Highly replicated evolution of parapatric ecotypes

James, M. E.[1], Arenas-Castro, H[1], Groh, J. S.[1,2], Engelstädter, J.[1], and D. 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.

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. In contrast, when ecotypes arise once, expand their range and colonise similar environments, their populations cluster by ecology and not geography. However, discriminating between these scenarios is difficult because gene flow upon secondary contact can create the appearance of multiple origins despite a true single origin history. Here, we convincingly demonstrate multiple origins within the Dune and Headland ecotypes of an Australian wildflower, *Senecio lautus*. We observed phylogenetic clustering by geography and strong genetic structure between populations. There was surprisingly little gene flow between parapatric ecotypes, which is not high enough to obscure a single origin history. Overall, our work highlights the importance of demonstrating that populations have arisen repeatedly and independently within studies of parallel evolution.

Introduction

Governed by natural selection, parallel evolution occurs when populations evolve similar traits after repeatedly and independently colonising similar habitats (Schluter & Nagel, 1995). The patchy distribution of phenotypically similar populations means they frequently occur next to other contrasting forms (e.g., plant species inhabiting serpentine and non-serpentine soils in Scandinavia (Berglund *et al.*, 2003), and marine snails adapted to crab predators or wave action along the rocky coasts of Spain (Johannesson *et al.*, 2010). Parallel evolution by natural selection creates consistent patterns of phenotypic similarity and divergence that can extend to morphological (Elmer *et al.*, 2010; Ravinet *et al.*, 2013; Perreault-Payette *et al.*, 2017), behavioural (York & Fernald, 2017), and reproductive (Smith & Rausher, 2011) 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 (Orr, 2005; Stern & Orgogozo, 2009; Rosenblum *et al.*, 2014; Lenormand *et al.*, 2016; Stoltzfus & McCandlish, 2017; Blount *et al.*, 2018; Yeaman *et al.*, 2018). However, it is surprisingly rare for studies of parallel evolution to convincingly demonstrate that populations have arisen in an independent and repeated fashion (hereafter multiple origin). Ruling out alternative demographic scenarios, such as a single origin of forms followed by gene flow upon secondary contact, is seldomly performed (but see Quesada *et al.*, 2007; Bierne *et al.*, 2013; Butlin *et al.*, 2014; Pérez-Pereira *et al.*, 2017, and see Ostevik *et al.*, 2012 for a critical review of the evidence in plants). In light of this, researchers may incorrectly assume a parallel colonisation history, leading to inaccurate inferences about the prevalence of parallel evolution in nature.

Typically, researchers of parallel evolution by natural selection ask whether phylogenetic clustering of populations coincides with the geography and not with the ecology of populations (Allender *et al.*, 2003; Quesada *et al.*, 2007; Johannesson *et al.*, 2010; Butlin *et al.*, 2014; Trucchi *et al.*, 2017). This is because genetic clustering of geographically close populations implies dispersal might be geographically restricted (i.e., isolation by distance; Wright, 1943), and colonisation of contrasting and neighbouring habitats might have occurred independently many times. However, alternative historical scenarios could also lead to clustering of populations by geography, and must be ruled out before examining the evolution of traits in light of parallel evolution (Endler, 1977; Barton & Hewitt, 1985; Coyne

& Orr, 2004; Bierne *et al.*, 2013). To understand this problem, first consider a scenario where an ancestral population gives rise to two locally adapted populations that occupy distinct yet geographically proximate habitats (hereafter ecotypes, Figure 1A). These two populations migrate and colonise new localities, where the same contrasting habitats are geographically close each time. This scenario of a single split followed by range expansion of two ecotypes does not have a parallel colonisation and adaptation history because each ecotype only arose once (rather than multiple independent times after independent colonisation of contrasting habitats). Because gene flow is either not possible after the original ecotypic split, or does not homogenise adjacent populations after range expansion, populations sharing the same ecology form reciprocally monophyletic clades in a phylogeny (Figure 1A).

Nevertheless, if there is sufficient gene flow between geographically close populations from two ecotypes that originated only once, the original phylogenetic signal of reciprocal monophyly can be eroded (Endler, 1977; Barton & Hewitt, 1985; Coyne & Orr, 2004; Bierne *et al.*, 2013). In other words, as the original signal of a single origin disappears, populations become most related to their neighbouring population and not to the other populations of the same ecotype. Therefore, gene flow can result in grouping of populations by geography rather than ecology if many loci are homogenised (Figure 1B). This phylogenetic signal is identical to that of true parallel evolution (a multiple origin scenario), where populations from two ecotypes arise multiple independent times (Figure 1C). Gene flow dynamics can thus fundamentally alter our interpretation of parallel evolution, to the extent that we can mistakenly infer parallel evolution in systems where secondary contact after range expansion of a single origin fused the history of locally adapted populations (Endler, 1977; Barton & Hewitt, 1985; Coyne & Orr, 2004; Bierne *et al.*, 2013).

However, not all levels of gene flow have the same equivocal effect in the genetic record of colonisation history (Bierne *et al.*, 2013). This makes it difficult to distinguish single from multiple origins of ecotypes. Systems of parallel evolution frequently detect gene flow between populations, especially when contrasting ecotypes are in close geographic proximity (i.e. parapatry). However, only few studies comprehensively model the demographic history of populations (Quesada *et al.*, 2007; Bierne *et al.*, 2013; Butlin *et al.*, 2014; Meier *et al.*, 2017; Pérez-Pereira *et al.*, 2017; Trucchi *et al.*, 2017), and even fewer have used simulations to address whether the levels of gene flow can obscure the observed phylogeny (Bierne *et al.*, 2013; Pérez-Pereira *et al.*, 2017). The system that has perhaps most clearly demonstrated the parallel origins of forms 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 (Quesada *et al.*, 2007; Johannesson *et al.*, 2010; Bierne *et al.*, 2013; Butlin *et al.*, 2014; Pérez-Pereira *et al.*, 2017). Also, an obvious extreme case of multiple origins arises when parallel evolution occurs between geographically distant populations where lack of gene flow cannot obscure the history of colonisation (e.g., threespine stickleback populations that colonise separate continents (Magalhaes *et al.*, 2019) . However, in other systems where gene flow is moderate between ecotypes (Rougemont *et al.*, 2015; Le Moan *et al.*, 2016; Rougeux *et al.*, 2017, 2019; Herman *et al.*, 2018), it remains unclear to what extent gene flow contributed to the signal of parallel evolution.

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 relied on different mutations in the same gene, suggesting this adaptive trait has arisen multiple independent times (Chan *et al.*, 2010). Conversely, in systems where the exact same mutation is repeatedly involved in adaptation (Colosimo *et al.*, 2005), it is more difficult to identify whether the trait was repeatedly selected for (i.e. via standing genetic variation), rather than arising once followed by the repeated colonisation of similar environments (Lee & Coop, 2019). 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, particularly because genetic experiments are not feasible or the genetic architecture of adaptation is highly polygenic (Tiffin & Ross-Ibarra, 2014; Yeaman, 2015).

The above considerations suggest we need to characterise the origin and colonisation history of forms as well as the repeated evolution of traits in systems where populations have adapted to similar environments. Such an approach will clarify the possible role of natural selection in shaping diversity in systems with broad geographic and ecological ranges, thus paving the way for understanding the molecular basis of adaptation and its implications for our understanding of predictability and repeatability in evolution. In this work, we characterise the origin and colonisation 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 (Roda *et al.*, 2013; Melo *et al.*, 2014). The two forms differ in their growth habit: the Dune ecotype is erect and colonises sand dunes, and the Headland ecotype is prostrate, forming matts on the ground of rocky headlands (Ali, 1964; Radford *et al.*, 2004; Thompson, 2005). These locally adapted populations (Richards & Ortiz-Barrientos, 2016; Walter *et al.*, 2016) are separated by strong extrinsic reproductive isolation (Melo *et al.*, 2014; Richards *et al.*, 2016) and exhibit similar morphology of each ecotype across populations (James *et al.*, 2020). 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 populations.

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 (Roda *et al.*, 2013). 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 colonisations and origins of Dune and Headland populations. Here, we directly estimate patterns of gene flow within and between Dune and Headland ecotypes, as well as other demographic parameters important for characterising the colonisation history of this system. We use estimates of demographic parameters in forward population genetic simulations to explore the conditions that would favour a phylogenetic transition from clustering by their ecology to clustering by their geography, thus helping us gain further confidence on our conclusions about parallel parapatric divergence in this system. Our results illustrate the way we understand parallel evolution and pave the way for analyses of parallel trait evolution driven by natural selection in plants, where cases of parallelism remain understudied.

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 ($n_{mean} = 58$, $n_{total} = 1338$; Figure 2A, Table S1). We sampled mature (flowering) plants evenly across the geographic range of each population, ensuring that sampled plants were at least one metre apart. DNA was extracted using a modified CTAB protocol (Clarke, 2009) 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 standardised to 10ng/uL.

GBS library construction

We created reduced representation libraries by using a two-enzyme Genotyping-by-Sequencing (GBS) approach (modified from Poland *et al.*, 2012). We created seven libraries, 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 (Faircloth & Glenn, 2011; Rohland & Reich, 2012). 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 (Schmieder *et al.*, 2010). 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; Table S2). Reads were mapped to the *S. lautus* reference PacBio genome v1.0 (Wilkinson, 2019) with *BWA-MEM* v0.7.15 (Li & Durbin, 2009; Li, 2013). On average, 86% of reads (SD = 15) mapped to the reference genome, and 81% (SD = 15) mapped properly with their paired read (Table S2). *PicardTools* v2.7.0 (Broad Institute, 2019) 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

(Garrison & Marth, 2012). Because SNPs were separately called for each of the 23 populations, we first normalised the 23 VCF files before merging them together. This was achieved by first using BCFtools v1.4.1 (Li *et al.*, 2009) to split multiallelic sites into biallelic records. Each file was then normalised by re-joining biallelic sites into multiallelic records. We then left-aligned and normalised indels, and used *vt* (Tan *et al.*, 2015) 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 (Puritz *et al.*, 2014a; b), including iterative filtering to maximise the number of sampled SNPs (O'Leary *et al.*, 2018). Using *VCFtools* v0.1.15 (Danecek *et al.*, 2011), 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* (Garrison, 2016). 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* (Raj *et al.*, 2014). We used *PLINK* v1.9 (Purcell *et al.*, 2007) 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 (Nguyen *et al.*, 2015) using the polymorphisms-aware phylogenetic model (Schrempf *et al.*, 2016). We first used ModelFinder (Kalyaanamoorthy *et al.*, 2017) 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 Schrempf *et al.*, 2016). 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 (Hoang *et al.*, 2018), and 10,000 replicates of SH-aLRT (Guindon *et al.*, 2010).

We further explored broad patterns of population structure using the variational Bayesian framework, *fastSTRUCTURE* v1.0 (Raj *et al.*, 2014). Here, we implement *fastSTRUCTURE* as extra evidence for whether populations genetically cluster by geography or ecotype. We do not infer specific historical admixture scenarios from *fastSTRUCTURE*, as different demographic scenarios can give rise to indistinguishable structure plots (Lawson *et al.*, 2018). The *fastSTRUCTURE* algorithm assigns individuals into genetic clusters (K) by minimising departures from Hardy-Weinberg equilibrium and inferring individual ancestry proportions to each genetic cluster. 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 maximised the marginal likelihood of the data. Results were summarised and plotted in the R package *pophelper* v2.2.7 (Francis, 2017).

Is there gene flow across the system?

To explore patterns of gene flow in a phylogenetic context, we used *TreeMix* v1.13 (Pickrell & Pritchard, 2012). *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, 1000 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 (Pickrell & Pritchard, 2012).

To more formally test for admixture, we used the *threepop* function in *TreeMix* to calculate *f3*-statistics (Reich *et al.*, 2009). 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*, so *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* (Thompson, 2005), except the prostrate *S. lautus* Headland ecotype and a few Alpine populations, suggesting these prostrate forms are derived. We tested for isolation by distance (IBD; Wright, 1943) in the ancestral and derived ecotypes to evaluate similarities in their dispersal dynamics (Slatkin, 1993). We tested for IBD using migration rates (2Nm) inferred in *fastsimcoal2* (see below) as well as Slatkin's $\widehat{M}$, $(1 / F_{ST} - 1)/4$, as a proxy for gene flow (Slatkin, 1993). For Slatkin's $\widehat{M}$, we used the dataset excluding the western Australia populations (D09 and D35), with 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*acos(sin(lat1)*sin(lat2)+cos(lat1)*cos(lat2)*cos(long1-long2))*, 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 (R Core Team, 2017), 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* in the *vegan* package (Blanchet *et al.*, 2018).

Is there gene flow between parapatric populations?

We examined levels of admixture between parapatric populations with *STRUCTURE* v2.3.4 (Pickrell & Pritchard, 2012). *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 (Falush *et al.*, 2003) 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 (Evanno *et al.*, 2005), 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, summarise 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 (Excoffier *et al.*, 2013). 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 (Hahn, 2018). We tested eight demographic models (Figure 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 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 and eleven Headland-Headland pairs (Table S2). To determine the baseline level of gene flow inferred by *fastsimcoal2* between isolated populations, namely the null gene flow expectation, we estimated migration rates for three very divergent allopatric populations (>1,500 km apart, between the eastern and south-eastern clades; D03-D32, D03-H12, and H02-H12), and took the highest detected migration rate from these allopatric comparisons as the baseline rate.

As above, the western Australia populations (D09 and D35) were excluded from this dataset to increase the number of sampled SNPs. For each pair, we filtered for a minor allele count of one (MAC1), retaining between 6,679 and 19,951 variable sites per pair (mean = 12,155 SNPs, SD = 3,316). By using a MAC1 and a relatively high number of samples per population (mean = 57, SD = 15), we retain rare alleles that are informative about migration events between the populations (Slatkin, 1985b). Since we cannot distinguish ancestral from derived alleles, we used the minor allele SFS (folded SFS). We used an *ad hoc* approach to estimate the number of monomorphic sites (see Supplementary Methods). Gene flow estimates were robust to varying the number of monomorphic sites (data not shown). We used custom R functions (modified from Liu *et al.*, 2018) to generate the joint folded SFS per population pair without downsampling.

We performed 50 independent *fastsimcoal2* runs per model per population pair. Each run consisted of 100,000 coalescent simulations and 40 expectation-maximisation cycles for parameter optimisation. We used a mutation rate of 1.0×10^{-8} based on Asteraceae EST sequence comparisons and fossil calibrations (Strasburg & Rieseberg, 2008). We ranked the models based on the Kullback–Leibler information value which was estimated from the AIC scores of the best run per model. Here, the normalisation of the difference between the AIC scores of a particular model and the best model in the set provides a measure of the degree of support for a particular model, namely model likelihood (w_i) (Thomé & Carstens, 2016). Since the use of linked-SNPs might lead to pseudo-replication issues when comparing models based on *fastsimcoal2* likelihood values (Bagley *et al.*, 2017) and the SFS discards linkage information, we verified SNPs were largely unlinked by calculating linkage-disequilibrium in PLINK (data not shown).

As *fastsimcoal2* uses simulations to approximate the likelihood values, there is variance in the likelihood estimates. To test whether the best model significantly differs from alternative models with negligible gene flow (2Nm = 0.01) but the same values at other parameters, we compared their likelihood distributions based on 100 expected SFS from 100,000 coalescent simulations per model (Bagley *et al.*, 2017). If likelihood distributions overlap, there is no significant differences between the fit of both models (Meier *et al.*, 2017). 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-maximisation 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 (Mangiafico, 2015).

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

To ask under what conditions gene flow can erode a signal of phylogenetic monophyly of each ecotype, we ran forward simulations of neutral polymorphism in *SLiM2* (Haller & Messer, 2017), see Supplementary Methods. *SLiM2* simulates diploid genomes using a Wright-Fisher model, and tracks derived mutations within simulated genomes. We mimicked a model of a single-origin scenario, where an ancestral population splits into two populations (Figure 5A). 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 splits again and one population of one ecotype exchanges genes with the other ecotype (representing one parapatric pair at *location 1*, and this also occurs at *location 2* to represent the other parapatric pair). We varied the following parameters: population size, migration rate, time from the present to the second split (*T1*), and time from the second split to the split of the ancestral population (*T2*). In addition, an outgroup population was retained after the first split, in order to construct a rooted tree (see below). Each model used a heuristic burn-in period of 10 x N (population size) generations to reach mutation-drift balance in the ancestral population. After each simulation, 30 individuals per population were sampled and output in a VCF file.

We calculated a distance matrix between individual genotypes and constructed a rooted neighbour-joining tree within the *ape* R package (Paradis *et al.*, 2004; Popescu *et al.*, 2012). We calculated the genealogical sorting index (GSI; Cummings *et al.*, 2008), see Supplementary Methods. GSI is a measure of how monophyletic an arbitrary set of tips are on a tree. If all the tips form a monophyletic group, GSI will be 1, whereas if the tips are dispersed throughout the phylogeny, then GSI will be closer to 0. We calculated GSI for four sets of tips on each tree: for all 'Dunes', for all 'Headlands', for both populations from *location 1*, and for both populations from *location 2*. We took the average of the first two, and the average of the second two, then the log of the ratio of these two GSI values. When positive, the GSI ratio indicates a false signal of parallel origins, as the populations that are parapatric appear as each other's closest relatives. Conversely, when the GSI ratio is negative, it indicates that the 'true' signal of the single origin is stronger.

We also asked where our observed data fall in the parameter space, and whether the inferred migration rates between the Dune and Headland of each replicate pair is high enough to obscure the signal of a single origin. To first estimate which panel of the parameter space the *S. lautus* system is located, we estimated the internal branch (T1 of Figure 5A) by averaging the direct estimates of divergence times within ecotypes (calculated from *fastsimcoal2*, see above for details). We assumed a similar divergence time for the internal branch (T2 of Figure 5A). We then asked whether our observed D-H migration rates fall within the region where the phylogeny is not distorted.

To further explore the effect of gene flow in phylogenetic distortion, we compared the relative node order of the observed phylogeny (where the topology is estimated in the absence of gene flow) to the *fastsimcoal2* models (where gene flow is taken into account). We did this for four population pairs (D04-H05 and D05-H06; D14-H15 and D32-H12). Specifically, if the observed phylogeny represents a true parallel origin scenario, then isolation-with-migration models which jointly estimate gene flow and divergence time should infer deeper divergence times for populations of the same ecotype, compared to comparisons between putative sister populations of divergent ecotypes. We used divergence times estimated in *fastsimcoal2* to compare divergence within and between ecotypes, and asked whether these estimated divergence times were in accordance with the topology of a phylogeny.

Results

Populations cluster by geography and not by ecology

Phylogenetic inference reveals that neither ecotype forms a monophyletic clade, providing evidence against a single origin scenario (Figure 2B). Parapatric Dune-Headland populations are also often sister-taxa, giving evidence for the multiple origin of ecotypes. To visualise the major genetic structure within *fastSTRUCTURE*, we plotted the lowest K-values that capture the major structure in the data (Pritchard *et al.*, 2000; Lawson *et al.*, 2018; 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 (Figure 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 (Roda *et al.*, 2013; Melo *et al.*, 2019).

Minimal gene flow across the system

In the absence of migration, the *TreeMix* phylogeny explained 95.9% of the data, with the 24 additional migration events augmenting this value to 98.9 % (Figure S1). Figure 3A shows the first migration event ($P < 2.2\text{x}10^{-308}$) with a migration weight (w) of 0.40. Although the 24 other migration events were also significant ($P_{average} = 2.92\text{x}10^{-3}$, SD = 0.0062), their individual weightings were small (see Figure 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 colonisation history including long distance yet rare migration events, these P-values should be treated with caution. This is because model comparisons in *TreeMix* suffers from multiple testing, a large number of parameters, and the estimated graph can be inaccurate (Pickrell & Pritchard, 2012). We therefore tested the robustness of this inference using *f3-statistics*. All *f3-statistics* were positive (Figure 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; Figure 3B). A strong IBD trend exists between ecotypes for the eight pairs studied here ($\widehat{M}$: $F_{1,6} = 0.55$, P = 0.05661, multiple $R^2 = 0.48$, Figure 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 $R^2 = 0.08$, Figure 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 (Figure S4B) and each population is genetically distinct, suggesting that *S. lautus* has a simple demographic history with limited admixture (Lawson *et al.*, 2018). Specifically, the K-value that best explained the structure in the data was 22, the rate of change in the likelihood of each K-value (Figure S4C) was negligible for K = 24-28, and the K-value that maximised the marginal likelihood of the data was 28, together suggesting that the optimal K-value is around 23 (Figure S4A,B). The *fastSTRUCTURE* results for K=23 show that each population forms a distinct genetic cluster (Figure 3D), suggesting very little, if any, admixture between them, further implying that each sampled population has been separated from other populations long enough to be genetically distinct (see pairwise F_{ST} values in Table S3) and with insufficient levels of gene flow to homogenise their genomes (Lawson *et al.*, 2018). 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_{mean} = 11.5) compared to other populations (n_{mean} = 62). For groups with fewer samples, genetic clustering programs such as *fastSTRUCTURE* are likely to assign them as mixtures of multiple populations rather than their own distinct population (Lawson *et al.*, 2018). This is evident for K = 22, where the Tasmania populations appear admixed (Figure S4A).

Minimal gene flow between parapatric ecotypes and distant populations

We observed very few admixed individuals between the parapatric Dune-Headland populations at each locality within the *STRUCTURE* analysis for K = 2 (Figure 4B). On average, 9.36% of individuals were admixed per population, although their admixture proportions were on average less than 1% (mean = 0.008, SD = 0.018). This suggests that gene flow between parapatric populations might have stopped back in the past, and lineage sorting of many alleles has already taken place. For all pairs, the best K-value based on the Evanno method (Evanno *et al.*, 2005) was K = 2 (Figure S5). Demographic modelling in *fastsimcoal2* revealed the most likely divergence model for all population pairs within and between ecotypes was bidirectional gene flow after secondary contact (w_i > 0.99; Figure S6). However, for parapatric Dune-Headland population comparisons, direct measurements of migration rates revealed that most migration rates were very low (2Nm < 1.00), with the exception of D04-H05 and D32-H12 (Figure 3B upper section, 4B; Table S4, S5). For Dune-Dune population comparisons we also detected very low migration rates ($2Nm_{mean}$ = 0.23, SD = 0.09), with all pairs containing 2Nm < 1.00. For Headland-Headland comparisons we again detected very low migration rates ($2Nm_{mean}$ = 0.57, SD = 1.01), with all pairs containing 2Nm < 1.00, with the exception of H12-H12A (Figure 3B upper section, 4B; Table S4, S5). 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 1,500km (i.e. the null expectation; 2Nm = 0.39; Figure 3B). Three Dune-Headland pairs (D00-H00, D03-H02 and D12-H14) were also within this null range. Alternative models with negligible gene flow did not fit the data better with the exception of the D03-H02 pair (Figure S7), as previously detected in Melo *et al.*, (2019). Overall, the observed magnitude of gene flow and previous lines of evidence, make us conclude that most parapatric Dune-Headland populations in *S. lautus* are effectively allopatric.

Levels of gene flow do not obscure a single origin scenario

Simulation of a single origin scenario with gene flow (Figure 5A) in *SLiM* revealed how gene flow can erode the true signal of monophyly of each ecotype (true signal = Figure 5B blue regions, negative GSI ratios) leading to a distorted phylogeny where populations cluster by geography (distorted signal = Figure 5B red regions, positive GSI ratios). We detected this phylogenetic shift when *T1* was long and *T2* was short (Figure 5B top right panel) even for small amounts of gene flow. In contrast, we did not detect it when *T1* was short and *T2* was long (Figure 5B bottom left panel). For intermediate lengths of *T1* and *T2*, increasing migration rates lead to an increase in GSI ratios. In general, population size did not dramatically altered GSI ratios, except when both *T1* and *T2* were short (Figure 5B, upper left quadrats). Overall, these patterns suggest that as the speed of diversification increases (from long to short internal branches, *T2*), even small amounts of gene flow are likely to erode the true signal of a single origin where populations cluster by ecology and not by geography. This is because short internal branches are already likely to distort the phylogenetic signal due to high levels of ancestral polymorphism in each parapatric pair.

Our observed data are located in the bottom right panel of Figure 5B: the estimated divergence time within ecotypes (branch T1 of Figure 5A) ranged between 43,928 and 128,159 generations (mean = 84,559, SD = 22,916), which is closest to T1 = 100,000, also assuming a similar time between splitting events between parapatric pairs (branch T2 of

Figure 5A). None of our observed bidirectional migration rates (mean $m = 1.5\text{x}10^{-05}$, range = $1.7\text{x}10^{-06}$ to $5.2\text{x}10^{-05}$) fall within the region of the parameter space that would create full phylogenetic distortion (i.e. darkest red region of Figure 5B, bottom right panel). Most observed migration rates are low, bidirectional and less than $1.0\text{x}10^{-05}$ (Figure 5C, light grey), which are in regions of the parameter space that are unlikely to distort a single origin scenario. Only two population pairs (D04-H05 and D32-H12) had levels of gene flow that could have partially distorted a single origin scenario (Figure C, dark grey).

Divergence time estimations in *fastsimcoal2* (which considers gene flow) were in accordance with the observed phylogeny: we observed deeper divergence times for populations of the same ecotype compared to sister-taxa of different ecotypes. More specifically, for D04-H05 and D05-H06, the average divergence time between populations of the same ecotype (i.e. D04-D05 and H05-H06) was 79,801 (SD = 2,698), whereas the average divergence time between populations at each locality (i.e. D04-H05 and D05-H06) was 49,317 (SD = 26,319). This is also true for D14-H15 and D32-H12, where the average divergence time between populations of the same ecotype (i.e. D14-D32 and H15-H12) was 68,723 (SD = 17,526), and the average divergence time between populations at each locality (i.e. D14-H15 and D32-H12) was 43,318 (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 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, our results convincingly show that parapatric Dune and Headland populations have evolved multiple times repeatedly and independently, and that selection and drift, rather than gene flow, play a predominant role in the distribution of genetic diversity in this system. Below we discuss these results in light of parallel parapatric divergence in this highly replicated system of evolution.

Strong genetic structure between *Senecio lautus* coastal populations

The dispersal of gametes and seeds within a landscape depends upon the physical distance they can move and the availability of suitable habitats (Hansson, 1991). Within highly patchy environments, most gametes and seeds are restricted to disperse locally. Within coastal populations, where suitable habitats are largely limited, the dispersal kernel of a species highly restricts gene flow (Nathan *et al.*, 2012). In *S. lautus*, dispersal is governed by pollinators (including native bees, moths and butterflies) which can transport pollen grains up to 2 km in a day (White, 2008), and by wind that drives dispersal of seeds with pappi (Andersen, 1993). Although there is potential for long distance movement within the system both within and between ecotypes (Roda *et al.*, 2013), strong local adaptation prevents migrants and hybrids from effectively establishing beyond their local site (Richards & Ortiz-Barrientos, 2016; Walter *et al.*, 2016). This, coupled with a landscape of patchy environments along the coast, suggests that population structure in *S. lautus* is expected to be pronounced. Our findings are in accord with this expectation, and highlight various levels of population structure and history in the coastal system.

Genetic structure within *S. lautus* clusters populations according to their geographic distribution along the Australian coast, and not by the environment they occupy (Figure 2B, 2C). Within *fastSTRUCTURE*, the largest genetic groups within *S. lautus* encompass two clades (Figure 2C) which are called the eastern and south-eastern clades. Each clade can be further subdivided into two subclades (Figure 2C). These four clades are largely independent of each other, do not have evidence of long-distance gene flow between them (Figure 3A), and 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 (Figure 2B), and is consistent with our previous work using targeted sequencing of neutral genes (Melo *et al.*, 2019) and RADseq using pools of individuals (Roda *et al.*, 2013).

A further level of structure can be visualised at the locality (i.e. parapatric Dune-Headland populations), where each population is unique in this system: F_{ST} values are above 0.2 in each population comparison (Table S3), and *fastSTRUCTURE* supports K-values equal to the same number of populations sampled in this study (Figure 3D). Also, all parapatric pairs are fully differentiated with little admixture (Figure 4B), even those such as D04-H05 at Coffs Harbour (NSW) that have adjacent habitats, i.e. where the potential for gene flow between ecotypes is high. Previous ecological experiments in this population pair have demonstrated strong extrinsic reproductive isolation against migrants and hybrids (Richards *et al.*, 2016; Richards & Ortiz-Barrientos, 2016), and cline analyses revealed that the barrier to gene flow is complete (North, 2015). Finally, no single estimate of the *f3-statistic* for any population triad was negative (Figure S3), further supporting that there are negligible levels of gene flow between populations across the entire system.

There is a strong signal of isolation by distance (Wright, 1943) within each ecotype as well as the Dune-Headland parapatric pairs, where there is an increase in genetic differentiation between populations with increasing geographic distance. This pattern arises when populations are geographically restricted and are at an equilibrium of dispersal and drift. Isolation by distance also suggests that long distance dispersal within the system is not pervasive, and populations have colonised their habitats far enough in the past to approach an equilibrium under the current patterns of dispersal (Slatkin, 1993). Overall, a combination of strong selection and limited dispersal can explain why parapatric populations persist despite the opportunity for homogenising gene flow between them. Future ecological studies that directly estimate seed and pollen dispersal kernels will help clarify the relative contributions of movement and local adaptation to divergence in parapatry.

Parallel evolution of parapatric *S. lautus* ecotypes with minimal levels of gene flow

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 (Quesada *et al.*, 2007; Johannesson *et al.*, 2010; Bierne *et al.*, 2013; Butlin *et al.*, 2014; Martin *et al.*, 2015; Rougemont *et al.*, 2015; Le Moan *et al.*, 2016; Meier *et al.*, 2017; Pérez-Pereira *et al.*, 2017; Trucchi *et al.*, 2017; Rougeux *et al.*, 2019). A single origin scenario combined with high levels of gene flow can alter the phylogenetic relationships of populations, falsely suggesting multiple independent origins (Endler, 1977; Barton & Hewitt, 1985; Coyne & Orr, 2004; Bierne *et al.*, 2013). This is because genetic structure at neutral markers can be decoupled from colonisation history via introgression and incomplete lineage sorting. This needs careful scrutiny in our system: there are multiple parapatric divergences that have the potential for high gene flow due to their close geographic proximity. Surprisingly, we observed minimal levels of gene flow between parapatric *S. lautus* Dune-Headland pairs, similar levels of gene flow between parapatric and allopatric population pairs (Figure 3B), as well as linearly down the coast for populations within each ecotype (Figure 3B, 3C; Table S4, S5). However, two parapatric population pairs (D04-H05 and D32-H12) had an estimated number of migrants per generation above one (Figure 3B; Table S4, S5). This consistent with population genetic theory stating that, in the absence of divergent selection, these levels of gene flow would homogenise them (Slatkin, 1985a), potentially obscuring a single origin scenario. Therefore, we further addressed this problem by using forward simulations of the neutral divergence process to ask if the demographic parameters estimated in this study are likely to obscure the history of colonisation and divergence in *S. lautus*. This simulation approach is conservative because previous transplant experiments in the system (Melo *et al.*, 2014; Richards *et al.*, 2016; Richards & Ortiz-Barrientos, 2016; Walter *et al.*, 2016, 2018b; a) as well as clinal analyses (North, 2015), have shown that divergent natural selection is strong and creates extrinsic reproductive isolation between Dune and Headland populations.

In our simulations, we investigated the interaction of gene flow, incomplete lineage sorting, and drift on phylogenetic distortion. Although it is clear that there are regions in the parameter space that completely erode the signal of a single origin of ecotypes and falsely suggest their parallel origins (Figure 5B, red regions), there is also large fraction of the parameter space that does not (Figure 5B, blue regions). Although our simulations revealed that even small amounts of gene flow can distort the phylogeny, our observed levels of gene flow between parapatric Dune-Headland *S. lautus* populations are not high enough to distort the phylogenetic relationships amongst populations from different ecotypes. Even the two parapatric pairs (D04-H05 and D32-H12) that experience the most bidirectional gene flow do not fall within the region of parameter space where complete distortion of the phylogeny occurs (i.e. darkest red region of Figure 5B, bottom right panel). In addition, these population pairs are geographically and genetically distant from pairs at other localities, suggesting that their divergence likely occurred in parapatry. Further evidence that gene flow has not obscured a single origin scenario in *S. lautus* comes from comparing joint estimates of gene flow and divergence times (as implemented in isolation with migration models) between population pairs of the same ecotype and putative sister populations of divergent ecotypes. We observed that population pairs of the same ecotype and not those from different ecotypes show deeper divergence times. In addition, constructing the phylogeny taking into account gene flow did not alter the topology from its estimation in the absence of gene flow (Figure 3A), and parapatric pairs were not better explained by the presence of gene flow. Together, these results imply that phylogenetic distortion is highly unlikely in *S. lautus* and that such relationships reflect the true history of populations and ecotypes.

Overall, our results indicate that coastal Dune-Headland *S. lautus* populations are highly replicated, having originated multiple independent times in parapatry with limited levels of gene flow. Within the system we have high confidence for at least six (and potentially eight) independent parapatric Dune-Headland divergences. We treat the divergences at two localities (D04-H05 and D32-H12) with some caution: this is because their estimated number of migrants per generation is above one (which can lead to population homogenisation; Slatkin, 1985a), and they also fall within the simulated parameter space where there is some potential for phylogenetic distortion. Nevertheless, these pairs are from the two separate clades, and are genetically isolated from other such pairs, so even moderate levels of gene flow within each of these distant pairs will not substantially distort the phylogeny. In addition, the estimates of divergence times for these populations (taking into account gene flow) do not suggest homogenisation after secondary contact has occurred. Therefore, all eight parapatric divergences sampled within this study appear to have originated independently, thus evolving in parallel. Furthermore, in comparison to other systems (e.g., Le Moan *et al.*, 2016; Meier *et al.*, 2017; Trucchi *et al.*, 2017; Rougeux *et al.*, 2019), our observed migration rates between ecotypes of each pair are generally lower (*S. lautus* mean $m = 1.5 \times 10^{-5}$, range = 5.3×10^{-5} to 1.7×10^{-6}), yet are most similar to the *Littorina saxatilis* system (mean $m \sim 1.0 \times 10^{-6}$; Butlin *et al.*, 2014), which is perhaps the clearest example of parallel evolution in nature.

The limits of inference from parallel evolution

Parallel evolution allows the study of deterministic evolution in multiple ways. It not only helps us understand whether populations adapting to similar conditions evolve similar phenotypes, but whether this repeated adaptation is driven by the same or different genetic mechanisms in replicate populations (Lenormand *et al.*, 2016). Furthermore, in systems where reproductive isolation has evolved, we can begin to understand the relative contributions of prezygotic and postzygotic barriers to divergence (Nosil *et al.*, 2002; Rogers & Bernatchez, 2006; Stankowski, 2013). However, much remains unknown about the tempo and mode of adaptation and speciation and particularly whether the two processes share a common genetic basis. As such, studies of parallel evolution can help uncover further "rules of speciation" (Coyne & Orr, 1989), particularly with regard to the role of natural selection in creating diversity at different levels of organisation.

In coastal *S. lautus* we can start answering these questions. On one hand, our study helps us better interpret the multiple transplant experiments carried out in this system, and suggests that we can compare them as replicates of the adaptation and speciation process. For instance, it is common to four different coastal localities to find selection against migrants and hybrids in the field (Walter *et al.*, 2016), but very weak intrinsic reproductive isolation in F1 hybrids (Melo *et al.*, 2014; Richards *et al.*, 2016; Walter *et al.*, 2016), suggesting that parapatric ecological divergence is a major driver of diversification under natural conditions in the system. Nonetheless, two different studies have found strong intrinsic reproductive isolation in F2 hybrids (Richards *et al.*, 2016; Walter *et al.*, 2016), suggesting that genetic incompatibilities are indeed accumulating and segregating within populations. Polymorphic genetic incompatibilities have been discovered in many systems now (Scopece *et al.*, 2010; Cutter, 2012; Larson *et al.*, 2018), and perhaps monkeyflowers best illustrate how they are contributing to plant speciation (Lowry & Willis, 2010; Oneal *et al.*, 2014; Sweigart & Flagel, 2015; Zuellig & Sweigart, 2018). Within *S. lautus*, intrinsic reproductive isolation in F1 hybrids is almost complete between very divergent lineages (between the eastern and south-eastern clades), but also according to ecology: although Dune populations are interfertile despite half a million years of divergence, crosses between ecotypes, and crosses between divergent Headlands are almost fully infertile (Melo *et al.*, 2019). Given that these clades have independently evolved Dune and Headland forms, we can infer that adaptation to similar conditions is only concordant for certain habitats but not for others. This variable level of predictability might relate to the form of selection acting in each environment, or to the ruggedness of fitness landscapes across geography (Lenormand *et al.*, 2009, 2016; Salazar-Ciudad & Marín-Riera, 2013; de Visser & Krug, 2014; Blount *et al.*, 2018).

A major step forward to make sense of patterns of evolution across multiple replicates of parapatric divergence would be to isolate the actual genes responsible for adaptation and speciation, and to model their individual demographic history (Lee & Coop, 2017, 2019). This could help us better understand if adaptation arises from new mutations or from standing genetic variation, as well as reveal the nature of parallelism in a given system. For instance, we could describe parallel evolution in terms of repeated fixation of the same alleles (e.g., Colosimo *et al.*, 2005), or fixation of functionally equivalent alleles, as it might be plausible during polygenic adaptation (Berg & Coop, 2014; Tiffin & Ross-Ibarra, 2014; Yeaman, 2015), or when phenotypes arise from loss-of-function mutations (e.g., Chan *et al.*, 2010). Furthermore, to understand how patterns of evolution at the genotypic level manifest at the phenotype, studies of parallel evolution should directly link adaptive loci to phenotypic traits and further demonstrate that the trait(s) itself has been under repeated selection in independent populations (Storz & Wheat, 2010; Pardo-Diaz *et al.*, 2015; Hoban *et al.*, 2016). Nonetheless, our study demonstrates that studying neutral loci can uncover patterns of colonisation and migration that are consistent with parallel evolution, or even reveal alternate divergence scenarios (e.g., Roesti *et al.*, 2015). Given the strong correlation between coastal environment and growth habit in *S. lautus* (James *et al.*, 2020) and the results presented here, studying the genetics of adaptation across this highly replicated system will reveal the mode and tempo of adaptation and speciation. Our work also implies that previous discoveries in this system implicating divergence in hormone signalling, flowering, and stress-related pathways (Wilkinson *et al.*, 2019) are worth studying under the umbrella of parallel evolution thus helping us better frame divergence at different levels of organisation and development.

Finally, in our work we have unusually 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 allows me to sample of many rare variants and therefore better distinguish ancestral polymorphism from migration. Studies undertaking demographic modelling often sample 10-25 individuals per population (e.g., Roesti *et al.*, 2015; Kautt *et al.*, 2016; Trucchi *et al.*, 2017) and occasionally even less than 10 (e.g., Meier *et al.*, 2017), thus 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 (Slatkin, 1985b; Hey & Nielsen, 2007; Strasburg & Rieseberg, 2010; Cruickshank & Hahn, 2014). As our simulations reveal that even small amounts of gene flow can obscure a phylogenetic topology, studies that fail to detect gene flow with few numbers of individuals should treat results with caution.

Overall, here 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, with strong genetic structure between populations, isolation by distance, and surprisingly minimal gene flow between parapatric populations at each locality as well as the system as a whole. Simulations confirmed that gene flow levels are not high enough to obscure a single origin scenario. This makes *S. lautus* a highly replicated system of parapatric divergence and 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 (Foster *et al.*, 2007; Trucchi *et al.*, 2017; Cai *et al.*, 2019; Konečná *et al.*, 2019). Our work emphasises that researchers in the field of parallel evolution should strive to rule out a single origin scenario to demonstrate that populations within a system have arisen repeatedly and independently.

Acknowledgements

We are grateful to M.J. Wilkinson, A. Nguyen Vu and H.L. North for assisting with sample collection. 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 The University of Queensland and the Australian Research Council grants to DO, and by an Australian Postgraduate Award and a Graduate School International Travel Award to MEJ.

Author contributions

MEJ and DO conceived the project. MEJ and JE undertook sample collection. MEJ extracted DNA, prepared libraries, performed bioinformatics, and undertook the *IQ-Tree*, *fastSTRUCTURE*, *STRUCTURE* and *TreeMix* analyses. HA conducted the *fastsimcoal2* analyses. JSG performed the *SLiM* simulations. MEJ and DO wrote the paper with input from all authors. DO is the mentor and supervisor for the research program.

Conflicts of interest

We do not have any conflicts of interest.

Data archival

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

References

Ali, S. 1964. *Senecio lautus* complex in Australia. I. Taxonomic considerations and discussion of some of the related taxa from New Zealand. *Australian Journal of Botany* **12**: 282–291.

Allender, C.J., Seehausen, O., Knight, M.E., Turner, G.F. & Maclean, N. 2003. Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. *Proceedings of the National Academy of Sciences* **100**: 14074–14079.

Andersen, M.C. 1993. Diaspore morphology and seed dispersal in several wind-dispersed Asteraceae. *American Journal of Botany* **80**: 487–492.

Bagley, R.K., Sousa, V.C., Niemiller, M.L. & Linnen, C.R. 2017. History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*). *Molecular Ecology* **26**: 1022–1044.

Barton, N.H. & Hewitt, G.M. 1985. Analysis of hybrid zones. *Annual Review of Ecology and Systematics* **16**: 113–148.

Berg, J.J. & Coop, G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genetics* **10**: e1004412.

Berglund, A.B.N., Dahlgren, S. & Westerbergh, A. 2003. Evidence for parallel evolution and site-specific selection of serpentine tolerance in *Cerastium alpinum* during the colonization of Scandinavia. *New Phytologist* **161**: 199–209.

Bierne, N., Gagnaire, P.A. & David, P. 2013. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Current Zoology* **59**: 72–86.

Blanchet, G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., *et al.* 2018. *vegan: Community Ecology Package.* R package version 2.5-2. https://CRAN.R-project.org/package=vegan.

Blount, Z.D., Lenski, R.E. & Losos, J.B. 2018. Contingency and determinism in evolution: Replaying life's tape. *Science* **362**: eaam5979.

Broad Institute. 2019. *PicardTools*. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/.

Butlin, R.K., Saura, M., Charrier, G., Jackson, B., André, C., Caballero, A., *et al.* 2014. Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution* **68**: 935–949.

Cai, Z., Zhou, L., Ren, N.N., Xu, X., Liu, R., Huang, L., *et al.* 2019. Parallel speciation of wild rice associated with habitat shifts. *Molecular Biology and Evolution* **36**: 875–889.

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Shapiro, M.D., Brady, S.D., *et al.* 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**: 302–305.

Clarke, J.D. 2009. Cetyltrimethyl Ammonium Bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor Protocols* **2009**: Pdb.prot5177.

Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villarreal, G., Dickson, M., Grimwood, J., *et al.* 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**: 1928–1933.

Coyne, J.A. & Orr, H.A. 2004. *Speciation*. Sunderland, MA: Sinauer Associates.

Coyne, J.A. & Orr, H.A. 1989. Two rules of speciation. In: *Speciation and its consequences*. Sunderland, MA: Sinauer Associates.

Cruickshank, T.E. & Hahn, M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* **23**: 3133–3157.

Cummings, M.P., Neel, M.C. & Shaw, K.L. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* **62**: 2411–2422.

Cutter, A.D. 2012. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. *Trends in Ecology & Evolution* **27**: 209–218.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

de Visser, J.A.G.M. & Krug, J. 2014. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* **15**: 480–490.

Elmer, K.R., Kusche, H., Lehtonen, T.K. & Meyer, A. 2010. Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 1763–1782.

Endler, J. 1977. *Geographic variation, speciation, and clines*. Princeton, NJ: Princeton University Press.

Evanno, G., Regnaut, S. & Goudet, J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. & Foll, M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics* **9**: e1003905.

Faircloth, B. & Glenn, T. 2011. *Serapure*. https://ethanomics.files.wordpress.com/2012/08/serapure_v2-2.pdf.

Falush, D., Stephens, M. & Pritchard, J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Foster, S.A., McKinnon, G.E., Steane, D.A., Potts, B.M. & Vaillancourt, R.E. 2007. Parallel evolution of dwarf ecotypes in the forest tree *Eucalyptus globulus*. *New Phytologist* **175**: 370–380.

Francis, R.M. 2017. POPHELPER: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources* **17**: 27–32.

Garrison, E. 2016. *Vcflib, a simple C++ library for parsing and manipulating VCF files.* https://github.com/vcflib/vcflib.

Garrison, E. & Marth, G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv. arXiv:1207.3907v2 [q-bio.GN].

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307–321.

Hahn, M.W. 2018. *Molecular population genetics*. Sunderland, MA: Sinauer Associates.

Haller, B.C. & Messer, P.W. 2017. SLiM 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution* **34**: 230–240.

Hansson, L. 1991. Dispersal and connectivity in metapopulations. *Biological Journal of the Linnean Society* **42**: 89–103.

Herman, A., Brandvain, Y., Weagley, J., Jeffery, W.R., Keene, A.C., Kono, T.J.Y., *et al.* 2018. The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, *Astyanax mexicanus*. *Molecular Ecology* **27**: 4397–4416.

Hey, J. & Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences* **104**: 2785–2790.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**: 518–522.

Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., *et al.* 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist* **188**: 379–397.

James, M.E., Wilkinson, M.J., North, H.L., Engelstädter, J. & Ortiz-Barrientos, D. 2020. A framework to quantify phenotypic and genotypic parallel evolution. BioRxiv [Preprint].

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E. & Butlin, R.K. 2010. Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 1735–1747.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Kautt, A.F., Machado-Schiaffino, G. & Meyer, A. 2016. Multispecies outcomes of sympatric speciation after admixture with the aource population in two radiations of Nicaraguan crater lake cichlids. *PLoS Genet* **12**: e1006157.

Konečná, V., Nowak, M.D. & Kolář, F. 2019. Parallel colonization of subalpine habitats in the central European mountains by *Primula elatior*. *Scientific Reports* **9**: 3294.

Larson, E.L., Vanderpool, D., Sarver, B.A.J., Callahan, C., Keeble, S., Provencio, L.P., *et al.* 2018. The evolution of polymorphic hybrid incompatibilities in house mice. *Genetics* **209**: 845–859.

Lawson, D.J., van Dorp, L. & Falush, D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* **9**: 3258.

Le Moan, A., Gagnaire, P.A. & Bonhomme, F. 2016. Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology* **25**: 3187–3202.

Lee, K.M. & Coop, G. 2017. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics* **207**: 1591–1619.

Lee, K.M. & Coop, G. 2019. Population genomics perspectives on convergent adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**: 20180236.

Lenormand, T., Chevin, L.M. & Bataillon, T. 2016. Parallel evolution: what does it (not) tell us and why is it (still) interesting? In: *Chance in Evolution*. Chicago, USA: University Chicago Press.

Lenormand, T., Roze, D. & Rousset, F. 2009. Stochasticity in evolution. *Trends in Ecology & Evolution* **24**: 157–165.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN].

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liu, S., Ferchaud, A.L., Grønkjaer, P., Nygaard, R. & Hansen, M.M. 2018. Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Molecular Ecology* **27**: 4725–4743.

Lowry, D.B. & Willis, J.H. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology* **8**: e1000500.

Magalhaes, I.S., Whiting, J.R., D'Agostino, D., Hohenlohe, P.A., Mahmud, M., Bell, M.A., *et al.* 2019. Intercontinental genomic parallelism in multiple adaptive radiations. BioRxiv [Preprint]. https://dx.doi.org/10.1101/856344.

Mangiafico, S., S. 2015. *An R companion for the handbook of biological statistics*. https://rcompanion.org/rcompanion/.

Martin, C.H., Cutler, J.S., Friel, J.P., Dening Touokong, C., Coop, G. & Wainwright, P.C. 2015. 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.

Meier, J.I., Sousa, V.C., Marques, D.A., Selz, O.M., Wagner, C.E., Excoffier, L., *et al.* 2017. Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization. *Molecular Ecology* **26**: 123–141.

Melo, M.C., Grealy, A., Brittain, B., Walter, G.M. & Ortiz-Barrientos, D. 2014. Strong extrinsic reproductive isolation between parapatric populations of an Australian groundsel. *New Phytologist* **203**: 323–334.

Melo, M.C., James, M.E., Roda, F., Bernal-Franco, D., Wilkinson, M.J., Liu, H., *et al.* 2019. Evidence for mutation-order speciation in an Australian wildflower. BioRxiv [Preprint]. doi: https://doi.org/10.1101/692673.

Nathan, R., Klein, E., Robledo-Arnuncio, J.J. & Revilla, E. 2012. Dispersal kernels: review. In: *Dispersal ecology and evolution*. Oxford, UK: Oxford University Press.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

North, H.L. 2015. Divergent selection maintains parapatric ecotypes in effective panmixia. Honours Thesis. The University of Queensland.

Nosil, P., Crespi, B.J. & Sandoval, C.P. 2002. Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature* **417**: 440–443.

O'Leary, S.J., Puritz, J.B., Willis, S.C., Hollenbeck, C.M. & Portnoy, D.S. 2018. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* **27**: 3193–3206.

Oneal, E., Lowry, D.B., Wright, K.M., Zhu, Z. & Willis, J.H. 2014. Divergent population structure and climate associations of a chromosomal inversion polymorphism across the *Mimulus guttatus* species complex. *Molecular Ecology* **23**: 2844–2860.

Orr, H.A. 2005. The probability of parallel evolution. *Evolution* **59**: 216–220.

Ostevik, K.L., Moyers, B.T., Owens, G.L. & Rieseberg, L.H. 2012. Parallel ecological speciation in plants? *International Journal of Ecology* **2012**: 1–17.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Pardo-Diaz, C., Salazar, C. & Jiggins, C.D. 2015. Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution* **6**: 445–464.

Pérez-Pereira, N., Quesada, H. & Caballero, A. 2017. Can parallel ecological speciation be detected with phylogenetic analyses? *Molecular Phylogenetics and Evolution* **116**: 149–156.

Perreault-Payette, A., Muir, A.M., Goetz, F., Perrier, C., Normandeau, E., Sirois, P., *et al.* 2017. Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. *Molecular Ecology* **26**: 1477–1497.

Pickrell, J.K. & Pritchard, J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**: e1002967.

Poland, J.A., Brown, P.J., Sorrells, M.E. & Jannink, J.L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme Genotyping-by-Sequencing approach. *PLoS ONE* **7**: e32253.

Popescu, A.A., Huber, K.T. & Paradis, E. 2012. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**: 1536–1537.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., *et al.* 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559–575.

Puritz, J.B., Hollenbeck, C.M. & Gold, J.R. 2014a. *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**: e431.

Puritz, J.B., Matz, M.V., Toonen, R.J., Weber, J.N., Bolnick, D.I. & Bird, C.E. 2014b. Demystifying the RAD fad. *Molecular Ecology* **23**: 5937–5942.

Quesada, H., Posada, D., Caballero, A., Morán, P. & Rolán-Alvarez, E. 2007. Phylogenetic evidence for multiple sympatric ecological diversification in a marine snail. *Evolution* **61**: 1600–1612.

R Core Team. 2017. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Radford, I.J., Cousens, R.D. & Michael, P.W. 2004. Morphological and genetic variation in the *Senecio pinnatifolius* complex: are variants worthy of taxonomic recognition? *Australian Systematic Botany* **17**: 29–48.

Raj, A., Stephens, M. & Pritchard, J.K. 2014. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Ravinet, M., Prodöhl, P.A. & Harrod, C. 2013. Parallel and nonparallel ecological, morphological and genetic divergence in lake-stream stickleback from a single catchment. *Journal of Evolutionary Biology* **26**: 186–204.

Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. 2009. Reconstructing Indian population history. *Nature* **461**: 489–494.

Richards, T.J. & Ortiz-Barrientos, D. 2016. Immigrant inviability produces a strong barrier to gene flow between parapatric ecotypes of *Senecio lautus*. *Evolution* **70**: 1239–1248.

Richards, T.J., Walter, G.M., McGuigan, K. & Ortiz-Barrientos, D. 2016. Divergent natural selection drives the evolution of reproductive isolation in an Australian wildflower. *Evolution* **70**: 1993–2003.

Roda, F., Ambrose, L., Walter, G.M., Liu, H.L., Schaul, A., Lowe, A., *et al.* 2013. Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. *Molecular Ecology* **22**: 2941–2952.

Roesti, M., Kueng, B., Moser, D. & Berner, D. 2015. The genomics of ecological vicariance in threespine stickleback fish. *Nat Commun* **6**: 8767.

Rogers, S.M. & Bernatchez, L. 2006. The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (*Coregonus clupeaformis*). *Journal of Evolutionary Biology* **19**: 1979–1994.

Rohland, N. & Reich, D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* **22**: 939–946.

Rosenblum, E.B., Parent, C.E. & Brandt, E.E. 2014. The molecular basis of phenotypic convergence. *Annual Review of Ecology, Evolution, and Systematics* **45**: 203–226.

Rougemont, Q., Gaigher, A., Lasne, E., Côte, J., Coke, M., Besnard, A.L., *et al.* 2015. Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys. *Journal of Evolutionary Biology* **28**: 2248–2263.

Rougeux, C., Bernatchez, L. & Gagnaire, P.A. 2017. Modeling the multiple facets of speciation-with-gene-flow toward Inferring the divergence history of lake whitefish species pairs (Coregonus clupeaformis). *Genome Biology and Evolution* **9**: 2057–2074.

Rougeux, C., Gagnaire, P.A. & Bernatchez, L. 2019. Model-based demographic inference of introgression history in European whitefish species pairs'. *Journal of Evolutionary Biology* **32**: 806–817.

Salazar-Ciudad, I. & Marín-Riera, M. 2013. Adaptive dynamics under development-based genotype–phenotype maps. *Nature* **497**: 361–364.

Schluter, D. & Nagel, L.M. 1995. Parallel speciation by natural selection. *The American Naturalist* **146**: 292–301.

Schmieder, R., Lim, Y.W., Rohwer, F. & Edwards, R. 2010. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**: 341.

Schrempf, D., Minh, B.Q., De Maio, N., von Haeseler, A. & Kosiol, C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology* **407**: 362–370.

Scopece, G., Lexer, C., Widmer, A. & Cozzolino, S. 2010. Polymorphism of postmating reproductive isolation within plant species. *TAXON* **59**: 1367–1374.

Slatkin, M. 1985a. Gene flow in natural populations. *Annual Review of Ecology and Systematics* **16**: 393–430.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**: 264–279.

Slatkin, M. 1985b. Rare alleles as indicators of gene flow. *Evolution* **39**: 53–65.

Smith, S.D. & Rausher, M.D. 2011. Gene loss and parallel evolution contribute to species difference in flower color. *Molecular Biology and Evolution* **28**: 2799–2810.

Stankowski, S. 2013. Ecological speciation in an island snail: evidence for the parallel evolution of a novel ecotype and maintenance by ecologically dependent postzygotic isolation. *Molecular Ecology* **22**: 2726–2741.

Stern, D.L. & Orgogozo, V. 2009. Is genetic evolution predictable? *Science* **323**: 746–751.

Stoltzfus, A. & McCandlish, D.M. 2017. Mutational biases influence parallel adaptation. *Molecular Biology and Evolution* **34**: 2163–2172.

Storz, J.F. & Wheat, C.W. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* **64**: 2489–2509.

Strasburg, J.L. & Rieseberg, L.H. 2010. How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. *Molecular Biology and Evolution* **27**: 297–310.

Strasburg, J.L. & Rieseberg, L.H. 2008. 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.

Sweigart, A.L. & Flagel, L.E. 2015. Evidence of natural selection acting on a polymorphic hybrid incompatibility locus in *Mimulus*. *Genetics* **199**: 543–554.

Tan, A., Abecasis, G.R. & Kang, H.M. 2015. Unified representation of genetic variants. *Bioinformatics* **31**: 2202–2204.

Thomé, M.T.C. & Carstens, B.C. 2016. Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proceedings of the National Academy of Sciences* **113**: 8010–8017.

Thompson, I.R. 2005. Taxonomic studies of Australian *Senecio* (Asteraceae): 5. The *S. pinnatifolius/S. lautus* complex. *Muelleria* 23–76.

Tiffin, P. & Ross-Ibarra, J. 2014. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution* **29**: 673–680.

Trucchi, E., Frajman, B., Haverkamp, T.H.A., Schönswetter, P. & Paun, O. 2017. Genomic analyses suggest parallel ecological divergence in *Heliosperma pusillum* (Caryophyllaceae). *New Phytologist* **216**: 267–278.

Walter, G.M., Aguirre, J.D., Blows, M.W. & Ortiz-Barrientos, D. 2018a. Evolution of genetic variance during adaptive radiation. *The American Naturalist* **191**: E108–E128.

Walter, G.M., Wilkinson, M.J., Aguirre, J.D., Blows, M.W. & Ortiz-Barrientos, D. 2018b. Environmentally induced development costs underlie fitness tradeoffs. *Ecology* **99**: 1391–1401.

Walter, G.M., Wilkinson, M.J., James, M.E., Richards, T.J., Aguirre, J.D. & Ortiz-Barrientos, D. 2016. Diversification across a heterogeneous landscape. *Evolution* **70**: 1979–1992.

White, E. 2008. Indirect interactions between alien and native *Senecio* species as mediated by insects. PhD Thesis. Queensland University of Technology.

Wilkinson, M.J. 2019. The genetic basis of adaptive evolution and divergence in an Australian wildflower. PhD Thesis. The University of Queensland.

Wilkinson, M.J., Roda, F., Walter, G.M., James, M.E., Nipper, R., Walsh, J., *et al.* 2019. Divergence in hormone signalling links local adaptation and hybrid failure. BioRxiv [Preprint]. https://doi.org/10.1101/845354.

Wright, S. 1943. Isolation by distance. *Genetics* **28**: 114–138.

Yeaman, S. 2015. Local adaptation by alleles of small effect. *The American Naturalist* **186**: S74–S89.

Yeaman, S., Gerstein, A.C., Hodgins, K.A. & Whitlock, M.C. 2018. Quantifying how constraints limit the diversity of viable routes to adaptation. *PLOS Genetics* **14**: e1007717.

York, R.A. & Fernald, R.D. 2017. The repeated evolution of behavior. *Frontiers in Ecology and Evolution* **4**: 143.

Zuellig, M.P. & Sweigart, A.L. 2018. A two-locus hybrid incompatibility is widespread, polymorphic, and active in natural populations of *Mimulus*. *Evolution* **72**: 2394–2405.

Figures

Figure 1. The colonisation history and phylogenetic topology for alternate origin scenarios

Schematic diagram representing the colonisation 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 colonisation. 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 orange population. Populations phylogenetically cluster according to their geographic distribution, which can be indistinguishable from a single origin with gene flow scenario **(B)**.

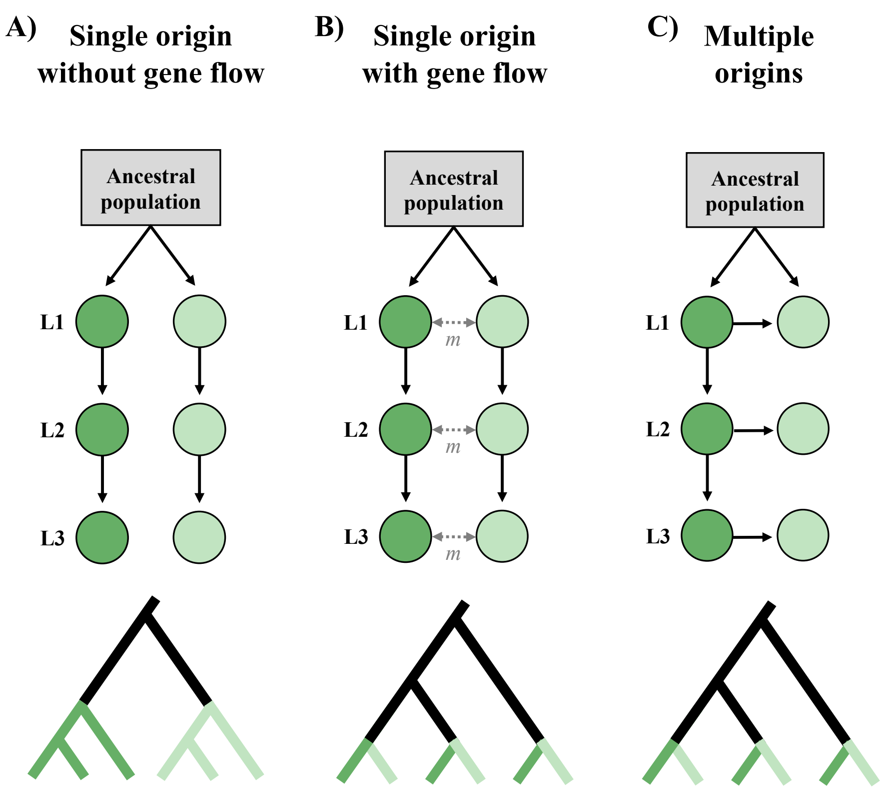


Figure 2. Sampling locations and genetic clustering of *Senecio lautus* populations

(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 colours representing ancestry proportions to each cluster. Populations are ordered according to their geographic distribution along the coast.

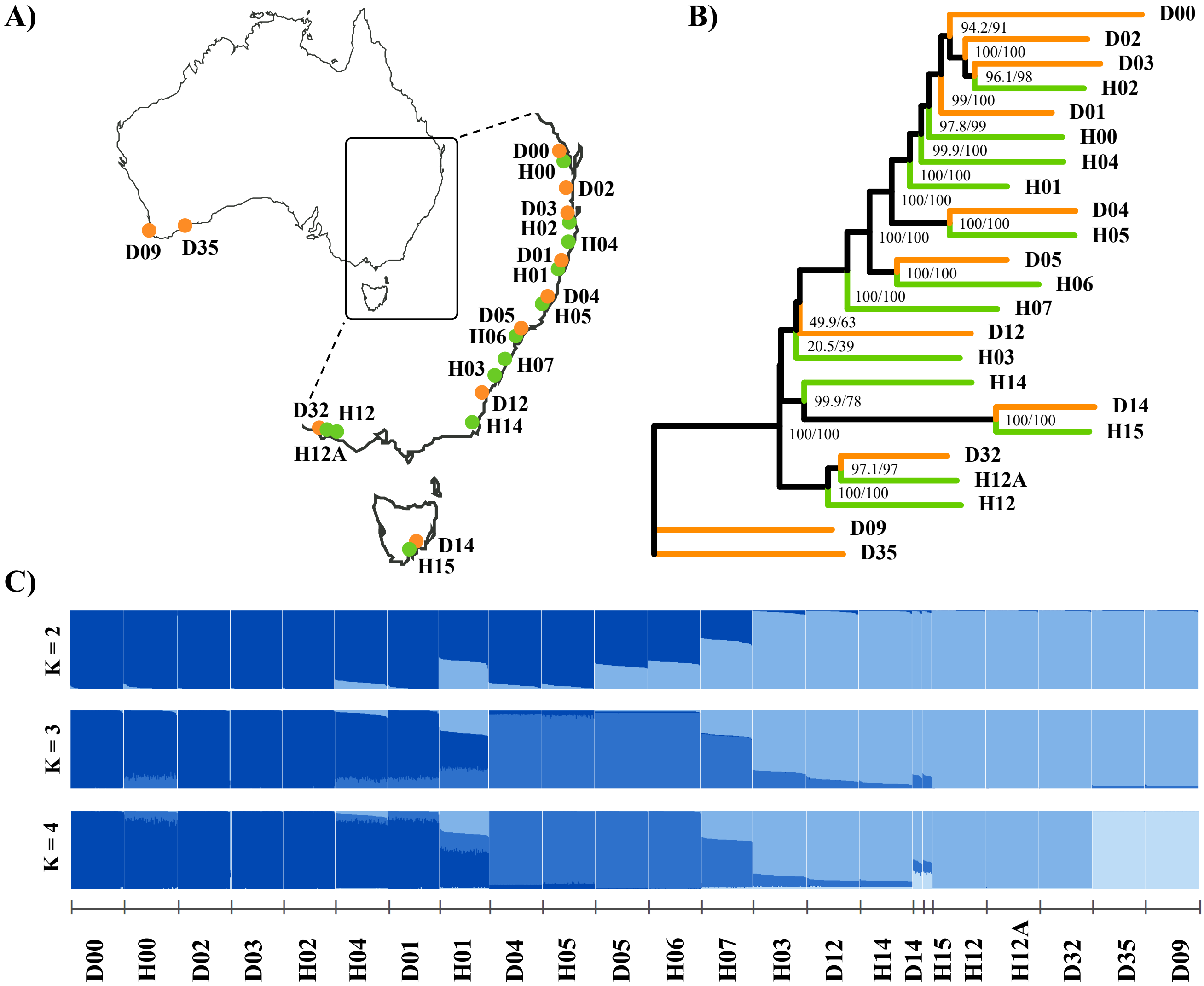


Figure 3. Patterns of long-distance gene flow, IBD, and genetic clustering

(A) Maximum likelihood tree with one migration event inferred in *TreeMix*, the x-axis representing genetic drift. The arrow represents the migration event (w = 0.40, P < 2.2×10^{-308}). **(B)** Patterns of isolation by distance across Dune and Headland populations for Dune-Headland pairs (black), Dune-Dune (orange) and Headland-Headland (green). Average migration rate is the mean bidirectional migration for each pair. Grey shading represents the null model for migration rates, inferred from the maximum migration value from three allopatric comparisons. Grey horizontal dashed line represents migration (2Nm) of one. Pairs falling above this line are labelled. Black dashed line represents the linear model for the DH comparisons. **(C)** Relationship between geographic distance and divergence time for parapatric Dune-Headland pairs (black), Dune-Dune (orange) and Headland-Headland (green). 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 colours representing ancestry proportions to each cluster. Populations are ordered according to their geographic distribution along the coast.

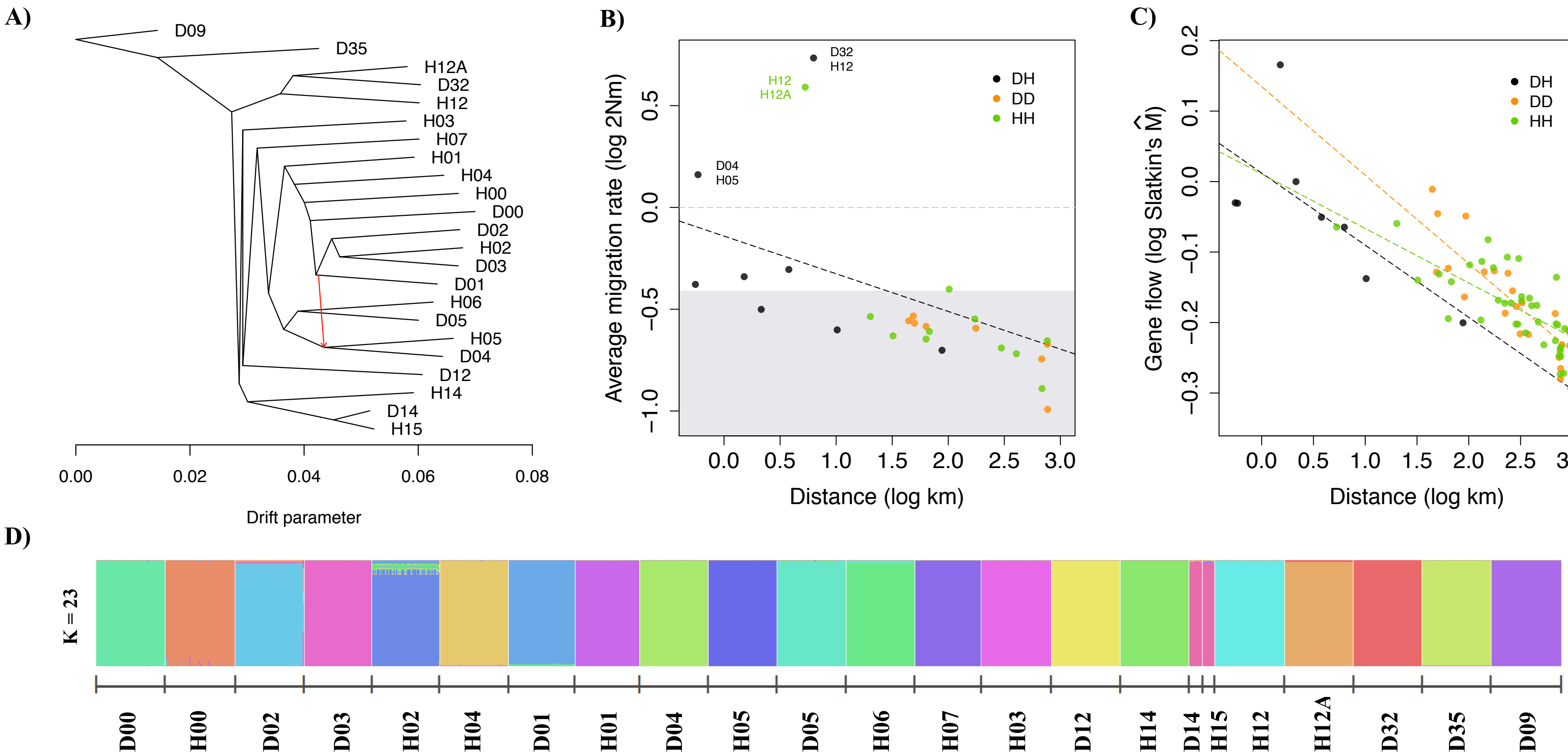


Figure 4. Patterns of gene flow and admixture between parapatric Dune-Headland populations

(A) Schematic diagram representing the eight demographic models and estimated parameters in *fastsimcoal2*: 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 (orange) and Headland (green) ecotypes at each locality. Each individual is depicted as a bar, with colours representing ancestry proportions to each cluster. Below are the migration rates (*m*) from the Dune to Headland, and Headland to Dune within each locality estimated within *fastsimcoal2*. Asterisks denote pairs with 2Nm > 1.

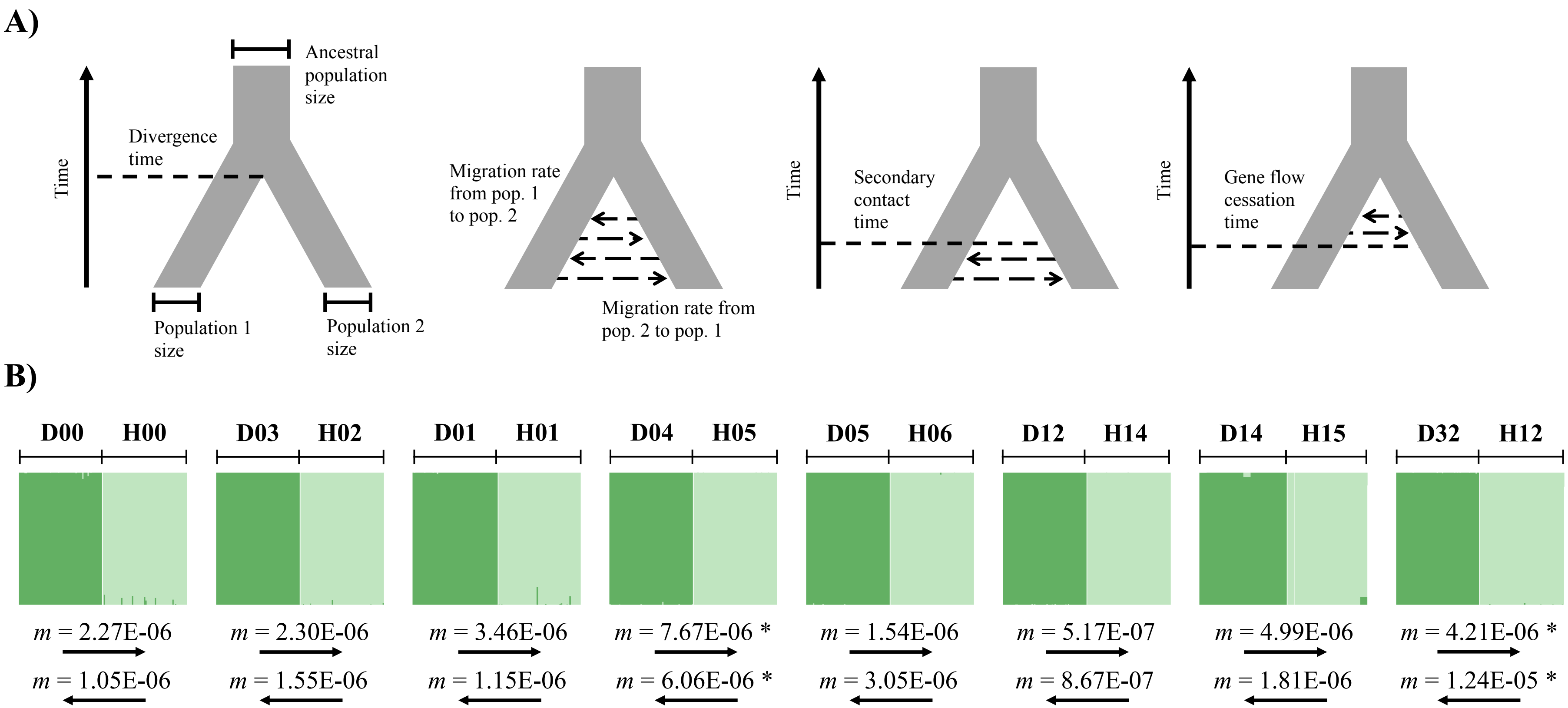
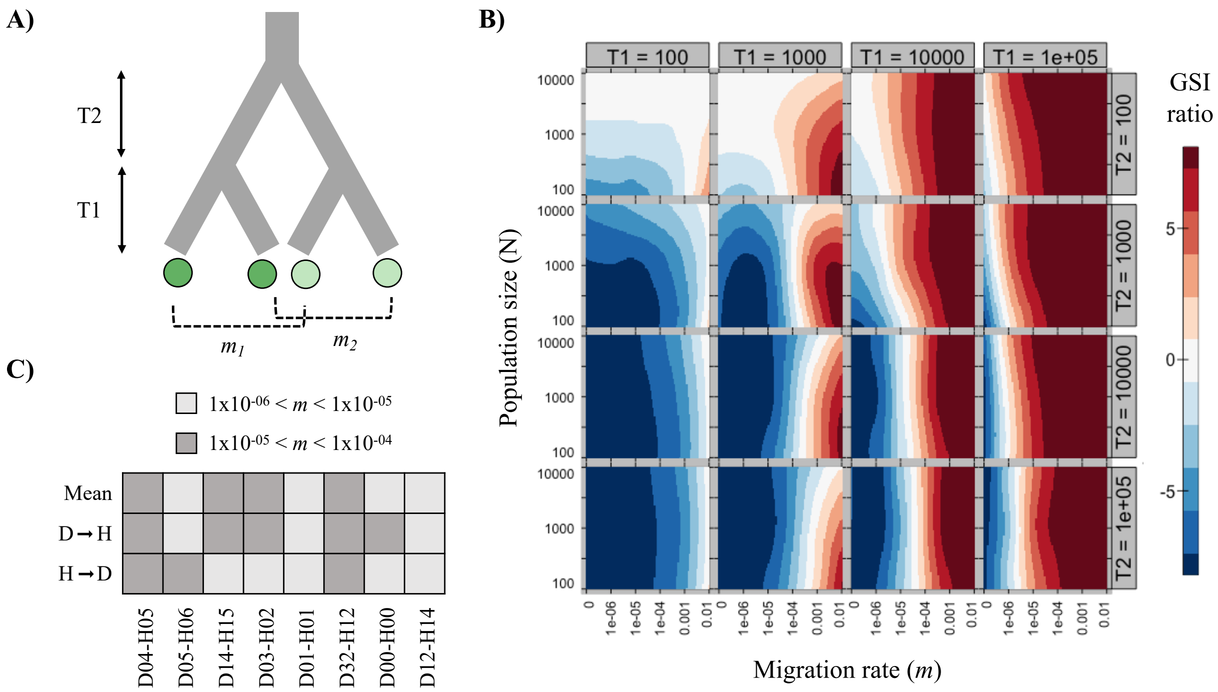


Figure 5. Forward population genetic simulations in *SLiM*

(A) Schematic diagram representing the single origin scenario simulated in *SLiM*. *T1* represents the time from the present to the second split, *T2* represents the time from the second split to the split of the ancestral population. Gene flow (denoted by dotted lines) occurs between populations within each locality. Light green and dark green circles represent populations from different ecotypes. **(B)** Logarithm of the ratio of mean Genealogical Sorting Index (GSI) calculated for parapatric populations, over mean GSI for sister populations of the same ecotype from a reconstructed phylogeny from simulation output in *SLiM2*. Increasingly positive GSI values (red) denote increasing levels of polyphyly due to distortion of the single origin scenario, whereas increasingly negative GSI values (blue) denote increasing levels of the true signal of monophyly due to lack of gene flow between ecotypes. **(C)** Summary of observed migration rates in *fastsimcoal2* for the eight replicate parapatric pairs for mean bidirectional migration (mean), Dune to Headland migration (D→H), and Headland to Dune migration (H→D). Light grey boxes indicate migration rate (*m*) values between $1x10^{-06}$ and $1x10^{-05}$, corresponding to regions of the parameter space in **(B)**, bottom right panel, where phylogeny distortion is unlikely. Dark grey boxes indicate migration rates between $1x10^{-05}$ and $1x10^{-04}$, corresponding to regions of the parameter space where there is some concern for phylogenetic distortion. Population pairs are ordered from left to right in accordance with increasing geographic distance between ecotypes.



Supplementary tables and figures

Table S1. 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 those 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	QLD: Lennox Head	Dune	D01-H01	S28° 46.858' E153° 35.655'	60
Eastern	H01	QLD: 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

Table S2. Sequencing and alignment summary for *Senecio lautus* individuals

Summary statistics for the 23 populations used within the study. Excluded from the table are the 19 individuals removed due to high missing data.

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)

Table S3. Pairwise F_{ST} for *S. lautus* populations

Pairwise FST values between all 21 populations of the south 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	-

Table S4. Estimation of gene flow rates and population 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. 2NmP1->P2: gene flow (2Nm) from population 1 to population 2. 2NmP2->P1: gene flow (2Nm) from population 2 to population 1. Values in bold represent 2Nm > 1.

Comparison	Populations	Asize	Pop1size	Pop2size	DivTime	SecTime	2NmP1->P2	2NmP2->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

Table S5. Bootstrap values for gene flow estimates inferred in *fastsimcoal2*

Populations: the two populations used for each comparison (population 1 is on the left, and population 2 on the right. 2NmP1->P2min and max: lower and upper 95% confidence intervals for gene flow from population 1 to population 2, respectively. 2NmP2->P1min and max: lower and upper 95% confidence intervals for gene flow from population 2 to population 1, respectively. Populations in bold represent 2Nm > 1.

Comparison	Populations	2NmP1->P2min	2NmP1->P2max	2NmP2->P1min	2NmP2->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

Figure 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 variance explain for each model for 1-25 migration events.

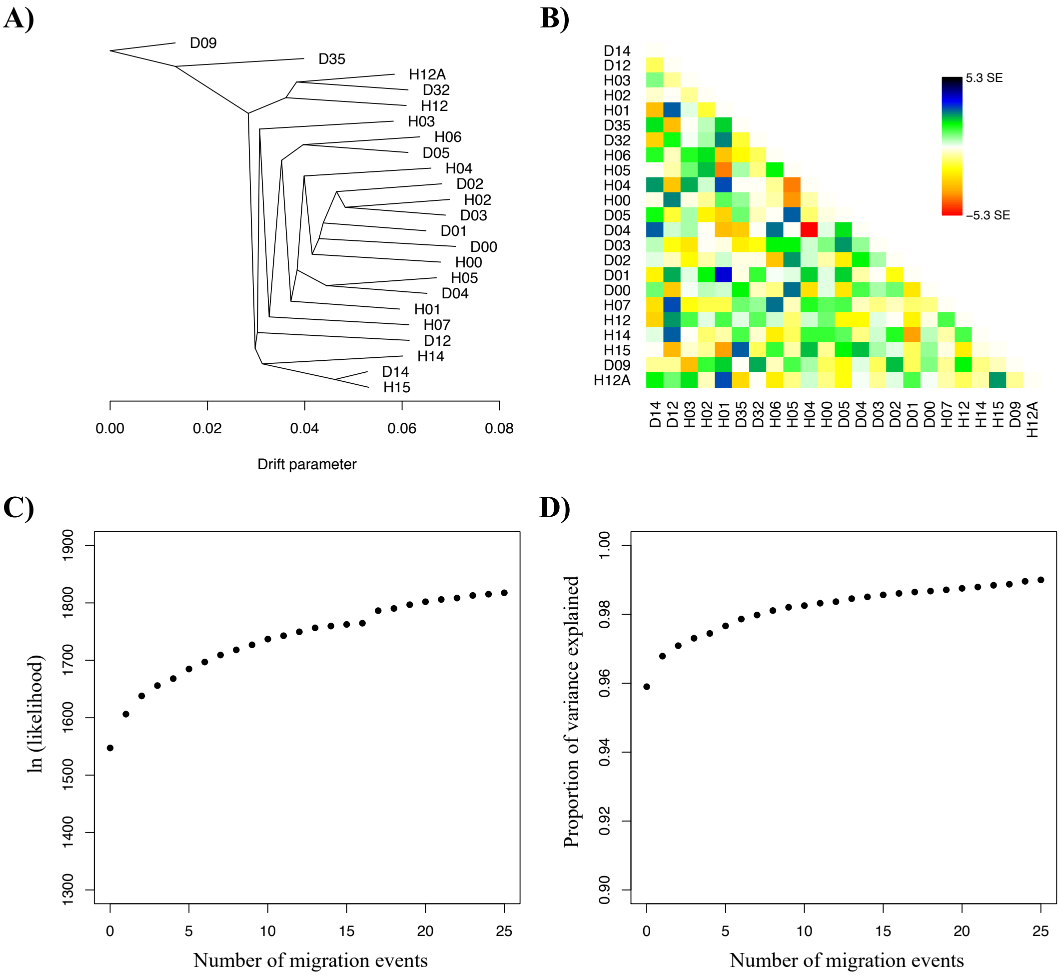

A)
B)
C)
D)
D09
D35
H12A
D32
H12
H03
H06
D05
H04
D02
H02
D03
D01
D00
H00
H05
D04
H01
H07
D12
H14
D14
H15
0.00
0.02
0.04
0.06
0.08
Drift parameter
5.3 SE
−5.3 SE
D14 D12 H03 H02 H01 D35 D32 H06 H05 H04 H00 D05 D04 D03 D02 D01 D00 H07 H12 H14 H15 D09 H12A
ln (likelihood)
1300
1400
1500
1600
1700
1800
1900
Number of migration events
0
5
10
15
20
25
Proportion of variance explained
0.90
0.92
0.94
0.96
0.98
1.00
Number of migration events

Figure S2. TreeMix migration events 1-10

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

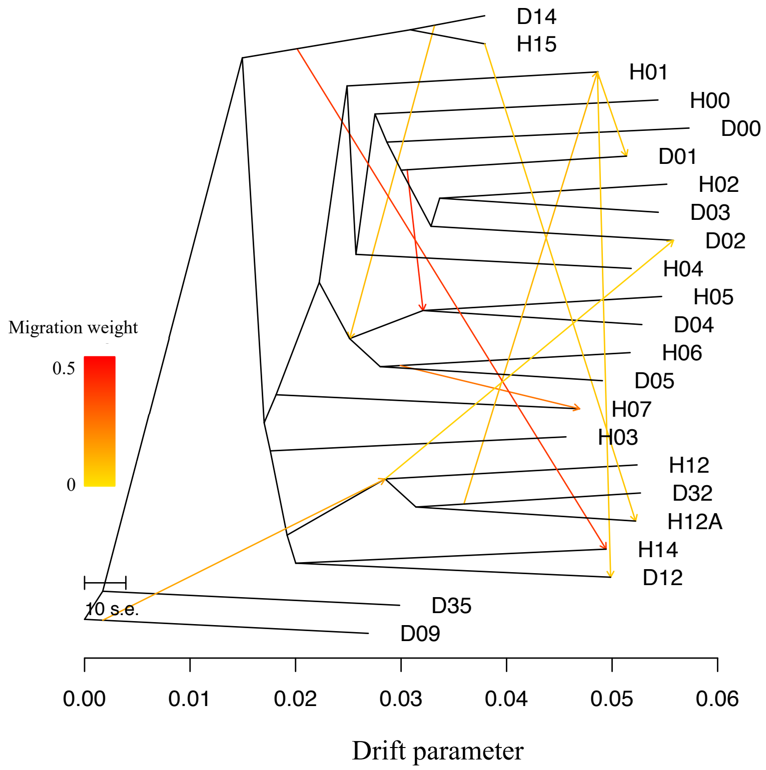


Figure S3. Frequency distribution of *f3*-statistics

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

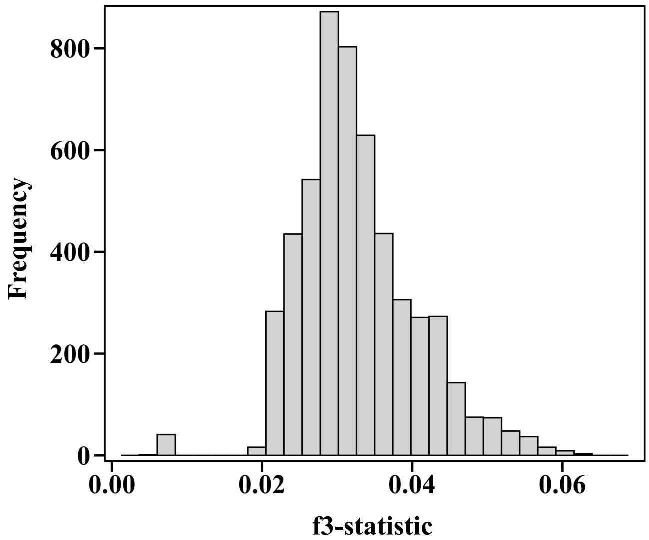


Figure S4. ***fastSTRUCTURE* K=22, K=28 and marginal likelihoods**

(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 colours 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 maximised 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 with negligible change in likelihood values.

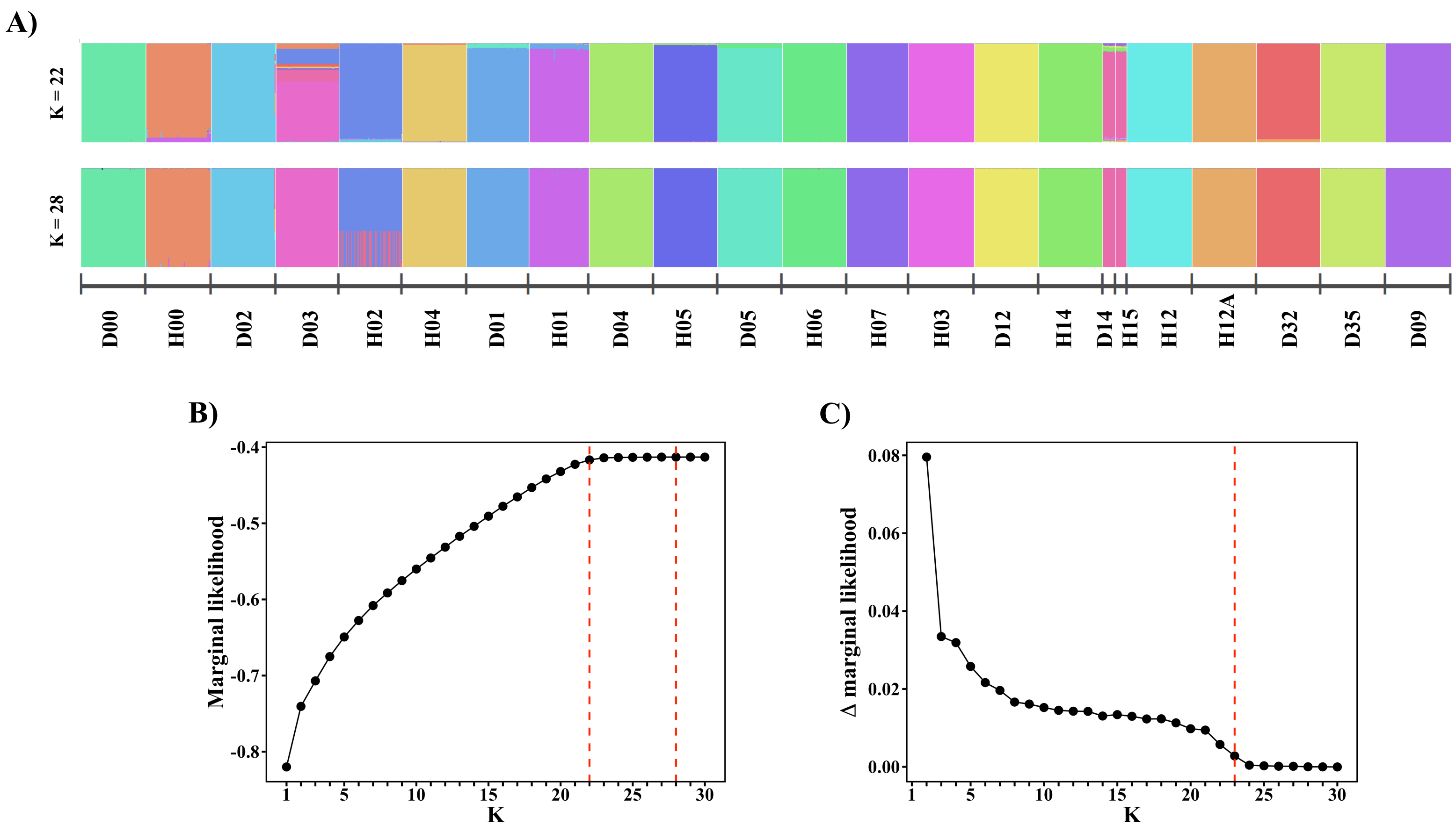


Figure S5. *STRUCTURE* best K-values for Dune-Headland pairs

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).

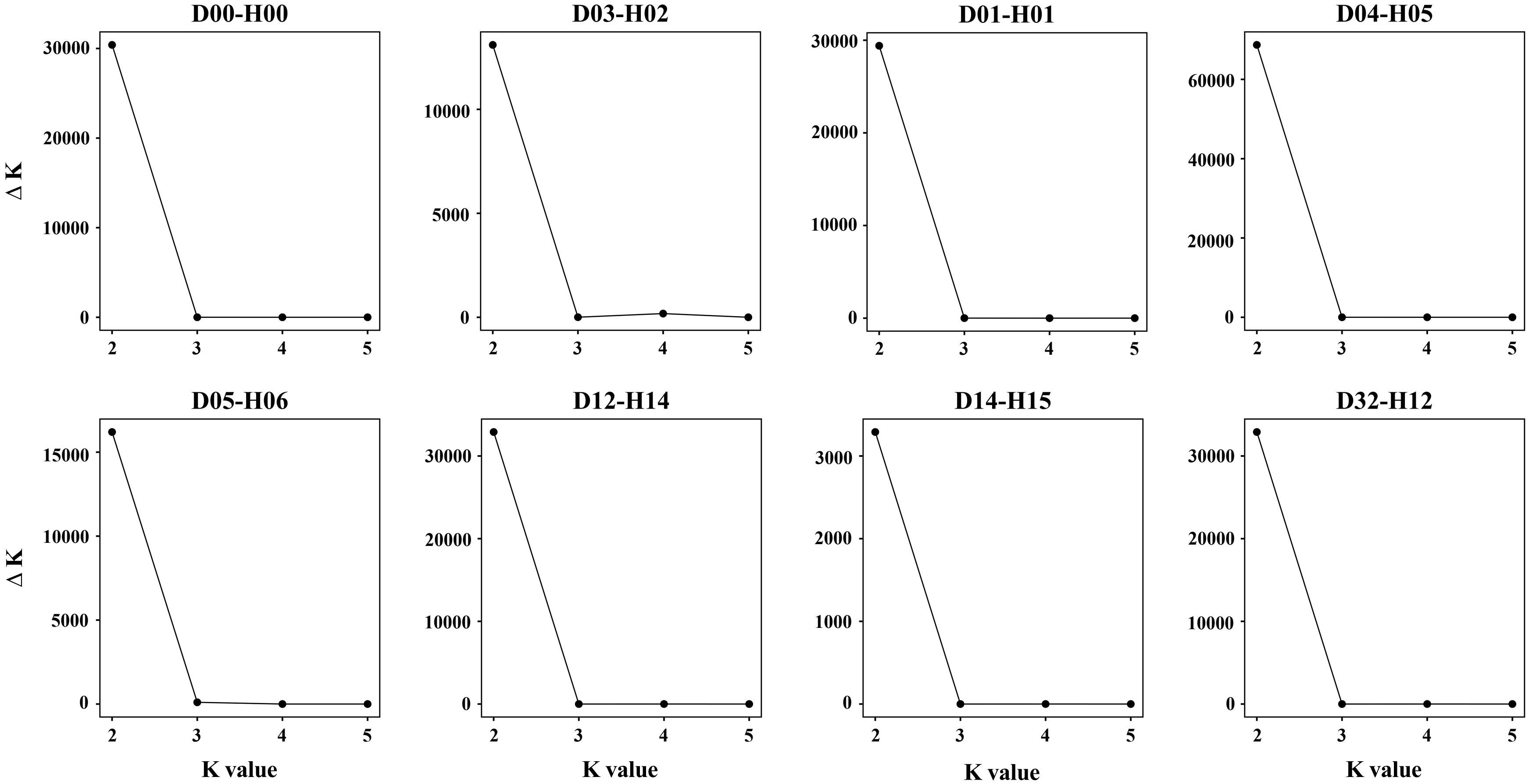


Figure S6. Log-likelihood values for the eight demographic models tested in *fastsimcoal2* per pair

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.

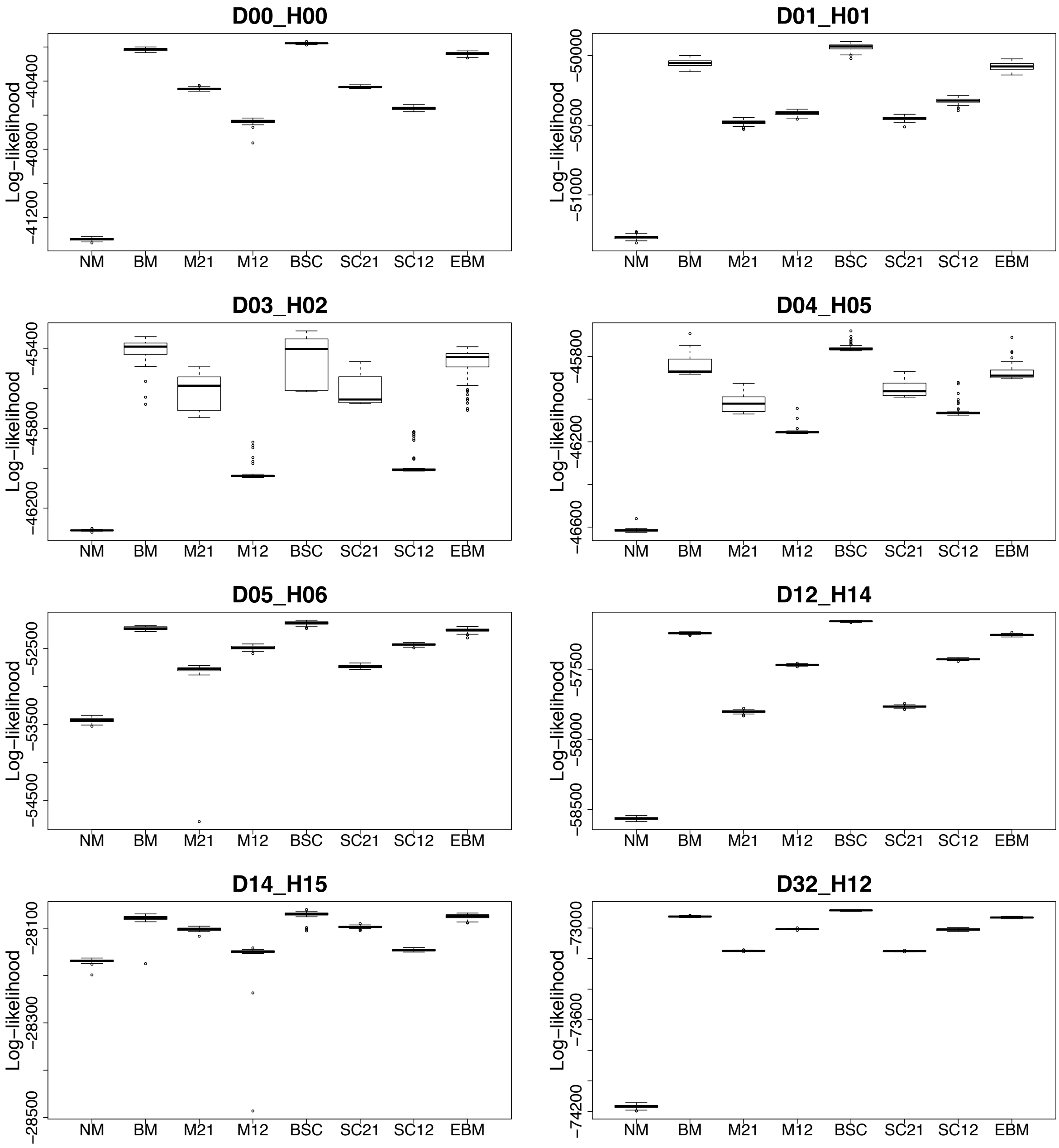


Figure S6 cont.

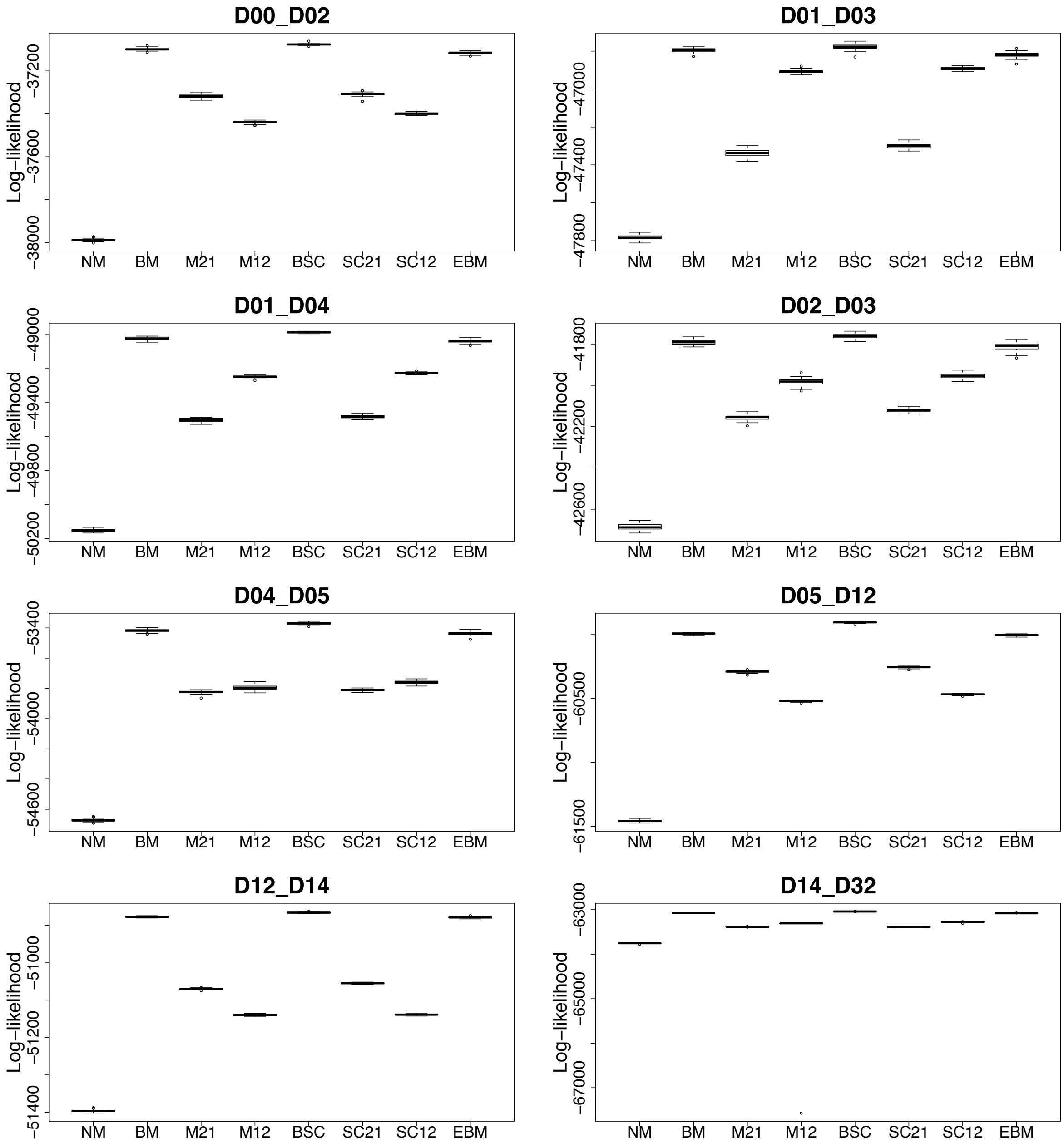


Figure S6 cont.

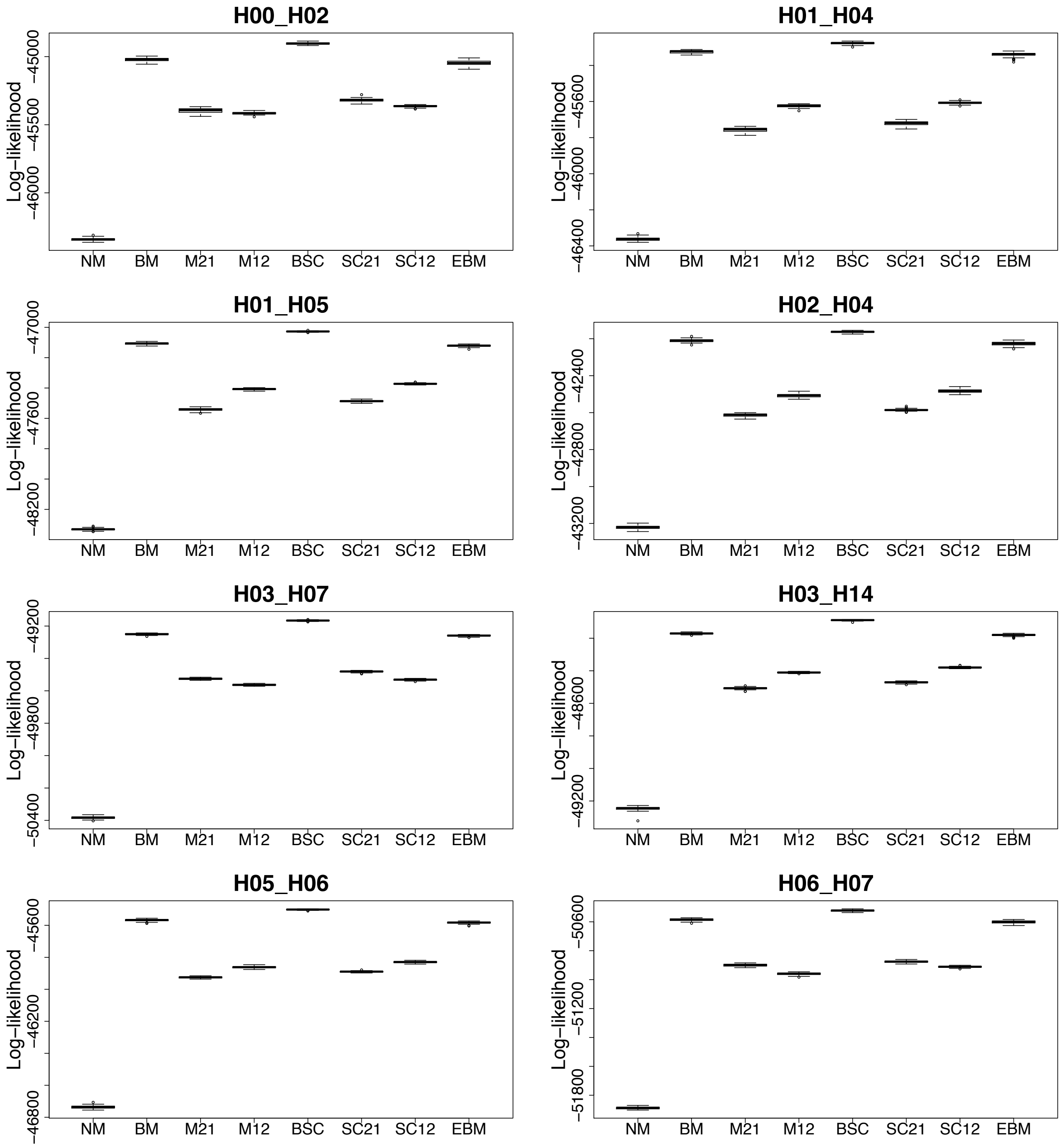


Figure S6 cont.

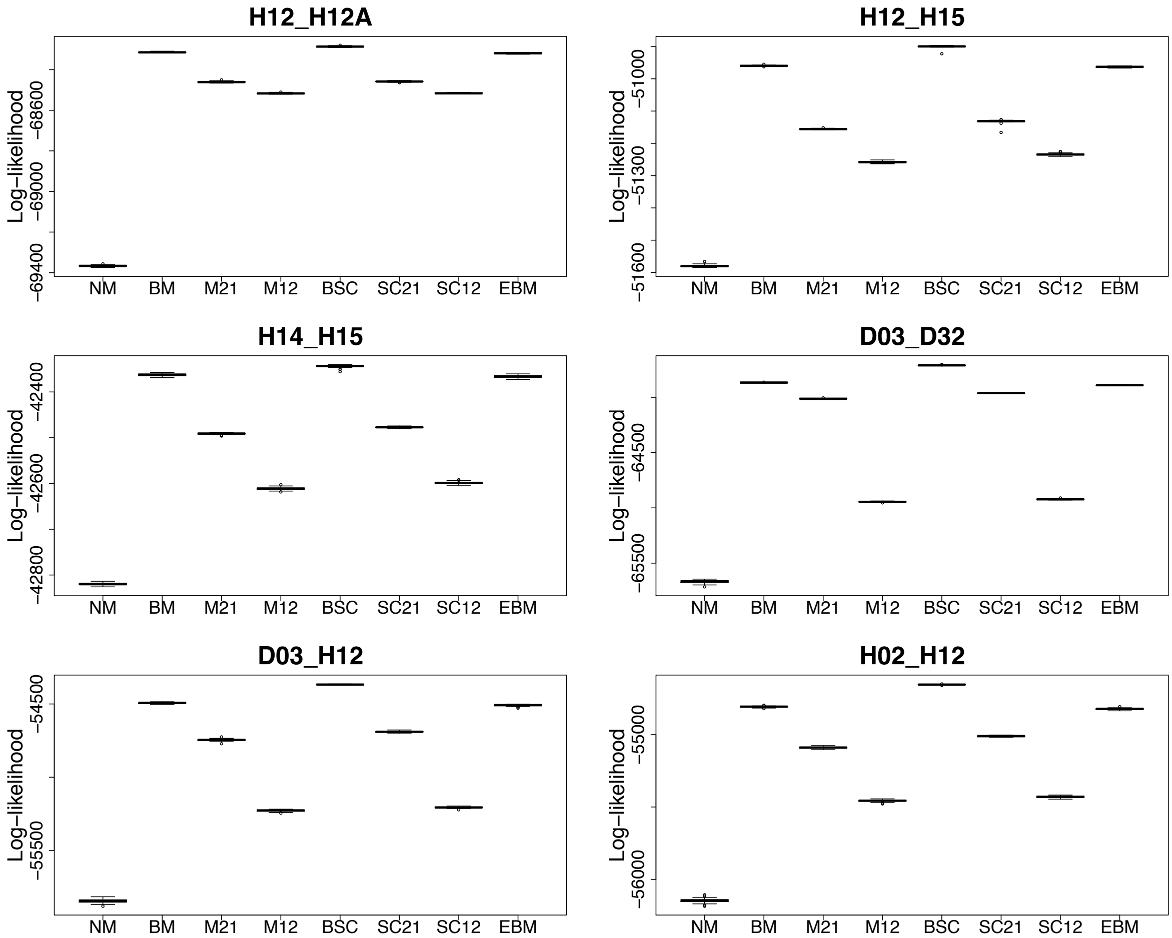


Figure S7. Likelihood values for testing if migration is significantly different from zero

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

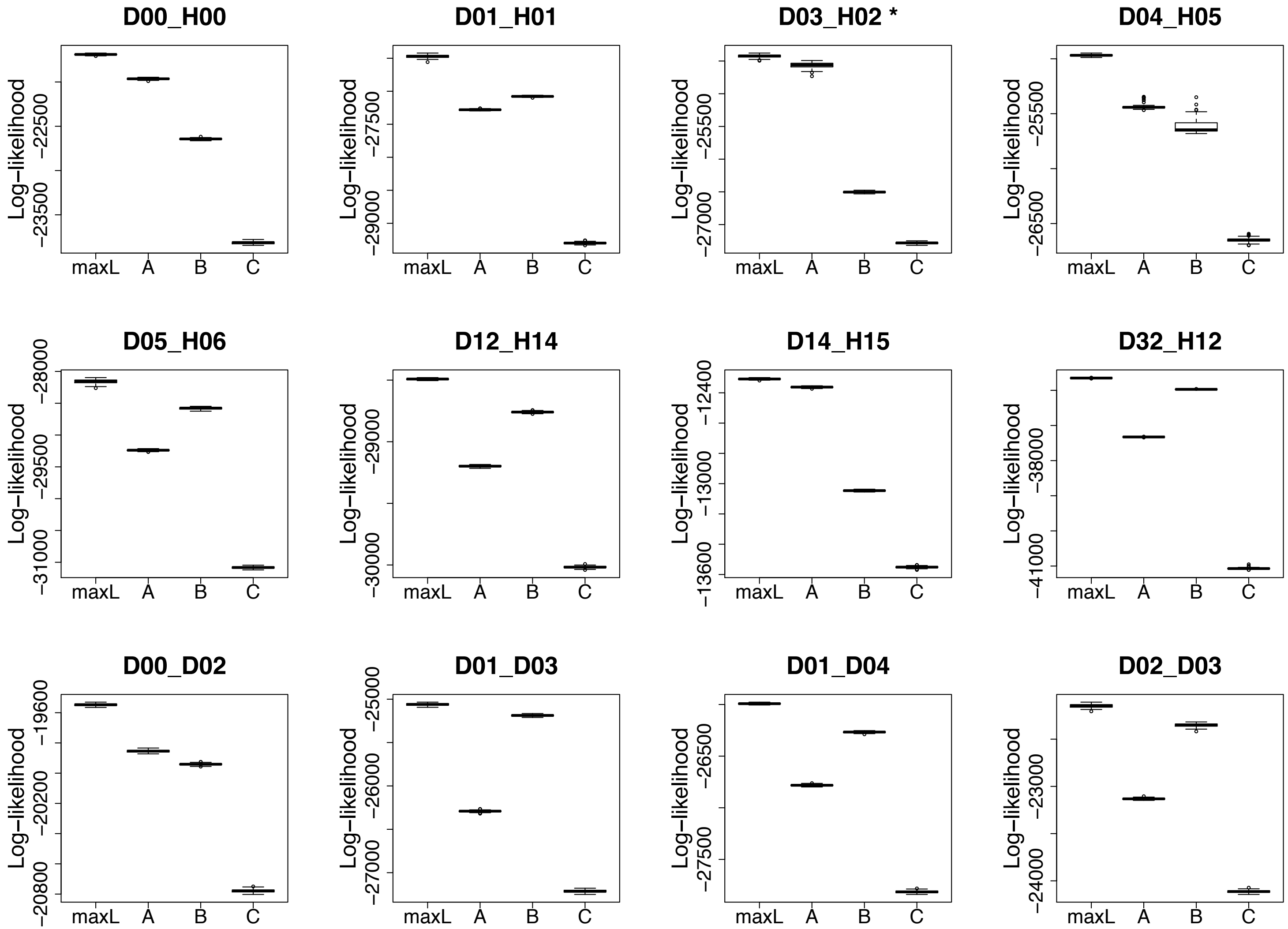


Figure S7 cont.

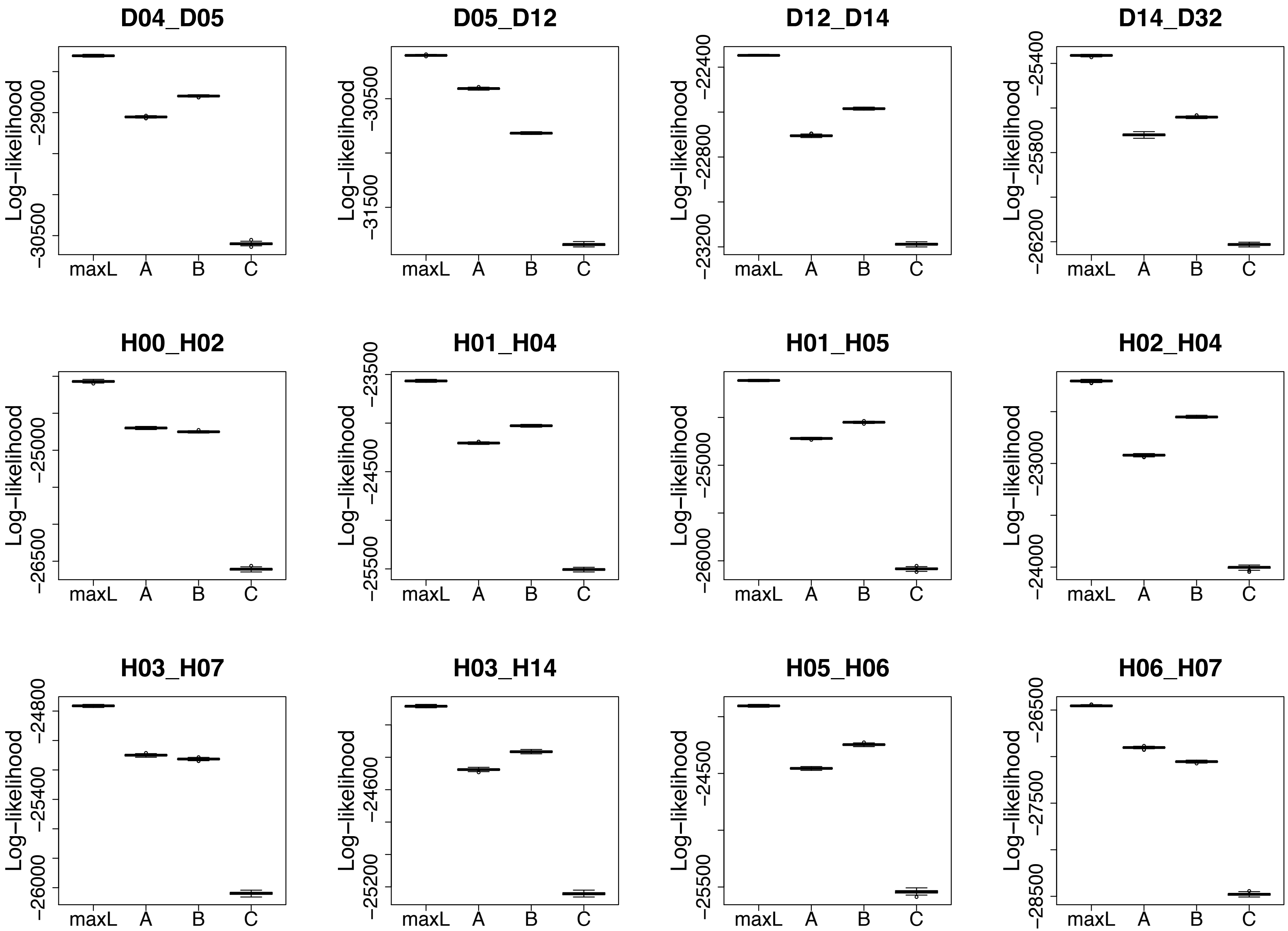
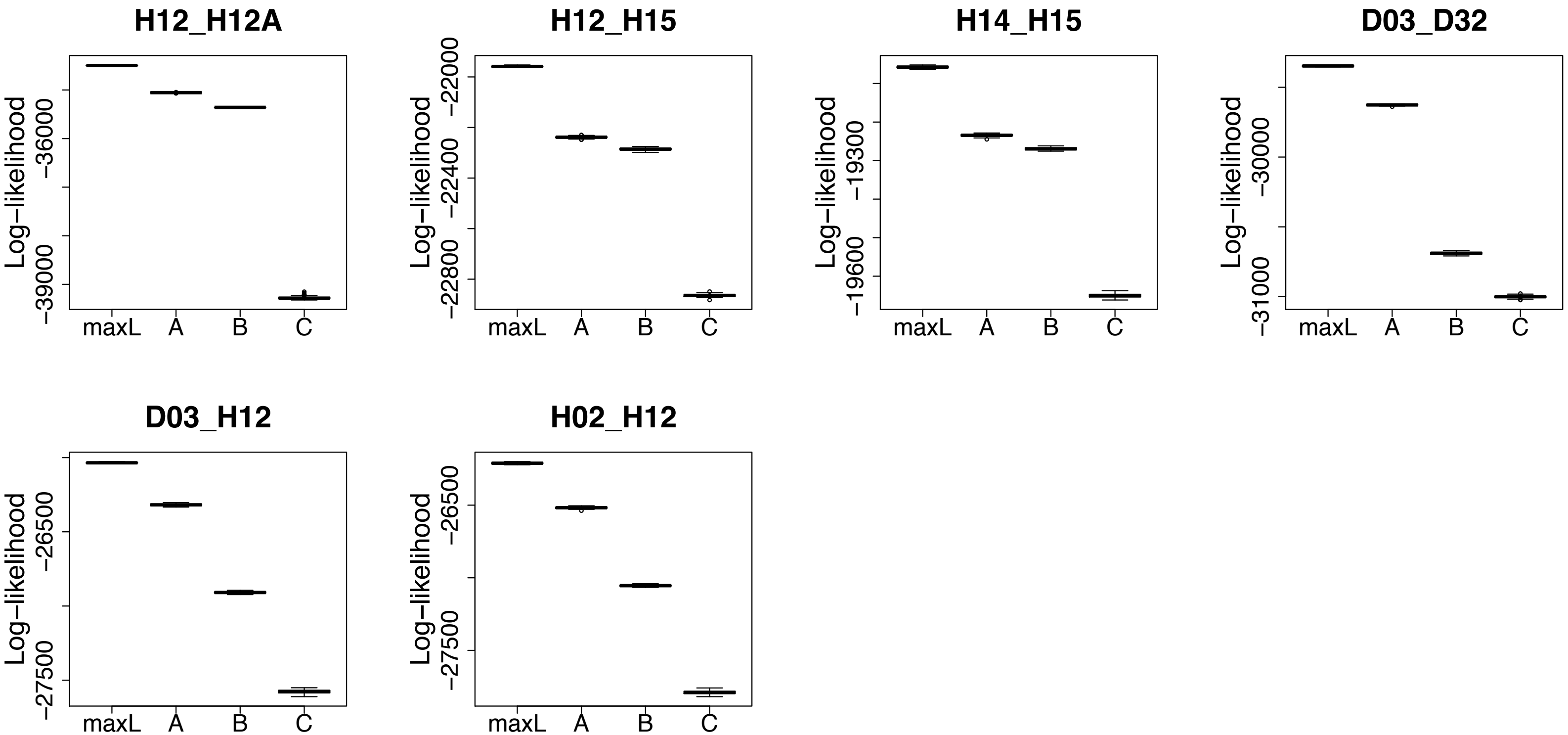


Figure S7 cont.



Supplementary methods

Estimation of the number of monomorphic sites per pair

To estimate the 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).

SLiM2 code for simulations

```
initialize() {
	if (exists("slimgui")) {
		defineConstant("seed", 1);
		defineConstant("mu", 1e-7);
		defineConstant("r", 1e-8);

		defineConstant("N", 100);
		defineConstant("t1", 10000);
		defineConstant("t2", 5000);
		defineConstant("mig", 0);
		defineConstant("tSec", 0.1);
		defineConstant("outPath", "~/workspace/PopGenSims/OriginScenarios");
	}

	setSeed(seed);
	initializeMutationRate(mu);
	initializeMutationType("m1", 0.5, "f", 0.0);
	initializeGenomicElementType("g1", m1, 1.0);
	initializeGenomicElement(g1, 0, 1e6 - 1);
	initializeRecombinationRate(r);
}

1 {

	// create ancestral population
	sim.addSubpop("p0", N);

	// schedule split and migration events based on parameter values
	t0 = 10*N;
	outGen = t0+t1+t2;

	sim.rescheduleScriptBlock(s1, start=t0+1, end=t0+1);
	sim.rescheduleScriptBlock(s2, start=t0+t1+1, end=t0+t1+1);
	sim.rescheduleScriptBlock(s4, start=outGen, end=outGen);

	if(mig == 0){
		sim.deregisterScriptBlock(s3);
	} else
```

```
		sim.rescheduleScriptBlock(s3, start=asInteger(t0+t1+1+round(t2*(1-tSec))), end=outGen);

}

s1 10 {

		// initial split of ecotypes here
		sim.addSubpopSplit("p10", N, p0);
		sim.addSubpopSplit("p100", N, p0);
		p0.setSubpopulationSize(10);
}

s2 20 {

		// output t1 vcf here
		outgroup = sample(p0.individuals, 1);
		ingroup = sapply(sim.subpopulations[sim.subpopulations != p0], "sample(applyValue.individuals, 30, replace = F);");
		set = c(outgroup, ingroup);
		set.genomes.outputVCF(filePath = paste(c(outPath, "/", "t1.replicate-", seed, ".vcf"), sep=""), outputMultiallelics = F);

		// subsequent split into populations here
		sim.addSubpopSplit("p11", N, p10);
		sim.addSubpopSplit("p101", N, p100);
}

s3 30 {

		// initialize migration here between parapatric divergent ecotypes
		p10.setMigrationRates(p100, mig);
		p100.setMigrationRates(p10, mig);
		p11.setMigrationRates(p101, mig);
		p101.setMigrationRates(p11, mig);
}

s4 40 late() {

		// output final vcf here
		outgroup = sample(p0.individuals, 1);
		ingroup = sapply(sim.subpopulations[sim.subpopulations != p0], "sample(applyValue.individuals, 30, replace = F);");
		set = c(outgroup, ingroup);
		set.genomes.outputVCF(filePath = paste(c(outPath, "/", "t2.replicate-", seed, ".vcf"), sep=""), outputMultiallelics = F);
}
```

R code for the genealogical sorting index (GSI) calculations

```
# Code modified from:  Ravinet, M. et al. The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. PLoS Genetetocs 14, e1007358 (2018).

# ===== Load Dependencies =====
library(ape)
suppressMessages(library(vcfR))
library(geiger)
suppressMessages(library(adegenet))

# ===== Read Input Data =====
args <- commandArgs(trailingOnly = TRUE)
filename <- args[1]

#filename <- "~/Dropbox (OL)/OriginScenarios-results/N-1000.t1-10000.t2-10000.tSec-0.25.mig-1e-06/t2.replicate-1.vcf"
vcf <- read.vcfR(filename, verbose = FALSE)


# ===== Define gsi Function =====
gsi <- function(tr, grp){
  n <- length(grp) - 1
  # only consider internal nodes (tips get index 1:Ntip(tr))
  internal.nodes <- seq(Ntip(tr)+1, Ntip(tr) + Nnode(tr))
  # For each internal node, what are the descendant tips
  node.descendants <- lapply(internal.nodes, function(n) tips(tr, n))

  # ----- Jeff Groh 10 May 2019 -----
  # Previous code contained an error in the following lines:
```

```
   # Which nodes have descendants in the group being considered?
   # required <- sapply(descendants, function(x) any(grp %in% x) )
  # The problem with this is that it selects *all* nodes which contain *any*
  # members of the focal group. However, in the denominator for the gsi calculation,
  # we are only interested in summing the degrees of nodes which belong to the minimum
  # subtree that contains all members of the focal group. The code below fixes this by
  # selecting the correct set of nodes.

  # find root of mimimum subtree containing all members of focal group
  # how many tips of the focal group are descended from each node
  n.focal.members <- sapply(node.descendants, function(x){ length(which(grp %in% x)) })
  # how many total tips are descended from each node
  n.total.members <- sapply(node.descendants, function(x){ length(x) })
  # to be a root of the minimum subtree, a node must contain at least all members of the focal group
  candidate.subtree.roots <- which(n.focal.members >= length(grp))
  # Out of these, the node with the least number of total descendants will be the subtree root
  candidates.total.members <- n.total.members[candidate.subtree.roots]
  winner <- candidate.subtree.roots[which(candidates.total.members == min(candidates.total.members))]
  root.node <- internal.nodes[winner]
  # find all tips which descend from the min subtree root node
  subtree.tips <- tips(tr, root.node)
  # find all nodes whose descendents include any of those tips
  nodes.with.focal.descendants <- sapply(node.descendants, function(x)any(subtree.tips %in% x))
  # but with fewer descendants than that of the subtree root node
  node.depths <- node.depth(tr)[internal.nodes]
  root.depth <- node.depths[winner]
  # select required nodes for calculation
  required.nodes <- nodes.with.focal.descendants == TRUE & node.depths <= root.depth
  # ----- End Correction -----

  # How many connections to those nodes have? (tree is not necessarily
  # dichotomous)
  degree <- table(tr$edge)[ internal.nodes[required.nodes] ]
  #Ape takes one connection off the root node, so if d=2 then treat it as d=3
  # (no other nodes can have d=2)
  obs.gs <- n / (sum( degree - 2 ) + sum(degree==2))
  #minGS (basically same procedure but for  whole tree)
  degree.total <- table(tr$edge)[seq(Ntip(tr)+1, Ntip(tr) + Nnode(tr))]
  min.gs <- n / (sum( degree.total - 2 ) + sum(degree.total==2))
  gsi <- (obs.gs - min.gs) / (1 - min.gs)
  return(gsi)
}


# ===== Calculate GSI From Phylogenetic Tree =====
# Calculate gsi with respect to environment, that is, high gsi should reflect
# apparent monophyly of groups from the same location (multiple origins)
# rather than monophyly of true clades (single origin).
# Also calculate gsi for true clades so these can be compared.
# In vcf output from slim, individuals are organized sequentially as such:
# p0 (1 individual), p10, p11, p100, p101 (30 individuals each)
# where there is gene flow between p10 & p100 and also p11 & p101 (parapatric pairs).
# Create vectors of names of individuals that belong to these groups.
# This will be used as input for the gsi calculation.

all.inds <- colnames(vcf@gt)[-c(1:2)] # this vector starts with i1 (excluding outgroup)
loc1 <- all.inds[c(1:30,61:90)]
loc2 <- all.inds[c(31:60,91:120)]
clade1 <- all.inds[c(1:60)]
clade2 <- all.inds[c(61:120)]

# Calculate gsi for entire chromosome (1Mb)
gen <- as.matrix(vcfR2genlight(vcf))
tr <- root(nj(dist(gen)), outgroup = "i0", resolve.root = TRUE)

gsi.clade1 <- gsi(tr, clade1)
gsi.clade2 <- gsi(tr, clade2)
gsi.loc1 <- gsi(tr, loc1)
gsi.loc2 <- gsi(tr, loc2)


# ===== Output GSI Values =====
cat(paste(c(gsi.clade1, gsi.clade2, gsi.loc1, gsi.loc2), sep="\t"))
cat("\n")
```