Evolution Letters, 2025, pp. 1–9

doi: DOI HERE

Genomic analyses in *Drosophila* do not support the classic allopatric model of speciation.

Leeban H. Yusuf,[1,*] Dominik R. Laetsch,[2] Konrad Lohse[2,*] and Michael G. Ritchie[1,*]

[1]Centre for Biological Diversity, School of Biology, University of St Andrews, St Andrews KY16 9TH, Scotland and [2]Institute of Ecology and Evolution, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, EH9 3FL, Scotland.
[*]Corresponding author. ly36@st-andrews.ac.uk, konrad.lohse@ed.ac.uk, mgr@st-andrews.ac.uk
K.L.(Konrad Lohse) and M.G.R (Michael G. Ritchie) contributed equally to this work.

Abstract

The allopatric model of speciation has dominated our understanding of speciation biology and biogeography since the Modern Synthesis. It is uncontroversial because reproductive isolation may readily emerge as a by-product of evolutionary divergence during allopatry unopposed by gene flow. Recent genomic studies have found that gene flow between species is common, but whether allopatric speciation is common has rarely been systematically tested across a continuum of closely-related species. Here, we fit a range of demographic models of evolutionary divergence to whole-genome sequence data from 93 pairs of *Drosophila* species to infer speciation histories and levels of post-divergence gene flow. We find that speciation with gene flow is common, even between currently allopatric pairs of species. Estimates of historical gene flow are not predicted by current range overlap. Whilst evidence for secondary contact is generally limited, a few sympatric pairs showed strong support for a secondary contact model. Our analyses suggest that most speciation processes involve some long-term gene flow, perhaps due to repeated cycles of allopatry and contact, without requiring an extensive allopatric phase.

Key words: Speciation, Gene flow, Biogeography

The idea that speciation processes can be classified into distinct geographic modes (allopatric, parapatric, and sympatric) emerged during the Modern Synthesis and continues to be a central theme of evolutionary research [48, 15, 47]. Allopatric speciation is thought to be common and is often considered a null model of speciation [15, 51]. Allopatric speciation appears uncontroversial both empirically given the widespread prevalence of 'geographic' varieties and species (which influenced Darwin and Mayr), and conceptually due to the ease with which the absence of gene flow facilitates the build-up of reproductive isolation (RI) due to genetic drift, natural, or sexual selection. Much of the debate over other modes of speciation revolves around the appearance of different forms of RI in the face of potentially homogenising gene flow. If and when species come into secondary contact after an allopatric phase, natural selection to reduce competition (but see [3]) or to prevent maladaptive hybridization ('reinforcement') may lead to complete isolation or the formation of stable hybrid zones [1].

Comparisons between currently sympatric and allopatric species pairs have provided apparently axiomatic evidence for reinforcement. In particular, analyses that contrast estimates of pre- and postzygotic RI between currently allopatric and sympatric species pairs of *Drosophila* demonstrated convincingly that mate discrimination evolves more rapidly between sympatric species [13, 14]. Alongside other comparative surveys [63, 64, 52, 23, 58], these patterns were interpreted as evidence that incompatibilities which arose during an allopatric phase lead to reinforcement after secondary contact [45]. However, classifying closely related species based on their current geographic ranges assumes that these in part reflect the geographical context during the key phases of their speciation history [39, 5, 29]. It has been acknowledged that allopatry and sympatry are ends of a spectrum [20, 44, 9], and an increasing number of genomic studies suggest that historical and contemporary gene flow is more prevalent than the architects of the modern synthesis had envisaged [56, 18, 21, 53, 17]. Additionally, gene flow may diminish gradually without the need for extended periods of allopatry during divergence [51]. It is therefore, unclear whether most speciation processes involve a strictly allopatric phase, and how common allopatric speciation versus speciation with gene flow is.

Here, we take advantage of the wealth of genomic data available for one of the best-studied groups of species, *Drosophila*. Population genetic analyses of *Drosophila* species pairs have spurred the development of inference methods that model the joint effects of incomplete lineage sorting and gene flow on local genealogies. These methods integrate over all possible genealogies sampled from short blocks of sequence across the genome, enabling maximum likelihood estimation of contrasting, nested models of species divergence and gene flow [31, 40, 33]. For small samples it is possible to calculate likelihoods analytically [61, 38], which allows efficient inference of simple demographic histories even from

© The Author 2025. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

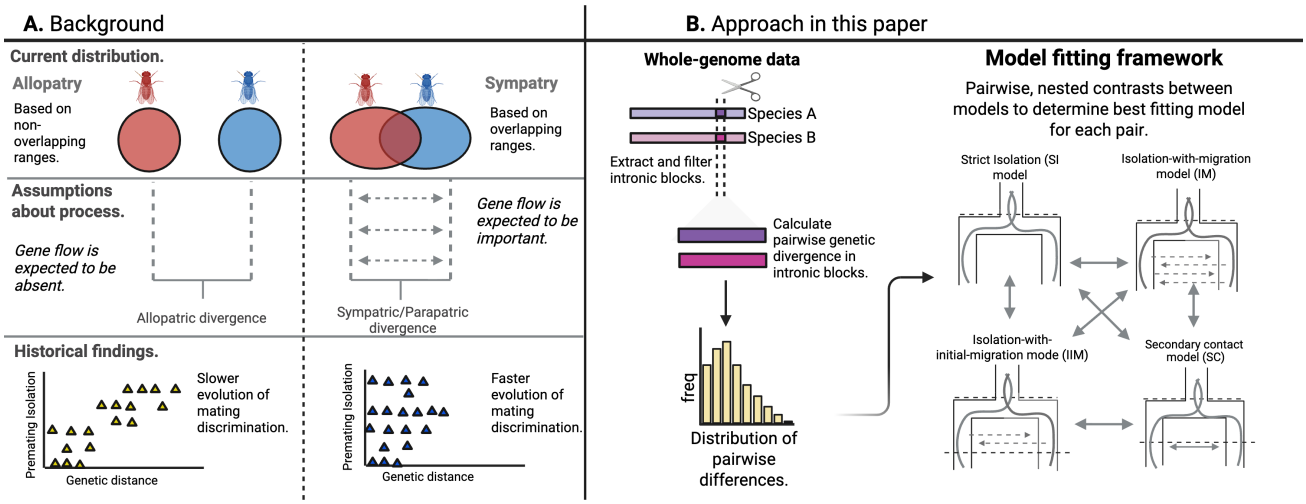


Fig. 1. A) Background and **B)** our approach to understanding biogeographic modes of speciation. **A)** Previous comparative surveys on the evolution of reproductive isolation categorized species pairs by geography. This assumes that current range overlap is informative about the underlying speciation process, e.g. sympatric pairs are more likely to have exchanged migrants than allopatric taxa. **B)** Our approach uses whole-genome data to test to what extent levels of long-term gene flow differ between taxon pairs with and without current range overlap. We summarized genome-wide divergence in short, intronic blocks in terms of the distributions of pairwise differences. We fitted demographic models of speciation representing strictly allopatric speciation and different models of speciation with gene flow, to each pair and assessed their relative support in a likelihood framework.

a single genome per species [61, 62]. While gene flow analyses in *Drosophila* have demonstrated its prevalence [56], we lack a systematic evaluation of the relationship between inferred historical demography and current range overlap in the genus.

We analyse the pairwise distribution of sequence differences (which is a function of the distribution of pairwise coalescence times) in putatively neutral, short intronic blocks in 93 *Drosophila* species pairs to estimate the relative support for allopatric species divergence and speciation with gene flow. Sampling a single intronic block per gene justifies the assumption of statistical independence between blocks and allows us to calculate support for alternative demographic models of species divergence and parameter estimates in a likelihood framework [61, 62]. Our focal species pairs are a subset of the species pairs originally studied by Coyne and Orr [13, 14] and include currently allopatric and sympatric species spanning a range of genomic divergence ($0.01 < d_{xy} < 0.05$) (Supplementary table 2, and Supplementary Fig. 1 and 2). While our choice of taxon pairs was determined by the availability of whole genome sequence data, we note that our focal taxa cover a wide taxonomic and geographic range within the *Drosophila* genus [19]. Furthermore, the classic result of Coyne and Orr [13, 14] (faster evolution of pre-zygotic isolation for sympatric pairs) holds for this subset (Supplementary Fig. 1). In fact, both premating and postzygotic RI are higher in sympatric compared to allopatric pairs (two-sample Wilcoxon tests: W = 638.5, p = 0.004 and W = 94.5, p = 0.009), a pattern already reported by [63, 46].

We fitted alternative demographic models of strict allopatric divergence versus divergence with gene flow and secondary contact [61, 62] (Fig. 1) to ask the following questions: (1) How much support for (historical or contemporary) gene flow is there and does this differ between currently sympatric or allopatric species pairs? (2) Do currently sympatric and allopatric pairs differ in their species divergence time and/or ancestral population size? (3) Finally, do sympatric pairs show greater support for secondary contact histories than allopatric pairs? We conclude that speciation involving a low rate of long-term gene flow is ubiquitous in *Drosophila*, and therefore that there is little support for allopatric speciation in this group.

Results

Speciation with gene flow is common across *Drosophila* pairs

To assess support for gene flow in each focal species pair of *Drosophila*, we compared four speciation models: (1) a strict isolation model (SI), where a pair of species diverge from an ancestral population of size N_e at some time T_0 and remain in strict allopatry, (2) an isolation-with-migration model (IM), where divergence occurred with symmetric migration at a constant rate $M = 2N_e m$ per generation from T_0 until the present and (3) a secondary contact model (SC) where divergence occurred without gene flow initially, but where the pair experiences an instantaneous burst of bidirectional gene flow that transfers a fraction f of lineages at time T_1 (see SI Appendix section 1.4 for details). The latter may reflect a 'reinforcement' scenario where gene flow only occurs transiently following secondary contact. We restrict analyses to these relatively simple models both because previous comparative surveys of speciation demography have considered the same scenarios [53] and because more complex models are unlikely to be identifiable from the pairwise distribution of differences. In particular, we also considered an isolation-with-initial-migration model (IIM) where gene flow occurred at the onset of divergence but stopped at time T_1. However, since only one taxon pair (*D. lummei* - *D. littoralis*) had significantly support for IIM, we concluded that this scenario is either not identifiable or not sufficiently supported by the data and so restricted comparison of demographic histories to SI, IM and SC models. Note also that our gene flow estimates are effective parameters which incorporate any selection against early generation hybrids (see Discussion).

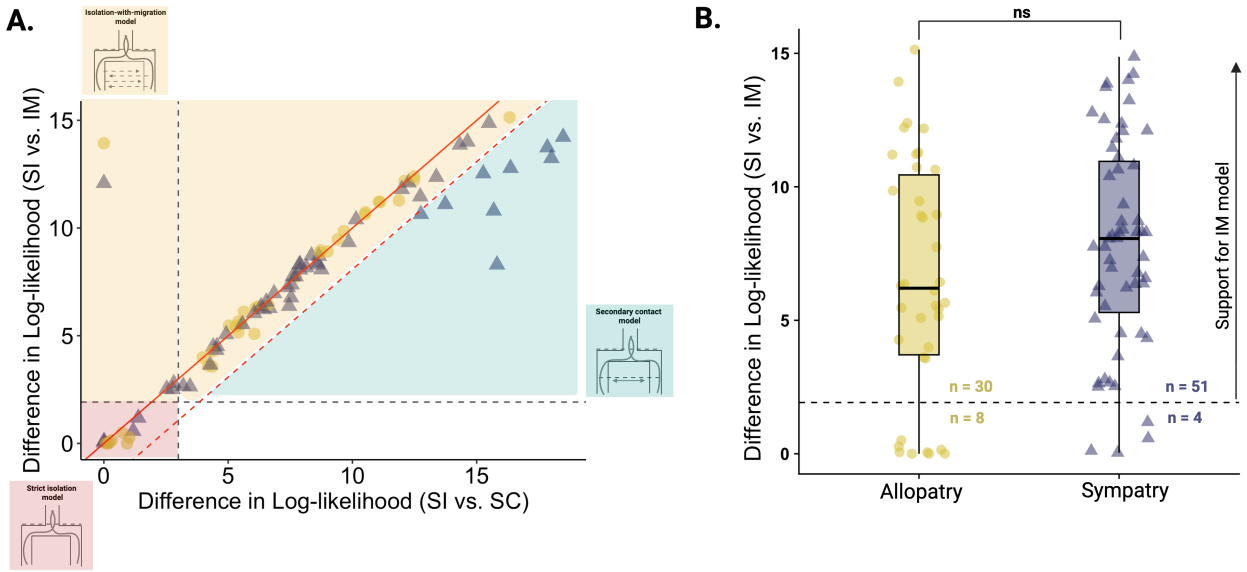


Fig. 2. Evidence for gene flow across allopatric and sympatric *Drosophila* pairs. This plot summarises which pairs best-fit the SC model (right bottom, teal) IM (top right, yellow) and SI (bottom corner, red). The vast majority of pairs fall into regions supporting gene flow. **A.** The difference in log-likelihood between strict isolation (SI) and isolation-with-migration (IM) histories against the difference in log-likelihood between strict isolation (SI) and secondary contact (SC) histories. Black-dashed lines are the critical value thresholds for the SI v.s. IM (horizontal line at $2\Delta \ln L > 3.84$, 1 d.f.) and SI v.s. SC (vertical line at $2\Delta \ln L > 5.99$, 2 d.f.). The red solid line is the line of equality (y=x), showing correlated support for gene flow models regardless of whether gene flow is continuous or an instantaneous event. Points to the right of the dashed red line represent pairs where support for the SC model exceeds (a critical value of $2\Delta \ln L > 3.84$) the support for an IM model, with the exception of a single pair which best fits an IIM history better than both the IM and SC models. Two pairs fit the IM model considerably better than the SC model (top left). Shaded areas denote best-fitting models in our comparative framework. **B.** The difference in relative support for an IM history, against a SI history, for currently sympatric and allopatric species pairs. The vertical dashed line indicates the critical value ($p < 0.05$). The numbers of species pairs that fit or do not fit an IM model significantly better than an SI model are shown on either side of the vertical line.

Assigning a best-fitting but minimally complex model to every species pair reveals overwhelming support for speciation histories that involve gene flow: out of the 93 *Drosophila* pairs, only 12 pairs best fit an SI model, i.e. neither IM nor SC give a significant improvement in model fit (Fig 2A), and four of these are sympatric. For the majority of species pairs (76) both IM and SC scenarios fit significantly better than a SI history and relative support for either model is highly correlated (Fig. 2A). Given that the SC model is more complex than the IM model (a total of four instead of three parameters), we only accepted it as the best and most parsimonious history when the relative support for it exceeded a critical value of $2\Delta \ln L > 3.841$. We also used this threshold (which is equivalent to assuming a χ^2 distribution with 1 d.f.) to compare nested models that differ by one parameter (IM vs SI and IIM vs IM). Given these model selection criteria, we find that an SC history is the best supported model for eight pairs and accept the IM model as the best-fitting history for 72 pairs (Fig. 2A).

No difference in support for historic gene flow between sympatric and allopatric *Drosophila* pairs

If current range is indicative of the mode of speciation, we would expect allopatric pairs to best fit a history of strict isolation (SI) with no significant improvement in fit when including gene flow. In contrast, we may expect currently sympatric pairs to fit an isolation-with-migration (IM) model better than an SI model. Alternatively, if sympatry is recent (and associated with secondary gene flow), we may expect sympatric pairs to be associated with histories of secondary contact (SC) ('recent sympatry' hypothesis).

We find no difference in relative support (as measured by $\Delta \ln L$) for an IM model over an SI model between allopatric and sympatric pairs (two-sample Wilcoxon test: W = 867, p = 0.165) (Fig. 2A). However, we do find that the 12 pairs that do not show support for any gene flow scenario, i.e. that best fit an SI history, include more allopatric than sympatric pairs (one-tailed Fisher exact test: p = 0.064). Interestingly, these SI pairs almost entirely belong to the *Drosophila* nasuta group, including the species pair *D. albomicans* and *D. nasuta*, which in previous analyses also did not show evidence for post-divergence gene flow [4]. All eight pairs that best fit a secondary contact (SC) model are currently sympatric species, consistent with our 'recent-sympatry' hypothesis (one-tailed Fisher exact test: p = 0.019) (Fig. 2A)

Current range overlap poorly predicts the demography of speciation

To test whether historical gene flow differs between allopatric and sympatric pairs, we converted estimates of the scaled effective rate of gene flow (i.e. the number of migrants per generation M) under the IM model, into a per lineage probability of gene flow: e.g. given an estimated duration of gene flow between species T, the probability that the ancestry of an individual haplotype at a random position in the genome is affected by migration is $1-e^{-MT_0}$. This conversion allows for a direct comparison between estimates of continuous (IM) and discrete (SC) gene flow. We find a lower long-term probability of gene flow for allopatric pairs (mean probability of gene flow: 66%) compared to sympatric pairs (mean probability of gene flow: 74%). When assuming the best-fitting gene flow model

for each pair, this difference is not significant (two-sample Wilcoxon test: W = 934, p = 0.387)(Fig. 3A). Hence, there is evidence of less gene flow between allopatric species, though the extent is still high. Note that the IM model, although accepted as the best supported model for 72 taxon pairs yielded nonsensical parameter estimates in 22 pairs, i.e. we obtained arbitrarily high estimates of either gene flow (M) or divergence time T_0. This suggests either poor model fit and/or parameter non-identifiability (see Discussion) and we have removed these estimates from subsequent analyses. Excluding the 22 pairs with non-identifiable migration parameters did not change the lack of difference in long-term gene flow probability between sympatric and allopatric pairs (two-sample Wilcoxon test: W = 603, p = 0.847).

Additionally, we find that currently allopatric and sympatric pairs do not differ in the onset of divergence (measured in generations), but do differ in ancestral effective population size (Fig. 3B and C). While sympatric pairs are younger on average than allopatric pairs, this difference is not significant (two-sample Wilcoxon test: W = 1219, p = 0.175) (Fig. 3B). However, sympatric pairs have larger ancestral effective population sizes than allopatric pairs (two-sample Wilcoxon test: W = 768, p = 0.03). Note that this difference is not significant when we exclude the 22 species pairs with non-identifiable parameters (two-sample Wilcoxon test: W = 458, p = 0.06).

Secondary contact pairs do not have greater pre-mating isolation

Reinforcement is frequently invoked to explain enhanced mating discrimination in sympatry relative to allopatry [49, 15]. We asked whether the eight pairs that fit a secondary contact (SC model) substantially better than an isolation-with-migration (IM) model ($2\Delta \ln L > 3.841$) show greater premating isolation, than pairs for which IM is the best fitting model ($\Delta \ln L > 0$; n = 72). We find no significant difference in premating isolation for SC pairs (mean pre-mating isolation:: 0.90) compared to pairs that best fit an IM model (mean pre-mating isolation: 0.88) (two-sample Wilcoxon test: W = 294.5, p = 0.816).

Discussion

Allopatric speciation is considered the most common mode of speciation because the evolution of RI is unhindered once gene flow has ceased completely [15]. Analysing genomic data for 93 *Drosophila* species pairs in a hierarchical modelling framework, we find that most pairs show evidence for post-divergence gene flow. Perhaps surprisingly, while we find some evidence for increased gene flow between currently sympatric pairs, our main finding is that some level of post-divergence gene flow is common, even between currently allopatric pairs. While sympatric pairs show overall greater support for secondary contact histories and very recent gene flow as expected, there is little or no difference in statistical support for histories involving gene flow between pairs that are currently sympatric or allopatric.

These results have implications for interpreting classic comparative surveys of speciation in *Drosophila* and other taxa. Mating discrimination has been shown to evolve more rapidly in currently sympatric *Drosophila* species pairs relative to allopatric pairs, and this observation has been interpreted as support for reinforcement [15, 45, 54, 50]. The fact that we find significantly greater support for secondary contact histories for sympatric compared to allopatric species pairs is compatible with a role of reinforcement in speciation and shows that our analyses are extracting genuine signals of gene flow. In order to understand which aspects of the data allow us to distinguish between SC and IM models, it is helpful to inspect their absolute fit to the observed distribution of pairwise differences: taxon pairs that best fit an SC history show a characteristic excess of monomorphic blocks ($S = 0$) in the data (Supplementary Figs 13-15). While this high frequency of monomorphic blocks can be explained by a recent burst of admixture upon secondary contact, it is incompatible with histories of continuous gene flow (i.e. the IM model).

Limits and robustness of demographic inference

Our core finding that overall levels of historic gene flow differ little between currently sympatric and allopatric *Drosophila* pairs is perhaps surprising. It is important to stress that our coalescence-based inference of demographic history is based on minimal sampling of a single haplotype per species and is therefore limited to a small set of simplistic models. The fact that IM and SC models, which assume symmetric migration and ignore heterogeneity in N_e, already provide a good absolute fit to the observed S distributions suggests that more realistic models that include multiple phases of gene flow or account for asymmetry are not identifiable. Future population genomic analyses that include data on intraspecific diversity will undoubtedly be able to fit more realistic demographic models of species divergence.

There are, however, fundamental limits to demographic inference that do not depend on the sample sizes and data summaries used: estimates of gene flow and ancestral N_e between taxon pairs are potentially confounded by ghost admixture into one or both focal taxa from a third taxon [6]. Furthermore, periods of high gene flow (or low N_e) erase the genomic footprints of older demographic events and inference is limited to long-term rates of gene flow that are sufficiently low ($M < 10$). Importantly, this means that periods of gene flow that are high enough for reinforcement selection to act may be indistinguishable from panmixia.

Given the simplicity of our data summary, an important question is to what extent our finding of pervasive historic gene flow in *Drosophila* hinges on the ability to accurately estimate the frequency of monomorphic blocks. This may be difficult for two reasons: firstly, intronic sequence may be incompletely annotated in some taxa and including conserved regulatory sequence in introns will inflate the frequency of monomorphic blocks. Secondly, many of the sequence data we analyse are from isofemale lines with a history of lab culture and it is perhaps feasible that in some pairs recent admixture may have occurred as a result of lab contamination. To assess the dependence of our results on the frequency of monomorphic blocks, we estimated the fit of SI, IM and SC models for all pairs when monomorphic blocks are excluded. This involves conditioning on only observing blocks that differ between species in a pair by maximizing the conditioned log likelihood $\ln \frac{P[S=k]}{1-P[S=0]}$. We find that while conditioning on $S > 0$ – unsurprisingly – decreases support for IM and SC histories (relative to SI) overall (see Supplementary Fig 13-15; SI model: 15 (allopatric) and 24 (sympatric) pairs; IM model: 23 (allopatric) and 31 (sympatric) pairs; SC model: 0; IIM model: 0), our main qualitative result of a signature of gene flow in most taxa (n = 54) without any systematic difference

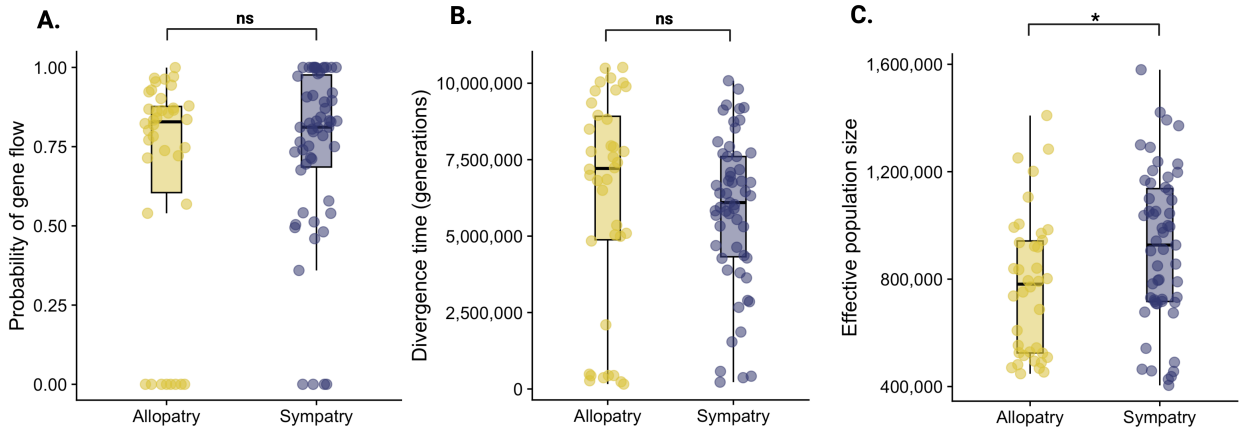


Fig. 3. Sympatric and allopatric pairs show similar divergence times, effective population size and levels of long-term gene flow. Current range overlap does not reflect inferred historical demography. **A.** We find no difference in long-term rates of gene flow between allopatric and sympatric pairs. Gene flow was calculated for each pair using parameter estimates from the best-fitting model. **B.** and **C.** We find no difference in scaled divergence time, but we find a difference in effective population size (the best-fitting model for each pair) between allopatric and sympatric pairs. 'ns' indicates no significant difference and '*' indicates $p < 0.05$ via Mann-Whitney U test.

between sympatric and allopatric pairs is unaffected ($\Delta \ln L$ between SI v.s. IM models; two-sample Wilcoxon test: $W = 1242$, $p = 0.124$).

Finally, a power analysis based on data simulated under plausible parameters for *Drosophila* shows that the overall greater support for speciation histories involving a low level of gene flow is unlikely to reflect biases that arise from the fact that our multilocus inference framework ignores recombination within blocks (see simulation results in Supplementary section 1.5 and Supplementary Fig. 8). Additionally, when we use a simple phylogenetic correction to account for the non-independence of some species pairs, we still find consistent evidence of gene flow across nodes in the *Drosophila* phylogeny [2](Supplementary section 1.6; Supplementary Fig. 11 and 12). Thus, there is a still pervasive signal of post-divergence gene flow in *Drosophila* pairs, even between currently allopatric pairs, and across all of our sensitivity analyses.

Speciation with gene flow is the rule not the exception

Our results strongly imply that speciation does not require an extended allopatric phase to allow the build-up of RI. Instead, a low level of gene flow at all stages of speciation appears to be relatively common in *Drosophila*. Numerous other studies have inferred gene flow between 'good' species without explicitly testing if this is greater in sympatric species [41, 65, 56, 53]. It appears that speciation in the face of gene flow is common and the traditional classification into distinct geographic modes of speciation, with allopatry taken as the default, is outdated [44, 9, 20, 42]. The importance of 'strict isolation' to the Biological Species Concept has diminished greatly over the last few decades, partly as a result of genomic analyses of introgression. Moreover, it is generally accepted that speciation is a continuous process [55, 43] and that complete RI is not necessary between species. Instead, our results suggest that speciation more often involves extended periods of genetic exchange and incomplete RI rather than an abrupt and complete cessation of gene flow. Of course, this widens the 'grey zone' where species barriers remain permeable long after species divergence has become irreversible [43, 7]. Genomic analyses of speciation with gene flow clearly demonstrate that levels of gene flow can vary widely across the genome [34, 22]. The extent to which this variation reflects clustered genetic architectures for barriers to gene flow that have arisen as a consequence of speciation with gene flow or simply pre-existing variation in recombination, perhaps owing to structural rearrangements such as inversions, is an open question. Nevertheless, it is clear that species barriers can build up and reach tipping points in the face of on-going gene flow without any extended period of strict allopatry.

However, it is important to emphasise that our suggestion that a strict allopatric model of speciation is probably uncommon in *Drosophila* does not suggest that allopatry is unimportant or plays no role. Studies of phylogeography, ancient DNA and niche modelling recognise that species ranges may be dynamic over timescales that are short relative to speciation processes. Within the limits of our modeling approach we have demonstrated that some amount of long term gene flow during species divergence is almost ubiquitous. However, estimating the timing and duration of individual episodes of gene flow is arguably a much more difficult task. The Pleistocene climate history has been dominated by cycles of ice ages and warmer interglacials. Thus, species both in the temperate zones [30, 28] and the tropics [26, 11] have undergone repeated periods of allopatry and secondary contact and gene flow. Such complex histories would be challenging to model explicitly (and are impossible to infer from the distribution of pairwise differences). By contrast, the IM model assumes a single long term (over the scale of N_e generations) rate of gene flow. This parameter therefore reflects a long term average of genetic exchange over many Pleistocene cycles of range shifts.

Another important consideration is how applicable our results are to other taxa. *Drosophila* is highly vagile, flighted and long distance dispersal has been demonstrated in some species [12]. However, many *Drosophila* species have specialised niches, so successful dispersal requires either movement of host plants, commensalism or switches to novel hosts. One may expect larger, less vagile, organisms to show less dynamic range

changes. However, phylogeographic patterns of diversity that imply rapid range shift in response to glacial cycles have been recognised in vastly diverse taxa, including plants and larger vertebrates [30]. We therefore see no reason to suppose that our result that current range overlap is essentially uncorrelated with historic gene flow during species divergence is limited to *Drosophila*.

We stress that demographic analyses of larger genomic datasets (preferably from wild-caught samples and known areas of range overlap) that make use of intraspecific variation will allow to fit more realistic models of divergence and may help to answer new questions. For example, rather than assuming that gene flow during species divergence is symmetric – as we have done here – it would be fascinating to test which traits and population genetic processes correlate with the direction of gene flow. There is clearly also a need for similar comparative analyses of speciation histories for taxa that have richer geographic range and life history information than *Drosophila* to contextualise the evolution of RI and the extent, timing and direction of gene flow at different points along the speciation continuum.

Methods

Data sampling and quality control

We collated data on genetic distance, reproductive isolation (RI) and range overlap for 93 pairs of Drosophila species from published datasets [63, 64, 13, 14] (see Supplementary table 5 for RI data). These pairs are relatively phylogenetically broad, as we include species pairs from the melanogaster, repleta, virilis, immigrans, obscura, and willistoni groups. According to a recent *Drosophila* phylogenies [32, 19], these lineages are well distributed across the genus. We augmented these with analyses of demographic history for each species pair using genomic data to test key assumptions about the relationship between current range overlap and speciation history. To do this, we obtained one genome assembly (based mostly on long read data) per species pair and whole-genome sequencing (WGS) data (Illumina, short read) for each species from publicly available datasets via NCBI (Supplementary table 3). We only included WGS data with similar coverage, removed pool-sequenced or experimentally manipulated samples, and prioritised data obtained from wild-caught individuals (rather than lab lines). For WGS datasets that met these criteria, we randomly assigned one resequencing dataset to represent each species. Full filtering details can be found in SI Appendix (section 1.1). Details on genome assemblies and WGS datasets sampled can be found in Supplementary table 3 and 4 in the SI Appendix.

Genome annotation, mapping and variant calling

We annotated each genome separately using BRAKER2 (v2.1.6) and *D. melanogaster* protein sequences as evidence [8]. Raw reads for each WGS dataset were trimmed using fastp [10], aligned using bwa-mem2 [36, 59], and finally sorted and filtered using sambamba [57] and picard. Both WGS datasets of each species pair were mapped to the annotated genome (belonging to one of the two species) in the pair. Variants were called using freebayes and filtered for missing genotypes, read mapping bias and depth using bcftools and gIMble[24, 35, 34, 16]. For full details on annotation, mapping and filtering see SI Appendix (section 1.2).

Sampling intronic blocks

We opted to use intronic sequences for demographic analyses since introns can be more reliably annotated than intergenic regions but show much less functional constraint than protein-coding genes. Functional constraint across introns can be variable, so we implemented a range of filtering strategies to keep intronic segments most likely to be selectively neutral[27, 25]. To ensure we minimised linkage between intronic regions, we only sampled one intronic block per gene. Intronic blocks of a fixed, species-specific length were sampled from callable variants after filtering. This means that blocks could span intronic sequence (within a single intron) that were excluded because quality or coverage filters. Our filtering strategy is described in full detail and can be found in the Supplementary Material (section 1.3 and Supplementary Fig. 9). As a sanity check on our filtering strategy we compared d_{XY} in the filtered intron dataset to d_{XY} calculated across all sites in the genome (Supplementary Fig. 10). This confirms that selective constraint on the intronic sequences included in our analyses is low compared to the genome-wide average.

Extracting genomic DNA for a single *Drosophila* sample for sequencing often requires pooling multiple individuals from an isofemale line. Even when a single individual can be sequenced, the heterozygosity in the resulting WGS data reflects both the effective size (N_e) at the species level and the history of lab culture. We therefore restrict our analysis to the most minimal sampling scheme of a single haplotype per species and base inference on the distribution of pairwise differences (S) between species. The S distribution in short intronic block is a vector of counts $\bar{k} = n_1, n_2 \ldots n_{k_{\max}}$). We assumed that the heterozygous sites in each block can be randomly assigned to haplotypes using a simple binomial sampling procedure (implemented in `Mathematica` v12.3) that considers all ways of phasing heterozygous sites: e.g. a block containing two fixed differences between species and a single heterozygous site contributes probabilities $\frac{1}{2}$ each to $S = 2$ and 3. Likewise, a block with two fixed differences and two heterozygous sites contributes probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ to $S = 2$, 3 and 4 respectively. To account for differences in the number of blocks between species pairs, we normalised S distributions by $500/n_i$ (where n_i is the number of blocks in pair i).

Modelling speciation histories across *Drosophila*

To understand demographic histories and the degree of gene flow for species pairs, we fitted a range of models of demographic history to the distribution of pairwise differences in short intronic blocks between species: (a) a strict isolation (SI) model (most consistent with strict allopatric speciation) characterised as an instantaneous split of an ancestral population at T_0 without gene flow, (b) isolation with migration (IM), where an ancestral population diverges with symmetric migration (M = $4N_e$m migrants per generation, where N_e is the ancestral population) between the time of divergence and the present, (c) an isolation with initial migration model (IIM), where an ancestral population diverges with an initial period of symmetric migration and gene flow ceases at T_1 and (d) a secondary contact (SC) model, where an ancestral population diverges in allopatry, and an instantaneous recent, pulse of gene flow, i.e. a total proportion (f) of the population is introgressed at time T_1. This secondary contact model differs from the other SC models in the literature in that gene flow is modeled as a short bidirectional pulse rather than continuous migration from T_1 to the present [53], reflecting the expectation that

reinforcement rapidly halts gene flow during secondary contact. In all cases, we assumed a single N_e parameter that is shared between the ancestral population and the daughter species. See Supplementary Figure 7 for visualisations of each model.

Analytic solutions derived in [61, 38, 62] allow efficient maximum likelihood estimation of parameters under the IM and IIM model from the S distribution. We use the expression for the probability of seeing k differences between a pair of sequences sampled from different populations under the IM model and IIM model $P[S = k]$ [61, eq. 24] and [62, eq. 29]. We assumed a single N_e parameter $\theta = 4N_e\mu$ for the ancestor and the two daughter species and a fixed mutation rate across all blocks. The likelihood expression for the secondary contact model was adapted from [37]. For each model and species pair we maximized the log likelihood across blocks:

$$\ln L[T_0, T_1, M, \theta|\bar{k}] = \sum_{k=0}^{k_{\max}} \ln P[S = k] \times n_k \qquad (1)$$

This likelihood calculation was implemented in `Mathematica`; we used the function FindMaximum to maximise eq. 1 and obtained maximum likelihood estimates (MLE) of parameters under each model and for each species pair. Given that the SI, IM and IIM models are nested, we used likelihood ratio tests (assuming the $2\Delta \ln L$ follows a χ^2 distribution) to determine relative model support for each species pair. Since the IIM and SC models have the same number of parameters, we compared them simply in terms of relative log-likelihood. A detailed description of our rationale for evaluation of model support can be found in Supplementary Material (section 1.4.1 and 1.4.2).

To scale parameter estimates, we used the spontaneous mutation rate estimate of $\mu = 3.32 \times 10^{-9}$ per base and generation for *D. melanogaster* and additionally included the 95% confidence intervals (2.52×10^{-9} - 4.30×10^{-9}) of this estimate (26).

Simulations

A key assumption of our inference framework is that it is possible to sample neutrally evolving variation in short block of sequence within which recombination can be ignored. Violations of this assumption can lead to biases in parameter estimates and could result in erroneous support for histories of gene flow[60]. To rule out this possibility, we simulated S distributions for each pair in our dataset using the maximum likelihood parameter estimates under a SI model and realistic rates of recombination ($r = 1.03$ cM/Mb) and mutation 3.32×10^{-9} (38). Comparing $\Delta \ln L$ between SI and IM histories for simulated datasets suggest a maximum false positive rate of 25 %, far below the proportion of species pairs that support a history of gene flow (87 %). Moreover, we find that the IM model has considerably higher relative support in the real data compared to data simulated under a null model with recombination but no gene flow. Full details can be found in SI Appendix (section 1.5).

Acknowledgments

We thank the reviewer and the associate editor for their feedback. We thank Sam Ebdon for bioinformatic advice. Additionally, we thank Roger K. Butlin and Jim Mallet for helpful feedback that improved the manuscript, and Cher Chow for assistance with figures. LY was supported by a University of St Andrews studentship. KL was supported by a fellowship from the Natural Environment Research Council (NERC, NE/L011522/1) and a European Research Council starting grant (ModelGenomLand 757648) which also supported DRL. MGR was supported by a grant from the Natural Environment Research Council (NE/V001566/1).

Author contributions

Conceptualisation: LHY, KL, MGR. Analysis: LHY, DRL, KL. Writing (original draft): LHY, KL, MGR. Writing (review and editing): LHY, KL, MGR.
The authors declare that they have no competing interests.

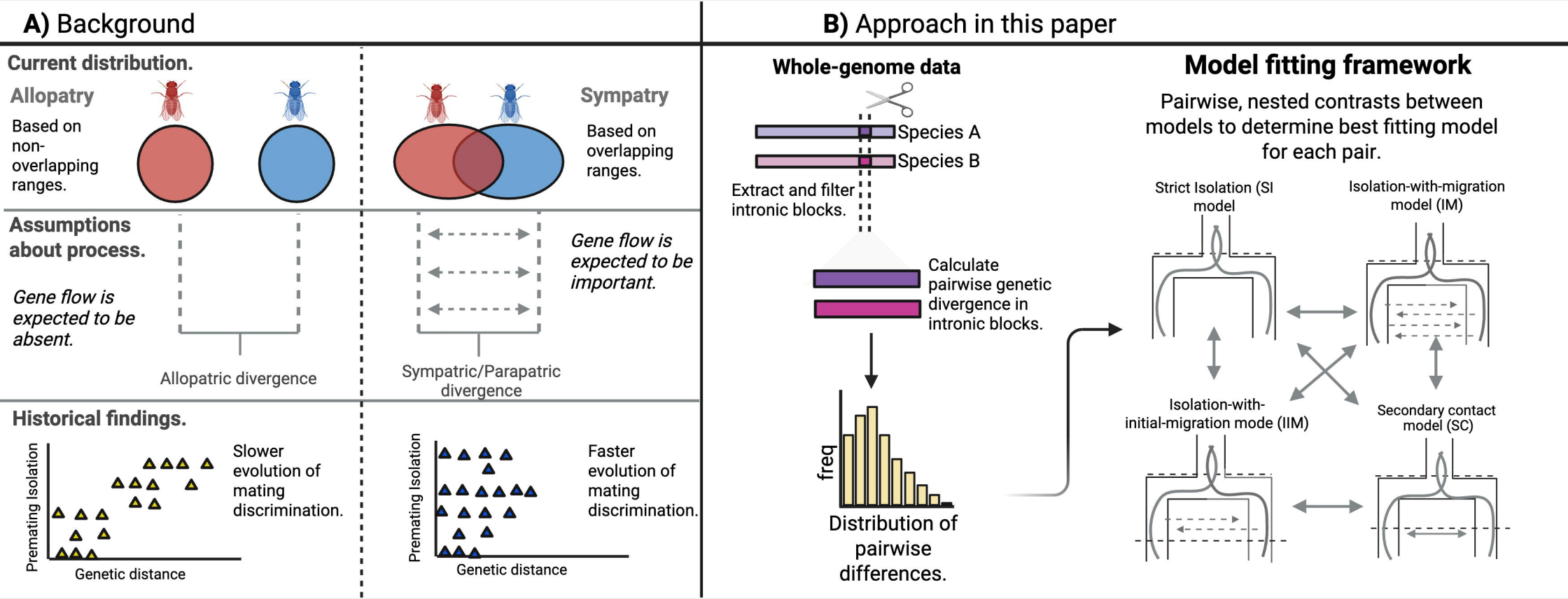
References

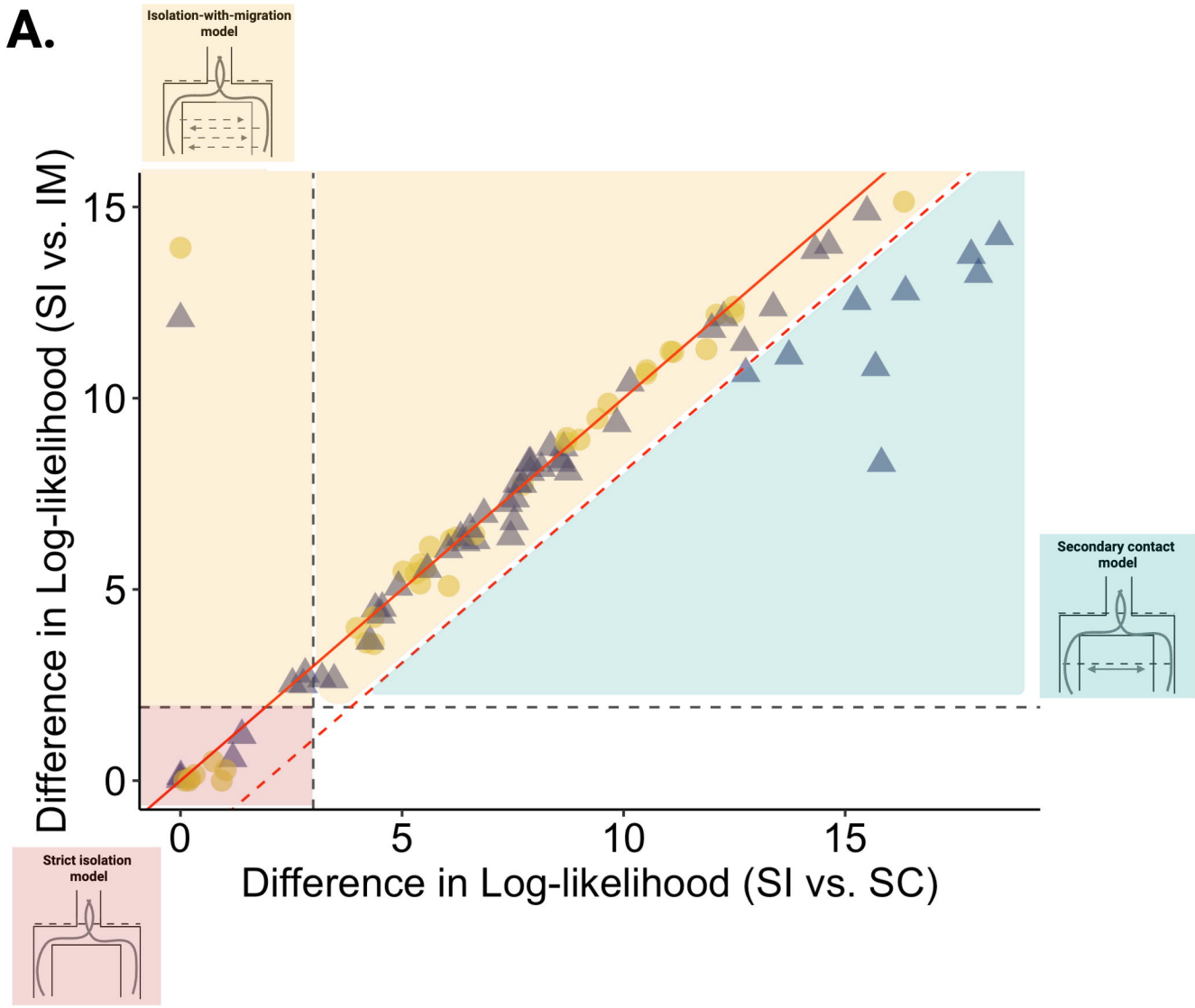
1. R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, and R. Buggs. Hybridization and speciation. *Journal of evolutionary biology*, 26(2):229–246, 2013. Publisher: Blackwell Publishing Ltd Oxford, UK.
2. S. A. Anderson, S. Kaushik, and D. R. Matute. The comparative analysis of lineage-pair traits. *Systematic Biology*, page syaf061, 2025.
3. S. A. Anderson and J. T. Weir. The role of divergent ecological adaptation during allopatric speciation in vertebrates. *Science*, 378(6625):1214–1218, 2022.
4. D. Bachtrog. The speciation history of the Drosophila nasuta complex. *Genetics Research*, 88(1):13–26, 2006. Publisher: Cambridge University Press.
5. T. G. Barraclough and A. P. Vogler. Detecting the Geographical Pattern of Speciation from Species-Level Phylogenies. *The American Naturalist*, 155(4):419–434, Apr. 2000.
6. P. Beerli. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*, 13(4):827–836, 2004.
7. G. Bisschop, D. Setter, M. Rafajlović, S. J. Baird, and K. Lohse. The impact of global selection on local adaptation and reproductive isolation. *Philosophical Transactions of the Royal Society B*, 375(1806):20190531, 2020.
8. T. Brna, K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky. Braker2: automatic eukaryotic genome annotation with genemark-ep+ and augustus supported by a protein database. *NAR genomics and bioinformatics*, 3(1):lqaa108, 2021.
9. R. K. Butlin, J. Galindo, and J. W. Grahame. Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506):2997–3007, 2008.
10. S. Chen, Y. Zhou, Y. Chen, and J. Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.
11. H. Cheng, A. Sinha, F. W. Cruz, X. Wang, R. L. Edwards, F. M. d'Horta, C. C. Ribas, M. Vuille, L. D. Stott, and A. S. Auler. Climate change patterns in amazonia and biodiversity. *Nature communications*, 4(1):1411, 2013.
12. J. A. Coyne, I. A. Boussy, T. Prout, S. H. Bryant, J. Jones, and J. A. Moore. Long-distance migration of drosophila. *The American Naturalist*, 119(4):589–595, 1982.
13. J. A. Coyne and H. A. Orr. PATTERNS OF SPECIATION IN DROSOPHILA. *Evolution*, 43(2):362–381, Mar. 1989.
14. J. A. Coyne and H. A. Orr. " Patterns of speciation in Drosophila" revisited. *Evolution*, pages 295–303, 1997. Publisher: JSTOR.

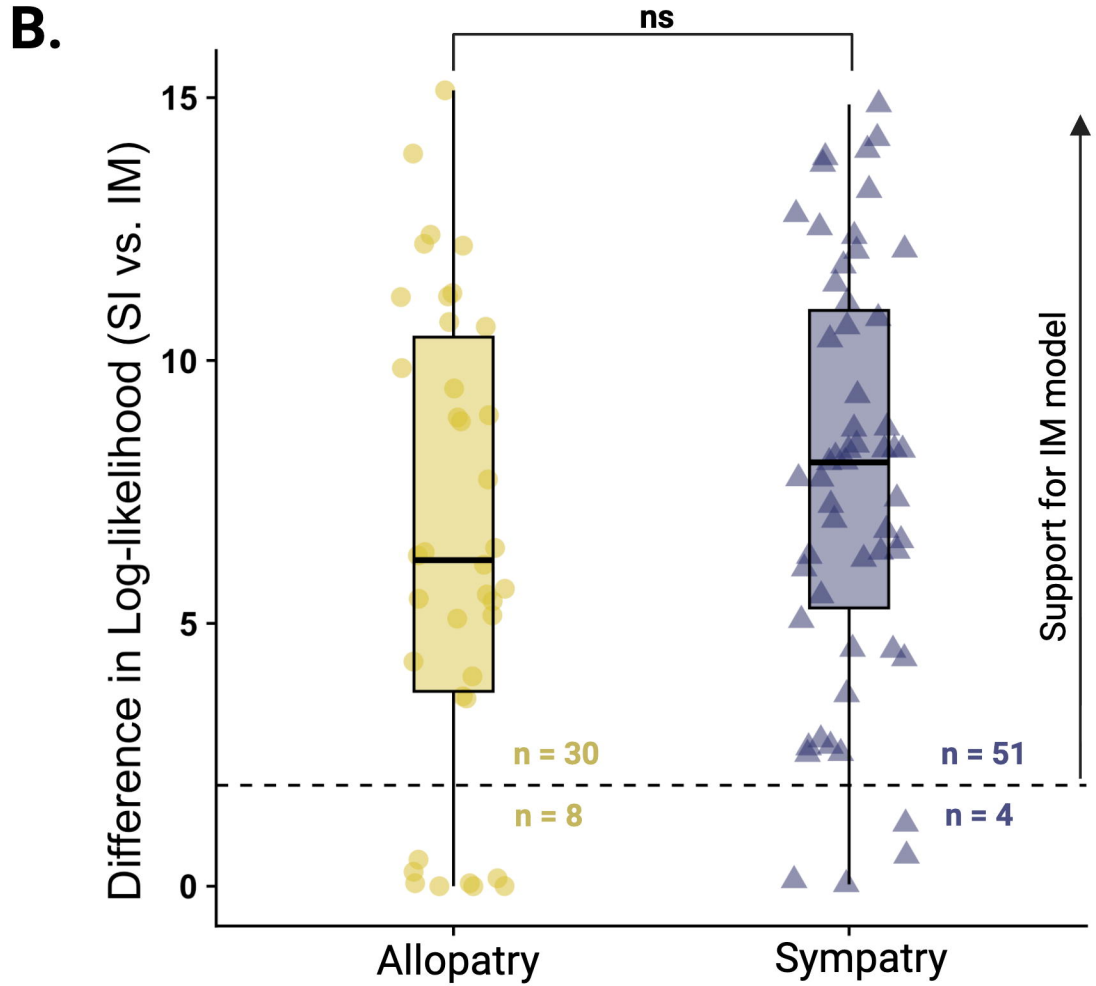
15. J. A. Coyne and H. A. Orr. *Speciation*. Sinauer Associates,, 2004.
16. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giab008, 2021.
17. A. De Jode, A. Le Moan, K. Johannesson, R. Faria, S. Stankowski, A. M. Westram, R. K. Butlin, M. Rafajlović, and C. Fraïsse. Ten years of demographic modelling of divergence and speciation in the sea. *Evolutionary Applications*, 16(2):542–559, Feb. 2023.
18. N. B. Edelman, P. B. Frandsen, M. Miyagi, B. Clavijo, J. Davey, R. B. Dikow, G. García-Accinelli, S. M. Van Belleghem, N. Patterson, D. E. Neafsey, R. Challis, S. Kumar, G. R. P. Moreira, C. Salazar, M. Chouteau, B. A. Counterman, R. Papa, M. Blaxter, R. D. Reed, K. K. Dasmahapatra, M. Kronforst, M. Joron, C. D. Jiggins, W. O. McMillan, F. Di Palma, A. J. Blumberg, J. Wakeley, D. Jaffe, and J. Mallet. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465):594–599, Nov. 2019.
19. C. Finet, V. A. Kassner, A. B. Carvalho, H. Chung, J. P. Day, S. Day, E. K. Delaney, F. C. De Ré, H. D. Dufour, E. Dupim, et al. Drosophyla: resources for drosophilid phylogeny and systematics. *Genome biology and evolution*, 13(8):evab179, 2021.
20. B. M. Fitzpatrick, J. A. Fordyce, and S. Gavrilets. Pattern, process and geographic modes of speciation. *Journal of evolutionary biology*, 22(11):2342–2347, 2009. Publisher: Blackwell Publishing Ltd Oxford, UK.
21. M. C. Fontaine, J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, S. N. Mitchell, Y.-C. Wu, H. A. Smith, R. R. Love, M. K. Lawniczak, M. A. Slotman, S. J. Emrich, M. W. Hahn, and N. J. Besansky. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524, Jan. 2015.
22. C. Fraïsse, I. Popovic, C. Mazoyer, B. Spataro, S. Delmotte, J. Romiguier, E. Loire, A. Simon, N. Galtier, L. Duret, et al. Dils: Demographic inferences with linked selection by using abc. *Molecular Ecology Resources*, 21(8):2629–2644, 2021.
23. D. J. Funk, P. Nosil, and W. J. Etges. Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences*, 103(9):3209–3213, Feb. 2006.
24. E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.
25. P. R. Haddrill, B. Charlesworth, D. L. Halligan, and P. Andolfatto. Patterns of intron sequence evolution in drosophila are dependent upon length and gc content. *Genome biology*, 6(8):R67, 2005.
26. J. Haffer. Speciation in amazonian forest birds: most species probably originated in forest refuges during dry climatic periods. *Science*, 165(3889):131–137, 1969.
27. D. L. Halligan and P. D. Keightley. Ubiquitous selective constraints in the drosophila genome revealed by a genome-wide interspecies comparison. *Genome research*, 16(7):875–884, 2006.
28. R. G. Harrison and E. L. Larson. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105(S1):795–809, 2014.
29. G. M. Hewitt. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological journal of the Linnean Society*, 58(3):247–276, 1996.
30. G. M. Hewitt. Genetic consequences of climatic oscillations in the quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1442):183–195, 2004.
31. J. Hey and R. Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, 167(2):747–760, 2004. Publisher: Oxford University Press.
32. B. Y. Kim, H. R. Gellert, S. H. Church, A. Suvorov, S. S. Anderson, O. Barmina, S. G. Beskid, A. A. Comeault, K. N. Crown, S. E. Diamond, S. Dorus, T. Fujichika, J. A. Hemker, J. Hrcek, M. Kankare, T. Katoh, K. N. Magnacca, R. A. Martin, T. Matsunaga, M. J. Medeiros, D. E. Miller, S. Pitnick, M. Schiffer, S. Simoni, T. E. Steenwinkel, Z. A. Syed, A. Takahashi, K. H.-C. Wei, T. Yokoyama, M. B. Eisen, A. Kopp, D. Matute, D. J. Obbard, P. M. O'Grady, D. K. Price, M. J. Toda, T. Werner, and D. A. Petrov. Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. *PLOS Biology*, 22(7):e3002697, July 2024.
33. R. M. Kliman, P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey. The population genetics of the origin and divergence of the drosophila simulans complex species. *Genetics*, 156(4):1913–1931, 2000.
34. D. R. Laetsch, G. Bisschop, S. H. Martin, S. Aeschbacher, D. Setter, and K. Lohse. Demographically explicit scans for barriers to gene flow using gimble. *PLoS genetics*, 19(10):e1010999, 2023.
35. H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.
36. H. Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.
37. K. Lohse and L. A. Frantz. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, 196(4):1241–1251, 2014. Publisher: Oxford University Press.
38. K. Lohse, R. J. Harrison, and N. H. Barton. A general method for calculating likelihoods under the coalescent process. *Genetics*, 189(3):977–987, 2011.
39. J. B. Losos and R. E. Glor. Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology & Evolution*, 18(5):220–227, 2003. Publisher: Elsevier.
40. C. A. Machado, R. M. Kliman, J. A. Markert, and J. Hey. Inferring the history of speciation from multilocus dna sequence data: the case of drosophila pseudoobscura and close relatives. *Molecular biology and evolution*, 19(4):472–488, 2002.
41. D. Mai, M. J. Nalley, and D. Bachtrog. Patterns of genomic differentiation in the drosophila nasuta species complex. *Molecular biology and evolution*, 37(1):208–220, 2020.
42. J. Mallet. Speciation in the 21st century. *Heredity*, 95(1):105–109, 2005.
43. J. Mallet. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal*

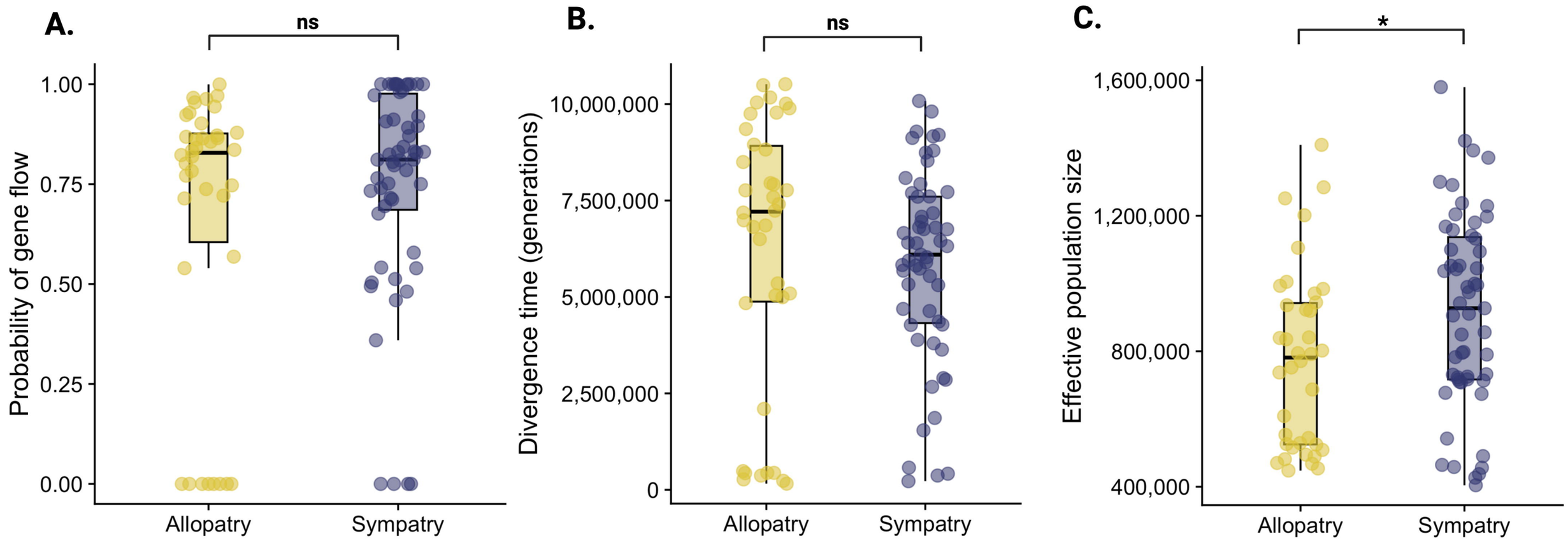
Society B: Biological Sciences, 363(1506):2971–2986, 2008.

44. J. Mallet, A. Meyer, P. Nosil, and J. L. Feder. Space, sympatry and speciation. *Journal of evolutionary biology*, 22(11):2332–2341, 2009.
45. D. R. Matute and B. S. Cooper. Comparative studies on speciation: 30 years since Coyne and Orr. *Evolution*, 75(4):764–778, Apr. 2021.
46. D. R. Matute and B. S. Cooper. Reinforcement alone does not explain increased reproductive isolation in sympatry. *bioRxiv*, pages 2021–05, 2021.
47. E. Mayr. *Animal Species and Evolution*. Harvard University Press, Feb. 1963.
48. E. Mayr. *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press, 1999.
49. M. A. Noor. Reinforcement and other consequences of sympatry. *Heredity*, 83(5):503–508, 1999. Publisher: Nature Publishing Group.
50. M. A. Noor. Reinforcement and other consequences of sympatry. *Heredity*, 83(5):503–508, 1999.
51. P. Nosil. *Ecological Speciation*. Oxford University Press, 2012.
52. P. Nosil. Degree of sympatry affects reinforcement in Drosophila. *Evolution*, 67(3):868–872, 2013. Publisher: Blackwell Publishing Inc Malden, USA.
53. C. Roux, C. Fraisse, J. Romiguier, Y. Anciaux, N. Galtier, and N. Bierne. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS biology*, 14(12):e2000234, 2016. Publisher: Public Library of Science San Francisco, CA USA.
54. S. Singhal, G. E. Derryberry, G. A. Bravo, E. P. Derryberry, R. T. Brumfield, and M. G. Harvey. The dynamics of introgression across an avian radiation. *Evolution Letters*, 5(6):568–581, 2021.
55. S. Stankowski and M. Ravinet. Defining the speciation continuum. *Evolution*, 75(6):1256–1273, 2021.
56. A. Suvorov, B. Y. Kim, J. Wang, E. E. Armstrong, D. Peede, E. R. D'agostino, D. K. Price, P. J. Waddell, M. Lang, and V. Courtier-Orgogozo. Widespread introgression across a phylogeny of 155 Drosophila genomes. *Current Biology*, 32(1):111–123, 2022. Publisher: Elsevier.
57. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.
58. M. Turelli, J. R. Lipkowitz, and Y. Brandvain. 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(4):1176–1187, 2014. Publisher: Blackwell Publishing Inc Malden, USA.
59. M. Vasimuddin, S. Misra, H. Li, and S. Aluru. Efficient architecture-aware acceleration of bwa-mem for multicore systems. In *2019 IEEE international parallel and distributed processing symposium (IPDPS)*, pages 314–324. IEEE, 2019.
60. J. D. Wall. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, 154(3):1271–1279, 2000.
61. H. M. Wilkinson-Herbots. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theoretical Population Biology*, 73(2):277–288, 2008. Publisher: Elsevier.
62. H. M. Wilkinson-Herbots. The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theoretical population biology*, 82(2):92–108, 2012. Publisher: Elsevier.
63. R. Yukilevich. Asymmetrical patterns of speciation uniquely support reinforcement in Drosophila. *Evolution*, 66(5):1430–1446, 2012. Publisher: Blackwell Publishing Inc Malden, USA.
64. R. Yukilevich. The rate test of speciation: estimating the likelihood of non-allopatric speciation from reproductive isolation rates in Drosophila. *Evolution*, 68(4):1150–1162, 2014. Publisher: Blackwell Publishing Inc Malden, USA.
65. L. H. Yusuf, V. Tyukmaeva, A. Hoikkala, and M. G. Ritchie. Divergence and introgression among the virilis group of drosophila. *Evolution Letters*, 6(6):537–551, 12 2022.


A) Background
Current distribution.
Allopatry
Based on non-overlapping ranges.
Sympatry
Based on overlapping ranges.
Assumptions about process.
Gene flow is expected to be absent.
Allopatric divergence
Gene flow is expected to be important.
Sympatric/Parapatric divergence
Historical findings.
Premating Isolation
Genetic distance
Slower evolution of mating discrimination.
Premating Isolation
Genetic distance
Faster evolution of mating discrimination.
B) Approach in this paper
Whole-genome data
Species A
Species B
Extract and filter intronic blocks.
Calculate pairwise genetic divergence in intronic blocks.
freq
Distribution of pairwise differences.
Model fitting framework
Pairwise, nested contrasts between models to determine best fitting model for each pair.
Strict Isolation (SI model
Isolation-with-migration model (IM)
Isolation-with-initial-migration mode (IIM)
Secondary contact model (SC)


A.
Isolation-with-migration model
Secondary contact model
Strict isolation model
Difference in Log-likelihood (SI vs. IM)
Difference in Log-likelihood (SI vs. SC)
0
5
10
15


B.
ns
Difference in Log-likelihood (SI vs. IM)
Support for IM model
n = 30
n = 8
n = 51
n = 4
Allopatry
Sympatry
0
5
10
15


A.
ns
Probability of gene flow
1.00
0.75
0.50
0.25
0.00
Allopatry
Sympatry
B.
ns
Divergence time (generations)
10,000,000
7,500,000
5,000,000
2,500,000
0
Allopatry
Sympatry
C.
*
Effective population size
1,600,000
1,200,000
800,000
400,000
Allopatry
Sympatry