Highly replicated evolution of parapatric ecotypes

Maddie E. James[1*], Henry Arenas-Castro[1], Jeffery S. Groh[1,2], Jan Engelstädter[1], Daniel Ortiz-Barrientos[1]

[1]The University of Queensland, School of Biological Sciences, St. Lucia QLD 4072, Australia. [2]Current address: University of California, Davis, Department of Evolution and Ecology, Davis, CA 95616, United States.

* Corresponding author

Abstract

Parallel evolution of ecotypes occurs when selection independently drives the evolution of similar traits across similar environments. The multiple origin of ecotypes is often inferred on the basis of a phylogeny which clusters populations according to geographic location and not by the environment they occupy. However, the use of phylogenies to infer parallel evolution in closely related populations is problematic due to the potential for gene flow and incomplete lineage sorting to uncouple the genetic structure at neutral markers from the colonization history of populations. Here, we demonstrate multiple origins within ecotypes of an Australian wildflower, *Senecio lautus*. We observed strong genetic structure as well as phylogenetic clustering by geography, and show this is unlikely due to gene flow between parapatric ecotypes, which is surprisingly low. We further confirm this analytically by demonstrating that phylogenetic distortion due to gene flow often requires higher levels of migration than those observed in *S. lautus*. Our results imply that selection can repeatedly create similar phenotypes despite the perceived homogenizing effects of gene flow.

Introduction

Parallel evolution occurs when populations evolve similar traits after repeatedly and independently colonizing similar habitats[1]. The patchy distribution of habitats means that phenotypically similar populations frequently occur next to other contrasting phenotypes (e.g., plant species adapted to serpentine and non-serpentine soils in Scandinavia[2], and marine snails adapted to crab predators or wave action along the rocky coasts of Spain[3]). Parallel evolution by natural selection creates consistent patterns of phenotypic similarity and divergence that can extend to morphological[4–6], behavioural[7], and reproductive[8] traits. The nature of parallel trait evolution largely depends on the demographic history of the system under investigation, where the interplay of geography, gene flow, and natural selection with the genetic architecture of traits determines its repeatability[9–15]. However, it is surprisingly rare for studies of parallel evolution to convincingly demonstrate that populations exhibiting similar phenotypes have arisen in an independent and repeated fashion ('multiple origin' scenario). Ruling out alternative demographic scenarios, such as a single origin of ecotypes followed by gene flow upon secondary contact, is seldom performed (but see refs. [16–19], and see ref. [20] for a critical review of the evidence in plants). In light of this, researchers may incorrectly assume a parallel demographic history, leading to inaccurate inferences about the prevalence of parallel evolution in nature.

Typically, researchers identify parallel evolution by natural selection by asking whether phylogenetic clustering of populations coincides with the geography and not with the ecology of populations[3,17,18,21,22]. This is because genetic clustering of geographically close populations implies dispersal might be geographically restricted (i.e., isolation by distance[23]), and colonization of contrasting and neighboring habitats might have occurred independently many times. The rationale for this argument is that the genome-wide phylogenetic history can be used as a proxy for understanding the history of adaptation across multiple populations. That is, if adaptation appears to have taken place on different genetic backgrounds, then the genetic changes that drove adaptation likely occurred independently. By genetic changes, we refer specifically to independent allele frequency changes driven by similar natural selection pressures, rather than the identity of the beneficial mutations themselves[24].

The above argument rests upon the assumption that the genome-wide pattern of relatedness accurately depicts the history of the loci underlying adaptation, though this is not necessarily the case. For example, alternative historical scenarios could also lead to clustering of

54 populations by geography, and must be ruled out before examining the evolution of traits in
55 light of parallel evolution[16,25–27]. To understand this problem, first consider a scenario where
56 an ancestral population gives rise to two locally adapted populations that occupy ecologically
57 distinct yet geographically proximate habitats (hereafter ecotypes, Fig. 1A). These two
58 populations migrate to new localities in parallel, where each time the same contrasting
59 habitats are geographically close. This scenario of a single split followed by range expansion
60 of two ecotypes does not involve a parallel adaptation history because each ecotype only
61 arose once (rather than multiple independent times after independent colonization of
62 contrasting habitats). Because gene flow is either not possible after the original ecotypic split,
63 or does not homogenize adjacent populations after range expansion, populations sharing the
64 same ecology form reciprocally monophyletic clades in a phylogeny (Fig. 1A).

65 Nevertheless, if there is sufficient gene flow between geographically close populations from
66 two ecotypes that originated only once, the original phylogenetic signal of reciprocal
67 monophyly can be eroded[16,25–27]. In other words, as the original genome-wide phylogenetic
68 signal of a single origin disappears, populations become most related to their neighboring
69 population and not to the other populations of the same ecotype. Therefore, gene flow can
70 result in grouping of populations by geography rather than ecology (Fig. 1B). This
71 phylogenetic signal is identical to that of true parallel evolution (a multiple origin scenario),
72 where the derived ecotype arises multiple independent times from the ancestral ecotype (Fig.
73 1C). Gene flow dynamics can thus fundamentally alter our interpretation of parallel
74 evolution, to the extent that we can mistakenly infer parallel evolution in systems where
75 secondary contact after range expansion of a single origin has obscured the history of locally
76 adapted populations[16,25–27]. We note that non-monophyly is not a requirement for parallel
77 evolution in a more general sense, but it is so in systems where parallel evolution coincides
78 with a patchy geographic distribution of populations (pairs of ecotypes in multiple
79 localities)[16], where the phylogenetic line of reasoning is commonly employed.

80 In systems of parallel evolution, gene flow is frequently detected between populations,
81 especially when contrasting ecotypes are in close geographic proximity (i.e., parapatry).
82 Although not all levels of gene flow have the same equivocal effect on the genetic record of
83 colonization history[16], gene flow between ecotypes must be taken into account when
84 demonstrating parallel evolution within a system. However, only very few systems have
85 thoroughly investigated the demographic history of populations, and even fewer have used

coalescent modelling or simulations to address whether the estimated levels of gene flow could have obscured the observed phylogeny. The system that has perhaps most clearly demonstrated the parallel origins of contrasting populations in the presence of gene flow is the marine snail *Littorina saxatilis*. Multiple lines of evidence suggest the wave and crab ecotypes have evolved multiple independent times along rocky coastlines[3,16–19]. Other systems providing clear evidence for parallel evolution include Lake Victoria cichlids[28] and alpine and montane *Heliosperma pusillum* ecotypes[22]. Also, an obvious case of multiple origins is when parallel evolution occurs between geographically distant populations where gene flow could not have obscured the phylogenetic signal and demographic history of populations (e.g., threespine stickleback populations that colonized freshwater environments on separate continents[29]). However, in other systems where gene flow is moderate between ecotypes[30–34], it remains unclear to what extent gene flow contributed to the signal of parallel evolution.

Despite the potential for gene flow to paint a false picture of the phylogenetic history of multiple populations adapted to the same environment, this false signal itself does not necessarily negate the argument for parallel evolution. For one, it is possible that two populations that are true sister groups in a phylogenetic sense –a single origin of the genetic background– did in fact adapt independently. Second, distortion of the phylogenetic topology by gene flow occurs under rather restrictive settings, namely high migration rates, as we show in our results. Thus, for the trait-environment association to have persisted in spite of constant reintroduction of maladaptive alleles implies that selection must have had to regenerate the optimal phenotype independently across populations. In other words, if gene flow was indeed sufficiently high to have distorted the phylogenetic topology in multiple cases, selection must have independently resisted the introduction of maladaptive alleles. In this manner, Lee and Coop's (ref. [24]) framing of independence as an overlap in selective deaths across populations can be extended to consider both the overlap in selective deaths during the initial sweep, as well as during a secondary phase of resisting maladaptive gene flow.

We must keep in mind that identifying the genetic basis of parallel trait evolution often provides unambiguous evidence for parallel evolution of ecotypes. For instance, in sticklebacks, the repeated evolution of pelvic loss in separate populations occurred via different mutations in the same gene[35], suggesting this adaptive trait has arisen and been

selected for multiple independent times. In contrast, in systems where the exact same mutation is repeatedly involved in adaptation (e.g., ref. [36]), it is challenging to identify whether the adaptive mutation was repeatedly and independently selected for in each population (either from de-novo mutations or via standing genetic variation[24,37]). Knowing the causal genes of adaptation is ideal as the demographic history of individual adaptive loci can be modelled, avoiding the complications of distinguishing between single and multiple origins using neutral polymorphisms (as described above). However, directly isolating the specific genes involved in adaptation is infeasible in most non-model organisms or when the genetic architecture of adaptation is highly polygenic[38,39].

The above considerations suggest that without knowing the genetic basis of parallel adaptation, we should carefully characterize and interpret the phylogeographic history to understand the level of independence in systems where populations are adapted to similar environments. Such an approach is necessary to demonstrate that natural selection has independently acted in separate populations during the repeated adaptation to similar environments. This knowledge paves the way for future research on dissecting the molecular basis of parallel adaptation, and its implications for our understanding of the predictability and repeatability of evolution. In this work, we characterize the phylogenetic and demographic history of *Senecio lautus*, an Australian wildflower that appears to have evolved multiple times in parapatry into two contrasting coastal forms called Dune and Headland ecotypes[40,41]. The two forms differ in their growth habit: the Dune ecotype is erect and inhabits sand dunes, and the Headland ecotype is prostrate, forming matts on the ground of rocky headlands[42–44]. These locally adapted populations[40,45–49] are separated by strong extrinsic reproductive isolation[40,47], and populations exhibit similar morphology within each ecotype[50,51]. With this work we hope to clearly illustrate how the demographic history of populations affects the evidence for the independent and repeated origins of parapatric ecotypes.

Previous work using pools of DNA sequences from multiple coastal, inland, alpine, and woodland *S. lautus* ecotypes found that strong isolation by distance separated all populations along the coast and that geography, not ecology, explained the phylogenetic clustering of its coastal populations[41]. Although these results suggest that the Dune and Headland ecotypes have evolved in parallel, it remains unclear if gene flow could be responsible for this pattern of ecotypic and geographic differentiation, thus potentially affecting our inferences on the

number of independent colonization and origin events of multiple Dune and Headland populations. Here, we directly estimate patterns of gene flow between 23 Dune and Headland populations, as well as other parameters important for characterizing the demographic history of this system. We create a coalescent model to explore the conditions that would erode a signal of phylogenetic monophyly of each ecotype, thus enabling us to gain further confidence in our conclusions about parallel parapatric divergence in this system. Our results enhance our understanding of the nature of parallel evolution and pave the way for analyses of parallel trait evolution driven by natural selection in plants, where cases of parallelism remain understudied.

Results

Populations cluster by geography and not by ecology

To ask whether populations cluster according to their geographic distribution, we explored broad patterns of genetic clustering across the 23 Dune and Headland *S. lautus* populations (Fig 2A). Phylogenetic inference in *IQ-TREE* provides evidence against a single origin scenario: neither ecotype forms a monophyletic clade, and parapatric Dune-Headland populations are often sister-taxa, giving evidence for the multiple origin of ecotypes (Fig. 2B). To visualize the major genetic structure within *fastSTRUCTURE*, we plotted the lowest K-values that capture the major structure in the data as suggested by refs. [52,53], although the "best" K-value across all populations was higher (see below). The clustering of populations into two genetic groups (K=2) revealed a striking correspondence to geography (Fig. 2C), where the eastern populations (dark blue) are separated from those populations further south and to the west (light blue). This strong genetic structuring into two main clades suggests there are at least two independent origins within the system. When three genetic groups (K=3) are considered, the eastern populations are further separated into two clusters, again largely corresponding to geography and reflecting the phylogenetic structure of the data; K=4 distinguishes the west Australia populations from those on the south-eastern coast. This genetic clustering of populations according to their geographic distribution provides further evidence against a single origin scenario, and is consistent with previous work in this system[41,54].

Minimal admixture across the system

To understand the role of gene flow in shaping the patterns of divergence across the system, we explored patterns of admixture in a phylogenetic context within *TreeMix* and formally tested for admixture using *f3-statistics* across the 23 Dune and Headland populations. In the absence of migration, the *TreeMix* phylogeny explained 95.9% of the data, with 24 additional migration events augmenting this value to 98.9 % (Supplementary Fig. S1). Fig. 3A shows the first migration event ($P < 2.2x10^{-308}$) with a migration weight (*w*) of 0.40. Although the 24 other migration events were also significant ($P_{average} = 2.92x10^{-3}$, SD = 0.0062), their individual weightings were small (see Supplementary Fig. S2 for 1-10 migration events), most of them were not between parapatric pairs, and the addition of these migration events did not substantially alter the topology from its estimation in the absence of gene flow. Although these results could suggest a potential complex colonization history including long distance yet rare migration events, these P-values should be treated with caution. This is because model comparisons in *TreeMix* suffer from multiple testing, a large number of parameters, and the estimated graph can be inaccurate[55]. We therefore tested the robustness of this inference using *f3-statistics*. All *f3-statistics* were positive (Supplementary Fig. S3), giving no evidence of admixture between any populations. Strong isolation by distance within each ecotype further supports this contention using $\widehat{M}$ as a proxy for migration rates (IBD within Dunes: Mantel test, r = -0.83, P = 0.0001; within Headlands r = -0.73, P = <0.0001; Fig. 3C). A strong IBD trend exists between ecotypes for the eight pairs studied here ($\widehat{M}$: $F_{1,6} = 0.55$, P = 0.05661, multiple $R^2 = 0.48$, Fig. 3C). Although the same trend was seen in the migration rate estimates from *fastsimcoal2*, it was not statistically significant, perhaps due to the low sample size (*fastsimcoal2*: $F_{1,6} = 0.53$, P = 0.4953, multiple $R^2 = 0.08$, Fig. 3B). Overall, this pattern of IBD implies that there is geographically restricted dispersal within the system and populations are evolving largely independently from one another.

The absence of admixture across the system is also supported by *fastSTRUCTURE* across all populations. The inferred value of K is close to the number of sampled populations (Supplementary Figs. S4B, S4C) and each population is genetically distinct (Supplementary Fig. S4A), suggesting that *S. lautus* has a simple demographic history with limited admixture[52]. Specifically, the K-value that best explained the structure in the data was 22, and the K-value that maximized the marginal likelihood of the data was 28 (Supplementary Fig. S4B). The rate of change in the likelihood of each K-value was negligible for K=24-28

(Supplementary Fig. S4C). Together, this suggests that the optimal K-value is around 23, which is the number of populations within our study. The *fastSTRUCTURE* results for K=23 show that each population forms a distinct genetic cluster (Fig. 3D), suggesting very little, if any, admixture between them. This further implies that each sampled population has been separated from other populations long enough to be genetically distinct (see pairwise F_{ST} values in Supplementary Table S1) and with insufficient levels of gene flow to homogenize their genomes[52]. Further, when we examine all K-values from 1-23, there is a distinct hierarchical structure that mirrors the phylogeny, suggesting that such structure is an accurate representation of the history of the populations. The Tasmania population pair (D14-H15) should be treated with caution due to the smaller sample size ($n_{Dune} = 12$, $n_{Headland} = 11$) compared to other populations ($n_{mean} = 62$, SD = 1.19). For groups with few samples, genetic clustering programs such as *fastSTRUCTURE* are likely to assign them as mixtures of multiple populations rather than their own distinct population[52]. This is evident for K=22, where the Tasmania populations appear admixed (Supplementary Fig. S4A).

Minimal gene flow between parapatric ecotypes and distant populations

We investigated whether the parapatric ecotypes at each locality have diverged in the face of gene flow by analyzing patterns of admixture in *STRUCTURE* and directly estimating levels of gene flow in *fastsimcoal2.* We observed very few admixed individuals between the parapatric Dune-Headland populations at each locality within the *STRUCTURE* analysis for K=2 (Fig. 4B). On average, 9.36% (SD = 5.48) of individuals were admixed per population, although their admixture proportions were on average less than 1% (mean = 0.8%, SD = 1.8). This suggests that gene flow between parapatric populations might have ceased in the past. For all pairs, the best K-value based on the Evanno method[56] was K=2 (Supplementary Fig. S5). Demographic modelling in *fastsimcoal2* revealed the most likely divergence model for all comparisons within and between ecotypes was bidirectional gene flow after secondary contact ($w_i > 0.99$; Supplementary Fig. S6). For Dune-Headland population pairs, direct measurements of migration rates were very low, with all Dune-Headland migration rates below one ($2Nm < 1.00$), except for D04-H05 and D32-H12 where $2Nm$ was slightly above one (Fig. 3B upper section, 4B; Supplementary Tables S2, S3). For Dune-Dune population comparisons we also detected very low migration rates ($2Nm_{mean} = 0.23$, SD = 0.09; Supplementary Tables S2, S3), with all comparisons containing $2Nm < 1.00$. Similarly, for Headland-Headland comparisons we again detected very low migration rates ($2Nm_{mean}$ =

0.57, SD = 1.01), with all comparisons containing $2Nm < 1.00$, with the exception of H12-H12A (Fig. 3B upper section, 4B; Supplementary Tables S2, S3). Across all comparisons, all Dune-Dune pairs and most Headland-Headland pairs exhibited gene flow levels lower than the maximum migration rate of allopatric populations separated by more than 1,500 km (i.e., the null gene flow expectation; $2Nm = 0.39$; Fig. 3B). Three Dune-Headland pairs (D00-H00, D03-H02 and D12-H14) were also within this null range. Alternative models that assumed negligible gene flow, while keeping other demographic parameters fixed, did not fit the data better with the exception of the D03-H02 pair (Supplementary Fig. S7). Although the most likely divergence scenario for all population comparisons was bidirectional gene flow after secondary contact, migration rates under all models were very low across all population pairs ($2Nm_{mean} = 0.58$, SD = 1.43, Supplementary Table S4). Thus, even if our choice of model was biased towards secondary contact[57], the extent of gene flow during the history of populations is consistently low and does not depend strongly upon the mode of divergence.

Potential for gene flow to obscure a single origin scenario

We analyzed a neutral coalescent model representing a single origin of the derived ecotype to investigate under which situations the history at a neutral locus unlinked to the selected site would indicate a pattern of non-monophyly of the derived ecotype, thus potentially supporting a false inference of parallel evolution. We found a clear influence of all examined parameters (Fig. 5A) on the probability of inferring non-monophyly, suggesting that certain demographic scenarios can lead to an observed phylogenetic signal that belies the history of a single origin of ecotypes. When internal branches ($t2 - t1$) are short, and ancestral polymorphism is expected to be elevated, the probability of distortion is high and relatively independent of migration rates in the terminal branches (Fig. 5B, lower section of graph). When internal branches are long, the probability of distortion is low and again relatively independent of migration rates (Fig. 5B, upper section of graph). Furthermore, when the terminal branches ($t1$) are long relative to the timing of the burst of migration (tm), phylogenetic distortion requires high levels of migration (Fig. 5C, upper section of graph). When the terminal branches are short relative to the timing of migration, very high levels of migration are required to distort the phylogeny (Fig. 5C, lower section of graph). Note that the probability of phylogenetic distortion when the terminal branches are long relative to the timing of migration is not as high as the distortion due to short internal branches (pay attention to the different scale of probability in Fig. 5C compared to Fig. 5B).

Although we cannot directly map our observed data for *S. lautus* in the modelled parameter space, we can nonetheless explore the likelihood of phylogenetic distortion by considering divergence time estimates from *fastsimcoal2* in combination with the phylogenetic topology estimated in *IQ-TREE*. If the paraphyly in our phylogeny is not an artifact of gene flow, we expect the order of divergence times estimated from *fastsimcoal2*, which accounts for gene flow, to be in accordance with the observed phylogeny, which does not account for gene flow. We observed deeper divergence times for populations of the same ecotype compared to sister-taxa of different ecotypes. For D04-H05 and D05-H06, the average divergence time between populations of the same ecotype (D04-D05 and H05-H06) was 79,801 generations (SD = 2,698), whereas the average divergence time between populations at each locality (D04-H05 and D05-H06) was 49,317 generations (SD = 26,319). This was also true for D14-H15 and D32-H12, where the average divergence time between populations of the same ecotype (D14-D32 and H15-H12) was 68,723 generations (SD = 17,526), and the average divergence time between populations at each locality (D14-H15 and D32-H12) was 43,318 generations (SD = 6,522). Overall, this gives further evidence that the phylogenetic topology (estimated in the absence of gene flow) has not resulted from gene flow distortion.

Discussion

We have used an array of complementary approaches to disentangle the demographic history of the coastal *Senecio lautus* ecotypes. In this system, many lines of evidence support a multiple origin scenario for the evolution of the parapatric Dune and Headland populations. The demographic history of this system reveals striking population structure and a strong effect of geography and restricted dispersal, to the extent that all populations are evolving largely independently from each other. Together with previous results from transplant experiments[40,45–49], our results convincingly show that selection and drift, rather than gene flow, play a predominant role in the genetic structure among ecotypic populations in this system. Below we discuss these results in light of parallel parapatric divergence in this highly replicated system.

Large scale genetic structure within *S. lautus* clusters populations according to their geographic distribution along the Australian coast, and not by the environment they occupy. Within *fastSTRUCTURE*, the largest genetic groups encompass two clades which are largely independent of each other, do not have evidence of long-distance gene flow between them, and also appear to contain multiple repeated instances of parapatric divergence. This genetic

structure, where populations group by geography and not ecology, is mirrored in the phylogeny, and is consistent with our previous work using targeted sequencing of 13 neutral genes[54] and RADseq using pools of individuals[41]. There is also a strong signal of isolation by distance[23] within each ecotype as well as across Dune-Headland pairs, implying long distance dispersal within the system is not pervasive, and populations are likely at an equilibrium between dispersal and drift[58]. This is perhaps not surprising given that Dune and Headland populations have restricted geographic ranges along the coast.

Fine scale genetic structure at the level of the locality (i.e., parapatric Dune-Headland populations) shows that each population is genetically distinct. F_{ST} values are above 0.2 for most population pair comparisons, and *fastSTRUCTURE* reveals that all parapatric pairs are fully differentiated with little admixture. Consistent with this, no single estimate of the *f3-statistic* for any population triad was negative, further suggesting that there are negligible levels of admixture between parapatric populations as well as across the entire system. Despite the high potential for gene flow between parapatric populations due to their close geographic proximity and relatively weak F1 intrinsic reproductive isolation[40,45,47], multiple transplant experiments in the system have shown that divergent natural selection is strong and creates extrinsic reproductive isolation between Dune and Headland populations at each locality[40,45–49]. Therefore our findings are in agreement with theoretical expectations, where parapatric divergence and speciation is favored when gene flow is limited and selection against immigrants and hybrids is strong[59,60]. Overall, a combination of strong selection and limited dispersal can explain why parapatric populations persist despite the opportunity for homogenizing gene flow between them.

A common doubt arising in purported cases of parallel evolution is whether gene flow is responsible for the grouping of populations by geography and not by ecology[3,16–19,22,28,30,31,34,61]. A single origin of ecotypes combined with high levels of gene flow between parapatric ecotypes at each locality can alter the phylogenetic relationships of populations, falsely suggesting multiple independent origins. This is because genetic structure at neutral markers can be decoupled from colonization history via introgression and incomplete lineage sorting[16,25–27]. This needs careful scrutiny in our system as previous work showed that genomic divergence was more heterogenous in parapatric populations compared to allopatric populations[41], a signature of divergence with gene flow[62,63]. However, this pattern can also arise due to ancestral polymorphism and incomplete lineage sorting if parapatric populations

are younger than allopatric[64], which our current work shows. Thus, processes unrelated to divergence in the face of high levels of gene flow could have also contributed to the patterns of genomic divergence in this system[41].

To help understand the role of gene flow during parapatric divergence in *S. lautus*, we first directly estimated rates of gene flow within *fastsimcoal2*. Unexpectedly, we observed minimal levels of gene flow within parapatric *S. lautus* Dune-Headland pairs. This reveals that previous patterns of genomic divergence among these populations[41] likely reflect a signature of increased genome-wide differentiation over time in allopatry, and incomplete lineage sorting in parapatric populations rather than heterogeneous divergence in the face of high levels of gene flow. Furthermore, we observed that most Dune-Headland levels of bidirectional gene flow were similar to populations from different clades and separated by >1,500km, suggesting that most Dune and Headland populations at each locality could be viewed as effectively allopatric. Although unmodelled changes in population size tend to favor secondary contact models of gene flow[57], our estimated migration rates were consistently very low across all models of gene flow (see Supplementary Table S4 for details) with the notable exceptions of D04-H05 and D32-H12 pairs. These populations experience levels of gene flow that would make them genetically undistinguishable ($2Nm > 1.00$)[65]. Thus, it is still possible that gene flow has altered the observed phylogenetic topology for these populations.

We therefore explored the conditions which are likely to obscure the history of colonization by modelling the neutral divergence process through coalescent analyses. We observed that the likelihood of phylogenetic distortion is accentuated with very short internal branches, which are expected to carry high levels of ancestral polymorphism and therefore increase the probability that true sister taxa do not remain monophyletic; this effect is largely independent of migration rates as gene flow will not contribute more to population similarity beyond to what ancestral polymorphism already does. Short internal branches are frequently detected in systems where diversification occurs rapidly, such as in cases of adaptive radiations[66], which seems to be the case in *S. lautus*. Our theoretical approach also reveals that increasing levels of gene flow increases the likelihood of phylogenetic distortion, especially when the time of migration is further from the population split. This is because when $t1 - tm$ is long, there is more time for a coalescent event to occur that produces a topology different from the true species tree. More importantly, in these cases of secondary contact, quite high levels of

371 migration are required to create the appearance of multiple origins. Although conventional
372 thinking highlights that even small amounts of gene flow have the potential to mix
373 populations and erode their history (e.g., refs. [25–27]), our work suggests that this might not be
374 true under all cases of secondary contact between diverged populations. As expected, our
375 work reveals that it is important to consider the joint contributions of gene flow as well as
376 ancestral polymorphism when inferring the likelihood of phylogenetic distortion.

377 Within *S. lautus*, even though the short internal branches and recent secondary contact have
378 the potential to obscure the phylogeny and falsely suggest parallel evolution, this is likely
379 circumvented by the extremely low rates of gene flow between most parapatric ecotypes. In
380 other words, it appears that higher amounts of gene flow would be needed to counteract the
381 divergence that has accumulated over time in the *S. lautus* system. We must also note that our
382 theoretical approach is conservative as we have ignored the effects of selection against
383 introgressed alleles. We expect that linkage to loci underlying local adaptation should act to
384 decrease the probability of a phylogeny topology switch at the locus considered. As such, a
385 polygenic basis of local adaptation could greatly reduce the probability of a topology switch
386 due to gene flow. Overall, when considering our theoretical work in combination with
387 patterns of gene flow and genetic structure in the system, there is strong evidence that *S.*
388 *lautus* populations have evolved multiple independent times. Below we outline other lines of
389 evidence from our empirical work that support this assertion.

390 First, further evidence that gene flow has not obscured a single origin scenario in *S. lautus*
391 comes from comparing joint estimates of gene flow and divergence times (as implemented in
392 isolation with migration models) between population pairs of the same ecotype and putative
393 sister populations of divergent ecotypes. We observed that populations of the same ecotype
394 consistently show deeper divergence times than those from different ecotypes, which reflects
395 the topology of the phylogeny estimated in the absence of gene flow. In addition,
396 constructing the phylogeny considering gene flow (in *TreeMix*) did not alter the topology
397 from its estimation in the absence of gene flow. Although the divergences at two localities
398 (D04-H05 and D32-H12) experience levels of gene flow high enough to potentially result in
399 phylogenetic distortion, their levels of differentiation are rather high. Furthermore, each of
400 these pairs is from a separate clade and are genetically isolated from other such pairs, so even
401 moderate levels of gene flow within each pair would not have distorted the phylogeny across
402 the entire system. Even if we treat the divergences at these two localities (D04-H05 and D32-

H12) with some caution, transplant experiments within D04-H05 and other population pairs[40,45–49] have revealed strong extrinsic reproductive isolation, suggesting the barrier to gene flow is very strong between parapatric ecotypes. Together, these results also imply that phylogenetic distortion is highly unlikely in *S. lautus* and that such relationships reflect the true history of populations and ecotypes.

Additional support for the parallel evolution of *S. lautus* populations comes when considering our results in combination with previous work. Even if gene flow was high enough to distort the phylogeny across multiple populations, the clear association between environment and phenotype in the system in both common garden[51] and field conditions[50] as well as the strong divergent selection within each population pair[40,45–49], suggests that natural selection may have independently resisted the introduction of maladaptive alleles across parapatric populations. Similar phenotypes across replicate populations have also arisen via mutations in different genes[49,50,67], indicating that natural selection has necessarily acted independently within each population to drive the evolution of similar phenotypes. This adds further strength to our argument that observed levels of gene flow in *S. lautus* are not strong enough to obscure the historical relationships of populations. Overall, our results indicate that Dune and Headland populations have originated multiple independent times in parapatry with limited levels of gene flow which makes *S. lautus* a highly replicated system of parapatric divergences.

The *S. lautus* system allows us to study the deterministic nature of parallel evolution in multiple ways. For instance, we can now begin to understand how genetic architectures vary and evolve during adaptation. In doing so, researchers can then demonstrate whether alleles, genes or pathways have been repeatedly selected for across replicate populations[24,68]. This will also help us understand whether adaptation arises from new mutations or standing genetic variation (e.g., ref. [68]), or from fixation of functionally equivalent alleles (such as during polygenic adaptation[38,39,69]), or loss-of-function mutations (e.g., ref. [35]). Once adaptive genes have been identified, studies of parallel evolution should directly link the adaptive loci to phenotypic traits and further demonstrate that the traits themselves have been under repeated selection in independent populations[70–72]. In systems where this is not feasible, our study demonstrates that studying genome-wide loci can uncover patterns of phylogeography and migration that are consistent with parallel evolution.

Finally, in our work we have unusual high power to detect gene flow, as the number of individuals sequenced in each population is large (N_{mean} = 57, $2N_{mean}$ > 100 chromosomes per population). This sampling regime allowed us to sample many rare variants and therefore better distinguish ancestral polymorphism from migration. Studies undertaking demographic modelling often sample 10-25 individuals per population (e.g., refs. [22,73,74]) and occasionally even less than 10 (e.g., ref. [28]). Thus these studies cannot easily distinguish shared variants due to gene flow from ancestral polymorphism, which can make results biased to detecting moderate to high levels of gene flow, especially for recently diverged populations and in underpowered studies[64,75–77]. As our coalescent modelling reveals that little gene flow can obscure a phylogenetic topology under certain conditions (e.g., during very young adaptive radiations), studies that fail to detect gene flow with many numbers of individuals and loci can treat results with confidence.

Overall, we provide strong evidence for multiple origins of parapatric Dune and Headland populations within *S. lautus*. Across this highly replicated system we observed phylogenetic clustering by geography, strong genetic structure between populations, isolation by distance, and surprisingly low gene flow between parapatric populations at each locality as well as the system as a whole. Coalescent modelling confirmed that levels of gene flow are likely not high enough to obscure a single origin scenario. Furthermore, the phylogenetic relationships of populations estimated in the presence of gene flow agree with the main phylogeny, which supports a multiple origin scenario. These results from our current work in combination with strong divergent selection between ecotypes[40,45–49], strong trait-environment association in the system[50,51] and adaptation across replicate populations occurring mainly via mutations in different genes[49,50,67], implies that selection has independently driven the parallel evolution of populations. This makes *S. lautus* one of the clearest examples of the parallel evolution of ecotypes discovered yet, adding to the increasing number of potential cases of parallel evolution in plants[22,78–80]. It also positions the species as a powerful system of replicated parapatric divergence to study the origin of adaptations and reproductive isolation.

Methods

Sample collection and DNA extraction

Leaf samples for DNA extraction were collected from 23 Dune and Headland populations of *Senecio lautus* along the coast of Australia, which included eight parapatric Dune-Headland

population pairs, three allopatric Headland populations, and three allopatric Dune populations (n_{mean} = 58, n_{total} = 1338; Fig. 2A, Supplementary Table S5). We sampled mature (flowering) plants evenly across the geographic range of each population, ensuring that sampled plants were at least one meter apart. DNA was extracted using a modified CTAB protocol[81] and cleaned with Epoch Life Sciences spin columns. We quantified sample concentration with the Invitrogen Quant-iT PicoGreen dsDNA Assay Kit, and used the BioTek Take3 Micro-Volume Plate to ensure DNA samples were pure. Samples were standardized to 10ng/uL.

GBS library construction

We created reduced representation libraries to obtain restriction site associated DNA (RAD) markers. Specifically, we used a two-enzyme Genotyping-by-Sequencing (GBS) approach (modified from ref. [82]). We created seven libraries from the 23 Dune and Headland populations, each containing 192 barcoded individuals. For each individual, genomic DNA was digested with the restriction enzymes Pst1-HF (New England Biosciences; NEB) and Msp1 (NEB). Forward and reverse barcodes were ligated to fragments from each sample, and subsequently cleaned with homemade Serapure beads[83,84]. For each sample we amplified the fragments and added Illumina sequencing primers via PCRs. Each sample was quantified with the Invitrogen Quant-iT PicoGreen dsDNA Assay Kit. We created seven equimolar pools (192 individuals per pool), ensuring each population was evenly distributed across the pools. Each pool was size-selected on the BluePippin (2% DF Marker V1, 300-500bp; Sage Science), and cleaned with the Monarch PCR & DNA cleanup kit (NEB). Pooled libraries were sent to Beijing Genomics Institute for sequencing on seven lanes of the HiSeq4000, with 100bp paired-end sequencing.

Bioinformatics

The Beijing Genomics Institute removed forward barcodes and quality filtered the raw reads to remove reads containing Illumina adaptors, low quality reads (> 50% of bases < Q10), and reads with > 10% Ns. We trimmed reverse barcodes with *TagCleaner* standalone v0.12[85]. We retained an average of 2,849,159 clean reads (SD = 827,036) across the 1,319 individuals (after the removal of 19 individuals with high missing data, see below; Supplementary Table S6). Reads were mapped to the *S. lautus* reference PacBio genome v1.0[49] with *BWA-MEM* v0.7.15[86,87]. On average, 86% of reads (SD = 15) mapped to the reference genome, and 81% (SD = 15) mapped properly with their paired-read (Supplementary Table S6). *PicardTools*

v2.7.0[88] was used to clean aligned reads and to add read groups (PCR duplicates were not marked for removal). We jointly called all variant and invariant sites for each population with *FreeBayes* v1.1.0[89]. Because SNPs were separately called for each of the 23 populations, we first normalized the 23 VCF files before merging them together. This was achieved by first using BCFtools v1.4.1[90] to split multiallelic sites into biallelic records. Each file was then normalized by re-joining biallelic sites into multiallelic records. We then left-aligned and normalized indels, and used *vt*[91] to decompose biallelic block substitutions into separate SNPs for each population. We then merged the 23 per-population VCF files into one large file for subsequent SNP filtering.

We largely followed the *dDocent* pipeline for SNP filtering[92,93], including iterative filtering to maximize the number of sampled SNPs[94]. Using *VCFtools* v0.1.15[95], we first retained sites if they were present in > 50% of individuals, had a minimum quality score of 30, and a minimum minor allele count of 1. We then filtered for a minimum depth of 3 for a genotype call. Individuals were removed if they contained > 40% missing data. We then filtered for a maximum mean depth of 100, and a minimum mean depth of 10. We filtered for missing data per population, removing sites if they contained > 50% of missing data within each population. We refiltered for an overall missing data of 20%. Indels were removed with *vcflib*[96]. We then filtered for population-specific Hardy Weinberg Equilibrium using the *filter_hwe_by_pop.pl* script within *dDocent*. See below for the minor allele frequency thresholds for each analysis.

Do populations cluster by geography or ecotype?

To explore the broad patterns of genetic clustering of populations, we performed two separate analyses: phylogeny construction and *fastSTRUCTURE*[97]. We used *PLINK* v1.9[98] to filter for a minor allele frequency of 0.05 and also to thin SNPs by retaining one unlinked SNP per RAD locus. This dataset contained 3,844 unlinked SNPs across the 1,319 individuals. We generated a maximum likelihood phylogeny within *IQ-TREE* v1.6.0[99] using the polymorphisms-aware phylogenetic model[100]. We first used ModelFinder[101] to determine the best-fit substitution model for the data (TVMe+FQ+P+N9+G4), and increased the virtual population size (N) to the maximum value of 19 (as recommended by ref. [100]). Default parameters were used for tree construction, with the western Australia D09 population assigned as the outgroup. To assess convergence, we undertook 10 separate runs of *IQ-TREE* and examined tree topology (which remained unchanged with 10 independent runs). We also

ensured that the log-likelihood values were stable at the end of each run. Branch support was performed using 10,000 replicates of UFboot[102], and 10,000 replicates of SH-aLRT[103].

We further explored broad patterns of population structure using the variational Bayesian framework, *fastSTRUCTURE* v1.0[97]. Here, we implement *fastSTRUCTURE* as extra evidence for whether populations genetically cluster by geography or ecotype. We did not infer specific historical admixture scenarios from *fastSTRUCTURE*, as different demographic scenarios can give rise to indistinguishable structure plots[52]. The *fastSTRUCTURE* algorithm assigns individuals into genetic clusters (K) by minimizing departures from Hardy-Weinberg equilibrium and inferring individual ancestry proportions to each genetic cluster. We followed the recommendations by refs. [104,105]. We ran the simple prior (K=1-30) with 100 independent runs per K-value. In order to determine the most likely number of genetic clusters (the optimal K), we used the *chooseK.py* script from *fastSTRUCTURE* to examine (1) the K-value that best explained the structure in the data (the smallest number of model components that accounted for almost all of the ancestry in the sample), and (2) the K-value that maximized the marginal likelihood of the data. Results were summarized and plotted in the R package *pophelper* v2.2.7[106].

Has gene flow shaped patterns of divergence across the system?

To explore patterns of gene flow in a phylogenetic context, we used *TreeMix* v1.13[55]. *TreeMix* constructs a bifurcating maximum likelihood tree, identifies populations that are poor fits to the model, and sequentially adds migration events that improve the fit of the data. We filtered our data for MAF 0.01, retaining 24,933 SNPs across the 1,319 individuals. We constructed an initial 25 maximum likelihood trees with no migration, 1,000 bootstrap replicates in blocks of 50 SNPs with D09 as the assigned outgroup, and selected the tree with the highest log-likelihood as the input tree for all subsequent analyses. We then tested between 1-25 migration events in blocks of 50 SNPs. Trees and migration events were robust to varying the size of the linkage blocks as well as the MAF threshold of the dataset (data not shown). To select the number of migration events, we examined the log-likelihoods and cumulative variance explained by each model, as well as performed jackknife estimates to obtain the standard error and significance of the weight of each migration event. However, the interpretation of these P-values should be treated with caution due to possible errors in the tree structure as well as the inference of incorrect migration events[55].

To more formally test for admixture, we used the *threepop* function in *TreeMix* to calculate *f3*-statistics[107]. The *f3*-statistic determines whether a particular population (*A*) is the result of admixture between two other populations (*B* and *C*). It measures the difference in allele frequencies between populations *A* and *B*, and populations *A* and *C*. Only when admixture is present, we expect the allele frequency of population *A* to be intermediate between the allele frequencies of populations *B* and *C*. In contrast, in the absence of gene flow, population *A* allele frequency should not be consistently intermediate between *B* and *C*. Therefore, *f3* can be interpreted as the amount of shared genetic drift between two populations from a common ancestor. In the absence of admixture, *f3* (*A*; *B*, *C*) will be positive, whereas a significantly negative value of *f3* provides evidence for *A* being admixed from *B* and *C*. We calculated *f3* for all triads of populations with jackknifing in blocks of 50 SNPs to obtain Z-scores for calculating statistical significance (Z-score < -3.8 = P < 0.0001).

The erect phenotype is common across Australian species of the genus *Senecio*[43], except for the prostrate *S. lautus* Headland ecotype and a few Alpine populations, suggesting these prostrate forms are derived. We tested for isolation by distance (IBD[23]) in the ancestral and derived ecotypes to evaluate similarities in their dispersal dynamics[58]. We tested for IBD using migration rates (2*Nm*) inferred in *fastsimcoal2* (see below) as well as Slatkin's $\widehat{M}$, $(1 / F_{ST} - 1)/4$, as a proxy for gene flow[58]. For Slatkin's $\widehat{M}$, we excluded the western Australia populations (D09 and D35), filtered for a MAF of 0.05, and calculated pairwise F_{ST} in VCFtools. We calculated pairwise geographic distances using the following formula, which uses the spherical law of cosines to consider the curvature of the earth: $6378137*\arccos(\sin(lat_1)*\sin(lat_2)+\cos(lat_1)*\cos(lat_2)*\cos(long_1-long_2))$, where 6378137 is earth's radius in meters, and *lat* and *long* are the latitude and longitude (in radians) of the two populations compared. For the *fastsimcoal2* migration rates, we tested for IBD between the Dune and Headland of each population pair using a linear model in R v3.4.2[108], using an average of the bidirectional gene flow rates for each pair (log-log scale). For Slatkin's $\widehat{M}$, we also tested for IBD between the Dune and Headland of each population pair (log-log scale) using a linear model in R, and tested for IBD within the Dunes, and within the Headlands (log-log scale) using Mantel tests with 9,999 permutations in R (*mantel* function in the *vegan* package[109]).

Is there gene flow between parapatric populations?

We examined levels of admixture between parapatric populations with *STRUCTURE* v2.3.4[55]. *STRUCTURE* is a Bayesian MCMC approach that assigns populations into genetic clusters (K) based on individual genotypes by assuming Hardy-Weinberg Equilibrium within a population. It assigns each individual an admixture coefficient to depict the proportion of the genome that originated from a particular K cluster. To increase the numbers of SNPs, we took a subset of the data by excluding the two populations from the west coast of Australia (D09 and D35). Excluding these most divergent populations decreased the amount of missing data and thus increased the number of common SNPs in the south-eastern populations. We used the same filtering procedure as above, filtered for MAF 0.05 and thinned SNPs in *PLINK* to retain one SNP per RAD locus. Each population pair was extracted and subsequently filtered for MAF 0.05. We retained between 837 and 2,606 unlinked SNPs per pair (mean = 1,905 SNPs; SD = 575). *STRUCTURE* analysis was run using the admixture model and the correlated allele frequency model[110] with 10 independent runs for K=1-6 (50,000 burn-in and 200,000 MCMC). We ensured convergence of all summary statistics. As we were specifically interested in detecting admixed individuals between the two ecotypes, we plot results for K=2. To explore any additional genetic structure within a pair, we also estimated the optimal K-value with the Evanno method[56], by examining the maximum value for ΔK (the second order rate of change in the log probability of data between successive K-values). The R package *pophelper* was used to calculate the ΔK, summarize results and plot the data.

We directly estimated levels of gene flow between population pairs from the site frequency spectrum (SFS) using the composite-likelihood method implemented in *fastsimcoal2* v2.6.0.3[111]. The joint SFS of two populations is sensitive to demographic processes. For instance, gene flow will result in more low-frequency shared polymorphisms than expected under a non-migration scenario[112]. We tested eight demographic models (Fig. 4A), and inferred migration rates, as well as other demographic parameters including current population sizes, ancestral population size, divergence time, time of secondary contact, and gene flow cessation time, for eight Dune-Headland (DH) population pairs. We additionally asked whether gene flow was occurring in a linear fashion down the coast within each ecotype by testing eight Dune-Dune (DD) and eleven Headland-Headland (HH) pairs. To determine the baseline level of gene flow inferred by *fastsimcoal2* between isolated

621 populations, namely the null gene flow expectation, we estimated migration rates for three
622 very divergent allopatric populations (>1,500km apart, between the eastern and south-eastern
623 clades; D03-D32, D03-H12, and H02-H12), and took the highest detected migration rate
624 from these allopatric comparisons as the baseline migration rate.

625 As above, the western Australia populations (D09 and D35) were excluded from this dataset
626 to increase the number of sampled SNPs. For each pair, we filtered for a minor allele count of
627 one (MAC1), retaining between 6,679 and 19,951 variable sites per pair (mean = 12,155
628 SNPs, SD = 3,316). By using a MAC1 and a relatively high number of samples per
629 population (mean = 57, SD = 15), we retained rare alleles that are informative about
630 migration events between the populations[75]. Since we cannot distinguish ancestral from
631 derived alleles, we used the minor allele SFS (folded SFS). We used an *ad hoc* approach to
632 estimate the number of monomorphic sites (see Supplementary Methods section "Estimation
633 of monomorphic sites per pair"). Gene flow estimates were robust to varying the number of
634 monomorphic sites (data not shown). We used custom R functions (modified from ref. [113]) to
635 generate the joint folded SFS per population pair without downsampling. See Supplementary
636 Table S4 for details on the number of SNPs, number of monomorphic sites and models tested
637 for each pair comparison.

638 We performed 50 independent *fastsimcoal2* runs per model per population pair. Each run
639 consisted of 100,000 coalescent simulations and 40 expectation-maximization cycles for
640 parameter optimization. We used a mutation rate of 1.0×10^{-8} based on Asteraceae EST
641 sequence comparisons and fossil calibrations[114]. We ranked the models based on the
642 Kullback–Leibler information value which was estimated from the AIC scores of the best run
643 per model. Here, the normalization of the difference between the AIC scores of a particular
644 model and the best model in the set provides a measure of the degree of support for a
645 particular model, namely model likelihood (w_i)[115]. Since the use of linked-SNPs might lead to
646 pseudo-replication issues when comparing models based on *fastsimcoal2* likelihood values[116]
647 and the SFS discards linkage information, we verified SNPs were largely unlinked by
648 calculating linkage-disequilibrium in PLINK (data not shown).

649 As *fastsimcoal2* uses simulations to approximate the likelihood values, there is variance in
650 the likelihood estimates. To test whether the best model significantly differs from alternative
651 models with negligible gene flow ($2Nm = 0.01$) but the same values at other parameters, we
652 compared their likelihood distributions based on 100 expected SFS from 100,000 coalescent

simulations per model[116]. If likelihood distributions overlap, there is no significant differences between the fit of both models[28]. To obtain confidence intervals for all demographic parameters, we performed parametric bootstrapping. Given the parameter values of the best run of the best model, we simulated 100 SFS and re-estimated the parameter values from them. Each run consisted of 100,000 coalescent simulations and 30 expectation-maximization cycles. The parameter values of the best run of the best model were specified as initial values of each bootstrapping run. We computed the 95% confidence intervals of all parameters with the *groupwiseMean* function of *rcompanion* R package[117].

Is gene flow high enough to obscure a single origin scenario?

To better understand under what conditions gene flow can erode a signal of phylogenetic monophyly of each ecotype, we created a coalescent model to represent a single origin scenario of ecotypes (see Supplementary Methods section "Probabilities of gene flow distorting phylogeny topology" for full details). We assumed a species tree consisting of four populations, with two sets of sister taxa to represent populations of the same ecotype (Fig. 5A). The ancestor of the ecotypes splits at time *t2* in the past, and we can think of this split as an initial single origin of the Dune and Headland ecotypes. To represent two parapatric population pairs, each of these two ecotypes further split simultaneously at time *t1* in the past. We considered an instantaneous burst of migration from the ancestral ecotype populations into the derived ecotype populations at each locality by assuming that a fraction *m* of alleles in each derived ecotype population (10%) is replaced by migrant alleles from a parapatric population at time *tm* in the past (50,000 generations ago). We then considered sampling an allele from each of the four populations at the present, and conditioned on the migrant status of the sampled alleles to calculate coalescent probabilities of gene tree topologies that result in a grouping in which the two populations of the derived ecotype do not appear as sister taxa in the gene genealogy. These methods recapitulate recent more formal treatments of the probability of hemiplasy (non-monophyly despite a single evolutionary origin) under incomplete lineage sorting and introgression[118,119], though we have considered a scenario involving four populations to reflect the nature of parapatric pairs. Moreover, our emphasis is placed on the implications of gene flow for the original inference of the species tree itself, rather than how it pertains to the history of a selected locus of interest under an inferred phylogeny.

Finally, although we cannot directly map where our observed data fall in the simulated parameter space, we can gain further confidence on the likelihood of phylogenetic distortion by considering divergence time estimates from *fastsimcoal2* in combination with the phylogenetic topology estimated in *IQ-TREE*. More specifically, we asked whether divergence times between populations from the same ecotype are deeper than between populations from different ecotypes. The estimation of divergence times in *fastsimcoal2* considers gene flow, thus if they are in accordance with relative node order of the *IQ-TREE* phylogeny (which is estimated without accounting for gene flow), it suggests that phylogenetic distortion within the system is unlikely. We thus compared the *fastsimcoal2* divergence times to the relative node order of the *IQ-TREE* phylogeny for four population pairs (D04-H05 and D05-H06; D14-H15 and D32-H12). We selected these pairs because they represent neighboring sister-taxa within the phylogeny.

Acknowledgements

We are grateful to M.J. Wilkinson, H.N. Nguyen Vu and H.L. North for assisting with sample collection. We extend our thanks to R.K. Bagley for assisting with *fastsimcoal2*. We thank Ortiz-Barrientos laboratory members for insightful comments on previous versions of this manuscript. S. Chenoweth and M. Blows provided very useful feedback on M.E. James' PhD dissertation. This work was funded by an Australian Research Council grant (DP190103039) to D.O. and J.E., and a University of Queensland Graduate School International Travel Award to M.E.J.

Author contributions

M.E.J and D.O. conceived the project. M.E.J. and J.E. undertook sample collection. M.E.J. extracted DNA, prepared libraries, performed bioinformatics, and undertook the *IQ-TREE*, *fastSTRUCTURE*, *STRUCTURE* and *TreeMix* analyses. H.A. conducted the *fastsimcoal2* analyses. J.S.G. performed the coalescent modelling with input from J.E. The paper was written by M.E.J. and D.O. with input from all authors. D.O. is the mentor and supervisor for the research program.

Competing interests

The authors declare no competing interests.

Data availability

Data will be uploaded to Dryad upon acceptance of the manuscript.

References

1. Schluter, D. & Nagel, L. M. Parallel speciation by natural selection. *Am. Nat.* **146**, 292–301 (1995).
2. Berglund, A. B. N., Dahlgren, S. & Westerbergh, A. Evidence for parallel evolution and site-specific selection of serpentine tolerance in *Cerastium alpinum* during the colonization of Scandinavia. *New Phytol.* **161**, 199–209 (2003).
3. Johannesson, K. *et al.* Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philos. Trans. R. Soc. B* **365**, 1735–1747 (2010).
4. Elmer, K. R., Kusche, H., Lehtonen, T. K. & Meyer, A. Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes. *Philos. Trans. R. Soc. B* **365**, 1763–1782 (2010).
5. Ravinet, M., Prodöhl, P. A. & Harrod, C. Parallel and nonparallel ecological, morphological and genetic divergence in lake-stream stickleback from a single catchment. *J. Evol. Biol.* **26**, 186–204 (2013).
6. Perreault-Payette, A. *et al.* Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. *Mol. Ecol.* **26**, 1477–1497 (2017).
7. York, R. A. & Fernald, R. D. The repeated evolution of behavior. *Front. Ecol. Evol.* **4**, 143 (2017).
8. Smith, S. D. & Rausher, M. D. Gene loss and parallel evolution contribute to species difference in flower color. *Mol. Biol. Evol.* **28**, 2799–2810 (2011).
9. Orr, H. A. The probability of parallel evolution. *Evolution* **59**, 216–220 (2005).
10. Stern, D. L. & Orgogozo, V. Is genetic evolution predictable? *Science* **323**, 746–751 (2009).
11. Lenormand, T., Chevin, L. M. & Bataillon, T. Parallel evolution: what does it (not) tell us and why is it (still) interesting? in *Chance in Evolution* (University Chicago Press, Chicago, 2016).
12. Stoltzfus, A. & McCandlish, D. M. Mutational biases influence parallel adaptation. *Mol. Biol. Evol.* **34**, 2163–2172 (2017).
13. Yeaman, S., Gerstein, A. C., Hodgins, K. A. & Whitlock, M. C. Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genet.* **14**, e1007717 (2018).

14. Rosenblum, E. B., Parent, C. E. & Brandt, E. E. The molecular basis of phenotypic convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**, 203–226 (2014).
15. Blount, Z. D., Lenski, R. E. & Losos, J. B. Contingency and determinism in evolution: Replaying life's tape. *Science* **362**, eaam5979 (2018).
16. Bierne, N., Gagnaire, P. A. & David, P. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Curr. Zool.* **59**, 72–86 (2013).
17. Quesada, H., Posada, D., Caballero, A., Morán, P. & Rolán-Alvarez, E. Phylogenetic evidence for multiple sympatric ecological diversification in a marine snail. *Evolution* **61**, 1600–1612 (2007).
18. Butlin, R. K. *et al.* Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution* **68**, 935–949 (2014).
19. Pérez-Pereira, N., Quesada, H. & Caballero, A. Can parallel ecological speciation be detected with phylogenetic analyses? *Mol. Phylogenet. Evol.* **116**, 149–156 (2017).
20. Ostevik, K. L., Moyers, B. T., Owens, G. L. & Rieseberg, L. H. Parallel ecological speciation in plants? *Int. J. Ecol.* **2012**, 1–17 (2012).
21. Allender, C. J., Seehausen, O., Knight, M. E., Turner, G. F. & Maclean, N. Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. *PNAS* **100**, 14074–14079 (2003).
22. Trucchi, E., Frajman, B., Haverkamp, T. H. A., Schönswetter, P. & Paun, O. Genomic analyses suggest parallel ecological divergence in *Heliosperma pusillum* (Caryophyllaceae). *New Phytol.* **216**, 267–278 (2017).
23. Wright, S. Isolation by distance. *Genetics* **28**, 114–138 (1943).
24. Lee, K. M. & Coop, G. Population genomics perspectives on convergent adaptation. *Philos. Trans. R. Soc. B* **374**, 20180236 (2019).
25. Coyne, J. A. & Orr, H. A. *Speciation*. (Sinauer Associates, Sunderland, 2004).
26. Barton, N. H. & Hewitt, G. M. Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* **16**, 113–148 (1985).
27. Endler, J. *Geographic variation, speciation, and clines*. (Princeton University Press, Princeton, 1977).
28. Meier, J. I. *et al.* Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization. *Mol. Ecol.* **26**, 123–141 (2017).

29. Magalhaes, I. S. *et al.* Intercontinental genomic parallelism in multiple adaptive radiations. Preprint at https://doi.org/10.1101/856344. (2019) doi:10.1101/856344.
30. Le Moan, A., Gagnaire, P. A. & Bonhomme, F. Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Mol. Ecol.* **25**, 3187–3202 (2016).
31. Rougemont, Q. *et al.* Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys. *J. Evol. Biol.* **28**, 2248–2263 (2015).
32. Herman, A. *et al.* The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, *Astyanax mexicanus*. *Mol. Ecol.* **27**, 4397–4416 (2018).
33. Rougeux, C., Bernatchez, L. & Gagnaire, P. A. Modeling the multiple facets of speciation-with-gene-flow toward Inferring the divergence history of lake whitefish species pairs (Coregonus clupeaformis). *Genome Biol. Evol.* **9**, 2057–2074 (2017).
34. Rougeux, C., Gagnaire, P. A. & Bernatchez, L. Model-based demographic inference of introgression history in European whitefish species pairs'. *J. Evol. Biol.* **32**, 806–817 (2019).
35. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–305 (2010).
36. Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**, 1928–1933 (2005).
37. Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W. & Berner, D. The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol.* **23**, 3944–3956 (2014).
38. Yeaman, S. Local adaptation by alleles of small effect. *Am. Nat.* **186**, S74–S89 (2015).
39. Tiffin, P. & Ross-Ibarra, J. Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* **29**, 673–680 (2014).
40. Melo, M. C., Grealy, A., Brittain, B., Walter, G. M. & Ortiz-Barrientos, D. Strong extrinsic reproductive isolation between parapatric populations of an Australian groundsel. *New Phytol.* **203**, 323–334 (2014).
41. Roda, F. *et al.* Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. *Mol. Ecol.* **22**, 2941–2952 (2013).
42. Radford, I. J., Cousens, R. D. & Michael, P. W. Morphological and genetic variation in the *Senecio pinnatifolius* complex: are variants worthy of taxonomic recognition? *Aust. Syst. Bot.* **17**, 29–48 (2004).

43. Thompson, I. R. Taxonomic studies of Australian *Senecio* (Asteraceae): 5. The *S. pinnatifolius/S. lautus* complex. *Muelleria* **21**, 23–76 (2005).
44. Ali, S. *Senecio lautus* complex in Australia. I. Taxonomic considerations and discussion of some of the related taxa from New Zealand. *Aust. J. Bot.* **12**, 282–291 (1964).
45. Walter, G. M. *et al.* Diversification across a heterogeneous landscape. *Evolution* **70**, 1979–1992 (2016).
46. Richards, T. J. & Ortiz-Barrientos, D. Immigrant inviability produces a strong barrier to gene flow between parapatric ecotypes of *Senecio lautus*. *Evolution* **70**, 1239–1248 (2016).
47. Richards, T. J., Walter, G. M., McGuigan, K. & Ortiz-Barrientos, D. Divergent natural selection drives the evolution of reproductive isolation in an Australian wildflower. *Evolution* **70**, 1993–2003 (2016).
48. Walter, G. M., Wilkinson, M. J., Aguirre, J. D., Blows, M. W. & Ortiz-Barrientos, D. Environmentally induced development costs underlie fitness tradeoffs. *Ecology* **99**, 1391–1401 (2018).
49. Wilkinson, M. J. *et al.* Divergence in hormone signaling drives local adaptation and intrinsic reproductive isolation. Preprint at https://doi.org/10.1101/845354. (2020).
50. James, M. E., Wilkinson, M. J., North, H. L., Engelstädter, J. & Ortiz-Barrientos, D. A framework to quantify phenotypic and genotypic parallel evolution. Preprint at https://doi.org/10.1101/2020.02.05.936450. (2020).
51. Walter, G. M., Aguirre, J. D., Blows, M. W. & Ortiz-Barrientos, D. Evolution of genetic variance during adaptive radiation. *Am. Nat.* **191**, E108–E128 (2018).
52. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* **9**, 3258 (2018).
53. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
54. Melo, M. C. *et al.* Evidence for mutation-order speciation in an Australian wildflower. Preprint at https://doi.org/10.1101/692673. (2019).
55. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
56. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

57. Momigliano, P., Florin, A.-B. & Merilä, J. Biases in demographic modelling affect our understanding of the process of speciation. Preprint at https://doi.org/10.1101/2020.06.03.128298. (2020) doi:10.1101/2020.06.03.128298.
58. Slatkin, M. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279 (1993).
59. Bank, C., Bürger, R. & Hermisson, J. The limits to parapatric speciation: Dobzhansky–Muller Incompatibilities in a continent–island model. *Genetics* **191**, 845–863 (2012).
60. Blanckaert, A. & Hermisson, J. The limits to parapatric speciation II: Strengthening a preexisting genetic barrier to gene flow in parapatry. *Genetics* **209**, 241–254 (2018).
61. Martin, C. H. *et al.* Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution* **69**, 1406–1422 (2015).
62. Feder, J. L. & Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **64**, 1729–1747 (2010).
63. Feder, J. L., Egan, S. P. & Nosil, P. The genomics of speciation-with-gene-flow. *Trends in Genetics* **28**, 342–350 (2012).
64. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
65. Slatkin, M. Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* **16**, 393–430 (1985).
66. Schluter, D. *The ecology of adaptive radiation*. (Oxford University Press, Oxford, 2000).
67. Roda, F. *et al.* Convergence and divergence during the adaptation to similar environments by an Australian groundsel. *Evolution* **67**, 2515–2529 (2013).
68. Lee, K. M. & Coop, G. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics* **207**, 1591–1619 (2017).
69. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
70. Storz, J. F. & Wheat, C. W. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* **64**, 2489–2509 (2010).
71. Pardo-Diaz, C., Salazar, C. & Jiggins, C. D. Towards the identification of the loci of adaptive evolution. *Methods Ecol. Evol.* **6**, 445–464 (2015).

72. Hoban, S. *et al.* Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* **188**, 379–397 (2016).
73. Roesti, M., Kueng, B., Moser, D. & Berner, D. The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **6**, 8767 (2015).
74. Kautt, A. F., Machado-Schiaffino, G. & Meyer, A. Multispecies outcomes of sympatric speciation after admixture with the source population in two radiations of Nicaraguan crater lake cichlids. *PLoS Genet.* **12**, e1006157 (2016).
75. Slatkin, M. Rare alleles as indicators of gene flow. *Evolution* **39**, 53–65 (1985).
76. Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *PNAS* **104**, 2785–2790 (2007).
77. Strasburg, J. L. & Rieseberg, L. H. How robust are 'Isolation with Migration' analyses to violations of the IM model? A simulation study. *Mol. Biol. Evol.* **27**, 297–310 (2010).
78. Foster, S. A., McKinnon, G. E., Steane, D. A., Potts, B. M. & Vaillancourt, R. E. Parallel evolution of dwarf ecotypes in the forest tree *Eucalyptus globulus*. *New Phytol.* **175**, 370–380 (2007).
79. Cai, Z. *et al.* Parallel speciation of wild rice associated with habitat shifts. *Mol. Biol. Evol.* **36**, 875–889 (2019).
80. Konečná, V., Nowak, M. D. & Kolář, F. Parallel colonization of subalpine habitats in the central European mountains by *Primula elatior*. *Sci. Rep.* **9**, 3294 (2019).
81. Clarke, J. D. Cetyltrimethyl Ammonium Bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harb. Protoc.* **2009**, Pdb.prot5177 (2009).
82. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme Genotyping-by-Sequencing approach. *PLoS ONE* **7**, e32253 (2012).
83. Faircloth, B. & Glenn, T. *Serapure*. (https://ethanomics.files.wordpress.com/2012/08/serapure_v2-2.pdf, 2011).
84. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
85. Schmieder, R., Lim, Y. W., Rohwer, F. & Edwards, R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**, 341 (2010).
86. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).

87. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
88. Broad Institute. *PicardTools*. (GitHub Repository. http://broadinstitute.github.io/picard/, 2019).
89. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv. arXiv:1207.3907v2 [q-bio.GN] (2012).
90. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
91. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
92. Puritz, J. B., Hollenbeck, C. M. & Gold, J. R. *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**, e431 (2014).
93. Puritz, J. B. *et al.* Demystifying the RAD fad. *Mol. Ecol.* **23**, 5937–5942 (2014).
94. O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M. & Portnoy, D. S. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* **27**, 3193–3206 (2018).
95. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
96. Garrison, E. *Vcflib, a simple C++ library for parsing and manipulating VCF files.* (https://github.com/vcflib/vcflib, 2016).
97. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
98. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
99. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
100. Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A. & Kosiol, C. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J. Theor. Biol.* **407**, 362–370 (2016).
101. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

102. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
103. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
104. Gilbert, K. J. *et al.* Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Mol Ecol* **21**, 4925–4930 (2012).
105. Janes, J. K. *et al.* The $K = 2$ conundrum. *Mol Ecol* **26**, 3594–3602 (2017).
106. Francis, R. M. POPHELPER: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2017).
107. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
108. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/, 2017).
109. Blanchet, G. *et al. vegan: Community Ecology Package.* (R package version 2.5-2. https://CRAN.R-project.org/package=vegan, 2018).
110. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
111. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
112. Hahn, M. W. *Molecular population genetics*. (Sinauer Associates, Sunderland, 2018).
113. Liu, S., Ferchaud, A. L., Grønkjaer, P., Nygaard, R. & Hansen, M. M. Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Mol. Ecol.* **27**, 4725–4743 (2018).
114. Strasburg, J. L. & Rieseberg, L. H. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*-large effective population sizes and rates of long-term gene flow. *Evolution* **62**, 1936–1950 (2008).
115. Thomé, M. T. C. & Carstens, B. C. Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *PNAS* **113**, 8010–8017 (2016).
116. Bagley, R. K., Sousa, V. C., Niemiller, M. L. & Linnen, C. R. History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*). *Mol. Ecol.* **26**, 1022–1044 (2017).

117. Mangiafico, S., S. *An R companion for the handbook of biological statistics*. (https://rcompanion.org/rcompanion/, 2015).
118. Guerrero, R. F. & Hahn, M. W. Quantifying the risk of hemiplasy in phylogenetic inference. *Proc Natl Acad Sci USA* **115**, 12787–12792 (2018).
119. Hibbins, M. S., Gibson, M. J. S. & Hahn, M. W. Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. Preprint at https://doi.org/10.1101/2020.04.15.043752. (2020).

Figures

Fig. 1 The colonization history and phylogenetic topology for alternate origin scenarios

Schematic diagram representing the colonization history and phylogenetic topology of two ecotypes (dark green and light green) from an ancestral population (grey) for three origin scenarios. Solid arrows depict the sequence of colonization. Double headed dotted arrows represent gene flow (*m*) between the ecotypes within each locality. L1, L2 and L3 represent three geographically distant localities, where a population of each ecotype resides. **(A)** Within a single origin scenario, the two ecotypes arise once from the ancestor, followed by range expansion. In the absence of gene flow, ecotypes form monophyletic clades within the phylogeny. **(B)** The single origin with gene flow scenario involves gene flow upon secondary contact between the ecotypes within each locality. Here, the observed phylogenetic topology shows populations clustering according to their geographic distribution. **(C)** Within a multiple origin scenario, the ancestral (dark green) ecotype arises once from the ancestor followed by range expansion, with the derived (light green) populations independently arising from each dark green population. Populations phylogenetically cluster according to their geographic distribution, which can be indistinguishable from a single origin with gene flow scenario **(B)**.

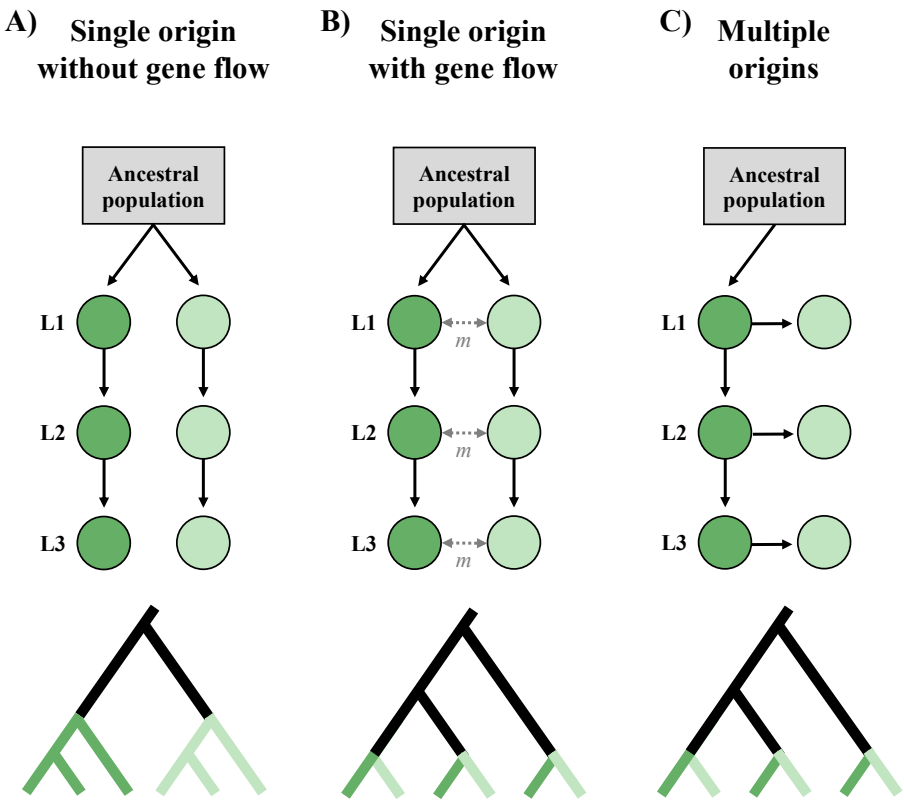


Fig. 2 Sampling locations and genetic clustering of *Senecio lautus* populations

(A) Sampling locations of the 23 Dune (orange) and Headland (green) *Senecio lautus* populations along the coast of Australia. **(B)** Maximum likelihood phylogeny of Dune and Headland populations implemented in *IQ-TREE*. Numbers on each node represent the SH-alRT support (%), followed by the ultrafast bootstrap support (%). **(C)** Bayesian assignment of individuals to genetic clusters within *fastSTRUCTURE* for K=2-4. Each of the 1,319 individuals is depicted as a bar, with colors representing ancestry proportions to each cluster. Populations are ordered according to their geographic distribution along the coast.

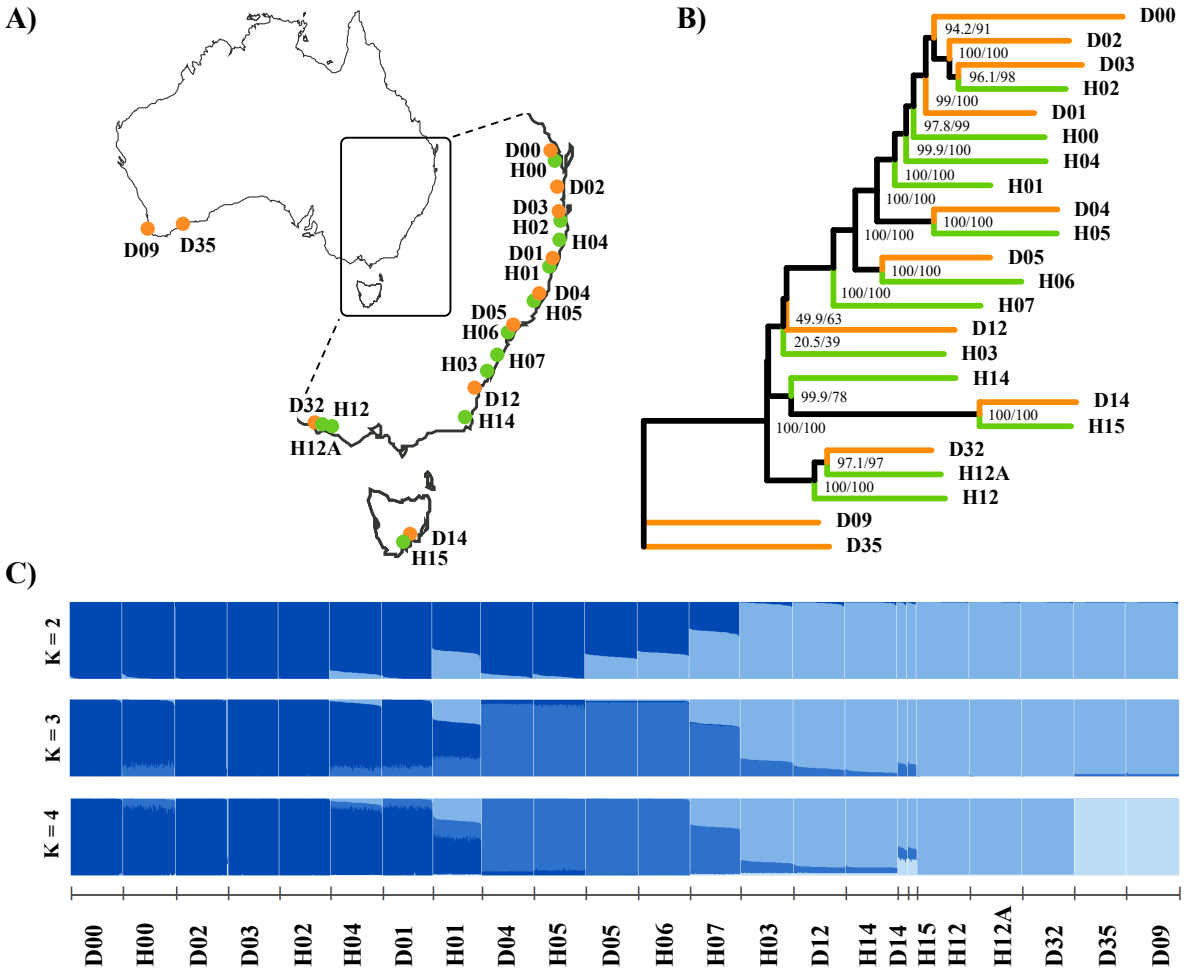


Fig. 3 Patterns of long-distance gene flow, IBD, and genetic clustering

(A) Maximum likelihood tree with one migration event inferred in *TreeMix*, the x-axis representing genetic drift. The red arrow represents the migration event ($w = 0.40$). **(B)** Patterns of isolation by distance across Dune and Headland populations for Dune-Headland (DH, black), Dune-Dune (DD, orange) and Headland-Headland (HH, green) pairs. Average migration rate is the mean bidirectional migration for each pair estimated in *fastsimcoal2*. Grey shading represents the null expectation for migration rates, inferred from the maximum migration value from three allopatric comparisons. Grey horizontal dashed line represents one migrant per generation ($2Nm = 1$). Pairs falling above this line are labelled. Black dashed line represents the linear model for the DH comparisons. **(C)** Patterns of isolation by distance using Slatkin's $\widehat{M}$ for parapatric Dune-Headland (black), Dune-Dune (orange) and Headland-Headland (green) pairs. Black, orange and green dashed line represent the linear model for the DH, DD and HH comparisons respectively. **(D)** Bayesian assignment of individuals to genetic clusters within *fastSTRUCTURE* for K=23. Each of the 1,319 individuals is depicted as a bar, with colors representing ancestry proportions to each cluster. Populations are ordered according to their geographic distribution along the coast.

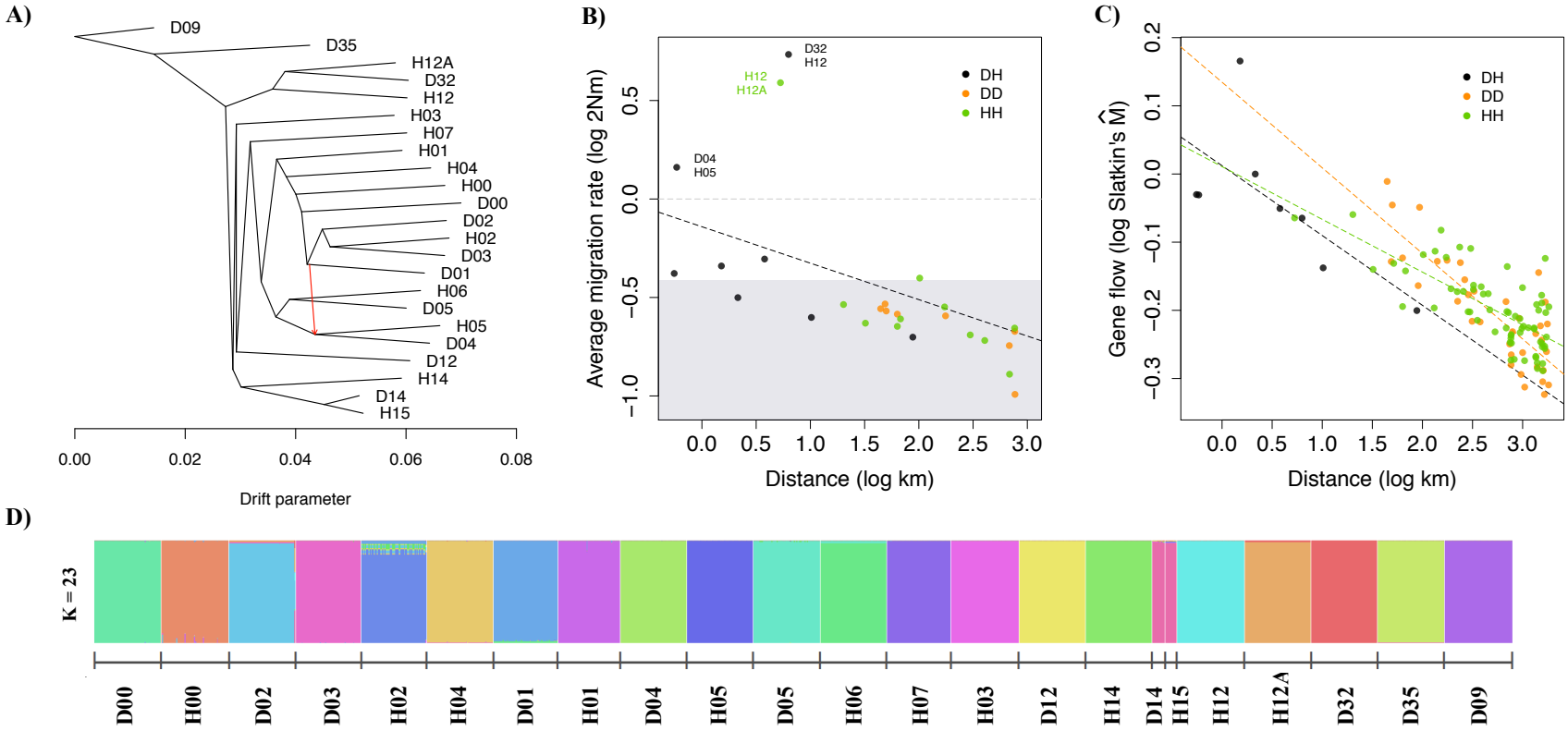


Fig. 4 Patterns of gene flow and admixture between parapatric Dune-Headland populations

(A) Schematic diagram representing the eight demographic models run in *fastsimcoal2* and their estimated parameters: no migration, bidirectional migration, Dune to Headland migration, Headland to Dune migration, bidirectional migration after secondary contact, Dune to Headland migration after secondary contact, Headland to Dune migration after secondary contact, bidirectional migration after population splitting with cessation of gene flow. **(B)** Bayesian assignment of individuals to genetic clusters within *STRUCTURE* for K=2 for the Dune (dark green) and Headland (light green) ecotypes at each locality. Each individual is depicted as a bar, with colors representing ancestry proportions to each cluster. Below are the migration rates (*m*, forward in time) from the Dune to Headland, and Headland to Dune within each locality estimated within *fastsimcoal2*. Asterisks denote pairs with $2Nm > 1$.

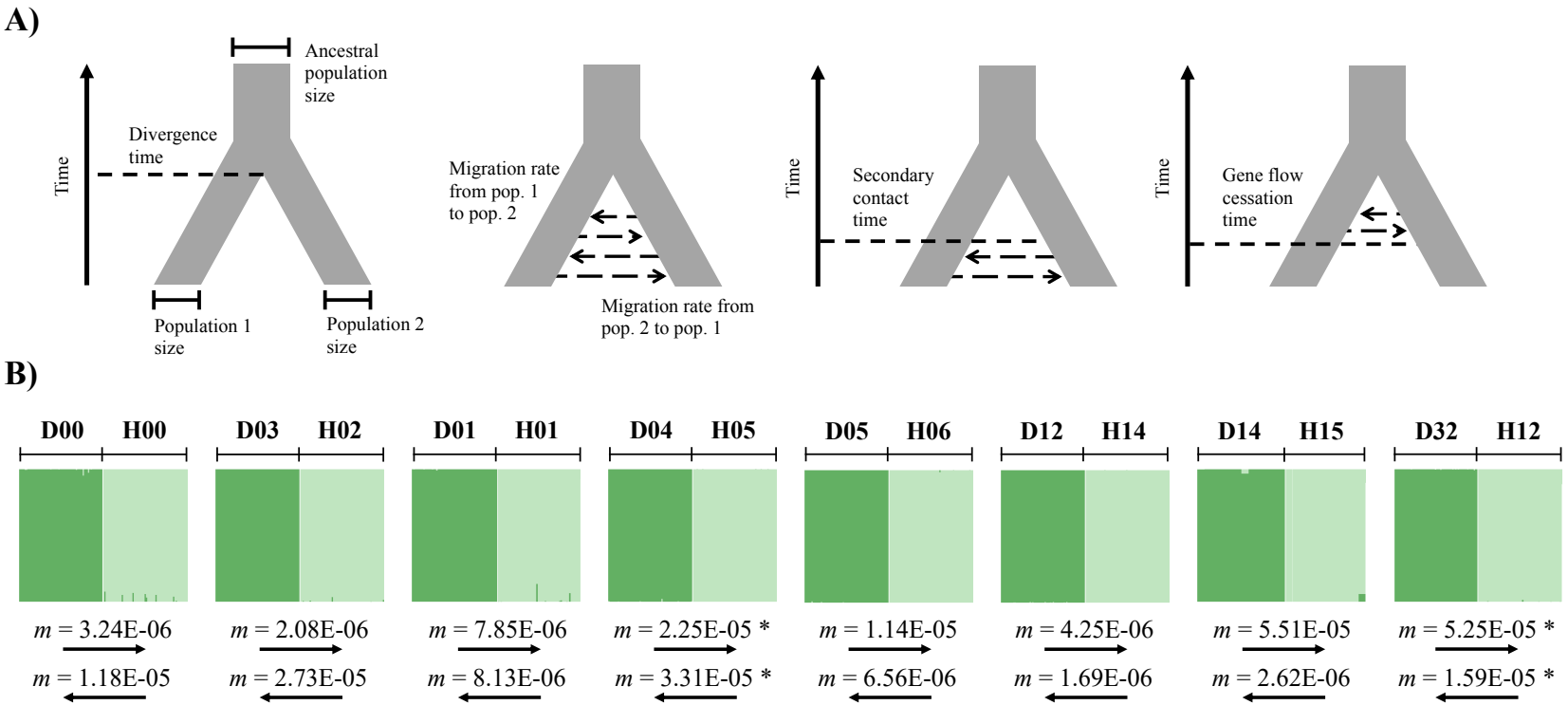
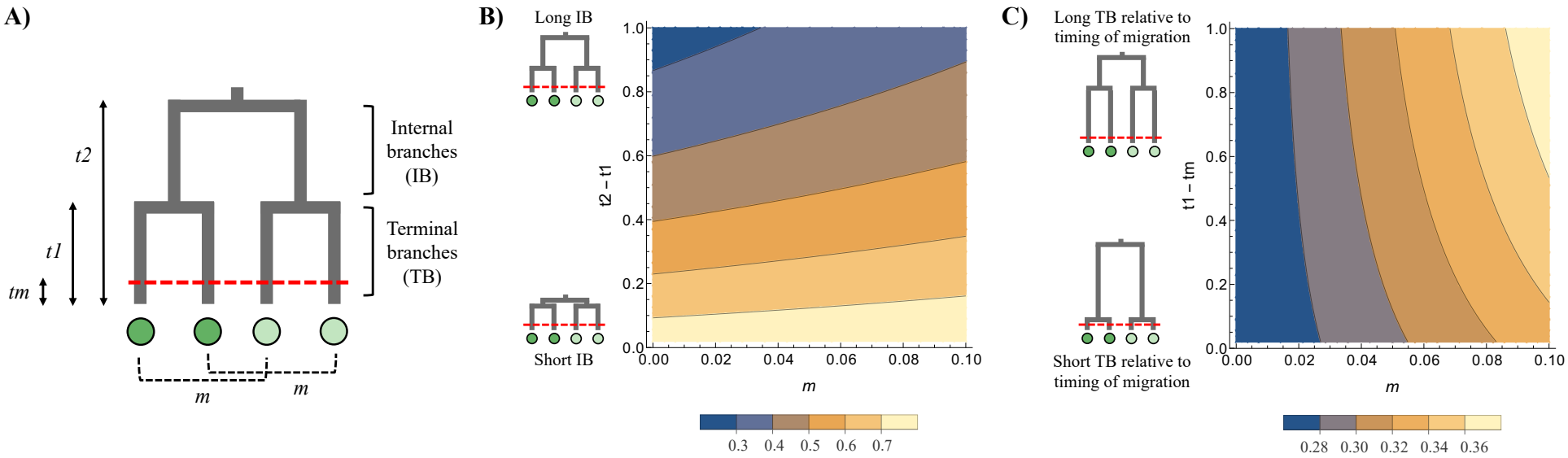


Fig. 5 Coalescent modelling to infer the probability of phylogenetic distortion

(A) Schematic diagram representing the modelled single origin scenario. *t2* represents the time to the split of the ancestral population (i.e., the initial origin of the Dune and Headland ecotypes). Each of these two ecotypes further split at time *t1* in the past. Parapatric populations at each locality are connected with dashed lines, and an instantaneous burst of migration (*m*) occurs at time *tm* in the past (dashed horizontal red line). In the model, all times are expressed in units of 2*N* generations. Light green and dark green circles represent populations from different ecotypes. **(B, C)** Probability that the phylogenetic topology of the single origin scenario is distorted, falsely suggesting the parallel evolution of ecotypes. Population size is set to 250,000 and *tm* is 0.1 (corresponding to 50,000 generations). Small phylogenies are schematic diagrams of the extreme values of the y-axis **(B)** High probability of phylogenetic distortion occurs when internal branches (IB) are short (lower yellow region). *t1* is set to 1 (corresponding to 500,000 generations). **(C)** The probability of phylogenetic distortion requires high migration and increases when length of terminal branches (TB) are longer prior to the burst of migration (right-hand yellow region). *t2* is set to 1 (corresponding to 500,000 generations). Note the different scale of probability in panel **(C)** compared to panel **(B)**.



Supplementary information

Supplementary Table S1. Pairwise F_{ST} values for *S. lautus* populations

Pairwise FST values between all 21 populations of the eastern and south-eastern clades.

	D00	**D01**	**D02**	**D03**	**D04**	**D05**	**D12**	**D14**	**D32**	**H00**	**H01**	**H02**	**H03**	**H04**	**H05**	**H06**	**H07**	**H12**	**H12A**	**H14**	**H15**
D00	-																				
D01	0.25	-																			
D02	0.25	0.22	-																		
D03	0.27	0.22	0.20	-																	
D04	0.29	0.25	0.26	0.28	-																
D05	0.29	0.25	0.27	0.27	0.25	-															
D12	0.34	0.29	0.31	0.33	0.31	0.28	-														
D14	0.34	0.28	0.29	0.31	0.29	0.26	0.32	-													
D32	0.34	0.30	0.33	0.34	0.32	0.30	0.30	0.32	-												
H00	0.26	0.23	0.24	0.25	0.26	0.26	0.29	0.28	0.31	-											
H01	0.26	0.22	0.25	0.25	0.25	0.24	0.26	0.25	0.28	0.23	-										
H02	0.26	0.21	0.21	0.20	0.27	0.27	0.31	0.30	0.33	0.25	0.25	-									
H03	0.33	0.29	0.31	0.32	0.30	0.27	0.28	0.30	0.30	0.29	0.25	0.31	-								
H04	0.28	0.23	0.25	0.26	0.27	0.26	0.30	0.29	0.31	0.24	0.22	0.26	0.29	-							
H05	0.29	0.25	0.27	0.28	0.21	0.26	0.31	0.30	0.32	0.27	0.25	0.27	0.30	0.27	-						
H06	0.30	0.27	0.28	0.29	0.27	0.21	0.30	0.28	0.31	0.27	0.24	0.28	0.28	0.27	0.28	-					
H07	0.31	0.27	0.28	0.29	0.28	0.24	0.28	0.28	0.30	0.27	0.24	0.29	0.27	0.27	0.28	0.26	-				
H12	0.35	0.31	0.33	0.34	0.33	0.30	0.31	0.32	0.22	0.31	0.28	0.33	0.30	0.31	0.33	0.32	0.30	-			
H12A	0.34	0.30	0.33	0.33	0.32	0.30	0.29	0.31	0.20	0.31	0.27	0.32	0.29	0.31	0.32	0.32	0.30	0.22	-		
H14	0.34	0.30	0.31	0.33	0.31	0.28	0.28	0.30	0.30	0.30	0.27	0.32	0.28	0.29	0.32	0.30	0.28	0.31	0.30	-	
H15	0.34	0.28	0.29	0.31	0.30	0.26	0.32	0.15	0.32	0.28	0.25	0.30	0.30	0.29	0.31	0.28	0.28	0.32	0.30	0.30	-

Supplementary Table S2. Estimation of gene flow and other demographic parameters in *fastsimcoal2*

Populations: the two populations used for each comparison (population 1 is on the left, and population 2 on the right). Asize: ancestral effective population size. Pop1size: effective population size of population 1. Pop2size: effective population size of population 2. DivTime: divergence time. SecTime: time since gene flow upon secondary contact. 2*Nm*P1->P2: migration rate (2*Nm*) from population 1 to population 2. 2*Nm*P2->P1: migration rate (2*Nm*) from population 2 to population 1. Migration rates are forward in time. Values in bold represent 2Nm > 1.

Comparison	Populations	Asize	Pop1size	Pop2size	DivTime	SecTime	2*Nm*P1->P2	2*Nm*P2->P1
Dune-Headland	D00-H00	100497	47926	134364	71945	18690	0.2176	0.2830
	D03-H02	88035	34637	152616	44190	15031	0.1590	0.4722
	D01-H01	72385	90270	159101	71918	13268	0.6241	0.3671
	D04-H05	87603	90859	123949	30707	6329	**1.3942**	**1.5024**
	D05-H06	97653	131873	70970	67927	16810	0.4049	0.4325
	D12-H14	56510	211701	103102	110018	11783	0.2188	0.1787
	D14-H15	102574	39573	143420	47929	39730	0.3952	0.5187
	D32-H12	56568	661726	212041	38706	11290	**5.5694**	**5.2694**
Dune-Dune	D00-D02	97055	63843	113624	52711	6436	0.3242	0.2617
	D01-D03	99168	142624	51613	58652	23772	0.3280	0.2119
	D01-D04	92638	121936	74257	66970	11319	0.2901	0.2200
	D02-D03	93440	116800	48492	54857	23163	0.3514	0.2029
	D04-D05	78638	86635	110138	77895	18111	0.3178	0.2027
	D05-D12	35322	103044	223346	128159	21689	0.1562	0.2041
	D12-D14	22223	259172	38450	118024	35758	0.0991	0.1046
	D14-D32	47348	27002	641721	56330	12595	0.3179	0.1074
Headland-Headland	H00-H02	97070	107516	94259	81290	9622	0.3920	0.4012
	H01-H04	78784	171269	86431	81443	19633	0.2561	0.3261
	H01-H05	61893	173061	89468	95145	17016	0.2546	0.3116
	H02-H04	78009	107213	91698	87055	20222	0.2913	0.1768
	H03-H07	57850	147099	125603	109197	12874	0.1904	0.1921
	H03-H14	63207	157400	119559	108683	9099	0.2068	0.2012
	H05-H06	84257	109077	88382	81710	10907	0.2559	0.1953
	H06-H07	67117	89121	141737	88627	14422	0.2532	0.2387
	H12-H12A	52196	286574	322443	43928	14082	**3.7584**	**4.0315**
	H12-H15	46457	362929	51902	81116	9800	0.1921	0.2501
	H14-H15	46091	168257	72243	101092	18596	0.1396	0.1181
Allopatric	D03-D32	35657	53174	566115	76621	5201	0.3873	0.3238
	D03-H12	37227	67316	333665	99511	8840	0.1984	0.2642
	H02-H12	33876	78181	313687	111278	9257	0.1884	0.2707

Supplementary Table S3. Confidence intervals for gene flow estimates inferred in *fastsimcoal2*

95% confidence intervals (CI) for migration rates inferred from 100 bootstrap runs in *fastsimcoal2*. Populations: the two populations used for each comparison; population 1 (*P1*) is on the left, and population 2 (*P2*) on the right. *2NmP1->P2min* and *max* are the lower and upper 95% CI for migration rate (2N*m*) from *P1* to *P2*, respectively. *2NmP2->P1min* and *max* are the lower and upper 95% CI for migration rate from *P2* to *P1*, respectively. Migration rates are forward in time. Populations in bold represent 2N*m* > 1.

Comparison	Populations	2*Nm*P1->P2min	2*Nm*P1->P2max	2*Nm*P2->P1min	2*Nm*P2->P1max
Dune-Headland	D00-H00	0.2181	0.2281	0.2769	0.2865
	D03-H02	0.1511	0.1611	0.4576	0.4758
	D01-H01	0.5991	0.6174	0.3554	0.3655
	D04-H05	**1.3120**	**1.3648**	**1.4347**	**1.4903**
	D05-H06	0.3924	0.4054	0.4172	0.4321
	D12-H14	0.2174	0.2236	0.1775	0.1840
	D14-H15	0.3776	0.4010	0.5001	0.5215
	D32-H12	**4.9046**	**5.0810**	**5.2088**	**5.3999**
Dune-Dune	D00-D02	0.3186	0.3344	0.2576	0.2686
	D01-D03	0.3183	0.3305	0.2040	0.2137
	D01-D04	0.2818	0.2911	0.2148	0.2226
	D02-D03	0.3417	0.3563	0.2000	0.2096
	D04-D05	0.3106	0.3228	0.2008	0.2078
	D05-D12	0.1533	0.1584	0.2020	0.2077
	D12-D14	0.0976	0.1011	0.1037	0.1068
	D14-D32	0.3109	0.3274	0.1050	0.1106
Headland-Headland	H00-H02	0.3828	0.3952	0.3901	0.4038
	H01-H04	0.2504	0.2574	0.3183	0.3290
	H01-H05	0.2485	0.2558	0.3033	0.3139
	H02-H04	0.2892	0.2978	0.1741	0.1802
	H03-H07	0.1882	0.1945	0.1917	0.1975
	H03-H14	0.2029	0.2089	0.1976	0.2038
	H05-H06	0.2526	0.2610	0.1926	0.1992
	H06-H07	0.2460	0.2537	0.2310	0.2382
	H12-H12A	**3.6292**	**3.7754**	**3.9901**	**4.1278**
	H12-H15	0.1878	0.1954	0.2484	0.2574
	H14-H15	0.1357	0.1415	0.1163	0.1202
Allopatric	D03-D32	0.3699	0.3893	0.3174	0.3273
	D03-H12	0.1932	0.2013	0.2604	0.2680
	H02-H12	0.1832	0.1900	0.2645	0.2731

Supplementary Table S4. Parameter estimates for all tested models in *fastsimcoal2*

(see excel file for Supplementary Table S4)

Supplementary Table S5. Sampling locations

Sampling locations of the 23 *Senecio lautus* Dune and Headland populations. Coordinates represent the mid-point of each population. N corresponds to the final number of individuals after removing individuals with low coverage. Parapatric pairs in bold are sister-taxa within the phylogeny. H12A is a population found within an ecotone between the Dune (D32) and Headland (H12) at this locality.

Clade	Population code	Location	Ecotype	Pair	Coordinates	N
Eastern	D00	QLD: Stradbroke Island	Dune	D00-H00	S27° 31.153' E153° 30.189'	62
Eastern	H00	QLD: Stradbroke Island	Headland	D00-H00	S27° 26.140' E153° 32.749'	63
Eastern	D02	QLD: Southport	Dune	-	S27° 56.846' E153° 25.736'	62
Eastern	D03	NSW: Cabarita	Dune	**D03-H02**	S28° 19.794' E153° 34.264'	61
Eastern	H02	NSW: Cabarita	Headland	**D03-H02**	S28° 21.013' E153° 34.676'	61
Eastern	H04	NSW: Byron Bay	Headland	-	S28° 38.060' E153° 38.268'	62
Eastern	D01	NSW: Lennox Head	Dune	D01-H01	S28° 46.858' E153° 35.655'	60
Eastern	H01	NSW: Lennox Head	Headland	D01-H01	S28° 48.813' E153° 36.313'	58
Eastern	D04	NSW: Coffs Harbour	Dune	**D04-H05**	S30° 18.946' E153° 08.142'	62
Eastern	H05	NSW: Coffs Harbour	Headland	**D04-H05**	S30° 18.741' E153° 08.676'	62
Eastern	D05	NSW: South West Rocks	Dune	**D05-H06**	S30° 53.027' E153° 04.037'	62
Eastern	H06	NSW: South West Rocks	Headland	**D05-H06**	S30° 52.710' E153° 04.549'	62
South-eastern	H07	NSW: Port Macquarie	Headland	-	S31° 28.526' E152° 56.219	60
South-eastern	H03	NSW: Kiama	Headland	-	S34° 40.301' E150° 51.704'	63
South-eastern	D12	NSW: Bermagui	Dune	D12-H14	S36° 28.346' E150° 03.581'	62
South-eastern	H14	NSW: Green Cape	Headland	D12-H14	S37° 15.748' E150° 02.991'	62
South-eastern	D32	VIC: Cape Bridgewater	Dune	**D32-H12**	S38° 19.631' E141° 23.772'	62
South-eastern	H12	VIC: Cape Bridgewater	Headland	**D32-H12**	S38° 22.728' E141° 22.018'	63
South-eastern	H12A	VIC: Cape Bridgewater	Intermediate	-	S38° 20.282' E141° 23.896'	62
South-eastern	D14	TAS: Port Arthur	Dune	**D14-H15**	S43° 10.550' E147° 51.267'	12
South-eastern	H15	TAS: Port Arthur	Headland	**D14-H15**	S43° 11.240' E147° 50.672'	11
Western	D35	WA: Isthmus Hill	Dune	-	S35° 05.885' E117° 59.182'	62
Western	D09	WA: Leeuwin-Naturaliste National Park	Dune	-	S33° 46.239' E114° 59.541'	63

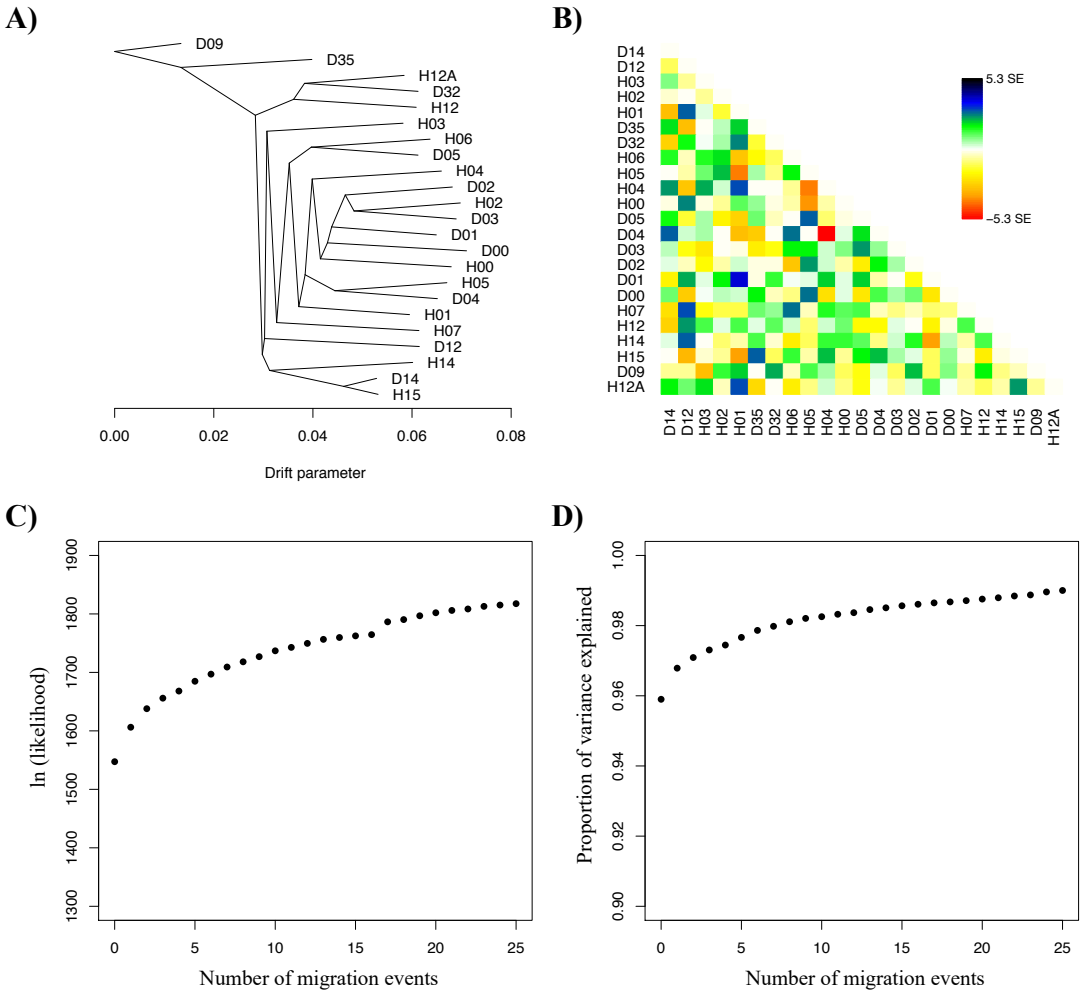
Supplementary Table S6. Sequencing and alignment summary for *Senecio lautus* populations

Summary statistics for the 23 populations used within the study. The 19 individuals removed due to high missing data are not included.

Population code	Mean # clean reads (range)	Mean % mapped reads (range)	% mapped reads properly paired (range)
D00	2,138,896 (971,466 - 3,506,240)	94 (62 - 98)	92 (61 - 96)
H00	3,075,580 (1,528,536 - 6,198,407)	81 (16 - 97)	79 (16 - 95)
D02	2,714,361 (895,858 - 5,258,091)	80 (18 - 96)	76 (17 - 94)
D03	3,160,935 (2,015,566 - 8,748,545)	84 (21 - 97)	78 (20 - 95)
H02	2,772,081 (1,408,465 - 4,192,718)	85 (34 - 96)	83 (33 - 94)
H04	3,176,210 (1,695,120 - 5,950,574)	90 (72 - 97)	79 (60 - 95)
D01	3,061,253 (1,318,262 - 4,548,766)	96 (83 - 98)	90 (72 - 96)
H01	2,770,561 (1,105,881 - 6,164,034)	93 (42 - 98)	91 (36 - 96)
D04	2,922,712 (2,146,253 - 3,718,635)	91 (62 - 98)	83 (61 - 96)
H05	2,866,233 (1,754,603 - 4,696,562)	92 (71 - 97)	85 (67 - 95)
D05	2,854,456 (1,554,814 - 4,156,601)	93 (48 - 97)	87 (44 - 94)
H06	2,112,573 (1,253,010 - 3,538,428)	84 (37 - 97)	82 (36 - 95)
H07	3,116,096 (1,646,581 - 10,437,355)	82 (27 - 98)	73 (21 - 96)
H03	2,795,169 (1,593,958 - 5,514,042)	77 (15 - 97)	76 (14 - 95)
D12	2,700,235 (1,448,045 - 5,032,607)	90 (45 - 98)	83 (39 - 94)
H14	3,033,007 (1,661,205 - 8,349,758)	71 (11 - 96)	67 (11 - 95)
D32	2,854,449 (1,517,908 - 5,609,011)	79 (19 - 97)	76 (17 - 95)
H12	2,892,473 (1,220,369 - 4,774,451)	83 (34 - 97)	80 (33 - 94)
H12A	2,614,734 (1,509,934 - 8,120,979)	85 (27 - 98)	82 (27 - 95)
D14	2,894,283 (1,704,586 - 4,893,613)	94 (75 - 98)	85 (58 - 95)
H15	3,229,783 (1,823,447 - 4,958,055)	90 (33 - 97)	84 (29 - 94)
D35	2,987,725 (1,754,767 - 6,004,276)	90 (62 - 98)	78 (44 - 95)
D09	3,008,471 (1,794,627 - 4,826,686)	67 (21 - 96)	63 (20 - 92)

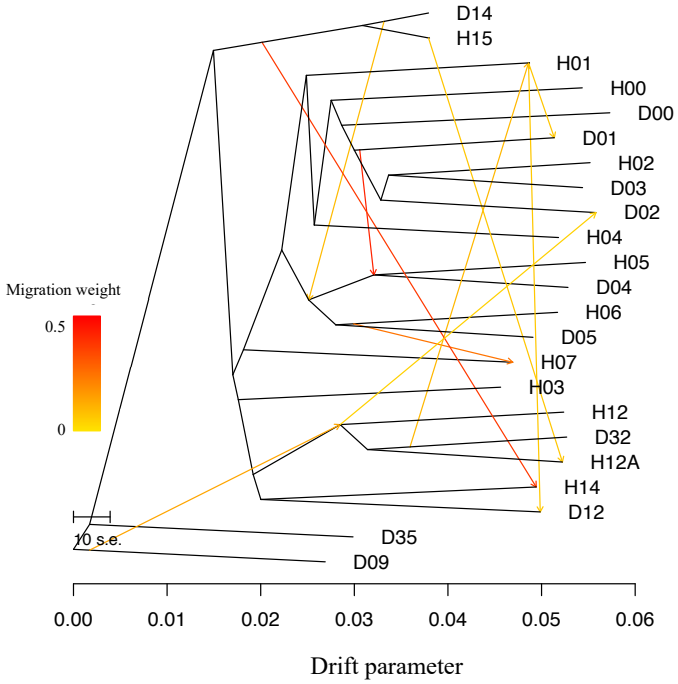
Supplementary Fig. S1 Summary of TreeMix runs

(A) Maximum likelihood tree with no migration. **(B)** Residuals for the no migration tree. **(C)** Log-likelihoods for each model for 1-25 migration events. **(D)** Proportion of variance explained by each model for 1-25 migration events.



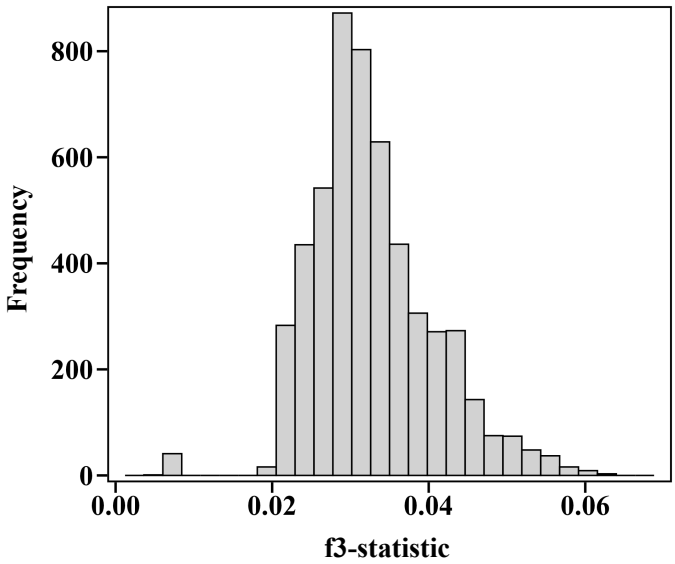
Supplementary Fig. S2 TreeMix migration events 1-10

Maximum likelihood tree with 10 migration events. Colored arrows denote the intensity and direction of migration events.



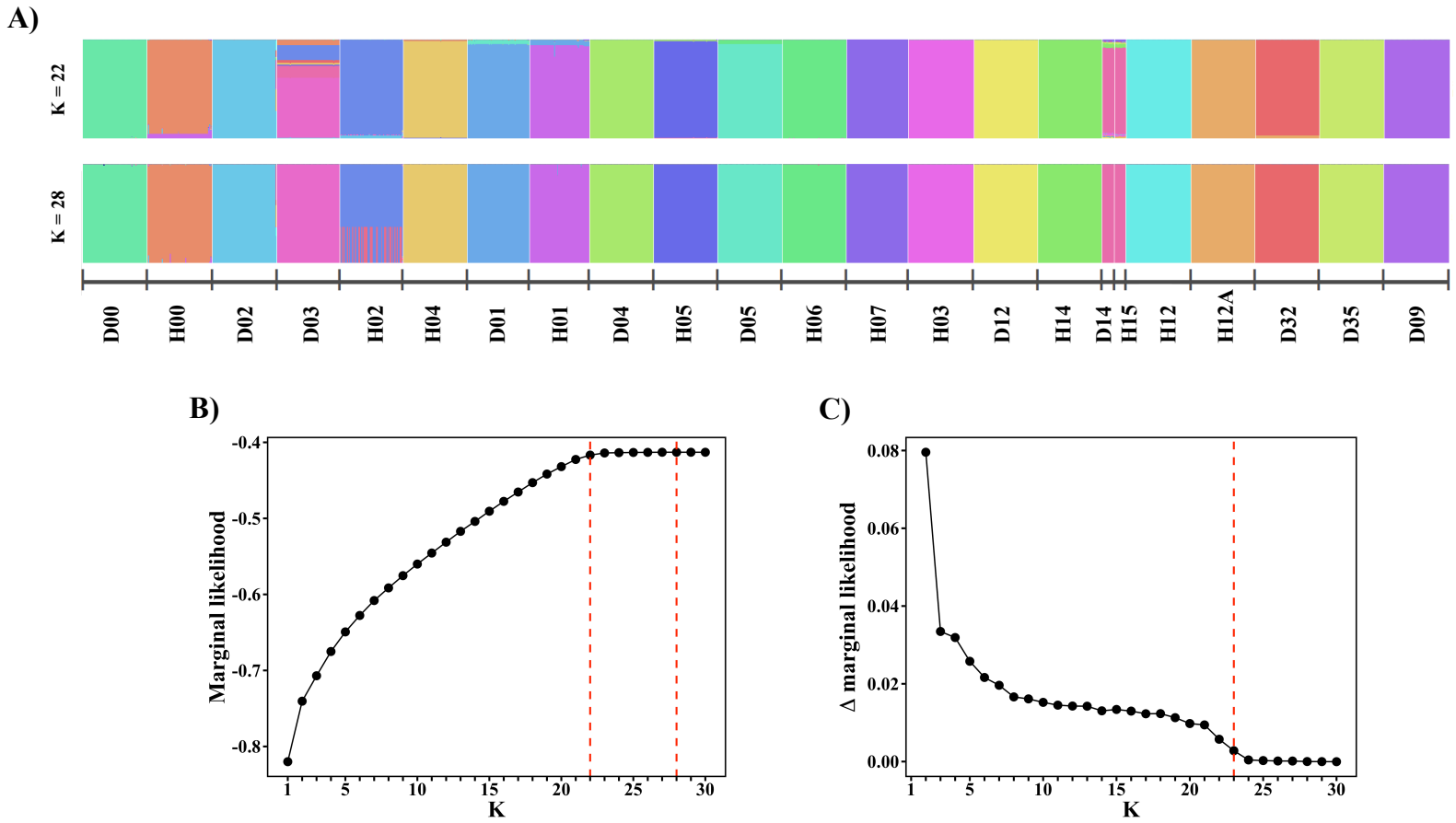
Supplementary Fig. S3 Frequency distribution of *f3*-statistics

Frequency distribution of *f3*-statistics calculated in *TreeMix* across all populations.



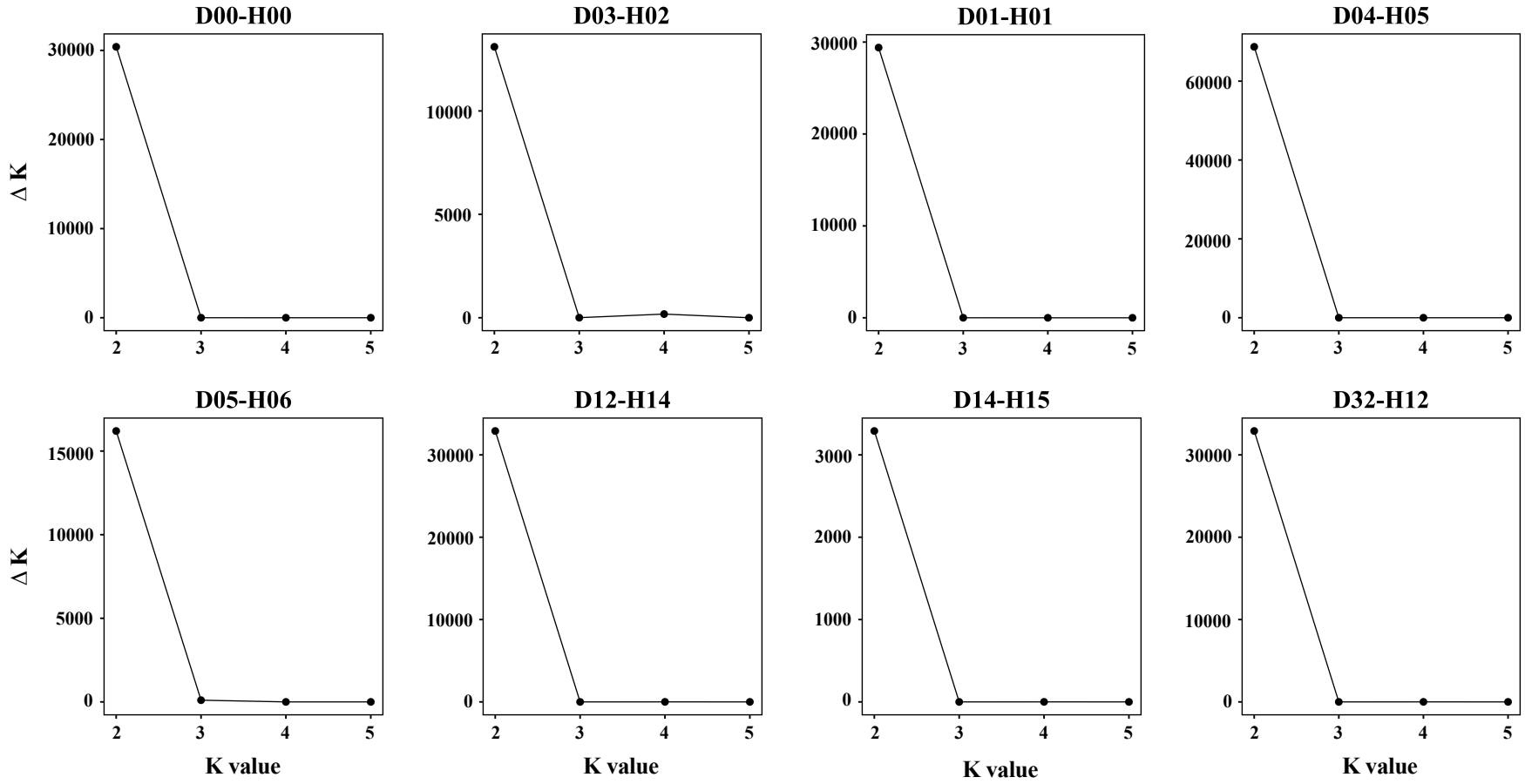
Supplementary Fig. S4 *fastSTRUCTURE* K=22, K=28 and marginal likelihoods

(A) Bayesian assignment of individuals to genetic clusters within *fastSTRUCTURE* for K=22 and K = 28. Each of the 1,319 individuals is depicted as a bar, with colors representing ancestry proportions to each cluster. Populations are ordered according to their geographic distribution along the coast. **(B)** Marginal likelihood values for successive K-values within *fastSTRUCTURE*. Red dashed lines denote the K-value that best explained the structure in the data (K = 22), as well as the K-value that maximized the marginal likelihood of the data (K = 28). **(C)** Change in marginal likelihoods from *fastSTRUCTURE* for successive K-values. Red dashed line denotes K = 23, higher K-values produce a negligible change in likelihood values.



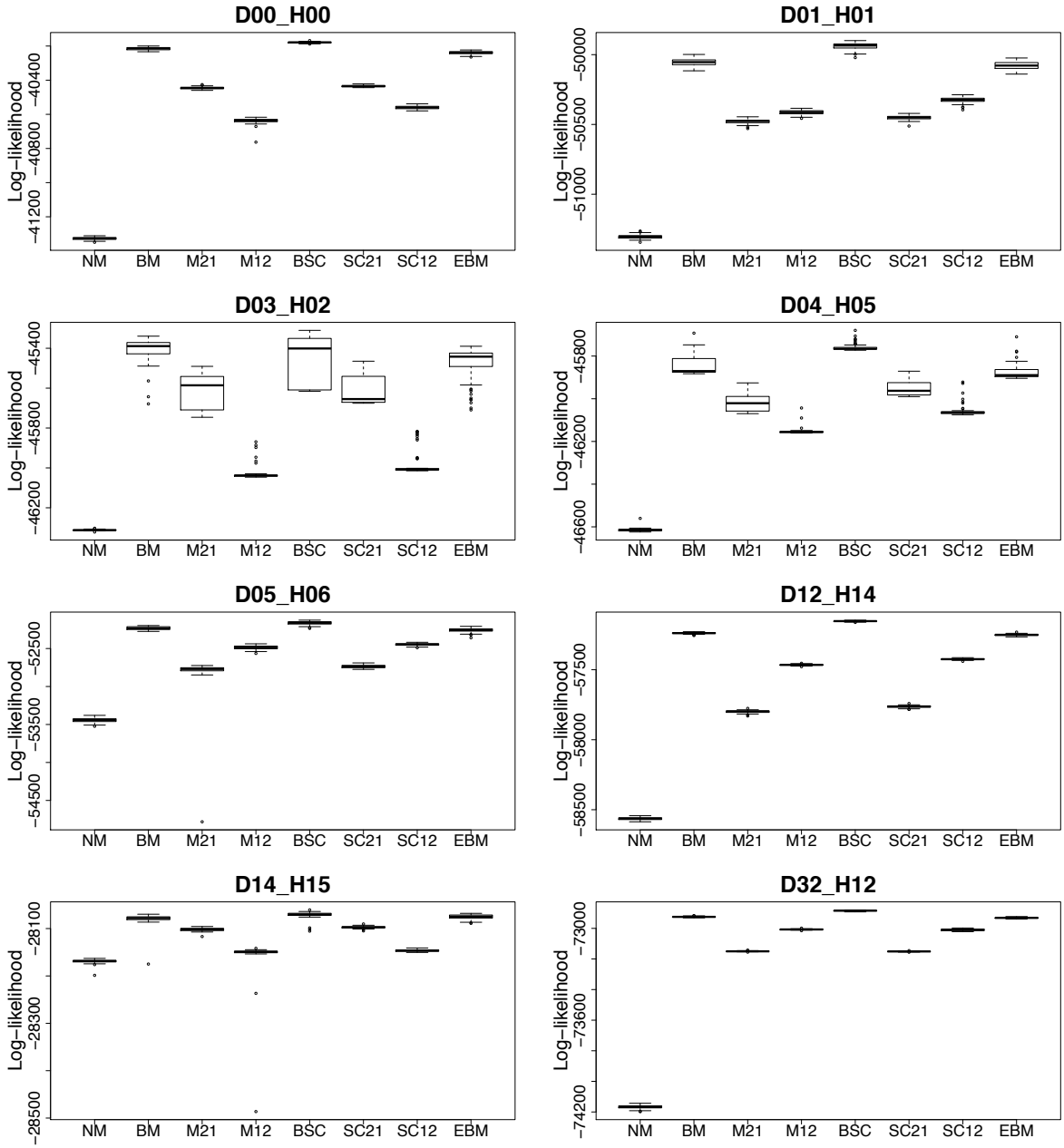
Supplementary Fig. S5 *STRUCTURE* best K-values for the Dune-Headland pairs

STRUCTURE best K-values for the eight Dune-Headland replicate pairs based on the maximum value for ΔK (the second order rate of change in the log probability of data between successive K-values).

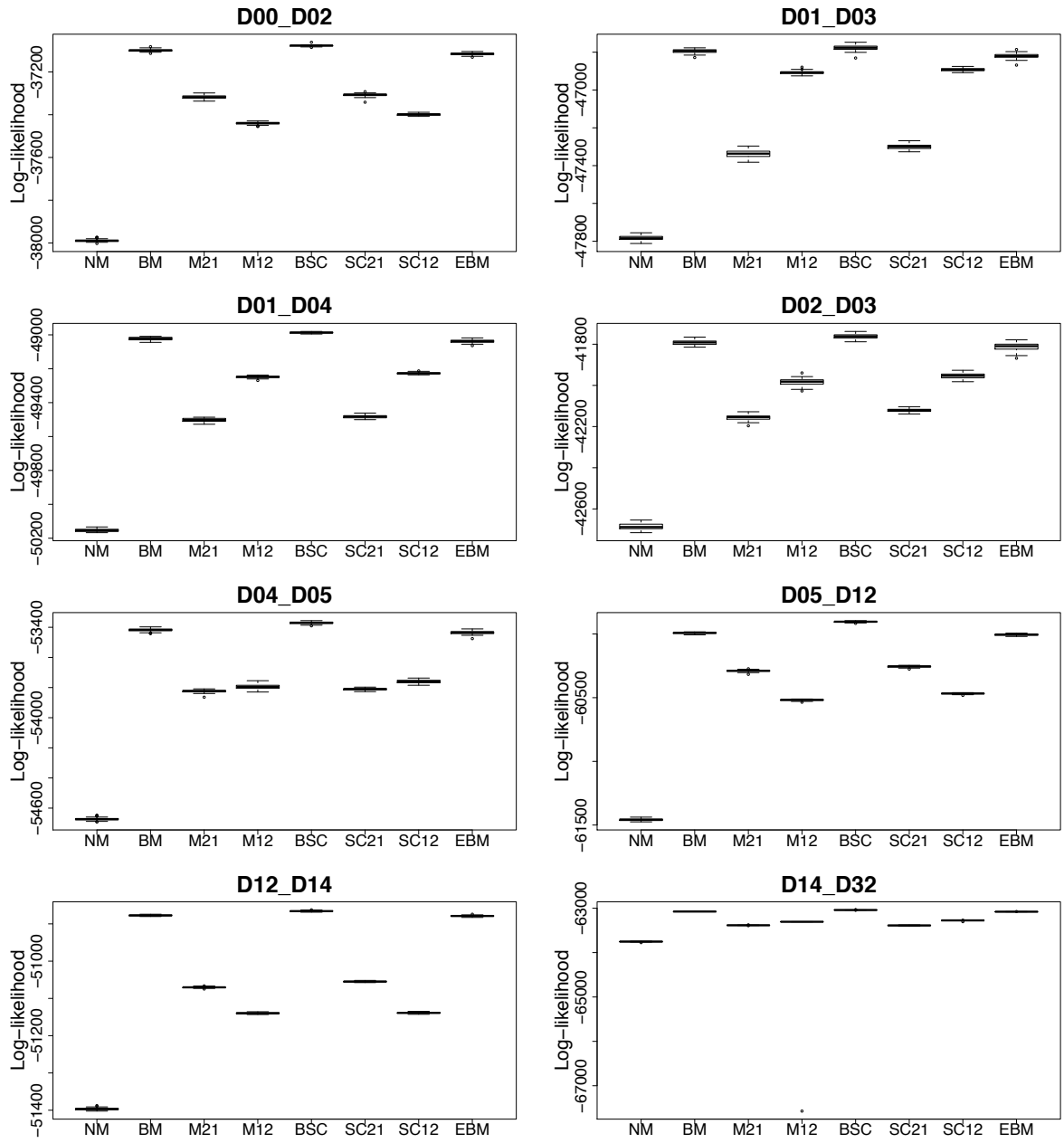


Supplementary Fig. S6 Log-likelihood values for the eight demographic models tested in *fastsimcoal2* per pair

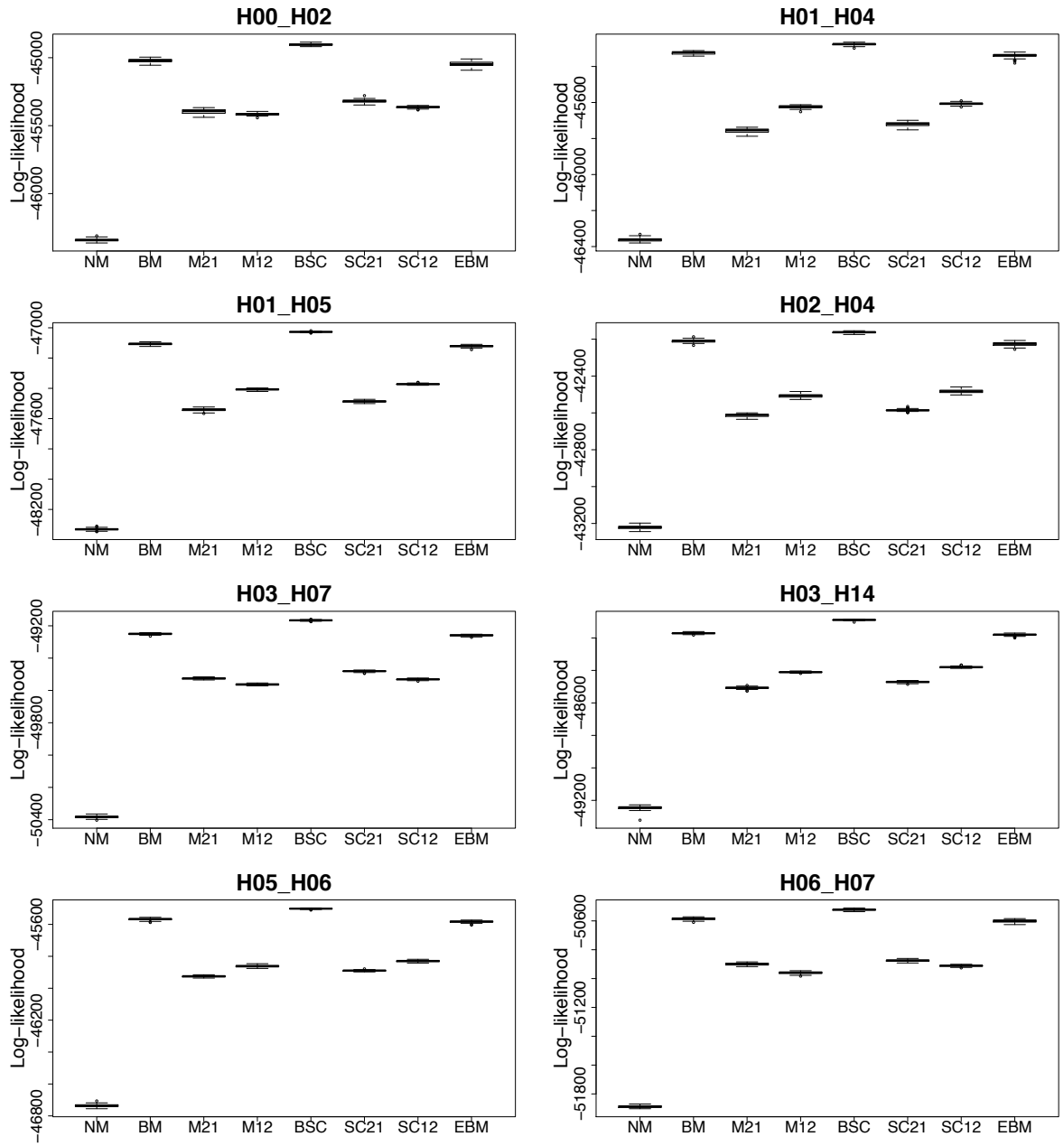
NM: no migration. BM: bidirectional migration. M21: migration from population 2 to 1. M12: migration from population 1 to 2. BSC: bidirectional migration after secondary contact. SC21: migration from population 2 to 1 after secondary contact. SC12: migration from population 1 to 2 after secondary contact. EBM: bidirectional migration after population splitting with cessation of gene flow.



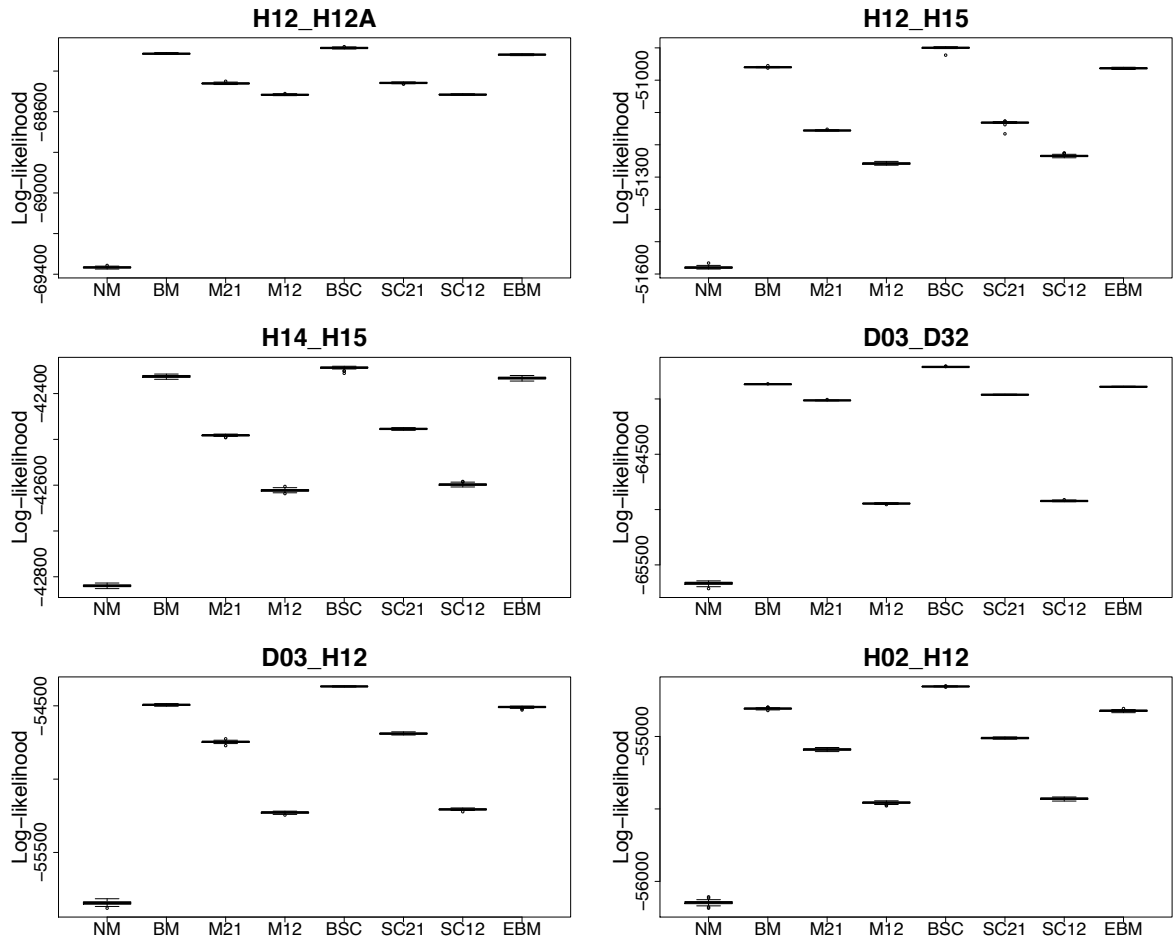
Supplementary Fig. S6 cont.



Supplementary Fig. S6 cont.

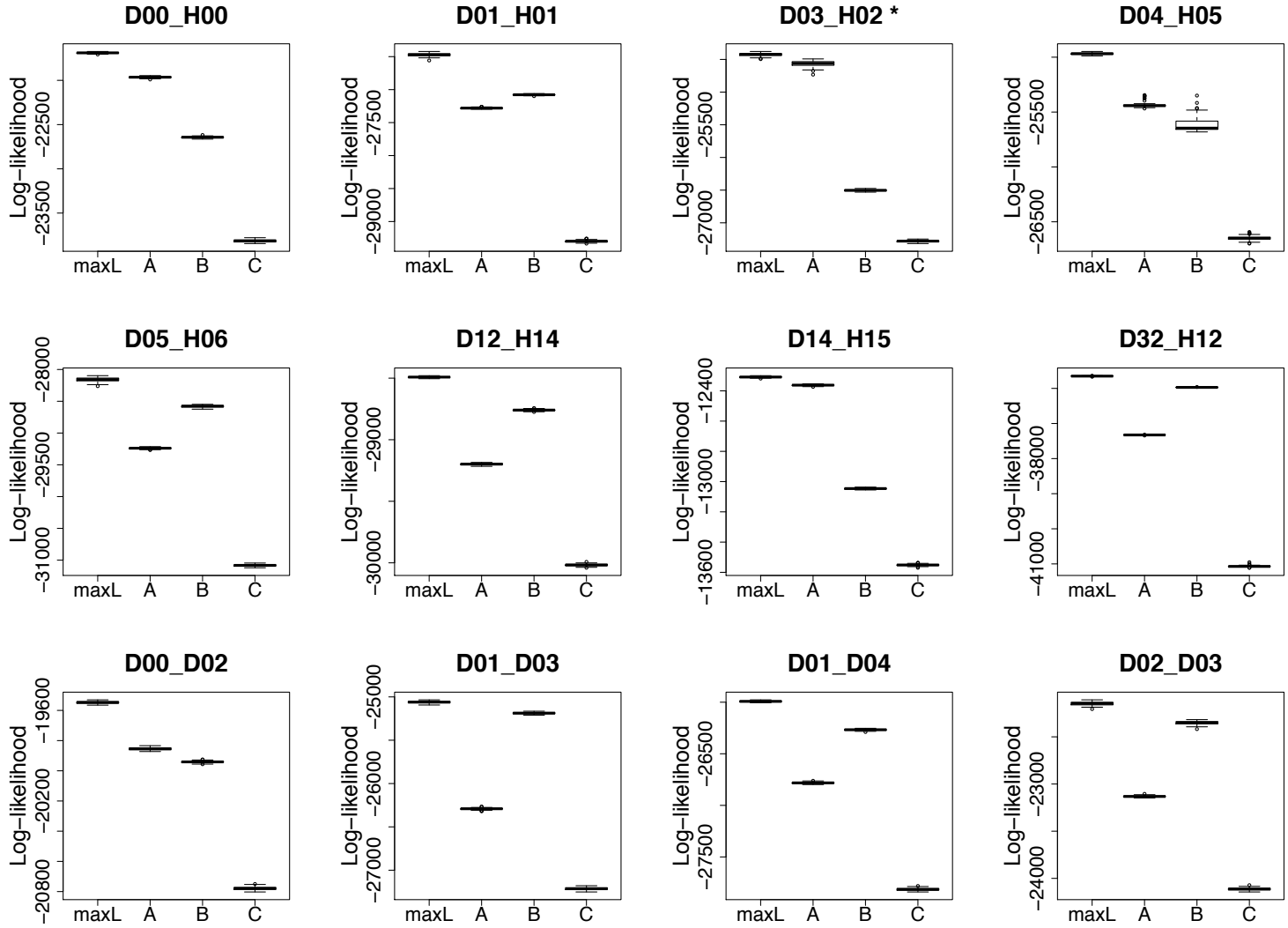


Supplementary Fig. S6 cont.

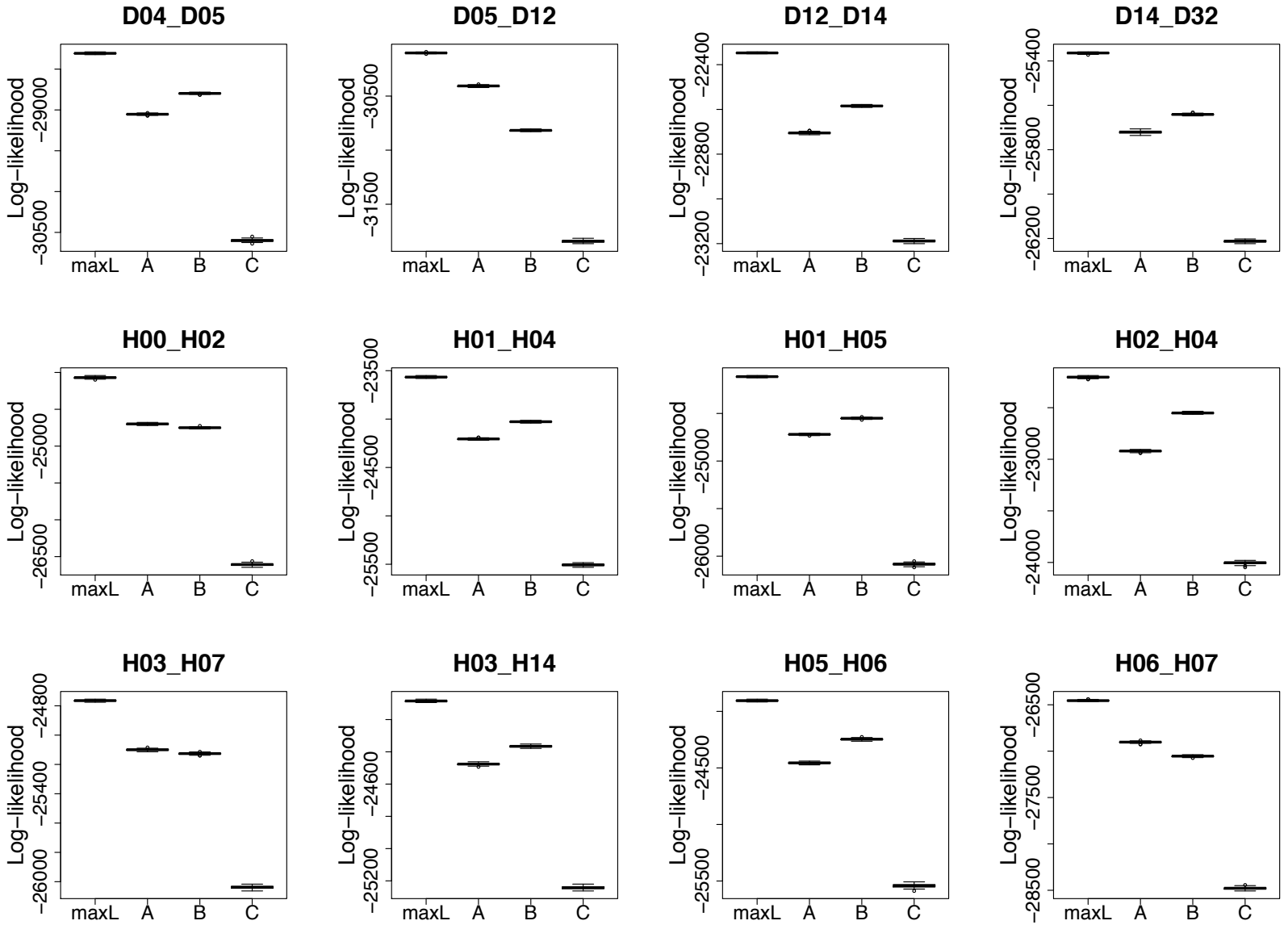


Supplementary Fig. S7 Likelihood values for testing whether gene flow is negligible

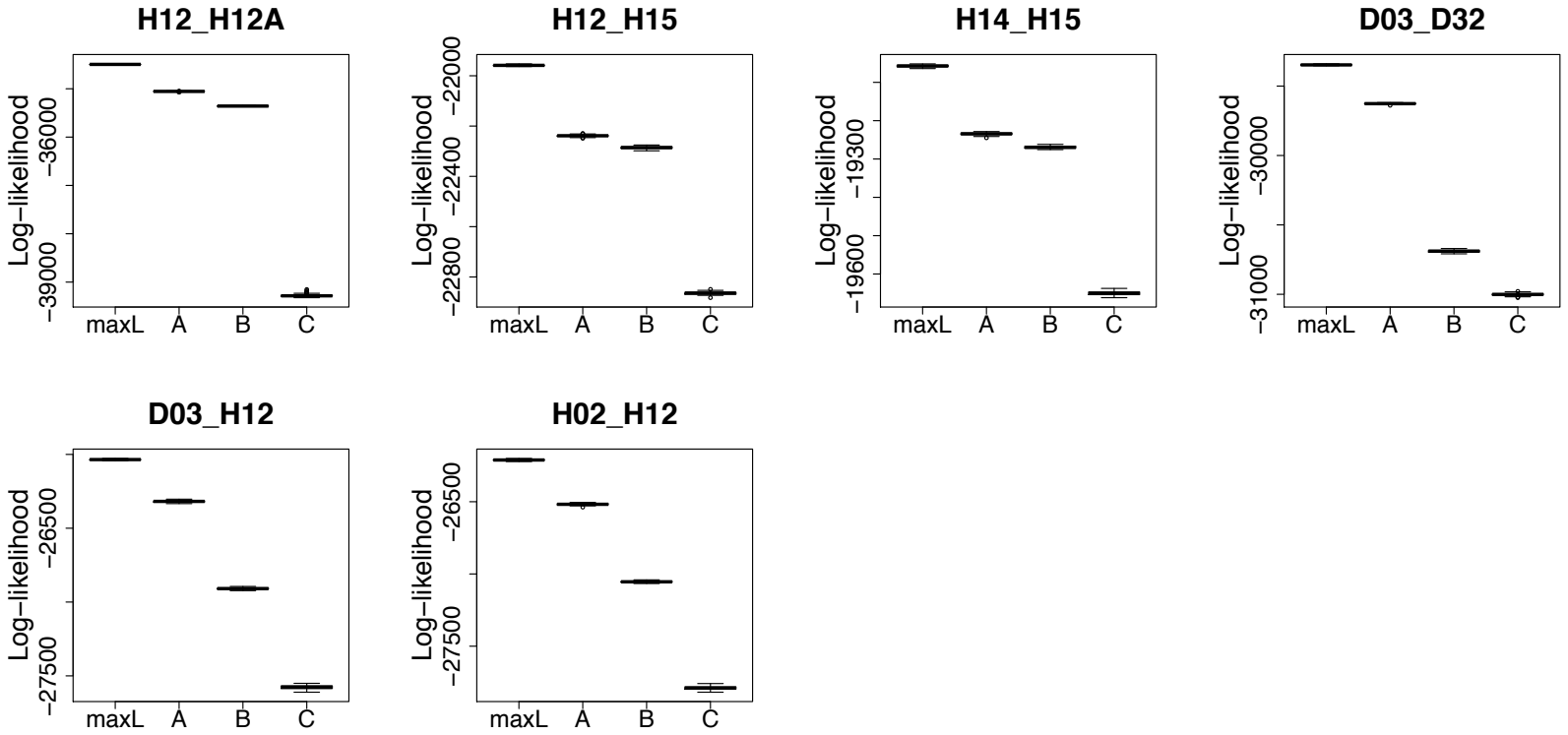
Max L: maximum likelihood for the best run from the best model. *A*: fixed negligible gene flow ($2Nm = 0.01$) from population 2 (on the right) to population 1 (on the left). *B*: fixed negligible gene flow from population 1 to 2 ($2Nm = 0.01$). *C*: fixed negligible gene flow in both directions ($2Nm = 0.01$). The asterisk denotes the pair where at least one of the migration rates is not significantly different from $2Nm = 0.01$.



Supplementary Fig. S7 cont.



Supplementary Fig. S7 cont.



Supplementary methods

Estimation of monomorphic sites per pair

To estimate the number of monomorphic sites per pair we first calculated the number of RAD loci by using *PLINK* to thin for one SNP per RAD locus. The total read length of each RAD locus was (on average) 190bp (taking into account the length of the sequencing read after removal of barcodes/indexes). We used the following formula to calculate the number of monomorphic sites per pair:

Monomorphic sites = (read length x number RAD loci) - number variable sites

Here, we may be slightly overestimating the number of monomorphic sites as we are assuming all sites without a called SNP are monomorphic, although some could be actual variants that were not called due to not passing filtering requirements. Nevertheless, the parameter estimates (especially the migration rates) were robust to varying the number of monomorphic sites (data not shown).

Probabilities of gene flow distorting phylogeny topology

Consider a phylogeny of four populations, referred to as P1, P2, P3, P4. We have observed a species/population topology in which populations P3 and P4 are not sister. If we assume that the true species/population tree has the topology ((P1, P2), (P3, P4)), what is the probability that P3 and P4 are not sister under a model containing gene flow between P1 and P3 and between P2 and P4?

Alleles x1, x2, x3, x4 represent four lineages which are sampled from the respective populations at the present (t = 0). The lineages will be traced backwards in time, tracking coalescent events. The MRCA common ancestor of the sampled lineages is denoted x1,2,3,4. Other common ancestors are denoted likewise. Migration occurs from P1 to P3 and from P2 to P4 at tm. Migration occurs as an instantaneous burst and the parameter m represents the fraction of alleles in the recipient populations replaced by migrant alleles from the donor population. Divergence between P1 and P2 and between P3 and P4 occurs at time t1. The ancestor of P1 and P2 diverges from the ancestor of P3 and P4 at time t2.

The approach below will sum coalescent probabilities of all mutually exclusive ways in which (3,4) are not closest relatives. The probability of each scenario will be assigned to a variable and these will be summed at the end. Probabilities will be conditioned on whether x3 and x4 are descended from migrant alleles.

Condition on x3 being descended from a migrant, x4 not descended from a migrant

S1. x1 and x3 coalesce between tm and t1

In[1]:=
```
S1 = 1 - E^(-(t1 - tm))
```

Out[1]=
$$1 - e^{-t1+tm}$$

S2. x1 and x3 don't coalesce between tm and t1. In the ancestral lineage containing exclusively x1, x2, and x3, a single coalescent event occurs between t1 and t2. That coalescent event is either (x3 with x1) or (x3 with x2).

In[2]:=
```
t = t2 - t1
OneCoal = Integrate[3 E^(-3 s) (E^(-(t - s))), {s, 0, (t)}]
S2 = (1 - S1) OneCoal (2/3)
```

Out[2]=
$$-t1 + t2$$

Out[3]=
$$\frac{3}{2} e^{t1-3\,t2}\left(-e^{2\,t1} + e^{2\,t2}\right)$$

Out[4]=
$$e^{-3\,t2+tm}\left(-e^{2\,t1} + e^{2\,t2}\right)$$

S3. x1 and x3 don't coalesce between tm and t1. In the ancestral lineage containing exclusively x1, x2, and x3, a single coalescent event occurs between t1 and t2. That coalescent event is (x1 with x2). In the ancestral lineage containing x1,2, x3, and x4, the first coalescent event that occurs is either (x1,2 with x3) or (x1,2 with x4).

In[5]:=
```
S3 = (1 - S1) OneCoal (1/3) (2/3)
```

Out[5]=
$$\frac{1}{3} e^{-3\,t2+tm}\left(-e^{2\,t1} + e^{2\,t2}\right)$$

S4. x1 and x3 don't coalesce between tm and t1. In the ancestral lineage containing exclusively x1, x2, and x3, two coalescent events occur, the first is (x1 with x2) and the second is (x1,2 with x3)

In[6]:=
```
TwoCoal = Integrate[3 E^(-3 s) (1 - E^(-(t - s))), {s, 0, t}]
S4 = (1 - S1) TwoCoal (1/3)
```

Out[6]=
$$\frac{1}{2}\left(2 + e^{3\,t1-3\,t2} - 3\,e^{t1-t2}\right)$$

Out[7]=
$$\frac{1}{6}\,e^{-t1+tm}\left(2 + e^{3\,t1-3\,t2} - 3\,e^{t1-t2}\right)$$

S5. x1 and x3 don't coalesce between tm and t1. In the ancestral lineage containing exclusively x1, x2, and x3, zero coalescent events occur between t1 and t2. In the ancestral lineage containing x1, x2, x3, and x4, the first coalescent event that occurs is one of (x3 with x1), (x3 with x2), (x1 with x4), (x2 with x4).

In[8]:=
```
S5 = (1 - S1) E^(-3 (t2 - t1)) (4/6)
```

Out[8]=
$$\frac{2}{3}\,e^{-t1-3\,(-t1+t2)+tm}$$

S6. x1 and x3 don't coalesce between tm and t1. In the ancestral lineage containing exclusively x1, x2, and x3, zero coalescent events occur between t1 and t2. In the ancestral lineage containing x1, x2, x3, and x4, the first coalescent event that occurs is (x1 with x2). The next coalescent event to occur is either (x1,2 with x3) or (x1,2 with x4).

In[9]:=
```
S6 = (1 - S1) E^(-3 (t2 - t1)) (1/6) (2/3)
```

Out[9]=
$$\frac{1}{9}\,e^{-t1-3\,(-t1+t2)+tm}$$

Putting this together, the probability that x3 is descended from a migrant, x4 is not descended from a migrant, and that conditional on this scenario x3 and x4 do not appear as sister in the phylogeny is:

In[10]:=
```
Prob1 = m (1 - m) (S1 + S2 + S3 + S4 + S5 + S6)
```

Out[10]=
$$\left(1 - e^{-t1+tm} + \frac{7}{9}\,e^{-t1-3\,(-t1+t2)+tm} + \frac{1}{6}\,e^{-t1+tm}\left(2 + e^{3\,t1-3\,t2} - 3\,e^{t1-t2}\right) + \frac{4}{3}\,e^{-3\,t2+tm}\left(-e^{2\,t1} + e^{2\,t2}\right)\right)(1 - m)\,m$$

Condition on x4 being descended from a migrant, x3 not descended from a migrant

By symmetry, the probability that x4 is descended from a migrant, x3 is not descended from a migrant, and that conditional on this scenario x3 and x4 not being each other's closest relatives is:

In[11]:=
```
Prob2 = Prob1
```

Out[11]=
$$\left(1 - e^{-t1+tm} + \frac{7}{9}\,e^{-t1-3\,(-t1+t2)+tm} + \frac{1}{6}\,e^{-t1+tm}\left(2 + e^{3\,t1-3\,t2} - 3\,e^{t1-t2}\right) + \frac{4}{3}\,e^{-3\,t2+tm}\left(-e^{2\,t1} + e^{2\,t2}\right)\right)(1 - m)\,m$$

Condition on neither x3 nor x4 being descended from migrants

S7. x3 and x4 don't coalesce between t1 and t2. x1 and x2 don't coalesce between t1 and t2. In the ancestral lineage containing x1, x2, x3, and x4, the first coalescent event is one of (x3 with x1), (x3 with x2), (x1 with x4), (x2 with x4).

In[12]:=
```
S7 = (E^(-(t2 - t1)))^2 (4/6)
```

Out[12]=
$$\frac{2}{3}\,e^{2\,t1-2\,t2}$$

S8. x3 and x4 don't coalesce between t1 and t2. x1 and x2 don't coalesce between t1 and t2. In the ancestral lineage containing x1, x2, x3, and x4, the first coalescent event is (x1 with x2). The next coalescent event is one of (x1,2 with x3), (x1,2, x4).

In[13]:= S8 = (E^(−(t2 − t1)))^2 (1/6) (2/3)

Out[13]= $\frac{1}{9} e^{2\,t1-2\,t2}$

S9. x3 and x4 don't coalesce between t1 and t2. x1 and x2 coalesce between t1 and t2. In the ancestral lineage containing x1,2, x3, and x4, the first coalescent event is one of (x1,2 with x3), (x1,2, x4).

In[14]:= S9 = E^(−(t2 − t1)) (1 − E^(−(t2 − t1))) (2/3)

Out[14]= $\frac{2}{3} e^{t1-t2} \left(1 - e^{t1-t2}\right)$

Putting this together, the probability that neither x3 nor x4 are descended from migrants, and conditional on this x3 and x4 not being each other's closest relatives is:

In[15]:= Prob3 = (1 − m)^2 (S7 + S8 + S9)

Out[15]= $\left(\frac{7}{9} e^{2\,t1-2\,t2} + \frac{2}{3} e^{t1-t2} \left(1 - e^{t1-t2}\right)\right) (1 - m)^2$

Condition on both x3 and x4 being descended from migrants.

S10. x1 and x3 coalesce between tm and t1. x2 and x4 do not coalesce between tm and t1.

In[16]:= S10 = (1 − E^(−(t1 − tm))) E^(−(t1 − tm))

Out[16]= $e^{-t1+tm} \left(1 - e^{-t1+tm}\right)$

S11. x1 and x3 do not coalesce between tm and t1. x2 and x4 coalesce between tm and t1. By symmetry, this is the same as above

In[17]:= S11 = S10

Out[17]= $e^{-t1+tm} \left(1 - e^{-t1+tm}\right)$

S12. x1 and x3 coalesce between tm and t1. x2 and x4 coalesce between tm and t1.

In[18]:= S12 = (1 − E^(−(t1 − tm)))^2

Out[18]= $\left(1 - e^{-t1+tm}\right)^2$

S13. x1 and x3 do not coalesce between tm and t1. x2 and x4 do not coalesce between tm and t1. In the ancestral lineage containing x1, x2, x3, and x4, the first coalescent event is one of (x3 with x1), (x3 with x2), (x1 with x4), (x2 with x4).

In[19]:= S13 = (E^(−(t1 − tm)))^2 (4/6)

Out[19]= $\frac{2}{3} e^{-2\,t1+2\,tm}$

S14. x1 and x3 do not coalesce between tm and t1. x2 and x4 do not coalesce between tm and t1. In the ancestral lineage containing x1, x2, x3, and x4, the first coalescent event is (x1 with x2). The next coalescent event is either (x1,2 with x3) or (x1,2 with x4).

In[20]:= S14 = (E^(−(t1 − tm)))^2 (1/6) (2/3)

Out[20]= $\frac{1}{9} e^{-2\,t1+2\,tm}$

Putting this together, the probability of x3 and x4 both being descended from migrant alleles, and conditional on this x3 and x4 not being each other's closest relatives is:

In[21]:=
```
Prob4 = m^2 (S10 + S11 + S12 + S13 + S14)
```

Out[21]=
$$\left(\frac{7}{9} e^{-2\,t1+2\,tm} + 2\,e^{-t1+tm}\left(1 - e^{-t1+tm}\right) + \left(1 - e^{-t1+tm}\right)^2\right) m^2$$

Then from the law of total probability, we have that the probability of x3 and x4 not being each other's closest relatives is:

In[22]:=
```
P = Prob1 + Prob2 + Prob3 + Prob4
```

Out[22]=
$$\left(\frac{7}{9} e^{2\,t1-2\,t2} + \frac{2}{3} e^{t1-t2}\left(1 - e^{t1-t2}\right)\right)(1-m)^2 +$$
$$2\left(1 - e^{-t1+tm} + \frac{7}{9} e^{-t1-3\,(-t1+t2)+tm} + \frac{1}{6} e^{-t1+tm}\left(2 + e^{3\,t1-3\,t2} - 3\,e^{t1-t2}\right) + \frac{4}{3} e^{-3\,t2+tm}\left(-e^{2\,t1} + e^{2\,t2}\right)\right)(1-m)\,m +$$
$$\left(\frac{7}{9} e^{-2\,t1+2\,tm} + 2\,e^{-t1+tm}\left(1 - e^{-t1+tm}\right) + \left(1 - e^{-t1+tm}\right)^2\right) m^2$$

Visuals

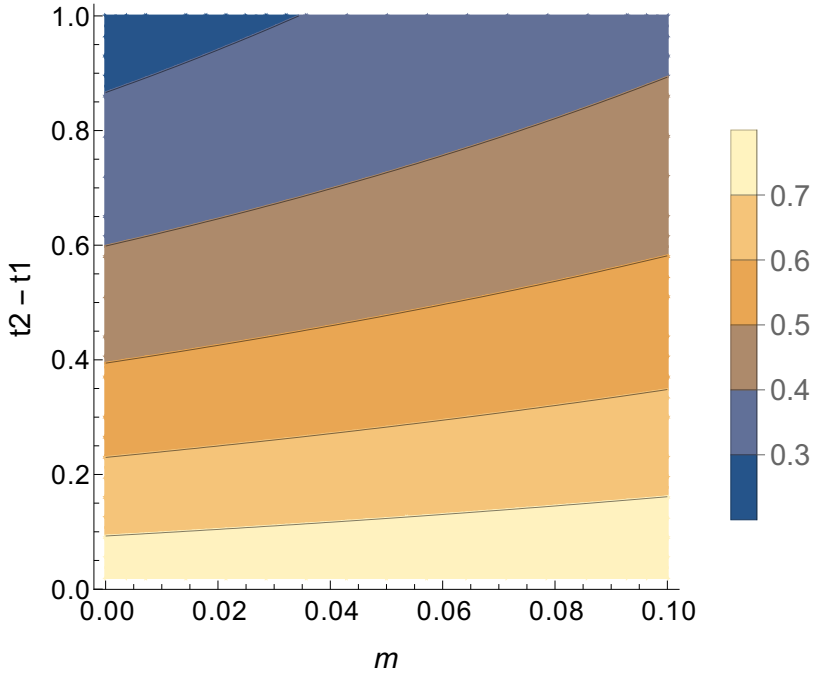
In[23]:=
```
PPlot1 = P /. t2 → (t3 + t1) /. t1 → 1 /. tm → .1
PPlot2 = P /. t2 → t1 + 1 /. t1 → tm + t4 /. tm → .1
```

Out[23]=
$$\left(\frac{7}{9} e^{2-2\,(1+t3)} + \frac{2}{3} e^{-t3}\left(1 - e^{-t3}\right)\right)(1-m)^2 +$$
$$2\left(0.59343 + \frac{7}{9} e^{-0.9-3\,t3} + \frac{4}{3} e^{0.1-3\,(1+t3)}\left(-e^2 + e^{2\,(1+t3)}\right) + 0.0677616\,\left(2 - 3\,e^{-t3} + e^{3-3\,(1+t3)}\right)\right)(1-m)\,m +$$
$$0.963267\ m^2$$

Out[24]=
$$\left(\frac{2\left(1 - \frac{1}{e}\right)}{3\,e} + \frac{7}{9} e^{2\,(0.1+t4)-2\,(1.1+t4)}\right)(1-m)^2 +$$
$$2\left(1 + \frac{7\,e^{-3-t4}}{9} - e^{-t4} + \frac{4}{3} e^{0.1-3\,(1.1+t4)}\left(-e^{2\,(0.1+t4)} + e^{2\,(1.1+t4)}\right) + \frac{1}{6} e^{-t4}\left(2 - \frac{3}{e} + e^{3\,(0.1+t4)-3\,(1.1+t4)}\right)\right)(1-m)\,m +$$
$$\left(\frac{7}{9} e^{0.2-2\,(0.1+t4)} + 2\,e^{-t4}\left(1 - e^{-t4}\right) + \left(1 - e^{-t4}\right)^2\right) m^2$$

In[25]:=
```
ContourPlot[PPlot1, {m, 0, .1}, {t3, ((1/50)), 1}, PlotLegends → Automatic,
  Axes → False, Frame → {True, True, False, False}, FrameLabel → {m, t2 − t1},
  LabelStyle → Directive[FontSize → 16], FrameTicksStyle → Directive[FontSize → 14]]
ContourPlot[PPlot2, {m, 0, .1}, {t4, ((1/50)), 1}, PlotLegends → Automatic,
  Axes → False, Frame → {True, True, False, False}, FrameLabel → {m, t1 − tm},
  FrameTicksStyle → Directive[FontSize → 14], LabelStyle → Directive[FontSize → 16]]
```

Out[25]=



Out[26]=