Population structure and introgression among recently differentiated \textit{Drosophila melanogaster} populations

J.M. Coughlan\textsuperscript{1,}*, A.J. Dagilis\textsuperscript{1*}, A. Serrato-Capuchina\textsuperscript{2}, H. Elias\textsuperscript{1}, D. Peede\textsuperscript{3,4}, K. Isbell\textsuperscript{1}, D.M. Castillo\textsuperscript{5}, B.S. Cooper\textsuperscript{6} and D.R. Matute\textsuperscript{1,}\%

*Denotes equal contribution

\textsuperscript{1} Biology Department, University of North Carolina
\textsuperscript{2} Arnold Arboretum, Harvard University
\textsuperscript{3} Department of Ecology, Evolution, and Organismal Biology, Brown University
\textsuperscript{4} Center for Computational Molecular Biology, Brown University
\textsuperscript{5} School of Biological Sciences, University of Utah
\textsuperscript{6} Division of Biological Sciences, University of Montana
\textsuperscript{7} Biology Department
\textsuperscript{8} University of North Carolina
\textsuperscript{9} 250 Bell Tower Drive, Genome Sciences Building
\textsuperscript{10} Chapel Hill, NC 27510

Email: dmatute@email.unc.edu, jcoug@email.unc.edu

Running title: \textit{Diversity and genetic structure in a model system}

Keywords: Female mate choice, Genetic differentiation, Genetic incompatibility, Hybridization, Polygenic adaptation, Reproductive isolation

Author contributions

Experimental design: JMC, AJD, DMC, and DRM designed all experiments and data analyses.

Data collection: JMC, ASC, HE, DP, KI, BSC, and DRM collected the data, including field work, mate choice experiments, and DNA extractions.
Data analyses: JMC and AJD performed all data analyses.

Manuscript writing: JMC, AJD, DMC, BSC, and DRM wrote the manuscript.
ABSTRACT

Understanding the factors that produce and maintain genetic variation is a central goal of evolutionary biology. Despite a century of genetic analysis, the evolutionary history underlying patterns of exceptional genetic and phenotypic variation in the model organism *Drosophila melanogaster* remains poorly understood. In particular, how genetic and phenotypic variation is partitioned across global *D. melanogaster* populations, and specifically in its putative ancestral range in Subtropical Africa, remains unresolved. Here, we integrate genomic and behavioral analyses to assess patterns of population genetic structure, admixture, mate preference, and genetic incompatibility throughout the range of this model organism. Our analysis includes 174 new accessions from novel and under-sampled regions within Subtropical Africa. We find that while almost all Out of Africa genomes correspond to a single genetic ancestry, different geographic regions within Africa contain multiple distinct ancestries, including the presence of substantial cryptic diversity within Subtropical Africa. We find evidence for significant admixture—and variation in admixture rates—between geographic regions within Africa, as well as between African and non-African lineages. By combining behavioral analysis with population genomics, we demonstrate that female mate choice is highly polymorphic, behavioral types are not monophyletic, and that genomic differences between behavioral types correspond to many regions across the genome. These include regions associated with neurological development, behavior, olfactory perception, and learning. Finally, we discovered that many individual pairs of putative incompatibility loci likely evolved during or after the expansion of *D. melanogaster* out of Africa. This work contributes to our understanding of the evolutionary history of a key model system, and provides insight into the distribution of reproductive barriers that are polymorphic within species.
**INTRODUCTION**

*Drosophila melanogaster* has remained one of the most powerful genetic systems to understand the molecular underpinnings of animal biology since its development in the early 20th century [1–3]. The species is distributed globally and is commonly associated with human settlements [4,5]. Recent sampling efforts have strongly suggested that *D. melanogaster* originated in the African mopane forest and initially bred on marula fruits, with a transition to human commensalism within Africa approximately 10,000-13,000 years ago [4,5]. Yet, much remains unknown about the natural history, distribution, and evolutionary history of *D. melanogaster*, particularly in its ancestral range. Given the importance of model systems, like *D. melanogaster*, to our understanding of the genetic basis of morphological [6,7], physiological [8], and behavioral [9] traits, as well as our understanding of different evolutionary processes in both natural and experimental contexts [10–12], it remains critical to understand how genetic and phenotypic variation evolved and is maintained in the ancestral range.

Although significant population genetic structure between Africa and non-African populations has long been recognized in *D. melanogaster* [13], and the existence of population genetic structure within Africa has more recently been suggested [14–16], the partition of genetic diversity within the ancestral range is largely still unresolved. Early multilocus surveys found limited to modest structure within Africa [14,17], and supported distinct West and East African clades [15]. Most recent efforts suggest modern day remnants of ancestral *D. melanogaster* lineages exist as isolated, genetically unique populations within the putative ancestral range [5,16,18]. Human-aided migration following the transition to human commensalism is thought to have
contributed to both the within Africa expansion [4,5,19], and subsequent global
expansion, the latter of which likely resulted from a single out of African event [5,20–
22], with multiple bottlenecks [23,24]. However, despite housing the vast majority of
genetic diversity, the demographic processes shaping population genetic structure in
the ancestral range remain largely unknown (though see [4,5]).

Post expansion, multiple historical events have also created opportunities for
human-mediated admixture between genetically distinct lineages of *D. melanogaster*. It
has previously been suggested that the opening of western commercial routes and the
‘Scramble for Africa’ facilitated hybridization between local African populations of *D.
melanogaster* and invading non-African *D. melanogaster* individuals, particularly in more
urban areas [25,26]. Indeed, the extent of non-African ancestry in Africa is widely
variable between populations [18], with some evidence for more pronounced signatures
of admixture in urban populations, [25,26]. Second, human migration associated with
slavery roughly 400 years ago produced a secondary contact zone between African and
non-African populations of *D. melanogaster* in the southeastern United States and the
Caribbean [27,28]. However, it is unknown whether cryptic genetic lineages within Africa
[16] have differentially contributed African ancestry outside of Africa [29]. Furthermore,
the extent to which admixture between Out of Africa and African lineages contributes to
the modern-day genetic composition of *D. melanogaster* within Africa is relatively
unresolved (but see [18,25,26]). Lastly, patterns of admixture between African
populations are almost entirely unexplored, despite their relevance for unraveling the
evolutionary history of *D. melanogaster*.

One mechanism that could contribute to differentiation among *D. melanogaster*
populations is behavioral variation between populations which results in assortative mating. Behavioral surveys of female mate choice within *D. melanogaster* from Subtropical Africa revealed a potential case of incipient speciation within *D. melanogaster* ([13,30–32]. At least two more cases of region-specific female mate preference have also been described [26,33]). Cosmopolitan (denoted ‘M’) flies are globally distributed, including within Subtropical Africa, while a second lineage is largely restricted to Zambia and Zimbabwe (denoted ‘Z’). While Z females show significant preference for Z males, M females show no preference, and males of both lineages court both types of females at similar rates [30,32,34]. Early attempts to map the genetic basis of female preference found the trait to be highly polygenic, involving loci on all major chromosome arms, and significant non-additive effects [30,31]. Despite the contributions of *D. melanogaster* to the field of speciation [35,36], it remains unknown whether the frequencies of these behavioral types covary with patterns of *D. melanogaster* population structure within Subtropical Africa. Understanding if/how behavioral isolation is associated with genetic structuring can allow us to determine if divergence in complex, quantitative traits—such as mate choice—can contribute to stable genetic divergence within a species, a hypothesis which is still widely debated [37–39].

In addition to variation in female mate choice, negative epistatic interactions between polymorphic loci, similar to hybrid incompatibilities, can cause substantial variation in fitness, including the production of low fitness individuals within a species [40,41]. While it has been hypothesized that these incompatibilities are geographically structured across the *D. melanogaster* range [40], very little is known about the
geographic origins or distributions of these alleles. In particular, it remains unknown if these alleles correspond to previously described behavioral and/or genetic lineages (i.e. [13,16,32]). Determining how these forms of reproductive isolation correspond with population structure and patterns of gene flow can contribute to our understanding of the early stages of population differentiation and speciation.

Here we address the extent of genetic differentiation and admixture in \(D.\) melanogaster, relate patterns of genetic structure to behavioral differences in the putative ancestral range, and describe the geographic distribution of putative incompatibility alleles and their contributions to admixture and population structure. This work leverages genome-wide information from 420 individuals, including 174 samples from novel and previously undersampled geographic regions as well as extensive behavioral assays of flies from Subtropical Africa. Specifically, we ask: does \(D.\) melanogaster show population genetic structure in Africa, particularly within the ancestral range? What is the extent of gene flow between unique genetic lineages? And how do two forms of polymorphic RI (mate choice and genetic incompatibilities) influence population structure in this system? Our results help to clarify the demographic history of \(D.\) melanogaster and provide some insights into the persistence of genetically unique clades within \(D.\) melanogaster.

**METHODS**

**Sampling**

We collected \(D.\) melanogaster from seven novel locations in Zambia, Namibia, and Zimbabwe using a similar approach to previously described efforts ([5]; see Table S1 for
Our approach differed from other samplings in that we used multiple potential substrates. Each trap consisted of buckets adjacent to each other and separated by about 50cms. The buckets were filled with mashed bananas (purchased locally), Marula (*Sclerocarya birrea*) or muzinzila fruits (*Berchemia discolor*). In all cases, we removed the husk of the fruit, added yeast (Red Star Active Dry Yeast - 16 oz. #201265), and allowed them to ferment for about 24 hours. These traps were put underneath trees. We collected all flies in the bucket using a sweeping net (BioQuip; Rancho Domingo, CA) after 24, 48, and 72 hours. We then netted and aspirated flies with a poooter (1135A Aspirator–BioQuip; Rancho Domingo, CA)) and anesthetized within 20 minutes of collection using FlyNap (triethylamine, Carolina Biological Supply Co.). Females and males were separated. Males and individuals from other species were placed in ethanol; *D. melanogaster* females were placed in 30mL plastic vials with cornmeal food and allowed to oviposit. Of 339 collected females, we were able to establish 244 isofemale lines (i.e., shelf-stable lines derived from a single matriarchal lineage).

**DNA extraction and sequencing**

We extracted DNA from 20 individuals from each of the 174 unique isofemale lines using a Gentra Puregene Tissue Kit (Qiagen, Valencia, CA, USA) following the recommended tissue protocol with volumes of reagents as suggested for processing 5 - 10 mg of tissue (see Table S1 for collection details). To prepare the genomic DNA libraries we used KAPA HyperPrep kits (Roche Sequencing, Pleasanton, CA) with a target fragment size of 300-500 bp at the University of North Carolina (UNC) School of Medicine’s high-throughput sequencing facility. Next, we pooled individually barcoded
libraries into groups of ~10 individual libraries and each pool was sequenced on either a single lane of an Illumina HiSeq 4000 or a Novaseq6000S4XP platform, in both cases generating paired-end 150 bp reads. This sequencing strategy yielded between 2.6-23.1 billion bp of raw sequence data for each individual (See Table S1 for coverage information). Additionally, we sequenced 32 lines that were advanced generation isofemale lines collected in Malawi, Zimbabwe and Botswana (outlined in Table S1).

Public data

We obtained whole genome sequences for an additional 247 isolines via NCBI SRA. Of these publicly available genomes, 72 were of flies sampled outside of Africa and 175 were of flies collected from within Africa, 30 of which resided in the purported ancestral range in Subtropical Africa [16,18] (see Table S1 for details).

Variant calling

We aligned 420 genomes of *D. melanogaster* to the *D. melanogaster* v6.32 reference genome [42] using *bwa mem* function [43]. We then used *Picard Tools* ([http://broadinstitute.github.io/picard/](http://broadinstitute.github.io/picard/)) to clean, sort and dedupe individual files before individually genotyping them in *GATK4* [44] with the *HaplotypeCaller* function. All samples were then jointly genotyped in *GATK4* using the *GenotypeGVCFs* function, following *GATK* best practices [44]. The resultant VCF was filtered so that indels were removed, and only biallelic sites with a minimum quality score of 30, minimum coverage of 5X, minimum genotype quality of 30, a maximum of 50% missing data were kept. We additionally removed 10 individuals with poor quality genomes (e.g. less than 5X coverage).
average coverage). For analyses requiring an outgroup (such as our phylogenetic reconstruction, outlined below), we also included twelve *D. simulans* and one *D. yakuba* genomes (see Table S1 for SRA accession numbers). These sequences were processed as above, and VCF files were merged using *bcftools merge* function [45]. We used this VCF file to perform the population genomic analyses, outlined below.

**Lineage relationships: population structure, PCA, and phylogenetic reconstruction**

To better understand the relationships among a global sampling of *D. melanogaster*, we constructed maximum likelihood (ML) phylogenies for the autosomes and *X* chromosome separately using *iqtree* version 1.6.12 [46–48]. We generated ML trees for non-overlapping regions of 100KB using the model-finder and ultra-fast bootstrap approach with 1,000 bootstraps. We then used the resulting ML trees as input for *ASTRAL v5.1.1* [49] in generating a consensus tree for the autosomes and *X* chromosome independently (Figure S2).

To characterize fine scale population genetic structure we did a *K*-means clustering analysis and PCA using *PCAngsd* [50]. *PCAngsd* uses genotype likelihoods to first perform a genome-wide PCA, then assess population structure wherein the number of ancestry types (*K*) is defined as the number of significant PCs + 1. For these analyses, we used the deduped .bam files generated above. We only included sites with <10% missing data, a minimum mapping quality of 30, and a minimum base quality of 20. This generated a dataset of 461,147 sites for 420 individuals.
Given our phylogenetic, PC, and K-means clustering analysis, we identified six geo-genetic lineages which correspond to five geographic regions (all Out of Africa lines (OOA); all East African lines excluding lines from Ethiopia (East), all West African lines, all Ethiopian lines, nine lines from Subtropical Africa which form a unique ancestry group that we refer to as Hahare-Distinct (HD), and all other Subtropical African lines (Central)). We thus calculated all pairwise measures of divergence and differentiation between all lineages identified from our phylogenetic, PC, and K-means clustering analysis (e.g. $F_{ST}$ and $D_{xy}$; defined in Figure 1) and nucleotide diversity within each lineage (e.g. $\pi$) in 1KB windows across the genome using pixy [51] with default filtering expressions (i.e. DP$\geq$10,GQ$\geq$20,RGQ$\geq$20). For this analysis we used an all-sites VCF to include invariant sites, and performed the same filtering above, but retained sites with a maximum of 2 alleles, so as to include invariant and biallelic sites. We also calculated Tajima’s D using VCFTools [52], and estimated the Site-Frequency Spectra (SFS) using SweeD [53] for each geo-genetic lineage.

**Introgression Analyses**

To estimate broad patterns of gene flow between genetic lineages of $D. melanogaster$, we first calculated Patterson’s $D$ and $f_G$ (a Patterson’s $D$ derivative which more accurately estimates the proportion of the genome experiencing introgression [54]) using Dsuite [55] with $D. simulans$ and $D. yakuba$ as outgroups for all possible trios, given the following phylogeny: (((OOA, ((East, West, Ethiopia)),HD), Central). Significance of Patterson’s $D$ was determined using a standard block jackknife
procedure [55]. In specific cases (outlined below) we also quantified differences in the extent of introgression between genetic lineages that show significant Patterson’s $D$, using $f_{dm}$ in non-overlapping 20 SNP windows for focal trios (outlined below). $f_{dm}$ is a Patterson’s $D$ derivative that is more appropriate for windowed analysis and provides a more accurate estimate of the proportion of the genome that has experienced introgression than Patterson’s $D$ [55]. Finally, we bolstered our introgression analyses by assessing heterogeneity in the genome in the relationships between potentially introgressing groups by calculating tree topology weights using twisst [56]. For each focal trio (outlined below), we calculated the topology weight at each non-overlapping 100KB window for trees among four groups: $D. simulans$ individuals as an outgroup, the two potentially introgressing groups (outlined below) and the remaining African flies as a single group (AFR). While $f_{dm}$ calculates the proportion of shared derived variants between non-sister lineages, twisst assesses the proportion of phylogenetic trees that fit particular taxonomic relationships. These analyses thus provide complimentary, but uniquely informative quantifications of introgression.

Our goals were threefold: (1) assess if female mate choice in Subtropical Africa is associated with reduced gene flow with other African lineages, (2) quantify if urbanization in Subtropical Africa is associated with increased levels of gene flow with OOA, and (3) identify the source(s) of African ancestry in the SE United States. To address each of the specific questions listed above, we used specific trios.

Gene flow between Central Africa and other African lineages: we first assessed whether Patterson’s $D$ was significantly elevated from 0 between all trios involving HD or Central Africa as P3, and Ethiopia, West, and/or East Africa as P1 and P2. For trios that had
significant Patterson’s $D$, we then calculated $f_{dm}$ in 20 SNP windows across the genome to evaluate the extent of introgression. To determine whether the extent of introgression varied between trios, we used an ANOVA with windowed $f_{dm}$ as the response variable, and P3 (HD or Central Africa), P2 (East or West), and their interaction as independent variables with a Type III Sum of Squares using the car package in R [57]. In all cases, Ethiopia was used as P1, as there were no scenarios in which Ethiopia demonstrated significant introgression with either HD or Central Africa (Table S3; Figure 3).

Urbanization and introgression with OOA: we calculated Patterson’s $D$ for all trios in which Central Africa (composed largely of rural individuals) or HD (composed largely of urban individuals) was P3, and OOA and either Ethiopia, East, or West Africa was P1 or P2. Then, to determine if either HD or Central Africa has experienced more introgression with OOA, we calculated $f_{dm}$ in 20 SNP windows across the genome for both comparisons with Ethiopia as P1, OOA as P2, and either Central Africa or HD as P3. We assessed differences in windowed $f_{dm}$ using an ANOVA with Type III SS using the car package in R [57], with $f_{dm}$ as the independent variable and P3 (HD or Central Africa) as the independent variable.

Source(s) of African ancestry in the SE United States: we performed two sets of analyses. Again, we first calculated Patterson’s $D$ with either lines from the SE United States or all other OOA lines as P3, and Ethiopia, East and/or West Africa as P1 and P2. Secondly, we calculated Patterson’s $D$ with HD or Central Africa as P3 and either lines from the SE United States or all other OOA as P2, and each of Ethiopia, East and West Africa as P1. As flies from the SE United States are hypothesized to represent a contact zone between OOA and an unidentified African lineage, we expect that flies
from the SE United States should have increased levels of introgression with the source
African lineage(s), relative to the rest of OOA flies. Therefore, we again calculated \( f_{dm} \) in
20 SNP windows across the genome for trios with significant Patterson’s \( D \), and asked
which (if any) African lineages had higher \( f_{dm} \) with the SE-United States than with OOA
using ANOVAs with Type III SS using the \textit{car} package in R [57] with windowed \( f_{dm} \) as
the response variable and each OOA lineage (SE-United States or all other OOA lines)
as the independent variable.

Last, to assess heterogeneity in introgression across the genome, we leveraged
the windowed \( f_{dm} \) analyses described above and first asked do chromosomes vary in the
extent of introgression using an ANOVA with type III SS in the \textit{car} package in R [57].
For each focal trio we performed these analyses with \( f_{dm} \) as the response variable and
chromosome arm as the independent variable. We then used pairwise \( t \)-tests to assess
significant differences between chromosome arms. To determine what genomic regions
were specifically more likely to introgress in each of our focal comparisons, we also
defined introgression outliers as windows with the top 1% of \( f_{dm} \) values.

Assessing mate preference in Subtropical African samples

We next tested if patterns of diversity, differentiation, and population structure
correspond to Z and M mating types within \textit{D. melanogaster} (as described in [30,32]).
While Z mating behavior seems largely restricted to Subtropical Africa, previous work
has identified that female mate choice is likely variable in the ancestral range of \textit{D. melanogaster} [34]—however, the frequency of Z behavior within the ancestral range, as
well as its phylogenetic distribution among Subtropical African flies, is largely unknown.

To assess the prevalence and strength of female mate choice in Subtropical Africa, we performed a series of replicated choice experiments for 47 focal isofemale lines. These lines originate from four sites in Zambia (a total of 17 lines), three sites in Zimbabwe (a total of 21 lines), one population from Malawi (two lines), one population from Botswana (one line), and one population from Namibia (six lines). Each focal isofemale line was tested in a choice experiment where 7-10 day old virgin females were presented with a standard Z and M male in a vial containing cornmeal/Karo/agar medium. We used the lines ZS2 and RAL371 as representative Z and M lines, respectively. We note that while RAL lines have mixed African-European ancestry, RAL371 is 81.2% European ancestry, slightly above the average 80.2% for the population as a whole [40]. Vials were watched continually for up to three hours, and once mating began the unmated male was removed from the vial by aspiration.

As Z and M males are morphologically indistinguishable, males were transferred to vials containing cornmeal/Karo/agar medium that had been dyed with either red or blue food safe dye approximately one hour prior to the mate choice experiment, and were easily distinguished based on abdominal coloration. Across replicates, we switched which males were fed what color to control for any bias abdominal coloration might have on female preference, and we found no effect of food color on female preference (based on an ANOVA with type III SS, with proportion of Z males chosen as the response variable, and isoline and color of Z male as fixed effects: effect of color of Z male: $F=0.088, df=1, p=0.77$; effect of isoline: $F=1.57, df=47, p=0.039$). In total, we set up an average of 53 trials (range: 12-100) per focal isofemale line, and, on average,
37 trials per inbred line (i.e. ‘isoline’; range: 8-71) resulted in a successful mating (e.g. 1,724/2,507 total trials were successful). We tested for female mate preference using a Fisher’s exact test to determine if the observed ratio of Z:M successful matings significantly deviated from the expected under random mating (e.g. 1:1) for each isofemale line (we note that the results are qualitatively the same when mate preference is assessed using a binomial test). To assess population level differences, we also performed two additional models: first, to assess if populations varied in the proportion of females with strong preference, we did a logistic regression with the response variable being whether lines had mating preference and population as the independent variable. Second, to assess if the strength of preference differed between populations, we performed a generalized linear mixed effect regression (glmer) with a poisson distribution, with the count of Z males chosen as the dependent variable, and the isolate, replicate trial date, and total matings as random effects. For both of these analyses we used the lme4 package [58] and assessed significance of fixed effects using an Type III Wald’s X² test using the Anova function in the car package in R [57].

To test for genetic differences associated with female mate preference, we first calculated the population branch excess (PBE) statistic [59] in 1KB windows across the genome with Subtropical African lines with significant female mate choice as the focal population, and Subtropical African lines that do not have strong female mate choice and RAL lines as the non-focal populations. PBE is a derivative of the population branch statistic (i.e. PBS; [60]), but quantifies branch-specific evolution for a focal population relative to two non-focal populations [59]. We then defined behavioral PBE outliers as the top 1% of PBE values (368 windows). Using this list of outliers, we did a Gene
Ontology (GO) Enrichment Analyses with a Fisher’s Exact test with an FDR threshold of 0.05 using PANTHER 16.0 [61] to assess if behavioral PBE outliers were enriched for biological GO terms relating to behavior, sensory perception, memory, learning or neurological development, each of which may have specific function in female choosiness or male attractiveness [62].

Next, we studied whether female mate choice in Subtropical Africa is associated with global patterns of differentiation or introgression in D. melanogaster, as a test of the hypothesis that female mate choice may serve as a significant reproductive barrier in nature and contribute to global patterns of population genetic structure. To do this, we calculated whether behavioral PBE outliers had elevated differentiation relative to the rest of the genome for a broader sample of lines from Subtropical Africa. We used the aforementioned estimates of $F_{ST}$ calculated in 1KB windows between all Central Africa (including all lines, regardless of phenotype) and each OOA geographic region (Asia, Europe, North Africa, North America, and Tazmania), as well as $F_{ST}$ between Central Africa and each within-Africa geo-genetic lineage (West, East, Ethiopia, South, and HD). This analysis enabled us to determine whether Central Africa broadly shows increased differentiation for regions associated with female choosiness, in line with a polygenic genetic basis of female mate preference [30,31]. We then performed an ANOVA with Type III SS to assess whether $F_{ST}$ differed between locus type (genome-wide versus behavioral PBE outlier), between comparison (each geographic region), and their interaction using the car package in R [57]. We used the emmeans package to estimate significance of specific contrasts [63]. To determine whether female mate choice was associated with differential patterns of introgression at the level of individual
loci, we also ascertained whether windowed $f_{dM}$ between Central Africa and each of
West Africa and Europe differed between behavioral PBE outlier loci and the rest of the
genome, using the $f_{dM}$ datasets described above. To do this, we performed a $t$-test with
windowed $f_{dM}$ as the dependent variable, and locus type (behavioral PBE outlier vs
genome-wide) as the independent variable.

Finally, we asked whether isolines that had significant female mate choice also
showed an overall lower history of introgression than isolines with no strong female
mate choice. To do this, we recalculated Patterson’s $D$ and $f_G$ for trios involving OOA,
West Africa, East Africa, and/or Ethiopia as P1 and P2, and isolines of each behavioral
type as P3. We restricted these analyses to include only Central Africa flies for which
we had known behavioral phenotypes, and excluded phenotyped flies from the HD
lineage, as these lines have substantial population structure from Central Africa and HD
and Central Africa are not monophyletic (Figure 1; Figure S5). We then performed a $t$-
test with $f_G$ as the dependent variable and the identity of P3 (i.e. lines that showed
behavioral preference versus those that did not) as the dependent variable.

Determining the global distribution of previously identified incompatibilities

We characterized patterns of differentiation for loci that have previously been implicated
in genetic incompatibilities [40,41]. We note that while many these loci likely represent
genetic incompatibilities in the classical sense (i.e. Dobzhansky-Muller Incompatibilities,
[64,65]), some of these loci—particularly those from natural admixture zones—may be
involved in assortative mating or extrinsic reproductive isolation [66]. For simplicity, we
refer to these loci as putatively incompatible, and we focus on loci previously described
in two studies. First, [41] used a global panel of \textit{D. melanogaster} inbred lines to create
synthetically admixed populations from a series of round-robin matings followed by
continual inbreeding. This design enabled the identification of pairs of alleles that
appear less frequently than expected under random mating and Mendelian segregation
in their final recombinant inbred line population (i.e. Genotype Ratio Distortion). Using, a
similar premise, but in a naturally admixed population, Pool [40] used patterns of linkage
disequilibrium in the SE United States to assess pairs of alleles that occur together less
frequently than expected based on their allele frequencies (i.e. Ancestry Disequilibrium).
[40] also determined that many of these loci were highly differentiated between Africa
and Europe, using populations from West Africa and France, respectively.

Elevated differentiation of these putative incompatibility alleles between West
Africa and France may stem from multiple evolutionary scenarios, and differentiating
these scenarios can help elucidate the geographic distribution and potential origins of
putative incompatibilities within \textit{D. melanogaster}. Here, we aim to differentiate two
potential scenarios: First, putative incompatibilities between Europe and Africa may
have arisen with or after the Out of Africa expansion and thus may represent more
recently derived incompatibilities (i.e. within the last 10-23kya; [5,20,22]). Under this
scenario, we predict that differentiation at putative incompatibility loci should be low
between geo-genetic lineages in Africa, but high between Europe and all African
populations. Second, it is also plausible that putative incompatibilities between Europe
and West Africa are very old and also exist between West Africa and another geo-
genetic lineage within Africa. Shared ancestry or subsequent introgression could explain
the presence of these alleles in Europe. Under this scenario, we expect that
differentiation should be high between West Africa and both Europe and other African
geo-genetic lineages, but relatively low between Europe and other African geo-genetic
lineages.

To differentiate these scenarios, we used the $F_{ST}$ windows from above for all
pairwise comparisons of Central Africa, West Africa, and Europe. We first ask if putative
incompatibility loci have elevated divergence relative to the whole genome for any
specific geographic pair using an ANOVA with Type III SS using the \texttt{car} library in R [57].
Specifically, $F_{ST}$ was the dependent variable, and locus type (e.g. genome-wide, loci
identified by [41], or loci identified by [40]), population pair (Central Africa-West Africa,
Europe-Central Africa, or West Africa-Europe), and their interaction were the
independent variables. We then used the \texttt{emmeans} package in R [63] to determine
significance for specific contrasts. Second, we aim to assess the history and distribution
of individual putative incompatibility pairs and differentiate the two evolutionary
scenarios outlined above. For these analyses, we identified putative incompatibility pairs
in which both interacting loci have the predicted patterns of differentiation. We define
highly differentiated loci as those with $F_{ST}$ values within the top 2.5% of $F_{ST}$ for that
population pair.

\section*{RESULTS}

\textit{Diversity, divergence, and evolutionary relationships among populations of \textit{D. melanogaster}}
To understand the global distribution of diversity and population structure of *D. melanogaster*, we combined whole genome resequence data from 420 lines of *D. melanogaster* from around the world, including 174 newly sequenced genomes from under- and undersampled, rural locales within the proposed ancestral range in Subtropical Africa. PCA revealed that genetic variation within *D. melanogaster* is mainly structured between flies from Subtropical Africa and Out of Africa (OOA), as reflected by PC1 (which explains 46.7% of the variation). While somewhat intermediate along PC1, Ethiopia, as well as East and West Africa, are much more distinct from OOA and Subtropical Africa along PC2 (which explains 18.6% of the total variation). We further identified eight unique genetic ancestries using K-means clustering in *PCAngsd* [50].

Four of these ancestries predominantly occur in Subtropical Africa, three largely correspond to one each of Ethiopia, West Africa, and East Africa, and a final ancestry type is most common in all OOA accessions (including *D. melanogaster* from North Africa; Figure 1).

*K*-means clustering and PC analyses largely agree with our consensus phylogeny with *D. simulans* and *D. yakuba* as outgroups. Of the eight ancestry types identified, four correspond to largely monophyletic clades; OOA, Ethiopia, West Africa, and a unique ancestry type from Subtropical Africa that comprises nine individuals predominantly from Harare, Zimbabwe (denoted by gold in Figure 1). We refer to this lineage as Harare-Distinct (HD). Of the remaining four ancestries, individuals from East Africa cluster together in a PCA, but are split across our phylogeny, with some individuals sister to West Africa, and others sister to all OOA individuals. The remaining three ancestries are largely found in individuals from Subtropical Africa, but do not
correspond to monophyletic clades or unique sampling locales. They also do not exist in pure form; the vast majority of Subtropical African individuals comprise two, and sometimes three of these ancestries (denoted as yellow, red, and orange in Figure 1). While we find no strict correspondence between ancestry type and monophyly in these Subtropical African individuals, we note that the largest and most distantly related clade of Subtropical African flies has the largest proportion of one ancestry type (indicated by orange in Figure 1), and largely contain samples from more remote sampling locales. Combining these results, we define six geo-genetic lineages that we use for subsequent analyses: OOA, Ethiopia, East Africa, West Africa, HD, and all other Subtropical African flies (which we refer to as Central Africa). In some analyses, we also compare individuals from South Africa, Northern Africa, and the SE United States, but we note that these geographic groupings are mere subsets of individuals from our Central African (for South Africa) or OOA (for North Africa and the SE United States) geo-genetic lineages.

Despite substantial population genetic structure, global $F_{ST}$ estimates were relatively low in almost every comparison, with the lowest global $F_{ST}$ found between Central and South Africa ($F_{ST}=0.022$), and the highest between HD and OOA ($F_{ST}=0.246$; Table S2). Variation in pairwise global $F_{ST}$ is likely due to differences in nucleotide diversity, as all pairwise comparisons of $D_{xy}$ were very similar (Table S2), and the populations with the highest global $F_{ST}$ were those in which both populations also showed the lowest pairwise nucleotide diversity ($\pi$; Table S2; Figure 2).
FIGURE 1. Broad sampling of *Drosophila melanogaster* from Subtropical Africa reveals substantial cryptic genetic structure. (A) Geographic sampling of 420 genomes from a global distribution of *D. melanogaster*; pie charts represent the average ancestry determined by PCAngsd from that sampling local. (B) Zoom-in panel from (A), focusing on Subtropical African accessions. (C) PCA of all genomes, colors indicate the genetic clade the samples reside in based on the phylogeny. (D) Phylogeny of all accessions with an ancestry plot denoting their average genomic composition of each sample. Labels to the right of the ancestry plot indicate how samples cluster into the 6 genetic lineages that we identify using PCA, K-means clustering, and phylogenetic reconstruction.

In addition to substantial cryptic genetic structure, the Central African lineage contains substantially higher levels of genetic diversity than any other geo-genetic
lineage (Figure 2). This diversity is likely caused by an excess of rare alleles, as the Central African lineage also shows the most negative values of Tajima’s D, and a left-skewed Site-Frequency Spectrum (SFS), in line with this lineage having a much larger effective population size and a recent history of population expansion (as suggested in [22]). In contrast, two lineages show signals of recent population contraction and lower levels of diversity: OOA (including North Africa) and HD (Figure 2). South Africa shows both higher values of Tajima’s D, and a slightly less left-skewed SFS but does not show a substantial reduction of genetic diversity. Tajima’s D and the SFS also differ between chromosomes, although the direction of these differences depended on the geo-genetic lineage (based on an ANOVA with Type III SS: chromosome arm × geo-genetic lineage effect: $F=163.12$, $df=28$, $p<0.0001$). For five lineages, the X chromosome had lower values of Tajima’s D and a slightly less left-skewed SFS (Central Africa, East Africa, West Africa, HD, and South Africa), while Ethiopia, North Africa, and OOA show the opposite pattern (Figure 2; Figure S3). Overall, these results suggest that both different lineages, and different chromosome arms, of *D. melanogaster* have experienced different demographic histories and even within Subtropical Africa, the HD lineage has a significantly different demographic history than the Central African lineage.
FIGURE 2. Genome-wide statistics for each geo-genetic lineage. (A,B) Average pairwise diversity (π) and Tajima’s D. (C) Folded site frequency spectrum for each genetic clade/geographic lineage for autosomes (lighter, outlined in grey) and the X chromosome (darker, outlined in black). Data for the autosomes is averaged across 2L, 2R, 3L, 3R, and chromosome 4.

Patterns of gene flow throughout the range of D. melanogaster

We next evaluated the extent of gene flow among distinct lineages within a global sampling of D. melanogaster. Specifically, we focus on three potential cases to better understand the sources and dynamics of gene flow across the range of D. melanogaster: (1) between Subtropical Africa and other African lineages, (2) the extent of gene flow between OOA and both HD and Central Africa, and (3) the source(s) of
African ancestry in the SE United States (as proposed by [27,28]). We describe the results for each of these cases as follows.

First, we evaluated the extent of admixture among geo-genetic lineages within Africa. We find evidence for extensive gene exchange between Central Africa and both East and West Africa, but not between Central Africa and Ethiopia, nor between Central Africa and HD (Table S3, Figure 3). Although gene flow between Central and both East and West Africa is pervasive, the magnitude of gene flow is likely quite low, as only 1-3% of the genome is inferred to be admixed between Central Africa and each of East and West Africa, respectively (depending on whether \( f_G \) or \( f_{dM} \) is used; Table S3, Figure 3). Similarly, we find significant gene flow between HD and both East and West Africa (Table S3, Figure 3). Despite closer geographic proximity between most East African samples and Central African samples, introgression between West Africa and both Central Africa and HD is significantly elevated compared to introgression between East Africa and either Central Africa or HD (Figure 3; \( f_G \) between West Africa and either HD or Central Africa ranged from 0.02-0.033, while \( f_G \) between East Africa and either HD or Central Africa ranged from 0.017-0.022; we find no significant difference between Central Africa and HD in the extent of gene flow with either West or East Africa).

Overall, this suggests that the magnitude of admixture is not equivalent between Subtropical African and other African \( D. melanogaster \) lineages and may not simply correspond to geographic distance.

Second, we evaluated the extent of introgression between OOA and both Central Africa and HD. Specifically, we aimed to test whether back migration from OOA may be responsible for the genetically unique HD lineage (denoted in gold in Figures 1 and 2).
as a test of whether urbanization has facilitated introgression in Africa. Other genomic evidence suggests that this may be the case—the HD lineage shows substantially reduced diversity and shows genomic patterns of a bottleneck relative to Central Africa, despite no geographic separation (Figure 1, Figure 2), and mimics patterns of diversity and Tajima’s D seen in OOA populations (Figure 2). In line with this hypothesis, we find evidence of admixture between HD and OOA; \( f_{dm} \) is elevated between OOA and HD when compared to OOA and Central Africa (Figure 3), wherein 18-20% of OOA and HD genomes have introgressed, while only 15% of genomes between OOA and Central Africa have introgressed (depending on whether \( f_G \) or \( f_{dm} \) is used; Table S3, Figure 3).

However, tree topology weights generated by \textit{twisst} further reveal that the unique HD lineage does not appear to be simply a hybrid of flies from Central Africa and OOA (Figure 3C). While large proportions of the genome show near complete support for topologies in which HD is sister to OOA (purple, in Figure 3C), the second most common topology does not place HD sister to samples from Africa, but rather places all other African samples sister to OOA (teal, Figure 3C). This suggests that rather than being a patchwork of African and OOA ancestry (in which we expect yellow and purple topologies to be common in Figure 3C), HD carries distinct genetic variation, at least in genomic regions where the topology supports Africa sister to OOA (such as the center of 2L). Extensive introgression between other African lineages and OOA may also contribute to these patterns. In total, this suggests that while HD has likely experienced substantial introgression with OOA, it has also likely experienced a unique evolutionary history from either OOA or the rest of Central Africa.

Last, we studied the origin of the African alleles harbored in the SE United
States. We find significantly elevated Patterson’s $D$ between SE United States and each of Central Africa, West Africa, and HD which suggests these populations have contributed to $D. melanogaster$ from the SE United States. We do not observe a similar pattern for Ethiopia or East Africa (Table S3, Figure 3). Patterns of introgression are not unique to the SE United States, as we find elevated Patterson’s $D$ between all other OOA lines and each of Central Africa, West Africa, and HD (Table S3, Figure 3), suggesting that the contribution of these African populations might precede the split of different OOA populations. However, if $D. melanogaster$ from the SE United States experienced a second pulse of introgression unique from other OOA populations, then we predict that SE United States flies should show elevated signals of introgression with at least one African source relative to other OOA flies. In line with this prediction, both $f_G$ and $f_{dM}$ between West Africa and the SE United States are elevated relative to either $f_G$ and $f_{dM}$ between West Africa and other OOA populations (Table S3, Figure 3), while Central Africa and HD show no difference in either $f_G$ and $f_{dM}$ with the SE United States or other other OOA lines (Table S3, Figure 3). However, it is challenging to detect regions of the genome that show unique signals of introgression between West Africa and the SE United States relative to West Africa and OOA, as the landscape of $f_{dM}$ across the genome is highly correlated between these two comparisons ($r^2=0.59$, $p<0.001$; as it is between Central Africa and OOA and Central Africa and SE United States: $r^2=0.509$, $p<0.001$; and between HD and OOA and HD and SE United States: $r^2=0.607$, $p<0.0001$). Moreover, the weighted topologies with West Africa as sister to all OOA versus only accessions from the SE United States are highly similar (Figure S7).

In total, this suggests that all OOA populations have experienced some level of gene
flow with Central Africa, West Africa, and HD, but flies from the SE United States may have experienced a second, and independent pulse of West African ancestry (as has been hypothesized by [27,28]). The lack of distinction may also be caused by rapid purging of introgressed alleles, reducing any signals of introgression in a short time-span [67].

We next examined how patterns of introgression vary across the genome within Africa and between African and non-African lineages. For these analyses, we focus on the three scenarios outlined above: (1) Introgression within Africa (as measured between West Africa and Central Africa), (2) introgression between Out of Africa and the unique HD lineage, and (3) introgression between West Africa and the SE United States. In all instances, Ethiopia was used as P1 as we find little to no evidence of admixture between Ethiopia and any P3 used herein (Table S3). We find that chromosomes significantly differ in the extent of introgression for all comparisons (as measured by $f_{dm}$; West-SE-US: $F=47.11$, $df=5$, $p<0.001$; West-Cen: $F=18.36$, $df=5$, $p<0.001$; OOA-HD: $F=16.01$, $df=5$, $p<0.001$). For all three comparisons, the X chromosome showed significantly higher $f_{dm}$ values, than almost every autosome (Table S4). However, definitive evidence of increased introgression on the X-chromosome is much less apparent when using weighted topologies (Figure 3, Figure S7).

Many genes which may be involved in the transition to human commensalism fall within the top 1% of $f_{dm}$ windows (see Table S5 for full list). For example, these windows include several genes involved in insecticide resistance, (including Cyp6a18 (between OOA-HD), Cyp313a and ACE (West-SE-US), and Cyp12a4 and LRR (West-Central)), several genes related to metabolism, feeding behavior, and perception of and response
to food sources (including happyhour, pbx, for, and Gfat1 (West-Cen), NPFR, lovit, and Ald1 (West-SE-US), and Gr5a (OOA-HD)), as well as genes involved in immune function (Dcr-2, Rab4, DptA, DptB, and Tl (OOA-HD), Npc2h, Npc2g, and ben (West-Cen), and Charon and Ance-2 (West-Se-US)). Although causative connections are still needed, it is perhaps unsurprising that alleles putatively related to human commensalism may be overrepresented in three examples of potentially long-range, and potentially human-mediated admixture.

Lastly, we find several genes associated with mating behavior fall within the top 1% of introgressed regions including Desat2 (West-SENA; although we note that this gene may also be involved in desiccation resistance, as is the case with other cuticular hydrocarbons [68,69]), fru (West-CEN), and 5-HT1A, clt, and Oamb (OOA-HD). While these regions may not be involved in human commensalism per se, introgression at these loci may have implications for the distribution of female mate choice, and the distribution of female mate choice may be influenced by human commensalism; a possibility that we explore below.
FIGURE 3: Patterns of gene flow between different African and non-African lineages of D. melanogaster. (A) Boxplots of Patterson’s D for all trios given the phylogeny (((OOA, ((East, West),Ethiopia),HD), Central Africa). P3 is given on the X-axis, with the identity of the P2 population denoted by the color of the boxplot. Values of D represent the range of values over multiple P1 populations. Comparisons that yielded a significant Patterson’s D based on a standard block jackknife procedure [55] are denoted with an asterisk, otherwise they were deemed Non-Significant (NS). (B) Mean and standard errors of 20-SNP windows of $f_{st}$ from across the genome for each African geo-genetic lineage showing significant introgression with either other African or non-Africa lineages. The OOA lineage is split into those that heil from the SE United States (SE-US) and all others (OOA). Significance was determined by ANOVAs with Type III SS: NS = Not Significant, * = 0.01<p<0.05. (C) Weighted topologies for three configurations of OOA, HD, and Africa (denoted by colored phylogenies to the left).
To understand if and how strong female mate preference originally described in Subtropical Africa corresponds to any of the Subtropical African ancestries that we identify herein, we surveyed a subset of lines from Subtropical Africa for the existence of female mate choice using replicated choice experiments. Briefly, 47 focal isofemales from ten sampling locales in Subtropical Africa were presented with both a Subtropical Africa Z (ZS2) or OOA M (RAL371) isoline male. We quantified female mate preference as a significant deviation from random choice, wherein ZS2 males were more likely to copulate than RAL371 males. We find that female mate preference was common in our experiment, but far from fixed across Subtropical Africa (e.g. 19/47 lines had a preference for ZS2 males; Table S6, Figure S4). The frequency of lines showing female mate preference varied among sampling locales, with the average proportion of females exhibiting preference per population ranging from 0-100% (mean across all populations= 44%; \(\chi^2=19.04, df=9, p=0.025\); Figure 4; Table S6). Populations also varied in the strength of preference. Some populations had nearly complete preference for Z males, and others showed a mild aversion to Z males (\(\chi^2=18.63, df=9, p=0.029\); Table S6; Figure S4; average proportion Z males chosen per population ranged from 0.398-0.969). Lastly, we find that flies that with female mate choice are not monophyletic, nor do any major PCs ascribing genetic variation among phenotyped flies separate isolines with female preference from those that do not (Figure S5). Our results suggest that the dynamics of mate preference are complex with Subtropical Africa; both the presence and strength of female mate choice vary within and among populations, and this trait is not monophyletically distributed.

We leveraged the natural diversity in female mate choice among these 47 lines
from Subtropical Africa to determine what regions of the genome are associated with female mate choice. Given that isolines with strong preference do not show significant population structure relative to isolines with no strong preference (Figure S5), we instead quantified branch-specific evolution for lines with significant female mate preference. We calculated the population branch excess ($PBE$) statistic [59], with Subtropical African lines with strong female mate preference as the focal population and Subtropical African lines with no strong female mate choice and RAL as the non-focal populations. We defined the top 1% (368 windows) as outliers. These regions represent all 5 chromosomal arms, although there is a significant excess of outliers on 3L and on the X chromosome ($X^2=183.01$, df=4, $p<0.0001$). Of these 368 windows, 283 contain at least one gene, and 84 of the remaining 85 windows are proximal to at least one gene (of the windows in which genes are proximal, genes occur within 16.2KB on average, range: 2.5-100KB). A total of 407 unique genes are included either within these windows, or in close proximity (i.e. within 16.2KB on average). Gene Ontology overrepresentation analyses indicate that of these 407 genes, there is an excess of genes associated with learning, memory, cognition, chemotaxis, sensory perception, detection of stimulus, behavior, and several ontology categories associated with neurological development (as well as other biological functions; see Table S8 for details).

Furthermore, several individual genes are of potential interest for future study (Table S7). We find thirty genes that are known to influence behavior in the top 1% of $PBE$ outliers, including five genes specifically involved in courtship behavior ($Shep$, $Rdl$, $Tbh$, $lig$, and $eag$). Thirty-eight genes that are involved in sensory perception and
sensory learning are also within the top 1% of PBE values, particularly those involved in sensory perception of smell or olfactory learning, including three odorant receptor genes (Or22b, Or67c, and Or85e), four defective proboscis extension response genes (dpr8, dpr13, dpr14, dpr18), seven ionotropic receptor genes (Ir11a, Ir67b, Ir67c, Ir75a, Ir75b, Ir94f, Ir94h), Dop2R and vn (both involved in olfactory learning), as well as Ank2 and sif, which are involved in sensory perception of sound and visual perception, respectively.

Five genes which are known to influence male aggression fall in the top 1% of behavioral PBE values (Dop2R, Rpb6, CkIIalpha, Tbh, and Rdl). Z males tend to be substantially more aggressive during courtship than their M male counterparts, particularly to Z females (Figure S8). Lastly, ten genes that are involved in mushroom body development are also included in the top 1% of PBE values, including Frl, DAAM, PsGEF, rg, and rad. Mushroom bodies play a central role in learning and memory, particularly of olfactory perception, and therefore may be important in differentiating male mates. Furthermore, mutants with aberrant mushroom bodies and inhibition of mushroom bodies can cause virgin females to reject matings [70,71]. In total, we find that D. melanogaster from Subtropical Africa with significant mate preference differ in many regions across the genome, indicative of a polygenic basis of female mate choice.

We next sought to assess whether loci putatively involved in female mate choice were also more highly differentiated among a broader geographic sampling of D. melanogaster as a test of polygenic adaptation of this trait. Specifically, we asked whether behavioral PBE outliers were among the most highly differentiated loci between all Central Africa lines (regardless of phenotype) and other geographic regions. We find that $F_{ST}$ among behavior outliers is elevated between Central Africa and all other
genetic lineages relative to genome-wide $F_{ST}$ (locus type (e.g. genome-wide versus $PBE$ outlier): $F=7.57$, df=1, $p=0.0059$; geo-genetic lineage effect: $F=29085$, df=6, $p<0.0001$, locus type × lineage interaction: $F=1.29$, df=6, $p=0.26$, Figure 4, Figure S6).

Further analysis revealed that differences in $F_{ST}$ between behavioral $PBE$ outlier loci and genome-wide estimates are greater between Central Africa and OOA and between Central Africa and HD than between Central Africa and any other African lineage (Figure S6). However, despite this modest evidence for selection, we do not find that $PBE$ outlier loci are less likely to introgress than the rest of the genome, as would be expected if female mate choice served as a significant barrier to introgression ($f_{dM}$ for behavioral $PBE$ outliers is not significantly reduced relative to the rest of the genome between Central Africa and either West Africa: $t=-0.064$, $df=406.27$, $p=0.95$ or Central Africa and Europe: $t=-1.52$, $df=428.69$, $p=0.13$). Furthermore, individual isolines that differ in the strength of female mate preference show no difference in the history of introgression experienced (mean $f_G$ for lines with female mate choice = 0.097, mean $f_G$ for lines with no strong female mate choice= 0.112; $t=-0.27$, $df=9.7$, $p=0.79$). In total, this work suggests that while female mate choice may show modest signals of polygenic adaptation, it is likely not a very effective barrier to gene flow in natural $D. melanogaster$ populations.
FIGURE 4: Geographic and genetic dissection of female preference in Subtropical Africa. (A) Frequency of isofemale lines that show significant female mate choice in two-male choice experiments per sampling locale. (B) Average $F_{ST}$ is elevated in the top 1% PBE regions between the Central Africa lineage (regardless of mate preference phenotype) and all other geographic regions. (C) Cartoon illustrating that the PBE statistic can infer branch-specific evolution of the focal population (in this case, flies with significant female mate preference (red, “pref”), relative to flies with no preference (blue, “no pref”) and a third population of RAL)). (D) PBE of D. melanogaster lines with strong mate preference, measured in 1KB windows across the five major chromosome arms of D. melanogaster. Grey horizontal line indicates the top 1% threshold. Colored dots indicate a sample of genes with functions relevant to female mate preference.
Distribution of putative incompatibilities throughout *D. melanogaster*

We last assessed the distribution of alleles potentially involved in negative epistatic fitness effects in admixed populations (herein ‘incompatibility loci’) throughout the range of *D. melanogaster*. Although we refer to these loci as incompatibilities, they include any loci that are found in repulsion of one another, and thus may include traditional incompatibilities (i.e., involved in intrinsic postzygotic isolation), alleles involved in ecological hybrid breakdown, as well as loci involved in assortative mating [66]. We find that loci identified by [40], show slightly increased overall levels of differentiation than the genome-wide average between Central and West Africa, but significantly lower differentiation than the genome-wide average between Europe and each of Central and West Africa (Figure 5). In contrast, we find no significant difference in average $F_{ST}$ between loci identified by [41] and the genome-wide average for any population pair. Thus, while on average, putatively incompatible loci are within the most differentiated loci between West and Central Africa, they are not broadly differentiated on a global scale.

We next assessed whether specific pairs of incompatibility loci showed elevated differentiation, and used patterns of differentiation to characterize their potential geographic origins. Out of 445 putative incompatibility loci identified by [40] and 45 putative incompatibility loci identified by [41], we identified only eight pairs of interacting incompatibility alleles with high differentiation within Africa, and low differentiation between Central Africa and Europe; indicative of potentially older incompatibilities that divide genetic lineages within Africa. Within these eight loci, eight unique genes are included in the $F_{ST}$ outlier windows, two of which were also found in our *PBE* behavioral
analysis (Abl, and eag; Table S9). In contrast, we identified 71 pairs of loci with a signature of more recent origin (i.e. low differentiation in Africa, high differentiation between Europe and both West and Central Africa). We find that the $F_{ST}$ peaks within these 71 pairs of loci contain 133 unique genes (Table S9), 13 of which are also behavioral PBE outliers. These include DAAM, Rbp6, beat-IIIC, Doa, Lasp, luna, Eip75, Mp, olf413, Pex7, pk, Svil, and eag (we note this is a separate 1KB window than the window which shows a signal of an putatively older incompatibility, described immediately above). Many of these genes are involved in neurological development and/or behavior (DAAM, Rbp6, eag, beat-IIIC, Eip75, pk, and Mb), and reproduction (Doa, Lasp). Thus, many putative incompatibilities show a signature of more recent evolution, and a significant proportion of these loci may also be associated with behavioral differences within Subtropical Africa.

**FIGURE 5: Global differentiation of putative incompatibility loci.** (A) Distribution of $F_{ST}$ for incompatibility loci from CD [41], Pool [40] and genome-wide for three comparisons: Central Africa vs Europe, West Africa vs Central Africa, and West Africa vs Europe. (B,C) Two zoomed in windows representing one incompatibility pair, as identified by Pool [40]. In panel (B) divergence is elevated between Europe-Central Africa and West-Europe (but not West-Central Africa) for the length of the gene Piezo, indicating this allele may be more recently derived in OOA populations. In (C) we show the corresponding locus: again, divergence is elevated between Europe-Central Africa and West-Europe (but not West-Central Africa) in a window containing no genes, but slight up and down stream of comm and comm2, respectively. We also highlight the
approximate position of AGO2, the hypothesized interacting partner (based on [40]), and show there is no increased divergence in this region. $F_{ST}$ is plotted as a chromosome-level Z score, so as to more easily compare the three measures of divergence.

**DISCUSSION**

The evolutionary history of genetic model systems has been the target of extensive research, including *D. melanogaster* [4,5,16,18]. Nonetheless, gaps in the sampling of *D. melanogaster* across regions of its range have left crucial aspects of its history unexplored. We collect samples and explore the partitioning of genetic diversity and patterns of gene flow across the global range of *D. melanogaster*, and evaluate the distribution of two sources of polymorphic reproductive isolation that have been previously described (mate choice and putative genetic incompatibilities). We find that flies from Subtropical Africa harbor previously unknown genetic diversity and population genetic structure. We also find that *D. melanogaster* has experienced a complex history of gene flow, particularly within Africa. Finally, we find that behavioral isolation and hybrid incompatibilities segregating within the species have a multilayered history, but overall, do not explain patterns of population genetic structure in Subtropical Africa. These results contribute not only to our understanding of the natural and evolutionary history of a powerful model system, but also in our understanding of how reproductive isolation influences global patterns of population genetic structure. We discuss each of these implications in the following paragraphs.
Population genetic structure in the ancestral range of Drosophila melanogaster

*Drosophila melanogaster* individuals from its purported ancestral range of Subtropical Africa show the highest genetic diversity and the lowest values of Tajima’s D, in line with a demographic history of population expansion and/or strong purifying selection (in agreement with [13,15,16,20,22]). While these flies, originating largely from Zimbabwe, Zambia, Malawi and Namibia cluster together in a genome-wide PCA, our phylogenetic analyses indicate that Subtropical African individuals comprise a non-monophyletic clade that is most distantly related to all other *D. melanogaster*, as would be expected if the mopane forest in countries like Zambia and Zimbabwe are the likely origin of *D. melanogaster* variation [4,5]. However, we also describe a strong signal of structure among individuals from the ancestral range. We identify at least four unique ancestries using K-means clustering within Subtropical Africa, only one of which is monophyletic in our phylogeny. This monophyletic ancestry group—which we refer to as Harare Distinct (HD)—comprise a distinct cluster in a whole-genome PCA (Figure 1). HD has substantially reduced nucleotide diversity and elevated Tajima’s D, indicative of a population bottleneck or recent introgression (or less likely, balancing selection; [72]). These individuals also have evidence of elevated introgression with OOA relative to Central Africa. Given that seven of nine individuals in this lineage are derived from the urban center of Harare, HD may be a product of human-assisted migration of OOA individuals into urban centers in Subtropical Africa, in line with earlier work which used microsatellites [25], and similar to patterns shown in West Africa [26]. While our results are partially in line with this hypothesis, we also show that HD is likely not simply an admixed population between OOA and the Central African lineage. Although much
remains unknown about the HD clade, the presence of previously undescribed diversity within the ancestral range of arguably one of the most well studied model organisms highlights the need for further sampling and genetic interrogating of flies from Subtropical Africa.

*Drosophila melanogaster* from other regions within Africa also show strong population genetic structure, with West Africa, East Africa, and Ethiopia largely being identified as three distinct ancestry types, each clustering uniquely in a PCA, and West Africa and Ethiopia being largely monophyletic in our phylogenetic analyses. Individuals from these areas also show intermediate levels of diversity and Tajima’s D, in line with the hypothesis that, while these flies likely do not represent ancestral populations, they were also likely some of the first steps of range expansion from Subtropical Africa (as suggested by [5]). We find evidence of gene flow between both East and West Africa and Subtropical Africa. In contrast to East and West Africa, Ethiopia does not show any introgression with Subtropical Africa and represents a genetically unique lineage from other East Africa flies (in agreement with [16,18]). Many of the Ethiopian samples herein originate from the high elevation populations from the Ethiopian Highlands; an environment that might not be permissive to migration from the lowlands [73–75]. Substantial geographic and/or ecological barriers to migration may thus isolate individuals from the Ethiopian Highlands from the rest of Africa.

In contrast, individuals from outside of Africa show substantially reduced complexity; all OOA flies cluster together in a genome-wide PCA, and largely ascribe to a single ancestry. As a whole, these samples have reduced diversity and elevated Tajima’s D, which is suggestive of a bottleneck, likely accompanying an out of Africa
expansion (as has been found previously [13,20,22]). We find that all OOA individuals are sister to a subset of East African individuals, specifically those from Kenya, but as a whole, OOA lines do not show evidence of introgression with East Africa. It is therefore possible that OOA lineages are derived from East Africa, and specifically from the ancestors of modern Kenyan flies. Additionally, we find that African lineages have differentially contributed to genetic variation within OOA via post-expansion introgression. We find that all OOA flies have similar levels of introgression with the Central Africa and HD lineages, while individuals from the SE United States show slightly elevated levels of gene flow with flies from West Africa, relative to other OOA individuals. It has previously been suggested that flies from the SE United States represent a secondary contact zone between individuals of European and African ancestry [16,18,27,40], here we show that different African lineages have contributed differentially to *D. melanogaster* populations outside of Africa.

Taken together, these broad-scale analyses suggest that *D. melanogaster* has substantial population genetic structure within the ancestral range in Subtropical Africa, including the presence of a unique, and potential admixed lineage found in Harare, Zimbabwe. Patterns of gene flow within Africa are complex, as there is at least some degree of gene flow among most geo-genetic lineages; highland Ethiopian lines remaining relatively distinct and unconnected. Gene flow between African and non-African lineages is also variable, depending on the African and non-African population in question, with relatively even patterns of introgression between Subtropical Africa and OOA populations, but an increase in West African ancestry in SE United States populations, relative to other OOA populations.
Gene flow is elevated in genes related to human-commensalism

In addition to variation in broad-scale patterns of gene flow between distinct geo-genetic lineages of D. melanogaster, we also find a dynamic landscape of introgression across the genome. First, using \( f_{dM} \), we find that for all comparisons introgression is higher on the \( X \) chromosome than autosomes (Table S4). This is unusual as the sex chromosomes are often depleted for introgression [76–81], and indeed even within D. melanogaster, OOA populations often lack African ancestry on the \( X \) (for example see [40]), although several examples of systems with high levels of gene flow on the \( X \) also exist [82–85]. As the \( X \) chromosome also contains an excess of alleles associated with \( Z/M \) female mate choice, the finding that the \( X \) chromosome exhibits the highest levels of introgression has implications for the efficacy of female mate choice as a barrier to introgression, a possibility that we explore below.

Genomic windows with the highest levels of introgression contain an excess of genes associated with human commensalism, specifically genes involved in insecticide resistance, metabolism and immune system function. Humans present organisms with novel, often adverse conditions. Human habitat change is a primary driver of hybridization in many systems [86]. While the loss of local adaptation due to introgression of domesticated traits is often considered to be one of the threats of hybridization [87], wild populations without certain domesticated traits may simply be unable to survive in human proximity. The genes we identify as outliers in introgression analyses are often associated with the unique pressures of human commensalism (see [88] for an overview)— insecticide resistance [5,89,90] and diet [91,92] are the most salient cases. Increased population densities associated with human habitation may
also favor changes to the immune system [93], and genes involved with metabolism have been identified as potentially important to domestication in several cases [94–96]. Moreover, previous work has shown that *D. melanogaster* from non-human associated collection sites do show elevated, branch specific evolution for many immunity related genes [5], suggesting that the transition to human commensalism in *D. melanogaster* specifically may involve allelic changes in immunity.

We also find that several genes involved in mate choice and courtship behavior fall within the top 1% of introgressed regions. One potential explanation is that strong female mate preference comes at a fitness cost, potentially by limiting mating potential or delaying reproduction. Thus, it is possible that the cost:benefit ratio of female choice may differ under scenarios of human commensalism, ultimately favoring the loss of female choice in these environments. Under this hypothesis, subsequent introgression of less-choosy alleles back into Subtropical Africa may be favored. This may be particularly true in more urban areas, which may suffer from both continual introgression from non-choosy donor populations, and potentially a higher cost of female choosiness. Cost in urban areas may be mediated by selection for faster reproduction via increased resource competition in regions with higher *D. melanogaster* density. Although much more work is needed to further test this hypothesis, previous work has shown the loss of female preference in more urban populations of *D. melanogaster* from Brazzaville, Congo [26]. Similarly, we find that the urban HD clade (which is potentially admixed) encompasses fewer lines with strong female mate choice than the rest of our sample from Subtropical Africa, despite close geographic proximity (Figure 4, Figure S5).
Female mate choice is polymorphic and likely highly polygenic

Female mate preference in Subtropical African *D. melanogaster* is a classic example of incipient speciation since its original discovery in the mid 1990s [30–32,34,97–99]. Here we revisit this classic system to survey flies from Subtropical Africa for the presence and strength of female mate choice, particularly in previously unsampled, remote locales in Zambia, Zimbabwe, and Namibia. We find that female mate choice is relatively common, but highly polymorphic both within and among sampling locales, and not monophyletically distributed on our phylogenetic tree. These results are in line with previous work which has shown substantial variation in a female mate preference for a smaller number of isolines [34], as well as recent work highlighting the variation in male courtship behavior within Subtropical Africa [30–32,34,97–99]. Three evolutionary scenarios may explain this polymorphism: (1) female mate choice is an ancestral polymorphism which is maintained in Subtropical Africa, (2) strong female mate choice is the ancestral condition, and introgression from outside of this region has eroded its prevalence within Subtropical Africa, or (3) female mate choice is a derived character within Subtropical Africa and has either been eroded by introgression or never fully fixed throughout this region. While we cannot fully disentangle these scenarios with our current dataset, which of these scenarios has occurred in this system has implications for the stability of reproductive isolation in natural populations, as well as the potential evolutionary costs and benefits of female choosiness.

One potential caveat to the above female mate choice work is that we used a standard Z and M line for all assays as a way to standardize the female preference tests. However, one possibility is that female mate choice is more common than we
estimate, but preference is specific to local males (e.g. not all Z males are equivalent; [97–99]). Several lines of evidence suggest that this may be true for some populations; for example, the population with one of the highest levels of female preference was the population in which the focal male was derived from (e.g. ZS). Additionally, work that has used a standard M line and a male of the same isoline as the focal female generally showed greater values of female mate choice [34]. Lastly, African males vary in both cuticular hydrocarbon profiles and mating displays [99,100]. Therefore, while we present the most widespread survey of female mate choice from Subtropical Africa to date, more research will be needed to more precisely characterize the dynamics of female choice in this system.

We leveraged natural variation in the presence of female mate choice within Subtropical Africa to assess genomic differences associated with female pickiness and male attractiveness. Although genomic outliers exist on all major chromosome arms, chromosomes 3L and the X are highly enriched for outliers. These results are largely congruent with previous analyses [30,31], which found that the third chromosome had the largest effect on both female preference and male attractiveness, including a large contribution from 3L. In contrast, [30], found a much smaller contribution of the X chromosome in single chromosome replacement experiments, but substantial epistasis between the X chromosome and both major autosomes.

Behavioral PBE outlier regions are significantly enriched for genes that are involved in male courtship behavior, olfactory perception, learning, memory, and neurological development. Several individual genes present interesting candidates for future work, including genes involved in mushroom body development and sensory
perception. Mushroom body ablation in virgin females results in higher rejection rates [70,71], and plays a role in male memory and courtship effort [101]. Additionally, while the precise male traits controlling attractiveness are unknown, cuticular hydrocarbons likely play an important role [97,99,102], and therefore, evolution of sensory perception-particularly sensory perception of smell and taste may also help regulate female mate preference (see [103] for an example of sensory perception of taste in male mate choice). Lastly, we find that two genes—shep and eag—which have been shown directly to affect mating behavior are highly differentiated in Subtropical African lines with significant female preference. Shep is responsible for neuron remodeling during metamorphosis, and whose loss-of-function results in increased rejection of male mates in virgin females [104]. Eag regulates potassium permeability, and mutants of this gene have been shown to influence the amount of time males spend courting [62,105]. Thus, while we present some intriguing candidate genes for future functional analyses, overall our work suggests that the genetic basis of mate preference is likely highly polygenic.

Lastly, while we find that PBE behavioral outliers show elevated differentiated between a broader sampling of the Central African lineage and non-Central African populations (particularly those from OOA), we also find that these behavioral PBE outliers do not show decreased levels of introgression relative to the rest of the genome, nor do lines with strong female preference show a reduced history of introgression. Taken together, these results could arise from a scenario in which female mate preference has experienced polygenic local adaptation, but this asymmetric mate preference is insufficient to noticeably dampen gene flow between mating types in Subtropical Africa. In particular, female mate choice may still be maintained if the
genetic basis of mate choice involves genetic redundancy (e.g. multiple alleles are sufficient to cause female mate choice, and therefore the loss of any given allele by introgression does not dampen female mate choice). Earlier mapping efforts [30,31] suggest the existence of genetic redundancy in female mate preference and male attractiveness, though more detailed genetic analyses are needed.

Putative incompatibilities are highly differentiated, and many are likely recently derived

We lastly studied the distribution of previously identified putative incompatibility loci across our global sampling of *D. melanogaster*. We find that while, on average, these loci show slightly elevated differentiation between Central and West Africa. However, while some pairs of loci show a signature of differentiation within Africa, there are substantially more pairs of loci that are highly differentiated between Europe and both West and Central Africa. Moreover, we find that of the putative incompatibility loci that are highly differentiated (either between Europe and both African lineages or within Africa), 10-25% of genes within these windows overlap with our behavioral PBE outliers. There are two explanations for the high overlap between these two datasets. First, some of the loci identified by [40] may represent loci involved in assortative mating, as many of these putative incompatibilities were ascertained by detecting pairs of alleles with ancestry disequilibrium in naturally admixed populations [66]. Genes identified by these two approaches may constitute alleles involved in female mate choice, including *eag*, which has been shown to directly influence courtship behavior, *DAAM*, which is involved in mushroom body development, and *rbp6* which influences male aggressive
behavior. Second, while PBE analyses are an effective approach to detect branch
specific evolution for a focal population [5,59], by definition, these statistics cannot
differentiate branch specific evolution which is specific to a trait of interest versus subtle
population structure and/or branch specific evolution of non-focal traits. Therefore, if
some incompatibility alleles exist at higher frequencies in isolines with strong female
choice relative to isolines with no strong female mate choice, they too may be included
in our PBE outlier analysis (even though these alleles may not directly influence female
mate choice). Regardless, our results suggest that putative incompatibility alleles do
show some genetic structure across the global distribution of D. melanogaster, and
better understanding the function of these alleles can better our understanding of both
reproductive isolation and barriers to gene flow within this model system.

CONCLUSION

Our work contributes to our understanding of the evolutionary history of D.
*melanogaster*, including the distribution of genetic diversity and putative reproductive
isolation within this model organism. We combined population genomics to assess how
genetic and phenotypic diversity is partitioned across the globe in *D. melanogaster*,
especially within its putative ancestral range in Subtropical Africa. While we find
significant population genetic structuring throughout the range, we also find that
polymorphic reproductive barriers are largely decoupled from broader patterns of
population genetic structure in this system. Furthermore, this work provides insights into
the complexity of genetic changes underlying female mate choice, and highlights many
genes of interest for future study. In total, our work contributes to our understanding of
the factors influencing how genetic diversity is structured across the range in a model
system.

REFERENCES


1158  15, 250–255.
1159
1161
1163
1165
1167
1169
1171
1173
1175
1177
1179
1181
1183


78. Cortés-Ortiz L, Nidiffer MD, Hermida-Lagunes J, García-Orduña F, Rangel-Negrín


91. Ravinet M, Elgvin TO, Trier C, Aliabadian M, Gavrilo A, Sætre G-P. 2018


1332 Chen D, Qu C, Bjorum SM, Beckingham KM, Hewes RS. 2014 Neuronal
1333 Remodeling During Metamorphosis Is Regulated by the alan shepard (shep) Gene
1335 Griffith LC, Wang J, Zhong Y, Wu CF, Greenspan RJ. 1994 Calcium/calmodulin-
1336 dependent protein kinase II and potassium channel subunit eag similarly affect
**SUPPLEMENTAL MATERIALS**

**Table S1:** Sample information for all individuals in this study. For each accession, the unique ID (sample), species (each are species names of the genus *Drosophila*), NCBI SRA code (new samples will be given a unique code upon data upload), the sequencing platform from which the data came, the geo-genetic lineage which the sample belongs to, collection location information, and the average depth of coverage are given.

**Table S2:** $F_{ST}$ (blue; lower triangle), $D_{xy}$ (green; upper triangle), and $\pi$ (diagonal) for all geo-genetic lineages of *D. melanogaster* and *D. simulans*.


**Table S4:** Mean $f_{dm}$ values for each chromosome arm by each focal comparison. Significant differences between chromosome arms for each comparison are indicated by superscript letters.

**Table S5:** Introgression Outliers. Top 1% of $f_{dm}$ outliers for each of three comparisons: West Africa and Central Africa (WestCen), HD and OOA (OOAHD), and West Africa and the SE United States (WestSena). Genes within the outlier window and those proximal (as well as their proximity in KB) are given.

**Table S6:** Summary of behavioral survey. Average proportion of Z males chosen across trials for each isolate phenotyped in our behavioral survey, as well as the total trials set up, the total matings recorded, and the proportion of successful trials (mating proportion). We then defined each line as a behavioral type (no preference= NP; preference= P) based on both a Fisher’s Exact and binomial test (these analyses agreed completely on type definition with a significance cutoff of p<0.05).

**Table S7:** Behavioral PBE outliers. Top 1% of PBE outliers with Central African lines that show significant mate preference as the focal population, and Central African lines with no strong mate preference and RAL lines as the non-focal populations. Names of genes directly within the outlier window, as well as names of genes in close proximity are listed for each window, along with their proximity in KB up/downstream.

**Table S8:** Behavioral PBE GO analyses. GO analyses of the genes in or proximal to PBE outliers. Performed using PANTHER v.16.0.
Table S9: Putative incompatibilities with \(F_{st}\) outliers. A list of putative incompatibilities identified by [40] and [41] in which both interacting loci show signatures of older or more recent derivation (denoted in the Popgen signature). Window positions were translated between V.5 and V.632 of the \textit{D. melanogaster} reference genome, as different reference genomes were used between [40] and the current study. As some windows identified by pool contain multiple \(F_{ST}\) outliers, the number of outliers contained for each locus window, plus the genes that fall within that outlier are listed.
**Figure S1**: PCs 3 and 4 of a genome-wide PCA. *Points are colored based on the geographic lineage. Percent of variance explained by each PC is indicated in brackets.*
Figure S2: Full ML phylogeny of all samples based on (A) autosomes and (B) the X chromosome only. Tip colors correspond with the geo-genetic lineages defined herein, plus *D. simulans* and *D. yakuba* as outgroups.
Figure S3: Tajima’s D by chromosome for each geo-genetic lineage. Letters denote significantly different groups.
Figure S4: Strength of female mate preference differs by population. Proportion of Z males copulated in a two-way choice trial between Z and M males and female flies from 10 sampling locales in Subtropical Africa. Points represent individual trials for individual isolines.
Figure S5: No genetic structure between lines with or without strong mate preference; denoted in red and blue respectively. (A) PCA of all Central African lines that were phenotype for mate preference. Substantial genetic structure between HD and Central clades, but no structuring between lines that do/do not show significant female mate preference. Moreover, there is no genetic structure between behavioral types across the first 8 PCs. Numbers in brackets denote the percent of variance explained by PC1 and PC2, respectively. (B) ML consensus phylogeny built from non-overlapping regions of 100KB using the rapid-hill climbing algorithm in RAxML, with 50 bootstraps per tree. Consensus phylogeny was built using these trees as input to ASTRAL. Each genetic clade is outlined with a thick vertical bar, with colors replicating the represented clades in Figure 1. We also include six lines from France (denoted by the blue OOA section) for context. We note that only 17 of 19 individuals with strong preference and have sequenced genomes are included in the phylogeny as 2 genomes did not pass quality thresholds for our phylogenetic analyses (LA66 and LA69).
Figure S6: Behavioral outliers are highly differentiated between Central Africa and other geo-genetic lineages. PBE outliers for behavior show increased $F_{st}$ relative to the rest of the genome for all comparisons (genome-wide versus outlier: $F=7.5732$, df=1, $p=0.0059$). However, when contrasting $F_{st}$ between Central Africa and specific geo-genetic lineages, differences between $F_{st}$ for genome-wide versus PBE behavior outliers is only significant for Central Africa versus OOA, North Africa, and HD.
Figure S7: Weighted topologies between West Africa and OOA. Highly similar landscapes of weighted topologies between West Africa and each of all OOA lines and just those from the SE-United States (SE-US) suggest a largely shared landscape of introgression.
**Figure S8: Courtship effort differs between males of different genotypes.**

Courtship effort (defined as the number of courtship attempts divided by latency) per behavioral type of male (M=Ral371, Z=ZS2) for each female in a two-way choice experiment. Z males attempt more courtings, particularly in the context of Z females.