Chromosome-length genome assemblies of cactophilic *Drosophila* illuminate links between structural and sequence evolution.

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

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

[2]*Department of Biology, Austin Peay State University, Clarksville, TN, USA*

[3]*Department of Zoology, Cambridge University, Cambridge, UK*

[4] *Université Côte d'Azur, INRAE, CNRS, UMR ISA, 06000 Nice, France*

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

[6]*Department of Biology, Colgate University, Hamilton, NY, USA*

[7]*BIO5 Institute, University of Arizona, Tucson, AZ, USA*

Corresponding authors: Kyle M. Benowitz, Department of Biology, Austin Peay State University, 681 Summer St., Clarksville TN, 37040. E-mail: benowitzk@apsu.edu. Luciano M. Matzkin, Department of Entomology, University of Arizona, 1104 E. South Campus Dr., Tucson, AZ, 85721. E-mail: lmatzkin@arizona.edu

Running Title: *Genomes of cactophilic Drosophila*

Abstract

A thorough understanding of adaptation and speciation requires model organisms with both a history of ecological and phenotypic study as well as a robust set of genomic resources. For decades, the cactophilic *Drosophila* species of the southwestern US and northern Mexico have fit this profile, serving as a crucial model system for understanding ecological adaptation, particularly in xeric environments, as well as the evolution of reproductive incompatibilities and speciation. Here, we take a major step towards gaining a complete molecular description of this system by assembling and annotating seven chromosome-length *de novo* genomes across the three species *D. mojavensis*, *D. arizonae*, and *D. navojoa*. Using this data, we present the most accurate reconstruction of the phylogenetic history of this clade to date. We further demonstrate a relationship between structural evolution and coding evolution both within and between species in this clade, and use this relationship to generate novel hypotheses for adaptation genes. All of 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.

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 top-down 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 cryptic or otherwise understudied phenotypes that may be contributing to ecological adaptation and speciation (Benowitz et al. 2020). With this type of bottom-up approach, genomic data may be repurposed to benefit studies of organismal natural history (Holmes et al. 2016, Sherman et al. 2016).

What genomic data is precisely needed for these purposes? For much of the genomic era, understanding genome-wide variation specifically meant understanding variation at the level of the gene. Molecular evolution at the level of the gene remains, and always will remain, a fundamental aspect of evolutionary genomic practice. However, evidence is incontrovertible that structural chromosomal variants play essential roles in adaptation and speciation. Gene duplication has long been known to be a major driver of phenotypic adaptation (Ohno 1975), while large chromosomal inversions play fundamental roles in adaptation and speciation (Noor et al. 2001; Kirkpatrick and Barton 2006). More recently, though, it is increasingly recognized that a broader variety of genomic rearrangements, including sequence gain, loss and transposition via transposable elements (Casacuberta and González 2013; Schrader and Schmitz 2019), microinversions (Redmond et al. 2020; Connallon and Olito 2021), and chromosomal fusions (Wellband et al. 2019) may also contribute to adaptation. Smaller structural variants, including insertions, deletions, and transpositions have also been increasingly shown to be implicated in speciation (Zhang et al. 2021). Following this, efforts are ongoing to create reproducible approaches to identify all types of

structural variation and quantifying their evolution across species and populations (Chakraborty et al. 2018; Wala et al. 2018; Goel et al. 2019; Heller and Vingron 2019; O'Donnell and Fischer 2020).

One challenge presented by the focus on more nuanced types of structural variation is that the fragmented, short-read assemblies that have been predominant in the world of non-model genomics may no longer suffice. Although these assemblies have been instrumental in facilitating gene expression studies and answering a wide variety of otherwise inaccessible questions regarding molecular evolution and the evolution of gene family content across a broad taxonomic range (Ellegren 2014), they offer an incomplete insight into gene duplication and none into the presence of larger structural variants (Chakraborty et al. 2018; Pollard et al. 2018; van Dijk et al. 2018). For this reason, the past several years have seen an increased emphasis on producing highly contiguous or chromosome-length genome assemblies for a broader range of organisms (Hotaling et al. 2021; Kim et al. 2021; Rhie et al. 2021). Fortunately, these efforts are being aided by both decreasing costs of long-read sequencing and the further development of methodologies to improve long-read assemblies (Amarasinghe et al. 2020; Jaworski et al. 2020; De Coster et al. 2021; Whibley et al. 2021).

Given the two-way street between genomic information and phenotypic and ecological information, we propose that the most promising study organisms will be those wherein hypotheses in both directions can be effectively leveraged; in other words, ecologically rich, tractable genomic systems with substantive empirical foundations in both areas. The cactophilic *Drosophila* within the *mulleri* complex of the *repleta* group neatly fit this description. 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 1982) proliferating in the necrosis, which is highly toxic (Kircher 1982; Fogleman and Heed 1989; Fogleman and Danielson 2001). As predicted, the transition to a cactophilic life-history has required adaptations to harsh environmental conditions including high temperatures (Schnebel and Grossfield 1984; Krebs 1999; Fasolo and Krebs 2004; MacLean et al. 2019;

Shaible and Matzkin 2022), low humidity (Gibbs and Matzkin 2001; Matzkin et al. 2007; Matzkin et al. 2009), and high toxicity (Guillén et al. 2015).

In addition to the novel colonization of their habitat, there has also been extensive ecological divergence within the cactophilic group. One subclade that has received particular attention is the *Drosophila mojavensis* species cluster, consisting of the three species *D. mojavensis*, *D. arizonae,* and *D. navojoa* (Matzkin 2014). This clade, which has diversified within the last few million years (Russo et al. 1995; Matzkin and Eanes 2005; 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 (Matzkin 2014; Etges 2019): one in the Sonoran Desert inhabiting organ pipe cactus (*Stenocereus thurberi*), one in Baja California inhabiting agria (*Stenocereus gummosus*), one in the Mojave Desert inhabiting red barrel cactus (*Ferrocactus cylindraceus*), and one on Santa Catalina Island (CA) 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 (*O. wilcoxii*; Heed 1982). These distinctions within and between species have proved to be an extremely fruitful substrate for hypotheses regarding phenotypic adaptation in many forms, including heat tolerance (Diaz et al. 2021a), desiccation resistance (Matzkin et al. 2007; Rajpurohit et al. 2013), chemical adaptation (Starmer et al. 1977); life-history (Etges 1990), olfaction (Date et al. 2013; Crowley-Gall 2016; 2019; Nemeth et al. 2018; Ammagarahalli et al. 2021), and behavior (Newby and Etges 1998; Coleman et al. 2018). Additionally, the recent divergence within and between species in the *D. mojavensis* species complex has made this clade into a model system for speciation and the evolution of reproductive incompatibilities (Markow 1981; Pantazidis and Zouros 1988; Zouros et al. 1988; Etges 1992; Knowles and Markow 2001; Miller et al. 2003; Pitnick et al. 2003; Reed and Markow 2004; Massie and Markow 2005; Kelleher and Markow 2007; Markow et al. 2007; Bono et al. 2011, 2015; Hardy et al. 2011; Richmond et al. 2012; Richmond

2014; McGirr et al. 2017; Diaz et al. 2021b; 2022)

The relatively close phylogenetic relationship to *D. melanogaster* has provided the *D. mojavensis* cluster with several advantages as a burgeoning genomic model system. The ability to detect chromosomal inversions via the analysis of polytene chromosomes, pioneered in *D. melanogaster*, allowed for early investigations into clinal variation as well as interpopulation and interspecific variation in inversions (Mettler 1963; Johnson 1980). More recently, the *D. mojavensis* population from Santa Catalina Island was among the first non-model *Drosophila* genomes sequenced (*Drosophila* 12 Genomes Consortium 2008), 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 *D. melanogaster* has allowed for clear interpretation of gene-level results as compared to more distantly related insects. This has been leveraged in a slew of candidate gene studies (Krebs 1999; Matzkin and Eanes 2003; Matzkin 2004; 2005; 2008; Guillén and Ruiz 2012; Diaz et al. 2018), whole-genome studies of molecular evolution (Guillén et al. 2015; 2019; Allan and Matzkin 2019; Rane et al. 2019), transcriptomics (Matzkin et al. 2006; Bono et al. 2011; Matzkin and Markow 2009; 2013; Matzkin 2012; Rajpurohit et al. 2013; Smith et al. 2013; Etges et al. 2015, 2017; Crowley-Gall et al. 2016; Nazario-Yepiz et al. 2017; Mateus et al. 2019; Benowitz et al. 2020; Banho et al. 2021ab; Diaz et al. 2021b; 2022), and functional analysis via CRISPR derived transgenics (Khallaf et al. 2020).

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, in spite of the outsized role that these populations have played in understanding molecular adaptation to variable host environments. Although this genome assembly is excellent, and has capably facilitated genetic mapping studies (Etges et al. 2007, 2009, 2010; Benowitz et al. 2019), several chromosomes, notably the X chromosome, remain far from contiguous. Outside of *D. mojavensis*, there are currently only two highly fragmented genome assemblies, one each from *D.*

133 *navojoa* and *D. arizonae* (Sanchez-Flores et al. 2016; Vanderlinde et al. 2019).

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 US	2002	15081-1352.22	15081-1352.22
	MOV	Anza-Borrego State Desert Park, California US	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

134

135 Here, we take a major step towards addressing this gap and lack of genomic resources by first re-
136 scaffolding and polishing the existing *D. mojavensis* genome from Santa Catalina Island, substantially
137 improving this assembly. We then build *de novo* genome assemblies for strains from the other three *D.*
138 *mojavensis* populations, two strains of *D. arizonae* collected from opposite ends of its range, and a single
139 strain of *D. navojoa* (Fig. 1; Table 1). By using a hybrid assembly approach combining short- and long-
140 read sequencing technologies, we are able to construct entire chromosomes for each genome, leading to
141 some of the most complete assemblies throughout the *Drosophila* clade. We first use these assemblies to
142 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 allows us novel insight into the rates of each type of evolutionary divergence in this clade, and also provides the power to test fundamental hypotheses on the relationship between structural and coding evolution. Lastly, in order to facilitate 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).

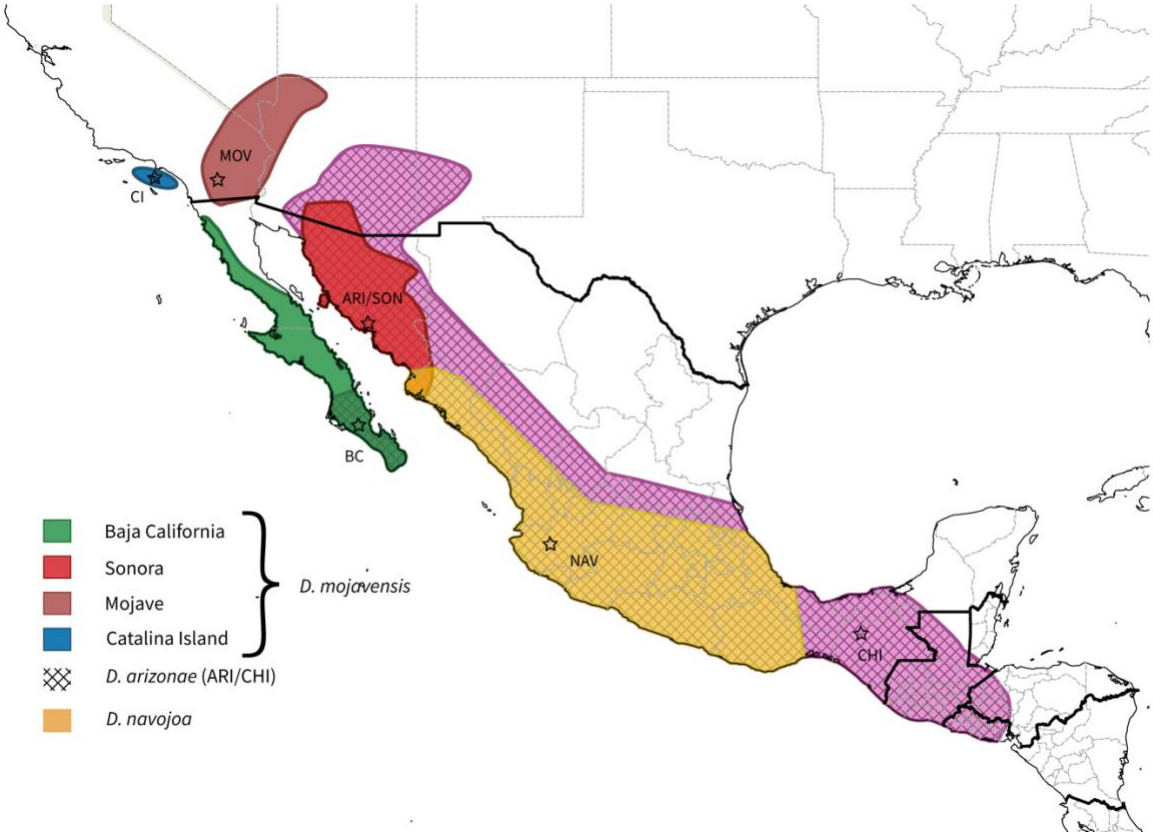

Figure 1. Ranges of the species and populations sequenced in this study. Hatched regions represent the range of *D. arizonae*. No discrete geographical boundary is known to separate the ARI and CHI populations sequenced here. Stars show the location of collection of the genome lines. Ranges are estimated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

Results

Genome assembly and annotation

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 (Mb)	12.26	0	0	0	0	0	0
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	13,755	13,358	13,388	13,426	13,408	13,321	13,203
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	99.3	98.9	98.8	99.0	98.9	98.9	99.1
Single Copy	98.8	98.4	98.4	98.5	98.4	98.4	98.6
Duplicated	0.5	0.5	0.4	0.5	0.5	0.5	0.5
Fragmented	0.3	0.5	0.5	0.5	0.5	0.5	0.4
Missing	0.4	0.6	0.7	0.5	0.6	0.6	0.5

155 Details of the strains used for genome sequencing can be found in Table 1. Genome assembly and
156 annotation statistics can be found in Table 2. All six *de novo* assemblies were highly contiguous, with all
157 six major chromosomes assembled with only a handful of gaps. The re-scaffolding of the original CI
158 genome also resulted in higher contiguity, although the assembly still has a higher percentage of gaps as
159 well as repeats, which could indicate the presence of redundant scaffolds.

160 Genome size was consistent across the three new *D. mojavensis* genomes, with both *D. arizonae*
161 exhibiting larger genomes and *D. navojoa* slightly smaller. There is no evidence that these evolutionary
162 patterns in genome size are driven by expansions of repeats or TEs, as these did not display such
163 consistent trends between the species.

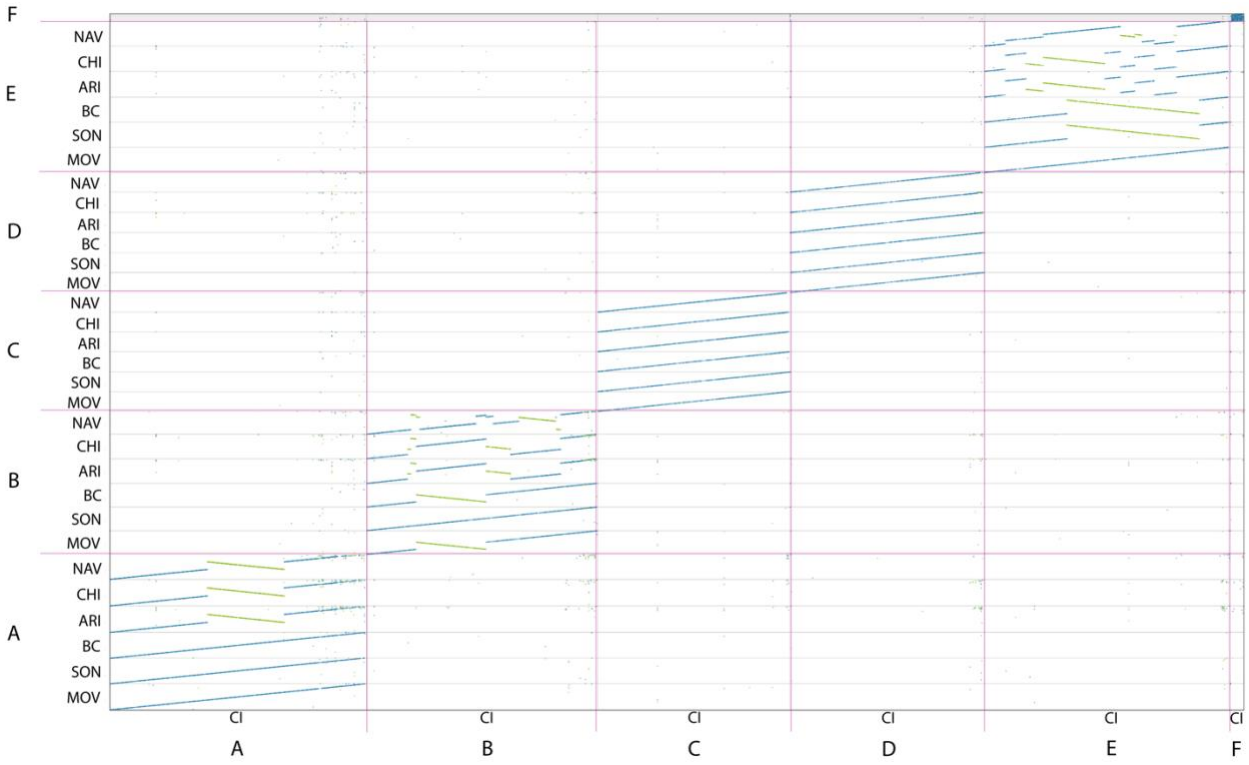

Figure 2. Syntenic conservation of the six *de novo* genomes sequenced in this study as compared to the reassembled CI genome. Letters A-F indicate Muller elements. Green lines indicate inversions.

164

165 Our genome assemblies confirmed previous findings on fixed chromosomal inversions between these
166 species and populations (Fig. 2; Supplemental Figs. S1, S2), 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. Some inversion breakpoints were associated with windows of high repeat and TE content (Fig. 3; Supplemental Fig. S3), but this pattern was not ubiquitous.

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.

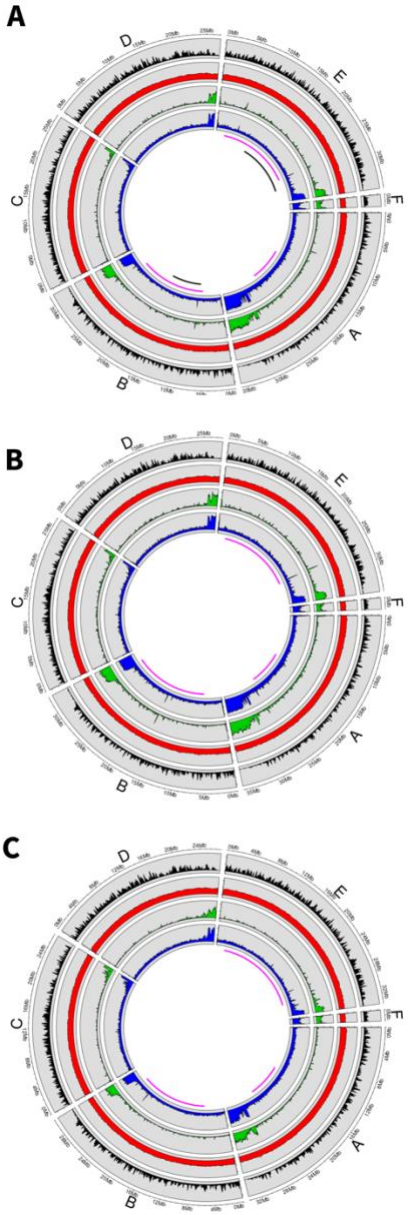

Figure 3. Genome statistics for: (*A*) Mojave *D. mojavensis*, (*B*) Chiapas *D. arizonae*, and (*C*) *D. navojoa*. From outside to in, circles represent gene content, GC content, TE content, and total repeat content. Pink bars below the circles represent the regions covered by interspecific inversion polymorphisms, and black bars represent regions covered by inversion polymorphisms within *D. mojavensis*.

Phylogenomics and divergence time estimation

Both the topology of the phylogeny as well as the divergence time estimates differed when using nuclear (Fig. 4) versus mitochondrial genes (Supplemental Fig. S4). 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. 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 were most closely related, although other aspects of the topology within *D. mojavensis* also differed.

Although the timing of divergence within the *D. arizonae*/*D. mojavensis* clade was identical between the two datasets, the mitochondrial phylogeny gave a twice as old split of *D. navojoa* from this group.

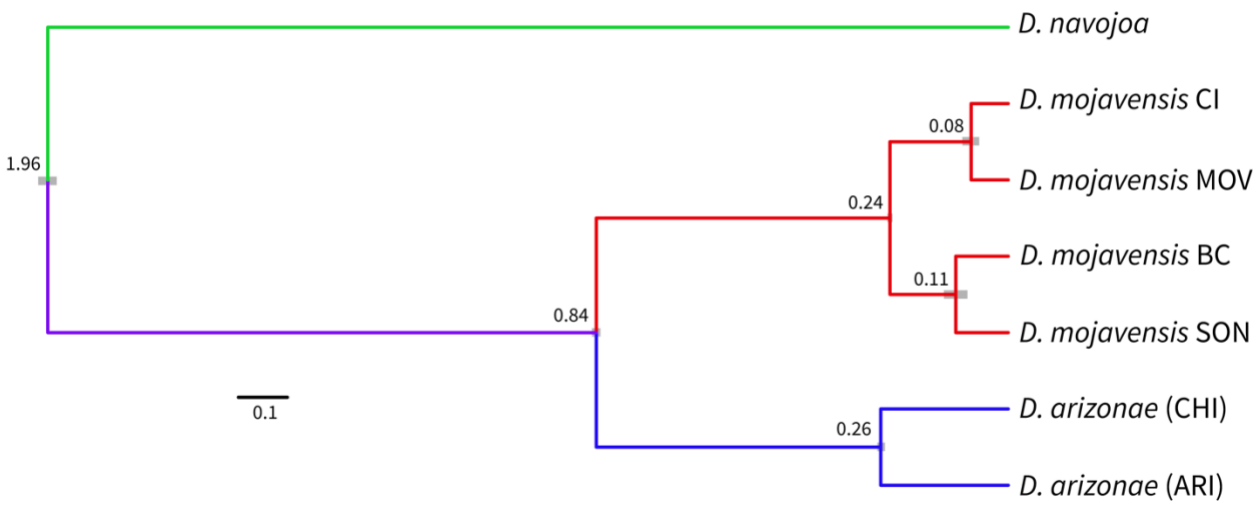

Figure 4. Phylogeny and divergence times (mya) as estimated by 12,218 single copy nuclear genes. Colors represent the accepted species identities and grey bars represent 95% confidence intervals for divergence time estimates.

Syntenic evolution

We defined syntenic (or collinear) regions of the genome as those displaying one-to-one conservation of sequence as called by SyRI (Goel et al. 2019) and syntenic divergence as the percentage of non-syntenic

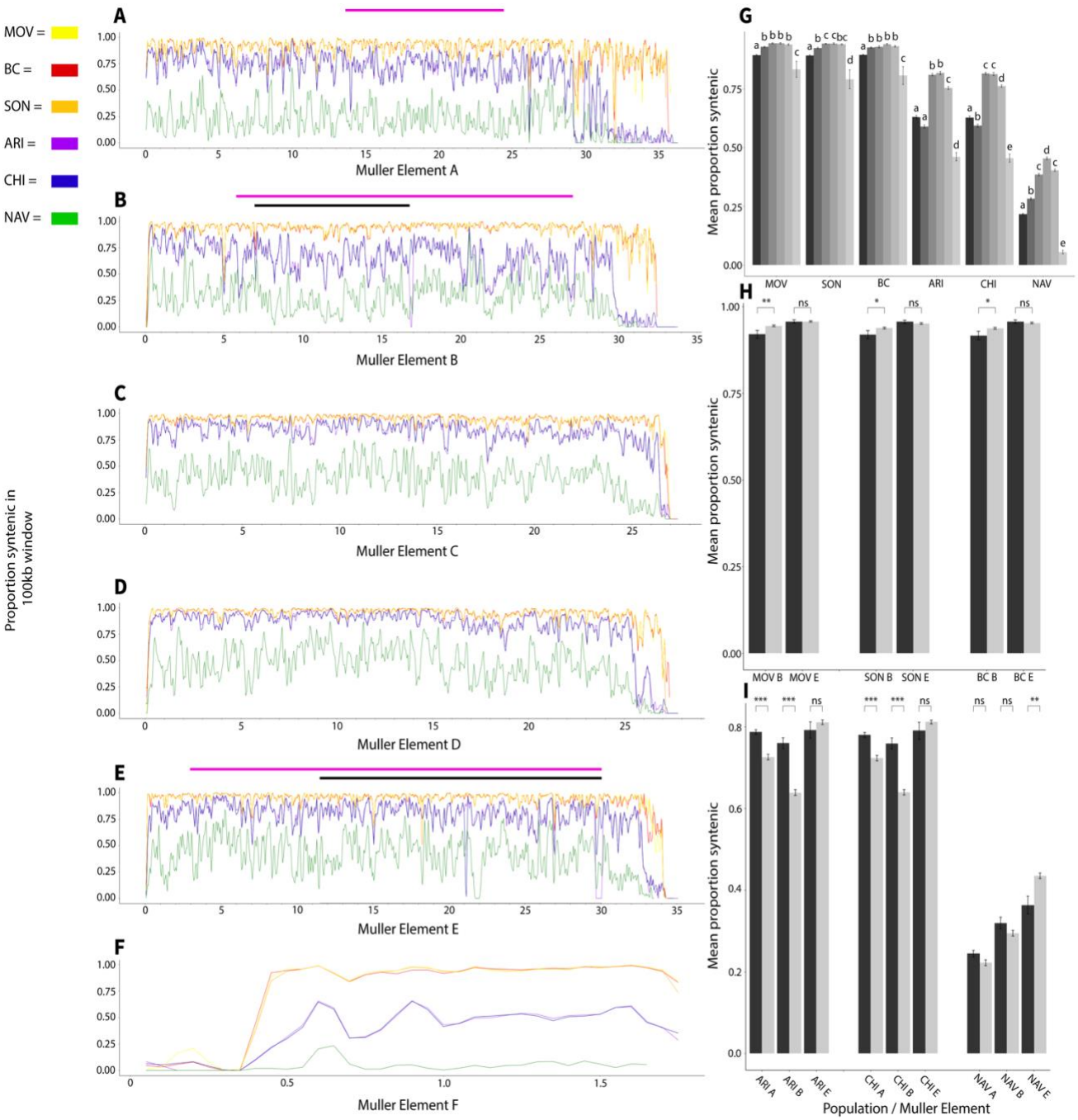

Figure 5. (*A-F*) The synteny score across 100kb windows for Muller elements A-F. MOV, yellow; SON, orange; BC, red; ARI, purple; CHI, blue; NAV, green. Pink and blue bars above plots indicate inter- and intraspecific inversion polymorphisms as in Figure 2. Numbers on x axis indicate position in Mb. (*G*) Mean synteny 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 synteny scores before (dark grey) and within (light grey) inverted regions of elements B and E for the three *D. mojavensis* populations. (*I*) Mean synteny scores before (dark grey) and within (light grey) 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 $p < 0.0001$ (***).

genome content between two genomes in a given region. Syntenic divergence of each of the six *de novo* sequenced populations from CI recapitulated the divergence patterns as found in the nuclear phylogeny. As expected, *D. navojoa* had by far the greatest mean syntenic divergence, while both *D. arizonae* populations had nearly identical levels of divergence. MOV had slightly higher synteny compared to BC and SON. Overall, the breakdown of synteny over evolutionary time was found to be linear, with a loss of roughly 33.67% of genome collinearity per million years (Supplemental Fig. S5)

Independently of chromosomal inversions, which were rearranged in all seven genomes to match the CI karyotype prior to analysis, significant variation in syntenic divergence was present between chromosomes. Muller elements A and F showed reduced synteny in all six genomes. In *D. arizonae,* Muller elements B and E, which also carry inversions, displayed lower synteny than C and D, which do not carry inversions. Patterns in *D. navojoa* were similar apart from a reduction in synteny on Muller element C compared to D (Fig. 5A-G).

Within chromosomes bearing inversions, the relative rates of syntenic divergence inside and outside (measured here as only the region on the centromeric side of the inversion due to low synteny 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 (we did not compare regions after the inversions due to major reductions of synteny in telomeres; Fig. 5A-F) in all three populations, including the SON population, which is homokaryotypic with CI (Fig. 5H). However, synteny in Muller element E was consistent before and within the inversion. Interspecific syntenic 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. 5I).

Molecular evolution

Lists of genes found to be under positive selection via BUSTED and codeml analyses can be found in

224 Supplemental Tables S1 and S2. A comparison of gene families previously hypothesized to be involved in
225 adaptation to variable cactus environments showed no elevated rates of positive selection in these gene
226 families via the codeml analysis (Fig. 6). However, higher rates of positive selection were found within
227 reproductive genes as well as orphan genes absent from *D. melanogaster*.

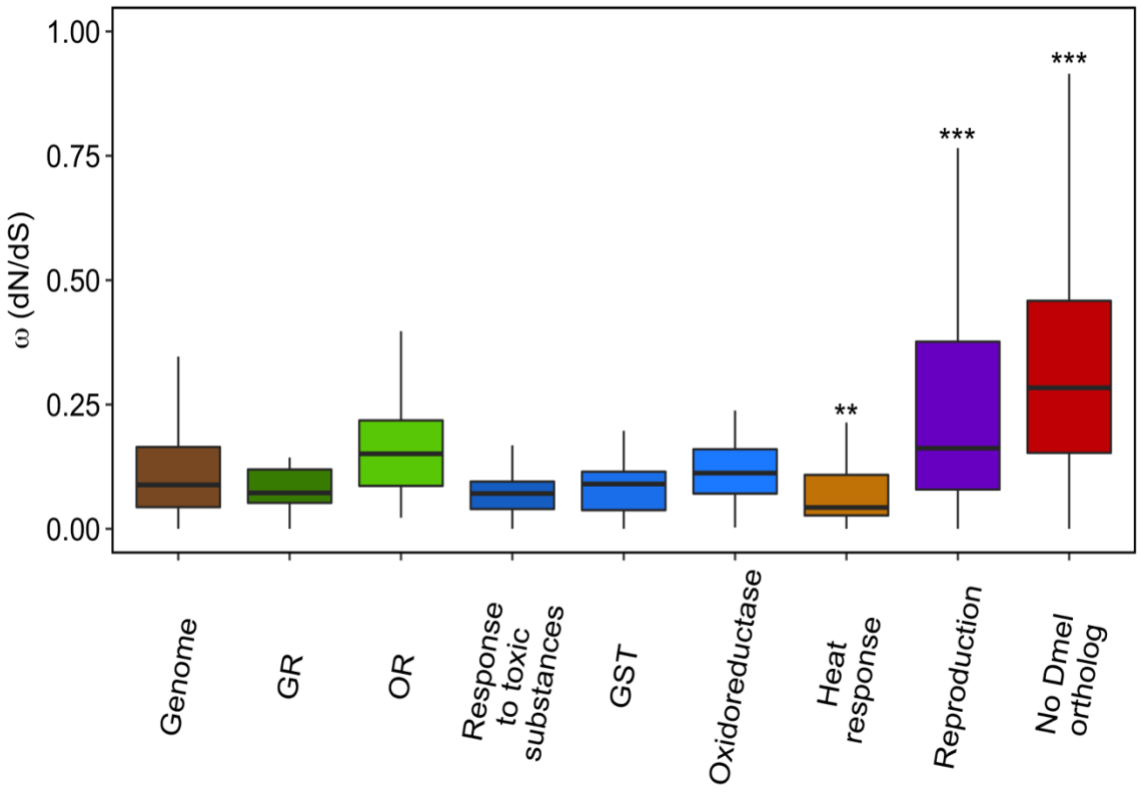

Figure 6. Comparison of omega values for different gene families and GO categories. Asterisks indicates significance at the level of $p < 0.001$ (**) or $p < 0.0001$ (***).

228

229 We found no evidence that genes surrounding either the breakpoints within *D. mojavensis* ($F_{1,12185}$ =
230 0.017, p = 0.68) nor the breakpoints in the clade as a whole ($F_{1,12185}$ = 0.33, p = 0.57) displayed elevated
231 evolutionary rates. Rates of positive selection were significantly negatively corelated to the synteny score
232 from CI to NAV of the sliding window containing the gene (Fig. 7). This pattern held for the synteny from
233 CI to the mean of the *D. arizonae* populations ($F_{1,12185}$ = 44.80, p = 2.28 x 10^{-11}) as well as the synteny
234 from CI to the mean of the other three *D. mojavensis* populations ($F_{1,12185}$ = 19.89, p = 8.26 x 10^{-6}).

235

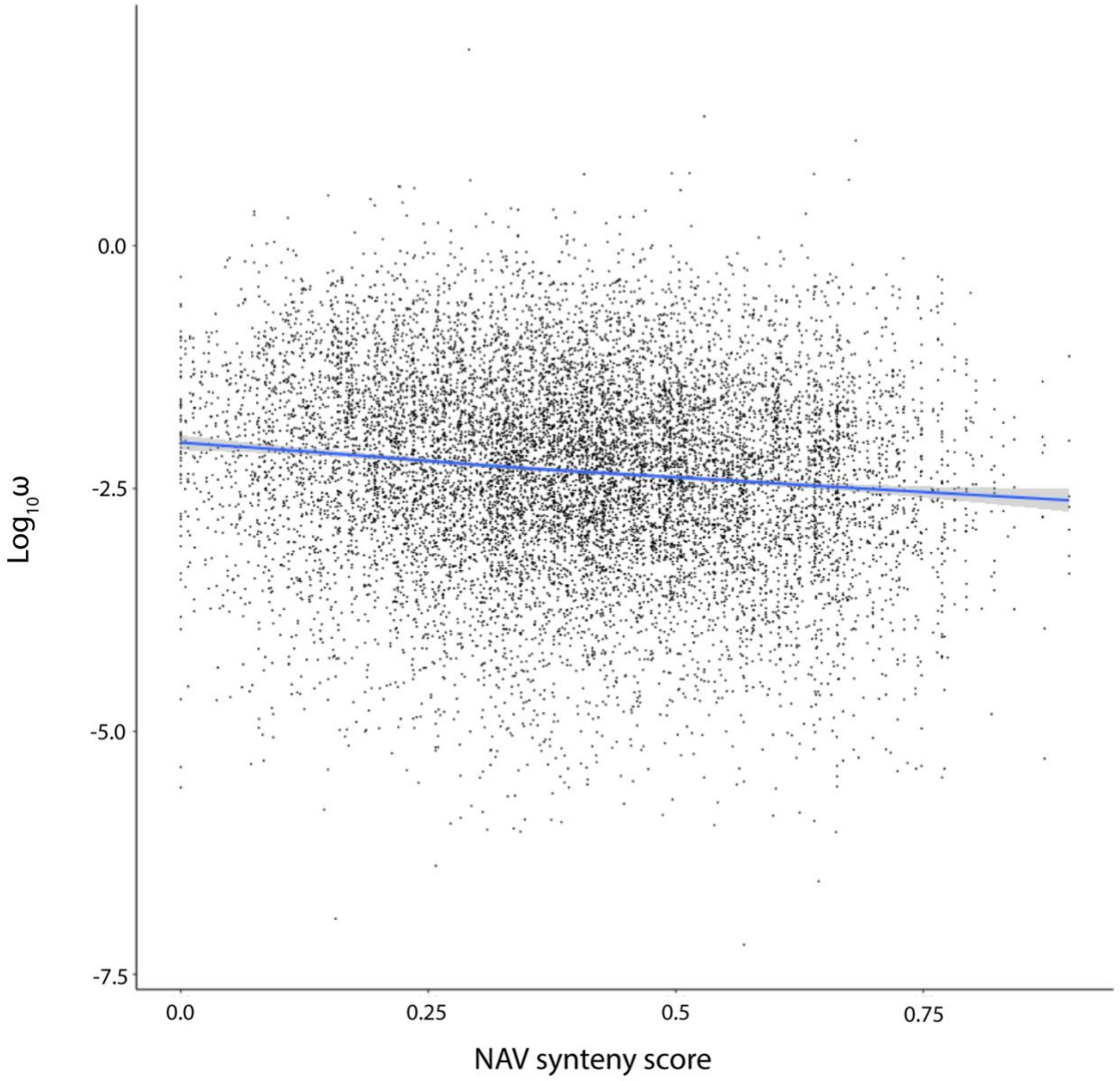

Figure 7. The genome-wide relationship between synteny (from *D. mojavensis* Catalina Island to *D. navojoa*) and molecular evolutionary rate across the phylogeny. The regression line represents a linear regression (y = -0.36x – 0.98), with 95% confidence intervals shaded ($F_{1,12153} = 94.64$, $p < 2.2x10^{-16}$).

236

237

Discussion

The assembly of these seven *de novo* genomes represents one of the largest and most contiguous genomic datasets in any *Drosophila* clade. To date, we can identify 21 genomes that have been assembled at or near chromosome level in *Drosophila* (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/Drosophila). However, these are heavily clustered in clades closely related to *D. melanogaster* and *D. pseudoobscura*. While there has been a clear logic to concentrating sequencing effort on a small handful of clades, the establishment of generalizable patterns of genomic evolution requires that analyses be replicated across taxa. Our genomes add to this endeavor. Furthermore, the inclusion of separate conspecific populations of two species allows for deeper resolution in the calculation of fundamental evolutionary patterns.

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. 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). These studies also presented variable divergence times; nuclear studies (Machado et al. 2007; 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 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 analysis in previous mitochondrial work (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 syntenic 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). However, we consider this discrepancy to be more likely due to noise, given the nearly 10,000-fold difference in genes analyzed (12,218 vs. 13) in the nuclear genome compared to the mitochondrial genome. Thus, our data best support the topology of the earlier nuclear phylogeny of Machado et al. (2007), with divergence times broadly in agreement with both previous nuclear studies (Machado et al. 2007; Smith et al. 2012).

Our results regarding divergence times between species significantly contradicted estimates from earlier work. These estimates have ranged widely, ranging from 0.66 mya 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* (reviewed in 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 three reasons. First, our analysis utilized more than twice as many loci, as we analyzed more than 12,000 single copy orthologs. Second, 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. Third, 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 the cyclic climatic fluctuations of the past few million years and the accompanying shifts in host cactus distribution. On the contrary, major geological events such as the raising of the Trans-Mexican Volcanic Belt, that have been previously 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.

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. Chakraborty et al. (2021) found that 15% of sequence did not align between *D. 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) estimate 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, amounts to approximately 13% divergence per million years. Jiao and Schneeberger (2020) report around 10% syntenic divergence between *Arabidopsis thaliana* accessions using the same software and methodology here, but cannot present a phylogenetic timeline of breakdown. Here, we observe that synteny decays in a linear fashion, with about 33% of genome collinearity lost per million years. Although this conclusion is undoubtedly sensitive to the particular parameters used to define a syntenic block, we hope this can serve as a baseline for future studies quantifying the evolution of genome structure both within *Drosophila* as well as in other taxa. Given increasing implication of structural variants in speciation (Zhang et al. 2021) and adaptation (Mérot et al. 2020), we are curious to see how these results stand out in an even broader comparative framework. Is this a fast or slow rate of structural divergence? Do structural variants accumulate more rapidly in highly speciose taxa, or those undergoing adaptive radiations?

To begin to address these questions within our dataset, we asked what factors might explain local variation in synteny within the *D. mojavensis* group. The best predictor of syntenic 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 (Supplemental Fig. S2). 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 (Hoffman et al. 2004; Kirkpatrick and Barton 2006; Wellenreuther and Bernatchez 2018). Non-adaptive 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. Further supporting that this relationship is correlational, we see no evidence that inversions cause additional decreases in synteny, 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 synteny was negatively linked to omega. In other words, genes with greater rates of protein-coding evolution were more likely to occur in regions of decreased collinearity. 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 changes (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 resulting in alterations to gene regulation alter protein function, which subsequently leads to a relaxation of purifying selection on amino acid sequences.

Given this relationship, we consider genes with elevated rates of positive selection in regions of low collinearity to be especially strong 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 synteny score in the 6th percentile or lower in all three species. 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 *D. 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. (2021b). 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 220X coverage. All other short-read libraries were built and sequenced on Illumina HiSeq 2000 at the HudsonAlpha Genome Sequencing Center (Huntsville, AL, USA) at 75X 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 sequenced on a PacBio Sequel at the Arizona Genomics Institute (Tucson, AZ, USA).

406 The assembly of the six *de novo* genomes largely followed the hybrid assembly strategy described in
407 Jaworski et al. (2020), wherein a detailed description of sequencing and assembly methods can be found.
408 Briefly, we used Platanus 1.2.4 (Kajitani et al. 2014) and DBG2OLC (Ye et al. 2016) to produce hybrid
409 assemblies of the short- and long-read data. We also used Canu 1.7 (Koren et al. 2017) for long-read only
410 assembly with the correctedErrorRate parameter set to 0.039 for the primary assembly though this was
411 increased to 0.065 to produce a less stringent assembly used for bridging and extending primary contigs.
412 We used Quickmerge 2.0 (Chakraborty et al. 2016) to merge these two assemblies into a draft assembly.
413 We then manually merged contigs based on whole genome alignments from Mauve (Darling et al. 2004)
414 and Nucmer (Delcher et al. 2002) including using the less stringent assembly in Geneious Prime
415 (Biomatters, Auckland, NZ). Where contigs could not be merged, we manually joined them based on
416 alignment with the other genomes and connected with an N-gap of 100bp. We aligned all remaining
417 contigs not assigned to a chromosome with Minimap2 (Li 2018) and subsequently discarded all contigs
418 with a match of over 80% to a chromosome scaffold. We polished each genome three times with Pilon
419 1.23 (Walker et al. 2014). During manual curation of our annotations with the help of RNA-seq data (see
420 below), we identified several small insertion/deletion errors in each genome that led to frameshift errors
421 causing problems with gene structure, and subsequently fixed these errors manually in Geneious Prime.
422 We noticed that the *D. arizonae* Chiapas genome had substantially more of these errors than the others
423 and therefore polished it a fourth time with Pilon 1.23 before fixing remaining errors manually as for the
424 other genomes. We also performed additional polishing of the gene-containing regions of *D. navojoa*
425 using majority consensus in Geneious Prime.

426 We reassembled the *D. mojavensis* Santa Catalina Island genome (hereafter, CI) in order to provide better
427 comparisons with the six *de novo* assemblies. We first polished the most current FlyBase assembly
428 (version r1.04) twice with Pilon 1.23. We then manually scaffolded by aligning contigs from the existing
429 Nanopore data (Miller et al. 2018) to the polished reference using Mauve and joining in Geneious Prime.
430 We filled all N-gaps over 20kb with contigs from the Nanopore dataset. Lastly, we filled all N-gaps

regardless of size if they occurred within 100bp 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 ~1.3Mb of sequence that was syntenic to sequence in Muller element A (X chromosome). We therefore split the CI Muller element F into two pieces: we kept bp 1-2,135,734 as chromosome F, while we joined bp 2,139,764-3,406,379 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 2008) 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 hours post-laying) and late (26 hours 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. 2010) 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 (*Drosophila* 12 Genomes Consortium 2008) 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. 2019) 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 (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 100kb sliding windows overlapping by 50kb, 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. We then re-ran the software on the remaining genes using stricter parameters, and repeating this procedure twice. In this way, we were able to capture genes that may have duplicated recently but before the common ancestor of the three species, and therefore still useful for our analyses.

We then performed codon-aware alignments of all single-copy orthologs using PRANK (Löytynoja 2014), and extracted fourfold degenerate sites from each alignment. We generated individual, unrooted gene trees using RAxML (Stamatakis 2014), and used these trees as input for consensus tree building using ASTRAL-III (Zhang et al. 2018) and MP-EST (Liu et al. 2010). All programs were run using default parameters.

After establishing a consensus tree topology, we used BPP (Flouri et al. 2018) on the entire genome and = (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 divergence time in *Drosophila* (Obbard et al. 2012) and previous work on *D. mojavensis* (Smith et al. 2012; Lohse et al. 2015), we used a neutral mutation rate of 3.5×10^{-9} (Keightley et al. 2009) and a rate of six generations per year 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 fourfold 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

523 years.

524 *Analysis of structural genome evolution*

525 We aligned all seven genomes using NUCmer in order to identify breakpoints and visualize previously
526 identified chromosomal inversions on Muller elements A, B, and E. We made figures of genome wide
527 synteny using Dot (https://github.com/marianattestad/dot). Prior to analyzing structural variation
528 quantitatively, we used these breakpoints to manually create "uninverted" chromosomes, wherein we
529 forced all chromosomes to be homokaryotypic with CI. This allowed us to compare synteny inside and
530 outside of major inversions in an unbiased manner. We re-ran NUCmer on the "uninverted" genome
531 assemblies and used this output as input for identification of structural variation and syntenic genome
532 regions using SyRI (Goel et al. 2019). Using the CI genome as our template, we followed Jiao and
533 Schneeberger (2020) in quantifying the percentage of syntenic sequence in 100kb regions of the genome
534 over 50kb sliding windows using custom bash scripts. We compared synteny across chromosomes within
535 each genome using ANOVA. For Muller element F, we calculated chromosome-wide synteny after
536 removing ~350 kb at the centromeric end of the CI chromosome, which may be a misassembly as it has
537 no corresponding region on any of the six *de novo* assemblies. For each chromosome with an inversion,
538 we additionally compared the synteny outside the inversion on the centromeric end to the synteny within
539 the inversion region using ANOVA. The region outside the inversion only included the region on the
540 centromeric side of the inversion. We did not compare the non-inverted region on the telomeric end due
541 to the extreme degradation of synteny near the telomere, especially in the interspecific comparisons.
542

543 *Analysis of molecular evolution*

544 For molecular evolutionary analyses, we used the same set of aligned single-copy orthologs as used above
545 in phylogenomic analyses, and used the best supported phylogeny from the analysis above. We analyzed

546 dN/dS of each sequence across the entire phylogeny using Codeml (PAML; Yang 2007) with models 0, 7,
547 and 8.

548 We considered two hypotheses regarding the relationship between structural and coding sequence
549 evolution. First, we predicted that genes proximal to the inversion breakpoints would be more likely to
550 experience positive selection. We tested this prediction by comparing the proportion of significantly
551 positively selected genes within 1Mb on either end of a breakpoint to the rest of the genes in the genome.
552 Second, we predicted that genes in regions of low synteny would be more likely to display signatures of
553 positive selection. To examine this prediction, we performed linear regression to examine the relationship
554 between the $\log_{10}\omega$ value of each gene and the synteny score between CI and NAV of the 100kb window
555 containing the gene. We chose to display NAV as the source of syntenic data due to the fact that it
556 displays the greatest variation in synteny while remaining correlated with structural variation in the other
557 genomes ($r_{NAV\text{-}MOJ} = 0.48$, $r_{NAV\text{-}ARI} = 0.73$). However, we additionally performed the same analysis on the
558 mean synteny scores of the two *D. arizonae* genomes and the three remaining *D. mojavensis* genomes to
559 confirm this pattern. We performed all statistical analyses in R 3.6.3 (R Core Team 2020).

560

561 **Data access**

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

565

566 **Competing interest statement**

567 The authors declare no competing interests.

Acknowledgments

We thank D. Kudrna for his work to produce the PacBio sequences. We thank N. Sage for assistance with genome annotations. This work was supported by the National Science Foundation (IOS-1557697 to L.M.M.). 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 with input from all authors.

References

Allan CW, Matzkin LM. 2019. Genomic analysis of the four ecologically distinct cactus host populations of *Drosophila mojavensis*. *BMC Genom* **20:** 732.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritche ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genom Biol* **21:** 30.

Ammagarahalli B, Layne JE, Rollmann SM. 2021. Host plant shift differentially alters olfactory sensitivity in female and male *Drosophila mojavensis*. *J Ins Physiol* **135:** 104312.

Arendsee ZW, Li L, Wurtele ES. 2014. Coming of age: orphan genes in plants. *Trends Plant Sci* **19:** 698-

589 708.

590 Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et
591 al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*
592 **587:** 246-251.

593 Banho CA, Mérel V, Oliveira TY, Carareto CM, Vieira C. 2021a. Comparative transcriptomics between
594 *Drosophila mojavensis* and *D. arizonae* reveals transgressive gene expression and
595 underexpression of spermatogenesis-related genes in hybrid testes. *Sci Rep* **11:** 9844.

596 Banho CA, Oliveira DS, Haudry A, Fablet M, Vieira C, Carareto CMA. 2021b. Transposable element
597 expression and regulation profile in gonads of interspecific hybrids of *Drosophila arizonae* and
598 *Drosophila mojavensis wrigleyi*. *Cells* **10:** 3574.

599 Barrett RDH, Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev*
600 *Genet* **12:** 767-780.

601 Bell EA, Butler CL, Oliveira C, Marburger S, Yant L, Taylor MI. 2022. Transposable element annotation
602 in non-model species: The benefits of species-specific repeat libraries using semi-automated
603 EDTA and DeepTE de novo pipelines. *Mol Ecol Res* **22:** 823-833.

604 Benowitz KM, Coleman JM, Matzkin LM. 2019. Assessing the architecture of *Drosophila mojavensis*
605 locomotor evolution with bulk segregant analysis. *G3* **9:** 1767-1775.

606 Benowitz KM, Coleman JM, Allan CW, Matzkin LM. 2020. Contributions of *cis*- and *trans*-regulatory
607 evolution to transcriptomic divergence across populations in the *Drosophila mojavensis* larval
608 brain. *Genome Biology and Evolution* **12:** 1407-1418.

609 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data.

Bioinformatics **30:** 2114-2120.

Bono JM, Markow TA. 2009. Post-zygotic isolation in cactophilic *Drosophila*: larval viability and adult life-history traits of *D. mojavensis/D. arizonae* hybrids. *J Evol Biol* **22:** 1387-1395.

Bono JM, Matzkin LM, Kelleher ES, Markow TA. 2011. Postmating transcriptional changes in reproductive tracts of con- and heterospecifically mated *Drosophila mojavensis* females. *Proc Natl Acad Sci USA* **108:** 7878-7883.

Bono JM, Matzkin LM, Hoang K, Brandsmeier L. 2015. Molecular evolution of candidate genes involved in post-mating-prezygotic reproductive isolation. *J Evol Biol* **28:** 403-414.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature* **512:** 393-399.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, et al. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genom Biol* **17:** 66.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. *Curr Prot Bioinf* **48:** 4.11.1-4.11.39.

Card DC, Adams RH, Schield DR, Perry BW, Corbin AB, Pasquesi GIM, Row K, Van Kleeck MJ, Daza JM, Booth W, et al. 2019. Genomic basis of convergent island phenotypes in boa constrictors. *Genom Biol Evol* **11:** 3123-3143.

Casacuberta C, González J. 2013. The impact of transposable elements in environmental adaptation. *Mol Ecol* **22:** 1503-1517.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo

assembly of metazoan genomes with modest long read coverage. *Nuc Ac Res* **44:** e147.

Chakraborty M, VanKuren NW, Zhao R, Zhang W, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet* **50:** 20-25.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Comms* **10:** 4872.

Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. 2021. Evolution of genome structure in the *Drosophila simulans* species complex. *Genom Res* **31:** 380-396.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10:** 563-569.

Coleman JM, Benowitz KM, Jost AG, Matzkin LM. 2018. Behavioral evolution accompanying host shifts in cactophilic *Drosophila* larvae. *Ecol Evol* **8:** 6921-6931.

Connallon T, Olito C. 2021. Natural selection and the distribution of chromosomal inversion lengths. *Mol Ecol* DOI: 10.1111/mec.16091.

Crowley-Gall A, Date P, Han C, Rhodes N, Andolfatto P, Layne JE, Rollmann SM. 2016. Population differences in olfaction accompany host shift in *Drosophila mojavensis*. *Proc R Soc Lond B* **283:** 20161562.

Crowley-Gall A, Shaw M, Rollmann SM. 2019. Host preference and olfaction in *Drosophila mojavensis*. *J Hered* **110:** 68-79.

Darling ACE, Mau B, Blatner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic

652 sequence with rearrangements. *Genom Res* **14:** 1394-1403.

653 Date P, Dweck HK, Stensmyr MC, Shann J, Hansson BS, Rollmann SM. 2013. Divergence in olfactory
654 host plant preference in *D. mojavensis* in response to cactus host use. *PLoS ONE* **8:** e70027.

655 De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing.
656 *Nat Rev Genet* **22:** 572-587.

657 Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment
658 and comparison. *Nuc Ac Res* **30:** 2478-2483.

659 Delprat A, Guillén Y, Ruiz A. 2019. Computational sequence analysis of inversion breakpoint regions in
660 the cactophilic *Drosophila mojavensis* lineage. *J Hered* **110:** 102-117.

661 Diaz F, Allan CW, Matzkin LM. 2018. Positive selection at sites of chemosensory genes is associated with
662 the recent divergence and local ecological adaptation in cactophilic *Drosophila*. *BMC Ecol Evol*
663 **18:** 144.

664 Diaz F, Kuijper B, Hoyle RB, Talamantes N, Coleman JM, Matzkin LM. 2021a. Environmental
665 predictability drives adaptive within- and transgenerational plasticity of heat tolerance across life
666 stages and climatic regions. *Func Ecol* **35:** 153-166.

667 Diaz F, Allan CW, Markow TA, Bono JM, Matzkin LM. 2021b. Gene expression and alternative splicing
668 dynamics are perturbed in female head transcriptomes following heterospecific copulation. *BMC*
669 *Genom* **22:** 359.

670 Diaz F, Allan CW, Chen X, Coleman JM, Bono JM, Matzkin LM. 2022. Divergent evolutionary
671 trajectories shape the postmating transcriptional profiles of conspecifically and heterospecifically
672 mated cactophilic *Drosophila* females. *Communications Biology* **5:** 842 doi:10.1038/s42003-022-

03758-2

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genom Res* **13:** 2213-2219.

Drosophila 12 Genomes Consortium. 2008. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203-218.

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elsik CG, Lewis SE. 2019. Apollo: democratizing genome annotation. *PLoS Comp Bio* **15:** e1006790.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol* **29:** 51-63.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genom Biol* **20:** 238.

Etges WJ. 1990. Direction of life history evolution in *Drosophila mojavensis*. In Barker JSF, Starmer WT, MacIntyre RJ (eds.) *Ecological and evolutionary genetics of drosophila*. Springer, New York, NY. pp. 37-56.

Etges WJ. 1992. Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis. Evolution* **46:** 1945-1950.

Etges WJ. 2019. Evolutionary genomics of host plant adaptation: insights from *Drosophila*. *Curr Op Ins Sci* **36:** 96-102.

693 Etges WJ, de Oliveira CC, Gragg E, Ortíz-Barrientos D, Noor MA, Ritchie MG. 2007. Genetics of
694 incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and
695 genotype x environment interactions. *Evolution* **61:** 1106-1119.

696 Etges WJ, de Oliveira CC, Ritchie MG, Noor MA. 2009. Genetics of incipient speciation in *Drosophila*
697 *mojavensis*. II. Host plants and mating status influence cuticular hydrocarbon QL expression and
698 GxE interactions. *Evolution* **63:** 1712-1730.

699 Etges WJ, de Oliveira CC, Ritchie MG, Noor MA. 2010. Genetics of incipient speciation in *Drosophila*
700 *mojavensis*. III. Life-history divergence in allopatry and reproductive isolation. *Evolution* **64:**
701 3549-3569.

702 Etges WJ, Trotter MV, de Oliveira CC, Rajpurohit S, Gibbs AG, Tuljapurkar S. 2015. Deciphering life
703 history transcriptomes in different environments. *Mol Ecol* **24:** 151-179.

704 Etges WJ, de Oliveira CC, Rajpurohit S, Gibbs AG. 2017. Effects of temperature on transcriptome and
705 cuticular hydrocarbon expression in ecologically differentiated populations of desert *Drosophila*.
706 *Ecol Evol* **7:** 619-637.

707 Fasolo AG, Krebs RA. 2004. A comparison of behavioural change in *Drosophila* during exposure to
708 thermal stress. *Biol J Linn Soc* **83:** 197-205.

709 Feder JL, Nosil P. 2009. Chromosomal inversions and species differences: when are genes affecting
710 adaptive divergence and reproductive isolation expected to reside within inversions? *Evolution*
711 **63:** 3061-3075.

712 Fellows DP, Heed WB. 1972. Factors affecting host plant selection in desert-adapted cactophilic
713 *Drosophila*. *Ecology* **53:** 850-858.

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

717 Fogleman JC, Danielson PB. 2001. Chemical interactions in the cactus-microorganism-*Drosophila* model
718 system of the Sonoran desert. *Am Zool* **41:** 877-889.

719 Fogleman JC, Starmer WT, Heed WB. 1981. Larval selectivity for yeast species by *Drosophila*
720 *mojavensis* in natural substrates. *Proc Natl Acad Sci USA* **78:** 4435-4439.

721 Fogleman JC, Starmer WT, Heed WB. 1982. Comparisons of yeast florae from natural substrates and
722 larval guts of southwestern *Drosophila*. *Oecologia* **52:** 187-191.

723 Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, Gordon D, Earl D,
724 Keane T, Eichler EE, et al. 2018. Comparative Annotation Toolkit (CAT) - simultaneous clade and
725 personal genome annotation. *Genom Res* **28:** 1029-1038.

726 Findlay GD, Yi X, MacCoss MJ, Swanson WJ. 2008. Proteomics reveals novel *Drosophila* seminal fluid
727 proteins transferred at mating. *PLoS Biol* **6:** e178.

728 Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species tree inference with BPP using genomic sequences and
729 the multispecies coalescent. *Mol Biol Evol* **35:** 2585-2593.

730 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for
731 automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117:**
732 9451-9457.

733 Gibbs AG, Matzkin LM. 2001. Evolution of water balance in the genus *Drosophila*. *J Exp Biol* **204:**
734 2331-2338.

735 Gilbert DG. 2007. DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nuc Ac Res* **35:**
736 D480-D485.

737 Goel M, Sun H, Jiao W-B, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local
738 sequence differences from whole-genome assemblies. *Genom Biol* **20:** 277.

739 Gu Z. 2014. *circlize* implements and enhances circular visualization in R. *Bioinformatics* 30: 2811-2812.

740 Guillén Y, Ruiz A. 2012. Gene alterations at Drosophila inversion breakpoints provide *prima facie*
741 evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genom*
742 **13:** 53.

743 Guillén Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, Casillas S, Ràmia M, Egea R, Negre B, et al.
744 2015. Genomics of ecological adaptation in cactophilic *Drosophila*. *Genom Biol Evol* **7:** 349-366.

745 Guillén Y, Casillas S, Ruiz A. 2019. Genome-wide patterns of sequence divergence of protein-coding
746 genes between *Drosophila buzzatii* and *D. mojavensis*. *J Hered* **110:** 92-101.

747 Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. 2008. Direct estimation
748 of the mitochondrial mutation rate in *Drosophila melanogaster*. *PLoS Biol* **6:** e204.

749 Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch
750 DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal
751 transcript alignment assemblies. *Nuc Ac Res* **31:** 5654-5666.

752 Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, Tiffin P. 2021. Genomic structural
753 variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the
754 chocolate tree. *Proc Natl Acad Sci USA* **118:** e2102914118.

755 Hardy RW, Lougheed A, Markow TA. 2011. Reproductive tract and spermatid abnormalities of hybrid

males from reciprocal crosses between *Drosophila mojavensis* and *D. arizonae*. *Fly* **5:** 76-80.

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

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

Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35:** 2907-2915.

Hoffmann AA, Sgrò CM, Weeks AR. 2004. Chromosomal inversion polymorphisms and adaptation. *Trends Ecol Evol* **19:** 482-488.

Holmes MW, Hammond TT, Wogan GOU, Walsh RE, Labarbera K, Wommack EA, Martins FM, Crawford JC, Mack KL, Bloch LM, et al. 2016. Natural history collections as windows on evolutionary processes. *Mol Ecol* **25:** 864-881.

Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB. 2021. Long reads are revolutionizing 20 years of insect sequencing. *Genom Biol Evol* 13: evab138.

Immarigeon C, Frei Y, Delbare SYN, Gligorov D, Almeida PM, Grey J, Fabbro L, Nagoshi E, Billeter J-C, Wolfner MF, et al. 2021. Identification of a micropeptide and multiple secondary cell genes that modulate *Drosophila* male reproductive success. *Proc Natl Acad Sci USA* **118:** e2001897118.

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

e66405.

Jasper WC, Linksvayer TA, Atallah J, Friedman D, Chiu JC, Johnson BR. 2015. Large-scale coding sequence change underlies the evolution of postdevelopmental novelty in honey bees. *Mol Biol Evol* **32:** 334-346.

Jaworski CC, Allan CW, Matzkin LM. 2020. Chromosome-level hybrid de novo genome assemblies as an attainable option for nonmodel insects. *Mol Ecol Res* **20:** 1277-1293.

Jiao W-B, Schneeberger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Comms* **11:** 989.

Johnson, WR. 1980. Chromosomal polymorphism in natural populations of the desert adapted species, *Drosophila mojavensis*. University of Arizona, Tucson, AZ. https://www.proquest.com/docview/303026452?pq-origsite=gscholar&fromopenview=true

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

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* mutation accumulation lines. *Genom Res* **19:** 1195-1201.

Kelleher ES, Markow TA. 2007. Reproductive tract interactions contribute to isolation in *Drosophila*. *Fly* **1:** 33-37.

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. 2020. Mate discrimination among subspecies through a

798 conserved olfactory pathway. *Sci Adv* **6:** eaba5279.

799 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping
800 with HISAT2 and HISAT-genotype. *Nat Biotech* **37:** 907-915.

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

804 Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation, and speciation. *Genetics* **173:**
805 419-434.

806 Knowles LL, Markow TA. 2001. Sexually antagonistic coevolution of a postmating-prezyotic
807 reproductive character in desert *Drosophila*. *Proc Natl Acad Sci USA* **98:** 8692-8696.

808 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate
809 long-read assembly via adaptive k-mer weighting and repeat separation. *Genom Res* **27:** 722-736.

810 Kosakovsky-Pond SL, Poon AF, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR,
811 Bouvier D, Nekrutenko A, et al. 2020. HyPhy 2.5 – a customizable platform for evolutionary
812 hypothesis testing using phylogenies. *Mol Biol Evol* **37:** 295-299.

813 Krebs RA. 1999. A comparison of Hsp70 expression and thermotolerance in adults and larvae of three
814 *Drosophila* species. *Cell Stress Chaperon* **4:** 23-249.

815 Lewontin RC. 1974. *The genetic basis of evolutionary change*. Columbia University Press, New York,
816 NY.

817 Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and
818 population genetical parameter estimation from sequencing data. *Bioinformatics* **27:** 2987-2993.

819 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094-3100.

820 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
821 *Bioinformatics* **25:** 1754-1760.

822 Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees
823 under the coalescent model. *BMC Evol Biol* **10:** 302.

824 Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, *et al.* 2020. Pan-genome of
825 wild and cultivated soybeans. *Cell* **182:** 162-176.

826 Lohse K, Clarke M, Ritchie MG, Etges WJ. 2015. Genome-wide tests for introgression between
827 cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution* **69:** 1178-1190.

828 Long E, Evans C, Chaston J, Udall JA. 2018. Genomic structural variations within five continental
829 populations of *Drosophila melanogaster*. *G3* **8:** 3247-3253.

830 Machado CA, Matzkin LM, Reed LK, Markow TA. 2007. Multilocus nuclear sequences reveal intra- and
831 interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*.
832 *Mol Ecol* **16:** 3009-3024.

833 MacLean HJ, Overgaard J, Kristensen TN, Lyster C, Hessner L, Olsvig E, Sørensen JG. 2019.
834 Temperature preference across life stages and acclimation temperatures investigated in four
835 *Drosophila* species. *J Therm Biol* **86:** 102428.

836 Markow TA. 1981. Courtship behavior and control of reproductive isolation between *Drosophila*
837 *mojavensis* and *Drosophila arizonensis*. *Evolution* **35:** 1022-1026.

838 Markow TA, Castrezana S, Pfeilier E. 2007. Flies across the water: genetic differentiation and
839 reproductive isolation in allopatric desert *Drosophila*. *Evolution* **56:** 546-552.

840 Massie KR, Markow TA. 2005. Sympatry, allopatry and sexual isolation between *Drosophila mojavensis*
841 and *D. arizonae*. *Hereditas* **142:** 51-55.

842 Mateus RP, Nazario-Yepiz NO, Ibarra-Laclette E, Ramirez Loustalot-Laclette M, Markow TA. 2019.
843 Developmental and transcriptomal responses to seasonal dietary shifts in the cactophilic
844 *Drosophila mojavensis* of North America. *J Hered* **110:** 58-67.

845 Mathers TC, Wouters RHM, Mugford ST, Swarbreck D, van Oosterhout C, Hogenhout SA. 2021.
846 Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and
847 long-term conservation of the X chromosome. *Mol Biol Evol* **38:** 856-875.

848 Matzkin LM. 2004. Population genetics and geographic variation of alcohol dehydrogenase (*Adh*)
849 paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. *Mol Biol*
850 *Evol* **21:** 276-285.

851 Matzkin LM. 2005. Activity variation in alcohol dehydrogenase paralogs is associated with adaptation to
852 cactus host use in cactophilic *Drosophila*. *Mol Ecol* **14:** 2223-2231.

853 Matzkin LM. 2008. The molecular basis of host adaptation in cactophilic *Drosophila*: molecular evolution
854 of a glutathione S-transferase gene (*GstD1*) in *Drosophila mojavensis*. *Genetics* **178:** 1073-1083.

855 Matzkin LM. 2012. Population transcriptomics of cactus host shifts in *Drosophila mojavensis*. *Mol Ecol*
856 **21:** 2428-2439.

857 Matzkin LM. 2014. Ecological genomics of host shifts in *Drosophila mojavensis*. *Adv Exp Med Biol* **781:**
858 233-247.

859 Matzkin LM, Eanes WF. 2003. Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in
860 cactophilic *Drosophila*. *Genetics* **163:** 181-194.

Matzkin LM, Markow TA. 2009. Transcriptional regulation of metabolism associated with the increased desiccation resistance of the cactophilic *Drosophila mojavensis*. *Genetics* **182:** 1279-1288.

Matzkin LM, Markow TA. 2013. Transcriptional differentiation across the four subspecies of *Drosophila mojavensis*. In Michalak P (ed.) *Speciation: natural processes, genetics, and biodiversity*. Nova Science Publishers, New York, NY. pp. 119-135.

Matzkin LM, Watts TD, Markow TA. 2007. Desiccation resistance in four *Drosophila* species: sex and population effects. *Fly* **1:** 268-273.

Matzkin LM, Watts TD, Markow TA. 2009. Evolution of stress resistance in *Drosophila*: interspecific variation in tolerance to desiccation and starvation. *Func Ecol* **23:** 521-527.

Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA. 2006. Functional genomics of cactus host shifts in *Drosophila mojavensis*. *Mol Ecol* **15:** 4635-4643.

McGirr JA, Johnson LM, Kelly W, Markow TA, Bono JM. 2017. Reproductive isolation among *Drosophila arizonae* from geographically isolated regions of North America. *Evol Biol* **44:** 82-90.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol Evol* **35:** 561-572.

Mettler LE. 1963. *Drosophila mojavensis baja*, a new form in the mulleri complex. *Dros Inf Serv* **38:** 57.

Miller GT, Starmer WT, Pitnick S. 2003. Quantitative genetic analysis of among-population variation in sperm and female sperm-storage organ length in *Drosophila*. *Genet Res* **81:** 213-220.

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly contiguous genome assemblies of 15 *Drosophila* species generated using Nanopore sequencing. *G3* **8:** 3131-3141.

881 Moreyra NN, Almeida FC, Allan C, Frankel N, Matzkin LM, Hasson E. 2022. Phylogenomics provides
882 insights into the evolution of cactophily and host plant shifts in *Drosophila*. *bioRxiv*
883 doi:10.1101/2022.04.29.490106.

884 Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP,
885 Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol* **32:** 1365-
886 1371.

887 Nazario-Yepiz NO, Ramirez Loustalot-Laclette M, Carpinteyro-Ponce J, Abreu-Goodger C, Markow TA.
888 2017. Transcriptional responses of ecologically diverse *Drosophila* species to larval diets
889 differing in relative sugar and protein ratios. *PLoS ONE* **12:** e0183007.

890 Nemeth DC, Ammagarahalli B, Layne JE, Rollmann SM. 2018. Evolution of coeloconic sensilla in the
891 peripheral olfactory system of *Drosophila mojavensis*. *J Ins Physiol* **110:** 13-22.

892 Newby BD, Etges WJ. 1998. Host preferences among populations of *Drosophila mojavensis* (Diptera:
893 Drosophilidae) that use different host cacti. *J Ins Behav* **11:** 691-712.

894 Noor MAF, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive
895 isolation of species. *Proc Natl Acad Sci USA* **98:** 12084-12088.

896 Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence
897 dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol* **29:** 3459-3473.

898 O'Donnell S, Fischer G. 2020. MUM&Co: Accurate detection of all SV types through whole genome
899 alignment. *Bioinformatics* **36:** 3242-3243.

900 Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ. 2012. Monophyly,
901 divergence times, and evolution of host plant use inferred from a revised phylogeny of the

Drosophila repleta species group. *Mol Phylogenet Evol* **64:** 533-544.

Orr HA. 2005. The genetic theory of adaptation: a brief history. *Nat Rev Genet* **6:** 119-127.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Santiago C, Lugo B, Elliott TA, Ware D, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genom Biol* **20:** 275.

Pantazidis AC, Zouros E. 1988. Location of an autosomal factor causing sterility in *Drosophila mojavensis* males carrying the *Drosophila arizonensis* Y chromosome. *Heredity* **60:** 299-304.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: algorithms for genome multiple sequence alignment. *Genom Res* **21:** 1512-1528.

Pertea M, Pertea GM, Antonescu CM, Chang TC. Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotech* **33:** 290-295.

Pitnick S, Miller GT, Schneider K, Markow TA. 2003. Ejaculate-female coevolution in *Drosophila mojavensis*. *Proc R Soc Lond B* **270:** 1507-1512.

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. *Hum Mol Genet* **27:** R234-R241.

Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, Chowdhary H, Pieniak I, Maynard LJ, Gibbins MA, et al. 2019. Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol Biol Evol* **36:** 2922-2924.

R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

922 Rajpurohit S, Oliveira CC, Etges WJ, Gibbs AG. 2013. Functional genomic and phenotypic responses to
923 desiccation in natural populations of a desert drosophilid. *Mol Ecol* **22:** 2698-2715.

924 Rampasso AS, Markow TA, Richmond MP. 2017. Genetic and phenotypic differentiation suggests
925 incipient speciation within *Drosophila arizonae* (Diptera: Drosophilidae). *Biol J Linn Soc* 122:
926 444-454.

927 Rane RV, Pearce SL, Li F, Coppin C, Schiffer M, Shirriffs J, Sgrò CM, Griffin PC, Zhang G, Lee SF, et al.
928 2019. Genomic changes associated with adaptation to arid environments in cactophilic
929 *Drosophila* species. *BMC Genom* **20:** 52.

930 Ravi Ram K, Wolfner MF. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay
931 between males and females during reproduction. *Int Comp Biol* **47:** 427-445.

932 Redmond SN, Sharma A, Sharakhov I, Tu Z, Sharakhova M, Neafsey DE. 2020. Linked-read sequencing
933 identifies abundant microinversions and introgression in the arboviral vector *Aedes aegeypti*.
934 *BMC Biol* **18:** 26.

935 Reed LK, Markow TA. 2004. Early events in speciation: polymorphism for hybrid male sterility in
936 *Drosophila*. *Proc Natl Acad Sci USA* **101:** 9009-9012.

937 Reed LK, Nyboer M, Markow TA. 2007. Evolutionary relationships of *Drosophila mojavensis* geographic
938 host races and their sister species *Drosophila arizonae*. *Mol Ecol* **16:** 1007-1022.

939 Reed LK, LaFlamme BA, Markow TA. 2008. Genetic architecture of hybrid male sterility in *Drosophila*:
940 analysis of intraspecies variation for interspecies isolation. *PLoS ONE* **3:** e3076.

941 Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
942 Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all

943 vertebrate species. *Nature* **592:** 737-746.

944 Richmond MP. 2014. The role of aedeagus size and shape in failed mating interactions among recently
945 diverged taxa in the *Drosophila mojavensis* species cluster. *BMC Evol Biol* **14:** 255.

946 Richmond MP, Johnson S, Markow TA. 2012. Evolution of reproductive morphology among recently
947 diverged taxa in the *Drosophila mojavensis* species cluster. *Ecol Evol* **2:** 397-408.

948 Riddle NC, Elgin SCR. 2018. The *Drosophila* dot chromosome: where genes flourish amidst repeats.
949 *Genetics* **210:** 757-772.

950 Rius N, Delprat A, Ruiz A. 2013. A divergent P element and its associated MITE, BuT5, generate
951 chromosomal inversions and are widespread within the *Drosophila replete* species group. *Genom*
952 *Biol Evol* **5:** 1127-1141.

953 Ross CL, Markow TA. 2006. Microsatellite variation among diverging populations of *Drosophila*
954 *mojavensis*. *J Evol Biol* **19:** 1691-1700.

955 Ruiz A, Heed WB, Wasserman M. 1990. Evolution of the *Mojavensis* cluster of cactophilic *Drosophila*
956 with descriptions of two new species. *J Hered* **81:** 30-42.

957 Russo CAM, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of drosophilid
958 species. *Mol Biol Evol* **12:** 391-404.

959 Sanchez-Flores A, Peñaloza F, Carpinteyro-Ponce J, Nazario-Yepiz N, Abreu-Goodger C, Machado CA,
960 Markow TA. 2016. Genome evolution in three species of cactophilic *Drosophila*. *G3* **6:** 3097-
961 3105.

962 Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente
963 VLS, Aguadé M, Anderson WW, et al. 2008. Polytene chromosomal maps of 11 *Drosophila*

964 species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics* **179:**
965 1601-1655.

966 Schnebel EM, Grossfield J. 1984. Mating-temperature range in *Drosophila*. *Evolution* **38:** 1296-1307.

967 Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. *Mol Ecol* **28:**
968 1537-1549.

969 Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation
970 completeness. In: Kollmar M (ed.) Gene Prediction. Methods in Molecular Biology, vol 1962.
971 Humana, New York, NY. pp. 227-245.

972 Shaible TM, Matzkin LM. 2022. Physiological and life history changes associated with seasonal
973 adaptation in the cactophilic *Drosophila mojavensis*. *Biology Open. in press*.

974 Sherman CDH, Lotterhos KE, Richardson MF, Tepolt CK, Rollins LA, Palumbi SR, Miller AD. 2016.
975 What are we missing about marine invasions? Filling in the gaps with evolutionary genomics.
976 *Mar Biol* **163:** 198.

977 Smith G, Lohse K, Etges WJ, Ritchie MG. 2012. Model-based comparisons of phylogeographic scenarios
978 resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*. *Mol Ecol* **21:** 3293-
979 3307.

980 Smith G, Fang Y, Liu X, Kenny J, Cossins AR, de Oliveira CC, Etges WJ, Ritchie MG. 2013.
981 Transcriptome-wide expression variation associated with environmental plasticity and mating
982 success in cactophilic *Drosophila mojavensis*. *Evolution* **67:** 1950-1963.

983 Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, Larracuente AM. 2020. Dynamic evolution
984 of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the *simulans*

clade. *Mol Biol Evol* **37:** 2241-2256.

Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312-1313.

Starmer WT, Heed WB, Rockwood-Sluss ES. 1977. Extension of longevity in *Drosophila mojavensis* by environmental ethanol: differences between subraces. *Proc Natl Acad Sci USA* **74:** 387-391.

van Dyke E, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The third revolution in sequencing technology. *Trends Genet* **34:** 666-681.

Vanderlinde T, Dupim EG, Nazario-Yepiz NO, Carvalho AB. 2019. An improved genome assembly for *Drosophila navojoa*, the basal species in the *mojavensis* cluster. *J Hered* **110:** 118-123.

Von Grotthuss M, Ashburner M, Ranz JM. 2010. Fragile regions and not functional constraints predominate in shaping gene organization in the genus *Drosophila*. *Genom Res* **20:** 1084-1096.

Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yap X, et al. 2018. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28:** 581-591.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9:** e112963.

Wellband K, Mérot C, Linnansaari T, Elliott JAK, Curry RA, Bernatchez L. 2019. Chromosomal fusion and life-history associated genomic variation contribute to within-river local adaptation of Atlantic salmon. *Mol Ecol* **28:** 1439-1459.

Wellenreuther M, Bernatchez L. 2018. Eco-evolutionary genomics of chromosomal inversions. *Trends*

Ecol Evol **33:** 427-440.

Whibley A, Kelley JL, Narum SR. 2021. The changing face of genome assemblies: guidance on achieving high-quality reference genomes. *Mol Ecol Res* **21:** 641-652.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.

Ye C, Ma ZS. 2016. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ*, **4:** e2016.

Ye C, Hill CM, Wu S, Ruan J, Ma ZS. 2016. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep* **6:** 31900.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf* **19:** 153.

Zhang L, Reifová R, Halenková Z, Gompert Z. 2021. How important are structural variants for speciation? *Genes* **12:** 1084.

Zouros E, Lofdahl K, Martin PA. 1988. Male hybrid sterility in *Drosophila*: interactions between autosomes and sex chromosomes in crosses of *D. mojavensis* and *D. arizonensis*. *Evolution* **42:** 1321-1331.