies have definitively identified multiple linked adaptive genes within an inversion. Alternatively, according to the breakpoint mutation hypothesis, the inversion mutation itself could result in adaptive phenotypic effects if it disrupts genes or alters regulation of nearby genes. Here, we evaluate both of these hypotheses using new long-read sequencing-based genomes of the yellow monkey flower, *Mimulus guttatus*. Our results provide support for both the supergene and breakpoint mutation hypotheses of adaptive inversion evolution and suggest that functional molecular studies will be required to definitively test each of these hypotheses. We also identified an ancient large inversion nested within a well-established adaptive inversion. This finding suggests that the supergene mechanism may occur in phases, with an expansion of the region of suppressed recombination capturing an increasing number of adaptive loci over time.

Significance statement (50 - 120 words)

Large regions of chromosomes can become reversed in orientation due to mutations known as inversions. These inversions are often involved in the evolution of major organismal adaptations, yet we know little about this process. In this study, we show that the evolution of a large adaptive inversion in monkeyflowers could be due to both the trapping of beneficial genes into a supergene as well as changes in gene expression due to the inversion mutation itself. Further, we found that this inversion has trapped an even older inversion, which may itself have

been adaptive. Thus, adaptive inversions on autosomes may evolve in sequence, much the same way that sex chromosomes evolve.

Introduction

Chromosomal inversions have been implicated in evolutionary adaptations since classic studies in *Drosophila* found that inversion polymorphisms are frequently correlated with environmental conditions over geographic space and through time (1–6). Large chromosomal inversions rearrange the structure of genomes and can have major phenotypic consequences (7–12). Inversions strongly suppress genetic recombination in heterokaryotic (inversion heterozygotes) individuals because recombinant gametes are unbalanced (2, 13, 14). Researchers have thus hypothesized that inversions could act as supergenes by holding together long haplotype blocks containing multiple adaptive polymorphisms through suppressed recombination (2, 3, 13, 15–18).

The realization that inversions could spread rapidly as a result of trapping adaptive alleles at multiple linked loci led Kirkpatrick and Barton to state a decade ago: “If the local adaptation mechanism is so powerful, why are inversions not everywhere? One possibility is that they in fact are, and that their frequency has been greatly underestimated (17).” Since Kirkpatrick and Barton’s 2006 paper, it has become clear that inversion polymorphisms truly are everywhere and are frequently associated with evolutionary adaptations. At least 49 unique inversions have now been linked to adaptive phenotypic traits in *Anopheles* mosquitoes (19). Inversions have been shown to control the phenotypes underlying mimicry in butterflies (15, 16, 20) as well as mating and migratory behaviors in birds (21–24). Shifts in the frequency of inversion polymorphisms have been linked to adaptation to forest versus prairie habitats in deer mice (25) and to global climate change in *Drosophila* (26, 27). Inversions appear to play a prominent role in adaptation and domestication of important crops, including maize, wheat, and peaches (28–31). Humans and chimpanzee genomes differ by nine major chromosomal inversions and >100 smaller ones

(32, 33), which may suggest they played a key role in speciation from our closest relatives (34, 35). Given the apparent major importance of chromosomal inversions in adaptation, and the prevalent involvement of inversions in human disease (11, 12, 36–41), it is surprising that we still know so little about how inversions are involved in evolutionary adaptations and produce their important adaptive phenotypic effects.

While the supergene hypothesis has been championed by evolutionary biologists in recent years, it is not the only mechanism by which inversions can contribute to adaptive phenotypic variation. Inversion mutations can also break genes or alter gene expression through mutations in promoters/enhancers or by changing the local chromatin architecture near breakpoints (7–9, 42, 43). This breakpoint hypothesis of adaptive inversion evolution and the supergene hypothesis are not mutually exclusive, and there may be adaptive inversions that spread by combining breakpoint mutation effects with the trapping of multiple adaptive loci through suppressed recombination (6). Despite decades of theoretical modeling and speculation, and the recent realization that adaptive inversions are everywhere, we still lack a molecular mechanistic understanding of the evolution of any adaptive inversion. This is perhaps unsurprising, given that (1) suppressed recombination within inversions has thwarted efforts to fine-map functional genetic variation; and (2) the breakpoints of inversions frequently fall in repetitive, transposon-rich regions of the genome. Further, only with the recent invention of long-read sequencing technologies has it become feasible to map inversion breakpoints and characterize their features (13, 44). Identification of the mechanistic causes of inversion evolution requires new efforts in systems where inversions can be definitively linked to adaptation in nature, where genome sequencing results in well assembled genomes, and functional tools are available to test mechanistic hypotheses that are generated by genomic analyses.

An ideal system to evaluate the supergene and breakpoint mutation hypothesis of inversion evolution is the yellow monkeyflower, *Mimulus guttatus* (syn. *Erythranthe guttata*). Over a decade ago, a chromosomal inversion polymorphism on chromosome 8 underlying a quantitative trait locus (*DIV1*) in this system was definitively linked to local adaptation through a field reciprocal transplant experiment (45). To our knowledge, this is the only inversion polymorphism that has been definitively linked to local adaptation in nature. The inversion at the *DIV1* locus, which we henceforth refer to as inv_chr8A, has large phenotypic effects on a suite of adaptive traits associated with the transition from an annual to perennial life history, including flowering time, branching, allocation to vegetative growth, and herbivore resistance (45, 46). The inv_chr8A inversion polymorphism is widespread across the geographic range of *Mimulus guttatus*, with one orientation of the inversion commonly found in perennial populations, and the other in annual populations (45). The inversion most strikingly contributes to the divergence of coastal perennial populations, which experience cooler conditions and moisture supplied by summer sea fog, and nearby inland annual populations, which complete their life cycle quickly due to diminishing water availability during summer drought. .

Recently, Coughlin and Willis (47) provided evidence to support the hypothesis that the inv_chr8A trapped pre-existing adaptive alleles in crosses between annual *M. guttatus* and a perennial sister species, *M. tilingii*, which share the same ancestral chromosomal inversion orientation. The orientation of inv_chr8A found in perennial *M. guttatus* is derived, even though perenniality itself is thought to be the ancestral state in this system (47, 48). Further, Coughlin and Willis (47) were able to exploit free recombination within the inv_chr8A region to show that there are at least two quantitative trait loci (QTL) for morphological divergence in crosses between annual *M. guttatus* and perennial *M. tilingii*. While this result provides circumstantial evidence for the supergene hypothesis, the QTL in annual *M. guttatus* x perennial *M. tilingii* cross did not affect key morphological traits, like stolon production, that are controlled by the

inv_chr8A inversion (45, 47–49). This inconsistency in phenotypic effects suggests that the QTL effects found in the annual *M. guttatus* x perennial *M. tilingii* cross are caused by alleles that arose independently in *M. tillingii,* and thus may not be the same as those that cause phenotypic effects in annual *M. guttatus* x perennial *M. guttatus* crosses. Regardless, definitive evidence for the supergene hypothesis will require identification of the exact genes and alleles underlying the phenotypic effects of the inversion polymorphism.

Identification of the genes that cause the phenotypic effects of inversion polymorphism in *M. guttatus* have been difficult because available genome assemblies derived from short read sequencing are fragmented (50). As a result, the full complement of genomic DNA and the breakpoints of the adaptive inv_chr8A and other known inversions have remained unidentified.

In this study, we report the sequencing, assembly, and annotation of new genomes for coastal perennial and inland annual lines of *M. guttatus* using Oxford Nanopore long-read sequencing. With these genomes, we were able to overcome previous challenges and identify the breakpoints of these inversions as well as the full complement of genes within the inversions, including inv_chr8A. To identify candidate genes for the adaptive phenotypic effects of the inversions, as well as explore possible phenotypic consequences of the inversion breakpoint mutations themselves, we combined our new genome assemblies with a set of existing population genetic and gene expression datasets. Our goal was to evaluate the evidence for the breakpoint and supergene hypotheses of inversion evolution and set the stage for future functional studies to conclusively test these hypotheses. We also made a few unexpected discoveries, including the identification of another large chromosomal inversion (inv_chr8B) nested within inv_chr8A. Overall, this study brings us closer to understanding how chromosomal inversions evolve and why genomes rearrange dramatically over long periods of evolutionary history.

Results

De novo assembly of M. guttatus inland annual and coastal perennial lines

We focused our sequencing, assembly, and annotation efforts on two inbred lines: SWB-S1, a coastal perennial, and LMC-L1, an inland annual. These lines have been used in multiple studies (45, 46, 48, 51) and are derived from population collections from Mendocino County in Northern California. While the presence of large chromosomal rearrangements can be identified by population genetics methods with short-read sequencing (52), determining the full complement of DNA within an inversion and the locations of inversion breakpoints requires long-read sequencing. To localize inversion breakpoints of known inversions that differentiate annuals and perennials and to detect previously unknown inversions, we conducted long-read sequencing using Oxford Nanopore Technologies (ONT).

The final, polished genome assemblies were approximately 277 Mbps spanning 1,198 contigs for LMC-L1 and 278 Mbps spanning 707 contigs for SWB-S1. Both assemblies demonstrated high contiguity with a contig N50 of 5.83 Mbps (LMC-L1) and 4.90 Mbps (SWB-S1) and a scaffold N50 of 18.2 Mbps (LMC-L1) and 18.8 Mbps (SWB-S1). Based on genetic maps and synteny, we assembled 93.0% of the LMC-L1 bps and 93.7% of the SWB-S1 sequence into 14 chromosomes (50). Our genomes recovered more than 98% of the eudicot set of BUSCO (53) orthologs (eudicots_odb10), an improvement from 96.9% in the previously available IM62 v2 assembly (50). The LTR Assembly Index (LAI), an independent assessment of assembly quality based on the completeness of assembled LTR retrotransposons (54), was much higher in both LMC-L1 (13.47) and SWB-S1 (16.28) than the IM62 v2 reference (8.79), indicating more complete assembly of repetitive regions (supplemental table 1).

Genes were annotated using a combination of *MAKER* (55, 56) and *liftoff* (57). We annotated a total of 27,583 genes in LMC-L1 and 26,876 genes in SWB, in comparison to 28,140 genes in IM62 v2 (50) (supplemental table 1). We compared the LMC-L1 and SWB-S1 genomes using *GENESPACE* (58), identifying 19,245 syntenic genes between LMC-L1 and SWB-S1 (supplemental table 2). In LMC-L1, we found 5,105 presence absence variants (PAVs), 3,026 pseudogenes, and 1,411 tandem duplicates. In SWB-S1, we found 4,707 PAVs, 3,110 pseudogenes, and 1,401 tandem duplicates.

Genome wide SNPs and structural variants

To identify larger structural variants (SVs) between the LMC-L1 and SWB-S1 genomes, we used *MUM&CO,* a whole genome alignment method (59). The number of SVs between the LMC-L1 and SWB-S1 genomes are reported in supplemental table 3. Including large inversions, which can contain hundreds of genes, 12,013 genes were associated with structural variants in LMC-L1 (variants affecting genes: 1,379 insertions, 8,493 deletions, 72 duplications, and 2,069 inversions), while 11,582 genes were associated with structural variants in the SWB-S1 (variants affecting genes: 1,375 insertions, 8,300 deletions, 51 duplications, and 1,856 inversions).

To identify SNPs between the LMC-L1 and SWB-S1 accessions, we aligned whole genome sequencing (WGS) from one accession to the other accession's reference genome. From these alignments, we focused on the evolution of coding regions by identifying high impact SNPs, which include the following variant types: stop gain (2,787 in SWB-S1 and 2,638 in LMC-L1), stop loss (723 in SWB-S1 and 700 in LMC-L1), frameshift deletion (7,299 in SWB-S1 and 7,541 in LMC-L1), frameshift insertion (7,002 in SWB-S1 and 6,951 in LMC-L1), and splicing (3,986 in SWB-S1 and 3,392 in LMC-L1). We also identified all other synonymous and nonsynonymous SNPs. All coding region SNPs are reported for both genomes in supplemental tables 4 and 5.

The evolution of large chromosome inversions

Determining the breakpoints of inversions lays the necessary foundation for testing hypotheses of inversion evolution. Previous efforts to identify inversion breakpoints in *M. guttatus* have been thwarted by the inability to sequence across highly repetitive breakpoint regions. Our new genome assemblies allowed us to fully assemble and localize breakpoint regions for the first time in the inv_chr8A inversion on chromosome 8. We were unable to define the start and stop position of each inversion to the exact nucleotide due to the complex repetitive nature of the region surrounding the breakpoints. Therefore, we defined the breakpoints as regions spanning from the last gene in the collinear syntenic block upstream of the inversion to the second gene of the inverted syntenic block. We included the second gene in the inverted block as a conservative estimate. On the opposite end of the inversion, we defined the breakpoint region as the second to last gene of the inverted syntenic block to the first gene of the collinear syntenic block upstream of the inversion (supplemental figure 13). It is important to note that, for all estimates of sequence level differences, we used the first base pair of the first gene of the inverted syntenic block to the last gene of the inverted syntenic block. In addition to inv_chr8A, we were able to localize the boundaries of another previously known inversion on chromosome 5 (inv_chr5A), as well as two previously unknown large inversions and many small inversions. These newly identified inversions include a smaller inversion nested within inv_chr8A (inv_chr8B) and a large inversion on chromosome 14 (inv_chr14A). Inv_chr8A was found to span 6.75 Mb in LMC-L1 and 5.6 Mb in SWB-S1 (figure 1A, supplemental figure 1, supplemental table 6). The smaller inversion nested within inv_chr8A, inv_chr8B, spans 213,792 bps in LMC-L1 and 261,302 bps in SWB-S1 (figure 1 B, supplemental figure 1, supplemental table 6). The fully assembled inv_chr5A is approximately 4.2 Mb long in LMC-L1 and 4.0 Mb in SWB-S1 (figure 1A, supplemental figure 1, supplemental table 6). While we have successfully captured inv_chr5A in our genome assemblies, it is located near the end of the

chromosome arm and there are gaps within the inversion in one of the assemblies. Therefore, it is unclear if the second breakpoint region is close to the end of inv_5A. The fourth large inversion, inv_chr14A, was previously unknown, with an approximate size of 2.4 Mb in LMC-L1 and 2.8 Mb in SWB-S1 (figure 1 A, supplemental figure 1, supplemental table 6).

We next examined the gene content differences between the inversion orientations. Inv_chr8A (including inv_chr8B) has 177 LMC-L1-specific genes, 125 SWB-S1-specific genes, and shares 613 genes between the two genomes. Inv_chr8B shares 36 genes between the genomes, but has 9 LMC-L1-specific genes and 8 SWB-S1-specific genes. Inv_chr5A has a total of 113 LMC-L1 specific genes and 90 SWB-S1 specific genes, but shares 396 genes between the LMC-L1 and SWB-S1. Lastly, the chromosome 14 inversion has 247 LMC-L1-specific genes,308 SWB-S1-specific genes, and 1,133 shared genes (supplemental table 7). Many of these large gene content differences are due to PAVs and copy number variants (CNVs).

While previous studies have demonstrated the role of inv_chr8A in local adaptation (45), and implicated inv_chr5A (60), our improved assemblies now capture the full complement of genes within these inversions and make it possible to identify candidate genes that could have driven the geographic distributions of these inversions. In the following section, we highlight genetic differences between inversion orientations. CNVs, including PAVs, could explain both differences in inversion size and phenotypic differences. Within inv_chr8A, there were 338 CNVs with 302 being PAVs; in inv_chr8B, there were 19 CNVs with 17 being PAVs. Within inv_chr5A, there were 227 CNVs, with 203 being PAVs. In inv_chr14A, there were 100 CNVs, with 84 being PAVs. Some of these CNVs in the inversions could be explained by genes becoming pseudogenized within the inversion. In inv_chr8A, including inv_chr8B, there were 90 pseudogenes in LMC-L1 and 74 pseudogenes in SWB-S1. Inv_chr8B included 6 pseudogenes in LMC-L1 and 8 pseudogenes in SWB-S1. In inv_chr5A, there were 54 pseudogenes in LMC-

L1 and 43 pseudogenes in SWB-S1. Lastly, in inv_chr14A there were 22 pseudogenes in LMC-L1 and 19 pseudogenes in SWB-S1. We report SNPs (all SNPs, high impact SNPs, and synonymous and nonsynonymous SNPs) falling within the chromosome inversions in supplemental tables 4 and 5. These SNPs were identified from WGS comparison of the SWB-S1 and LMC-L1 genomes.

Allelic differentiation is elevated in large inversions

To determine which inversion orientation is ancestral, we implemented GENESPACE (58) using *Mimulus lewisii,* a distant relative, as the reference genome (http://mimubase.org/FTP/Genomes/) (supplemental figure 2). From this analysis, we reasoned that inversions sharing the same orientation as *M. lewisii* are in the ancestral orientation. We found the orientations of inv_ch5A (supplemental figure 3), inv_chr8A (supplemental figure 4), and inv_chr8B (supplemental figure 4) are ancestral in LMC-L1. SWB-S1 has the ancestral orientation for inv_chr14A (supplemental figure 5). The results for inv_chr8A are consistent with the results found by Coughlan et al (61). Given that the orientation of inv_chr8B is shared between ecotypes and that inv_chr8B is on opposite ends of inv_chr8A, it is likely that inv_chr8B is derived in SWB-S1 but formed prior to inv_chr8A. To test this hypothesis, we used the rate of synonymous substitutions (synonymous SNPs / the number of mappable bases) to estimate the relative age of the chromosome inversions. Here, we see that the rate of synonymous SNPs is higher in inv_chr8B (LMC-L1: 0.0027 SWB-S1: 0.0023, supplemental table 5) than in inv_chr8A (LMC-L1: 0.0016, SWB-S1: 0.0013, supplemental table 5). This suggests that inv_chr8B arose prior to inv_chr8A.

A previous population genetics study found that population genetic summary statistics for inverted regions were significantly different from genome-wide averages (60). That study compared population genetic parameters for whole genome pooled sequences of 101 coastal

perennial accessions versus 92 inland annual accessions (60). We reanalyzed this data using our new LMC-L1 and SWB-S1 genome assemblies. Given the improvements in the genome assemblies over the previous reference IM62 v2 genome (50), we anticipated potentially new patterns in sequence diversity and differentiation. This new analysis also allowed us to evaluate how reference bias in alignments can affect population genetic summary statistic estimation.

As in other studies of *M. guttatus*, genetic diversity (50, 62) was very high, but estimates varied slightly based on which reference genome was used for analyses. For example, when reads were aligned to the SWB-S1 genome, the coastal perennial pool's nucleotide diversity (π) is 0.032, whereas the inland annual pool is 0.040. When reads were aligned to the LMC-L1 genome, the coastal perennial pool's π estimate is 0.034, whereas the inland annual pool remains 0.040. Regardless of the reference genome, the inland annual ecotype had slightly higher nucleotide diversity, which is consistent with prior studies (60) (supplemental figure 6).

To better understand the evolution of the large inversions, we compared the nucleotide diversity (π) and allelic differentiation (G statistic) within each inversion to the rest of the collinear genome. We first conducted a genome-wide analysis of the ratio of π (π Annual/π Perennial). We found the ratios of π to be similar across inversions and collinear regions of the genome; however, inv_chr8B showed a higher π in the inland annual pool than in the coastal perennial pool (figure 2, supplemental figure 6 and supplemental table 8). Similar to Gould et al (60), we identified inverted regions to be highly differentiated in comparison to the rest of the collinear genome (figure 2, supplemental figure 7 and supplemental table 8). Interestingly, inv_chr8B was even more differentiated (G statistic) than the other inversions, including the inv_chr8A inversion it is nested within.

Multiple candidate genes could contribute to inversion supergene effects

With clear resolution of the breakpoints for inv_chr8A and inv_chr8B, we were able to localize candidate genes that could underlie the phenotypic effects of inversions. We identified genes affected by divergence as those with outlier (top 1%) windows for the G statistic. The top windows for *G* contained 753 LMC-L1 and 635 SWB-S1 genes and/or their promoters and transcription start site (TSS) regions or 3' untranslated regions (UTR). Many of these outlier genes, promoters and TSS regions, and 3' UTRs were syntenic between the LMC-L1 and SWB-S1 genomes, including193 genes, 57 promoters and TSS regions, and 90 3' UTRs. Some of these outlier regions also fall within the chromosome inversions (supplemental table 9). Outlier windows falling within genes, promoters and TSS regions, and 3' UTRs are interesting because these could contribute to phenotypic differences between the ecotypes. While these results do not suggest a direct cause in phenotypic variation, it does suggest these outlier genes are differentiated and warrant further investigation.

One such gene is *GA20ox2*, which codes for a structure gene in the gibberellic acid synthesis pathway. Foliar application of gibberellic acid growth hormone to coastal perennial plants has been shown to increase internode elongation and convert branches from vegetative stolons to aerial flowering branches in a manner that mimics inland annual growth architecture (Lowry et al. 2019). *GA20ox2* falls within inv_chr8A and may contribute to adaptive divergence between the ecotypes. *GA20ox2* (MgS1_08g02280) was an outlier for G and π ratio in our reanalysis using the SWB-S1 reference genome, similar to what was found in Gould et al (60) when using IM62 as the reference genome. Comparison of the sequence for this gene in each of the three genomes showed that the model for this gene in IM62, Migut.H00683, is missing a large segment, including the entire oxoglutarate/iron-dependent dioxygenase (DIOX_N) domain. Both MgS1_08g02280 and MgL1_08g0964, the gene names for *GA20ox2* in each of the two reference genomes, appear to be intact genes with ORFs. This was confirmed by Sanger sequencing, and both genes are expressed based on qRT-PCR (figure 3). In LMC-L1,

expression of *GA20ox2* was ~1.7-fold higher in floral shoot apex compared to SWB-S1. Other GA-related genes were found to be outliers when using the LMC-L1 genome as the reference. *GAST1 protein homolog 5* (*GASA5*, MgL1_08g23640) was an FST outlier for both the gene and promoter region. MgL1_05g11230, a GRAS family domain gene, was differentially expressed, as well as an outlier for the G statistic and π ratio. MgL1_08g08120, another GRAS family domain gene, was an outlier for FST but not differentially expressed.

Another likely set of genes underlying phenotypes that differentiate ecotypes are a tandem array of *R2R3-MYB* genes within inv_chr8A (63, 64). Previous work showed that several calyx, corolla, and leaf anthocyanin pigmentation traits differentiating the ecotypes also map to the region containing inv_chr8A (64). It has been difficult to resolve the number of tandem copies of *R2R3-MYB* with previously available genome assemblies (i.e. IM62, TOL, and NONTOL; phytozome-next.jgi.doe.gov/). Based on our CNV analysis, LMC-L1 and SWB-S1 have different numbers of copies of these subgroup 6 *R2R3-MYB* genes. The LMC-L1 genome has 5 partial or complete copies of this gene, with 3 being flagged as pseudogenes (genes without valid open reading frames) in our analysis, while the SWB-S1 genome appears to have 6 partial or complete copies of this gene, none of which were called as pseudogenes (figure 4). Interestingly, all of these appear syntenic with conserved directionality, except for MgL1_08g06980, which is in the opposite orientation of all the other MYBs. This could indicate that it was duplicated independently. In investigating why some LMC-L1 copies were called pseudogenes, we found that MgL1_08g07004 matched perfectly to an annotation in IM62 (H00280), which is not annotated as an R2R3-MYB gene because it is missing distinguishing domains. However, MgL1_08g07004 almost perfectly matches MgL1_08g07000, especially in its intronic sequence, indicating that it is a partially duplicated gene.

Genes near inversion breakpoints were differentially expressed

The identification of inversion breakpoints allows for the evaluation of whether those breakpoints have disrupted genes in ways that could contribute to phenotypic effects. If an inversion breakpoint occurs within a gene, it would disrupt and likely eliminate its function. Breakpoints occurring near genes can also cause changes in gene expression by disrupting or changing the position of *cis*-regulatory elements. To establish whether gene expression might have been disrupted by the inversion breakpoint mutations, we evaluated transcript abundances within the breakpoint regions. We focused on the first two and last two genes of each inversion. We also evaluated transcript abundance for the first genes on either side just outside of the inverted region (supplemental figure 13).

We identified multiple genes which have changes that could have resulted from the inversion breakpoint mutations. In the SWB-S1 genome, MgS1_08g08940 (unknown function) falls just inside of inv_chr8B and is likely a pseudogenized orthologue of the MgL1_08g02150 (cysteine oxidase) gene found in LMC-L1. MgS1_08g08940 has an intact open reading frame and a highly similar nucleotide sequence to MgL1_08g02150. However, the predicted peptides for MgS1_08g0894 and MgL1_08g02150 are quite divergent due to a frameshift mutation that drastically alters the open reading frame of MgL1_08g02150. In the SWB-S1 inv_chr8B breakpoint region, there is a duplicated WRKY TF (MgS1_08g08950 and MgS1_08g08960) that is not duplicated near the LMC-L1 inv_chr8B breakpoint (MgL1_08g02160). We also found numerous gene content differences in the breakpoint regions of all inversions. However, our annotation differences could account for some of the differences, thus inflating gene content differences. Interestingly, in comparison to the other three major inversions, inv_chr14A seemed to have fewer gene content differences at the breakpoint regions. This is likely because inv_chr14A has had less time to accumulate mutational differences than the older three large inversions evaluated in this study. We also investigated these breakpoint region genes by

looking at differentially expressed genes from RNAseq data (51) generated from a reciprocal transplant experiment (figure 5, supplemental table 9). We found that a total of 7 genes within the breakpoint regions across 4 inversions had genetic differences in terms of expression levels.

DISCUSSION

While chromosomal inversions are now well recognized as key factors in adaptive evolution, elucidation of the molecular mechanisms underlying their adaptive phenotypic effects that cause them to spread has remained elusive. Our current study brings us one step closer to clarifying these molecular mechanisms for a chromosomal inversion that has been demonstrated to contribute to local adaptation in nature (45). Isolation of the breakpoints for this inversion allowed us to finally evaluate potential consequences of the original inversion mutation and clarified the divergent gene content for annual and perennial orientations of the inversion. Beyond the inv_chr8A inversion, long-read sequencing allowed us to localize three other large chromosomal inversions, each of which illustrates a key component of the evolution of inversions within species. Our results provide support for both the supergene and breakpoint mutation hypotheses of adaptive inversion evolution and indicate that functional molecular studies will be required to definitively test each of these hypotheses.

Previous research identified two of the four inversions we identified in our genome assemblies. The chromosome 8 inversion was initially reported in (45) and confirmed to have different orientations between the ecotypes. Additionally, the chromosome 8 inversion and its role in adaptive divergence have been well documented in *M. guttatus.* The chromosome 5 inversion has not been as well studied, but was initially reported in a cross between inland annual and coastal perennial lines (65). The chromosome 5 inversion is suspected to play a role in adaptation based on elevated allelic differentiation of this region between coastal perennial and

inland annual populations (51), which is a similar pattern to that for the known adaptive inversion on chromosome 8. Interestingly, in our analyses, inv_chr8A contained an internal smaller inversion. Previous studies did not identify the smaller internal inversion (inv_chr8B) on chromosome 8, and further research is needed to clarify its prevalence in other populations. The inv_chr14A inversion was mentioned in Flagel et al (66) as a 4 MB region of suppressed recombination. We were able to confirm that this region is in fact an inversion. Based on our analyses, the chromosome 14 inversion was less differentiated between the ecotypes, suggesting that it may not be as geographically widespread as the chromosome 5 and 8 inversions. Overall, the assembly of these genomes sets the stage to develop methodologies to more rapidly genotype the inversion polymorphisms and understand their evolution.

The role of suppressed recombination in the evolution of adaptive chromosomal inversions

Inversions are thought to play a major role in adaptation because they suppress recombination in heterokaryotic individuals, while allowing for free recombination in homokaryotic individuals. This dynamic means that locally adaptive alleles at distant loci on the same chromosome can be maintained in the same haplotypes for populations located in divergent habitats. At the same time, free recombination within each habitat allows for purging of deleterious and maladapted alleles for the same genomic region within habitats (67). In this way, inversions likely have evolutionary advantages over other types of long regions of suppressed recombination, like pericentromeric regions, where selection against maladapted alleles will lead the entire genomic region to be dragged along by linkage hitchhiking with selection.

While the idea that maintenance of coadapted alleles through suppressed recombination is a primary reason for the association of inversions with local adaptation, definitive evidence of this

through identification of the underlying genes has not been presented. With the full assembly of the well-established adaptive inversion on chromosome 8, we are now well positioned to finally test whether adaptive inversions trap co-adapted alleles across multiple genes into the same haplotype. Genes within this inversion involved in the gibberellin (GA) hormone pathway are of particular interest because of the discovery that the addition of gibberellin (GA3) to coastal perennial plants leads them to develop more like inland annual plants by converting vegetative stolons into upcurved flowering branches (46). Introgressing the inland annual orientation of the inversion into the coastal perennial background resulted in nearly identical phenotypic effects on branching and internode elongation as spraying coastal plants with GA3 (45). The GA genes *GA20ox2* and *GASA5* are top candidate genes for contributing to inv_chr8A's phenotypic effects. Not only were these genes allele frequency outliers in comparisons of perennial and annual populations, but *GA20ox2* was expressed at a higher level in the shoot apex of LMC-L1 than SWB-S1. *GA20ox2* expression is tissue- and time-specific and we were not able to find evidence that it was expressed when analyzing RNA-seq data from leaf tissue (60). In other systems (68), *GA20ox2* is typically expressed in the floral shoot apex, similar to our findings in *M. guttatus.* Gould et al *(60)* collected tissue only from leaves, which did not detect differential expression of *GA20ox2* in the floral shoot apex, which we found here with RT-qPCR.

While our study, as well as ongoing work by our group, implicates a role for *GA20ox2* in the phenotypic effects of inv_chr8A, the supergene hypothesis requires that we identify multiple genes within the inversion that contribute to adaptive phenotypic divergence. A promising candidate for also contributing to adaptive phenotypic effects is the cluster of R2R3-MYB genes within inv_chr8A. A previous study mapped variation in five vegetative anthocyanin traits to the region inv_chr8A (64), and coastal perennial and inland annual populations do differ in anthocyanin pigmentation. While this gene cluster is promising, it has been difficult in general to link vegetative anthocyanin polymorphisms to evolutionary adaptations (69–74).

The assembly of long-read chromosome scale annual and perennial genomes reported in this study has now provided the full complement of genes within the adaptive chromosome 8 inversion. This development has occurred at the same time that it is becoming much more feasible to conduct functional genetic research in multiple monkeyflower systems (75). We are currently in the process of developing overexpression and knockout lines for multiple genes within the chromosome 8 inversion. These studies will be able to definitely establish whether this inversion has trapped multiple co-adapted alleles into distinct haplotypes, as predicted by the supergene hypothesis. It should be mentioned that the potential for an inversion's spread is greater if it captures alleles at loci that are farther apart from each other in terms of recombination distance (76, 77). Thus, we plan to focus our future efforts on identifying adaptive alleles within inversions that are at loci distant from each other along chromosomes.

The role of chromosomal breakpoints in adaptive inversion evolution

The supergene hypothesis has gained considerable attention by evolutionary biologists in recent years. However, the role of the original inversion breakpoint mutations should not be ignored as potential causes of their phenotypic effects. The simplest way that inversion breakpoints can result in disruption of gene function is if they break genes. We found no evidence in this study that the large inversions disrupted any genes. Instead, breakpoints were located primarily in gene-poor, repeat-rich regions of the genome. While we found no evidence that the large inversions in our study broke any genes, it is possible that inversion mutations could have resulted in changing gene expression of nearby genes by disrupting promoter/enhancer elements or changing local chromatin landscapes. Indeed, we did find that many of the genes located within the inversion breakpoints regions on chromosome 5 and 8 had significantly different levels of transcript abundance between SWB-S1 and LMC-L1 plants.

The evolution of complex nested inversions

The biggest surprise of our study was the discovery of a relatively large inversion (214-261 kb; depending on genome) nested within the even larger adaptive chromosome 8 inversion. The evolution of nested inversions is not unprecedented, but is also not generally discussed in the context of locally adaptive inversions. Instead, nesting and clustering of multiple inversions in close proximity is typically associated with the evolution of sex chromosomes. For sex chromosomes, the accumulation of successive inversions and other rearrangements leads to contrasting levels of divergence (i.e. strata) for different regions of sex chromosomes, depending on the age of each region of suppressed recombination (78–81). This process eventually leads to distinct X and Y or Z and W chromosomes that no longer contain pseudoautosomal regions that recombine at all at meiosis when heterozygous (80, 82). Beyond sex chromosomes, studies in both maize and sunflowers have found evidence for nested large inversions (52, 83, 84). In maize, nested inversions have been important for the evolution of chromosomal knobs that spread by meiotic drive (83, 84). The significance of the autosomal nested inversions in sunflowers is unclear (L. Rieseberg personal communication).

There are multiple possible hypotheses for how and why the nested inversions arose on chromosome 8. The entire region that includes both inversions is known to be involved in local adaptation to coastal perennial versus inland annual habitats. Given the geographic distribution of the larger inversion, it is also likely involved in adaptation to wetter versus drier habitats within inland regions (45, 49, 85). Therefore, natural selection is intimately involved in the spread of the larger inversion. The question remains as to whether the smaller inversion also confers habitat-dependent fitness advantages.

There are two alternative hypotheses about the order in which this complex of nested inversions evolved. Either the larger chromosomal 8 inversion (inv_chr8A) evolved first or the smaller inversion (inv_chr8B) evolved first. If the smaller inversion had evolved first, we would expect a greater level of nucleotide divergence for the region of the smaller inversion relative to the rest of the larger inversion, just like different strata in sex chromosomes (80). This is exactly what we found. Had the larger of the two inversions evolved first, we should have observed an equivalent or lower level of differentiation for inv_chr8B when compared to the rest of inv_chr8A. The finding that the internal inversion evolved first suggests the possibility that the supergene model of inversion evolution could occur in phases with expansion of the region of suppressed recombination over time through successively larger inversions, much the same way that recombination suppression is expanded over time on sex chromosomes.

Conclusion

Our in-depth study of two chromosome level assembly genomes has further clarified the evolution of large adaptive chromosomal inversions in *M. guttatus.* The inversion mutations themselves are likely facilitated by highly repetitive regions of the genome. While none of these inversion mutations has resulted in the breaking of genes, they may have directly resulted in changes in gene expression of neighboring genes. The evolution of the known adaptive inversion on chromosome 8 likely involved the capture of and/or accumulation of locally adaptive alleles throughout the region into distinct annual and perennial haplotypes that have accumulated greater divergence over time. The results from this study suggest that the large inversion may have also captured an ancient large inversion (~250kb), which itself could have been a locally adaptive inversion polymorphism. Future research, much of which will be facilitated by these genomes, is needed to functionally evaluate the supergene hypothesis, establish if changes in chromatin structure contribute to gene expression changes near breakpoints, and to establish the dynamics that gave rise to nested adaptive inversions on

chromosome 8. Future research should also attempt to clarify which phenotypic effects are conferred by the chromosome 5 inversions, which increasingly appears to be associated with inland annual versus coastal perennial life history variation. Studies like this one are beginning to flesh out the mechanisms of how and why genome structure evolves over deep evolutionary time.

Materials and Methods

Tissue harvesting and DNA extraction

Our genome sequencing efforts focused on one inland annual line, LMC-L1 (Yorktown, CA, N 38.863983, W -123.083917), and one coastal perennial line, SWB-S1 (Irish Beach, CA, N 39.035983, W -123.690467). Prior to genome sequencing, LMC-L1 was inbred for 6 generations and SWB-S1 was inbred for 14 generations in the Duke University greenhouses and the Michigan State University growth facilities. The fewer number of generations inbreeding of LMC-L1 is due to the line having a high level of sterility-based inbreeding depression. Floral buds were collected for DNA extractions from plants that were grown in growth chambers in the following conditions: 16 hour day length, 22°C day and 18°C night temperatures, with 60% relative humidity, and a light intensity of 460mE. DNA extracted from this tissue was used for both long read nanopore and short read Illumina sequencing. For nanopore sequencing, nuclei were extracted from the floral bud tissue following Lu et al (86), with minor modifications. We ground 1-1.5 g of floral bud tissue into a coarse powder in liquid nitrogen. Ground tissue was resuspended in LB01 buffer (15 mM Tris pH 7.5, 2 mM EDTA (Ethylenediaminetetraacetic acid), 80 mM KCL, 20 mM NaCl, 15 mM 2-Mercaptoethanol, 0.15% Triton-X 100 and 0.5 mM Spermine) on ice with intermittent mixing. Tissue homogenate was then filtered through four-layers of miracloth (EMD Millipore Corp) followed by filtration through a 20 µm cell strainer (pluriStrainer®). Filtrate was carefully pipetted on top of equal volume Density Gradient

Centrifugation Buffer (1.7 M Sucrose, 10 mM Tris-HCl pH 8.0. 2 mM MgCl2, 5 mM 2-Mercaptoethanol, 1 mM EDTA, 0.15% Triton-X 100) in a 50 ml falcon tube and centrifuged at 2500g for 30 mins at 4℃. After centrifugation, supernatant was decanted and the nuclei pellet was used to isolate high-molecular weight (HMW) DNA using the Nanobind Plant Nuclei Big DNA kit (Circulomics, Baltimore, MD, Cat # NB-900-801-01).

Oxford Nanopore PromethION sequencing

Nanopore libraries were prepared using the Oxford Nanopore SQK-LSK110 Ligation Sequencing Kit following the manufacturer's recommendation. Purified libraries were loaded onto a FLO-PRO002 (vR9.4.1) flow cell. We performed sequencing using an Oxford Nanopore Technologies' PromethION sequencer (release 21.02.7) running for 96 hours. Resulting FAST5 files were basecalled using the Nanopore Guppy basecaller software (v4.3.4). Overall, we sequenced 12,580,640 (120.74 GB) and 13,106,600 (133.36 GB) nanopore reads for LMC-L1 and SWB-S1, respectively. The median read length for LMC-L1 was 6.1 kb and 7.7 kb for SWB-S1, with a read length N50 > 19.6 kb for both. All analyses incorporated quality trimmed reads with a mean quality score > 7 and read lengths of > 300 bp for a total of 9,255,001 reads in LMC-L1 (337x coverage) and 9,423,361 reads in SWB-S1 (360x coverage). 73.6% of the total reads for LMC-L1 and 71.9% of the total reads for SWB-S1 passed this filtering step.

Illumina Sequencing

We incorporated short read sequencing for polishing the genome assembly and SNP calling. Using unopened floral buds, we used a DNeasy Plant Mini Kit (Qiagen) to extract DNA. DNA was prepared for sequencing using the Illumina TruSeq Nano DNA Library prep kit with IDT for Illumina Unique Dual Index adapters following manufacturer's recommendations. We loaded the libraries onto one lane of an Illumina NovaSeq 6000 S4 flow cell and conducted sequencing in a

2x150 bp paired end format using a NovaSeq v1.5 300 cycle reagent kit. Base calling was done by Illumina Real Time Analysis (RTA) v3.4.4 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcf2fastq v2.20.0.

Oxford Nanopore PromethION assembly, polishing, and scaffolding

Adapters were removed from the reads using Porechop v0.2.5 (https://github.com/rrwick/Porechop) and quality filtered using NanoLyse and Nanofilt (-q 0 and -l 300). We removed possible contaminants using a microbe genebank (genbank-k31.sbt.json). Denovo genome assembly was conducted with Flye v2.9.1 (87), followed by two rounds of polishing using Racon (v1.5.0) (88) and a single round of Medaka (https://github.com/nanoporetech/medaka). The genome was finished with two rounds of Pilon (89) polishing using the Illumina sequencing. We checked for misassembly using Tigment-long (90). Scaffolding into chromosomes was initially performed using Allmaps (91) by combining genetic map data (45) and synteny to the IM62, TOL, and NONTOL genomes (phytozome-next.jgi.doe.gov). Additional contigs were added to the chromosomes through an iterative process using ragtag (92) and lrscaf (93), combined with manual inspection and correction. Genome assembly statistics were calculated using assembly-stats (*v1.0.1*, https://github.com/sanger-pathogens/assembly-stats). Completeness of the genome assembly was verified several times throughout the pipeline using BUSCO v2.0 (53) and the LTR Assembly Index (LAI) pipeline (54).

Gene annotations

Genes were annotated using the MAKER pipeline (55, 56). Support for gene models was provided by transcripts assembled from RNA-seq and protein alignments. Illumina reads were trimmed for adapter sequences using trimmomatic (v0.39; (94) and aligned to their respective

genome using HiSAT2 (v2.2.1; (95)). PacBio reads were aligned by minimap2 (v2.24; (96, 97)). Transcripts were assembled by StringTie2 (v2.2.1; (98)). For protein alignments, protein sequences from *Arabidopsis thaliana* (Araport11;(99), *Oryza sativa* (v7; (100)), *Solanum lycopersicum* (ITAG4.0; (101)), and UniProtKB/Swiss-Prot plants (release-2022_02; (102)) were aligned to each genome using Exonerate protein2genome (103). Repeatmasker (v4.1.1; (104)) was used to soft-mask the genome using TE sequences identified by Extensive *de-novo* TE Annotator (EDTA, see TE Annotations). An initial round of MAKER was run using the soft-masked genome and gff files from the transcriptome assembly and protein alignment. Annotations from this initial round of MAKER were then used to train SNAP (v2013_11_29; (105) and AUGUSTUS (v3.4.0; (106) ref). To prevent overfitting, only 600 randomly sampled MAKER annotations were used to train AUGUSTUS. MAKER was then run a second time using models from SNAP and AUGUSTUS. To identify putative genes missed by MAKER in either genome, we next used Liftoff (v1.6.3; (57)) to transfer annotations from one genome onto the other. Lifted annotations were then filtered using GffCompare (v0.11.2; (107)) to remove annotations with significant overlap to existing MAKER annotations. Remaining lifted annotations with valid open reading frames (ORF; “valid_ORFs=1” or “valid_ORF=True” in the GFF attributes column) were considered as putative genes, while those without a valid ORF (“valid_ORFs=0” or “valid_ORF=False” in the GFF attributes column) were considered putative pseudogenes. This process was repeated using annotations from the reference IM62 (v2.0;(50)) genome annotation. Genes were flagged as potentially misannotated transposons as described previously (108), with the exclusion of searching of Gypsy HMM profiles, as we found that this led to too many real genes being excluded. Given the often blurry line between genes and transposons, we chose to retain these gene models in the annotation and simply flag them as possible transposons.

Gene Function

To assign gene function, genes were first searched for matches to PFAM domains using InterProScan (v5.57-90.0) (109). Next DIAMOND BLASTX (110) was used to identify the top five hits against the Arabidopsis TAIR10 proteins with an e-value cutoff of 1e-6. These results were then integrated using create_functional_annotation_file.pl (custom script from Kevin Childs located in https://github.com/niederhuth/mimulus-assembly).

Structural variant and SNP calling and annotation

We used several approaches for structural variant calling including a whole genome alignment approach and a read based alignment approach (supplemental methods). Here, we report structural variants called via the whole genome alignment approach using MUM&Co (*version 3*) (59). Pseudo-chromosome assemblies were aligned to each other, switching the reference and query, for calling SVs for each line. SVs were filtered to include inversions, deletions, insertions, and duplications larger than or equal to 50 bp with a passing quality score. Because both SWB-S1 and LMC-L1 were inbred lines, we removed heterozygous variants as those are most likely false positives. Lastly, we removed called insertions and deletions with a continuous strand of "NNN," as these segments represent gaps created by the RagTag scaffolding program and are likely false positives.

To call SNPs between the two genomes, we aligned the Illumina sequences (2x150 bp Novaseq 6000) to each of the assembled genomes. We quality trimmed forward and reverse reads using Trimmomatic v0.39 (94), trimming the reads to a minimum length of 50 bps and a quality of phred33. We used BWA-MEM2 (111) to align SWB-S1 reads to the inland annual genome (98.63% alignment rate) and LMC-L1 reads to the coastal perennial (98.3% alignment rate) genome. SNPs were called using GATK (112). We marked duplicates using PICARDTOOLS' *MarkDuplicates* function and then used GATK's *HaplotypeCaller* to call SNPs in individual

samples. We then genotyped the SNPs using GATK's *GenotypeGVCF* and filtered them using *VariantFIltration*.

SNPs were annotated using ANNOVAR (113) and filtered to generate a list of high impact mutations (frameshift insertions, frameshift deletions, stop loss, stop gain, and splicing). We also used ANNOVAR (113) to identify synonymous and nonsynonymous SNPs. Structural variants were annotated using BEDtools *intersect* (114) to identify overlapping genes and structural variants.

Copy number variants (CNVs) and presence absence variants (PAVs) were identified with a custom script that used the outputs from GENESPACE. To identify PAVs and CNVs, we included the *M. guttatus* IM62 v2 genome (50) as well as SWB-S1 and LMC-L1 genomes.

Estimating SNP-depleted areas of the genome

To determine regions of depleted SNPs, we incorporated sequencing of coastal perennial (101 accessions) and inland annual (92 accessions) pools from Gould et al (60). Each pool was sequenced across two lanes of v2 Rapid lanes on the Illumina High-Seq 2500 platform. To analyze this data with our new genomes, we followed the bioinformatics pipeline published in Gould et al (60).

Synteny, PAVs, CNVs, and inversion breakpoints

To identify a stringent set of syntenic genes between ecotypes, we implemented GENESPACE (58), a pipeline that uses both Orthofinder (115, 116) and *blastp*. We used blocks of syntenic genes with reverse orientation to identify the inversion breakpoint regions for inversions on

chromosomes 5, 8, and 14. For all of these inversions, the repetitive nature of the region around the inversion breakpoints made it challenging to define the exact nucleotide of the breakpoint mutation. While we could not identify exact breakpoints, we could clearly define which genes were closest to both sides of each breakpoint.

RNA Re-ANALYSIS

To evaluate patterns of transcript expression for inversion genes, we reanalyzed RNA-seq data sets from Gould et al (51) in the context of our newly assembled genomes. These data included RNA samples from a field reciprocal transplant experiment, where LMC-L1 and SWB-S1 individuals were grown at both inland and coastal field sites (51). The tissue collected only includes leaf tissue and thus, our analysis of differentially expressed genes is limited to leaf tissue. Sequence data were mapped with STAR (117) using a 2-pass mapping method to both the LMC-L1 and SWB-S1 reference genomes.

We used DESeq2 (118) to determine differential gene expression between genotypes and between genotypes at differing field sites. We included genotype, field site (inland or coastal), and the interaction of genotype and field site as fixed effects for each gene. We calculated a Wald test statistic per gene to find significant differences. P-values were corrected using DESeq2's default settings, which incorporates Benjamini- Hochberg corrections and FDR <= 0.1 for a gene to be identified as differentially expressed. Additionally, we applied a log2 fold change cutoff of 0.05.

qRT-PCR of *GA20ox2* (AT5G51810, Migut.H00683)

To determine whether expression differences in *GA20ox2* might contribute to phenotypic differences between ecotypes we conducted qRT-PCR because this gene is expressed at a

level too low to be analyzed with RNA-seq. We extracted RNA from three biological replicates of leaf tissue and the floral shoot apex from wild-type LMC-L1 and SWB-S1 plants using the Spectrum Plant Total RNA Kit (Sigma). cDNA was synthesized for each sample using GoScript Reverse Transcription Mix, Oligo(dT) (Promega). We performed qRT-PCR using Power SYBR Green PCR master mix (Applied Biosystems) on a CFX96 touch real-time PCR machine (Bio-Rad). The PCR program was as follows: 40 cycles at 95°C for 15 s and 60°C for 30 s. We determined amplification efficiencies for each primer pair using a dilution series (1:2, 1:4, 1:8, 1:16) of pooled cDNA samples. We used *UBC*, the ubiquitin-conjugating enzyme gene, as a reference gene to calculate the relative expression of *GA20ox2* using the formula $(E_{ref})^{CP(ref)}/(E_{target})^{CP(target)}$ (119). No differential expression was found for leaf tissue and these results are not reported (two-tailed Student's t-test, $p > 0.05$).

Data, materials, and software availability

Code for genome assembly can be found at https://github.com/niederhuth/mimulus-assembly. Raw ONT sequencing files are available through the NCBI Sequence Read Archive under project identifier PRJNA1047892 – accessions SAMN38597945 (LMC-L1) and SAMN38597946 (SWB-S1).

Acknowledgments

This study was funded by the National Science Foundation (NSF) Plant Genome Research Program 2010769 to LMK and 2109560 to LS, Michigan State University Plant Biology start-up funds to CEN, and by two National Science Foundation Division of Integrative Organismal Systems Grants to DBL (IOS-1855927 and IOS-2153100). We would like to thank Oxford Nanopore Technologies and the MSU Genomics Core for providing reagents and services

that generated sequencing data that made this work possible. We would also like to thank Kevin Childs for guidance on genome annotation, Eleanore Ritter for guidance in detecting structural variants, and Lane Vitek for lab work.

References

1. T. Dobzhansky, *Genetics and the Origin of Species (3rd ed.)* (Columbia University Press: New York, NY, USA, 1951).

2. T. Dobzhansky, *Genetics of the Evolutionary Process* (Columbia University Press, 1970).

3. M. Kirkpatrick, A. Kern, Where's the Money? Inversions, Genes, and the Hunt for Genomic Targets of Selection. *Genetics* **190**, 1153–1155 (2012).

4. A. A. Hoffmann, L. H. Rieseberg, Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).

5. J. R. Adrion, M. W. Hahn, B. S. Cooper, Revisiting classic clines in Drosophila melanogaster in the age of genomics. *Trends Genet.* **31**, 434–444 (2015).

6. R. Villoutreix, *et al.*, Inversion breakpoints and the evolution of supergenes. *Mol. Ecol.* **30**, 2738–2755 (2021).

7. M. Puig, S. Casillas, S. Villatoro, M. Cáceres, Human inversions and their functional consequences. *Brief. Funct. Genomics* **14**, 369–379 (2015).

8. S. C. Elgin, G. Reuter, Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb. Perspect. Biol. 5: a017780 (2013).

9. H. J. Muller, Types of visible variations induced by X-rays in Drosophila. *J. Genet.* **22**, 299–334 (1930).

10. T. Dobzhansky, Position effects on genes. *Biol. Rev. Camb. Philos. Soc.* **11**, 364–384 (1936).

11. J. R. Lupski, Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).

12. C. R. Marshall, *et al.*, Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).

13. K. Huang, L. H. Rieseberg, Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front. Plant Sci.* (2020).

14. L. H. Rieseberg, Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).

15. M. Joron, *et al.*, Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).

16. M. J. Thompson, C. D. Jiggins, Supergenes and their role in evolution. *Heredity* **113**, 1–8 (2014).

17. M. Kirkpatrick, N. Barton, Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).

18. M. Kirkpatrick, B. Barrett, "Chromosome inversions, adaptive cassettes and the evolution of species' ranges" in *Invasion Genetics*, (John Wiley & Sons, Ltd, 2016), pp. 175–186.

19. D. Ayala, A. Ullastres, J. González, Adaptation through chromosomal inversions in Anopheles. *Front. Genet.* **5**, 129 (2014).

20. K. Kunte, *et al.*, doublesex is a mimicry supergene. *Nature* **507**, 229–232 (2014).

21. S. Lamichhaney, *et al.*, Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* **48**, 84–88 (2016).

22. E. M. Tuttle, *et al.*, Divergence and Functional Degradation of a Sex Chromosome-like Supergene. *Curr. Biol.* **26**, 344–350 (2016).

23. H. Jeong, *et al.*, Dynamic molecular evolution of a supergene with suppressed recombination in white-throated sparrows. *Elife* **11** (2022).

24. M. Lundberg, A. Mackintosh, A. Petri, S. Bensch, Inversions maintain differences between migratory phenotypes of a songbird. *Nat. Commun.* **14**, 452 (2023).

25. E. R. Hager, *et al.*, A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science* **377**, 399–405 (2022).

26. A. R. Anderson, A. A. Hoffmann, S. W. McKechnie, P. A. Umina, A. R. Weeks, The latitudinal cline in the In(3R)Payne inversion polymorphism has shifted in the last 20 years in Australian Drosophila melanogaster populations. *Mol. Ecol.* **14**, 851–858 (2005).

27. R. V. Rane, L. Rako, M. Kapun, S. F. Lee, A. A. Hoffmann, Genomic evidence for role of inversion 3RP of Drosophila melanogaster in facilitating climate change adaptation. *Mol. Ecol.* **24**, 2423–2432 (2015).

28. Z. Fang, *et al.*, Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics* **191**, 883–894 (2012).

29. T. Pyhäjärvi, M. B. Hufford, S. Mezmouk, J. Ross-Ibarra, Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* **5**, 1594–1609 (2013).

30. Z. Zhang, H. Zhu, B. S. Gill, W. Li, Fine mapping of shattering locus Br2 reveals a putative chromosomal inversion polymorphism between the two lineages of Aegilops tauschii. *Theor. Appl. Genet.* **128**, 745–755 (2015).

31. H. Zhou, *et al.*, A 1.7-Mb chromosomal inversion downstream of a PpOFP1 gene is responsible for flat fruit shape in peach. *Plant Biotechnol. J.* **19**, 192–205 (2021).

32. T. Marquès-Bonet, *et al.*, Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees. *Trends Genet.* **20**, 524–529 (2004).

33. T. L. Newman, *et al.*, A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356 (2005).

34. A. Navarro, N. H. Barton, Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* **300**, 321–324 (2003).

35. F. J. Ayala, M. Coluzzi, Chromosome speciation: humans, Drosophila, and mosquitoes. *Proc. Natl. Acad. Sci. U. S. A.* **102 Suppl 1**, 6535–6542 (2005).

36. D. Castermans, *et al.*, Identification and characterization of the TRIP8 and REEP3 genes on chromosome 10q21.3 as novel candidate genes for autism. *Eur. J. Hum. Genet.* **15**, 422–431 (2007).

37. F. Antonacci, *et al.*, Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009).

38. M. L. Jones, *et al.*, Disruption of AP3B1 by a chromosome 5 inversion: a new disease mechanism in Hermansky-Pudlak syndrome type 2. *BMC Med. Genet.* **14**, 42 (2013).

39. G. J. Anger, *et al.*, X-linked deafness-2 (DFNX2) phenotype associated with a paracentric inversion upstream of POU3F4. *Am. J. Audiol.* **23**, 1–6 (2014).

40. F. A. M. Maggiolini, *et al.*, Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.* **30**, 1680–1693 (2020).

41. D. Porubsky, *et al.*, Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).

42. M. J. Vogel, *et al.*, High-resolution mapping of heterochromatin redistribution in a Drosophila position-effect variegation model. *Epigenetics Chromatin* **2**, 1 (2009).

43. M. Berloco, G. Palumbo, L. Piacentini, S. Pimpinelli, L. Fanti, Position effect variegation and viability are both sensitive to dosage of constitutive heterochromatin in Drosophila. *G3* **4**, 1709–1716 (2014).

44. B. Burssed, M. Zamariolli, F. T. Bellucco, M. I. Melaragno, Mechanisms of structural chromosomal rearrangement formation. *Mol. Cytogenet.* **15**, 23 (2022).

45. D. B. Lowry, J. H. Willis, A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8** (2010).

46. D. B. Lowry, D. Popovic, D. J. Brennan, L. M. Holeski, Mechanisms of a locally adaptive shift in allocation among growth, reproduction, and herbivore resistance in Mimulus guttatus. *Evolution* **73**, 1168–1181 (2019).

47. J. M. Coughlan, J. H. Willis, Dissecting the role of a large chromosomal inversion in life history divergence throughout the Mimulus guttatus species complex. *Mol. Ecol.* **28**, 1343–

1357 (2019).

48. J. Friedman, Genetic determinants and epistasis for life history trait differences in the common monkeyflower, Mimulus guttatus. *J. Hered.* **105**, 816–827 (2014).

49. A. D. Twyford, J. Friedman, Adaptive divergence in the monkey flower Mimulus guttatus is maintained by a chromosomal inversion. *Evolution* **69**, 1476–1486 (2015).

50. U. Hellsten, *et al.*, Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19478–19482 (2013).

51. B. A. Gould, Y. Chen, D. B. Lowry, Gene regulatory divergence between locally adapted ecotypes in their native habitats. *Mol. Ecol.* **27**, 4174–4188 (2018).

52. M. Todesco, *et al.*, Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).

53. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

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

55. B. L. Cantarel, *et al.*, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).

56. C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

57. A. Shumate, S. L. Salzberg, Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2020).

58. J. T. Lovell, *et al.*, GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11** (2022).

59. S. O'Donnell, G. Fischer, MUM&Co: accurate detection of all SV types through whole-genome alignment. *Bioinformatics* **36**, 3242–3243 (2020).

60. B. A. Gould, Y. Chen, D. B. Lowry, Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Mol. Ecol.* (2017) https:/doi.org/10.1111/mec.13881.

61. J. M. Coughlan, M. W. Brown, J. H. Willis, The genetic architecture and evolution of life-history divergence among perennials in the Mimulus guttatus species complex. *Proc. Biol. Sci.* **288**, 20210077 (2021).

62. Y. Brandvain, A. M. Kenney, L. Flagel, G. Coop, A. L. Sweigart, Speciation and Introgression between Mimulus nasutus and Mimulus guttatus. *PLoS Genet.* (2014) https:/doi.org/10.1371/journal.pgen.1004410.

63. A. M. Cooley, J. L. Modliszewski, M. L. Rommel, J. H. Willis, Gene duplication in Mimulus underlies parallel floral evolution via independent trans-regulatory changes. *Curr. Biol.* **21**,

700–704 (2011).

64. D. B. Lowry, C. C. Sheng, J. R. Lasky, J. H. Willis, Five anthocyanin polymorphisms are associated with an R2R3-MYB cluster in Mimulus guttatus (Phrymaceae). *Am. J. Bot.* **99**, 82–91 (2012).

65. L. M. Holeski, *et al.*, A high-resolution genetic map of yellow monkeyflower identifies chemical defense QTLs and recombination rate variation. *G3* **4**, 813–821 (2014).

66. L. E. Flagel, *et al.*, GOOGA: A platform to synthesize mapping experiments and identify genomic structural diversity. *PLoS Comput. Biol.* **15**, e1006949 (2019).

67. E. L. Berdan, A. Blanckaert, R. K. Butlin, C. Bank, Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS Genet.* **17**, e1009411 (2021).

68. F. Andrés, *et al.*, SHORT VEGETATIVE PHASE reduces gibberellin biosynthesis at the Arabidopsis shoot apex to regulate the floral transition. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2760–9 (2014).

69. D. W. Lee, J. B. Lowry, B. C. Stone, Abaxial Anthocyanin Layer in Leaves of Tropical Rain Forest Plants: Enhancer of Light Capture in Deep Shade. *Biotropica* **11**, 70–77 (1979).

70. N. M. Hughes, K. L. Carpenter, J. G. Cannon, Estimating contribution of anthocyanin pigments to osmotic adjustment during winter leaf reddening. *J. Plant Physiol.* **170**, 230–233 (2013).

71. H. M. Schaefer, G. Rolshausen, Plants on red alert: do insects pay attention? *Bioessays* **28**, 65–71 (2006).

72. Y. Manetas, Why some leaves are anthocyanic and why most anthocyanic leaves are red? *Flora - Morphology, Distribution, Functional Ecology of Plants* **201**, 163–177 (2006).

73. J.-H. B. Hatier, K. S. Gould, Foliar anthocyanins as modulators of stress signals. *J. Theor. Biol.* **253**, 625–627 (2008).

74. N. M. Hughes, S. Lev-Yadun, Review: Why do some plants have leaves with red or purple undersides? *Environ. Exp. Bot.* **205**, 105126 (2023).

75. M. Liang, *et al.*, Taxon-specific, phased siRNAs underlie a speciation locus in monkeyflowers. *Science* **379**, 576–582 (2023).

76. B. Charlesworth, N. H. Barton, The spread of an inversion with migration and selection. *Genetics* **208**, 377–382 (2018).

77. T. Connallon, C. Olito, Natural selection and the distribution of chromosomal inversion lengths. *Mol. Ecol.* **31**, 3627–3641 (2022).

78. L.-J. L. Handley, H. Ceplitis, H. Ellegren, Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics* **167**, 367–376 (2004).

79. C. Olito, J. K. Abbott, The evolution of suppressed recombination between sex chromosomes and the lengths of evolutionary strata. *Evolution* **77**, 1077–1090 (2023).

80. D. Charlesworth, Why and how do Y chromosome stop recombining? *J. Evol. Biol.* **36**, 632–636 (2023).

81. D. A. Filatov, Recent expansion of the non-recombining sex-linked region on Silene latifolia sex chromosomes. *J. Evol. Biol.* **35**, 1696–1708 (2022).

82. R. Bergero, D. Charlesworth, The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* **24**, 94–102 (2009).

83. R. J. Mroczek, J. R. Melo, A. C. Luce, E. N. Hiatt, R. K. Dawe, The maize Ab10 meiotic drive system maps to supernumerary sequences in a large complex haplotype. *Genetics* **174**, 145–154 (2006).

84. R. K. Dawe, The maize abnormal chromosome 10 meiotic drive haplotype: a review. *Chromosome Res.* **30**, 205–216 (2022).

85. E. Oneal, D. B. Lowry, K. M. Wright, Z. Zhu, J. H. Willis, Divergent population structure and climate associations of a chromosomal inversion polymorphism across the Mimulus guttatus species complex. *Mol. Ecol.* **23**, 2844–2860 (2014).

86. Z. Lu, B. T. Hofmeister, C. Vollmers, R. M. DuBois, R. J. Schmitz, Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res.* **45**, e41–e41 (2017).

87. M. Kolmogorov, *et al.*, metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

88. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

89. B. J. Walker, *et al.*, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

90. S. D. Jackman, *et al.*, Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19**, 393 (2018).

91. H. Tang, *et al.*, ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).

92. M. Alonge, *et al.*, Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).

93. M. Qin, *et al.*, LRScaf: improving draft genomes using long noisy reads. *BMC Genomics* **20**, 955 (2019).

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

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

96. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

97. *Mimulus guttatusvar. IM767* v1.1" DOE-JGI, https://phytozome.jgi.doe.gov/info/Mguttatusvar_IM767_v1_1

98. S. Kovaka, *et al.*, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

99. C.-Y. Cheng, *et al.*, Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* **89**, 789–804 (2017).

100. Y. Kawahara, *et al.*, Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).

101. P. S. Hosmani, *et al.*, An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv*, 767764 (2019).

102. M. Schneider, A. Bairoch, C. H. Wu, R. Apweiler, Plant protein annotation in the UniProt Knowledgebase. *Plant Physiol.* **138**, 59–66 (2005).

103. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

104. , RepeatMasker Home Page (March 13, 2023).

105. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

106. M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–7 (2005).

107. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Res.* **9** (2020).

108. M. J. Bowman, J. A. Pulman, T. L. Liu, K. L. Childs, A modified GC-specific MAKER gene annotation method reveals improved and novel gene predictions of high and low GC content in Oryza sativa. *BMC Bioinformatics* **18**, 522 (2017).

109. P. Jones, *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

110. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

111. M. Vasimuddin, S. Misra, H. Li, S. Aluru, Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, (2019), pp. 314–324.

112. A. McKenna, *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

113. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

114. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

115. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

116. D. M. Emms, Kelly S. 2015a. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*

117. A. Dobin, *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

118. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

119. M. W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45 (2001).

Figures and Tables (In text)

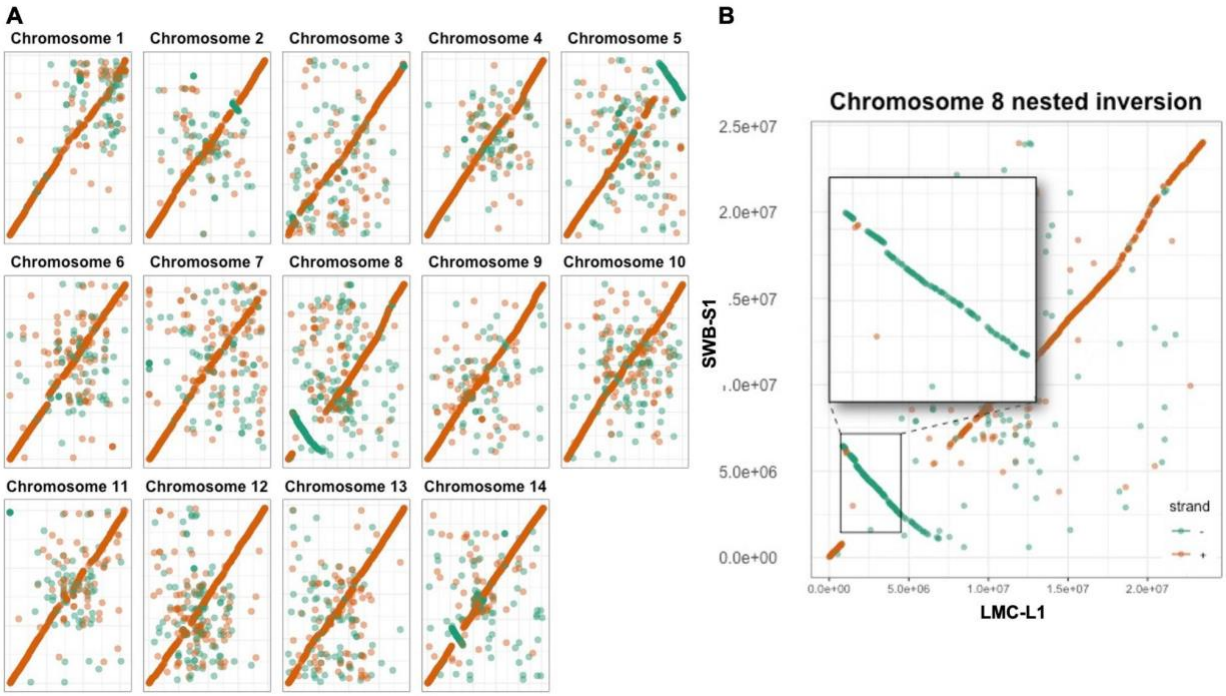

Figure 1. A. Whole genome alignment of all 14 chromosomes between the SWB-S1 and LMC-L1 genomes. Large inversions on chromosomes 5, 8, and 14 are visible in the dotplots, where the teal dots represent the negatively oriented strands while orange represent positively oriented strands. **B.** The inv_chr8B captured within inv_chr8A.

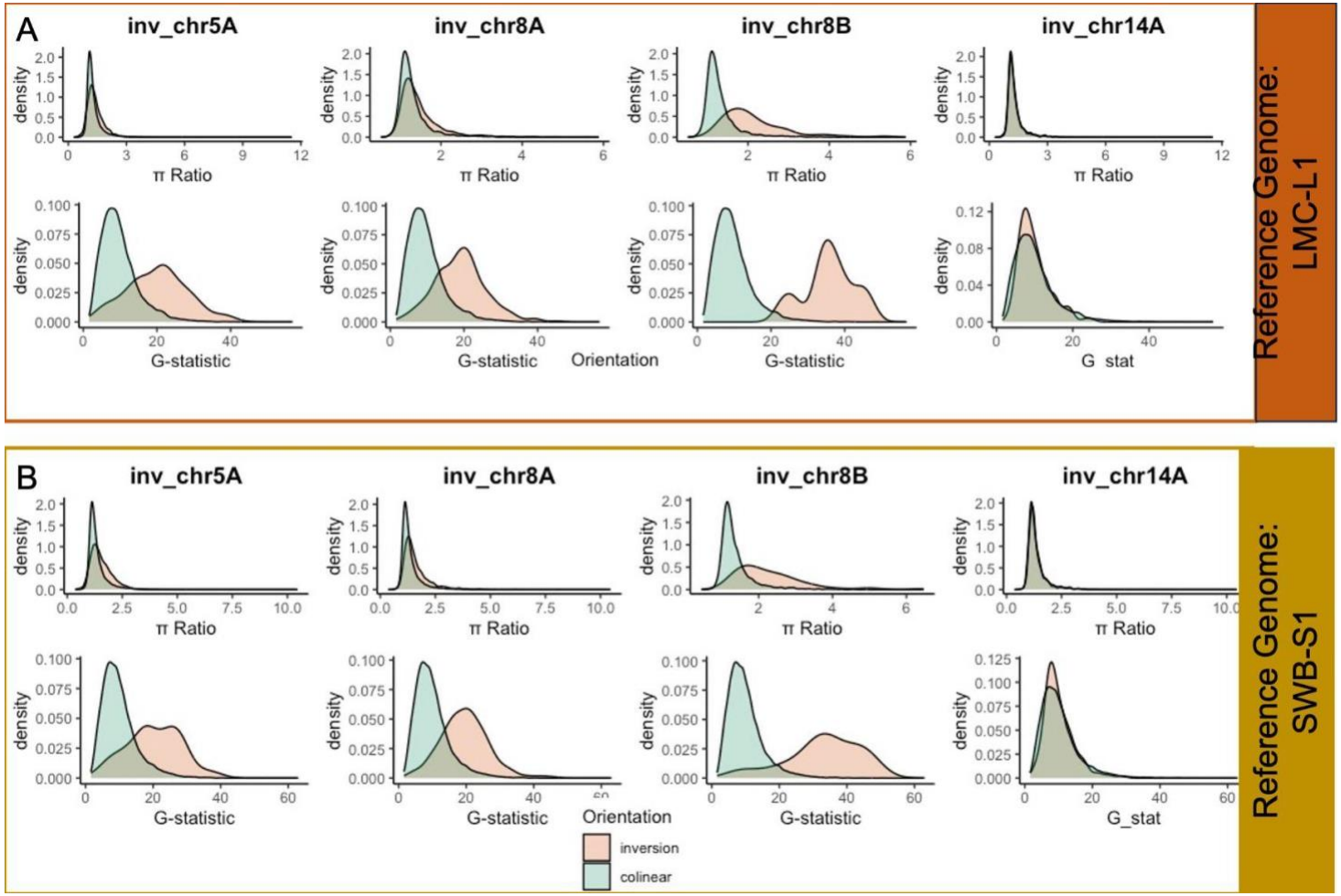

Figure 2. *G* statistic and π ratio for each of the inversions (salmon) compared to the non-inverted (teal) sequence of the same chromosome when aligned to the LMC-L1 Genome **A** and to the SWB-S1 genome **B**.

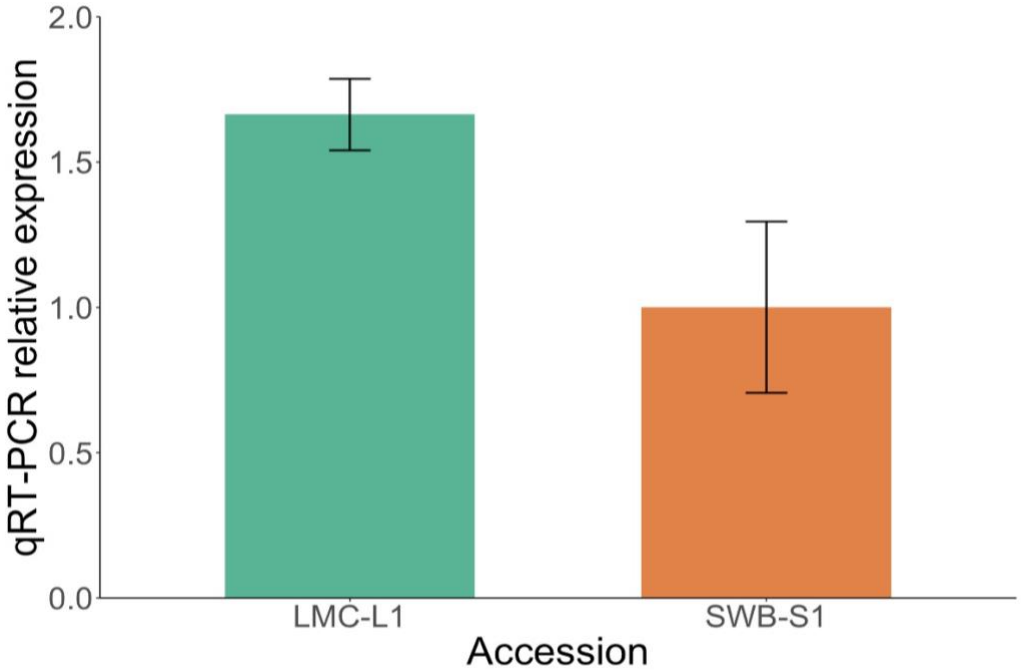

Figure 3. Quantitative RT-PCR showing expression of *GA20ox2* in wild-type LMC-L1 and SWB-S1 floral shoot apex (three biological replicates). Error bars show standard deviation.

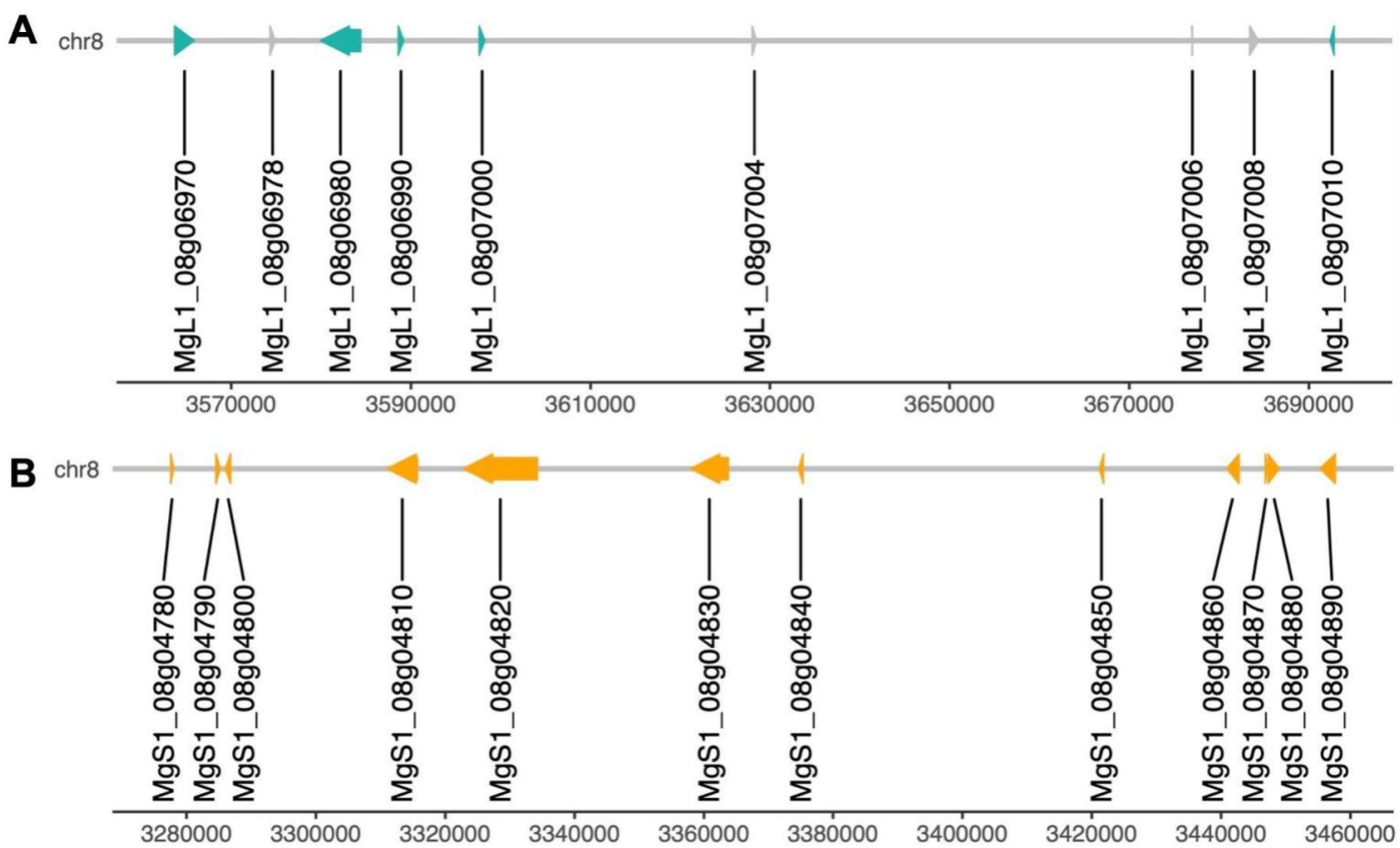

Figure 4. The MYB region along chromosome 8 in LMC-L1 (**A**) and SWB-S1 (**B**). Each arrow (teal in LMC-L1 and orange in SWB-S1) represents a gene and orientation. The gray arrows represent pseudogenes.

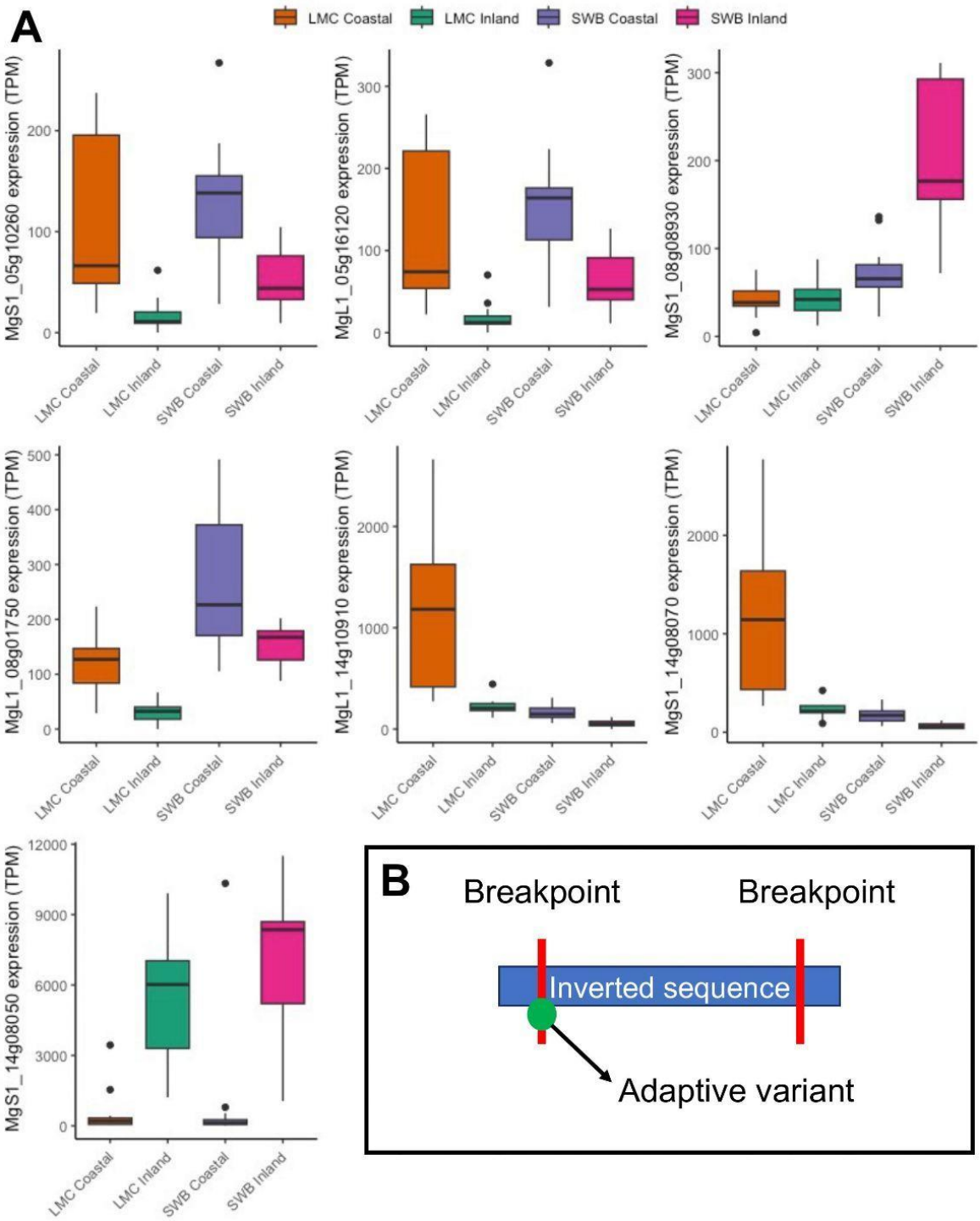

Figure 5. A. Differential expression represented in transcripts per million (TPM) in genes at the inversion breakpoints. Note, some plots are nearly identical because they are LMC-L1 (labeled LMC) and SWB-S1 (labeled SWB) orthologs (MgL14g10910 and MgS1_14g08070, MgS1_05g10260 and MgL1_05g16120). **B.** A diagram depicting the inversion breakpoint hypothesis.