bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Can secondary contact following range expansion be distinguished from barriers to gene flow?

Johanna Bertl[1,*] Harald Ringbauer[2,3]
Michael G. B. Blum[4]

June 21, 2018

[1] Department of Molecular Medicine (MOMA), Aarhus University, Aarhus, Denmark
[2] Institute of Science and Technology Austria, Klosterneuburg, Austria
[3] Department of Human Genetics, University of Chicago, Chicago, United States
[4] Univ. Grenoble Alpes, CNRS, Laboratoire TIMC-IMAG, Université Grenoble Alpes, CNRS, Grenoble, France
[*] *corresponding author*

Key words Isolation-by-distance, secondary contact, Quaternary glaciation, hybrid zone, barrier of gene-flow, landscape genetics

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Secondary contact is the reestablishment of gene flow between sister populations that have diverged. For instance, at the end of the Quaternary glaciations in Europe, secondary contact occurred during the northward expansion of the populations which had found refugia in the southern peninsulas. With the advent of multi-locus markers, secondary contact can be investigated using various molecular signatures including gradients of allele frequency, admixture clines, and local increase of genetic differentiation. We use coalescent simulations to investigate if molecular data provide enough information to distinguish between secondary contact following range expansion and an alternative evolutionary scenario consisting of a barrier to gene flow in an isolation-by-distance model. We find that an excess of Linkage Disequilibrium and of genetic diversity at the suture zone is a unique signature of secondary contact. We also find that the directionality index ψ, which was proposed to study range expansion, is informative to distinguish between the two hypotheses. However, although evidence for secondary contact is usually conveyed by statistics related to admixture coefficients, we find that they can be confounded by isolation-by-distance. We recommend to account for the spatial repartition of individuals when investigating secondary contact in order to better reflect the complex spatio-temporal evolution of populations and species.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

Hybrid zones are narrow regions in which genetically distinct populations meet, mate, and produce hybrids (Barton and Hewitt, 1985). Hybrid zones induced by secondary contact have often been observed in connection with the Quaternary glaciations (Hewitt, 2000). For instance, molecular markers suggest that the southern peninsulas of Europe were major ice age refugia of the European biota and that secondary contact occurred during the northward expansion which followed the last glacial maximum (Taberlet et al., 1998; Hewitt, 1999). With the advent of multi-locus molecular markers such as microsatellite or SNP data, hybrid zones can be investigated using various molecular signatures including gradients of allele frequency, admixture clines, and local increase of genetic differentiation (Nielsen et al., 2003; Adams et al., 2006; Strand et al., 2012; Bermond et al., 2012). Molecular or morphological clinal patterns provide evidence for secondary contact in various plant and animal species such as *Arabidopsis thaliana* (Huber et al., 2014), *Silene vulgaris* (Keller and Taylor, 2010), the grasshopper *Oedaleus decorus* (Kindler et al., 2012), the European hare *Lepus europaeus* (Antoniou et al., 2013) or the parrotbill bird *Paradoxornis webbianus* (Qu et al., 2012) to name just a few examples.

However, typical molecular signatures of secondary contact zones can also occur under other evolutionary scenarios. For instance, admixture clines can be observed under pure isolation-by-distance models where nearby populations are connected through gene flow (Engelhardt and Stephens, 2010). Additionally, an increase of genetic differentiation can occur in isolation-by-distance models when there are barriers to dispersal (Barton and Bengtsson, 1986; Riley et al., 2006). With the advent of landscape genetics, the search for barriers to gene flow has attracted considerable attention (Manel et al., 2003; Storfer et al., 2010). Although secondary-contact zones can occur at barriers to gene flow, the two models convey different evolutionary paradigms. Models of barriers to gene flow are usually based on isolation-by-distance settings where neighboring populations are connected through dispersal (Nagylaki, 1988; Safner et al., 2011; Blair et al., 2012). Around the barrier to gene flow, dispersal is lowered because of geographical or anthropogenic obstacles (Riley et al., 2006; Zalewski et al., 2009). By contrast, models of secondary contact include an initial phase of evolutionary divergence between two populations or between two sets of populations. The phase of evolutionary divergence is followed by a phase of gene flow between the two divergent units at the secondary contact zone (Murray and Hare, 2006; Durand et al., 2009). The fact that patterns of genetic differentiation can be attributed to different demographic factors is a recurrent problem when using molecular markers. Genetic structure may represent past or contemporary processes and it is notoriously difficult to disentangle between the two possible explanations (Epps and Keyghobadi, 2015).

Here, we use coalescent simulations to investigate to what extent molecular data provide information to distinguish secondary contact following range expansion from barriers to gene flow. We consider neutral simulations only and the barrier to gene flow is modeled by a reduced migration rate. Models of sec-

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ondary contact of neutral markers are non-equilibrium models that converge to a migration-drift equilibrium (Endler, 1977; Bierne et al., 2013), whereas locally adaptive loci would rather converge to a selection-migration balance (Barton, 1979). For both evolutionary scenarios, we simulate a one-dimensional stepping stone model as shown in figures 1 and 2.

For comparing the molecular signal left by the two distinct scenarios, we consider statistical measures that have been developed to provide evidence for different demographic processes. The first set of summary statistics, which is used to detect hybrid zones, contains measures of individual admixture coefficients between the parental source populations (Nielsen et al., 2003; Durand et al., 2009), and a measure of LD as an increase of LD is expected in admixed populations (McVean, 2002).

The second set of summary statistics contains measures to detect range expansion because secondary contact is frequently induced by geographical expansions of the ancestral populations (Hewitt, 2000). We consider the directionality index ψ as it is sensitive to the occurrence of recent range expansion and it should distinguish between equilibrium and non-equilibrium models (Peter and Slatkin, 2013). The properties of the directionality index have not been studied yet when there are introgressive events. Furthermore, we include genetic diversity which has been shown to decrease along the expansion direction (Austerlitz et al., 1997).

The last set of summary statistics pertains to isolation-by-distance and barriers to gene flow. First, we include the decay of correlation between allele frequencies as a function of distance as it provides evidence of isolation-by-distance (Hardy and Vekemans, 1999). To detect barriers to gene flow, many numerical methods such as 'wombling' identify zones of sharp changes in allele frequencies (Barbujani and Sokal, 1990; Manel et al., 2003). Here, we use local F_{ST} defined as F_{ST} per unit of spatial distance for this purpose. Local F_{ST} can be provided with georeferenced data by the software *LocalDiff* and we expect them to be larger around the barrier to gene flow (Duforet-Frebourg and Blum, 2014). The development of the software *LocalDiff* questioned the possibility of distinguishing between the two evolutionary scenarios under consideration. When studying patterns of differentiation in various alpine plants with *LocalDiff*, we found a region of larger local genetic differentiation in the Italian Aosta valley that was shared across alpine plant species. Both secondary contact zone following postglacial expansions or a barrier to gene flow in a equilibrium stepping-stone process were putative explanations (Duforet-Frebourg and Blum, 2014).

2 Methods

2.1 Simulation Models

We consider secondary contact in a one-dimensional nearest-neighbor stepping-stone model consisting of 100 demes (figure 1). Range expansion is modeled as a series of founder events with moderate bottlenecks. Time is given in coalescent

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

units before present time, i. e. in units of $4N$ generations where N is the diploid population size per deme at present time. Accordingly, all parameters are scaled with $4N$.

Phase 1 (ancestral population) The ancestral population is a random-mating population of size $2N$. At time t_S, it splits in two populations of size N.

Phase 2 (separate refugia) From time t_S to time t_E, the two populations are in separate refugia (demes 1 and 100, respectively), the population sizes are constant, and there is no gene flow.

Phase 3 (expansion) Starting at time t_E, both populations expand towards each other in the stepping-stone geometry. At time points $t_E, t_E - d, t_E - 2d$ etc., 10% of the individuals of the deme at the expansion front colonize a new deme. Instantaneously, the size of both demes increases to N again and migration occurs at rate m between neighboring demes.

Phase 4 (secondary contact) From $t_C = t_E - 48d$ until the present time, a stepping-stone model with 100 demes of size N is maintained with a migration rate m among the neighboring demes and a migration rate $m_B \leq m$ between demes 50 and 51 where secondary contact occurs. If secondary contact occurs at a barrier, $m_B < m$, otherwise $m_B = m$.

As an alternative demographic model we investigate a nearest-neighbor stepping-stone model with a constant range of 100 demes and with reduced gene flow in the center (figure 2). Again, the barrier to gene flow is modeled by a lower migration rate $m_B \leq m$ between demes 50 and 51.

DNA data of 20 haploid (or 10 diploid) individuals per deme is simulated with the coalescent simulator `ms` (Hudson, 2002). For each of them, we simulate 100 unlinked sequences consisting of 100,000 base pairs each. A sequence contains 100 SNPs and the scaled recombination rate within the sequence is 4.

In the secondary contact model we simulate data with parameters $t_S = 19$, $d = 1/8$ and different durations since secondary contact occurred (from $t_C = 0, 1, \ldots, 5$ until present time). In both models the scaled migration rate between neighboring demes is $m = 20$. In the stepping-stone model we consider different barrier permeabilities ($m_B/m = 0.002, 0.01, 0.02, 0.1, 0.2, 1$; a value of 1 denotes no barrier). To provide means and standard errors of the summary statistics, each simulation is repeated 100 times.

The ms command lines and the simulated data are available in the Figshare repository Bertl et al. (2018), `https://doi.org/10.6084/m9.figshare.4986545.v2`.

2.2 Summary statistics

In the following, we describe the summary statistics we calculated to disentangle the two evolutionary scenarios.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Hybrid zone summary statistics

Admixture coefficient Based on the first principal component we compute an admixture coefficient for the pool of the five demes left of the barrier (Paschou et al., 2007). The pools of the 5 leftmost and 5 rightmost demes are used as proxies for the two source populations. The admixture coefficient is defined as the average relative location of individuals in the putative admixed population with respect to the two source populations on the axis of the first principal component (Bryc et al., 2010). Let $\bar{y}_i$ be the score of the first principal component, averaged over the individuals in population $i = l, r, a$ (left, right source population, admixed population). Then, we define the admixture coefficient

$$\frac{|\bar{y}_a - \bar{y}_l|}{|\bar{y}_r - \bar{y}_l|}.$$

It takes values between 0 and 1 and is proportional to the fraction of genetic material inherited from the right source population through admixture.

The principal component analysis is conducted with the `R` function `prcomp` (R Core Team, 2012).

Linkage disequilibrium (LD) We average the squared correlation coefficient between 1000 randomly drawn pairs of SNPs within the same sequence over all unlinked sequences. We compute LD for each deme.

Range expansion summary statistics

Directionality index ψ The directionality index ψ has been developed to detect range expansion and infer its origin (Peter and Slatkin, 2013). It is based on the fact that populations further away from the origin of an expansion have experienced more genetic drift. The index $\psi_{i,j}$ is a pairwise measure between demes i and j that compares the average allele frequencies in the two demes: stronger drift yields higher differences in allele frequencies.

Denote the allele frequencies in deme i by the vector $f^{(i)} = \left(f_1^{(i)}, \ldots, f_L^{(i)}\right)$, where L is the total number of SNPs. Then, the directionality index for demes i and j, from each of which a sample of size M has been drawn, is defined as

$$\psi_{ij} = \frac{1}{LM} \sum_{l=1}^{L} \left(f_l^{(i)} - f_l^{(j)}\right).$$

Given that a range expansion has occurred, $\psi_{i,j}$ should be negative if deme i is closer to the origin of the expansion than j, and positive otherwise. If $\psi_{i,j} \approx 0$, both demes should be equally close to the origin of the expansion, or no range expansion has occurred. We compute values of $\psi_{26,j}$ with $j = 27, \ldots, 50$.

Genetic variability We measure genetic variability in each deme by averaging the number of pairwise nucleotide differences between all pairs of sequences, denoted by Δ.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Summary statistics for isolation-by-distance and barriers to gene flow

Allele frequency correlogram The Pearson correlation between the allele frequencies of demes i and j is denoted by $r_{i,j}$ and defined as

$$r_{i,j} = \frac{\sum_{l=1}^{L}(f_l^{(i)} - \bar{f}^{(i)})(f_l^{(j)} - \bar{f}^{(j)})}{\sqrt{\sum_{l=1}^{L}(f_l^{(i)} - \bar{f}^{(i)})^2 \sum_{l=1}^{L}(f_l^{(j)} - \bar{f}^{(j)})^2}}.$$

We compute the correlogram $r_{26,j}$ for $j = 27, \ldots, 75$.

Local $\mathbf{F_{ST}}$ Local values of F_{ST} correspond to pairwise F_{ST} between neighboring demes (Duforet-Frebourg and Blum, 2014). Here, we use Weir and Cockerham's estimator for multiple loci and random union of gametes (Weir and Cockerham, 1984, p. 1363).

2.3 Theoretical context

The simulated scenarios are partly amenable to theory. In particular, continuous spatial diffusion models have proven to be powerful approximations to discrete stepping stone models (Wright, 1943; Nagylaki, 1978). These converge quickly to their continuous diffusive counterparts, especially in one spatial dimension (Nagylaki, 1988; Barton, 2008; Forien, 2017). Recently, such continuous diffusion models have been applied to scenarios of barriers to gene flow (Ringbauer et al., 2018), as well as secondary contact (Sedghifar et al., 2015). Here, we summarize relevant findings and provide predictions for our simulations where available.

Diffusion Predictions

In the modern coalescence framework, the spatial distribution of ancestral lineages back in time is often modeled by a diffusion process (Wilkins and Wakeley, 2002). A central parameter is the variance of spatial displacement in one time unit, σ^2. In a stepping stone model this variance is $\sigma^2 = 2m$ (Nagylaki, 1988), thus in our simulations $\sigma^2 = 40$.

Diffusion theory gives rise to several predictions which we can test our simulations against. For free diffusion, the probability that an ancestral lineage is spatially displaced by distance Δx at time t back is given by a Gaussian with mean 0 and variance $\sigma^2 t$. In case of secondary contact, the expected fraction of ancestry that traces back to the left side of the contact zone is the probability of finding the ancestral lineage to the left of the barrier at time of secondary contact t_c. Integrating over all possible positions to the left of a barrier yields a cumulative Gaussian $\Phi(l/(\sigma\sqrt{t_C}))$, where l is the geographic distance from the contact zone (Sedghifar et al., 2015). This model predicts that the extent of significantly admixed ancestry covers approximately $\pm 2\sigma\sqrt{t_C}$ around the point of initial contact, and that on average clines dissipate as a cumulative normal distribution (Fig. S4).

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Another prominent but less straightforward prediction based on a diffusive limit is that pairwise F_{ST} increases approximately linearly with pairwise distance, with slope $1/(4N\sigma^2)$ in a homogenous linear habitat (Rousset, 1997).

Linkage Disequilibrium

A pulse of admixture produces linkage disequilibrium (LD), i.e. non-random associations of markers. This admixture LD decays approximately as $\exp(-\rho t_C)$ in a panmictic population, where ρ is the recombination rate between the two involved loci (Chakraborty and Weiss, 1988). In contrast, in spatially structured populations concordant admixture clines keep producing LD. Barton (1982) derived that in recombination-migration model, a common measure for LD, $D = \mathbb{E}(PQ) - \mathbb{E}(P)\mathbb{E}(Q)$, equilibrates to

$$D = p'q'\frac{\sigma^2}{\rho},$$

where p' and q' are the spatial slopes of alleles P and Q. This formula is valid if the slopes of the clines stay approximately constant over the timescale set by recombination ($\approx 1/\rho$). This signal of admixture LD is typically much stronger than background LD produced by random associations due to drift (Hill, 1981; Neel et al., 2013). Sedghifar et al. (2015) used a diffusion model to predict patterns of admixture LD in a one-dimensional habitat with recent secondary contact, similar to the scenario simulated here. They derived formulas for patterns of covariance of ancestry, and deduced from them that admixture LD in a one-dimensional secondary contact model breaks down slower than $\exp(-\rho t_C)$, with a peak at the center of the contact zone.

Genetic variability

Diversity decreases along an expanding wave of colonization (Austerlitz et al., 1997; Edmonds et al., 2004), at a rate which depends on the speed of advance and the shape of the wave front (Hallatschek and Nelson, 2008). In admixed zones after secondary contact diversity can be largely increased, since coalescence of lineages originating from different sides is pushed back to before the initial split. According to the diffusion model, the spatial extent of this area of significant admixture will be $\approx \pm 2\sigma\sqrt{t_C}$ around the contact zone. The increase of diversity will depend on the time of isolation of the two populations, and the population structure before expansion.

In contrast, a barrier is not expected to influence within-population diversity substantially. The invariance principle (Nagylaki, 1998) states that the (correctly weighted) mean within-deme coalescence time, and thus the mean within-deme diversity, is independent of the migration matrix. While a barrier can markedly influence the recent coalescence time distribution (Barton, 2008), it will not strongly influence mean coalescence times. Therefore, one expects a stable Δ across the range of demes in the barrier model.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Pairwise Differentiation

In case of a secondary contact pairwise differentiation, here measured by F_{ST}, will be initially increased across the barrier. However, F_{ST} for neighboring demes equilibrates quickly, with rate $\approx m + 1/2N$ (Slatkin and Barton, 1989), and the initial effect of secondary contact will not persist long on small geographic scales.

On the other hand, a barrier can distort equilibrium patterns of identity by descent (Nagylaki, 1988), and therefore patterns of pairwise F_{ST}. A barrier will only have a significant effect if the barrier is strong enough to markedly influence the spread of ancestry ($m_B/(m\sigma^2) \ll 1$) (Barton, 2008; Ringbauer et al., 2018).

3 Results

To test our simulations, we first compare our results to theoretical predictions from the diffusion model. Reassuringly, we find a close agreement. Average clines are accurately predicted by neutral diffusion expectations (supplementary figure S4). In the equilibrium case without a barrier, pairwise F_{ST} increases approximately linearly with slope $1/(4N\sigma^2)$ (supplementary figure S5), as predicted by (Rousset, 1997).

With the validity of the simulations confirmed, we compare the secondary contact model with no barrier at the contact zone ($m_B = m$) to the stepping stone model with a barrier. We plot the summary statistics either as a function of time since secondary contact or of the intensity of gene flow across the barrier (figure 3; see supplementary section S5 for the underlying values). For summary statistics computed per deme (genetic diversity Δ, LD) or per pair of neighboring demes (F_{ST}), we consider the pattern along the whole range of demes. Many important features are captured by the ratio between the values at the barrier or the suture zone, respectively, and the values to the left and right of it (see supplementary figures S1–S3 for examples of the pattern along the whole range of demes).

First, we consider the average admixture coefficient for the 5 populations that are located on the left-hand side of the barrier (demes 45–50). For the isolation-by-distance model with a barrier, these 5 populations are found to be admixed to an extent depending on the barrier permeability: when increasing the barrier permeability, admixture coefficients of individuals on the left-hand side of the barrier approach 50%. As expected, the populations are also found to be admixed in the secondary contact model without barrier (between 35% and 50%) except for the scenario where data is collected just before secondary contact occurs ($t_C = 0$).

The ratio between LD at the center (demes 49–52) and on both sides of the range (demes 24–27 and demes 74–77; demes closer to the edge of the range are skipped to avoid the edge effect; see supplementary figures S1–S3) shows that LD is homogeneous along the whole range of demes for different barrier permeabilities in the stepping stone model with a slight increase at the

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

barrier for low values of m_B/m. However, in the secondary contact model, LD is considerably increased in the secondary contact zone, at the scale of mixed ancestry predicted by diffusion theory, $\pm 2\sigma\sqrt{t_C}$ (supplementary figures S1–S2). The excess of LD ranges from a more than 2 fold to an approximately 1.3 fold increase and decreases as time since secondary contact increases.

Apart from random fluctuations, the directionality index ψ is constant for the stepping-stone model with constant migration rate ($m_B/m = 1$) as well as for old secondary contact ($t_C = 5$). More recent secondary contact results in a U-shaped pattern. The pairwise statistics $\psi_{26,i}$, $i = 27, \ldots, 50$ first decreases as expected when moving away from the origin of the expansion, but increases again towards the location of secondary contact. For the barrier model with a moderate or strong barrier ($m_B/m \geq 0.1$), $\psi_{26,i}$ remains constant for most of the range, but decreases slightly close to the barrier.

In the stepping stone simulations the number of pairwise differences (Δ) stays approximately constant over the range of demes (apart from an edge effect) and is hardly affected by the barrier (see also supplementary figure S3), as expected by the invariance principle. Conversely, in the secondary contact model Δ increases in the suture zone, again at the spatial scale predicted by diffusion theory (supplementary figures S1–S2). Only when secondary contact has not occurred yet ($t_C = 0$), the statistic Δ captures the effect of range expansion, and decreases when moving away from the origin.

When considering the decay of allele frequency correlation as a function of distance, we find a sharp decrease around the suture zone or around the barrier, respectively. In the barrier model, the correlation decreases linearly with distance and drops sharply at the barrier. In the secondary contact model we observe a more sigmoid shape. For older secondary contact, the sigmoid decay converges towards the linear decay of the equilibrium stepping-stone model predicted by Rousset (1997).

Pairwise F_{ST} between neighboring demes is increased at the barrier to gene flow; the less permeable the barrier, the larger the ratio of F_{ST} at the barrier compared to the rest of the range. In the secondary contact model, local F_{ST} is increased at the center when measured just before secondary contact ($t_C = 0$), but it remains constant along the range of demes when secondary contact is already established ($t_C \geq 1$).

To assess the robustness of these results, we performed simulations of less extreme scenarios. Also with more moderate founder events, a lower expansion speed and higher migration rate between demes, we find that the pattern of LD, genetic diversity and the directionality index remain distinctive summary statistics (supplementary section S3). However, we also observe that the footprint of secondary contact is more difficult to detect for very slow expansions or high migration rates between neighboring demes (supplementary figures S7 and S8). But even with these parameter settings, the directionality index ψ remains a discriminant statistic.

If secondary contact occurs at a barrier to gene flow, the difficulty of detecting the secondary contact from molecular data increases. We consider additional simulations where secondary contact occurs in a region where gene flow is re-

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

duced by a factor of 10 ($m_B/m = 0.1$). In many respects, we see an intermediate pattern between the two previously considered scenarios, yet, genetic diversity and the directionality index ψ still provide evidence for secondary contact and Linkage Disequilibrium is even more increased at the contact zone (figure 4).

Smaller subsets of the data are used to study the impact of the number of loci on the variance of the summary statistics. Analyses of smaller genomes with 10 and 20 unlinked sequences instead of 100 are presented in supplementary figures S9 and S10, respectively. The confidence intervals of all summary statistics are considerably larger, while the spatial pattern of the directionality index and the correlogram is preserved and visible, even through larger random fluctuations. The summary statistic least affected by the data reduction is local F_{ST}. LD and genetic diversity Δ provide a distinctive pattern for very recent secondary contact ($t_C = 1$) when the data consists of at least 20 loci, but not for 10 loci.

We also consider other sampling schemes where the amount of sampled data is reduced. When sampling 4 genomes per deme instead of 20, our results remain unchanged (supplementary figure S11). When sampling only from every 5th deme instead of every deme, they provide similar results for all but one summary statistic (supplementary figure S12). When reducing spatial sampling, it becomes more difficult to detect the spatial pattern of the directionality index ψ (supplementary figure S12).

4 Discussion

We find that admixture coefficients alone do not provide sufficient evidence for secondary contact in the presence of isolation-by-distance. Some other summary statistics of genetic data such as local values of F_{ST} or decay of correlation with distance were not more informative to identify the occurrence of secondary contact either.

By contrast, both an excess of LD and of genetic diversity at the suture zone are found to be unique signatures of secondary contact and also to be informative about the timing of secondary contact. In our simulations we observe an increase of these statistics in areas with mixed ancestry as predicted by the diffusion model ($\pm 2\sigma\sqrt{t_C}$). Several of these findings are well supported by previous theoretical considerations, as outlined above. In particular, admixture LD after secondary contact has been thoroughly studied before (Sedghifar et al., 2015). We do not detect such a marked increase of LD near the barrier to gene flow in our stepping stone model simulations, even for very strong barriers. It seems that the increase of differentiation across the barrier (which increases admixture LD) is roughly balanced by reduced migration across the barrier (which reduces admixture LD). This effect is not well documented in the literature, and warrants further theoretical investigation.

Although a peak of diversity can occur in glacial refugia (Hewitt, 2000), it has been observed before that the genetically most diverse populations were not located in southern Europe but at intermediate latitudes resulting from

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the admixture of divergent lineages that had expanded from separate refugia (Petit et al., 2003). Many hybrid zones have been formed by such postglacial secondary contact. In some cases, the divergence between hybridizing populations has evolved over very long timescales, and may have arisen already in primary contact. We still expect increased LD and admixture near the suture zones as unique signals of recent secondary contact then, since these signals are generated by allele frequency differences immediately before secondary contact. They do not depend on the evolutionary force that created this divergence.

We expect that excess of LD and of genetic diversity due to admixture would be similarly observed in a two-dimensional model. Qualitative patterns of these signals will be affected by the movement along the second dimension, but the underlying theoretical considerations due to admixture LD and increased diversity due to deeper coalescence time remain the same. The edge effects observed here, i.e. reduced diversity and excess of LD because of shorter coalescence times (Wilkins and Wakeley, 2002), will be most pronounced at the four corners of the species range then (Wilkins, 2004).

We also find that the directionality index ψ conveys a signature of secondary contact following expansion. Under range expansion, the ψ statistic is a monotonous function of the distance from the origin of the expansion (Peter and Slatkin, 2013) (see also figure 3 for $t_C = 0$). When secondary contact follows range expansion, it has a distinctive U-shaped pattern (figure 3). The distinctive U-shaped pattern is found only for recent enough secondary contact ($t_C = 1$) but is robust to a wide range of bottleneck intensity and expansion speed (supplementary figures S6 and S7). This statistic adds to the toolbox of population geneticists and provides a promising attempt to distinguish between equilibrium and non-equilibrium spatial processes.

The simulation setting was designed to mimic the evolutionary history of species that have undergone a population split during the Quaternary glaciations with subsequent expansion and secondary contact. Assuming a generation time of 1 year and 1,000 diploid organisms per deme, it includes the time frame of expansion and secondary contact after the last glacial maximum in Europe. Species that had spent the last glacial period in southern refugia started to expand northwards around 16,000 years ago, and subsequently, many plants established a stable distribution around 6,000 years ago (Hewitt, 1999). We assume the ancestral population split up and started diverging 38,000 years ago ($t_S = 19$) and the onset of the expansion varies from 16,000 to 6,000 years ago and lasted 6,000 years. Finally, secondary contact is established on the range of 10,000 years ago ($t_C = 5$) to present time ($t_C = 0$; in this setting, both populations have expanded, but no gene-flow has occurred yet). Our simulations show that the molecular signal of secondary contact vanishes after approximately 10,000 years. To apply our results to a specific organism, parameters like effective population size, time of divergence and expansion rate need to be calibrated.

In additional simulations, we found that the same summary statistics are distinctive for a wide range of parameter values, and also in datasets with fewer samples. However, reducing the number of loci increases the variance of the

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

summary statistics considerably as it has been observed in coalescent models of single populations or isolation-migration models (Felsenstein, 2006; Wang and Hey, 2010).

Our findings are relevant when investigating modes of speciation using computational approaches (Becquet and Przeworski, 2009). Secondary contact following divergence without gene flow (allopatry) is often compared to models of speciation where species start diverging while exchanging migrants (sympatry or parapatry) (Becquet and Przeworski, 2009; Duvaux et al., 2011; Roux et al., 2013). The different frameworks to study speciation are based on isolation-and-migration models, which do not account for the spatial and potentially continuous repartition of individuals (Pinho and Hey, 2010). As shown in our simulation study, accounting for spatial processes provides additional information that can partly be caught with the ψ directionality index, which has power to reveal evolutionary events such as secondary contact and range expansions.

The fact that isolation-by-distance affects the ascertainment of population structure has already been documented (Novembre and Stephens, 2008; Frantz et al., 2009). Accounting for space is a general recommendation that also stands when studying admixture between divergent populations of the same species (Patterson et al., 2012). Although isolation-by-distance is usually perceived as a confounding factor (Meirmans, 2012), the spatial sampling of individuals is in fact a chance to develop more powerful statistical approaches in evolutionary biology. Accounting for continuous populations should also be possible when performing simulations to choose the most probable scenario of speciation (Duvaux et al., 2011). Numerical simulators of genetic variation that account for the spatial repartitions of individuals are now available (Ray et al., 2010; Kelleher et al., 2013). We hope that these developements will encourage researchers to study speciation models that reflect the complex spatio-temporal dynamics of realistic species' evolutionary histories (Alvarado-Serrano and Hickerson, 2015).

Acknowledgments

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC) and the GenomeDK HPC cluster at Aarhus University.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

Adams, S. M., J. B. Lindmeier, and D. D. Duvernell, 2006. Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. Mol. Ecol. 15:1109–1123.

Alvarado-Serrano, D. F. and M. J. Hickerson, 2015. Spatially explicit summary statistics for historical population genetic inference. Methods Ecol. Evol. DOI 10.1111/2041-210X.12489.

Antoniou, A., A. Magoulas, P. Platis, and G. Kotoulas, 2013. Assessing the genetic landscape of a contact zone: the case of the European hare in north-eastern Greece. Genetica 141:23–40.

Austerlitz, F., B. Jung-Muller, B. Godelle, and P.-H. Gouyon, 1997. Evolution of coalescence times, genetic diversity and structure during colonization. Theor. Popul. Biol. 51:148–164.

Barbujani, G. and R. R. Sokal, 1990. Zones of sharp genetic change in europe are also linguistic boundaries. Proceedings of the National Academy of Sciences 87:1816–1819.

Barton, N., 1979. Gene flow past a cline. Heredity 43:333–339.

———, 1982. The structure of the hybrid zone in *Uroderma bilobatum* (chiroptera: phyllostomatidae). Evolution 36:863–866.

———, 2008. The effect of a barrier to gene flow on patterns of geographic variation. Genetics Research 90:139–149.

Barton, N. and B. O. Bengtsson, 1986. The barrier to genetic exchange between hybridising populations. Heredity 57:357.

Barton, N. H. and G. M. Hewitt, 1985. Analysis of hybrid zones. Annu. Rev. Ecol. Syst. 16:113–148.

Becquet, C. and M. Przeworski, 2009. Learning about modes of speciation by computational approaches. Evolution 63:2547–2562.

Bermond, G., M. Ciosi, E. Lombaert, A. Blin, M. Boriani, L. Furian, S. Toepfer, and T. Guillemaud, 2012. Secondary contact and admixture between independently invading populations of the western corn rootworm, *Diabrotica virgifera virgifera* in Europe. PLoS One 7:e50129.

Bertl, J., M. Blum, and H. Ringbauer, 2018. Simulations of secondary contact after range expansion and a barrier to gene flow URL `https://figshare.com/articles/Simulations_of_secondary_contact_after_range_expansion_and_a_barrier_to_gene_flow/4986545`.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

Blair, C., D. E. Weigel, M. Balazik, A. T. Keeley, F. M. Walker, E. Landguth, S. Cushman, M. Murphy, L. Waits, and N. Balkenhol, 2012. A simulation-based evaluation of methods for inferring linear barriers to gene flow. Mol. Ecol. Resour. 12:822–833.

Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser, S. Williams, A. Froment, J.-M. Bodo, C. Wambebe, S. A. Tishkoff, and C. D. Bustamante, 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. U.S.A. 107:786–791.

Chakraborty, R. and K. M. Weiss, 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. U.S.A. 85:9119–9123.

Duforet-Frebourg, N. and M. G. B. Blum, 2014. Non-stationary patterns of isolation-by-distance: inferring measures of local genetic differentiation with Bayesian kriging. Evolution 68:1110–1123.

Durand, E., F. Jay, O. E. Gaggiotti, and O. François, 2009. Spatial inference of admixture proportions and secondary contact zones. Mol. Biol. Evol. 29:1963–1973.

Duvaux, L., K. Belkhir, M. Boulesteix, and P. Boursot, 2011. Isolation and gene flow: inferring the speciation history of European house mice. Mol. Ecol. 20:5248–5264.

Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza, 2004. Mutations arising in the wave front of an expanding population. Proceedings of the National Academy of Sciences of the United States of America 101:975–979.

Endler, J. A., 1977. Geographic Variation, Speciation, and Clines. Princeton University Press.

Engelhardt, B. E. and M. Stephens, 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 6:e1001117.

Epps, C. W. and N. Keyghobadi, 2015. Landscape genetics in a changing world: disentangling historical and contemporary influences and inferring change. Molecular ecology 24:6021–6040.

Felsenstein, J., 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences or more loci? Molecular Biology and Evolution 23:691–700.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Forien, R., 2017. Gene flow accross geographical barriers-scaling limits of random walks with obstacles. arXiv preprint arXiv:1711.02589 .

Frantz, A., S. Cellina, A. Krier, L. Schley, and T. Burke, 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? J. Appl. Ecol. 46:493–505.

Hallatschek, O. and D. R. Nelson, 2008. Gene surfing in expanding populations. Theoretical Population Biology 73:158–170.

Hardy, O. J. and X. Vekemans, 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. Heredity 86:145–154.

Hewitt, G., 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–913.

Hewitt, G. M., 1999. Post-glacial re-colonization of European biota. Biol. J. Linn. Soc. 68:87–112.

Hill, W. G., 1981. Estimation of effective population size from data on linkage disequilibrium. Genetics Research 38:209–216.

Huber, C. D., M. Nordborg, J. Hermisson, and I. Hellmann, 2014. Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. Mol. Biol. Evol. 31:3026–3039.

Hudson, R. R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Kelleher, J., N. H. Barton, and A. M. Etheridge, 2013. Coalescent simulation in continuous space. Bioinformatics 29:955–956.

Keller, S. and D. Taylor, 2010. Genomic admixture increases fitness during a biological invasion. J. Evolution. Biol. 23:1720–1731.

Kindler, E., R. Arlettaz, and G. Heckel, 2012. Deep phylogeographic divergence and cytonuclear discordance in the grasshopper *Oedaleus decorus*. Mol. Phylogenet. Evol. 65:695–704.

Manel, S., M. K. Schwartz, G. Luikart, and P. Taberlet, 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18:189–197.

McVean, G. A. T., 2002. A genealogical interpretation of linkage disequilibrium. Genetics 162:987–991.

Meirmans, P. G., 2012. The trouble with isolation by distance. Mol. Ecol. 21:2839–2846.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Murray, M. C. and M. P. Hare, 2006. A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. Mol. Ecol. 15:4229–4242.

Nagylaki, T., 1978. A diffusion model for geographically structured populations. Journal of Mathematical Biology 6:375–382.

———, 1988. The influence of spatial inhomogeneities on neutral models of geographical variation: I. Formulation. Theoretical Population Biology 33:291–310.

———, 1998. The expected number of heterozygous sites in a subdivided population. Genetics 149:1599–1604.

Neel, M. C., K. McKelvey, N. Ryman, M. W. Lloyd, R. S. Bull, F. Allendorf, M. Schwartz, and R. Waples, 2013. Estimation of effective population size in continuously distributed populations: there goes the neighborhood. Heredity 111:189–199.

Nielsen, E. E., M. M. Hansen, D. E. Ruzzante, D. Meldrup, and P. Grønkjær, 2003. Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. Mol. Ecol. 12:1497–1508.

Novembre, J. and M. Stephens, 2008. Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 10:646–649.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas, 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 3:e160.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Peter, B. M. and M. Slatkin, 2013. Detecting range expansions from genetic data. Evolution 67:3274–3289.

Petit, R. J., I. Aguinagalde, J.-L. de Beaulieu, C. Bittkau, S. Brewer, R. Cheddadi, R. Ennos, S. Fineschi, D. Grivet, M. Lascoux, A. Mohanty, G. Müller-Starck, B. Demesure-Musch, A. Palmé, J. P. Martin, S. Rendell, and G. G. Vendramin, 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. Science 300:1563–1565.

Pinho, C. and J. Hey, 2010. Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. S. 41:215–230.

Qu, Y., R. Zhang, Q. Quan, G. Song, S. H. Li, and F. Lei, 2012. Incomplete lineage sorting or secondary admixture: disentangling historical divergence from recent gene flow in the vinous-throated parrotbill (*Paradoxornis webbianus*). Mol. Ecol. 21:6117–6133.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Ray, N., M. Currat, and L. Excoffier, 2010. SPLATCHE2: a spatially-explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26:2993–2994. URL `http://www.splatche.com/`.

Riley, S. P., J. P. Pollinger, R. M. Sauvajot, E. C. York, C. Bromley, T. K. Fuller, and R. K. Wayne, 2006. FAST-TRACK: A southern California freeway is a physical and social barrier to gene flow in carnivores. Mol. Ecol. 15:1733–1741.

Ringbauer, H., A. Kolesnikov, D. L. Field, and N. H. Barton, 2018. Estimating barriers to gene flow from distorted isolation by distance patterns. Genetics Pp. genetics–300638.

Rousset, F., 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228.

Roux, C., G. Tsagkogeorga, N. Bierne, and N. Galtier, 2013. Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. Mol. Biol. Evol. 30:1574–1587.

Safner, T., M. P. Miller, B. H. McRae, M.-J. Fortin, and S. Manel, 2011. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. Int. J. Mol. Sci. 12:865–889.

Sedghifar, A., Y. Brandvain, P. Ralph, and G. Coop, 2015. The spatial mixing of genomes in secondary contact zones. Genetics 201:243–261.

Slatkin, M. and N. H. Barton, 1989. A comparison of three indirect methods for estimating average levels of gene flow. Evolution Pp. 1349–1368.

Storfer, A., M. A. Murphy, S. F. Spear, R. Holderegger, and L. P. Waits, 2010. Landscape genetics: where are we now? Mol. Ecol. 19:3496–3514.

Strand, A. E., L. M. Williams, M. F. Oleksiak, and E. E. Sotka, 2012. Can diversifying selection be distinguished from history in geographic clines? A population genomic study of killifish (*Fundulus heteroclitus*). PLoS One 7:e45138.

Taberlet, P., L. Fumagalli, A.-G. Wust-Saucy, and J.-F. Cosson, 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol. Ecol. 7:453–464.

Wang, Y. and J. Hey, 2010. Estimating divergence parameters with small samples from a large number of loci. Genetics 184:363–379.

Weir, B. S. and C. Cockerham, 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Wilkins, J. F., 2004. A separation-of-timescales approach to the coalescent in a continuous population. Genetics 168:2227–2244.

Wilkins, J. F. and J. Wakeley, 2002. The coalescent in a continuous, finite, linear population. Genetics 161:873–888.

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

Zalewski, A., S. B. Piertney, H. Zalewska, and X. Lambin, 2009. Landscape barriers reduce gene flow in an invasive carnivore: geographical and local genetic structure of American mink in Scotland. Mol. Ecol. 18:1601–1615.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

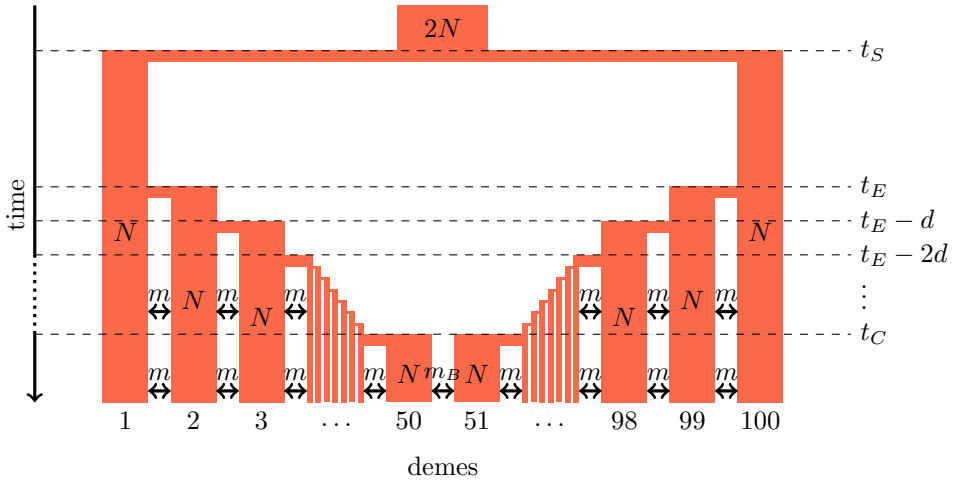

Figure 1: Secondary contact model in a 1-dimensional nearest-neighbor stepping-stone environment with 100 demes. Parameters: N: deme size; m: scaled migration rate; m_B: scaled migration rate at the barrier ($m_B \leq m$); t_S: time of the population split; t_E: time when the expansion starts, d: time between two expansion steps; t_C: time of secondary contact.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

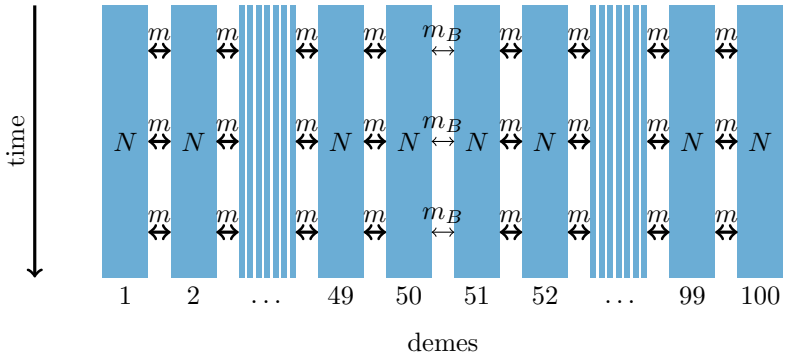

Figure 2: 1-dimensional model of a barrier to gene flow in a nearest-neighbor stepping-stone environment with 100 demes. Parameters: N: deme size; m: scaled migration rate; m_B: scaled migration rate at the barrier ($m_B \leq m$).

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

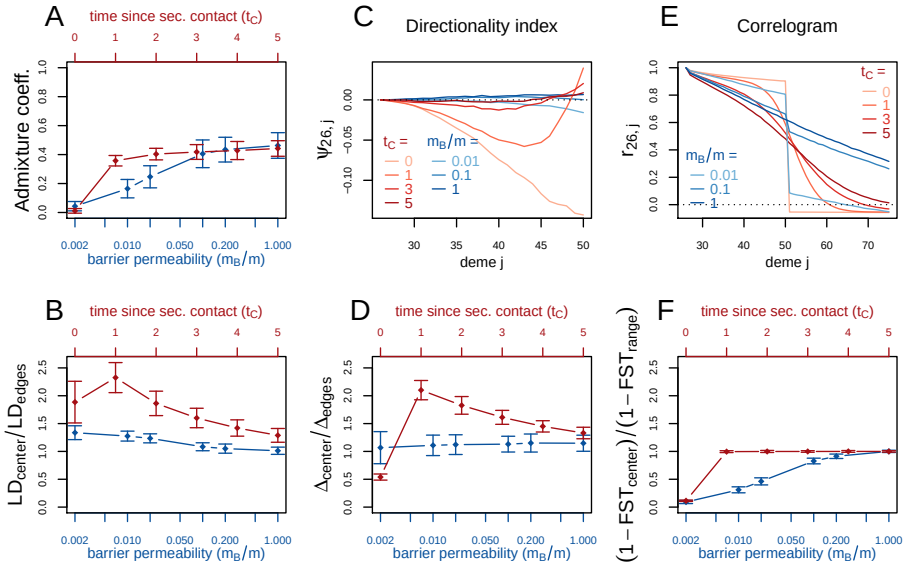

Figure 3: Red (lines and axes): secondary contact model with deme size $N = 1000$, scaled migration rate $m = 20$ (constant migration rate at secondary contact zone, $m_B = m$), time of the population split $t_S = 19$, time since secondary contact $t_C = 0, 1, \ldots, 5$, time between two expansion steps $d = 1/8$. Blue: stepping-stone model with barrier with migration rate $m = 20$ and barrier permeabilities $m_B/m = 0.002, 0.01, 0.02, 0.1, 0.2, 1$; a value of 1 denotes no barrier. The barrier permeability m_B/m is plotted on a logarithmic scale. The dots denote the mean and the error bars ± 2 standard errors, estimated from 100 replicates of the simulations. For Δ and LD, the subscript *center* denotes the mean over demes 49–52 and *edges* over demes 24–27 and 74–77. The admixture coefficient is computed for the 5 demes to the left of the contact zone, demes 46–50. For the F_{ST}, *center* denotes demes 50 and 51 and *range* the mean over the neighboring demes in 26–74 except demes 50 and 51. (For these statistics, the edges of the range are dismissed because of the edge-effect in the stepping-stone model.) For the allele frequency correlogram and the ψ statistic, only the mean is plotted.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

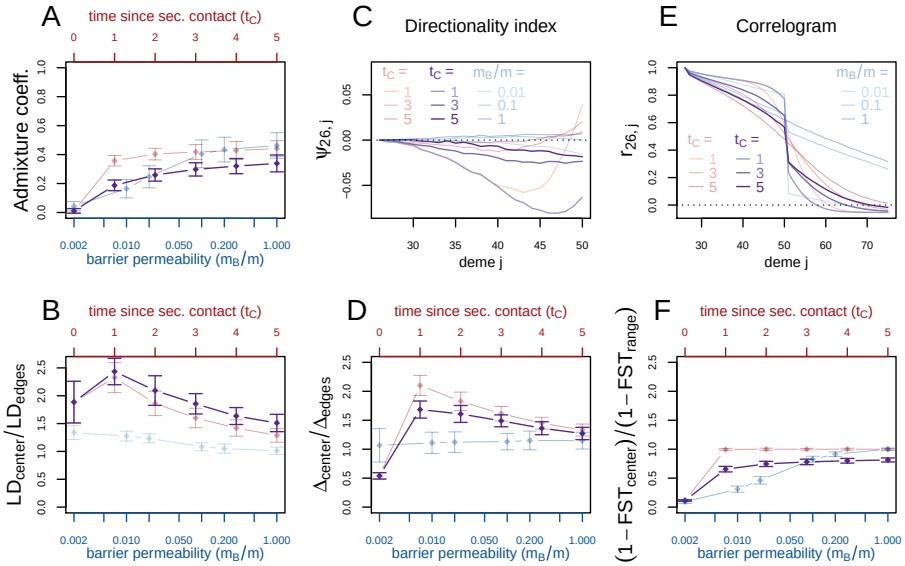

Figure 4: black: secondary contact model with a moderate barrier to gene flow at the secondary contact zone ($m_B = 2$, $m = 20$). The remaining parameters are the same as for the secondary contact model in figure 3. For comparison, the secondary contact model without barrier (faint red) and the stepping stone model (faint blue) are plotted again as in figure 3.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S1 Examples of summary statistics along the range of demes

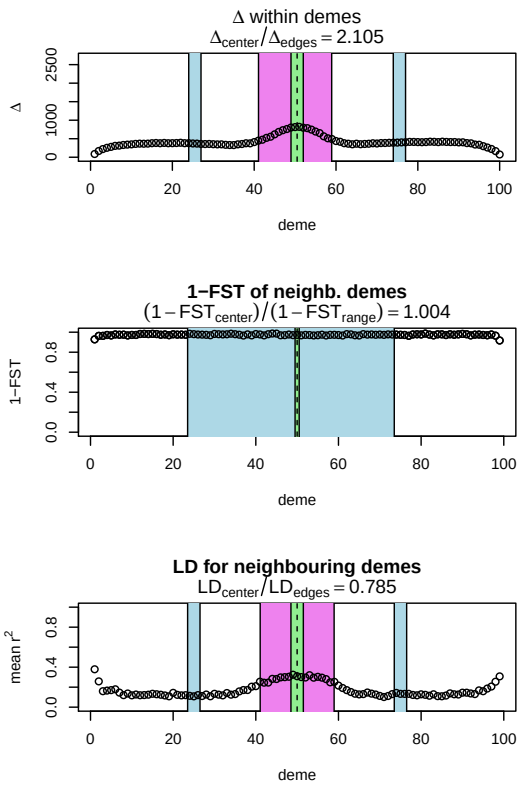

Figure S1: Simulation results for 1 replicate of the secondary contact model with $t_C = 1$. For Δ and LD, the subscript *center* denotes the mean over demes 49–52 (green zone) and *edges* over demes 24–27 and 74–77 (blue zone). For F_{ST}, *center* denotes demes 50 and 51 (green zone) and *range* the mean over the neighboring demes in 26–74 except demes 50 and 51 (blue zone). Pink: admixture zone; width $2\sigma\sqrt{t_C} = 8.94$ in either direction from the center of the zone.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

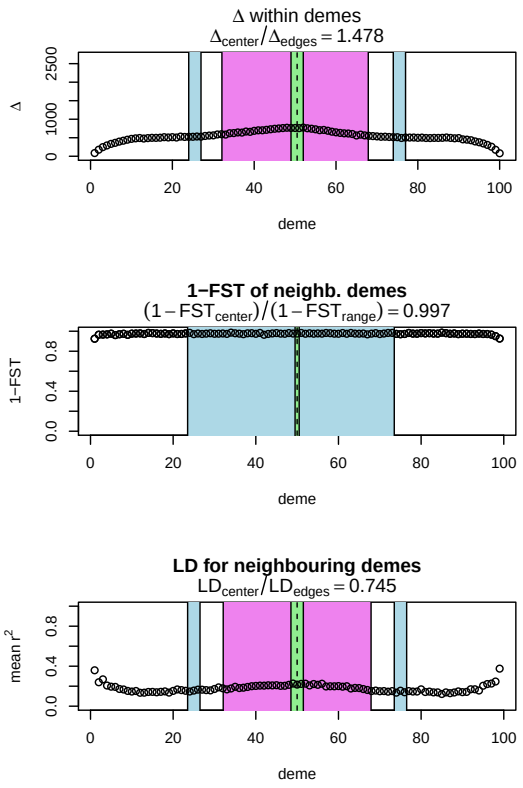

Figure S2: Simulation results for 1 replicate of the secondary contact model with $t_C = 4$. See fig. S1 for a description. Admixture zone width $2\sigma\sqrt{t_C} = 17.9$ in either direction from the center of the zone.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

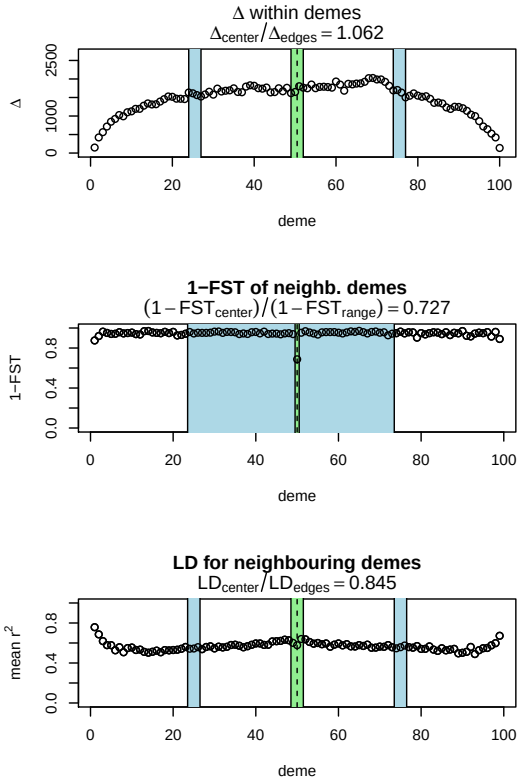

Figure S3: Simulation results for 1 replicate of the stepping stone model with $m_B = 2$. For Δ and LD, the subscript *center* denotes the mean over demes 49–52 (green zone) and *edges* over demes 24–27 and 74–77 (blue zone). For the F_{ST}, *center* denotes demes 50 and 51 (green zone) and *range* the mean over the neighboring demes in 26–74 except demes 50 and 51 (blue zone).

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S2 Comparison with expectations from the diffusion theory

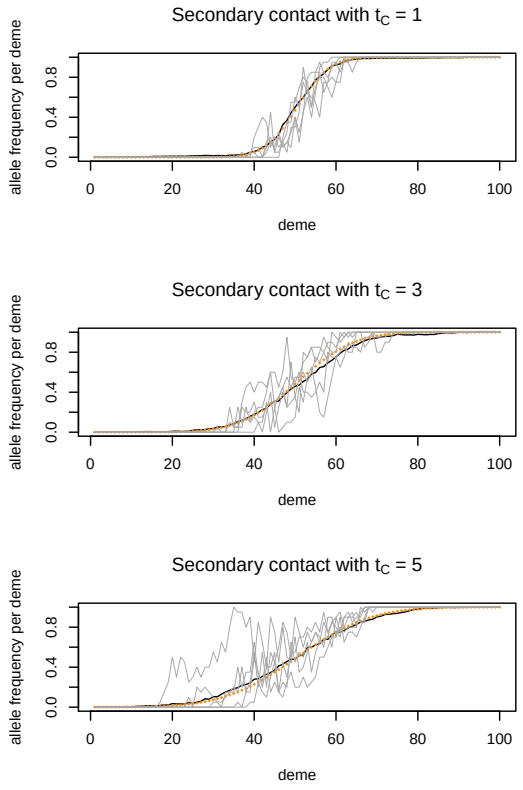

Figure S4: Comparison between observed and predicted allele frequency clines in the secondary contact model. Clinal SNPs with minor allele frequency below 0.05 in demes 1-10 and above 0.95 in demes 91-100 were selected. Gray lines: simulation results for 5 individual unlinked SNPs. Black line: mean over 100 unlinked SNPs. Orange dotted line: Predicted allele frequency cline from Sedghifar et al. (2015); cumulative normal distribution with mean 50.5 and variance $\sigma^2 t_C$.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

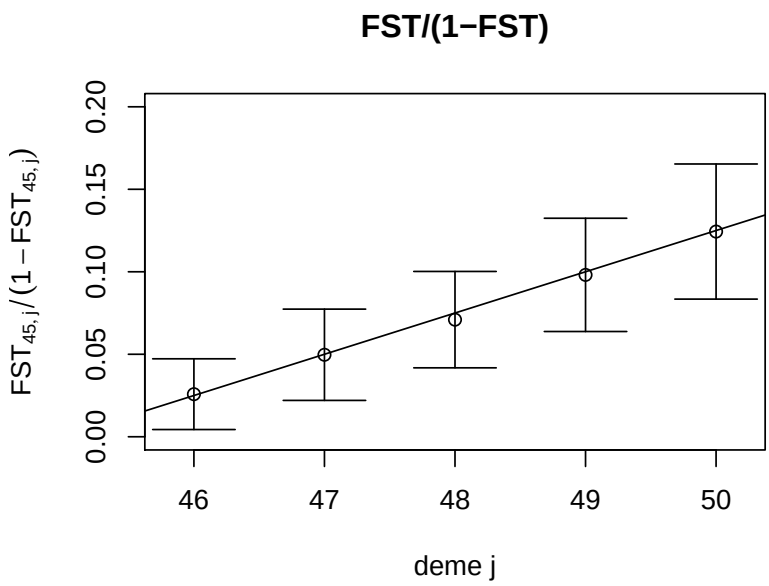

Figure S5: Comparison of the simulation results of the stepping stone model with no barrier ($m_B/m = 1$) with theory predictions. Rousset (1997) showed that for a one dimensional habitat pairwise $F_{ST}/(1 - F_{ST})$ increases linearly with geographic distance r for pairs of subpopulations with slope $1/(4N\sigma^2)$, as long as $\sqrt{2m}r/\sigma \ll 1$. Here, we compare this theoretical prediction to our simulation results. Dots: Mean over 100 replicates. Error bars: ± 2 standard deviations over 100 replicates. Straight lines depict theory predictions, with slope $1/20$ ($\sigma^2 = 2m = 20$, $4N = 1$).

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S3 Simulations at additional parameter combinations

To assess the sensitivity of the results with regard to the other parameters of the models, we resimulated under the secondary contact model with $t_C = 1$ and varied the bottleneck intensity (figure S6) and the expansion speed (d; figure S7). For comparison, the stepping stone model is shown with the same parameters as in the main text.

In addition, we vary the migration rate m simultaneously in both models, in the secondary contact model with $t_C = 1$ and in the stepping stone model with $m_B/m = 0.002$ (figure S8).

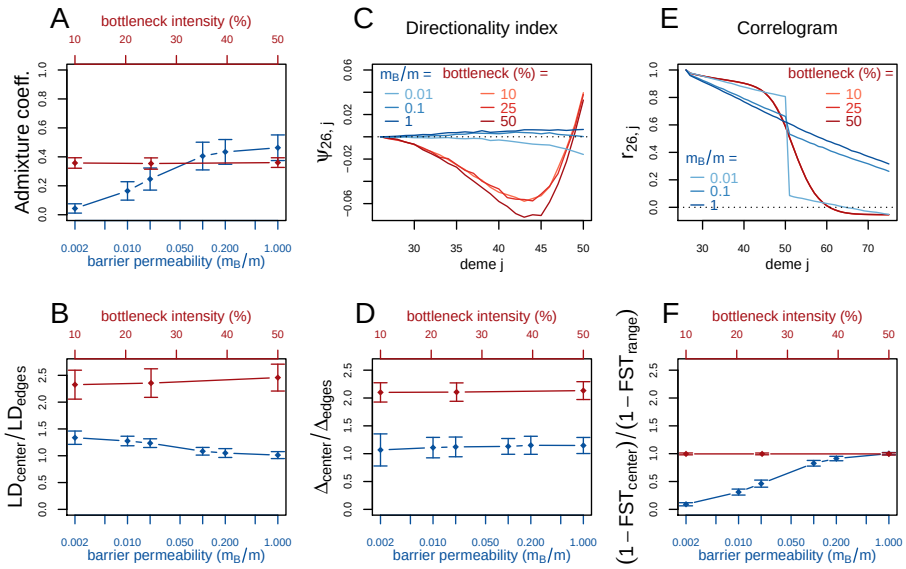

Figure S6: Red (lines and axis): secondary contact model with $t_C = 1$ and varying bottleneck intensity. Blue (lines and axis): stepping stone model with barrier. The dots denote the mean and the error bars ± 2 standard errors, estimated from 100 replicates of the simulations. The admixture coefficient is computed for the 5 demes to the left of the contact zone, demes 46–50. For Δ and LD, the subscript *center* denotes the mean over demes 49–52 and *edges* over demes 24–27 and 74–77. For the F_{ST}, *center* denotes demes 50 and 51 and *range* the mean over the neighboring demes in 26–74 except demes 50 and 51. (For these statistics, the edges of the range are dismissed because of the edge-effect in the stepping-stone model.) For the allele frequency correlogram and the ψ statistic, only the mean is plotted. Note that the correlogram is so similar for each of the three parameters that they are plotted on top of each other.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

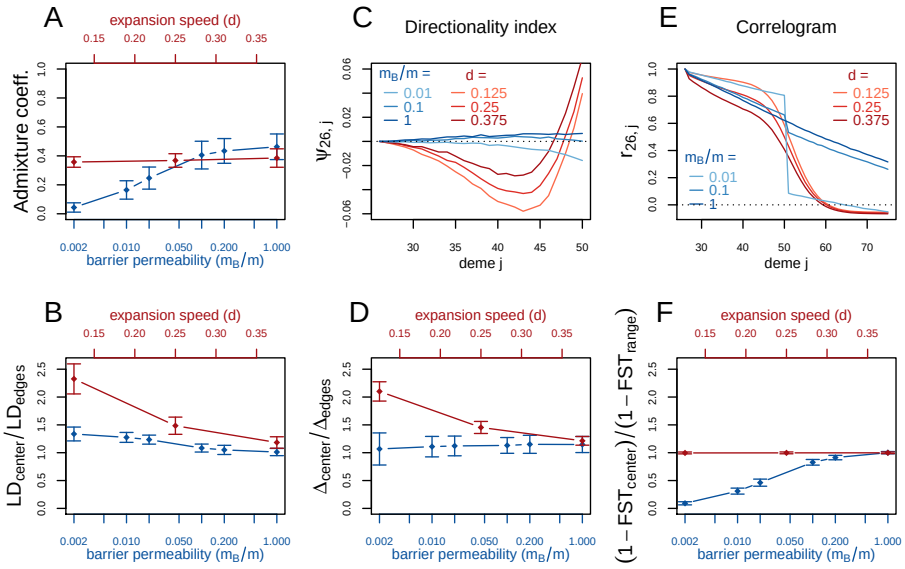

Figure S7: Red (lines and axis): secondary contact model with $t_C = 1$ and varying expansion speed (d). Blue (lines and axis): stepping stone model with barrier. Definitions of the summary statistics as in fig. S6.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

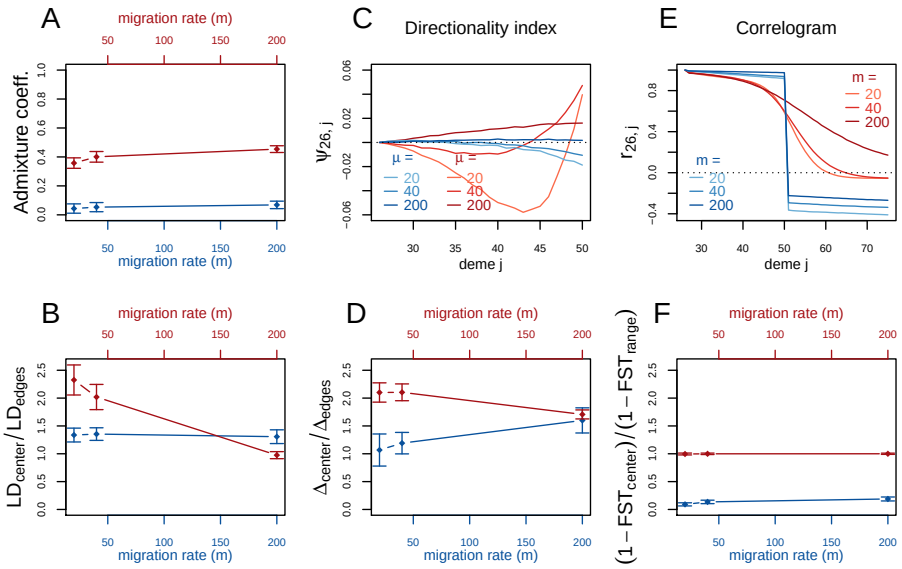

Figure S8: Red (lines and axis): secondary contact model with $t_C = 1$ and varying migration rate (m). Blue (lines and axis): stepping stone model with varying migration rate(m). The barrier permeability $m_B/m = 0.002$ is constant. Definitions of the summary statistics as in fig. S6.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S4 Different sampling schemes

We repeat the analysis with smaller datasets. Instead of 100 independent loci, we subsample the data and compute the summary statistics on 10 or 20 loci, respectively (fig. S9 and S10. We also simulate new data with only 4 genomes per deme (fig. S11 instead of 20). In addition, we compute the summary statistics for a dataset, where 20 genomes have been sampled from every 5th deme only (deme $5, 10, \ldots, 50, 51, 56, \ldots, 96$), so the data consists of DNA sequences from 20 demes (fig. S12).

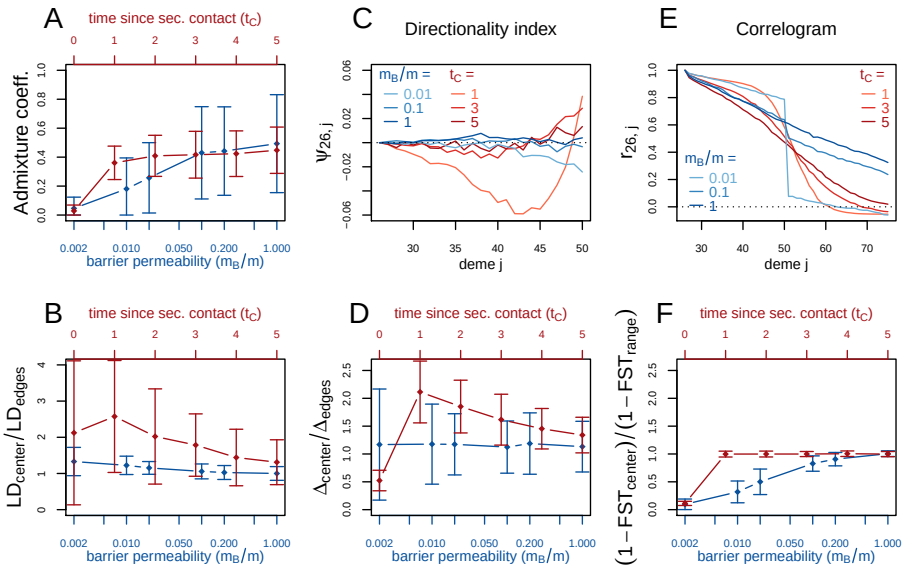

Figure S9: Smaller dataset: 10 independent loci. Red (lines and axes): secondary contact model; blue: stepping-stone model; both models as in fig. 3. Definitions of the summary statistics as in fig. S6. Note the rescaled y-axis in the bottom left panel (LD).

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

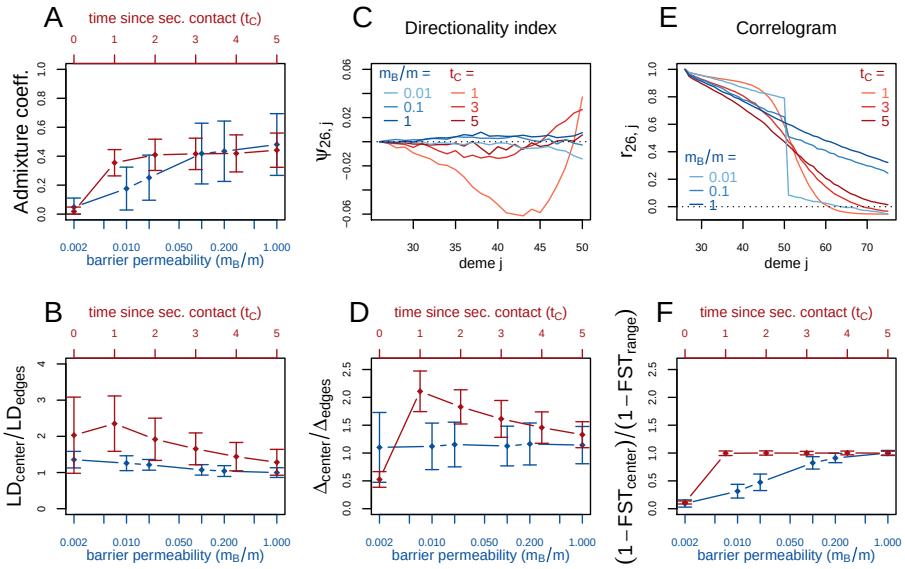

Figure S10: Smaller dataset: 20 independent loci. Red (lines and axes): secondary contact model; blue: stepping-stone model; both models as in fig. 3. Definitions of the summary statistics as in fig. S6. Note the rescaled y-axis in the bottom left panel (LD).

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

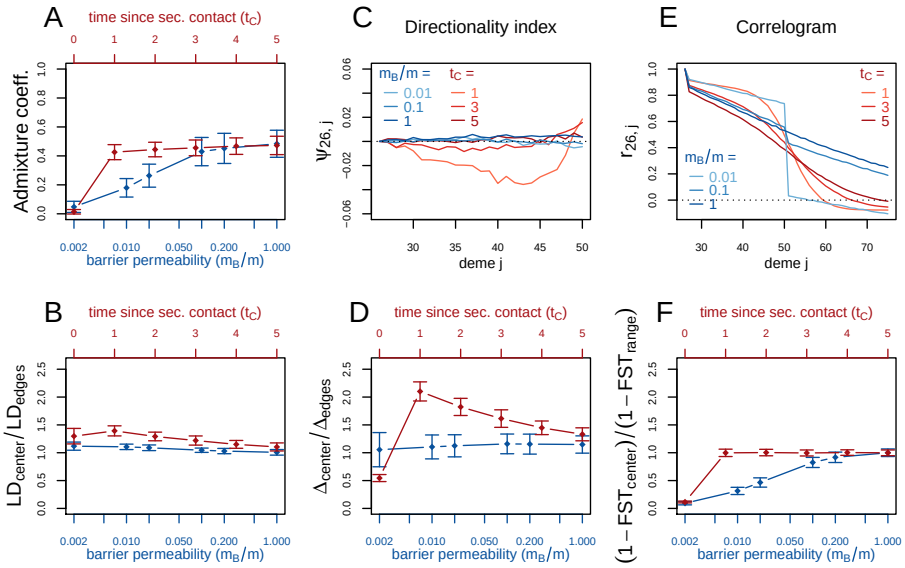

Figure S11: Smaller dataset: 4 genomes per deme. Red (lines and axes): secondary contact model; blue: stepping-stone model; both models as in fig. 3 (main text). Definitions of the summary statistics as in fig. S6.

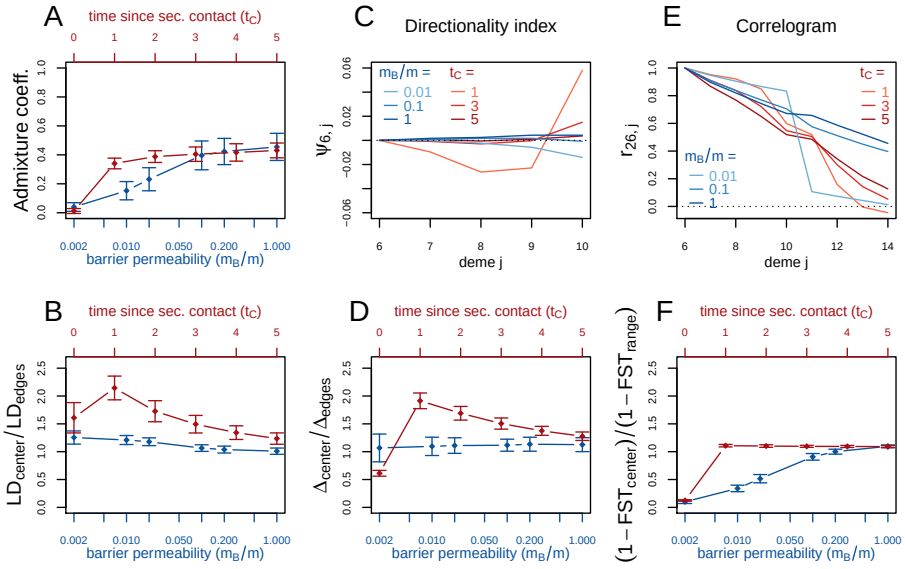

Figure S12: Smaller dataset: 20 demes. Red (lines and axes): secondary contact model; blue: stepping-stone model; both models as in fig. 3. Definitions of the summary statistics as in fig. S6.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S5 Tables of the results

This section contains the underlying values for the admixture coefficient, $LD_{\text{center}}/LD_{edges}$, $\Delta_{\text{center}}/\Delta_{\text{edges}}$ and $(1 - FST_{\text{center}})/(1 - FST_{range})$ of figures 3 and 4.

The tables S1–S4 have the following format: Left: Secondary contact model with different times since secondary contact (t_C); middle: secondary contact model with a moderate barrier to gene flow ($m_B/m = 0.1$) at the secondary contact zone; right: stepping-stone model with barrier to gene flow of different intensities (m_B/m; $m = 20$). The mean and standard error (se) of the admixture coefficient over 100 replicates of the simulations.

Secondary contact			with barrier		Stepping stone		
t_C	mean	se	mean	se	m_B/m	mean	se
0	0.011	0.008	0.011	0.008	0.002	0.044	0.016
1	0.358	0.018	0.187	0.018	0.010	0.165	0.032
2	0.403	0.020	0.259	0.021	0.020	0.247	0.038
3	0.418	0.026	0.298	0.023	0.100	0.405	0.048
4	0.428	0.031	0.320	0.026	0.200	0.434	0.043
5	0.442	0.027	0.339	0.029	1.000	0.463	0.044

Table S1: Admixture coefficient.

Secondary contact			with barrier		Stepping stone		
t_C	mean	se	mean	se	m_B/m	mean	se
0	1.886	0.187	1.886	0.187	0.002	1.337	0.062
1	2.326	0.135	2.435	0.118	0.010	1.275	0.045
2	1.863	0.110	2.094	0.133	0.020	1.235	0.041
3	1.601	0.087	1.856	0.091	0.100	1.084	0.036
4	1.421	0.073	1.636	0.076	0.200	1.050	0.041
5	1.288	0.062	1.510	0.078	1.000	1.012	0.032

Table S2: LD ratio.

bioRxiv preprint doi: https://doi.org/10.1101/043398; this version posted June 21, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Secondary contact			with barrier		Stepping stone		
t_C	mean	se	mean	se	m_B/m	mean	se
0	0.540	0.027	0.540	0.027	0.002	1.067	0.144
1	2.101	0.087	1.685	0.074	0.010	1.109	0.092
2	1.827	0.079	1.609	0.073	0.020	1.122	0.089
3	1.612	0.063	1.490	0.051	0.100	1.131	0.071
4	1.450	0.051	1.362	0.056	0.200	1.151	0.081
5	1.331	0.052	1.271	0.053	1.000	1.147	0.072

Table S3: Δ ratio.

Secondary contact			with barrier		Stepping stone		
t_C	mean	se	mean	se	m_B/m	mean	se
0	0.114	0.006	0.114	0.006	0.002	0.091	0.014
1	0.996	0.008	0.655	0.024	0.010	0.310	0.027
2	0.999	0.009	0.746	0.022	0.020	0.462	0.032
3	0.999	0.007	0.780	0.024	0.100	0.828	0.025
4	1.000	0.008	0.800	0.020	0.200	0.913	0.020
5	1.000	0.006	0.814	0.017	1.000	1.001	0.010

Table S4: FST ratio.