Widespread introgression across a phylogeny of 155 *Drosophila* genomes

Anton Suvorov[1,*], Bernard Y. Kim[2], Jeremy Wang[3], Ellie E. Armstrong[2], David Peede[3], Emmanuel R. R. D'Agostino[3], Donald K. Price[4], Peter Wadell[5], Michael Lang[6], Virginie Courtier-Orgogozo[6], Jean R. David[7,8], Dmitri Petrov[2], Daniel R. Matute[3,†], Daniel R. Schrider[1,†], and Aaron A. Comeault[9,†,*]

1. Department of Genetics, University of North Carolina, Chapel Hill, North Carolina. USA.
2. Department of Biology, Stanford University, Stanford, CA, USA
3. Biology Department, University of North Carolina, Chapel Hill, North Carolina. USA
4. School of Life Sciences, University of Nevada, Las Vegas. USA
5. School of Fundamental Sciences, Massey University, Palmerston North. New Zealand
6. CNRS, Institut Jacques Monod, Université de Paris. France
7. Laboratoire Evolution, Génomes, Comportement, Ecologie (EGCE) CNRS, IRD, Univ. Paris-sud, Université Paris-Saclay, Gif sur Yvette, France
8. Institut de Systématique, Evolution, Biodiversité, CNRS, MNHN, UPMC, EPHE, Muséum National d'Histoire Naturelle, Sorbonne Universités, Paris, France
9. School of Natural Sciences, Bangor University, Bangor, Gwynedd, LL57 2DGA, UK

*correspondence: antony.suvorov@gmail.com and a.comeault@bangor.ac.uk
†These authors contributed equally to this work.

Abstract

Genome-scale sequence data has invigorated the study of hybridization and introgression, particularly in animals. However, outside of a few notable cases, we lack systematic tests for introgression at a larger phylogenetic scale across entire clades. Here we leverage 155 genome assemblies, from 149 species, to generate a fossil-calibrated phylogeny and conduct multilocus tests for introgression across 9 monophyletic radiations within the genus *Drosophila*. Using complementary phylogenomic approaches, we identify widespread introgression across the evolutionary history of *Drosophila*. Mapping gene-tree discordance onto the phylogeny revealed that both ancient and recent introgression has occurred, with introgression at the base of species radiations being particularly common. Our results provide the first evidence of introgression occurring across the evolutionary history of *Drosophila* and highlight the need to continue to study the evolutionary consequences of hybridization and introgression in this genus and across the Tree of Life.

Introduction

The extent of gene exchange in nature has remained one of the most hotly debated questions in speciation genetics. Genomic data have revealed that introgression is common across taxa, having been identified in major groups such as fungi (Eberlein et al., 2019; Leducq et al., 2016; Tusso et al., 2019), vertebrates (Lamichhaney et al., 2015; Racimo et al., 2015; Schumer et al., 2018; Vanderpool et al., 2020), insects (Edelman et al., 2019; Lohse et al., 2015; Turissini and Matute, 2017), and angiosperms (Pease et al., 2018, 2016). The evolutionary effects of introgression are diverse, and are determined by multiple ecological and genomic factors (Rhymer and Simberloff, 1996; Taylor and Larson, 2019). Once thought to be strictly deleterious, it has become increasingly clear that introgression can serve as a source of genetic variation used during local adaptation (Hedrick, 2013; Suarez-Gonzalez et al., 2018) and adaptive radiation (Marques et al., 2019; Meier et al., 2017). While our understanding of introgression as a widespread phenomenon has clearly improved, it remains unclear how often it occurs across taxa. Ideally, determining the frequency of introgression across the Tree of Life would leverage the signal from systematic analyses of clade-level genomic data without an *a priori* selection of taxa known to hybridize in nature.

At the phylogenetic scale, hybridization has typically been explored at relatively recent timescales. For example, studies of hybridization between cats (Felidae; 10-12 My; ~40 species; (Li et al., 2016)), butterflies (*Heliconius*; 10-15 My; 15 species; (Edelman et al., 2019)), cichlid fishes from the African rift lakes (0.5-10 My; ~27 species; (Malinsky et al., 2018; Meier et al., 2017; Svardal et al., 2020)), and wild tomatoes (*Solanum*; ~4 My; ~20 species; (Pease et al., 2016)) all rejected a purely bifurcating phylogenetic history. In each of these systems introgression has occurred relatively recently, as the common ancestor for each species group occurred no more than 15 million years ago. A notable exception is evidence for introgression across much deeper phylogenetic timescales among vascular plants (Pease et al., 2018) and primates (Vanderpool et al., 2020). In some species, there is also evidence that introgression has been a source of adaptive genetic variation that has helped drive adaptation (e.g., (Chen et al., 2018; Eberlein et al., 2019; Jones et al., 2018; Platt et al., 2019; Richards and Martin, 2017)). These results therefore show how introgression has both (1) occurred in disparate taxonomic groups and (2) promoted adaptation and diversification in some. Notwithstanding the study by Pease et al. (2018), we still require systematic tests of introgression that use clade-level genomic

data that spans both deep and shallow phylogenetic time to better understand introgression's generality throughout evolution.

Species from the genus *Drosophila* remain one of the most powerful genetic systems to study animal evolution. Comparative analyses suggest that introgression might be common during speciation in the genus (Turelli et al., 2014). Genome scans of closely related drosophila species have provided evidence of gene flow and introgression (Brand et al., 2013; Dyer et al., 2018; Garrigan et al., 2012; Kang et al., 2017; Lohse et al., 2015; Mai et al., 2020; Schrider et al., 2018; Turissini and Matute, 2017). There is also evidence of contemporary hybridization (Kao et al., 2015; Matute and Ayroles, 2014; Sawamura et al., 2016) and stable hybrid zones between a handful of species (Cooper et al., 2018; Lachaise et al., 2000; Matute, 2010). These examples of hybridization and introgression show that species boundaries can be porous but cannot be taken as *prima facie* evidence of the commonality of introgression. Therefore, we still lack any systematic understanding of the relative frequency of hybridization and subsequent introgression across *Drosophila*. Here we analyze patterns of introgression across a phylogeny generated using 155 whole genomes derived from 149 species of *Drosophila*, and the genomes of four outgroup species. These species span over 50 million years of evolution and include multiple samples from nine major radiations within the family Drosophilidae. We used two different phylogenetic approaches to test whether introgression has occurred in each of these nine radiations. We found numerous instances of introgression across the entire evolutionary history of drosophilid flies, some mapping to early divergences within clades up to 20-25 Mya. Our results provide a taxonomically unbiased estimate of the prevalence of introgression at a macroevolutionary scale. Despite few known observations of current hybridization in nature, introgression appears to be a widespread phenomenon across the phylogeny of *Drosophila*.

Results

A high-confidence phylogeny of 155 *Drosophila* genomes

We first used genome-scale sequence data to infer phylogenetic relationships among species in our data set. To achieve this, we annotated and generated multiple sequence alignments for 2,791 Benchmarking Universal Single-Copy Orthologs (BUSCOs; v3; (Seppey et al., 2019; Waterhouse et al., 2017)) across 155 independently assembled *Drosophila* genomes together with four outgroups (3 additional species from Drosophilidae and *Anopheles gambiae*;

Supplementary Data). We used these alignments, totalling 8,187,056 nucleotide positions, and fossil calibrations to reconstruct a fossil-calibrated tree of *Drosophila* evolutionary history. Note that the inclusion of *Anopheles* as an outgroup allowed us to include a fossil of *Grauvogelia*, the oldest known dipteran, in our fossil calibration analysis, along with several *Drosophilidae* fossils and/or geological information (i.e., formation of Hawaiian Islands; see SI Appendix, Table S1). Our phylogenetic analyses (see Materials and Methods for details) using both maximum-likelihood (ML; IQ-TREE) and gene tree coalescent-based (ASTRAL) approaches with DNA data revealed well-supported relationships among nearly all species within our dataset. Phylogenies inferred using these two approaches only differed in a single relationship, where *D. villosipedis* was either recovered as a sister species to *D. limitata* + *D. ochracea* (ML topology) or as a sister to *D. limitata* + *D. ochracea* + *D. murphyi* + *D. sproati* (ASTRAL topology). The nodal supports were consistently high across both ML (Ultrafast bootstrap (UFBoot) = 100, an approximate likelihood ratio test with the nonparametric Shimodaira–Hasegawa correction (SH-aLRT) = 100, a Bayesian-like transformation of aLRT (aBayes) = 1) and ASTRAL (Local posterior probability (LPP) = 1) topologies with the exception of *D. limitata* + *D. ochracea* + *D. villosipedis* (UFBoot = 9, SH-aLRT = 81, aBayes = 1) and *D. carrolli* + *D. rhopaloa* + *D. kurseongensis* (UFBoot = 81.2, SH-aLRT = 81, aBayes = 1) on the ML tree, and *D. limitata* + *D. ochracea* + *D. murphyi* + *D. sproati* (LPP = 0.97) and *D. sulfugaster bilimbata* + *D. sulfugaster sulfurigaster* (LPP = 0.69) on the ASTRAL tree. Thus, the phylogeny we report here is the first of the genus *Drosophila* with almost all nodes resolved with high confidence—recent estimates of the *Drosophila* phylogeny lacked strong support throughout all tree depth levels (O'Grady and DeSalle, 2018; Russo et al., 2013; Yassin, 2013). Furthermore, an ML topology estimated from the dataset with more closely related outgroup species (see Materials and Methods) results in an identical topology with the aforementioned ML tree. The inferred phylogeny from the protein supermatrix showed only two incongruencies: (i) *D. villosipedis* was recovered as a sister species to *D. limitata* + *D. ochracea* + *D. murphyi* + *D. sproati* and (ii) *D. watanabei* + *D. punjabiensis* is sister to *D. bakoue* + *D. jambulina* clade. We performed further assessment of nodal support with Quartet Sampling (Pease et al., 2018), using the Quartet Concordance (QC) and Quartet Differential (QD) scores to identify quartet-tree species-tree discordance (Materials and Methods). At some nodes, an appreciable fraction of quartets disagreed with our inferred species tree topology (QC < 1; Supplementary Data), and in most of these cases this discordance was

skewed toward one of the two possible alternative topologies (i.e. QD < 1 but > 0) as is consistent with introgression. We formally explore this pattern below.

In order to estimate divergence times across the *Drosophila* phylogeny, we developed five calibration schemes (A, B, C, D and "Russo"; described in SI Appendix, Table S1). Overall, four of the five schemes yielded nearly identical age estimates with narrow 95 % credible intervals (CI), whereas scheme "Russo" (a fossil calibration strategy closely matching that from Russo et al. (2013)) showed slightly older estimates with notably wider 95% CIs (SI Appendix, Fig. S1; (Supplementary Data). Throughout this manuscript we use the time estimates obtained with scheme A. This calibration analysis estimated that extant members of the genus *Drosophila* branched off from the other Drosophilidae (*Leucophenga*, *Scaptodrosophila* and *Chymomyza*) ~53 Mya (95% CI: 50 - 56.6 Mya) during the Eocene Epoch of the Paleogene Period (Fig. 1). The same analysis inferred that the split between the two major lineages within *Drosophila*—the subgenera of *Sophophora* and *Drosophila*—occurred ~47 Mya (95% CI: 43.9- 49.9 Mya; Fig. 1; "A" and "B" clades, respectively); previous estimates of this time include ~32 Mya (95% CI: 25–40 Mya) as estimated by Obbard et al. (2012), ~63 Mya (95% CI: 39–87 Mya) by Tamura et al. (2004), and ~56 Mya (95% CI not available) by Russo et al. (2013). We also note that our divergence time estimates of the *Drosophila* subgenus (~34 Mya, 95% CI: 31.6 - 36.8 Mya; Clades 6 through 9) are somewhat younger than ~40 Mya, a previous estimate reported in Izumitani et al. (2016), although the latter had fairly wide confidence intervals (95% CI: 33.4 - 47.6 Mya).

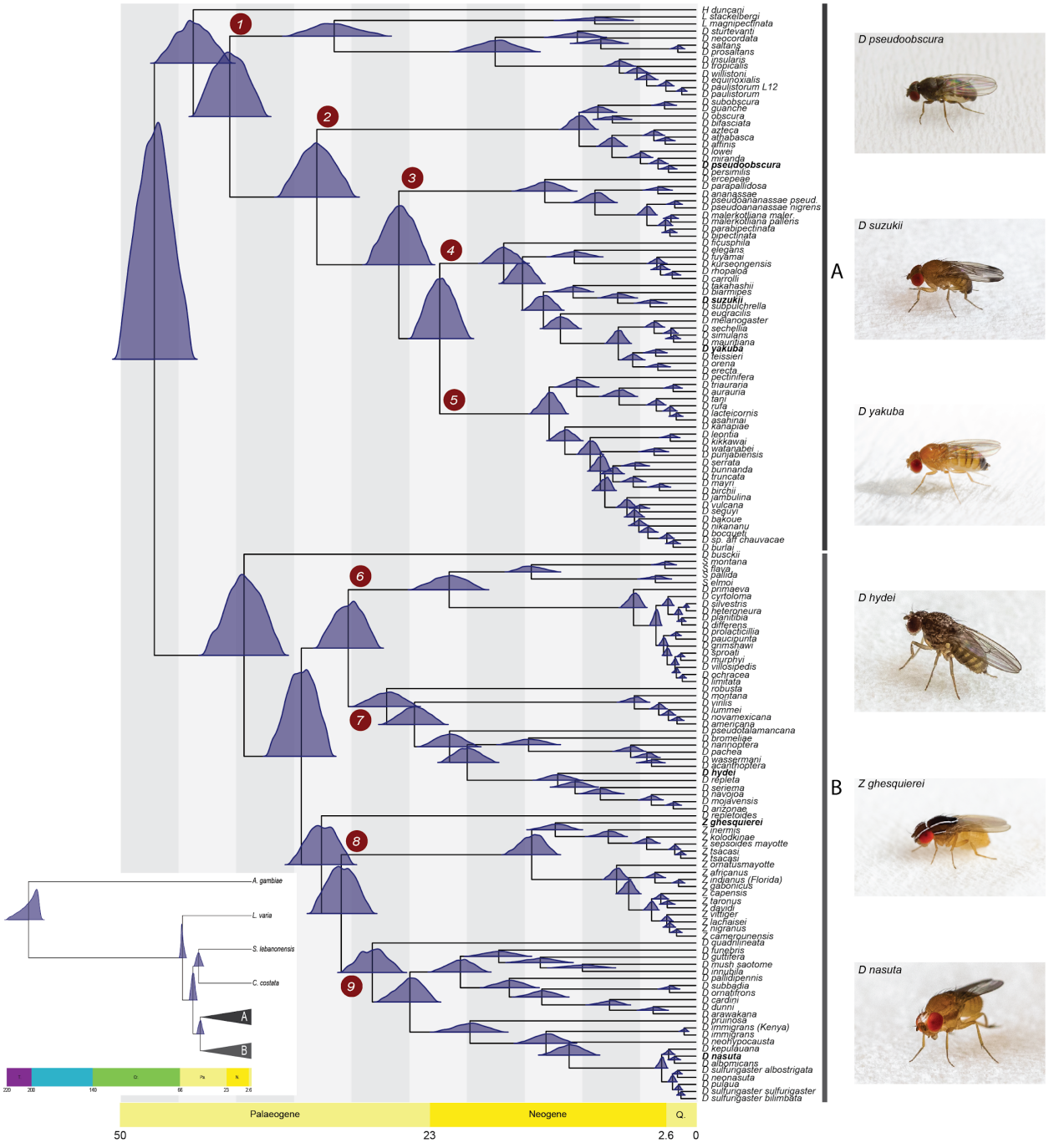

Fig. 1. Fossil calibrated maximum likelihood phylogenetic tree of the genus *Drosophila* inferred from a supermatrix of 2,791 BUSCO loci (total of 8,187,056 sites). The blue distributions at each divergence point on the tree represent nodal age posterior probabilities from MCMCTree. *Grauvogelia* and *Oligophryne* fossils were used to set priors on the age of the root of the tree, *Phytomyzites* and *Electrophortica succini* were used for priors for the root of the Drosophilidae family, and *Electrophortica succini* and *Scaptomyza dominicana* were used to set priors for the crown group "*Scaptomyza*", i.e. Most Recent Common Ancestor (MRCA) node of the *Scaptomyza* species (scheme A; SI Appendix, Table S1). The numbered red circles denote clades for which analyses of introgression were performed. Inset: the phylogenetic and temporal relationships between our distant outgroup *Anopheles gambiae*, more closely related outgroup species of Drosophilidae (*Leucophenga varia*, *Scaptodrosophila lebanonensis* and *Chymomyza costata*), and the *Drosophila* genus. A and B denote the two inferred major monophyletic groups within *Drosophila*.

Widespread signatures of introgression across the *Drosophila* phylogeny

To assess the prevalence of introgression across the *Drosophila* tree, we subdivided species into nine monophyletic lineages (herein referred to as clades 1 through 9; Fig. 1) and tested for introgression within each clade. These clades correspond to the deepest divergences within the genus, with most having an MRCA during the Paleogene. Clades 4 and 5 are the two exceptions, splitting from an MRCA later in the Neogene. Within each of the nine clades, the MRCA of all sampled genomes ranged from ~10 Mya (Fig. 1; clade 2) to ~32 Mya (Fig. 1; clade 1). We note that *Hirtodrosophila duncani*, *Drosophila busckii* and *Drosophila repletoides* were not included in these clade assignments as each of these species was the only sampled descendent of a deep lineage; additional taxon sampling is required to assign them to specific monophyletic species groups that could be tested for introgression.

We tested for introgression within each of these nine clades using two complementary phylogenomic methods that rely on the counts of gene trees inferred from the BUSCO loci that are discordant with the inferred trees (hereafter referred to as the discordant-count test or DCT) and the distribution of branch lengths for discordant gene trees (hereafter termed the branch-length test or BLT), respectively, among rooted triplets of taxa (both illustrated in SI Appendix, Figs. S2 and S10). These methods leverage information contained across a set of gene trees to differentiate patterns of discordance that are consistent with introgression from those that can be explained by incomplete lineage sorting alone (see Materials and Methods). Using these approaches, in 8 of our 9 clades we found at least one pair of species with evidence of introgression according to both DCT and BLT (i.e., the same pair of species showed evidence for introgression that was significant in both tests at an FDR-corrected *P*-value threshold of 0.05). Moreover, the overlap in species pairs with introgression detected by DCT and BLT was significant in 5 of the 9 clades analyzed ($P < 2.3 \times 10^{-7}$ in all cases; Figs. 2A,B and 3A,B and SI Appendix, Fig. S3). Because these two methods rely on independent analyses to detect introgression (i.e., counts of discordant trees and the distribution of branch lengths among trees, respectively), their highly significant overlap provides strong support for the presence of introgression across the *Drosophila* phylogeny. We found even stronger support for introgression across these clades using QuIBL (Edelman et al., 2019) (SI Appendix, Figs. S4C and S5C); however, we focus here on the overlap between DCT and BLT methods (after correcting each for

multiple testing), as this provides a more conservative estimate of the extent of introgression. We obtained similar results when less stringent criteria were used to define overlap between DCT and BLT (SI Appendix, Fig. S3), finding a significant number of pairs of species with evidence of introgression according to both tests in seven of the nine clades (all but Clades 1 and 3).

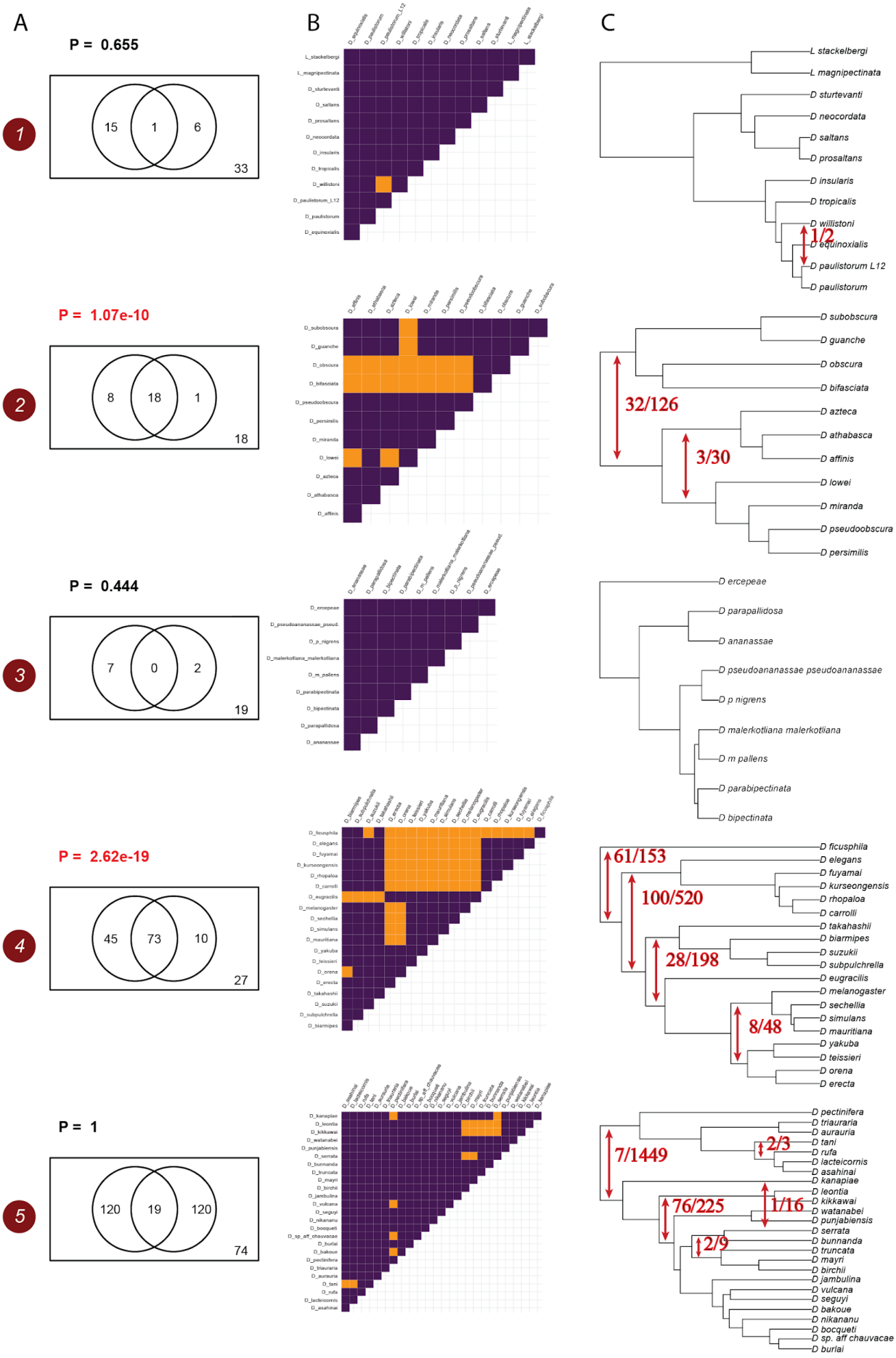

Fig. 2. Patterns of introgression inferred for the monophyletic clades 1-5 of the subgenus *Sophophora* (Species Group A in Fig. 1). (A) The Venn diagrams show agreement between DCT and BLT methods for identifying introgression. An introgression event between a pair of species was considered significant if at least one triplet used to test for introgression between this pair of species was significant for both DCT and BLT with FDR < 0.05. The *P*-value of the hypergeometric test used to assess whether there was a significant excess of species pairs with evidence for introgression from both tests. (B) Triangular matrices showing species pairs with evidence of introgression according to the same criterion as in (A) are shown in orange and all others in blue. (C) Introgression events mapped onto the corresponding clades. Ratios represent the number of triplets that support an inferred introgression event (red lines) over the total number of triplets we analysed that could have detected the event.

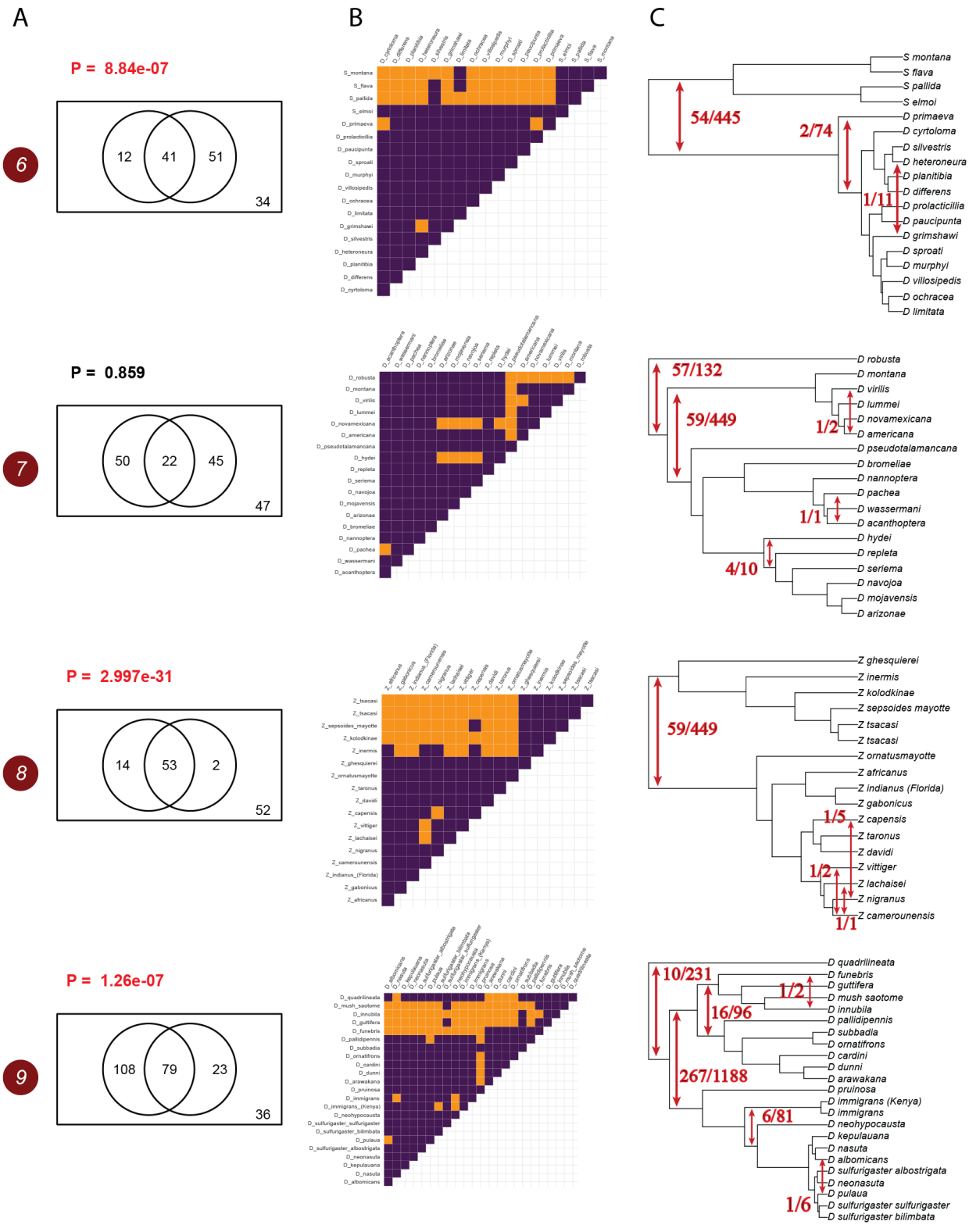

Fig. 3. Patterns of introgression inferred for the monophyletic clades 6-9 of the subgenus *Drosophila* (Species Group B in Fig. 1). Panels A-C summarize analysis as presented in Figure 2 (see caption for details).

In clade 1 there was only a single species pair for which the DCT and BLT were significant in the same triplet, and no such pairs in clade 3 (although both clades had several pairs significant according to one test or the other after FDR correction; Fig. S4). We asked whether the lower level of introgression detected in groups 1 and 3 was caused by their smaller sample size with respect to the other seven clades. We performed a power analysis by downsampling each clade to eight species (the smallest number in any of the nine clades minus one) and again tested for significant overlap between DCT and BLT (SI Appendix, Fig. S6). Results from this analysis show that the power to detect agreement between DCT and BLT in the downsampled datasets (i.e. 1 - P(Type II error)) is greater than 0.5, on average, in clades where significant overlap was detected in the full dataset. This suggests that, although our conservative analysis may have missed cases of introgression, if introgression were as common in Clades 1 and 3 as in the rest of the phylogeny we would have had reasonable power to detect it. These results were consistent regardless of the criteria used to define overlap (SI Appendix, Fig. S6).

The number of species pairs that show evidence of introgression is not equivalent to the number of independent introgression events among *Drosophila* species. This is because gene flow in the distant past can create evidence of introgression in multiple contemporary species pairs. For example, we found evidence for introgression between *D. eugracilis* and *D. biarmipes*, *D. takahashii*, *D. suzukii*, and *D. subpulchrella* (see Clade 4 heatmaps in Figure 2B). Rather than four independent instances of introgression between species, this pattern could reflect introgression between ancestral taxa that subsequently diverged into the contemporary species. Thus, we calculated how many species pairs showed overlapping DCT and BLT signals that spanned a shared ancestral node (i.e., spanned a shared MRCA). Cases where multiple species pairs each shared the same MRCA were considered to be indicative of a single ancestral introgression event between the branches that coalesce at this node, while those involving only a single species pair were considered to have resulted from introgression between the extant species pair (Figs. 2C and 3C). An example of the former can be seen in clade 6 where the evidence suggests introgression occurred between the Hawaiian *Scaptomyza* and *Drosophila* (Fig 3C) that are estimated to have diverged from each other more than 20 Mya. This ancient introgression may have occurred prior to the formation of Kauai island ~5 Mya which is now the oldest high island with extant species in these two groups (Magnacca and Price, 2015; Price and

Clague, 2002). In addition to producing a relatively conservative estimate of the minimum number of introgression events required to generate the patterns of discordance we observed in our data set, this approach yields a rough upper bound on the timing of ancestral introgression events. However, we note that our approach does recover some signatures of introgression that may be false positives (e.g., introgression between the relatively distantly related *D. serrata* and *D. kanapiae*; this event is supported by only one of the 16 triplets that could have detected it; Fig 2C).

According to the approach described above, a minimum of 30 independent introgression events are required to generate the distribution of discordant gene trees that our DCT-BLT method identified as consistent with introgression (red lines in Figs. 2C and 3C). For example, Clades 4, 6, 8, and 9 showed some of the strongest evidence of introgression, with 73, 41, 53, and 79 species pairs displaying signals of introgression based on both DCT and BLT methods (Fig. 2A and 3A). The minimum number of introgression events required to generate these signals is 4, 3, 4, and 6, respectively (Fig. 2C and 3C). All clades showed evidence of at least two introgression events except Clade 1 (1 event) and 3 (0 events). Therefore, our analyses provide evidence for multiple independent introgression events across at least 7 of the 9 clades with at least two to six independent events occurring within each clade generating the observed patterns of gene-tree discordance (Figs. 2C and 3C). Again, we stress that both our methods used to detect introgression (DCT and BLT) and our approach for counting introgression events are conservative, and thus the true number of events could be substantially greater. We also note that the majority of the introgression events predicted by our approach are quite ancient, although this could in part be a consequence of how we mapped introgression events onto the phylogeny. Careful examination of results in Figs 2B and 3B reveals that deep introgression events are clearly the best explanation for some of our patterns (e.g., the case from Clade 4 involving *D. eugracilis* described above), although more recent events have occurred as well (e.g. between *D. albomicans* and *D. pulaua*; Fig 3, clade 9).

We also used PhyloNet (Than et al., 2008; Wen et al., 2018) as an alternative approach to determine which branches exhibited the strongest signature of introgression within each of the nine monophyletic clades in our tree. To this end, within each clade we examined all possible network topologies produced by adding a single reticulation event to the species tree and determined which of the resulting phylogenetic networks produced the best likelihood score. We

note that networks with more reticulation events would most likely exhibit a better fit to observed patterns of introgression but the biological interpretation of complex networks with multiple reticulations is more challenging; thus, we limited ourselves to a single reticulation event even though this will produce false negatives in clades with multiple gene flow events. For all clades except 6 and 8, the networks with the highest likelihood scores from PhyloNet qualitatively agree with the inferred introgression patterns by DCT-BLT: the best-supported position of a reticulation event inferred by PhyloNet tended to occur in the same or similar locations on the tree as we inferred with our DCT-BLT analysis (SI Appendix, Figs. S7 and S8). On the other hand, PhyloNet inferred introgression events in Clades 6 and 8 that were not consistent with the admixture events inferred with DCT and BLT. The reasons for these differences are unclear but might be related to our limited inference of introgression with PhyloNet (i.e., forcing the occurrence of only one introgression event). Uncertainty over the precise history of introgression in clades 6 and 8 notwithstanding, PhyloNet is consistent with our other results in that introgression has occurred across the *Drosophila* phylogeny.

Discussion

A time-calibrated tree of drosophilid evolution

Drosophila, as a genus, remains a premier model in genetics, ecology, and evolutionary biology. With over 1,600 species (O'Grady and DeSalle, 2018), the genus has the potential to reveal why some groups are more speciose than others. Yet the phylogenetic relationships among the main groups in the genus have remained largely unresolved (reviewed in (O'Grady and DeSalle, 2018)). Here we estimated a robust time-calibrated phylogeny for the whole genus using multilocus genomic data and calibrated it using a fossil record.

Our results confirm that the genus *Drosophila* is paraphyletic, with the genera *Zaprionus*, *Scaptomyza*, *Leucophenga*, and *Hirtodrosophila* each nested within the larger genus *Drosophila*. Consistent with the subdivisions previously proposed by (Throckmorton, 1975) and (Yassin, 2013), clades 1-5 of our phylogeny contains species belonging to the monophyletic subgenus *Sophophora*, and includes species from the genus *Lordiphosa* (group A in Figure 1). Clades 6-9 of our phylogeny contains species belonging to the monophyletic subgenus *Drosophila* (group B in Figure 1) and include species from the Hawaiian *Drosophila* and the genera *Siphlodora*, *Phloridosa*, and *Zaprionus*. For more recent radiations within *Drosophila*, the topology we

present is largely congruent with previous studies (Izumitani et al., 2016; O'Grady and DeSalle, 2018) but two general observations are notable. First, our results confirm that *Lordiphosa* is closely related to the *saltans* and *willistoni* groups (Clade 1) and part of the *Sophophora* subgenus (consistent with (Katoh et al., 2000)). Second, we confirm that *Zaprionus* is related to the *cardini/qunaria/immigrans* group (consistent with (O'Grady and DeSalle, 2018) and Throckmorton (1975), but discordant with Russo et al. (2013)). Despite our well resolved phylogeny, comparisons with other studies emphasize the need to expand species sampling, especially given the potential to generate highly contiguous genomes at relatively low cost (Kim et al., 2020).

Our results from divergence time analysis suggest that the origin of *Drosophila* (including the subgenera *Sophophora* (clade A) and *Drosophila* (clade B)) occurred during the Eocene Epoch of the Paleogene, which is younger than estimates by Throckmorton (1975), Tamura et al. (2004), and Russo et al. (2013), but older than estimated by Obbard et al. (2012). These differences in divergence time estimates may be a result of different calibration information used, such as mutation rates, the time of formation of the Hawaiian Islands, and the fossil record. However, our comparison of various calibration schemes suggests that the choice of calibration information has a minor effect on age estimation (SI Appendix, Fig. S1). Additionally, credible intervals around our estimates tend to be notably narrower than in all of the aforementioned studies. In contrast to the previous studies, we used genome-scale multilocus data which would be expected to improve both the accuracy and precision of age estimates (Reis and Yang, 2013; Yang and Rannala, 2006).

The extent of introgression in *Drosophila*

Access to genome-scale data has reinvigorated the study of hybridization and introgression (Taylor and Larson, 2019). We used genome-scale sequence data to provide the first systematic survey of introgression across the phylogeny of drosophilid flies. Our complementary—and conservative—approaches identified overlapping evidence for introgression within eight of the nine clades we analyzed (Figs. 2 and 3). We conclude that at least 30 pairs of lineages have experienced introgression across *Drosophila*'s history. This number should be treated as an approximate lower bound and we cannot rule out the possibility that the true number is substantially higher. Studies in contemporary *Drosophila* species suggest

that selection may constrain the evolution of mixed ancestry, at least in naturally occurring (Cooper et al., 2018; Meiklejohn et al., 2018; Turissini and Matute, 2017) and experimental admixed populations (Matute et al., 2020). The results we have presented here utilized phylogenetic signals to show that introgression has nonetheless occurred and left a detectable signal within the genomes of many extant *Drosophila*.

In addition to providing an estimate of the extent of introgression, our results are informative about the timing of introgression among *Drosophila* lineages: the approach we used to estimate the number of introgression events, and map them onto the phylogeny produces a rough upper-bound estimate of the timing of these events. Thus, although many of the cases of introgression recovered by our approach appear to be relatively ancient, and map to early divergences within each of the nine clades we analyzed, some of these may in fact be a result of somewhat more recent gene flow events. As described in the Results, both our PhyloNet analyses and a careful examination of our DCT-BLT results are most consistent with ancient introgression events. We also find evidence for very recent events, and although our analyses did not search for gene flow between sister taxa, previous studies of closely related species in *Drosophila* have revealed evidence of introgression (Garrigan et al., 2012; Lohse et al., 2015; Mai et al., 2020; Schrider et al., 2018; Turissini and Matute, 2017). Studies that have taken phylogenomic approaches to detect introgression in other taxa have also reported evidence for introgression between both 'ancient' lineages (i.e., those that predate speciation events generating extant species) and extant species (Edelman et al., 2019; Li et al., 2016; Meier et al., 2017; Pease et al., 2016; Svardal et al., 2020). We conclude that introgression between *Drosophila* flies has similarly occurred throughout their evolutionary history.

Although the signal of introgression across our phylogeny provides evidence for widespread introgression in *Drosophila*, the evolutionary role of introgressed alleles remains to be tested. For example, the impact of hybridization and introgression on evolution can be diverse, from redistributing adaptive genetic variation (Anderson et al., 2009; Dasmahapatra, 2012; Jones et al., 2018) to generating negative epistasis between alleles that have evolved in different genomic backgrounds ((Fishman and Sweigart, 2018; Maheshwari and Barbash, 2011; Nosil and Schluter, 2011); reviewed in (Baack and Rieseberg, 2007; Hedrick, 2013; Moran et al., 2020; Suarez-Gonzalez et al., 2018)). The number of introgressed alleles that remain in a hybrid lineage depends on their selection coefficients (Harris and Nielsen, 2016; Kim et al., 2018;

Sachdeva and Barton, 2018), their location in the genome (i.e., sex chromosomes vs. autosomes, (Geraldes et al., 2006; Payseur et al., 2004; Storchová et al., 2010)), levels of divergence between the hybridizing species (Hamlin et al., 2020; Kronforst et al., 2013; Turissini and Matute, 2017), and recombination rates among loci (Martin et al., 2019; Schumer et al., 2018). Hybrids between species of *Drosophila* often show maladaptive phenotypes (Cooper et al., 2018; Coyne and Orr, 1997, 1989; Serrato-Capuchina et al., 2020; Turissini et al., 2018, 2017). Similarly, laboratory experiments studying hybrids generated from two independent species pairs of *Drosophila* have shown that hybrid swarms can quickly evolve to represent only one of their two parental species, while the genome of the second species is rapidly purged from the populations (Matute et al., 2020). These results show how hybrid *Drosophila* can be less fit than their parents, and further work is needed to determine the evolutionary effects of the introgression that we report here. Our results do, however, add to the growing body of literature that document a detectable phylogenetic signal of introgession left within the genomes of a wide range of species radiations that include *Drosophila*, other dipterans (Fontaine et al., 2015), lepidopterans (Dasmahapatra et al., 2012; Edelman et al., 2019; Martin et al., 2019), humans (Green et al., 2010; Juric et al., 2015; Racimo et al., 2015), fungi (Eberlein et al., 2019; Tusso et al., 2019), and angiosperm plants (Pease et al., 2018, 2016).

Caveats and future directions

We estimated the minimal number of events that explain the introgression patterns across the tree and in some cases those events were recovered as relatively ancient. However, our parsimonious approach favors older events over repeated and recent introgressions, and thus may bias the age of introgression towards ancient events and underestimate the true number of pairs of lineages that have exchanged genetic material. For example, introgression events we inferred at deeper nodes in our phylogeny tended to be supported by a subset of comparisons between species pairs that spanned those nodes (see 'ancient' introgression events in clades 2, 4-9; Figs. 2C and 3C). Some patterns we observe may therefore reflect scenarios where introgression has persisted along some lineages but been purged along others. Future efforts could also try to identify phylogenetic signatures of introgression between extant and extinct lineages (or lineages missing from our phylogeny), a pattern referred to as "ghost" introgression (Durvasula and Sankararaman, 2020; Ottenburghs, 2020).

Our analyses also do not identify the precise alleles that have crossed species boundaries or reveal the manner in which these alleles may have affected fitness in the recipient population (Baack and Rieseberg, 2007; Moran et al., 2020). Genome alignments, complete annotations, and/or population level sampling across the genus are required to determine whether certain genes or functional categories of genes are more likely to cross species boundaries than others. More complete taxonomic sampling, combined with methodological advances for inferring the number and timing of introgression events in large phylogenies, will increase our ability to identify the specific timing of introgression across *Drosophila*.

Conclusions

Speciation research has moved away from the debate of whether speciation can occur with gene flow to a more quantitative debate of how much introgression occurs in nature, and what are the fitness consequences of that introgression for the individuals in a population. Our well-resolved phylogeny and survey of introgression revealed that introgression has been a relatively common feature across the evolutionary history of *Drosophila*. Yet, identifying the specific consequences of introgression on fitness and the evolution of species and entire radiations within *Drosophila* and other systems remains a major challenge. Future research could combine the power of phylogenomic inference with population-level sampling to detect segregating introgression between sister species to further our understanding of the amount, timing, and fitness consequences of admixture for diversification.

Materials and Methods

Genome assemblies and public data

Genome sequences used by this work were obtained from concurrent projects and public databases. Genome sequencing and assembly for 84 genomes is described in (Kim et al., 2020). These data are available for download at NCBI BioProject PRJNA675888. For the remaining genomes: sequencing and assembly of 8 Hawaiian *Drosophila* were provided by E. Armstrong and D. Price, described in Armstrong et al. (in prep) and available at NCBI BioProject PRJNA593822; sequences and/or assemblies of five *nannoptera* group species were provided by M. Lang and V. Courtier-Orgogozo and are available at NCBI BioProject PRJNA611543; 44

were downloaded as assembled sequences from NCBI GenBank; *Z. sepsoides* and *D. neohypocausta* were sequenced as paired-end 150bp reads on Illumina HiSeq 4000 at UNC and assembled using SPAdes v3.11.1 with default parameters (Bankevich et al., 2012); and 15 were generated by assembling short read sequences downloaded from NCBI SRA. For sets of unassembled short reads, we used ABySS v2.2.3 (Jackman et al., 2017) with parameters 'k=64' with paired-end reads (typically 100-150bp) to assemble the reads. Finally, outgroup genome sequences (*A. gambiae*, *M. domestica*, *L. trifolii*, *C. hians*, and *E. gracilis*) were obtained from NCBI GenBank. See Table S2 for a full list of samples, strain information, accessions, and associated publications.

Orthology Inference

We identified single-copy orthologous genes in each genome using BUSCO (Benchmarking Universal Single-Copy Orthologs; v3.1.0; (Simão et al., 2015)). BUSCO was run with orthologs from the Diptera set in OrthoDB v.9 (odb9) using default parameters. For each species, all BUSCOs found in a single copy were used for phylogenetic analysis.

Phylogenetic reconstruction

Every DNA BUSCO locus was aligned with MAFFT v7.427 (Katoh et al., 2002) using the L-INS-i method (Supplementary Data). We removed sites that had fewer than three non-gap characters from the resulting multiple sequence alignments (MSAs). These trimmed MSAs were concatenated to form a supermatrix. We inferred a maximum likelihood (ML) phylogenetic tree (Supplementary Data) from the supermatrix (a.k.a. concatenated alignment) using IQ-TREE v1.6.5 (Nguyen et al., 2015), and treated the supermatrix as a single partition. IQ-TREE was run under GTR+I+G substitution model, as inference under any other substitution model will not necessarily lead to better accuracy of tree topology estimation (Abadi et al., 2019). To estimate the support for each node in this tree, we used three different reliability measures. We did 1,000 ultrafast bootstrap (UFBoot) replicates (Minh et al., 2013) and additionally performed an approximate likelihood ratio test with the nonparametric Shimodaira–Hasegawa correction (SH-aLRT) and a Bayesian-like transformation of aLRT (Anisimova et al., 2011). We used the ML gene trees obtained by IQ-TREE with a GTR+I+G substitution model for tree inference in ASTRAL (Sayyari and Mirarab, 2016). For the estimated ASTRAL tree (Supplementary Data)

we calculated the support of each node using local posterior probabilities (LPP) (Sayyari and Mirarab, 2016).

We did two additional analyses to verify the robustness of our topology inference. First, we inferred an ML tree using WAG+I+G substitution model from the protein supermatrix obtained from concatenation of protein BUSCO MSAs (Supplementary Data). MSAs based on amino acid sequences have been shown to have superior accuracy to DNA MSAs for distantly related species (Bininda-Emonds, 2005). Second, to verify that long branch attraction did not distort our tree topology, we inferred an ML tree under a GTR+I+G substitution model using a different set of outgroup species from the DNA supermatrix (Supplementary Data). Specifically, instead of distantly related *Anopheles gambiae*, we used *Musca domestica, Liriomyza trifolii*, *Curricula hians* and *Ephydra gracilis* together as our outgroup species (Supplementary Data).

Phylogenetic Support Analysis via Quartet Sampling

We used quartet sampling (QS) as an additional approach to estimate phylogenetic support (Pease et al., 2018). Briefly, QS provides three scores for internal nodes: (*i*) quartet concordance (QC), which gives an estimate of how sampled quartet topologies agree with the putative species tree; (*ii*) quartet differential (QD) which estimates frequency skewness of the discordant quartet topologies, and can be indicative of introgression if a skewed frequency observed, and (*iii*) quartet informativeness (QI) which quantifies how informative sampled quartets are by comparing likelihood scores of alternative quartet topologies. Finally, QS provides a score for terminal nodes, quartet fidelity (QF), which measures a taxon "rogueness". We did QS analysis using the DNA BUSCO supermatrix described above, specifying an IQ-TREE engine for quartet likelihood calculations with 100 replicates (i.e., number of quartet draws per focal branch).

Fossil Dating

We implemented the Bayesian algorithm of MCMCTREE (Yang, 2007) with approximate likelihood computation to estimate divergence times within *Drosophila* using several calibration schemes (SI Appendix, Table S1). First, we estimated branch length by ML and then the gradient and Hessian matrix around these ML estimates in MCMCTREE using the DNA supermatrix. Because large amounts of sequence data are not essential for accurate fossil

calibration (Anisimova, 2012), we performed dating analysis using a random sample of 1,000 MSA loci (out of 2,791) for the sake of computational efficiency. Thus, for this analysis the supermatrix was generated by concatenating 1,000 randomly selected gene-specific MSAs. Using fewer loci (10 and 100) for fossil calibration did not drastically affect nodal age estimation (SI Appendix, Fig. S9; Supplementary Data). We removed sites that had less than 80 non-gap characters from all these supermatrices. Second, we used the gradient and Hessian matrix, which constructs an approximate likelihood function by Taylor expansion (dos Reis and Yang, 2011), to perform fossil calibration in MCMC framework. For this step we specified a GTR+G substitution model with four gamma categories; birth, death and sampling parameters of 1, 1 and 0.1, respectively. To ensure convergence, the analysis was run for 7×10^6 generations (first 2×10^6 generations were discarded as burn-in), logging every 1,000 generations. We used the R package MCMCtreeR (Puttick, 2019) to visualize the calibrated tree.

Inferring Introgression Across the Tree

Triplet-based methods: In order to detect patterns of introgression we used three different methods that rely on the topologies of gene trees, and the distributions of their corresponding branch lengths, for triplets of species. If the true species tree is ((A, B), C), these tests are able to detect cases of introgression between A and C, or between B and C. These include two of the methods that we devised for this study, and which use complementary pieces of information—the counts of loci supporting either discordant topology, and the branch-length distributions of gene trees supporting these topologies, respectively—to test an introgression-free null model.

The first method we developed was the discordant-count test (DCT) (Supplementary Data), which compares the number of genes supporting each of the two possible discordant gene trees: ((A, C), B) or (A, (B, C)), similar in principle to the delta statistic from (Huson et al., 2005). Genes may support the two discordant topologies (denoted T_1 and T_2) in the presence of ILS and/or in the presence of introgression. In the absence of ancestral population structure, gene genealogies from loci experiencing ILS will show either topology with equal probability; ILS alone is not expected to bias the count towards one of the topologies. In the presence of introgression, one of the two topologies will be more frequent than the other because the pair of species experiencing gene flow will be sister lineages at all introgressed loci (illustrated in SI Appendix, Figs. S2 and S10). For example, if there is introgression between A and C, there will

be an excess of gene trees with the ((A, C), B) topology. The DCT identifies pairs of species that may have experienced introgression by performing a χ^2 goodness-of-fit test on the gene tree count values for a species triplet to determine whether their proportions significantly deviate from 0.5, the expected proportion for each gene genealogy under ILS. We used this test on all triplets extracted from BUSCO gene trees within each clade, and the resulting *P*-values were then corrected for multiple testing using the Benjamini-Hochberg procedure with a false discovery rate cutoff (FDR) of 0.05.

Second, we devised a branch length test (BLT) (Supplementary Data) to identify cases of introgression (illustrated in SI Appendix, Figs. S2 and S10). This test examines branch lengths to estimate the age of the most recent coalescence event (measured in substitutions per site). Introgression should result in more recent coalescences than expected under the concordant topology with complete lineage sorting, while ILS shows older coalescence events (Fontaine et al., 2015). Importantly, ILS alone is not expected to result in different coalescence times between the two discordant topologies, and this forms the null hypothesis for the BLT. For a given triplet, for each gene tree we calculated the distance *d* (a proxy for the divergence time between sister taxa) by averaging the external branch lengths leading to the two sister taxa under that gene tree topology. We calculated *d* for each gene tree and denote values of *d* from the first discordant topology d_{T1} and those from the second discordant topology d_{T2}. We then compared the distributions of d_{T1} and d_{T2} using a Mann-Whitney *U* test. Under ILS alone the expectation is that $d_{T1} = d_{T2}$, while in the presence of introgression $d_{T1} < d_{T2}$ (suggesting introgression consistent with discordant topology T_1) or $d_{T1} > d_{T2}$ (suggesting introgression with consistent with topology discordant T_2). The BLT is conceptually similar to the D3 test (Hahn and Hibbins, 2019), which transforms the values of d_{T1} and d_{T2} in a manner similar to the *D* statistic for detecting introgression (Green et al., 2010). As with the DCT, we performed the BLT on all triplets within a clade and used a Benjamini-Hochberg correction with a false discovery rate cutoff (FDR) of 0.05. We note that both the DCT and BLT will be conservative in cases where there is bidirectional introgression, with the extreme case of equal rates of introgression in both directions resulting in a complete loss of power.

Finally, we used QuIBL, an analysis of branch length distribution across gene trees to infer putative introgression patterns (Supplementary Data). Briefly, under coalescent theory internal branches of rooted gene trees for a set of 3 taxa (triplet) can be viewed as a mixture of

two distributions: one that generates branch lengths under ILS, and the other under introgression/speciation. Thus, the estimated mixing proportions (π_1 for ILS and π_2 for introgression/speciation; $\pi_1 + \pi_2 = 1$) of those distribution components show which fraction of the gene trees were generated through ILS or non-ILS processes. For a given triplet, QuIBL computes the proportion of gene trees that support the three alternative topologies. Then for every alternative topology QuIBL estimates mixing proportions along with other relevant parameters via Expectation-Maximization and computes Bayesian Information Criterion (BIC) scores for ILS-only and introgression models. For concordant topologies elevated values of π_2 are expected whereas for discordant ones π_2 values significantly greater than zero are indicative of introgression. To identify significant cases of introgression here we used a cutoff of $\Delta BIC < -30$ as in (Edelman et al., 2019). We ran QuIBL on every triplet individually under default parameters with the number of steps (the `numsteps` parameter) set to 50 and using *Anopheles gambiae* for triplet rooting; the branch length between *A. gambiae* and the triplet is not used for any of QuIBL's calculations.

We note that the DCT and BLT methods are potentially impacted by ancestral population structure: if the lineages leading to B and C were in subpopulations that were more likely to interbreed in the ancestral population, then the ((B, C), A) topology might be expected to be more prevalent than ((A, C), B), along with a shorter time back to the first coalescence. However, it is unclear how much of a concern ancestral population structure should be for this analysis, as it seems less likely that it would be a pair of lineages that diverged first (i.e., A and C or B and C) that interbred more frequently in the ancestral population instead of the two lineages that went on to be sister taxa (i.e., A and B). Moreover, we would not expect ancestral structure to impact the results of QuIBL, which should not be impacted by ancestral structure because this method searches for evidence of a mixture of coalescence times: one older time consistent with ILS and one time that is more recent than the split in the true species tree and that therefore cannot be explained by ancestral structure.

Phylogenetic networks: Introgression generates instances of reticulate evolution such that purely bifurcating trees cannot adequately represent evolutionary history; phylogenetic networks have been shown to provide a better fit to describe these patterns (Huson and Bryant, 2006; Solís-Lemus et al., 2016). We used PhyloNet (Than et al., 2008; Wen et al., 2018) to calculate

likelihood scores for networks generated by placing a single reticulation event (node) in an exhaustive manner, i.e., connecting all possible branch pairs within a clade. Because full likelihood calculations with PhyloNet can be prohibitively slow for large networks, for each of clades 1 through 9 we selected a subsample of 10 species in a manner that preserves the overall species tree topology. No subsampling was performed for clade 3 which has fewer than 10 species. Using these subsampled clade topologies, we formed all possible network topologies having a single reticulation node (with the exception of networks having reticulation nodes connecting sister taxa) (Supplementary Data). Because PhyloNet takes gene trees as input, for each clade we subsampled each gene tree to include only the subset of 10 species selected for the PhyloNet analysis (or all species in the case of clade 3); any gene trees missing at least one of these species were omitted from the analysis. Finally, we used the GalGTProb program (Yu et al., 2012) of the PhyloNet suite to obtain a likelihood score for each network topology for each clade. We report networks with the highest likelihood scores (Supplementary Data).

Acknowledgements: We thank M. Hahn, M. Turelli, and A. Yassin for helpful feedback on a previous draft. AS was supported by the NIH under award no. R00HG008696 and DRS was supported by the NIH under award nos. R00HG008696 and R35GM138286. BYK was supported by the NIH under award no. F32GM135998. JW was supported by the NIH under award no. K01DK119582. DP, ERRD, DRM and AAC were supported by NSF Dimensions of Biodiversity award 1737752 and NIH award R01GM121750. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability:

Supplementary file 1: Table S1 Tree calibration schemes for divergence time estimation analysis.

Supplementary file 2: Table S2 Full list of samples, strain information, accessions and associated publications.

Supplementary file 3: Figures S1-S10

Supplementary Data: dx.doi.org/10.6084/m9.figshare.13264697

Data analysis pipeline and code: https://github.com/SchriderLab/drosophila_phylogeny

Whole genome sequencing data generated for this study are available on NCBI (BioProject PRJNA675888, BioProject PRJNA593822, and BioProject PRJNA611543).

Competing interests: The authors declare that there is no conflict of interest

References:

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun* **10**:934. doi:10.1038/s41467-019-08822-w

Anderson TM, vonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, Bustamante CD, Ostrander EA, Tang H, Wayne RK, Barsh GS. 2009. Molecular and evolutionary history of melanism in North American gray wolves. *Science* **323**:1339–1343. doi:10.1126/science.1165448

Anisimova M, editor. 2012. Evolutionary Genomics: Statistical and Computational Methods, Volume 2, Methods in Molecular Biology. Springer Science+Business Media, LLC. doi:10.1007/978-1-61779-585-5

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O. 2011. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Syst Biol* **60**:685–699. doi:10.1093/sysbio/syr041

Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. *Curr Opin Genet Dev* **17**:513–518. doi:10.1016/j.gde.2007.09.001

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**:455–477. doi:10.1089/cmb.2012.0021

Bininda-Emonds OR. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* **6**:156. doi:10.1186/1471-2105-6-156

Brand CL, Kingan SB, Wu L, Garrigan D. 2013. A Selective Sweep across Species Boundaries in *Drosophila*. *Mol Biol Evol* **30**:2177–2186. doi:10.1093/molbev/mst123

Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, Hu S, Huang S, Zhang H, Zheng Z, Song W, Ma Z, Ma Y, Dang R, Zhang Z, Xu L, Jia Y, Liu S, Yue X, Deng W, Zhang X, Sun Z, Lan X, Han J, Chen H, Bradley DG, Jiang Y, Lei C. 2018. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat Commun* **9**:2337. doi:10.1038/s41467-018-04737-0

Cooper BS, Sedghifar A, Nash WT, Comeault AA, Matute DR. 2018. A Maladaptive Combination of Traits Contributes to the Maintenance of a *Drosophila* Hybrid Zone. *Curr Biol* **28**:2940-2947.e6. doi:10.1016/j.cub.2018.07.005

Coyne JA, Orr HA. 1997. "Patterns of speciation in *Drosophila*" revisited. *Evolution* **51**:295–303.

Coyne JA, Orr HA. 1989. Patterns of speciation in *Drosophila*. *Evolution* **43**:362–381.

Dasmahapatra KK. 2012. *Heliconius* genome supplementary information. *Nature*. doi:10.1038/nature11041

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin AV,

Hughes DST, Ferguson LC, Martin SH, Salazar C, Lewis JJ, Adler S, Ahn S-J, Baker D a., Baxter SW, Chamberlain NL, Chauhan R, Counterman B a., Dalmay T, Gilbert LE, Gordon K, Heckel DG, Hines HM, Hoff KJ, Holland PWH, Jacquin-Joly E, Jiggins FM, Jones RT, Kapan DD, Kersey P, Lamas G, Lawson D, Mapleson D, Maroja LS, Martin A, Moxon S, Palmer WJ, Papa R, Papanicolaou A, Pauchet Y, Ray D a., Rosser N, Salzberg SL, Supple M a., Surridge A, Tenger-Trolander A, Vogel H, Wilkinson P a., Wilson D, Yorke J a., Yuan F, Balmuth AL, Eland C, Gharbi K, Thomson M, Gibbs R a., Han Y, Jayaseelan JC, Kovar C, Mathew T, Muzny DM, Ongeri F, Pu L-L, Qu J, Thornton RL, Worley KC, Wu Y-Q, Linares M, Blaxter ML, Ffrench-Constant RH, Joron M, Kronforst MR, Mullen SP, Reed RD, Scherer SE, Richards S, Mallet J, Owen McMillan W, Jiggins CD. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**:94–98. doi:10.1038/nature11041

dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol Biol Evol* **28**:2161–2172. doi:10.1093/molbev/msr045

Durvasula A, Sankararaman S. 2020. Recovering signals of ghost archaic introgression in African populations. *Sci Adv* **6**:eaax5097. doi:10.1126/sciadv.aax5097

Dyer KA, Bewick ER, White BE, Bray MJ, Humphreys DP. 2018. Fine-scale geographic patterns of gene flow and reproductive character displacement in *Drosophila subquinaria* and *Drosophila recens*. *Mol Ecol* **27**:3655–3670. doi:https://doi.org/10.1111/mec.14825

Eberlein C, Hénault M, Fijarczyk A, Charron G, Bouvier M, Kohn LM, Anderson JB, Landry CR. 2019. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat Commun* **10**:923. doi:10.1038/s41467-019-08809-7

Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Belleghem SMV, Patterson N, Neafsey DE, Challis R, Kumar S, Moreira GRP, Salazar C, Chouteau M, Counterman BA, Papa R, Blaxter M, Reed RD, Dasmahapatra KK, Kronforst M, Joron M, Jiggins CD, McMillan WO, Palma FD, Blumberg AJ, Wakeley J, Jaffe D, Mallet J. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* **366**:594–599. doi:10.1126/science.aaw2090

Fishman L, Sweigart AL. 2018. When Two Rights Make a Wrong: The Evolutionary Genetics of Plant Hybrid Incompatibilities. *Annu Rev Plant Biol* **69**:707–731. doi:10.1146/annurev-arplant-042817-040113

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu Y-C, Smith HA, Love RR, Lawniczak MK, Slotman MA, Emrich SJ, Hahn MW, Besansky NJ. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**. doi:10.1126/science.1258524

Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC. 2012. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res* **22**:1499–511. doi:10.1101/gr.130922.111

Geraldes A, Ferrand N, Nachman MW. 2006. Contrasting Patterns of Introgression at X-Linked Loci Across the Hybrid Zone Between Subspecies of the European Rabbit (Oryctolagus cuniculus). *Genetics* **173**:919–933. doi:10.1534/genetics.105.054106

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof

A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Ž, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. 2010. A Draft Sequence of the Neandertal Genome. *Science* **328**:710–722. doi:10.1126/science.1188021

Hahn MW, Hibbins MS. 2019. A Three-Sample Test for Introgression. *Mol Biol Evol* **36**:2878–2882. doi:10.1093/molbev/msz178

Hamlin JAP, Hibbins MS, Moyle LC. 2020. Assessing biological factors affecting postspeciation introgression. *Evol Lett* **4**:137–154. doi:10.1002/evl3.159

Harris K, Nielsen R. 2016. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**:881–891. doi:10.1534/genetics.116.186890

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol* **22**:4606–4618. doi:10.1111/mec.12415

Huson DH, Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol* **23**:254–267. doi:10.1093/molbev/msj030

Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of Reticulate Networks from Gene Trees In: Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner PA, Waterman M, editors. Research in Computational Molecular Biology, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. pp. 233–249. doi:10.1007/11415770_18

Izumitani HF, Kusaka Y, Koshikawa S, Toda MJ, Katoh T. 2016. Phylogeography of the Subgenus *Drosophila* (Diptera: Drosophilidae): Evolutionary History of Faunal Divergence between the Old and the New Worlds. *PLOS ONE* **11**:e0160051. doi:10.1371/journal.pone.0160051

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res* **27**:768–777. doi:10.1101/gr.214346.116

Jones MR, Mills LS, Alves PC, Callahan CM, Alves JM, Lafferty DJR, Jiggins FM, Jensen JD, Melo-Ferreira J, Good JM. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science* **360**:1355–1358. doi:10.1126/science.aar5273

Juric I, Aeschbacher S, Coop G. 2015. The Strength of Selection Against Neanderthal Introgression. *PLoS Genet* **12**:e1006340. doi:doi:10.1371/journal.pgen.1006340

Kang L, Garner HR, Price DK, Michalak P. 2017. A Test for Gene Flow among Sympatric and Allopatric Hawaiian Picture-Winged *Drosophila*. *J Mol Evol* **84**:259–266. doi:10.1007/s00239-017-9795-7

Kao JY, Lymer S, Hwang SH, Sung A, Nuzhdin SV. 2015. Postmating reproductive barriers contribute to the incipient sexual isolation of the United States and Caribbean *Drosophila melanogaster*. *Ecol Evol* **5**:3171–3182. doi:https://doi.org/10.1002/ece3.1596

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**:3059–3066. doi:10.1093/nar/gkf436

Katoh T, Tamura K, Aotsuka T. 2000. Phylogenetic Position of the Subgenus *Lordiphosa* of the Genus Drosophila (Diptera: Drosophilidae) Inferred from Alcohol Dehydrogenase (Adh) Gene Sequences. *J Mol Evol* **51**:122–130. doi:10.1007/s002390010072

Kim BY, Huber CD, Lohmueller KE. 2018. Deleterious variation shapes the genomic landscape

of introgression. *PLOS Genet* **14**:e1007741. doi:10.1371/journal.pgen.1007741

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, ..., Matute DR, Petrov DA. 2020. Highly contiguous assemblies of 101 drosophilid genomes. *In prep*.

Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, Kapan DD, Mullen SP. 2013. Hybridization Reveals the Evolving Genomic Architecture of Speciation. *Cell Rep* 666–677.

Lachaise D, Harry M, Solignac M, Lemeunier F, Bénassi V, Cariou ML. 2000. Evolutionary novelties in islands: *Drosophila santomea*, a new melanogaster sister species from São Tomé. *Proc R Soc Lond B* **267**:1487–1495. doi:10.1098/rspb.2000.1169

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, Grant BR, Grant PR, Webster MT, Andersson L. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**:371.

Leducq J-B, Nielly-Thibault L, Charron G, Eberlein C, Verta J-P, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat Microbiol* **1**:15003. doi:10.1038/nmicrobiol.2015.3

Li G, Davis BW, Eizirik E, Murphy WJ. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res* **26**:1–11. doi:10.1101/gr.186668.114

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

Magnacca KN, Price DK. 2015. Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing *Drosophila* (Diptera: Drosophilidae). *Mol Phylogenet Evol* **92**:226–242. doi:10.1016/j.ympev.2015.06.014

Maheshwari S, Barbash DA. 2011. The Genetics of Hybrid Incompatibilities. *Annu Rev Genet* **45**:331–355. doi:10.1146/annurev-genet-110410-132514

Mai D, Nalley MJ, Bachtrog D. 2020. Patterns of Genomic Differentiation in the *Drosophila nasuta* Species Complex. *Mol Biol Evol* **37**:208–220. doi:10.1093/molbev/msz215

Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol* **2**:1940–1955. doi:10.1038/s41559-018-0717-x

Marques DA, Meier JI, Seehausen O. 2019. A Combinatorial View on Speciation and Adaptive Radiation. *Trends Ecol Evol* **34**:531–544. doi:10.1016/j.tree.2019.02.008

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biol* **17**:e2006288. doi:10.1371/journal.pbio.2006288

Matute DR. 2010. Reinforcement of gametic isolation in *Drosophila*. *PLoS Biol* **8**:e1000341. doi:10.1371/journal.pbio.1000341

Matute DR, Ayroles JF. 2014. Hybridization occurs between *Drosophila simulans* and *D. sechellia* in the Seychelles archipelago. *J Evol Biol* **27**:1057–68. doi:10.1111/jeb.12391

Matute DR, Comeault AA, Earley E, Serrato-Capuchina A, Peede D, Monroy-Eklund A, Huang W, Jones CD, Mackay TFC, Coyne JA. 2020. Rapid and Predictable Evolution of Admixed Populations Between Two *Drosophila* Species Pairs. *Genetics* **214**:211–230. doi:10.1534/genetics.119.302685

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun* **8**:14363. doi:10.1038/ncomms14363

Meiklejohn CD, Landeen EL, Gordon KE, Rzatkiewicz T, Kingan SB, Geneva AJ, Vedanayagam JP, Muirhead CA, Garrigan D, Stern DL, Presgraves DC. 2018. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. *eLife* **7**:e35468. doi:10.7554/eLife.35468

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**:1188–1195. doi:10.1093/molbev/mst024

Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. 2020. The genetic consequences of hybridization. *ArXiv201204077 Q-Bio*.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**:268–274. doi:10.1093/molbev/msu300

Nosil P, Schluter D. 2011. The genes underlying the process of speciation. *Trends Ecol Evol* **26**:160–167. doi:10.1016/j.tree.2011.01.001

Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating Divergence Dates and Substitution Rates in the *Drosophila* Phylogeny. *Mol Biol Evol* **29**:3459–3473. doi:10.1093/molbev/mss150

O'Grady PM, DeSalle R. 2018. Phylogeny of the Genus *Drosophila*. *Genetics* **209**:1–25. doi:10.1534/genetics.117.300583

Ottenburghs J. 2020. Ghost Introgression: Spooky Gene Flow in the Distant Past. *BioEssays* **42**:2000012. doi:https://doi.org/10.1002/bies.202000012

Payseur BA, Krenz JG, Nachman MW. 2004. Differential Patterns of Introgression Across the X Chromosome in a Hybrid Zone Between Two Species of House Mice. *Evolution* **58**:2064–2078. doi:https://doi.org/10.1111/j.0014-3820.2004.tb00490.x

Pease JB, Brown JW, Walker JF, Hinchliff CE, Smith SA. 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am J Bot* **105**:385–403. doi:10.1002/ajb2.1016

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLOS Biol* **14**:e1002379. doi:10.1371/journal.pbio.1002379

Platt RN, McDew-White M, Le Clec'h W, Chevalier FD, Allan F, Emery AM, Garba A, Hamidou AA, Ame SM, Webster JP, Rollinson D, Webster BL, Anderson TJC. 2019. Ancient Hybridization and Adaptive Introgression of an Invadolysin Gene in Schistosome Parasites. *Mol Biol Evol* **36**:2127–2142. doi:10.1093/molbev/msz154

Price JP, Clague DA. 2002. How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc R Soc B Biol Sci* **269**:2429–2435. doi:10.1098/rspb.2002.2175

Puttick MN. 2019. MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* **35**:5321–5322. doi:10.1093/bioinformatics/btz554

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet* **16**:359–371. doi:10.1038/nrg3936

Reis MD, Yang Z. 2013. The unbearable uncertainty of Bayesian divergence time estimation. *J Syst Evol* **51**:30–43. doi:https://doi.org/10.1111/j.1759-6831.2012.00236.x

Rhymer JM, Simberloff D. 1996. Extinction by Hybridization and Introgression. *Annu Rev Ecol Syst* **27**:83–109. doi:10.1146/annurev.ecolsys.27.1.83

Richards EJ, Martin CH. 2017. Adaptive introgression from distant Caribbean islands contributed to the diversification of a microendemic adaptive radiation of trophic specialist pupfishes. *PLOS Genet* **13**:e1006919. doi:10.1371/journal.pgen.1006919

Russo CAM, Mello B, Frazão A, Voloch CM. 2013. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). *Zool J Linn Soc* **169**:765–775. doi:10.1111/zoj.12062

Sachdeva H, Barton NH. 2018. Introgression of a Block of Genome Under Infinitesimal Selection. *Genetics* **209**:1279–1303. doi:10.1534/genetics.118.301018

Sawamura K, Sato H, Lee C-Y, Kamimura Y, Matsuda M. 2016. A Natural Population Derived from Species Hybridization in the *Drosophila ananassae* Species Complex on Penang Island, Malaysia. *Zoolog Sci* **33**:467–475. doi:10.2108/zs160038

Sayyari E, Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Mol Biol Evol* **33**:1654–1668. doi:10.1093/molbev/msw079

Schrider DR, Ayroles J, Matute DR, Kern AD. 2018. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLOS Genet* **14**:e1007341. doi:10.1371/journal.pgen.1007341

Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG, Przeworski M. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**:656–660. doi:10.1126/science.aar3684

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness In: Kollmar M, editor. Gene Prediction: Methods and Protocols, Methods in Molecular Biology. New York, NY: Springer. pp. 227–245. doi:10.1007/978-1-4939-9173-0_14

Serrato-Capuchina A, Schwochert TD, Zhang S, Roy B, Peede D, Koppelman C, Matute DR. 2020. Pure species discriminate against hybrids in the *Drosophila melanogaster* species subgroup. *bioRxiv* 2020.07.22.214924. doi:10.1101/2020.07.22.214924

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212. doi:10.1093/bioinformatics/btv351

Solís-Lemus C, Yang M, Ané C. 2016. Inconsistency of Species Tree Methods under Gene Flow. *Syst Biol* **65**:843–851. doi:10.1093/sysbio/syw030

Storchová R, Reif J, Nachman MW. 2010. Female Heterogamety and Speciation: Reduced Introgression of the Z Chromosome Between Two Species of Nightingales. *Evolution* **64**:456–471. doi:https://doi.org/10.1111/j.1558-5646.2009.00841.x

Suarez-Gonzalez A, Lexer C, Cronk QCB. 2018. Adaptive introgression: a plant perspective. *Biol Lett* **14**:20170688. doi:10.1098/rsbl.2017.0688

Svardal H, Quah FX, Malinsky M, Ngatunga BP, Miska EA, Salzburger W, Genner MJ, Turner GF, Durbin R. 2020. Ancestral Hybridization Facilitated Species Diversification in the Lake Malawi Cichlid Fish Adaptive Radiation. *Mol Biol Evol* **37**:1100–1113. doi:10.1093/molbev/msz294

Tamura K, Subramanian S, Kumar S. 2004. Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. *Mol Biol Evol* **21**:36–44. doi:10.1093/molbev/msg236

Taylor SA, Larson EL. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat Ecol Evol* **3**:170–177. doi:10.1038/s41559-

018-0777-y

Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9**:322. doi:10.1186/1471-2105-9-322

Throckmorton LH. 1975. The phylogeny, ecology, and geography of *Drosophila* In: King RC, editor. Handbook of Genetics, Vol 3. New York: Plenum Publishing Corp. pp. 421–469.

Turelli M, Lipkowitz JR, Brandvain Y. 2014. On the Coyne and Orr-Igin of Species: Effects of Intrinsic Postzygotic Isolation, Ecological Differentiation, X Chromosome Size, and Sympatry on Drosophila Speciation. *Evolution* **68**:1176–1187. doi:https://doi.org/10.1111/evo.12330

Turissini DA, Comeault AA, Liu G, Lee YCG, Matute DR. 2017. The ability of *Drosophila* hybrids to locate food declines with parental divergence. *Evolution* **71**:960–973. doi:10.1111/evo.13180

Turissini DA, Matute DR. 2017. Fine scale mapping of genomic introgressions within the Drosophila yakuba clade. *PLOS Genet* **13**:e1006971. doi:10.1371/journal.pgen.1006971

Turissini DA, McGirr JA, Patel SS, David JR, Matute DR. 2018. The Rate of Evolution of Postmating-Prezygotic Reproductive Isolation in *Drosophila*. *Mol Biol Evol* **35**:312–334. doi:10.1093/molbev/msx271

Tusso S, Nieuwenhuis BPS, Sedlazeck FJ, Davey JW, Jeffares DC, Wolf JBW. 2019. Ancestral Admixture Is the Main Determinant of Global Biodiversity in Fission Yeast. *Mol Biol Evol* **36**:1975–1989. doi:10.1093/molbev/msz126

Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny DM, Hibbins MS, Williamson RJ, Gibbs RA, Worley KC, Rogers J, Hahn MW. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLOS Biol* **18**:e3000954. doi:10.1371/journal.pbio.3000954

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. doi:10.1093/molbev/msx319

Wen D, Yu Y, Zhu J, Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet. *Syst Biol* **67**:735–740. doi:10.1093/sysbio/syy015

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Yang Z, Rannala B. 2006. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol Biol Evol* **23**:212–226. doi:10.1093/molbev/msj024

Yassin A. 2013. Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. *Syst Entomol* **38**:349–364. doi:10.1111/j.1365-3113.2012.00665.x

Yu Y, Degnan JH, Nakhleh L. 2012. The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection. *PLOS Genet* **8**:e1002660. doi:10.1371/journal.pgen.1002660