Fundamental Patterns of Structural Evolution Revealed by Chromosome-Length Genomes of Cactophilic *Drosophila*

Kyle M. Benowitz [1,4,*], Carson W. Allan [1], Coline C. Jaworski [1,5,6], Michael J. Sanderson [2], Fernando Diaz [1,7], Xingsen Chen [1], Luciano M. Matzkin [1,2,3,*]

[1]Department of Entomology, University of Arizona, Tucson, AZ, USA

[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

[3]BIO5 Institute, University of Arizona, Tucson, AZ, USA

[4]Present address: College of Integrative Sciences and Arts, Arizona State University, Mesa, AZ, USA

[5]Present address: Department of Zoology, Cambridge University, Cambridge, UK

[6]Present address: INRAE, CNRS, UMR ISA, Université Côte d'Azur, Nice, France

[7]Present address: Department of Life, Earth and Environmental Sciences, West Texas A&M University, Canyon, TX, USA

*Corresponding authors: E-mails: benowitz@asu.edu; lmatzkin@arizona.edu.

Accepted: August 26, 2024

Abstract

A thorough understanding of adaptation and speciation requires model organisms with both a history of ecological and phenotypic study as well as a complete set of genomic resources. In particular, high-quality genome assemblies of ecological model organisms are needed to assess the evolution of genome structure and its role in adaptation and speciation. Here, we generate new genomes of cactophilic *Drosophila*, a crucial model clade for understanding speciation and ecological adaptation in xeric environments. We generated chromosome-level genome assemblies and complete annotations for seven populations across *Drosophila mojavensis*, *Drosophila arizonae*, and *Drosophila navojoa*. We use these data first to establish the most robust phylogeny for this clade to date, and to assess patterns of molecular evolution across the phylogeny, showing concordance with a priori hypotheses regarding adaptive genes in this system. We then show that structural evolution occurs at constant rate across the phylogeny, varies by chromosome, and is correlated with molecular evolution. These results advance the understanding of the *D. mojavensis* clade by demonstrating core evolutionary genetic patterns and integrating those patterns to generate new gene-level hypotheses regarding adaptation. Our data are presented in a new public database (cactusflybase.arizona.edu), providing one of the most in-depth resources for the analysis of inter- and intraspecific evolutionary genomic data. Furthermore, we anticipate that the patterns of structural evolution identified here will serve as a baseline for future comparative studies to identify the factors that influence the evolution of genome structure across taxa.

Key words: cactophilic *Drosophila*, inversion, molecular evolution, phylogenomics, genome collinearity.

Significance

Although evidence is accumulating for the importance of structural mutations (inversions, duplications, transpositions, etc.) in adaptation and speciation, rates and patterns of structural evolution have not been elucidated as they have for nucleotide mutations or gene expression differences. Here, we use new contiguous genome assemblies from three desert-adapted *Drosophila* species to demonstrate that rates of structural evolution: (i) accumulate at constant rates over evolutionary time; (ii) are chromosome-specific; (iii) are related to protein-coding evolution; and (iv) vary between different clades of *Drosophila*. These results represent a key step in asking fundamental questions about the factors that influence the evolution of structural genetic variants both within and between taxa.

Introduction

The fundamental goal of evolutionary genetics is to link phenotypic adaptation to genomic variation (Lewontin 1974). Importantly, the causality of this link, for practical purposes, can be viewed as bidirectional. It is essential to use genomic approaches to ascribe genetic underpinnings to previously identified adaptive phenotypes. Such phenotype-to-genotype approaches are needed to answer fundamental questions regarding the type and number of genes underlying adaptation and the predictability of these processes, among others (Orr 2005; Barrett and Hoekstra 2011). On the other hand, it is equally as necessary to draw conclusions a posteriori from genomic comparisons to generate hypotheses about understudied phenotypes that may be contributing to ecological adaptation and speciation (Benowitz et al. 2020). With a genotype-to-phenotype approach, genomic data may be repurposed to provide further insights in studies of natural history (Holmes et al. 2016; Sherman et al. 2016). To serve these purposes, the number of sequenced genomes from non-model organisms has been increasing at a rapid rate.

Although all genomes are valuable, it is clear that genome assemblies must be of high quality to best contribute to the goal of connecting genotype to adaptation. The fragmented, short-read assemblies that have been common for most non-model organisms have been extremely useful in facilitating gene expression studies, studies of molecular evolution, and studies of gene family evolution. However, these assemblies do not allow for the accurate quantification of most aspects of structural variation and evolution. Evidence is mounting that structural variants, including gene duplications (Ohno 1970), large chromosomal inversions (Noor et al. 2001; Kirkpatrick and Barton 2006; Feder and Nosil 2009; Faria et al. 2019; Hager et al. 2022; Harringmeyer and Hoekstra 2022; Berdan et al. 2023), transposable element mutations (Casacuberta and González 2013; Schrader and Schmitz 2019), chromosomal fusions (Wellband et al. 2019; Liu et al. 2022), and small structural variants (Mérot et al. 2020; Zhang et al. 2021) can all be involved in adaptation and speciation. Thus, an increased focus on producing both highly contiguous genome assemblies (e.g. Hotaling et al. 2021; Kim et al. 2021; Rhie et al. 2021) and methods to detect structural variants (Corbett-Detig et al. 2012; Chakraborty et al. 2018; Wala et al. 2018; Goel et al. 2019; Heller and Vingron 2019; O'Donnell and Fischer 2020) are merited.

Here, we provide these genomic resources for a well-studied non-model system, the flies of the *Drosophila mojavensis* species cluster (Heed 1978, 1982). This system is a priority for increased sequencing effort because of the rich base of ecological knowledge that has accumulated over several decades. The *D. mojavensis* species cluster are cactophilic flies within the mulleri complex of the repleta group. Cactophilic flies have adapted to living in xeric environments by making a habitat of necrotic cactus tissue, where larvae develop and all life stages feed on yeasts (Fogleman et al. 1981, 1982) and bacteria (Fogleman and Foster 1989) proliferating in the necrosis, which is highly toxic (Kircher 1982; Fogleman and Heed 1989; Fogleman and Danielson 2001). Thus, cactophilic *Drosophila* present an excellent system for ecological adaptation both to novel chemical and nutritional environments in addition to hot and dry environments.

In addition to the novel colonization of their niche, there has also been extensive ecological divergence within the cactophilic group. The best studied of these is the *D. mojavensis* cluster, consisting of the three species *D. mojavensis*, *Drosophila arizonae*, and *Drosophila navojoa* (Matzkin 2014). This clade, which has diversified within the last few million years (Russo et al. 1995; Matzkin and Eanes 2003; Reed et al. 2007; Smith et al. 2012), inhabits a range of cactus hosts and habitat types (Matzkin 2014). Within *D. mojavensis*, there are four geographically and genetically distinct populations that largely (but not exclusively) inhabit single, distinct host cacti species (Matzkin 2014; Etges 2019): one in the Sonoran Desert (SON) inhabiting organ pipe cactus (*Stenocereus thurberi*), one in Baja California (BC) inhabiting agria (*Stenocereus gummosus*), one in the Mojave Desert (MOV) inhabiting red barrel cactus (*Ferocactus cylindraceus*), and one on Santa Catalina Island (CI) inhabiting prickly pear (*Opuntia littoralis*). Its sibling species, *D. arizonae*, is a generalist, inhabiting multiple cactus species within its range from Guatemala to southern

California (Fellows and Heed 1972; Heed 1978, 1982). The outgroup, *D. navojoa* from central Mexico, is a specialist on prickly pear (*Opuntia wilcoxii*; Heed 1982). These distinctions within and between species have formed the basis for testing many hypotheses regarding phenotypic adaptation both to the specific cactus host environment as well as the broader abiotic environment (reviewed in Matzkin 2014). Additionally, the recent divergence within and between species in the *D. mojavensis* species complex has made this clade into a fruitful system for speciation and the evolution of reproductive incompatibilities (reviewed in Mullen and Shaw 2014).

The close phylogenetic relationship to *D. melanogaster* has provided the *D. mojavensis* cluster with several advantages as a burgeoning genomic model system. The CI *D. mojavensis* population was among the first non-model *Drosophila* genomes sequenced (*Drosophila* 12 Genomes Consortium 2007; Gilbert 2007), giving the species of the *D. mojavensis* cluster a high-quality starting point and a template for further research. Additionally, the wealth of functional genomic knowledge in *Drosophila melanogaster* has allowed for clear interpretation of gene-level results as compared to more distantly related insects. This has been leveraged in many candidate gene studies (reviewed in Matzkin 2014) whole-genome studies of molecular evolution (e.g. Allan and Matzkin 2019; Guillén et al. 2019), transcriptomics (reviewed in Etges 2019), genetic mapping studies (e.g. Etges et al. 2007; Benowitz et al. 2019), and functional analysis via CRISPR derived transgenics (Khallaf et al. 2020; Wang et al. 2023).

Despite this extensive history of genomic research, the data needed to address many key hypotheses within this system remains unavailable. At present, there is only a de novo sequenced genome for one of the four *D. mojavensis* populations (*Drosophila* 12 Genomes Consortium 2007), in spite of the outsized role that these populations have played in understanding molecular adaptation to variable host environments. Outside of *D. mojavensis*, there are currently only two highly fragmented genome assemblies, one each from *D. navojoa* and *D. arizonae* (Sanchez-Flores et al. 2016; Vanderlinde et al. 2019).

Here, we take a major step toward addressing this gap and lack of genomic resources by re-scaffolding the best current assembly (CI) and generating de novo chromosome-level assemblies for the remaining three specialist populations of *D. mojavensis*, two generalist populations of *D. arizonae*, and one population of *D. navojoa* (Fig. 1; Table 1). We first use these assemblies to resolve longstanding questions regarding the phylogeny and divergence times within this group. We then assess protein-coding and structural evolution across all seven genomes. This provides novel insight into the rates of each type of evolutionary divergence in this clade, and also facilitates the power to test fundamental hypotheses on the relationship between structural and coding evolution. Lastly, in order to enable the use of these genomes as a resource for the communities of *Drosophila* biologists and ecological geneticists, we present a public database of the assemblies and annotations (cactusflybase.arizona.edu).

Results

Genome Assembly and Annotation

Our hybrid approach, which combined short and long-read sequencing data, produced highly contiguous genome assemblies for all six newly sequenced fly strains. These assemblies were nearly gap-free and contained all six expected Muller elements in single scaffolds (Table 2, Fig. 2). Each assembly also contained complete genomic content as indicated by BUSCO scores of over 99% (Table 2). Genome size was consistent across the three new *D. mojavensis* genomes (159 to 161 Mb), with both *D. arizonae* exhibiting slightly larger genomes (162 to 163 Mb) and *D. navojoa* a slightly smaller genome (156 Mb; Table 2). The re-scaffolding of the original CI genome also resulted in the merging of several scaffolds into full chromosomes, which closely match those of the de novo assemblies. However, this assembly still has a higher percentage of gaps as well as repeats, which likely indicates the presence of redundant scaffolds that could not be placed on a chromosome. This likely also explains the larger genome size. The reassembly of the CI genome also resulted in the relocation of approximately 1.3 MB of sequence from Muller element F to A.

Our genome assemblies confirmed previous findings (Ruiz et al. 1990; Delprat et al. 2019) on fixed chromosomal inversions between these species and populations (Fig. 2; supplementary figs. S1 to S3, Supplementary Material online), with a single inversion occurring at the base of *D. mojavensis* on Muller element A (X chromosome), and multiple overlapping inversions on Muller elements B and E.

To facilitate further study of these species, we have deposited the assemblies and annotations in a new public database at cactusflybase.arizona.edu. Users can download fasta and gff files directly, view annotations and underlying RNA-seq data via JBrowse (Buels et al. 2016), and BLAST the genome and proteome databases using SequenceServer (Priyam et al. 2019). Details of the species and populations sequenced and their husbandry are available as well. This database will be maintained and updated by the Matzkin Lab at the University of Arizona.

Phylogenomics and Divergence Time Estimation

We estimated phylogenies using RaxML and ASTRAL-III and divergence times using BPP. We used these methods in parallel for one dataset containing 13 mitochondrial genes and another containing 12,218 single-copy nuclear genes to provide comparisons with the many existing phylogenetic

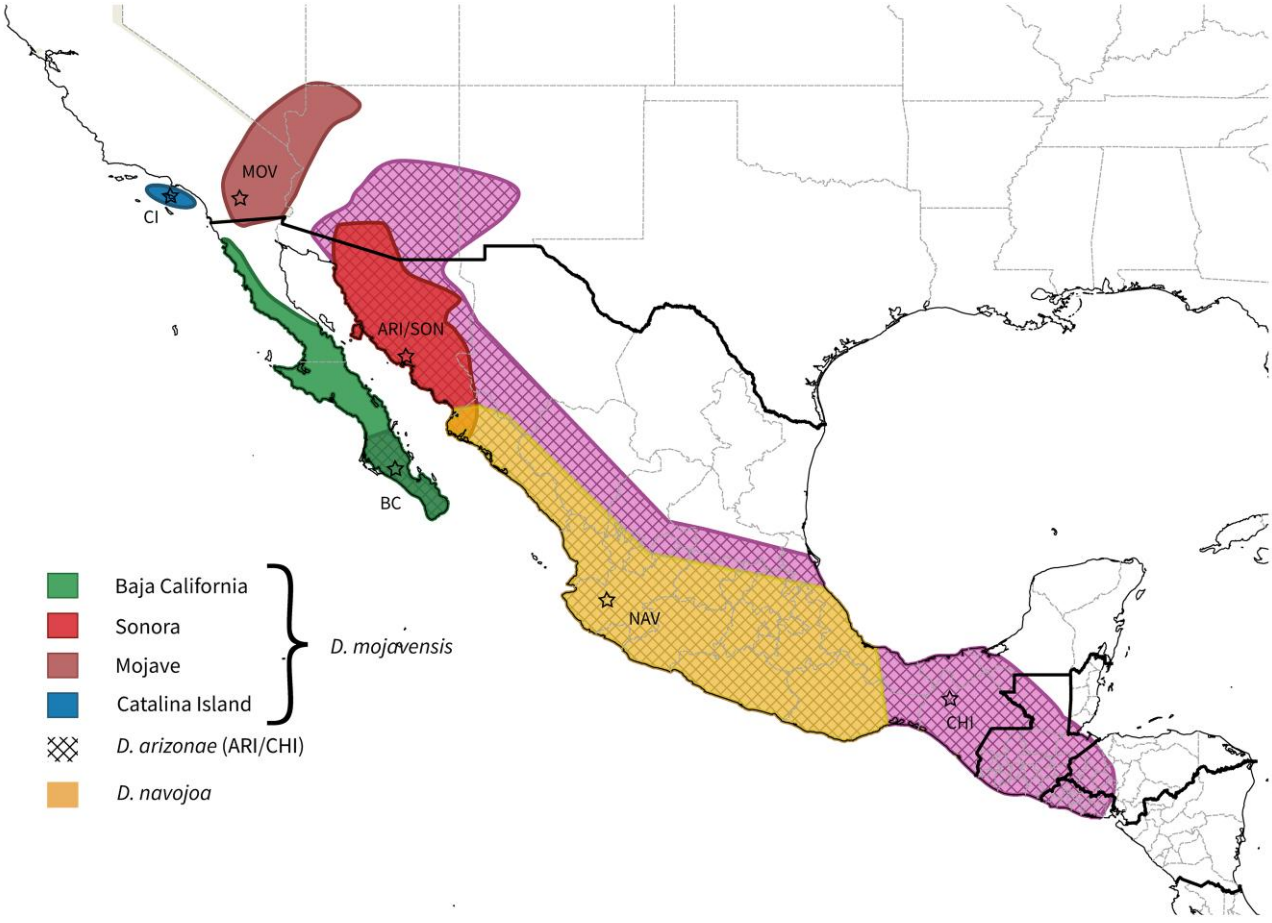

Fig. 1. Ranges of the species and populations sequenced in this study. Hatched regions represent the range of *D. arizonae*. No discrete geographical boundary is known to separate the ARI and CHI populations sequenced here. Stars show the location of collection of the genome lines. Ranges are estimated based on collection site and host plant ranges.

Table 1 Information on stocks, from the National *Drosophila* Species Stock Center (Cornell) used for genome sequencing in this study

Species	Population abbreviation	Location of collection	Date of collection	Stock center ID	Local ID
D. mojavensis	BC	La Paz, Baja California Mexico	2001	15081-1354.01	MJBC 155
	CI	Santa Catalina Island, California, USA	2002	15081-1352.22	15081-1352.22
	MOV	Anza-Borrego State Desert Park, California, USA	2002	15081-1353.01	MJANZA 402-8
	SON	Guaymas, Sonora Mexico	1998	15081-1355.01	MJ 122
D. arizonae	ARI	Guaymas, Sonora Mexico	2004	15081-1271.41	AR002
	CHI	Chiapas, Mexico	1987	15081-1271.14	AZ Chiapas 1B 13610
D. navojoa	NAV	Jalisco, Mexico	1997	15081-1374.11	15081-1374.11

studies in this group. Although here we only use single-gene trees to develop a species phylogeny, future analysis of discordance amongst these gene trees could reveal shared evolutionary patterns of genes within inversions or near breakpoints.

Both the topology of the species phylogeny as well as the divergence time estimates differed when using nuclear (Fig. 3) versus mitochondrial datasets (supplementary fig. S4, Supplementary Material online). The nuclear derived phylogeny placed the four *D. mojavensis* populations in a single clade, with the two *D. arizonae* populations as a sibling clade, and *D. navojoa* as an outgroup. ASTRAL-III gave posterior probabilities of 1 for each node in this phylogeny. While the mitochondrial phylogeny also had *D. navojoa* as an outgroup, it included the northern *D. arizonae* population as part of the *D. mojavensis* clade, with the Chiapas population an outgroup to that clade. Both phylogenies agreed that the BC and SON *D. mojavensis* populations

Table 2 Genome assembly statistics

...	CI	MOV	BC	SON	ARI	CHI	NAV
Genome size (Mb)	191.84	160.64	161.282	158.92	163.52	162.67	156.70
# of scaffolds	6,327	69	68	42	45	39	68
Scaffold N50 length (Mb)	32.37	32.40	32.47	32.28	33.82	33.67	31.27
# of contigs	10,611	71	69	46	51	43	69
Contig N50 length (Mb)	0.041	27.01	27.38	26.92	27.42	27.17	27.05
Gaps (%)	6.390	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
GC content (%)	39.48	39.66	39.67	39.64	39.7	39.65	39.95
Repeat content (%)	29.27	25.19	25.48	24.5	26.48	26.23	23.79
Number of proteins	13,675	13,295	13,327	13,364	13,342	13,264	13,144
Genome BUSCO (%)							
Complete	99.0	99.1	99.0	99.1	99.1	99.1	99.2
Single copy	98.6	98.8	98.6	98.8	98.8	98.8	98.8
Duplicated	0.4	0.3	0.4	0.3	0.3	0.3	0.4
Fragmented	0.4	0.3	0.5	0.4	0.2	0.3	0.4
Missing	0.6	0.6	0.5	0.5	0.7	0.6	0.4
Proteome BUSCO (%)							
Complete	99.7	99.9	99.8	99.9	99.8	99.9	99.9
Single copy	99.3	99.5	99.5	99.5	99.3	99.5	99.5
Duplicated	0.4	0.4	0.3	0.4	0.5	0.4	0.4
Fragmented	0.2	0.1	0.2	0.1	0.1	0.1	0.1
Missing	0.1	0	0.1	0	0.1	0	0

were most closely related, although other aspects of the topology within *D. mojavensis* also differed.

The timing of divergence within the *D. arizonae*/*D. mojavensis* clade was similar between the two datasets, at around 0.8 mya. However, the divergence between *D. navojoa* and the *D. arizonae*/*D. mojavensis* clade was estimated to be about twice as old in the mitochondrial phylogeny (3.93 mya) than the nuclear phylogeny (1.96 mya). Initial divergence within *D. mojavensis* was also estimated to be older from the mitochondrial data (0.46 mya) than from the nuclear data (0.24 mya).

Structural Evolution

We defined collinear regions of the genome as those displaying one-to-one conservation of sequence as called by SyRI (Goel et al. 2019). This software considers noncollinear regions as resulting from medium-sized indels, highly diverged regions, duplications, translocations, and inversions. Variants of less than 100 bp were not considered in the calculation of collinearity. We defined structural divergence as the percentage of noncollinear genome content between two genomes in a given region. Structural divergence of each of the six de novo sequenced populations from CI recapitulated the sequence divergence patterns as found in the nuclear phylogeny (Fig. 4g; supplementary fig. S5, Supplementary Material online). As the most diverged species, *D. navojoa* predictably had by far the greatest mean structural divergence, while both *D. arizonae* populations had nearly identical levels of divergence. MOV had slightly higher collinearity with CI compared to BC and SON. Overall, the divergence of structure over evolutionary time was found to be linear, with a loss of roughly 33.67% of collinearity per million years (supplementary fig. S5, Supplementary Material online). Structural divergence estimated using the same methodology in the *D. melanogaster* clade was also linear but slightly slower relative to sequence divergence (supplementary fig. S5, Supplementary Material online), suggesting that different taxa either accumulate or maintain structural mutations at different frequencies.

Independently of chromosomal inversions, which were rearranged in all seven genomes to match the CI karyotype prior to analysis, significant variation in structural divergence was present between chromosomes. Muller elements A and F showed reduced collinearity in all six genomes. In *D. arizonae*, Muller elements B and E, which also carry inversions, displayed lower collinearity than C and D, which do not carry inversions. Patterns in *D. navojoa* were similar apart from a reduction in collinearity on Muller element C compared to D (Fig. 4a to g).

Within chromosomes bearing inversions, the relative rates of structural divergence inside and outside (measured here as only the region on the centromeric side of the inversion due to low collinearity near telomeres) the inversion depended on the evolutionary distance and specific chromosome. Within *D. mojavensis*, Muller element B displayed greater divergence prior to the chromosomal inversion breakpoint (Fig. 4a to f) in all three populations, including the SON population, which is homokaryotypic with CI (Fig. 4h). However, collinearity in Muller element E was consistent before and within

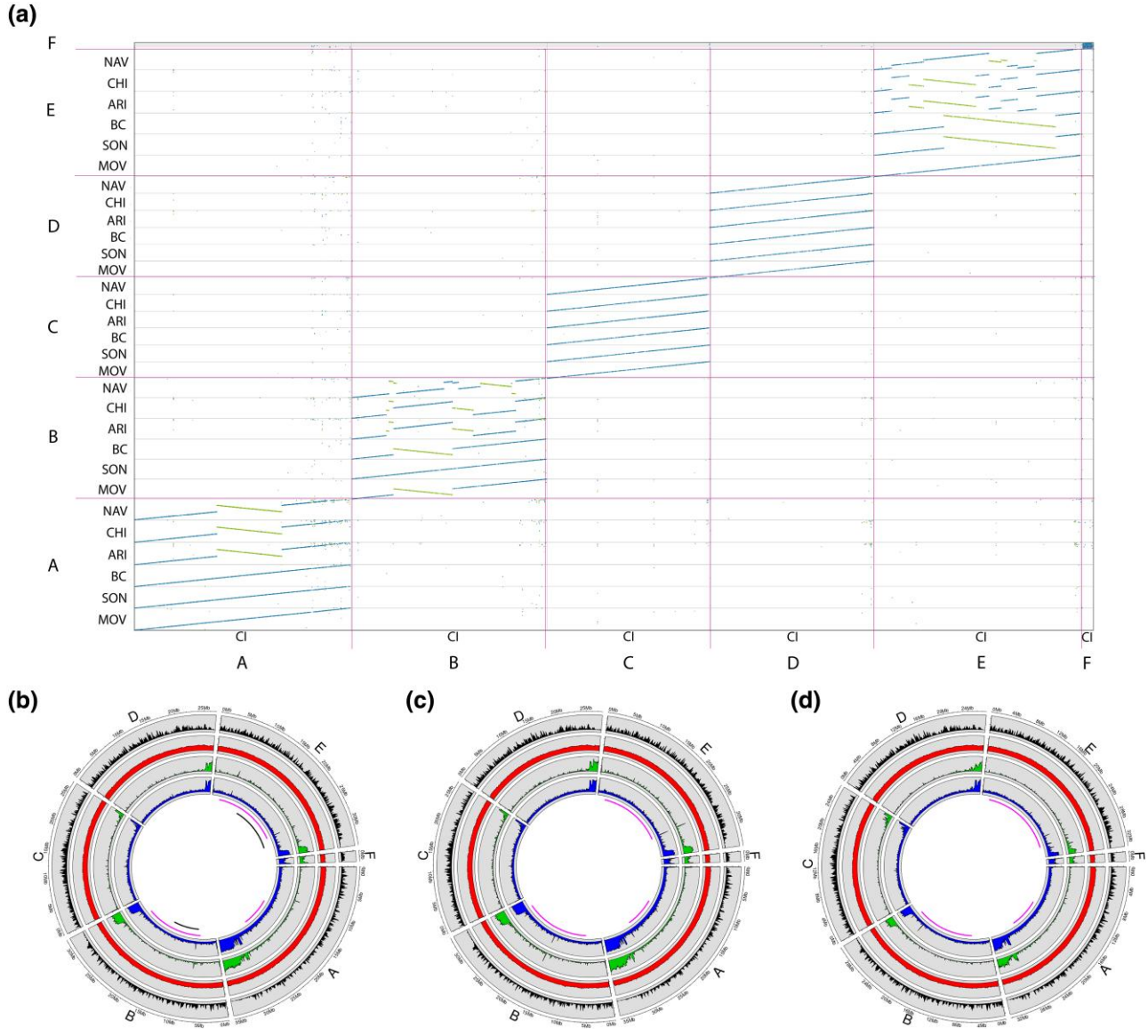

Fig. 2. Structure and composition of new genome assemblies. a) Chromosomal alignment of the six de novo genomes sequenced in this study as compared to the reassembled CI genome. Letters A to F indicate Muller elements. Green lines indicate major inversions. b to d) Genome statistics for: b) Mojave *D. mojavensis*, c) Chiapas *D. arizonae*, and d) *D. navojoa*. From outside to in, circles represent gene content, GC content, TE content, and total repeat content. Pink bars below the circles represent the regions covered by interspecific inversion polymorphisms, and black bars represent regions covered by inversion polymorphisms within *D. mojavensis*.

the inversion. Interspecific structural divergence, on the other hand, was greater within the inversion regions on Muller elements A and B, while the opposite was true on E (Fig. 4i).

Molecular Evolution

We analyzed molecular evolution of all 12,218 single-copy genes using BUSTED2 and codeml. We first examined evidence for elevated rates of positive selection amongst functional gene categories which have been previously associated with ecological adaptation or sexual selection in these species (Matzkin 2005, 2014; Matzkin and Markow 2013; Bono et al. 2015; Moreyra et al. 2022). A comparison of gene families previously hypothesized to be involved in adaptation to variable cactus environments, including odorant receptors ($z = 2.541$; $P = 0.398$), gustatory receptors ($z = -0.720$; $P = 1$), glutathione-S-transferases ($z = -1.342$; $P = 1$), toxic response genes ($z = -2.119$; $P = 1$), and oxidoreductases ($z = 1.389$; $P = 1$) showed no increased dN/dS in these gene families compared to background via the codeml analysis (Fig. 5). However, higher dN/dS values were found

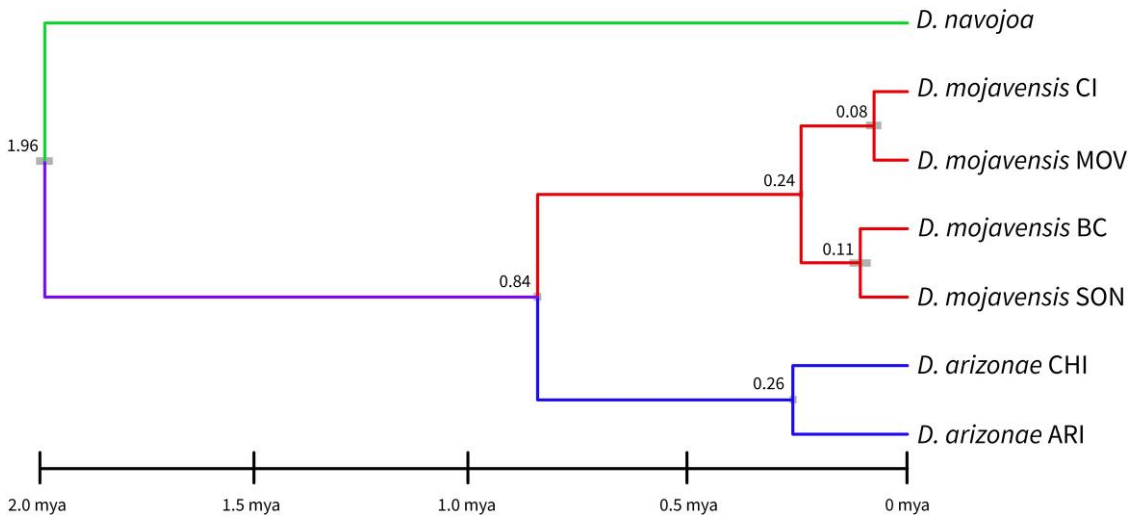

Fig. 3. Phylogeny and divergence times (mya) as estimated by BPP using 12,218 single-copy nuclear genes. Colors represent the accepted species identities and gray bars represent 95% confidence intervals for divergence time estimates.

within reproductive genes ($z = 7.153$; $P < 0.001$) as well as orphan genes absent from *D. melanogaster* ($z = 37.567$; $P < 0.001$). Genes involved in heat response displayed lower dN/dS values then the background gene set ($z = -3.566$; $P = 0.0130$). The full lists of genes found to be under positive selection via BUSTED and codeml analyses can be found in supplementary tables S1 and S2, Supplementary Material online.

We considered two hypotheses regarding the relationship between structural and coding sequence evolution. First, we predicted that genes proximal to the inversion breakpoints would be more likely to experience positive selection. This prediction stems from the hypothesis that genes around breakpoints experience a reduced rate of recombination, and it is this increase level of linkage disequilibrium that can lead to the co-segregation of adaptive mutations across these loci, sometimes refer to as supergenes (Villoutreix et al. 2021). Prior work in *D. melanogaster* indicates that long range linkage disequilibrium can level off to baseline around distances of approximately 1 Mb (Franssen et al. 2015), therefore we tested this prediction by comparing the proportion of significantly positively selected genes within 1 Mb on either end of a breakpoint to the rest of the genes in the genome. We found no evidence that genes adjacent to either the breakpoints within *D. mojavensis* ($F_{1,12185} = 0.017$, $P = 0.68$) nor the breakpoints in the clade as a whole ($F_{1,12185} = 0.33$, $P = 0.57$) displayed elevated evolutionary rates. Second, we predicted that genes in regions of low collinearity caused by any kind of structural variant would be more likely to display signatures of relaxed selection. This could reflect variation in constraint on coding and structural changes. Omega was significantly negatively correlated to the collinearity score from CI to NAV of the sliding window containing the gene (Fig. 6). This pattern held for the collinearity from CI to the mean of the *D. arizonae* populations ($F_{1,12185} = 44.80$, $P = 2.28 \times 10^{-11}$) as well as the collinearity from CI to the mean of the other three *D. mojavensis* populations ($F_{1,12185} = 19.89$, $P = 8.26 \times 10^{-6}$).

Discussion

The assembly of these seven complete de novo genomes adds to the tremendous genomic resources available within the *Drosophila* genus. These resources include 24 genomes that have been assembled at or near chromosome level in *Drosophila* (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/Drosophila) as well as many others assembled with chromosome-length scaffolds (Kim et al. 2021). Our genomes will of course allow for more robust comparative genomic analyses across *Drosophila*. However, our addition of assemblies for separate conspecific populations of two species allows for greater resolution in the calculation of fundamental evolutionary patterns on short timescales.

Accurate phylogenies and divergence times are critical both for quantifying rates of evolutionary change and for generating hypotheses on the phylogeographic causes of speciation events and adaptive radiations. Despite the extensive molecular investigation of the *D. mojavensis* species cluster, disagreement on both the topology and node ages of the phylogeny persists. Within *D. mojavensis*, three different trees have been supported. A mitochondrial study (Reed et al. 2007) found the Mojave Desert population as an outgroup while a nuclear study (Smith et al. 2012) placed the Baja California population as an outgroup. Two earlier nuclear studies (Ross and Markow 2006; Machado et al. 2007) found two clades, with Mojave–Catalina Island and Baja–Sonora as pairs of sibling species.

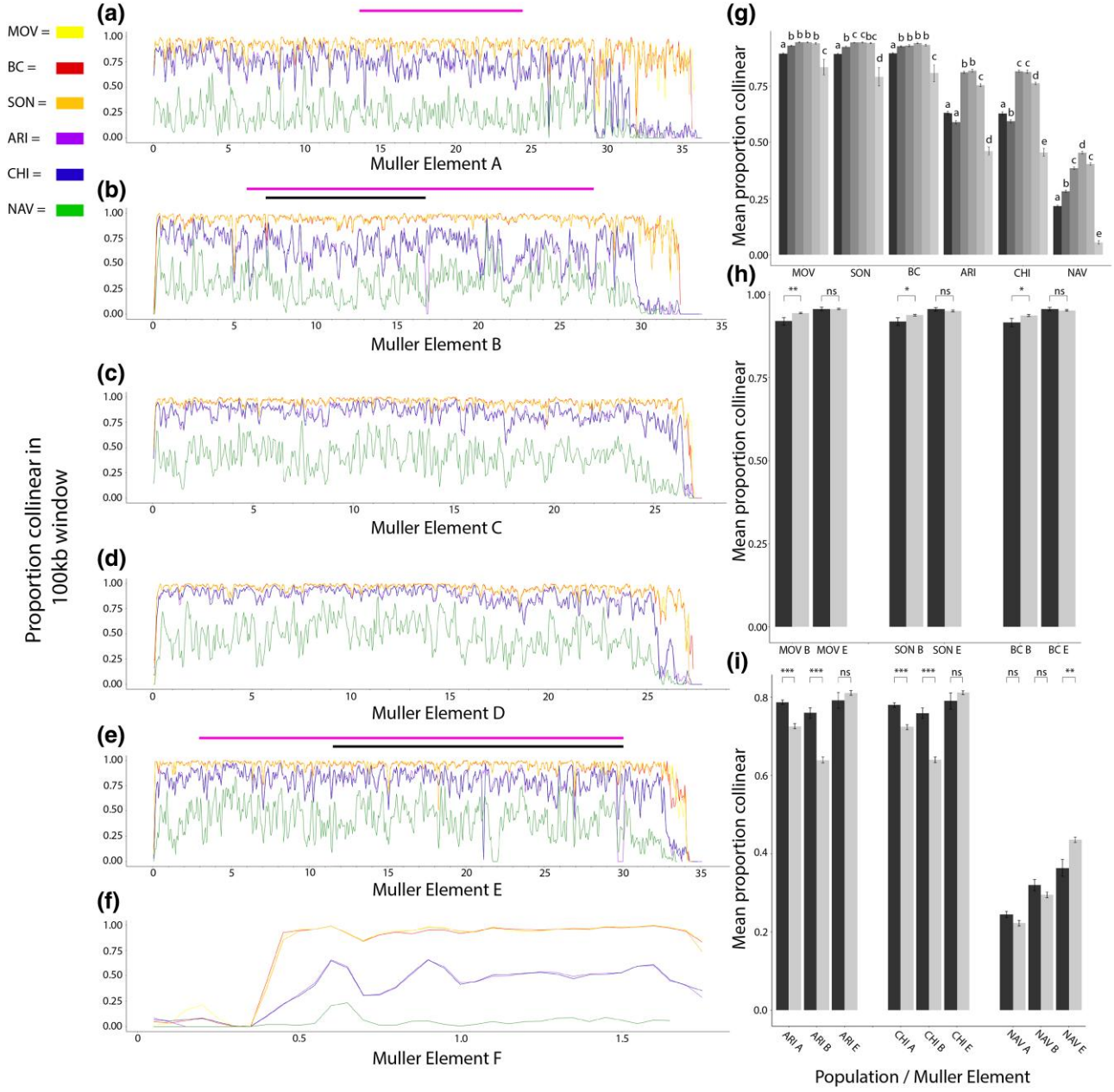

Fig. 4. a to f) The collinearity score (calculated relative to CI) across 100 kb windows for Muller elements A to F. MOV, yellow; SON, orange; BC, red; ARI, purple; CHI, blue; NAV, green. Pink and blue bars above plots indicate inter- and intraspecific inversion polymorphisms as in Fig. 2. Numbers on *x* axis indicate position in Mb with the centromeric end of the chromosome on the left. g) Mean collinearity score for each element and species. Letters indicate significant differences ($P < 0.05$) between elements for each species. Muller elements are arranged in order with A (darkest) at left and F (lightest) at right. h) Mean collinearity scores before (dark gray) and within (light gray) inverted regions of elements B and E for the three *D. mojavensis* populations. i) Mean collinearity scores before (dark gray) and within (light gray) inverted regions of elements A, B, and E for the *D. arizonae* and *D. navojoa* genomes. Asterisks in parts *H* and *I* indicate significance at the level of $P < 0.05$ (*), $P < 0.001$ (**), or $P < 0.0001$ (***).

Our mitochondrial data recapitulated the topology of the earlier mitochondrial tree, while our nuclear data supported the topology of the earlier studies (Ross and Markow 2006; Machado et al. 2007). We expect that these differences are due to a combination of sampling variance and extensive gene tree discordance, given that previous nuclear and mitochondrial studies used only a fraction of all loci, unlike the current analysis. These studies also presented variable divergence times; nuclear studies (Ross and Markow 2006; Smith et al. 2012) found the initial divergence within *D. mojavensis* to have occurred ~250,000 years ago with further divergence between 100,000 and 150,000 years

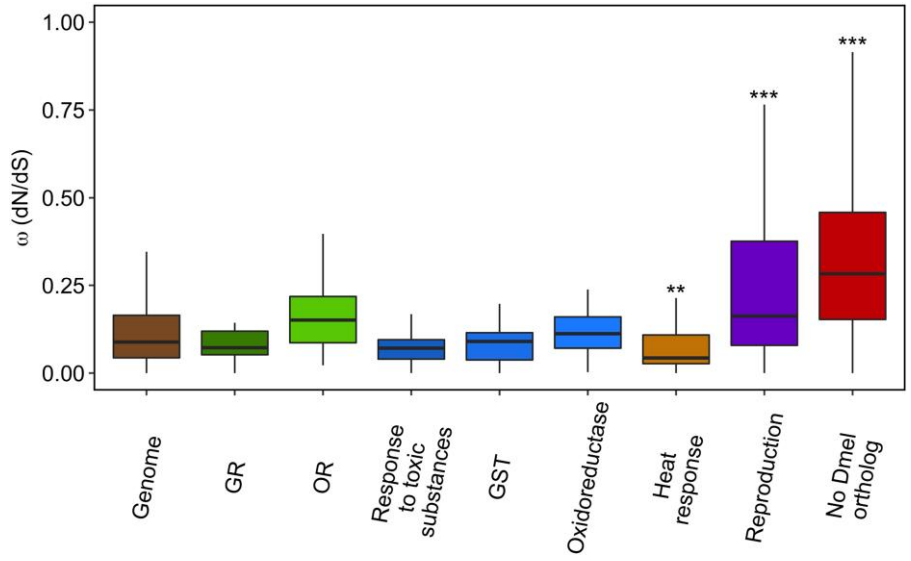

Fig. 5. Comparison of omega values for different gene families and GO categories. Asterisks indicates significant differences from the genome wide baseline (at left) at the level of $P < 0.01$ (**) or $P < 0.001$ (***).

ago, the mitochondrial study found node ages more than twice as old. Once again, our mitochondrial and nuclear data cleanly recover these differences.

Our mitochondrial and nuclear analyses also showed major differences in the relationships and divergence times across species. The biggest among these is the finding of paraphyly in *D. arizonae* in our mitochondrial dataset. Although one previous mitochondrial analysis (Reed et al. 2007) also failed to support *D. arizonae* as a clade, it differed in which population grouped with *D. mojavensis*. Here, our patterns of structural divergence align with the nuclear dataset in grouping the two *D. arizonae* populations as sibling taxa.

These consistent differences between mitochondrial and nuclear datasets could reflect a different demographic history for the mitochondrial genome, as previously suggested (Reed et al. 2007). Mitonuclear discordance is common in insects due to incomplete lineage sorting and hybridization (Toews and Brelsford 2012), both of which may be playing a role here. However, we also consider that noise may be important, given the massive difference in variable sites in the nuclear genome compared to the mitochondrial genome. Taken together, our data broadly support the topology of the earlier nuclear phylogeny of Machado et al. (2007), with divergence times within *D. mojavensis* largely in agreement with both previous nuclear studies (Ross and Markow 2006; Smith et al. 2012). However, further analyses of individual gene trees in relation to recombination rate variation could reveal more nuanced evolutionary patterns such as incomplete lineage sorting.

Our results regarding divergence times between species differed considerably from earlier estimates. These estimates have ranged widely, ranging from 0.66 to 4.2 mya for the split between *D. mojavensis* and *D. arizonae*, and from 2.9 to 7.8 mya for the divergence of these species to *D. navojoa* (Sanchez-Flores et al. 2016). In the most robust analysis to date, Sanchez-Flores et al. (2016) used over 5,000 nuclear loci to estimate an age of 5.86 mya for the split between *D. navojoa* and the rest of the *D. mojavensis* cluster and an age of 1.51 mya for the split between *D. arizonae* and *D. mojavensis*. Our nuclear analysis showed much younger divergence times of 1.96 mya for the split of *D. navojoa* and 0.84 mya for the split of *D. arizonae*. We argue that our results are more reliable for two reasons. First, our usage of multiple genomes for both *D. mojavensis* and *D. arizonae* further reduced the possibility for sampling error based on analyzing single genotypes per species. Second, our usage of the neutral mutation rate to calibrate the phylogeny is expected to be more accurate than using models calibrated from Hawaiian *Drosophila*, which have been found to inflate divergence times dramatically (Obbard et al. 2012). These younger estimates suggest that the speciation of this entire clade revolves around events including the cyclic climatic fluctuations of the past few million years and the accompanying shifts in host cactus distribution (Smith et al. 2012). On the contrary, major geological events such as the raising of the Trans-Mexican Volcanic Belt, that have also been hypothesized as possible causes of intra- and interspecific divergence (Machado et al. 2007; Rampasso et al. 2017) appear to be too ancient to have played a role here. However, if our estimation of six generations per year, which based on laboratory studies and phenological observations but not confirmed in wild populations (Matzkin and Eanes 2003; Smith et al. 2012; Lohse et al. 2015), is a significant overestimate, our divergence time estimates could be too recent.

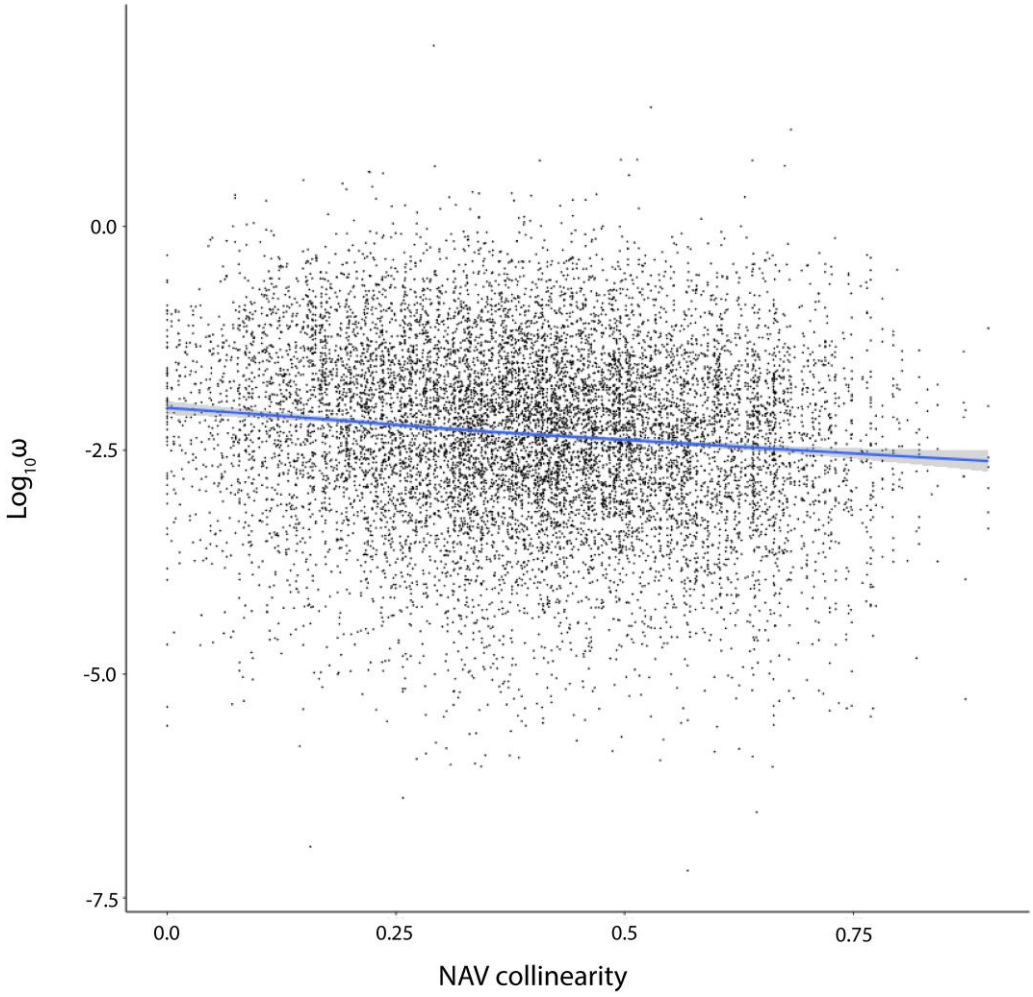

Fig. 6. The genome-wide relationship between collinearity (from *D. mojavensis* Catalina Island to *D. navojoa*) and molecular evolutionary rate across the phylogeny. The regression line represents a linear regression ($y = -0.36x - 0.98$), with 95% confidence intervals shaded ($F_{1,12153} = 94.64$, $P < 2.2 \times 10{-}16$).

Descriptions of the rates of sequence and expression evolution have served as foundational patterns of evolutionary genomics for decades. However, limited data relating to rates of accumulation of structural genomic variation have been published. Bhutkar et al. (2008) identify rates of 0.03 to 0.17 rearrangements/Mb/MY across the *Drosophila* genus, but did not consider structural variants other than micro- and macro-inversions. Chakraborty et al. (2021) found that 15% of sequence did not align between *Drosophila simulans* and *D. melanogaster*, which are diverged by about 3 million years, and noted that this was over twice the percentage of sequence variation between these species. Long et al. (2018) estimated a rate of 50 structural mutations per Mb per million years within *D. melanogaster*, which, given an average variant size of around 25 kb in their dataset, suggests that approximately 13% of the genome is diverging structurally per million years. Jiao and Schneeberger (2020) report around 10% structural divergence between *Arabidopsis thaliana* accessions using the same software and methodology here, but cannot present a phylogenetic timeline of breakdown. Here, we observe that in the *D. mojavensis* group structural similarity decays in a linear fashion, with about 33% of genome collinearity lost per million years. This rate is similar to, but clearly faster than, the rate of loss of collinearity in the *D. melanogaster* group, which we re-analyzed using the same methodology. This is consistent with the findings of Bhutkar et al. (2008), who found that rates of chromosomal inversions scale linearly with sequence divergence and are greater in subgenus Drosophila than in subgenus Sophophora. Thus, rates of structural change relative to that of sequence substitution vary between taxa. Given this fact, we hope to see more similar comparisons in other taxa both in and outside of the *Drosophila* clade. What ecological or demographic factors may influence rates of structural evolution? We expect genome size to positively correlate with structural divergence, due to the presence of more transposable elements and other sequences that

may tolerate structural mutations. However, we are also curious to see whether structural variants accumulate more rapidly in highly speciose taxa or those undergoing adaptive radiations. We anticipate that increasing numbers of high-quality genomes will allow for answers to these questions soon.

To begin to address related questions within our dataset, we asked what factors might explain local variation in collinearity within the *D. mojavensis* group. One strong predictor of structural divergence was chromosome. Although results varied slightly depending on the evolutionary distance, the dot chromosome (Muller element F) diverged most rapidly, followed by Muller elements A, B, and E. In nearly all comparisons, Muller elements C and D maintained the greatest collinearity. It is unlikely that this heterogeneity can be explained by a single factor. For Muller element F, although there is some evidence for relaxed constraint in *D. mojavensis* (Allan and Matzkin 2019), it is more likely that our results are explained by a genus-wide propensity for this chromosome to accumulate repeats and TEs, which has been attributed to a unique chromatin structure for this chromosome (Riddle and Elgin 2018). The consistent degradation of the X chromosome, on the other hand, appears to be linked to increased repeat but not TE content. This breakdown may be linked to the prevalence of rapidly evolving tandem repeats known to be common on *Drosophila* X chromosomes (Sproul et al. 2020).

No such variation in TE or repeat content is apparent amongst the four large autosomes. Instead, the variation in collinearity of these chromosomes is noteworthy for its association with the presence of major inversions. Muller elements B and E have inverted repeatedly in the *D. mojavensis* cluster, including multiple times at nearly identical breakpoints, whereas C and D have not (supplementary fig. S2, Supplementary Material online). Both adaptive and neutral hypotheses have been considered for the reuse of breakpoints. Adaptive explanations have focused on the potential for inversions to prevent recombination across genes involved in local adaptation, therefore maintaining positive combinations of alleles together (Hoffmann et al. 2004; Kirkpatrick and Barton 2006; Wellenreuther and Bernatchez 2018). These clusters of selected genes adjacent or within chromosomal inversions have been associated with a number of phenotypic traits, such as for example wing pigmentation patterns in *Heliconius* butterflies (Jay et al. 2022) and local adaptation in Atlantic salmon (Stenløkk et al. 2022). Nonadaptive explanations have considered that certain genomic regions may be susceptible to inversions due to variation in chromatin structure and genome fragility (von Grotthuss et al. 2010). Our results support the latter explanation for the *D. mojavensis* cluster, as Muller elements B and E appear to be more susceptible to a wide range of structural mutations beyond large inversions. P-elements have also been shown to be responsible for some inversions in *D. mojavensis* (Rius et al. 2013). Further supporting that this relationship is correlational, we see no evidence that inversions cause additional decreases in collinearity, as there was no consistent trend of increased collinearity outside of the inverted regions of these chromosomes. This does not exclude the possibility that specific breakpoints are relevant to adaptation; although we found no evidence that genes near breakpoints within the *D. mojavensis* cluster are more likely to display signatures of selection, the presence of some positively selected genes near breakpoints still reflects a potential link between inversions and adaptation. Furthermore, previous work (Guillén and Ruiz 2012) suggests that gene regulatory variation may be responsible for inversion associated adaptation in this system.

We also found that variation in overall genome collinearity caused by all types of structural changes was negatively linked to omega. As most omega values, even at the lowest levels of collinearity, were much less than one, we suggest that this trend reflects relaxed selection for genes in low collinearity regions. Two nonexclusive phenomena could help explain this pattern. First, genes already experiencing relaxed selection on protein function might better tolerate structural changes that may also influence splicing or expression (Hämälä et al. 2021), meaning that mutations near these genes are more likely to be maintained. Second, the causality could be reversed, and structural changes to genes could directly cause subsequent bouts of reduced constraint and relaxed selection. In many cases, this could be explained as a result of sub- or neofunctionalization following gene duplication. However, in our dataset, molecular evolution was only assessed for single-copy orthologs across all seven genomes. Thus, the relevant duplications would have occurred prior to the common ancestor of these species, and would not register as structural variants in this dataset. A more likely possibility is that structural changes result in alterations to gene regulation and phenotype, which subsequently leads to a relaxation of selection on amino acid sequences.

Although positive selection doesn't appear to be responsible for the genome-wide correlation with collinearity, relaxed constraint can still lead to positive selection by tolerating the substitution of potentially adaptive amino acids. We therefore consider genes experiencing positive selection in regions of low collinearity as interesting candidates for roles in adaptation and speciation. We are particularly interested in genes involved in reproduction, given the elevated rates of positive selection for genes in this category. One particularly interesting gene in this regard is GI18186; which has an omega of 1.166 and lies in a window with a collinearity score in the 6th percentile or lower in all three species comparisons. This gene is orthologous to the *D. melanogaster* gene CG13965 which is massively expressed in male accessory glands (Brown et al. 2014) and

has been localized to a small cluster of accessory gland proteins (Acps; Ravi Ram and Wolfner 2007). Furthermore, CG13965 protein is known to be transferred from males to females during mating, not only in *D. melanogaster* (Immarigeon et al. 2021) but in *D. simulans* and *Drosophila yakuba* as well (Findlay et al. 2008). Function of male protein in the female reproductive tract has been hypothesized as an important speciation mechanism between species and populations in the *D. mojavensis* cluster (Bono et al. 2011). Our results suggest that GI18186 is worthy of further attention, and that both changes to the expression and sequence of this gene may have contributed to pre-mating post-zygotic isolation leading to reproductive isolation, as is the case for Acps in *D. melanogaster* (Immarigeon et al. 2021). Given that the number of annotated Acps in *Drosophila* is in the hundreds, it is important to narrow down the list of possible relevant genes for more targeted studies. Thus, it is valuable that our integration of sequence and structural analysis allows us to make this prediction from single genome sequences alone.

Extending this, the second category of genes that were found to be overrepresented for positive selection are those without orthologs in *D. melanogaster*, and are therefore likely to be taxonomically restricted genes (TRGs) in at least the repleta group if not the *D. mojavensis* cluster. TRGs have been previously implicated in cactophilic *Drosophila* evolution (Moreyra et al. 2022) as well as many other taxa, and likely reflects both that TRGs are unlikely to have housekeeping functions and may be preferentially involved in novel traits and adaptations (Domazet-Loso and Tautz 2003; Arendsee et al. 2014; Jasper et al. 2015). In spite of their likelihood of relevance to adaptation, the lack of functional annotation for genes with no well-studied ortholog in a model organism represents a major issue in the biology of non-model organisms, and a systematic study of these genes is unlikely for the vast majority of taxa. Here, we find that most of the genes with evidence of positive selection in regions of low collinearity are TRGs. We argue that these genes should be prioritized in targeted investigations seeking to characterize the functions of currently unstudied genes.

Materials and Methods

Insect Strains, Genome Sequencing, and Assembly

Each strain used in this study (Table 1) was maintained as an inbred line in the Matzkin lab at the University of Arizona on a banana-molasses based diet (recipe in Coleman et al. 2018) through genome and RNA sequencing.

The original genomic scaffolds (*Drosophila* 12 Genomes Consortium 2007) as well as short- and long-read (Miller et al. 2018) sequence data for the Santa Catalina Island *D. mojavensis* assembly are previously published. The short-read Illumina data for the remaining *D. mojavensis* populations is described in Allan and Matzkin (2019). Long-read data for the Sonora *D. mojavensis* population is described in Jaworski et al. (2020). The short-read Illumina data for the *D. arizonae* Sonora population is described in Diaz et al. (2021). The short-read Illumina data for *D. navojoa* is described in Vanderlinde et al. (2019).

Briefly, for all short-read data, we extracted DNA from a pool of ten adult males and ten adult females using Qiagen DNeasy Blood & Tissue Kits (Qiagen, Hilden, Germany), and we constructed the *D. arizonae* Chiapas library using KAPA LTP Library Preparation Kit (Roche, Basel, Switzerland) kits. It was sequenced on an Illumina HiSeq 4000 at Novogene (Beijing, China) at 220× coverage. All other short-read libraries were built and sequenced on Illumina HiSeq 2000 at the HudsonAlpha Genome Sequencing Center (Huntsville, AL, USA) at 75× coverage. For all long-read data, we extracted high molecular weight DNA from a pool of 150 males and 150 females using a chloroform-based extraction, detailed method in Jaworski et al. (2020). PacBio libraries were built and CLR reads were sequenced on a PacBio Sequel at the Arizona Genomics Institute (Tucson, AZ, USA).

The assembly of the six de novo genomes largely followed the hybrid assembly strategy described in Jaworski et al. (2020), wherein a detailed description of sequencing and assembly methods can be found. Briefly, we used Platanus 1.2.4 (Kajitani et al. 2014) and DBG2OLC (Ye et al. 2016) to produce hybrid assemblies of the short- and long-read data. We also used Canu 1.7 (Koren et al. 2017) for long-read only assembly with the correctedErrorRate parameter set to 0.039 for the primary assembly though this was increased to 0.065 to produce a less stringent assembly used for bridging and extending primary contigs. We used Quickmerge 2.0 (Chakraborty et al. 2016) to merge these two assemblies into a draft assembly. We then manually merged contigs based on whole-genome alignments from Mauve (Darling et al. 2004) and Nucmer (Delcher et al. 2002) including using the less stringent assembly in Geneious Prime (Biomatters, Auckland, NZ). Where contigs could not be merged, we manually joined them based on alignment with the other genomes and connected with an N-gap of 100 bp. We checked each manual join against the alignments from the other assemblies as well as polytene chromosome maps (Schaeffer et al. 2008), and found that each was consistent. We are therefore confident in the arrangement of each chromosome, although of course uncertainties in the lengths of repeat regions remain where our merging approaches could not create scaffolds.

We aligned all remaining contigs not assigned to a chromosome with Minimap2 (Li 2018) and subsequently discarded all contigs with a match of over 80% to a chromosome scaffold. We polished each genome three times with Pilon 1.23 (Walker et al. 2014). During manual curation of our annotations with the help of RNA-seq data

(see below), we identified several small insertion/deletion errors in each genome that led to frameshift errors causing problems with gene structure, and subsequently fixed these errors manually in Geneious Prime. We noticed that the *D. arizonae* Chiapas genome had substantially more of these errors than the others and therefore polished it a fourth time with Pilon 1.23 before fixing remaining errors manually as for the other genomes. We also performed additional polishing of the gene-containing regions of *D. navojoa* using majority consensus in Geneious Prime.

We re-scaffolded the *D. mojavensis* Santa Catalina Island genome (hereafter, CI) in order to provide better comparisons of structure with the six de novo assemblies. We first polished the FlyBase assembly version r1.04 twice with Pilon 1.23. We then manually scaffolded by aligning contigs from the existing Nanopore data (Miller et al. 2018) to the polished reference using Mauve and joining in Geneious Prime. We filled all N-gaps over 20 kb with contigs from the Nanopore dataset. Lastly, we filled all N-gaps regardless of size if they occurred within 100 bp of a putative CDS feature identified during annotation. Similar to the other assemblies, annotation revealed several indel errors in coding regions the CI genome, which we fixed manually. In addition, to filtering duplicate scaffolds with Minimap2 we also removed scaffolds that previously had a gene annotation in the 1.04 release if those genes had strong BLAST hits to a gene on the chromosome scaffolds. Existing annotations were kept if no BLAST hit was found, all other annotations on unmapped scaffolds were removed.

We noticed that the previous assembly of Muller element F in CI was much larger than in our de novo assemblies, and contained ~1.3 Mb of sequence that was homologous to sequence in Muller element A (X chromosome). We therefore split the CI Muller element F into two pieces: we kept bp 1 to 2,135,734 as chromosome F, while we joined bp 2,139,764 to 3,406,379 to chromosome A based on alignments in Mauve and NUCmer. We confirmed this split based on separate mapping data from a cross of the CI, SON, and MOV *D. mojavensis* populations, which showed no genetic linkage across this breakpoint of the original chromosome F (K.M. Benowitz unpubl. Data). All other large scaffolds in CI were linked to a chromosome based on physical and genetic marker data from Schaeffer et al. (2008).

After finalizing the assemblies, we ran RepeatModeler (Flynn et al. 2020) on each genome before using USEARCH (Edgar 2010) with a 90% similarity cutoff to create a non-duplicated combined list of repetitive elements. We then ran RepeatMasker (http://www.repeatmasker.org) to generate masked versions of each assembly prior to annotation.

We generated mitochondrial assemblies for all six de novo genomes by mapping reads to the existing CI mitochondrial sequence (*Drosophila* 12 Genomes Consortium 2007) in Geneious Prime.

Genome Annotation

To help facilitate annotation, we performed a broad RNA-seq experiment designed to detect expression of as many genes as possible. In October 2020; we collected tissue from each of the seven genome strains during early (12 h post-laying) and late (26 h post-laying) embryonic stages, first, second, and third instar larvae, pupae, and male and female adults at varying ages post-eclosion. For each life stage, we ground tissue in 500 µL of Trizol reagent (Thermo Fisher Scientific, Waltham, MA, USA) prior to extracting RNA using a ZYMO Direct-zol RNA Miniprep Kit. We then quantified the RNA and pooled extractions for each life stage together to reach 1.5 µg of total RNA. We then built libraries using a KAPA stranded mRNA-Seq Kit for each strain and sequenced them on an Illumina HiSeq 4000 lane at Novogene. We trimmed all RNA reads using Trimmomatic (Bolger et al. 2014) and aligned each to its respective genome using Hisat2 (Kim et al. 2019) under default parameters.

We used the current annotation of the Catalina Island *D. mojavensis* genome as a starting point for our genome annotations. We first transferred these annotations to our new CI genome assembly using Mauve within Geneious Prime. We next aligned all seven genomes using Cactus 1.1 (Armstrong et al. 2020) before using the Comparative Annotation Toolkit (CAT 2.0; Fiddes et al. 2018) to transfer the annotations from the new CI genome to each of the other six genomes. Because these annotations were necessarily limited to genes that both existed and were annotated correctly in the original CI genome, we used two additional strategies to provide less biased annotations. First, we ran maker iteratively (Campbell et al. 2014; Card et al. 2019) to generate ab initio gene predictions for each genome, after initially training with a transcriptome generated by running StringTie (Pertea et al. 2015) on the aligned RNA-seq data and proteins taken from *D. mojavensis* and *D. melanogaster*. Second, we used PASA (Haas et al. 2003) within the funannotate pipeline (https://github.com/nextgenusfs/funannotate) to generate gene predictions after trimming, normalizing, and aligning the raw RNA-seq reads described above.

We determined a posteriori that the CAT annotations were by far the closest match to the raw RNA-seq data, and therefore chose to use these as our baseline for the final annotation. We next loaded GFF files from CAT, maker, and PASA, along with the raw RNA-seq alignments, into the Apollo genome annotation browser (Dunn et al. 2019) for manual curation. During manual curation we performed three tasks. First, we added new genes that were either unannotated in the original *D. mojavensis* genome or that the CAT pipeline did not add correctly. Second, we fixed genes that had either been incorrectly split or merged in the original annotation. Lastly, we fixed errors

that were introduced due to sequencing errors in either the original Catalina Island genome or one of the six new genomes, which generally required manually fixing both the genome (see above) and the corresponding annotation.

We analyzed both the completeness of our genome assemblies and our annotations by using BUSCO (Seppey et al. 2019) to compare our own gene content against the most recent database of conserved single-copy dipteran genes (Diptera_odb10).

We generated mitochondrial annotations by transferring existing annotations from the CI mitochondria to each of the other mitochondrial assemblies using Mauve.

We used results from RepeatModeler above to calculate repeat content for each genome and BBMap stats (https://sourceforge.net/projects/bbmap/) to calculate GC content. To estimate transposable element (TE) content, we used EDTA (Ou et al. 2019), which has been demonstrated to be effective in annotating non-model genomes (Bell et al. 2022). We used custom bash scripts to calculate the percentage of GC, repeats, TEs, and genes in 100 kb sliding windows overlapping by 50 kb, and plotted these percentages for each genome using the R package circlize (Gu 2014).

Phylogenomics and Divergence Time Estimation

We identified 12,218 single-copy orthologs across all seven genomes with OrthoFinder (Emms and Kelly 2019) using an iterative process. We first ran OrthoFinder under default parameters, separating single-copy orthologs from the remaining genes. After identifying 10,807 single-copy orthologs, we noticed that several gene clusters identified by OrthoFinder occurred in multiples of seven. This suggested that many genes, although part of ancestral duplications, were single-copy orthologs within the *D. mojavensis* group, and therefore still useful for our analyses. We therefore re-ran the software on the remaining genes that were not defined as single-copy orthologs using stricter parameters, and repeating this procedure twice. Using this approach, we were able to capture another 1,324 single-copy orthologs after one additional iteration of OrthoFinder, and a further 56 after a second iteration.

We then performed codon alignments of all single-copy orthologs using PRANK (Löytynoja 2014) with the "-codon" option, and extracted 4-fold degenerate sites using custom scripts from each alignment. We generated individual, unrooted gene trees using only the 4-fold degenerate sites using RAxML with the GTRCAT model (Stamatakis 2014), and used these trees as input for consensus tree building using ASTRAL-III (Zhang et al. 2018) and MP-EST. All programs were run using default parameters.

After establishing a consensus tree topology, we used BPP (Flouri et al. 2018) on all 12,218 single-copy orthologs for species tree estimation only (model 01) with 100,000 samples, a sampling frequency of 2; and a burn in of 10,000 samples, to estimate divergence times across the phylogeny. We altered the following parameters within BPP: thetaprior (3.0, 0.002) and tauprior (3.0, 0.003). All other parameters were left at default settings. Following recommendations for estimating divergence time in *Drosophila* (Obbard et al. 2012), we used a neutral mutation rate of 3.5×10^{-9} mutations/bp/generation (Keightley et al. 2009) and a rate of six generations per year (Matzkin and Eanes 2003; Smith et al. 2012; Lohse et al. 2015) to convert the substitution rate from BPP into age in years.

As several earlier estimates of divergence within this clade were made entirely (Reed et al. 2007) or in part (Oliveira et al. 2012) using mitochondrial data, we repeated the above analysis with the de novo mitochondrial genome assemblies. We first annotated thirteen known mitochondrial genes and extracted 4-fold degenerate sites before running BPP model 01 using the same parameters as above for the nuclear genes. We used the mitochondrial mutation rate of 6.2×10^{-8} per site per generation (Haag-Liautard et al. 2008) and a rate of six generations per year to calculate the BPP estimate of divergence in years.

Analysis of Structural Genome Evolution

We aligned all seven genomes using NUCmer in order to identify breakpoints and visualize previously identified chromosomal inversions on Muller elements A, B, and E. We made figures of genome wide collinearity using Dot (https://github.com/marianattestad/dot). Prior to analyzing structural variation quantitatively, we used these breakpoints to manually create "uninverted" chromosomes, wherein we forced all chromosomes to be homokaryotypic with CI. This allowed us to compare collinearity inside and outside of major inversions in an unbiased manner. This definition also allows us to compare chromosomes without overweighting the contribution of single, large inversion variants to the reduction in collinearity. We re-ran NUCmer on the "uninverted" genome assemblies and used this output as input for identification of structural variation and collinear genome regions using SyRI (Goel et al. 2019). Using the CI genome as our template, we followed Jiao and Schneeberger (2020) in quantifying the percentage of collinear sequence in 100 kb regions of the genome over 50 kb sliding windows using custom bash scripts. Thus, any structural variant present between populations or species other than a major inversion was considered a noncollinear region. We compared collinearity across chromosomes within each genome using ANOVA and Tukey's test for post hoc comparisons, using a collinearity score with overlapping windows removed. For Muller element F, we calculated chromosome-wide collinearity after removing ~350 kb at the centromeric end of the CI chromosome, which may be a misassembly as it has no corresponding region on any of the six de novo

assemblies. For each chromosome with an inversion, we additionally compared the collinearity outside the inversion on the centromeric end to the collinearity within the inversion region using ANOVA. The region outside the inversion only included the region on the centromeric side of the inversion. We did not compare the non-inverted region on the telomeric end due to the extreme degradation of collinearity near the telomere, especially in the interspecific comparisons.

To compare our results in a group of related *Drosophila* species, we downloaded chromosome-level genome assemblies for *D. melanogaster* (GCA_000001215.4; Hoskins et al. 2015), *Drosophila mauritiana* (GCA_004382145.1; Chakraborty et al. 2021), *Drosophila sechellia* (GCA_004382195.2; Chakraborty et al. 2021), *Drosophila simulans* (GCA_016746395.2; Chakraborty et al. 2021), *Drosophila yakuba* (GCA_016746365.2), *Drosophila teissieri* (GCA_016746235.2), and *Drosophila santomea* (GCA_016746245.2). We chose these species due to their assembly qualities and due to the fact that they have diverged recently (<3.5 mya), allowing for a clear comparison with our results from the *D. mojavensis* clade. As above, we aligned all species to the *D. melanogaster* assembly with nucmer, manually de-inverted any chromosomal inversions, and ran nucmer and syri on the de-inverted chromosomes to identify collinear regions. We compared the global collinearity scores within this clade to divergence times calculated with mutation rate calibration (Obbard et al. 2012) to match closely the methods that we used for the *D. mojavensis* clade.

Analysis of Molecular Evolution

For molecular evolutionary analyses, we used the same set of aligned single-copy orthologs as used above in phylogenomic analyses, and used the phylogeny from the analysis above. We analyzed the levels of selective pressure as the ratio of non-synonymous to synonymous substitutions (dN/dS) of each sequence across the entire phylogeny using Codeml (PAML; Yang 2007) by using models 0, 7, and 8. Whereas model 0 provided a descriptive, baseline value of dN/dS, a Fisher's Exact test comparing model 7 (which does not allow for positive selection) to model 8 (which allows for positive selection provided significance testing for positive selection of each gene). We also analyzed evidence for positive selection along the entire phylogeny using BUSTED2 (Murrell et al. 2015) within the HyPhy package (Kosakovsky-Pond et al. 2020). We present uncorrected *P*-values for all genes from both approaches as well as omega (dN/dS) values from codeml in supplementary table S1, Supplementary Material online, as well as the subset of genes considered significant by at least one approach in supplementary table S2, Supplementary Material online.

We identified orthologs for each gene in *D. melanogaster* by taking the best results from a blastp search run on all genes. The gene ontology terms of the *D. melanogaster* orthologs were then obtained from flybase (www.flybase.org). For the comparison of functional categories, loci were grouped by the following GO terms: Gustatory Receptors (GR) GO:0050909; Odorant Receptors (OR) GO:0050911; Response to toxic substance GO:0009636; Glutathione S-transferases (GST) GO:0006749; Oxidoreductases GO:0016705; Heat response GO:0009408 and Reproduction GO:0032504. The effect of GO category on dN/dS value was estimated with a van der Waerden test, which was followed by a post hoc nonparametric comparison between the background gene set and each GO category using the Dunn method for join ranking, performed in JMP 10.

We considered two hypotheses regarding the relationship between structural and coding sequence evolution. First, we predicted that given the increased linkage disequilibrium, gene clusters proximal to the inversion breakpoints would be more likely to experience as a group the effects of positive selection. This prediction stems from the hypothesis that adaptive genes may cluster around breakpoints due to the reduced likelihood of their disruption via recombination (Villoutreix et al. 2021). We tested this prediction by comparing the proportion of significantly positively selected genes within 1 Mb on either end of a breakpoint to the rest of the genes in the genome. Second, we predicted that genes in regions of low collinearity would be more likely to display signatures of positive selection. To examine this prediction, we performed linear regression to examine the relationship between the log10 ω value of each gene and the collinearity score between CI and NAV of the 100 kb window containing the gene. We chose to display NAV due to the fact that it displays the greatest variation in collinearity while remaining correlated with structural variation in the other genomes ($r_{NAV\text{-}MOJ} = 0.48$; $r_{NAV\text{-}ARI} = 0.73$). However, we additionally performed the same analysis on the mean collinearity scores of the two *D. arizonae* genomes and the three remaining *D. mojavensis* genomes to confirm this pattern. We performed all statistical analyses in R 3.6.3 (R Core Team 2020).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the National Science Foundation (IOS-1557697 and IOS-2220279 L.M.M.). We thank D. Kudrna for his work to produce the PacBio sequences. We thank N. Sage for assistance with genome annotations. We would like to dedicate this work to Bill Heed, Marvin Wasserman, and William Starmer whose foundational work on this system has been tremendously impactful.

Author Contributions

K.M.B., C.W.A., C.C.J., and L.M.M. conceived and designed the study. C.W.A. and C.C.J. assembled genomes. K.M.B., C.W.A., F.D., X.C., and L.M.M. annotated genomes. K.M.B., C.W.A., and L.M.M. performed analyses of genome structure. K.M.B. and M.J.S. performed analyses of phylogenomics and divergence time estimation. K.M.B., C.W.A., and L.M.M. performed molecular evolutionary analyses. K.M.B. and L.M.M. wrote the paper. All authors read and approved the final manuscript.

Data Availability

All raw genomic and transcriptomic sequence data have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) and are all associated with the accession number PRJNA593234. All scripts and other data are available at OSF (https://osf.io/mqvgh).

Literature Cited

Allan CW, Matzkin LM. Genomic analysis of the four ecologically distinct cactus host populations of *Drosophila mojavensis*. BMC Genom. 2019:20(1):732. https://doi.org/10.1186/s12864-019-6097-z.

Arendsee ZW, Li L, Wurtele ES. Coming of age: orphan genes in plants. Trends Plant Sci. 2014:19(11):698–708. https://doi.org/10.1016/j.tplants.2014.07.003.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020:587(7833):246–251. https://doi.org/10.1038/s41586-020-2871-y.

Barrett RDH, Hoekstra HE. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet. 2011:12(11):767–780. https://doi.org/10.1038/nrg3015.

Bell EA, Butler CL, Oliveira C, Marburger S, Yant L, Taylor MI. Transposable element annotation in non-model species: the benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. Mol Ecol Res. 2022:22(2):823–833. https://doi.org/10.1111/1755-0998.13489.

Benowitz KM, Coleman JM, Allan CW, Matzkin LM. Contributions of *cis*- and *trans*-regulatory evolution to transcriptomic divergence across populations in the *Drosophila mojavensis* larval brain. Genome Biol Evol. 2020:12(8):1407–1418. https://doi.org/10.1093/gbe/evaa145.

Benowitz KM, Coleman JM, Matzkin LM. Assessing the architecture of *Drosophila mojavensis* locomotor evolution with bulk segregant analysis. G3. 2019:9(5):1767–1775. https://doi.org/10.1534/g3.119.400036.

Berdan EL, Barton NH, Butlin R, Charlesworth B, Faria R, Fragata I, Gilbert KJ, Jay P, Kapun M, Lotterhos KE, et al. How chromosomal inversions reorient the evolutionary process. J Evol Biol. 2023:36(12):1761–1782. https://doi.org/10.1111/jeb.14242.

Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics. 2008:179(3):1657–1680. https://doi.org/10.1534/genetics.107.086108.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014:30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bono JM, Matzkin LM, Hoang K, Brandsmeier L. Molecular evolution of genes involved in post-mating pre-zygotic isolation in cactophilic *Drosophila*. J Evol Biol. 2015:28(2):403–414. https://doi.org/10.1111/jeb.12574.

Bono JM, Matzkin LM, Kelleher ES, Markow TA. Postmating transcriptional changes in reproductive tracts of con- and heterospecifically mated *Drosophila mojavensis* females. Proc Natl Acad Sci U S A. 2011:108(19):7878–7883. https://doi.org/10.1073/pnas.1100388108.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. Diversity and dynamics of the *Drosophila* transcriptome. Nature. 2014:512(7515):393–399. https://doi.org/10.1038/nature12962.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genom Biol. 2016:17(1):66. https://doi.org/10.1186/s13059-016-0924-1.

Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. Curr Prot Bioinf. 2014:48:4.11.1–4.11.39. https://doi.org/10.1002/0471250953.bi0411s48.

Card DC, Adams RH, Schield DR, Perry BW, Corbin AB, Pasquesi GIM, Row K, Van Kleeck MJ, Daza JM, Booth W, et al. Genomic basis of convergent island phenotypes in boa constrictors. Genom Biol Evol. 2019:11(11):3123–3143. https://doi.org/10.1093/gbe/evz226.

Casacuberta C, González J. The impact of transposable elements in environmental adaptation. Mol Ecol. 2013:22(6):1503–1517. https://doi.org/10.1111/mec.12170.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nuc Ac Res. 2016:44(19):e147. https://doi.org/10.1093/nar/gkw654.

Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. Evolution of genome structure in the *Drosophila simulans* species complex. Genom Res. 2021:31(3):380–396. https://doi.org/10.1101/gr.263442.120.

Chakraborty M, VanKuren NW, Zhao R, Zhang W, Kalsow S, Emerson JJ. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. Nat Genet. 2018:50(1):20–25. https://doi.org/10.1038/s41588-017-0010-y.

Coleman JM, Benowitz KM, Jost AG, Matzkin LM. Behavioral evolution accompanying host shifts in cactophilic *Drosophila* larvae. Ecol Evol. 2018:8(14):6921–6931. https://doi.org/10.1002/ece3.4209.

Corbett-Detig RB, Cardeno C, Langley CH. Sequence-based detection and breakpoint assembly of polymorphic inversions. Genetics. 2012:192(1):131–137. https://doi.org/10.1534/genetics.112.141622.

Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genom Res. 2004:14(7):1394–1403. https://doi.org/10.1101/gr.2289704.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002:30(11):2478–2483. https://doi.org/10.1093/nar/30.11.2478.

Delprat A, Guillén Y, Ruiz A. Computational sequence analysis of inversion breakpoint regions in the cactophilic *Drosophila mojavensis* lineage. J Hered. 2019:110(1):102–117. https://doi.org/10.1093/jhered/esy057.

Diaz F, Allan CW, Markow TA, Bono JM, Matzkin LM. Gene expression and alternative splicing dynamics are perturbed in female head transcriptomes following heterospecific copulation. BMC Genom. 2021:22(1):359. https://doi.org/10.1186/s12864-021-07669-0.

Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in *Drosophila*. Genom Res. 2003:13(10):2213–2219. https://doi.org/10.1101/gr.1311003.

Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 2007:450:203–218. https://doi.org/10.1038/nature06341.

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elsik CG, Lewis SE. Apollo: democratizing genome annotation. PLoS Comp Biol. 2019:15(2):e1006790. https://doi.org/10.1371/journal.pcbi.1006790.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010:26(19):2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genom Biol. 2019:20(1):238. https://doi.org/10.1186/s13059-019-1832-y.

Etges WJ. Evolutionary genomics of host plant adaptation: insights from *Drosophila*. Curr Opin Insect Sci. 2019:36:96–102. https://doi.org/10.1016/j.cois.2019.08.011.

Etges WJ, de Oliveira CC, Gragg E, Ortíz-Barrientos D, Noor MA, Ritchie MG. Genetics of incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and genotype × environment interactions. Evolution. 2007:61(5):1106–1119. https://doi.org/10.1111/j.1558-5646.2007.00104.x.

Faria R, Johannesson K, Butlin RK, Westram AM. Evolving inversions. Trends Ecol Evol. 2019:34(3):239–248. https://doi.org/10.1016/j.tree.2018.12.005.

Feder JL, Nosil P. Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? Evolution. 2009:63(12):3061–3075. https://doi.org/10.1111/j.1558-5646.2009.00786.x.

Fellows DP, Heed WB. Factors affecting host plant selection in desert-adapted cactophilic *Drosophila*. Ecology. 1972:53(5):850–858. https://doi.org/10.2307/1934300.

Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, Gordon D, Earl D, Keane T, Eichler EE, et al. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. Genom Res. 2018:28(7):1029–1038. https://doi.org/10.1101/gr.233460.117.

Findlay GD, Yi X, MacCoss MJ, Swanson WJ. Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating. PLoS Biol. 2008:6(7):e178. https://doi.org/10.1371/journal.pbio.0060178.

Flouri T, Jiao X, Rannala B, Yang Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol Biol Evol. 2018:35(10):2585–2593. https://doi.org/10.1093/molbev/msy147.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020:117(17):9451–9457. https://doi.org/10.1073/pnas.1921046117.

Fogleman JC, Danielson PB. Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran desert. Am Zool. 2001:41(4):877–889. https://doi.org/10.1093/icb/41.4.877.

Fogleman JC, Foster JL. Microbial colonization of injured cactus tissue (*Stenocereus gummosus*) and its relationship to the ecology of cactophilic *Drosophila mojavensis*. Appl Env Microbiol. 1989:55(1):100–105. https://doi.org/10.1128/aem.55.1.100-105.1989.

Fogleman JC, Heed WB. Columnar cacti and desert *Drosophila*: the chemistry of host-plant specificity. In: Schmidt J, editor. Special biotic relationships in the arid southwest. Albuquerque (NM): University of New Mexico Press; 1989. p. 1–24.

Fogleman JC, Starmer WT, Heed WB. Larval selectivity for yeast species by *Drosophila mojavensis* in natural substrates. Proc Natl Acad Sci U S A. 1981:78(7):4435–4439. https://doi.org/10.1073/pnas.78.7.4435.

Fogleman JC, Starmer WT, Heed WB. Comparisons of yeast florae from natural substrates and larval guts of southwestern *Drosophila*. Oecologia. 1982:52(2):187–191. https://doi.org/10.1007/BF00363835.

Franssen SU, Nolte V, Tobler R, Schlötterer C. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental *Drosophila melanogaster* populations. Mol Biol Evol. 2015:32(2):495–509. https://doi.org/10.1093/molbev/msu320.

Gilbert DG. DroSpeGe: rapid access database for new *Drosophila* species genomes. Nucleic Acids Res. 2007:35(Database):D480–D485. https://doi.org/10.1093/nar/gkl997.

Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genom Biol. 2019:20(1):277. https://doi.org/10.1186/s13059-019-1911-0.

Gu Z. *Circlize* implements and enhances circular visualization in R. Bioinformatics. 2014:30(19):2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

Guillén Y, Casillas S, Ruiz A. Genome-wide patterns of sequence divergence of protein-coding genes between *Drosophila buzzatii* and *D. mojavensis*. J Hered. 2019:110(1):92–101. https://doi.org/10.1093/jhered/esy041.

Guillén Y, Ruiz A. Gene alterations at Drosophila inversion breakpoints provide *prima facie* evidence for natural selection as an explanation for rapid chromosomal evolution. BMC Genom. 2012:13(1):53. https://doi.org/10.1186/1471-2164-13-53.

Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. Direct estimation of the mitochondrial mutation rate in *Drosophila melanogaster*. PLoS Biol. 2008:6(8):e204. https://doi.org/10.1371/journal.pbio.0060204.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003:31(19):5654–5666. https://doi.org/10.1093/nar/gkg770.

Hager ER, Harringmeyer OS, Wooldridge TB, Theingi S, Gable JT, McFadden S, Neugeborgen B, Turner KM, Jensen JD, Hoekstra HE. A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. Science. 2022:377(6604):399–405. https://doi.org/10.1126/science.abg0718.

Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, Tiffin P. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. Proc Natl Acad Sci U S A. 2021:118(35):e2102914118. https://doi.org/10.1073/pnas.2102914118.

Harringmeyer OS, Hoekstra HE. Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. Nat Ecol Evol. 2022:6(12):1965–1979. https://doi.org/10.1038/s41559-022-01890-0.

Heed, WB. Ecology and genetics of Sonoran desert *Drosophila*. In: Brussard PF, editor. Ecological genetics: the interface. New York (NY): Springer-Verlag; 1978. p. 109–126.

Heed, WB. The origin of *Drosophila* in the Sonoran Desert. In: Barker JSF, Starmer WT, editors. Ecological genetics and evolution: the cactus-yeast-Drosophila model system. New York (NY): Academic Press; 1982. p. 65–80.

Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019:35(17):2907–2915. https://doi.org/10.1093/bioinformatics/btz041.

Hoffmann AA, Sgrò CM, Weeks AR. Chromosomal inversion polymorphisms and adaptation. Trends Ecol Evol. 2004:19(9):482–488. https://doi.org/10.1016/j.tree.2004.06.013.

Holmes MW, Hammond TT, Wogan GOU, Walsh RE, Labarbera K, Wommack EA, Martins FM, Crawford JC, Mack KL, Bloch LM, et al. Natural history collections as windows on evolutionary processes. Mol Ecol. 2016:25(4):864–881. https://doi.org/10.1111/mec.13529.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, et al. The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genom Res. 2015:25(3):445–458. https://doi.org/10.1101/gr.185579.114.

Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB. Long reads are revolutionizing 20 years of insect sequencing. Genom Biol Evol. 2021:13(8):evab138. https://doi.org/10.1093/gbe/evab138.

Immarigeon C, Frei Y, Delbare SYN, Gligorov D, Almeida PM, Grey J, Fabbro L, Nagoshi E, Billeter J-C, Wolfner MF, et al. Identification of a micropeptide and multiple secondary cell genes that modulate *Drosophila* male reproductive success. Proc Natl Acad Sci U S A. 2021:118(15):e2001897118. https://doi.org/10.1073/pnas.2001897118.

Jasper WC, Linksvayer TA, Atallah J, Friedman D, Chiu JC, Johnson BR. Large-scale coding sequence change underlies the evolution of postdevelopmental novelty in honey bees. Mol Biol Evol. 2015:32(2):334–346. https://doi.org/10.1093/molbev/msu292.

Jaworski CC, Allan CW, Matzkin LM. Chromosome-level hybrid de novo genome assemblies as an attainable option for nonmodel insects. Mol Ecol Res. 2020:20(5):1277–1293. https://doi.org/10.1111/1755-0998.13176.

Jay P, Leroy M, Le Poul Y, Whibley A, Arias M, Chouteau M, Joron M. Association mapping of colour variation in a butterfly provides evidence that a supergene locks together a cluster of adaptive loci. Philos Trans R Soc B. 2022:377(1856):20210193. https://doi.org/10.1098/rstb.2021.0193.

Jiao W-B, Schneeberger K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun. 2020:11(1):989. https://doi.org/10.1038/s41467-020-14779-y.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genom Res. 2014:24(8):1384–1395. https://doi.org/10.1101/gr.170720.113.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three *Drosophila melanogaster* mutation accumulation lines. Genom Res. 2009:19(7):1195–1201. https://doi.org/10.1101/gr.091231.109.

Khallaf MA, Auer TO, Grabe V, Depetris-Chauvin A, Ammagarahalli B, Zhang D-D, Lavista-Llanos S, Kaftan F, Weissflog J, Matzkin LM, et al. Mate discrimination among subspecies through a conserved olfactory pathway. Sci Adv. 2020:6(25):eaba5279. https://doi.org/10.1126/sciadv.aba5279.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotech. 2019:37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4.

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ERR, Pelaez J, et al. Highly contiguous assemblies of 101 drosophilid genomes. eLife. 2021:10:e66405. https://doi.org/10.7554/eLife.66405.

Kircher HW. Chemical composition of cacti and its relationship to Sonoran desert *Drosophila*. In: Barker JSF, Starmer WT, editors. Ecological genetics and evolution: the cactus-yeast-Drosophila model system. New York (NY): Academic Press; 1982. p. 143–158.

Kirkpatrick M, Barton N. Chromosome inversions, local adaptation, and speciation. Genetics. 2006:173(1):419–434. https://doi.org/10.1534/genetics.105.047985.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genom Res. 2017:27(5):722–736. https://doi.org/10.1101/gr.215087.116.

Kosakovsky-Pond SL, Poon AF, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, et al. Hyphy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. Mol Biol Evol. 2020:37(1):295–299. https://doi.org/10.1093/molbev/msz197.

Lewontin RC. The genetic basis of evolutionary change. New York (NY): Columbia University Press; 1974.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Liu Z, Roesti M, Marques D, Hiltbrunner M, Saladin V, Peichel CL. Chromosomal fusions facilitate adaptation to divergent environments in threespine stickleback. Mol Biol Evol. 2022:39(2):msab358. https://doi.org/10.1093/molbev/msab358.

Lohse K, Clarke M, Ritchie MG, Etges WJ. Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. Evolution. 2015:69(5):1178–1190. https://doi.org/10.1111/evo.12650.

Long E, Evans C, Chaston J, Udall JA. Genomic structural variations within five continental populations of *Drosophila melanogaster*. G3. 2018:8(10):3247–3253. https://doi.org/10.1534/g3.118.200631.

Löytynoja A. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Multiple sequence alignment methods. Totowa (NJ): Humana Press; 2014. p. 155–170.

Machado CA, Matzkin LM, Reed LK, Markow TA. Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. Mol Ecol. 2007:16(14):3009–3024. https://doi.org/10.1111/j.1365-294X.2007.03325.x.

Matzkin LM. Activity variation in alcohol dehydrogenase paralogs is associated with adaptation to cactus host use in cactophilic *Drosophila*. Mol Ecol. 2005:14(7):2223–2231. https://doi.org/10.1111/j.1365-294X.2005.02532.x.

Matzkin LM. Ecological genomics of host shifts in *Drosophila mojavensis*. Adv Exp Med Biol. 2014:781:233–247. https://doi.org/10.1007/978-94-007-7347-9_12.

Matzkin LM, Eanes WF. Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. Genetics. 2003:163(1):181–194. https://doi.org/10.1093/genetics/163.1.181.

Matzkin LM, Markow TA. Transcriptional differentiation across the four cactus host races of *Drosophila mojavensis*. In Michalak P, editor. Speciation: natural processes, genetics, and biodiversity. New York (NY): Nova Science Publishers Inc; 2013. p. 119–135.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends Ecol Evol. 2020:35(7):561–572. https://doi.org/10.1016/j.tree.2020.03.002.

Miller DE, Staber C, Zeitlinger J, Hawley RS. Highly contiguous genome assemblies of 15 *Drosophila* species generated using Nanopore sequencing. G3. 2018:8(10):3131–3141. https://doi.org/10.1534/g3.118.200160.

Moreyra NN, Almeida FC, Allan C, Frankel N, Matzkin LM, Hasson E. Phylogenomics provides insights into the evolution of cactophily and host plant shifts in *Drosophila*. Mol Phylogenet Evol. 2022:178:107653. https://doi.org/10.1016/j.ympev.2022.107653.

Mullen SP, Shaw KL. Insect speciation rules: unifying concepts in speciation research. Annu Rev Entomol. 2014:59(1):339–361. https://doi.org/10.1146/annurev-ento-120710-100621.
Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. Gene-wide identification of episodic selection. Mol Biol Evol. 2015:32(5):1365–1371. https://doi.org/10.1093/molbev/msv035.
Noor MAF, Grams KL, Bertucci LA, Reiland J. Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci U S A. 2001:98(21):12084–12088. https://doi.org/10.1073/pnas.221274498.
Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. Mol Biol Evol. 2012:29(11):3459–3473. https://doi.org/10.1093/molbev/mss150.
O'Donnell S, Fischer G. MUM&Co: accurate detection of all SV types through whole genome alignment. Bioinformatics. 2020:36(10):3242–3243. https://doi.org/10.1093/bioinformatics/btaa115.
Ohno S. Evolution by gene duplication. New York (NY): Springer-Verlag; 1970.
Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. Mol Phylogenet Evol. 2012:64(3):533–544. https://doi.org/10.1016/j.ympev.2012.05.012.
Orr HA. The genetic theory of adaptation: a brief history. Nat Rev Genet. 2005:6(2):119–127. https://doi.org/10.1038/nrg1523.
Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Santiago C, Lugo B, Elliott TA, Ware D, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genom Biol. 2019:20(1):275. https://doi.org/10.1186/s13059-019-1905-y.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotech. 2015:33(3):290–295. https://doi.org/10.1038/nbt.3122.
Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, Chowdhary H, Pieniak I, Maynard LJ, Gibbins MA, et al. Sequenceserver: a modern graphical user interface for custom BLAST databases. Mol Biol Evol. 2019:36(12):2922–2924. https://doi.org/10.1093/molbev/msz185.
Rampasso AS, Markow TA, Richmond MP. Genetic and phenotypic differentiation suggests incipient speciation within *Drosophila arizonae* (Diptera: Drosophilidae). Biol J Linn Soc. 2017:122(2):444–454. https://doi.org/10.1093/biolinnean/blx073.
Ravi Ram K, Wolfner MF. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. Int Comp Biol. 2007:47(3):427–445. https://doi.org/10.1093/icb/icm046.
R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2020. https://www.R-project.org.
Reed LK, Nyboer M, Markow TA. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. Mol Ecol. 2007:16(5):1007–1022. https://doi.org/10.1111/j.1365-294X.2006.02941.x.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:592(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.
Riddle NC, Elgin SCR. The *Drosophila* dot chromosome: where genes flourish amidst repeats. Genetics. 2018:210(3):757–772. https://doi.org/10.1534/genetics.118.301146.
Rius N, Delprat A, Ruiz A. A divergent P element and its associated MITE, BuT5, generate chromosomal inversions and are widespread within the *Drosophila repleta* species group. Genom Biol Evol. 2013:5(6):1127–1141. https://doi.org/10.1093/gbe/evt076.
Ross CL, Markow TA. Microsatellite variation among diverging populations of *Drosophila mojavensis*. J Evol Biol. 2006:19(5):1691–1700. https://doi.org/10.1111/j.1420-9101.2006.01111.x.
Ruiz A, Heed WB, Wasserman M. Evolution of the *Mojavensis* cluster of cactophilic *Drosophila* with descriptions of two new species. J Hered. 1990:81(1):30–42. https://doi.org/10.1093/oxfordjournals.jhered.a110922.
Russo CAM, Takezaki N, Nei M. Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol. 1995:12(3):391–404. https://doi.org/10.1093/oxfordjournals.molbev.a040214.
Sanchez-Flores A, Peñaloza F, Carpinteyro-Ponce J, Nazario-Yepiz N, Abreu-Goodger C, Machado CA, Markow TA. Genome evolution in three species of cactophilic *Drosophila*. G3. 2016:6(10):3097–3105. https://doi.org/10.1534/g3.116.033779.
Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, et al. Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics. 2008:179(3):1601–1655. https://doi.org/10.1534/genetics.107.086074.
Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol. 2019:28(6):1537–1549. https://doi.org/10.1111/mec.14794.
Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, editors. Gene prediction. Methods in molecular biology. Vol. 1962. New York (NY): Humana; 2019. p. 227–245.
Sherman CDH, Lotterhos KE, Richardson MF, Tepolt CK, Rollins LA, Palumbi SR, Miller AD. What are we missing about marine invasions? Filling in the gaps with evolutionary genomics. Mar Biol. 2016:163(10):198. https://doi.org/10.1007/s00227-016-2961-4.
Smith G, Lohse K, Etges WJ, Ritchie MG. Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*. Mol Ecol. 2012:21(13):3293–3307. https://doi.org/10.1111/j.1365-294X.2012.05604.x.
Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, Larracuente AM. Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the *simulans* clade. Mol Biol Evol. 2020:37(8):2241–2256. https://doi.org/10.1093/molbev/msaa078.
Stamatakis A. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014:30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
Stenløkk K, Saitou M, Rud-Johansen L, Nome T, Moser M, Árnyasi M, Kent M, Barson NJ, Lien S. The emergence of supergenes from inversions in Atlantic salmon. Phil Trans R Soc Lond Ser B Biol Sci. 2022:377(1856):20210195. https://doi.org/10.1098/rstb.2021.0195.
Toews DPL, Brelsford A. The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol. 2012:21(16):3907–3930. https://doi.org/10.1111/j.1365-294X.2012.05664.x.
Vanderlinde T, Dupim EG, Nazario-Yepiz NO, Carvalho AB. An improved genome assembly for *Drosophila navojoa*, the basal species in the *mojavensis* cluster. J Hered. 2019:110(1):118–123. https://doi.org/10.1093/jhered/esy059.
Villoutreix R, Ayala D, Joron M, Gompert Z, Feder JL, Nosil P. Inversion breakpoints and the evolution of supergenes. Mol Ecol. 2021:30(12):2738–2755. https://doi.org/10.1111/mec.15907.
von Grotthuss M, Ashburner M, Ranz JM. Fragile regions and not functional constraints predominate in shaping gene organization in the

genus *Drosophila*. Genom Res. 2010:20(8):1084–1096. https://doi.org/10.1101/gr.103713.109.

Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yap X, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018:28(4):581–591. https://doi.org/10.1101/gr.221028.117.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014:9(11):e112963. https://doi.org/10.1371/journal.pone.0112963.

Wang Z, Pu J, Richards C, Giannetti E, Cong H, Lin Z, Chung H. Evolution of a fatty acyl-CoA elongase underlies desert adaptation in *Drosophila*. Sci Adv. 2023:9(35):eadg0328. https://doi.org/10.1126/sciadv.adg0328.

Wellband K, Mérot C, Linnansaari T, Elliott JAK, Curry RA, Bernatchez L. Chromosomal fusion and life-history associated genomic variation contribute to within-river local adaptation of Atlantic salmon. Mol Ecol. 2019:28(6):1439–1459. https://doi.org/10.1111/mec.14965.

Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends Ecol Evol. 2018:33(6):427–440. https://doi.org/10.1016/j.tree.2018.04.002.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007:24(8):1586–1591. https://doi.org/10.1093/molbev/msm088.

Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep. 2016:6(1):31900. https://doi.org/10.1038/srep31900.

Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 2018:19(S6):153. https://doi.org/10.1186/s12859-018-2129-y.

Zhang L, Reifová R, Halenková Z, Gompert Z. How important are structural variants for speciation? Genes. 2021:12(7):1084. https://doi.org/10.3390/genes12071084.

Associate editor: Shu-Dan Yeh