JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.12425

Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a *Mytilus* hybrid zone

C. ROUX*†‡, C. FRAÏSSE*†, V. CASTRIC§, X. VEKEMANS§, G. H. POGSON¶ & N. BIERNE*†

*Université Montpellier 2, Montpellier, France
†CNRS-UMR5554 Institut des Sciences de l'Evolution, Station Méditerranéenne de l'Environnement Littoral, Sète, France
‡Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
§Université Lille Nord de France, USTL, GEPV, CNRS, FRE 3268, Villeneuve d'Ascq, France
¶Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, USA

Keywords:

introgression;
molecular evolution;
natural selection;
population genetics;
simulation;
speciation.

Abstract

The use of molecular data to reconstruct the history of divergence and gene flow between populations of closely related taxa represents a challenging problem. It has been proposed that the long-standing debate about the geography of speciation can be resolved by comparing the likelihoods of a model of isolation with migration and a model of secondary contact. However, data are commonly only fit to a model of isolation with migration and rarely tested against the secondary contact alternative. Furthermore, most demographic inference methods have neglected variation in introgression rates and assume that the gene flow parameter (*Nm*) is similar among loci. Here, we show that neglecting this source of variation can give misleading results. We analysed DNA sequences sampled from populations of the marine mussels, *Mytilus edulis* and *M. galloprovincialis,* across a well-studied mosaic hybrid zone in Europe and evaluated various scenarios of speciation, with or without variation in introgression rates, using an Approximate Bayesian Computation (ABC) approach. Models with heterogeneous gene flow across loci always outperformed models assuming equal migration rates irrespective of the history of gene flow being considered. By incorporating this heterogeneity, the best-supported scenario was a long period of allopatric isolation during the first three-quarters of the time since divergence followed by secondary contact and introgression during the last quarter. By contrast, constraining migration to be homogeneous failed to discriminate among any of the different models of gene flow tested. Our simulations thus provide statistical support for the secondary contact scenario in the European *Mytilus* hybrid zone that the standard coalescent approach failed to confirm. Our results demonstrate that genomic variation in introgression rates can have profound impacts on the biological conclusions drawn from inference methods and needs to be incorporated in future studies.

Introduction

A number of approaches have been recently developed to reconstruct the history of divergence and gene flow between populations of closely related taxa using molecular data (Hey & Nielsen, 2004; Becquet & Przeworski, 2007). One attractive feature of these methods is their potential to distinguish between secondary contact and primary differentiation, a long-standing controversy in speciation research (Endler, 1977; Barton & Hewitt, 1989; Nosil, 2008). However, this task is particularly challenging because the true history of population divergence is often more complex than the

Correspondence: Camille Roux, Department of Ecology and Evolution, University of Lausanne, 1015, Lausanne, Switzerland.
Tel.: +4121 692 4263; fax: +4121 692 4265;
e-mail: camille.roux.1@unil.ch

models fitted to the data. Another problem that has received insufficient attention is that the genetic barriers to gene flow observed between parapatric populations undergoing speciation with gene flow, or diverged populations experiencing secondary contact, are often semipermeable (Harrison, 1993; Wu, 2001). This leads to genomewide heterogeneity (GWH) in the effective levels of gene flow due to the direct effect of selection on isolation genes as well as indirect effects on neutral loci depending on their linkage to selected genes (Barton, 1979; Barton & Bengtsson, 1986; Harrison, 1993; Charlesworth *et al.*, 1997; Nosil & Feder, 2012). The indirect effects of selection produce patterns of gene flow ranging from low rates of introgression in the neighbourhood of barrier loci (the so-called genomic islands of differentiation) to basal introgression rates in regions devoid of selected loci and maximal introgression around loci that have experienced the fixation of an unconditionally favourable allele (i.e. adaptive introgression (Pialek & Barton, 1997; Fraisse *et al.*, 2014)). Methods to infer the history of divergence and gene flow between closely related species from DNA sequence data have flourished during the last decade (Liu *et al.*, 2003; Hey, 2006; Becquet & Przeworski, 2009; Gutenkunst *et al.*, 2009; Pinho & Hey, 2010) with a progressive increase in the complexity of the underlying scenarios. The first and most frequently used method is the isolation with migration (IM) model, which considers the divergence of two populations from T generations in the past that continue to exchange genes at a fixed rate, Nm (Nielsen & Wakeley, 2001; Hey & Nielsen, 2004, 2007). Although these models have proven very useful, they may be sensitive to violations of certain assumptions, most importantly the absence of intragenic recombination and the interruption of gene flow during the divergence process (Strasburg & Rieseberg, 2010). Although more complex scenarios have been proposed (Becquet & Przeworski, 2009; Bazin *et al.*, 2010), the IM model often remains popular and many papers discuss the estimation of divergence times, levels and asymmetry of gene flow, without formally testing alternative models. Combined with a tendency in the speciation literature to incorporate secondary contact in the broader category of models of speciation with gene flow (Butlin, 1987; Smadja & Butlin, 2011), and the emerging popularity of ecological speciation that champions primary differentiation (Feder & Nosil, 2010; Nosil, 2012), it is unclear how the widespread use of the IM model over the past decade may have resulted in a biased perception of the relative importance of primary vs. secondary differentiation during speciation. More studies are needed that evaluate more complex models of incomplete reproductive isolation that incorporate formal statistical tests of alternative scenarios.

It is surprising that the assumption of similar levels of gene flow among loci has rarely been questioned even though interspecific genetic barriers to gene flow have long been known to be semipermeable (Barton & Hewitt, 1989; Harrison, 1993). Two recent studies have proposed inference methods that explicitly account for this heterogeneity of gene flow among loci (Roux *et al.*, 2013b; Sousa *et al.*, 2013). The approach of Sousa *et al.* (2013) extend the IM model by considering two or more groups of loci with different demographic parameters (migration rates and effective population sizes) and assign each locus to a given group using a Bayesian method. The approach of Roux *et al.* (2013b) takes advantage of the flexibility offered by Approximate Bayesian Computation (ABC) to investigate alternative demographic scenarios and considers the migration rate parameter for each locus as a random variable drawn itself from a distribution estimated from the data (according to a hierarchical Bayesian approach with hyperparameters). Both methods enable investigation of GWH, but only the method of Roux *et al.* (2013b) allows for model comparisons and a specific test of the scenario of secondary contact against a model of uninterrupted gene flow during divergence. This method was first applied on highly diverged sea squirt species whose genomes are nearly hermetic to interspecific gene flow (Roux *et al.*, 2013b), and its performance has not yet been evaluated on species exhibiting more permeable barriers to gene flow.

A good system to test scenarios of speciation allowing GWH in introgression rates is the hybrid zone between the marine mussels *Mytilus edulis* and *M. galloprovincialis* where a semipermeable barrier to gene flow has been previously described (Skibinski *et al.*, 1983; Bierne *et al.*, 2003a,b). The geographic structure of the zone is a mosaic of parental and hybrid populations along the Atlantic coasts of France (Bierne *et al.*, 2003a,b; Hilbish *et al.*, 2012) and the British Isles (Skibinski *et al.*, 1983). The interspecific barriers to gene flow are caused by a number of pre- and post-zygotic, intrinsic and extrinsic, isolating mechanisms including spawning asynchrony (Secor *et al.*, 2001), assortative fertilization (Bierne *et al.*, 2002), habitat choice (Bierne *et al.*, 2003a,b), local adaptation (Gardner & Skibinski, 1988; Hilbish *et al.*, 2002) and hybrid fitness depression (Bierne *et al.*, 2006). First studied during the golden age of allozyme markers, this hybrid zone was thought to result from secondary contact because this was the prevailing hypothesis at that time (Barton & Hewitt, 1985). However, the secondary contact hypothesis has not received any statistical support from rigorous analyses of gene genealogies at multiple loci. Furthermore, the *Mytilus* system shares many similarities with some popular systems of the ecological speciation literature such as selection against migrants in the fine-grained high shore–low shore microhabitat heterogeneity (Billard *et al.*, 2010; Johannesson *et al.*, 2010) and post-glacial colonization of new regions released after glacier retreat (Schluter & Rambaut, 1996; Pereyra *et al.*, 2009). In addition, one can question the plausibility of allopatric isolation in species with wide geographic distributions and high dispersal capabilities through a long planktonic larval phase (Palumbi, 1992). A re-evaluation of the secondary contact hypothesis for the European *Mytilus* hybrid zone is thus necessary and overdue.

In this study, we use DNA sequences data from eight nuclear loci to reconstruct the history of divergence and gene flow between *Mytilus edulis* and *M. galloprovincialis*. To allow for heterogeneity in migration rates among loci, we used the hierarchical ABC approach with hyperparameters developed by Roux *et al.* (2013b), which enabled us to fit a secondary contact model with GWH in introgression rates. We compared nested models with homogeneous or heterogeneous migration rates and evaluate alternative scenarios of speciation, including secondary introgression and parapatric primary differentiation. Our results provide strong support for the hypothesis of introgression following secondary contact and document the advantages of incorporating GWH in testing models of speciation.

Materials and methods

DNA polymorphism

We used *Mytilus* spp. samples collected at two localities that previous publications had confirmed are representative of monospecific patches of *Mytilus edulis*, WS (Wadden Sea, Holland; 48 diploid individuals) and *M. galloprovincialis*, FA (Faro, Algarve, Portugal; 48 diploid individuals). The genetic composition of these samples has been analysed previously with DNA fragment length polymorphism and AFLP markers (Bierne *et al.*, 2003a,b; Faure *et al.*, 2008; Boon *et al.*, 2009; Gosset & Bierne, 2013). In addition to the previously published nucleotide sequence data at three loci (Faure *et al.*, 2008; Boon *et al.*, 2009), we obtained data from five new loci with an average fragment length of ~900 bp. PCR primers are described in Table S1. With the exception of locus mc125, which consisted exclusively of coding sequence (Addison *et al.*, 2008), all other loci targeted a fragment of noncoding DNA (intron or intergenic). A standard protocol was used for the PCRs using the PromegaGoTaq® DNA polymerase (Promega, Madison, WI, USA). Sequences were cloned following the mark–recapture (MR)-cloning protocol (Bierne *et al.*, 2007; Faure *et al.*, 2007, 2008; Boon *et al.*, 2009). Individual PCRs were labelled with unique molecular tags using 5′-tailed primers. Tagged PCR products of similar quantities were mixed together and cloned into a pGEM-T vector using Promega pGEM-T cloning kits and sequenced with the universal primers SP6 and T7 flanking the insert at the Genoscope platform (http://www.genoscope.cns.fr/). To avoid sampling bias and to minimize the number of artefactual mutations produced during PCR, cloning and sequencing, we used a single allele per individual, chosen as the most frequently captured variant.

Data analyses

Only silent positions (i.e. synonymous polymorphisms in coding regions and noncoding polymorphisms in introns or intergenic regions) were used to study the demographic history of *Mytilus* populations. For each observed and simulated data set, we computed an array, S_{obs} and S_{sim}, of summary statistics related to polymorphism and divergence widely used in the literature for demographic inferences (Wakeley & Hey, 1997; Becquet & Przeworski, 2007; Ross-Ibarra *et al.*, 2008, 2009; Roux *et al.*, 2011). We computed classical diversity estimators (nucleotide diversity, π and Watterson's θ_W) (Watterson, 1975; Tajima, 1983), between-species differentiation measured by F_{ST} (computed as $1-\pi_s/\pi_T$ where π_s is the average pairwise nucleotide diversity within population and π_T is the total pairwise nucleotide diversity of the pooled sample across populations) and the departure of the site-frequency spectrum from mutation/drift equilibrium by Tajima's *D* (Tajima, 1989) using a routine written in C (MScalc, available from http://www.abcgwh.sitew.ch/; (Roux *et al.*, 2011, 2013a,b)). In addition, we classified the observed polymorphic sites into four distinct categories: (i) polymorphisms exclusive to *M. edulis* noted Sx_{edu}, (i.e. polymorphic sites for which only one allele was found in *M. galloprovincialis*, but two alleles segregate in *M. edulis*); (ii) polymorphisms exclusive to *M. galloprovincialis* noted Sx_{gal}; (iii) fixed differences between species (noted *Sf*); and (iv) shared polymorphic sites (noted *S*s) (i.e. sites for which the same two alleles were segregating in both species). We tested for the presence of intragenic recombination in our alignments using a composite-likelihood approach (McVean *et al.*, 2002) implemented in the PAIRWISE program of the LDhat 2.1 package. We assessed intralocus recombination separately for each species after 1000 permutations of the physical positions of segregating sites.

Inferring ancestral demography

Coalescent simulations

We used an ABC framework (Tavaré *et al.*, 1997; Beaumont *et al.*, 2002) to investigate three scenarios of speciation with gene flow (isolation with migration, IM; Ancient Migration, AM; and Secondary Contact, SC; Fig. 1) and one without gene flow (Strict Isolation, SI). All scenarios assume an instantaneous split of an ancestral population into two daughter populations of constant size. The IM model assumes continuous gene flow between populations. In the AM scenario, the migration events are restricted to the initial phase of speciation, whereas in the SC scenario, the two daughter populations begin to evolve in strict isolation forward in time

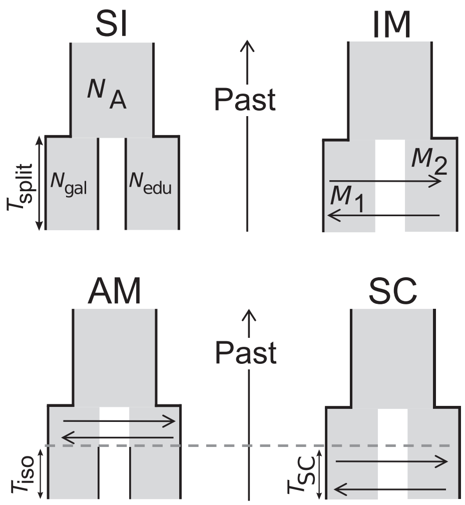

Fig. 1 Alternative scenarios of speciation for *M. edulis* and *M. galloprovincialis.* Four classes of models with different temporal pattern of migration are compared: strict isolation (SI), constant migration (IM), isolation with migration (AM) and secondary contact (SC). Four parameters are shared by all models: T_{split} is the number of generations since the speciation time; N_A, N_{gal} and N_{edu} are the number of effective individuals in the ancestral population, *M. galloprovincialis* and *M. edulis*, respectively. T_{iso} is the number of generations as the two nascent species stopped exchanging migrants in the AM model. T_{SC} is the number of generations as the two daughter species experienced secondary contact after a period of isolation in the SC model. The asymmetrical migration rates M_1 and M_2 are expressed in 4.*N.m* units, where *m* is the proportion of a population made up of migrants from the other population per generation.

and then experience secondary contact. For models with gene flow, we used the scaled migration rates $M = 4.Nm$ (with M_1 the migration rate from *M. galloprovincialis* to *M. edulis* and M_2 the migration rate to *M. galloprovincialis*, time being defined forward), where *m* is the fraction of the population that is composed of migrants from the other population each generation. For the IM, AM and SC scenarios, two alternative models were compared representing the hypotheses of identical vs. variable effective migration rates among loci (‘homogeneous’ vs. ‘heterogeneous’ models, respectively). We thus compared a total of seven models. Following Roux *et al.* (2013b), the ‘heterogeneous’ models consisted of hierarchical Bayesian models with migration rate parameters for each locus drawn from a scaled Beta distribution characterized by three hyperparameters (the alpha- and Beta-shape parameters of the Beta distribution and a scalar ‘c’ to which the Beta distribution is multiplied). This distribution accommodates a large variety of distinct shapes, while avoiding the pitfalls of over-parametrization. We also evaluated an alternative model where two groups of loci have different migration rates, with a proportion *p* of loci having a reduced gene flow when compared to the other group. This alternative model was mainly used to be reassured that the prior distribution used to account for GWH in effective levels of gene flow has little consequence on the comparison between ‘homogeneous’ and ‘heterogeneous’ models. Five million multilocus simulations were performed for each model. We used large uniform prior distributions for all parameters, with identical prior distributions for parameters common to all models. Prior distributions for $\theta_{edu}/\theta_{ref}$,$\theta_{gal}/\theta_{ref}$ and θ_A/θ_{ref} were uniform on the interval 0–20 with $\theta_{ref} = 4.N_{ref}.\mu$. N_{ref} is the effective number of individuals of a reference population used in coalescent simulations, arbitrarily fixed at 100 000, and μ the mutation rate of 2.763×10^{-8} per bp per generation. This rate was estimated from analysis of divergence between *M. californianus* and species from the *Mytilus edulis* complex, assuming a divergence time of 7.6 MY (Ort & Pogson, 2007) and a generation time of 2 years. Although we used LDhat to test for intralocus recombination, we assumed a recombination/mutation ratio equal to one for the following reasons: (i) although the departure from the demographic model used by LDhat is unlikely to bias the detection of recombination events (McVean *et al.*, 2002), it should bias the estimate of population recombination rate; (ii) when using haplotype-free summary statistics, implementing intralocus recombination in coalescent simulations does not require an exact rate as the most important feature is to allow for it (Molly Przeworski, personal communication). θ_{edu},θ_{gal} and θ_A are values for θ of the *M. edulis*, *M. galloprovincialis* and ancestral populations, respectively. We sampled $T_{split}/4.N_{ref}$ from the interval 0–25, 0–10^7 generations in demographic units. The parameters T_{iso} and T_{SC} were drawn from a uniform distribution on the interval 0-T_{split}. In the ‘homogeneous’ models, values of the two migration rate parameters M_1 and M_2 were randomly sampled from the uniform prior interval 0–20 for all loci. For the alternative ‘heterogeneous’ models, a single combination of the shape parameters from the Beta distribution was first randomly and independently sampled for each multilocus simulation from the uniform intervals 0–5 for alpha and 0–200 for Beta. The multiplier ‘c’ was randomly sampled from the uniform prior interval 0–20, once per multilocus simulation. Then, for each locus, the two migration rate parameters M_1 and M_2 were randomly sampled from the scaled Beta distribution. For the alternative two-group parametrization of GWH, we used an uniform prior 0–1 on *P*. Prior distributions were computed using a modified version of the Priorgen software (Ross-Ibarra *et al.*, 2008), and coalescent simulations were run using Msnsam (Ross-Ibarra *et al.*, 2008), a modified version of the ms program (Hudson, 2002) that allows for different sample sizes at each locus under an infinite site mutation model.

Model testing

To statistically evaluate alternative models of speciation, we followed a two-step hierarchical procedure (Fagundes *et al.*, 2007). First, for each scenario allowing migration (IM, AM and SC), we evaluated posterior probabilities for the two alternative models (homogeneous and heterogeneous). Next, we compared the best models from these scenarios in addition to the SI scenario. Posterior probabilities for each candidate model were estimated using a feed-forward neural network implementing a nonlinear multivariate regression by considering the model itself as an additional parameter to be inferred under the ABC framework using the R package 'abc' (Csillery *et al.*, 2012). The 2000 × n replicate simulations nearest to the observed values for the summary statistics were selected (where n is the number of compared models), and these were weighted by an Epanechnikov kernel that reaches a maximum when $S_{\text{obs}} = S_{\text{sim}}$. Computations were performed using 50 trained neural networks and 15 hidden networks in the regression. As model selection can be sensitive to prior choice, we implemented a parametric bootstrap evaluation of the probability of type I error by replicating the ABC inference on thousand pseudo-observed data sets, according to the procedure described in Fagundes *et al.* (2007). We randomly sampled 1000 replicates from the five million simulations performed for each model and used them as pseudo-observed data sets. For each data set, we applied the same model choice procedure to compute the posterior probabilities of each of the compared models. The relative distributions of these probabilities over the 1000 replicates were then used to compute the probability that the best-supported model (i.e. the one with the highest value of the posterior probability obtained from the simulated data set) is indeed the true simulated model (Fagundes *et al.*, 2007; Cornuet *et al.*, 2008; Roux *et al.*, 2011). Hence, one minus this probability gives the probability of type I error (i.e. the probability of rejecting a true hypothesis = P-value).

Parameter estimation

We first estimated parameters shared by all loci and then inferred migration rates for each locus under the heterogeneous models. Parameters were log-tangent-transformed (Hamilton *et al.*, 2005), which allows minimizing heteroscedasticity problems during regressions (the transformed and adjusted parameter values are then back-transformed to obtain the posteriors). Only the 2000 replicate simulations with the smallest associated Euclidean distance $\delta = ||S_{\text{obs}}\text{-}S_{\text{sim}}||$ were considered. The joint-posterior distribution of parameters describing the best model was then obtained by weighted nonlinear multivariate regressions of the parameters on the summary statistics (Blum & François, 2009). For each regression, 50 feed-forward neural networks and 15 hidden networks were trained using the R package 'abc' (Csillery *et al.*, 2012). When the best model involved heterogeneous gene flow, we then estimated the locus-specific migration rate parameters M_1 and M_2. Hence, for each locus, we ran 1.5 × 10^6 random coalescent simulations using parameter values sampled in the joint-posterior distribution for the five parameters common to all loci (N_{A}, N_{B}, N_{anc}, T_{split} and T_{SC}) obtained using the procedure described above. Finally, we applied the described rejection/regression analysis to the simulations performed for each locus to jointly estimate both effective migration rate parameters.

Results

Levels of DNA polymorphism and distribution of variable sites

Twelve to 24 multiply captured individual sequences in *M. edulis* and 14–20 in *M. galloprovincialis* were obtained from the MR-cloning experiment resulting in ~5.3 Kb of alignable silent sites including 737 biallelic positions (Table 1). Both species exhibited similar levels of silent nucleotide diversity when measured with either π ($\pi_{\text{edu}} = 0.0213$ and $\pi_{\text{gal}} = 0.0256$, Wilcoxon signed-rank test V = 17, $P = 0.9453$) or Watterson's θ ($\theta_{\text{edu}} = 0.0256$ and $\theta_{\text{gal}} = 0.0317$, V = 8; $P = 0.3525$). Silent segregating sites were mostly specific to each species (246 and 295 sites were exclusively polymorphic in *M. edulis* and *M. galloprovincialis,* respectively), but a large proportion of the polymorphic positions were shared by the two species (196 sites). Remarkably, the two species exhibited no fixed silent differences, but the distributions of pairwise nucleotide divergence between species were clearly not unimodal (Fig. S1; unimodality was rejected for each locus by the Hartigan's DIP-test; (Hartigan & Hartigan, 1985)). For several loci, the distributions of pairwise nucleotide divergence between species appeared to be bimodal (Fig. S1), reflecting that some coalescence events firstly occurred between alleles from different species rather than within gene pools because of secondary contact.

Variation in migration rates among loci and the timing of gene flow between *M. edulis* and *M. galloprovincialis*

Using an ABC approach with explicit modelling of intralocus recombination (as measured using the LDhat 2.1 package (McVean *et al.*, 2002), Table S2), we first applied a model choice procedure for each demographic scenario implementing migration (Fig. 1) to evaluate alternative scenarios of gene flow and test whether the heterogeneous model explained the data better than the homogeneous model. For all three scenarios incorporating gene flow, we observed unambiguous support in favour of the heterogeneous migration over the commonly used homogeneous alternative (Table 2). The posterior probabilities of the heterogeneous models were always higher than the homogeneous models (Bayes factors, BF, were 221 and 37 for the IM and SC models,

Table 1 Single locus observed statistics.

Loci	n *M. edulis**	n *M. galloprovincialis**	L†	π_{edu}‡	π_{gal}‡	θ^{f}_{edu} §	θ_{gal}§	D_{edu}¶	D_{gal}¶	Sf	Sx_{edu}	Sx_{gal}	Ss	FST	netdivAB
EF1	20	20	639	0.0195	0.0146	0.0278	0.0278	−1.9317	0.0575	0	49	49	14	0.5491	0.0405
EF2	20	20	1133	0.0026	0.0090	0.0070	0.0109	−0.6937	0.0117	0	28	44	0	0.3426	0.0059
Glucanase	18	14	468	0.0211	0.0126	0.0230	0.0188	−1.4043	0.0184	0	20	11	17	0.0712	0.0015
mac1	12	20	825	0.0043	0.0164	0.0040	0.0174	−0.2290	0.0195	0	6	47	4	0.3633	0.0091
Mannanase2	20	18	564	0.0194	0.0223	0.0255	0.0304	−1.1121	0.0222	0	30	38	21	0.0296	0.0014
mc125	19	18	108	0.0260	0.0737	0.0371	0.0705	0.1777	0.0586	0	6	19	8	0.0931	0.0088
mgd2	24	16	1100	0.0387	0.0339	0.0411	0.0433	−0.9419	0.0392	0	78	67	91	0.0495	0.0029
MytilinB	16	16	535	0.0385	0.0223	0.0394	0.0344	−1.4939	0.0308	0	29	20	41	0.0064	0.0004
Average	–	–	–	0.0215	0.0272	0.0256	0.0317	−0.9536	0.0322	–	–	–	–	0.1881	0.0088
Standard deviation	–	–	–	0.0145	0.0220	0.0141	0.0188	0.6917	0.0180	–	–	–	–	0.2018	0.0132
Sum	–	–	5372	–	–	–	–	–	–	0	246	295	196	–	–

*Total number of sequences.

†Silent length excluding all gaps from the total alignment.

‡Average number of pairwise differences.

§Watterson's θ based on the number of segregating positions.

¶Tajima's D.

respectively), and analyses using pseudo-observed data sets obtained by simulations indicated that these differences were highly significant. This result was not altered by the use of the alternative two-group parametrization of GWH (BFs were 72 and 32 for the IM and SC models, respectively). Figure 2 shows that the ABC approach had considerable power to detect a semipermeable barrier to gene flow for each demographic scenario: 100%, 99.9% and 100% of pseudo-observed data sets simulated under the IM, AM and SC scenarios with heterogeneous gene flow were correctly supported by our model choice procedure (i.e. associated with posterior probabilities above 0.5). Indeed, supporting either of the two alternative models did not require a very high posterior probability (Fig. S2). We then studied the temporal pattern of migration by applying the model choice procedure to comparisons between the SI scenario and the IM, AM and SC scenarios with heterogeneous migration (Table 2). Scenarios allowing for ongoing migration (IM and SC) had an elevated cumulative posterior probability (0.92), which strongly rejected the hypothesis that *M. edulis* and *M. galloprovincialis* represent fully isolated species. Among the four scenarios, SC was the best supported by our approach (Table 2), suggesting that migration between the two species is more likely a recent evolutionary event through secondary contact following a period of allopatric isolation rather than a continuous process since the ancestral split. We found a better support for the SC model than for other models whether GWH was modelled with a Beta distribution or with two groups of loci. Because the BF in favour of the SC scenario is small in the two analysis (BF = 1.3 using a Beta distribution, BF = 4.3 using two groups of loci), we tested for the robustness of this result by simulating 1000 pseudo-observed data sets under the different scenarios to empirically measure the rate of false positives, that is, the proportion of IM data sets supported as SC. We then applied to each simulated data set the same ABC model comparison procedure as for the *Mytilus* data set in order to have the distribution of posterior probabilities of SC for each scenario. Using these distributions, we estimated the probability that SC was the correct scenario given the observed posterior probability was 0.957 (P = 0.043 Fig. S3) with the Beta distribution and 0.97 (P = 0.03) with two groups of loci. Thus, even *a priori* low BFs (e.g. 1.3) were sufficient to support the SC scenario. It is worth emphasizing that the statistical distinction between the SC and IM scenarios was only found when migration rates were allowed to vary among loci; support for the SC scenario disappeared when gene flow was assumed to be homogeneous among loci.

Inference of the historical parameters describing the best-supported SC scenario

The length of time spent in allopatry relative to the initial time of divergence is an important factor determining

Table 2 Model classification.

	Within scenarios	Between scenarios
SI	x	0.010
IM		
Hetero (Beta)	0.99	0.40
Hetero (2G)	0.98	0.19
AM		
Hetero (Beta)	0.68	0.07
Hetero (2G)	0.81	0.01
SC		
Hetero (Beta)	0.97	**0.52**
Hetero (2G)	0.97	**0.80**

Hetero (Beta) and hetero (2G) correspond to the two different ways we used to account for GWH in introgression rates, with a scaled Beta-distribution or with two groups of loci differing by their Nm respectively. The first column corresponds to the compared models. Reported values in the second column correspond to the relative posterior probabilities of the hetero (Beta) and hetero (2G) when they are independently compared to their homogeneous alternative within the three scenarios with gene flow. Reported values in the third column correspond to the relative posterior probabilities for each of the four scenarios and for the different Beta and 2G analysis. The best-supported scenario is in bold.

the global strength of barriers to gene flow under the SC scenario of speciation for the two mussel species. To explore the timing of these events, we first inferred the five parameters common to all loci (N_{edu}, N_{gal}, N_{anc}, T_{split} and T_{SC}) and then estimated the parameters of the genomic distribution of migration rates. The joint-posterior distribution from 2000 accepted simulations was strongly differentiated from the prior of each parameter suggesting that our data provide sufficient information and that we explored the correct parameter space (Fig. 3). In Table 3, we report the 95% highest posterior density (95HPD) interval for the five parameters shared by all loci as well as the mode and median of each posterior distribution. Our estimates of the effective population size of *M. edulis* (195 000, 95HPD: 77 000–360 000) is slightly, but not significantly, lower than *M. galloprovincialis* (318 000, 95HPD: 78 000–1 225 000), but our analysis suggests that the ancestral population was substantially larger than the two daughter populations (965 000, 95HPD: 527 000–1 429 000). Interestingly, the less supported 'homogeneous' scenario is less informative about the demographic history with the exception of the two current population sizes which are the only two parameters well differentiated from their prior distributions (Fig. 3, Table 3). Parameter estimates of the best-supported scenario suggest an ancestral subdivision about 2.5 MY ago (95HPD: 1–6.5) followed by secondary contact beginning around 0.7 MY ago (95HPD: 0.4–1.1; Table 3). For both events, the 95HPD were considerably wide and partially overlapped each other. Identifying with accuracy, the timing of migration events is indeed a general caveat in historical inferences (Sousa *et al.*,

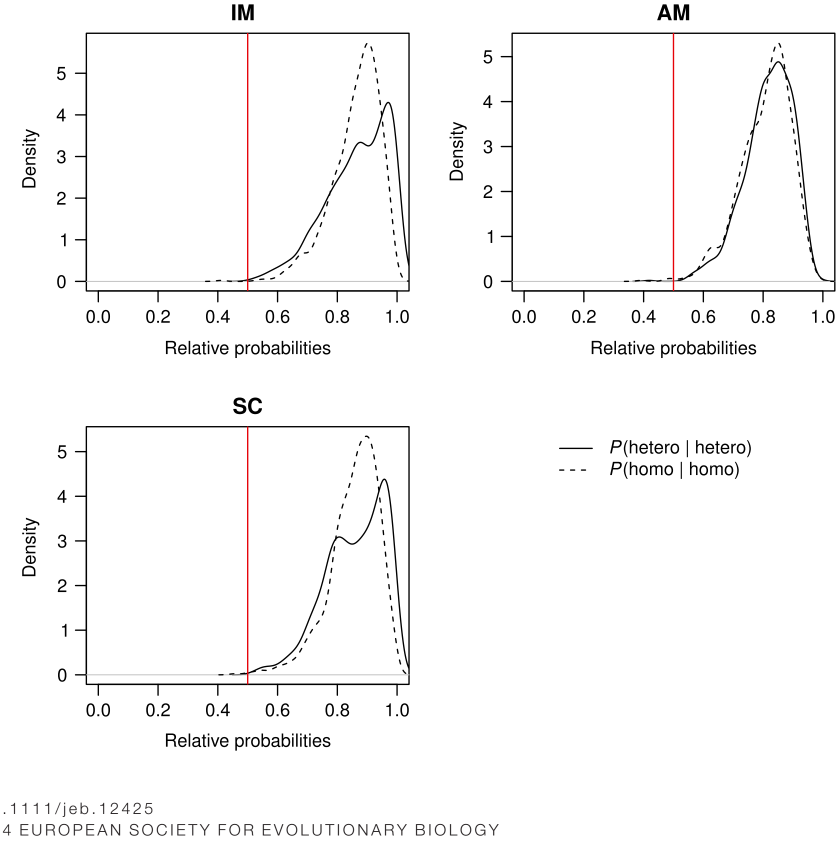

Fig. 2 Empirical distributions of estimated relative posterior probabilities in 'homo' vs. 'hetero' model comparisons. Each distribution was obtained from ABC analysis of 1000 simulated pseudo-observed data sets. The area under each curve above 0.5 represents the fraction of times that the true model is correctly recovered by our estimation procedure.

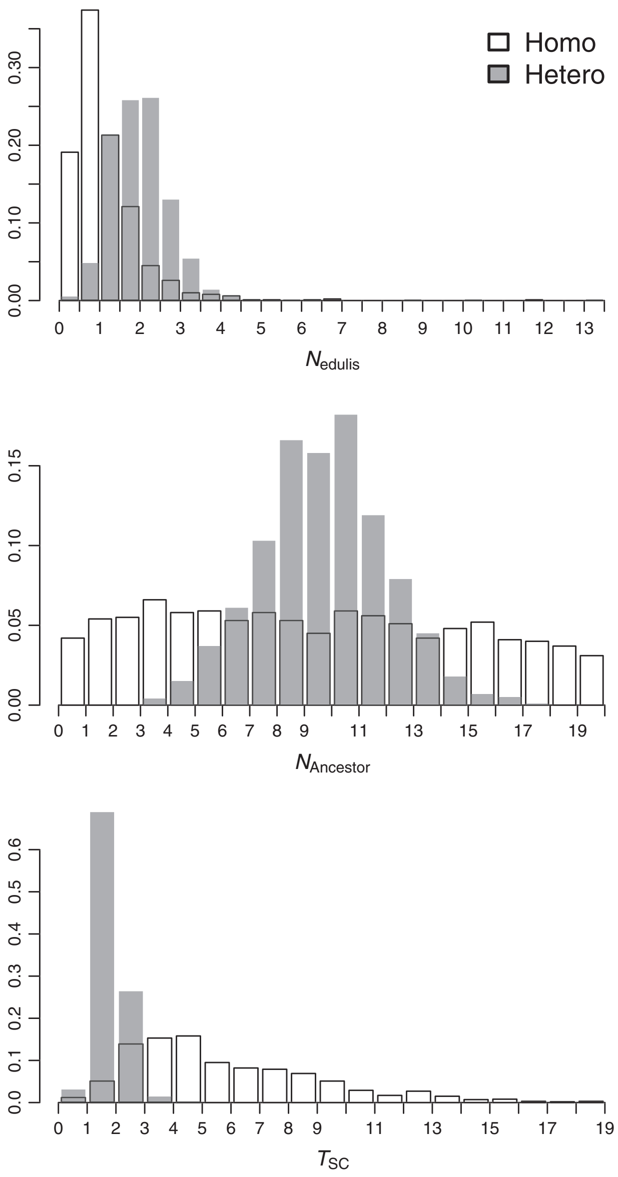

Fig. 3 Posterior distributions of parameters for the two homogeneous and heterogeneous alternative SC models. Homo- and hetero-posterior distributions are represented by open and shaded symbols, respectively.

2011; Strasburg & Rieseberg, 2011), and their interpretation should be taken with more caution than model comparison. Under this scenario, both species would have remained isolated for approximately three-quarters of their history. We then investigated the predicted distributions of genomic introgression rates from *M. galloprovincialis* into *M. edulis* (M_1) and from *M. edulis* into *M. galloprovincialis* (M_2) by 2×10^6 random samples from rescaled Beta distributions using the estimated joint-shape parameters (Fig. 4). According to our estimates, introgression into *M. edulis* occurred at a lower rate (average $M_1 = 0.85$) than introgression into *M. galloprovincialis* (average $M_2 = 1.22$), but the two distributions overlap considerably. We then compared observed and simulated summary statistics under a goodness-of-fit procedure using the joint-posterior distributions and found that the SC scenario with heterogeneous migration fit the data very well except for variation among loci in F_{ST}, which was slightly underestimated by the scenario (Table S3).

Finally, we obtained locus-specific estimates of both migration rates (Table 4). We note that posterior distributions for these two parameters were informative only for a few loci (Fig. 5). Therefore, locus-specific inferences

Table 3 Demographic and historical parameters estimated under the SC scenario.

Parameter	Migration model	Median	Mode	95HPD
Current *M. edulis*	Hetero (Beta)	195 000	200 000	77 000–360 000
population size	Homo	87 600	59 000	26 000–321 000
Current *M. galloprovincialis*	Hetero (Beta)	318 000	222 000	78 000–1 225 000
population size	Homo	208 500	160 000	87 000–547 000
Size of the ancestral	Hetero (Beta)	965 000	1 059 000	527 000–1 429 000
population	Homo	903 000	415 000	71 000–1 912 000
T_{split}	Hetero (Beta)	2 525 000	2 084 000	1 041 000–6 414 000
	Homo	3 470 000	2 594 000	773 000–8 625 000
T_{SC}	Hetero (Beta)	676 000	589 000	390 000–1 153 000
	Homo	1 965 000	1 477 000	573 000–5 539 000

The estimates were calibrated by assuming a generation time of 2 years and a mutation rate of 2.763×10^{-8} per bp per generation.

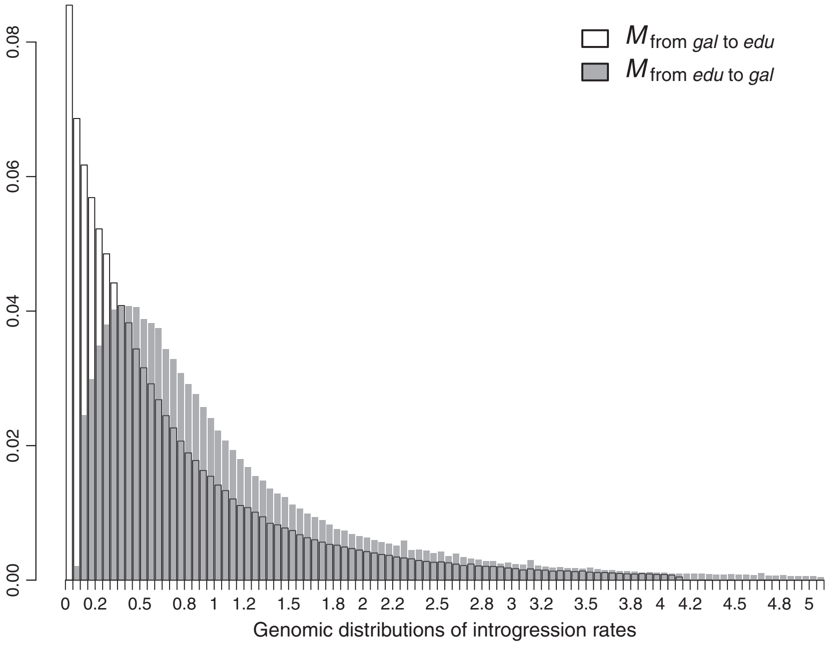

Fig. 4 Estimated genomic distributions of introgression rates into *M. edulis* and *M. galloprovincialis*. Distributions of introgression rates expressed in *Nm* are obtained after randomly sampling 1000 values from each 2000 Beta distributions retained by ABC analysis.

of introgression rates are qualitatively informative but do not allow precise quantification of the mean number of migrants per generation. Nevertheless, heterogeneity in migration rates across loci was clearly apparent, ranging from below one for *EF1α* to values substantially greater than one for *mytilin B*, *mc125* and *glucanase* (Fig. 5; Table 4). Consistent with the multilocus inference, most locus-specific introgression rates tend to be close to the lower bound of the prior distribution, with the introgression spectrum into *M. galloprovincialis* deviating slightly towards higher values than into *M. edulis*.

Discussion

Detecting variation in introgression rates

Models of divergence with gene flow have increased in sophistication since their initial development by Wakeley and Hey (Wakeley & Hey, 1997) and have provided important insights into the process of speciation (Feder & Nosil, 2010; Pinho & Hey, 2010). Here, we document how incorporating heterogeneous gene flow among loci within the ABC framework provided considerable power in detecting a semipermeable barrier to introgression between two mussel species (*M. edulis* and *M. galloprovincialis*) across a well-characterized hybrid zone. Models incorporating variable migration rates among loci strongly outperformed models assuming equal levels of gene flow by comparing posterior probabilities of the alternative models using a model choice procedure. A model-checking procedure from pseudo-observed simulated data sets (Fagundes *et al.*, 2007; Cornuet *et al.*, 2008) provided statistical support for a model of allopatric isolation for the first three-quarters of time since divergence that was followed by secondary contact and introgression. Our simulations comparing homogeneous

Table 4 Locus-specific estimates of the migration rates 4.*Nm*.

	$M_{\text{from gal to edu}}$		$M_{\text{from edu to gal}}$	
	Median	95% HPD	Median	95% HPD
EF1	0.5	0.1–2.1	0.2	0–0.5
EF2	0.3	0.08–1.2	2.3	0.6–5.5
Glucanase	8.8	0.7–19	6	0.3–15.8
mac1	1.3	0.1–11.6	3	0.7–12
Mannanase2	5.4	0.1–19	4.7	0.4–16.2
mc125	8.6	3.5–16.1	5.9	3.6–10.4
mgd2	4.2	0.9–12.7	4.2	1.1–8.5
MytilinB	12	4.2–18.7	9.6	1.9–17.6

and heterogeneous models identified very small rates of false positives and false negatives and showed how the approach can be applied to a flexible number of alternative speciation scenarios. Two recent studies have attempted to test for heterogeneity in migration rates across loci using related approaches. Sousa *et al.* (2013) proposed a modified version of the IM method (Hey & Nielsen, 2004, 2007) that allows the clustering of loci into distinct groups defined by their effective migration rates. This method also allows for groups of loci to experience different levels of genetic drift, which is an important improvement because different genomic regions may not share the same effective population size (Charlesworth, 2009). The method relies on the outlier approach in assuming that the majority of loci behave as neutral markers and only a few loci are affected by selection (hitchhiking within populations, and genetic barriers to gene flow between populations). However, the theory of hybrid zones has established that for a genetic barrier to be truly effective, a large proportion of marker needs to be closely linked to a barrier locus (Barton & Bengtsson, 1986) and data support this prediction (Barton & Hewitt, 1985; Harrison, 1993; Bierne *et al.*, 2013). Sousa *et al.* (2013) illustrated their method by testing GWH in migration rates between two subspecies of the European rabbit (*Oryctolagus cuniculus spp.*) for which a bimodal distribution of F_{ST}-values had been previously described (Geraldes *et al.*, 2008) and a strong association was observed between the levels of differentiation and the assignment to two groups of loci with shared migration parameters. The method of Sousa *et al.* (2013) thus proved effective in detecting GWH in introgression rates but does not allow further evaluation of the best-supported model of speciation (primary vs. secondary differentiation). The mode of speciation between the two *Oryctolagus* subspecies was not explicitly tested, and it remains unclear whether any gene flow occurred after initial divergence between lineages or whether sec-

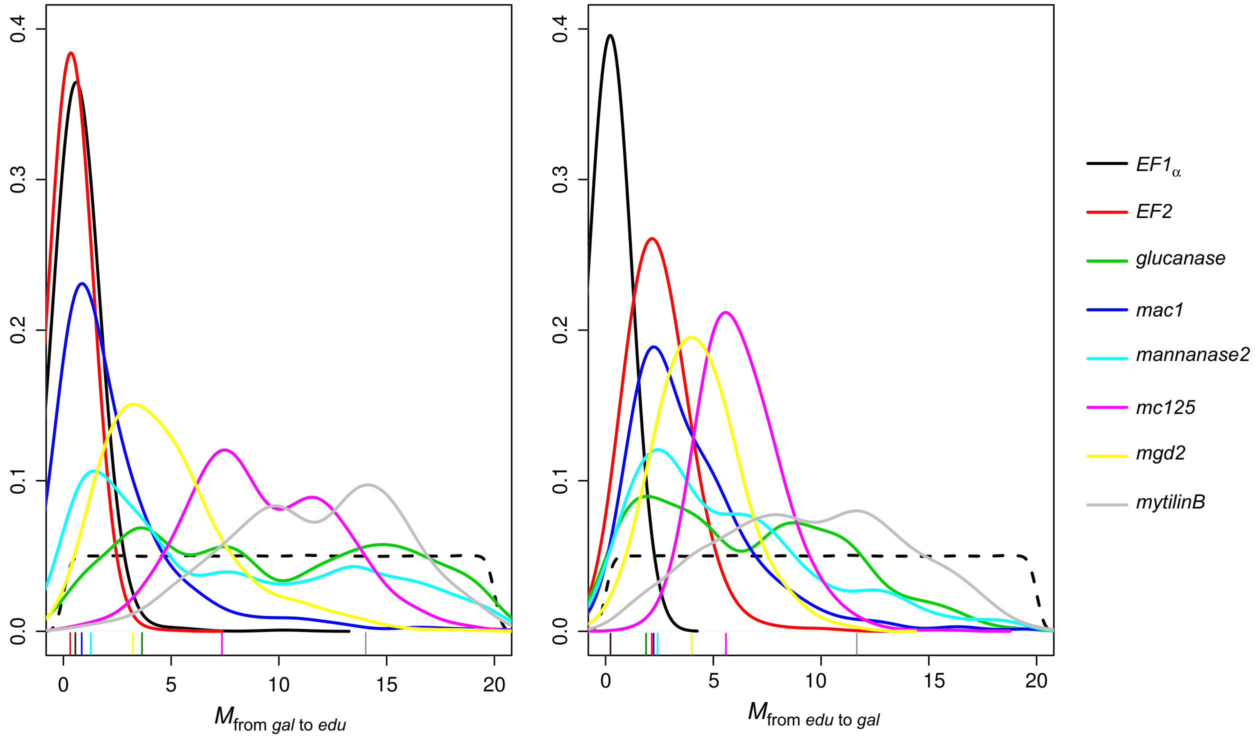

Fig. 5 Locus-specific estimates of introgression rates 4.*Nm* for the sequenced loci. The eight coloured lines represent posterior distributions for the sequenced loci. The dotted line represents the prior distribution.

ondary introgression occurred after a period of strict isolation.

Using an ABC-based model choice procedure, Roux *et al.* (2013b) investigated GWH in introgression rates between two highly divergent *Ciona intestinalis* species (≈14.4% of synonymous divergence). Due to the hierarchical Bayesian design, the procedure enabled estimation of the shape of the genomic distribution of migration rates (determined by the values of the hyperparameters *alpha* and *Beta*). A leptokurtic distribution was observed with the majority of the genome impeded by interspecific genetic barriers and only a small subset of the genome showing signals of introgression. Given the leptokurtic distribution, a large number of loci were necessary to identify regions introgressing between *Ciona* species.

For the *Mytilus* data set, the distribution of migration rates was L-shaped suggesting a predominance of genomic regions loosely permeable to introgression. As a result, GWH was identified in the *Mytilus* hybrid zone with reasonable power using only 8 loci. This result is consistent with a previous study investigating the genetic basis of post-zygotic isolation between the two mussel species that suggested the presence of a large number of recessive Dobzhansky–Muller incompatibilities across the genome (Bierne *et al.*, 2006). With broader genomic coverage, it might be possible to determine whether introgression acts over a large fraction of the genome or is restricted to small genomic regions similar to the genomic hot spots of introgression identified in *Ciona* (Roux *et al.*, 2013b).

Accounting for genomic variation in introgression rates when inferring the history of speciation

Detecting variation in introgression rates from multilocus gene genealogies represents an important advancement in the study of speciation, as previously emphasized by Sousa *et al.* (2013) and Roux *et al.* (2013b). However, one of the most important results of the present study is that by allowing GWH in introgression rates, we obtained statistical support for the secondary contact scenario in a model hybrid zone that the standard approach failed to support. Our simulations suggested that the subdivision of the ancestral mussel population occurred ~2.5 MY ago and was followed by a ~1.8 MY long period during which both *Mytilus* lineages remained completely isolated. This long period of allopatry is favourable for the accumulation of genetic incompatibilities (Navarro & Barton, 2003; Matute *et al.*, 2010; Moyle & Nakazato, 2010; Nachman & Payseur, 2012), and it is likely that a majority of the multifarious barriers to gene flow became fixed during this time. Following secondary contact, it is unclear whether gene flow has been continuous or intermittent due to distributional range shifts caused by glacial oscillations. Although the latter seems likely, our data set does not provide sufficient power to test for intermittent gene flow since secondary contact.

Our ABC approach supported the SC scenario, but only when we allowed heterogeneity in introgression rates among loci. Although the Bayes factor was moderate (BF = 1.3 with the Beta distribution and BF = 4.3 with two groups of loci with different gene flow), and one could have worried that model selection could be sensitive to prior choice, a parametric bootstrap procedure from thousand pseudo-observed data sets (Fagundes *et al.*, 2007) retrieved an estimate of the probability of type I error of 0.04 and 0.03 for the Beta distribution and the two-group parametrization of GWH, respectively. On the contrary, alternative models with homogeneous migration rates consistently led to ambiguous results. Neglecting genomic variation in introgression rates failed to distinguish between the IM and SC scenarios, and parameter estimates exhibited large variances in the posterior distributions of biologically relevant parameters (i.e. the times of initial divergence and secondary contact). At present, the ABC approach is the only method with sufficient flexibility to simultaneously incorporate GWH and formally test alternative scenarios of speciation (most importantly secondary contact vs. parapatric differentiation) with intralocus recombination. Although GWH has been incorporated in a full-likelihood approach (Sousa *et al.*, 2013), rigorous testing of the secondary contact model has been problematic, which has impeded testing of alternative geographic modes of speciation. Support for the IM model over the SI model has occasionally been taken as evidence for parapatric speciation even though secondary contact could account for the data equally well. Our results suggest that the conclusions drawn from some studies that have assumed similar levels of gene flow among loci may need to be revisited. Although introducing the additional complexity of GWH in introgression rates represents one improvement, it remains unclear how variation in other model parameters could affect the performance of speciation models (e.g. the divergence time, the mutation and recombination parameters). However, this is a general problem shared by all approaches developed to date, and the advantage of the model-based strategy rather lies in the comparison of nested models that specifically allow one to explore the effect of parameters that vary between models, even if the models compared are oversimplifications (Yang, 1997). It is likely that including additional complexities might improve our inferences, but they are unlikely to refute GWH and secondary contact in *Mytilus spp*. Since variable patterns of gene flow have been widely documented in numerous taxa including *Helianthus* sunflowers (Whitney *et al.*, 2010), *Heliconius* butterflies (Pardo-Diaz *et al.*, 2012), *Mus* mice (Song *et al.*, 2011) and *Ficedula* flycatchers (Ellegren *et al.*, 2012), it might prove useful to test these case studies with methods that explicitly account for GWH. It is clear that rigorous simulation studies are needed to further explore the effects of neglecting genomic heterogeneity of other parameters. However, we believe that the statistical evaluation of alternative models proposed in the hierarchical ABC framework should help to strengthen or refute modes of speciation that have remained intractable with full-likelihood approaches. This is especially pertinent for the model of secondary contact that may have often be too hastily dismissed by not taking into account variation in migration rates among loci.

Acknowledgments

Numerical results presented in this article were carried out using the ISEM computing cluster, and the authors highly appreciated the helpful support of its staff. We also highly appreciated and thank the technical staff of the CRI-Lille 1 centre. Molecular data were produced through the ISEM platform 'Génomique marine' at the Station Méditerranéenne de l'Environnement Littoral (OSU OREME), and we are very grateful to the National Sequencing Centre (Genoscope at Evry) and the Plateforme 'Génomique Environnementale' du Labex CeMEB for providing access to their sequencing platforms. We acknowledge Matthieu Faure for technical assistance and Nicolas Galtier for useful discussions. This study was supported by the Agence National de la Recherche (HySea project ANR-12-BSV7-0011) and the project Aquagenet (SUDOE, INTERREG IV B). This is article 2014-045 of Institut des Sciences de l'Evolution de Montpellier.

References

Addison, J.A., Ort, B.S., Mesa, K.A. & Pogson, G.H. 2008. Range-wide genetic homogeneity in the California sea mussel (*Mytilus californianus*): a comparison of allozymes, nuclear DNA markers, and mitochondrial DNA sequences. *Mol. Ecol.* **17**: 4222–4232.

Barton, N.H. 1979. The dynamics of hybrid zones. *Heredity (Edinb)* **43**: 341–359.

Barton, N. & Bengtsson, B.O. 1986. The barrier to genetic exchange between hybridising populations. *Heredity (Edinb)* **57**: 357–376.

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

Barton, N.H. & Hewitt, G.M. 1989. Adaptation, speciation and hybrid zones. *Nature* **341**: 497–503.

Bazin, E., Dawson, K.J. & Beaumont, M.A. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* **185**: 587–602.

Beaumont, M.A., Zhang, W. & Balding, D.J. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

Becquet, C. & Przeworski, M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**: 1505–1519.

Becquet, C. & Przeworski, M. 2009. Learning about modes of speciation by computational approaches. *Evolution (N.Y.)* **63**: 2547–2562.

Bierne, N., David, P., Boudry, P. & Bonhomme, F. 2002. Assortative fertilization and selection at larval stage in the mussels *Mytilus edulis* and *M. galloprovincialis*. *Evolution (N.Y.)* **56**: 292–298.

Bierne, N., Borsa, P., Daguin, C., Jollivet, D., Viard, F., Bonhomme, F. *et al.* 2003a. Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. *Mol. Ecol.* **12**: 447–461.

Bierne, N., Bonhomme, F. & David, P. 2003b. Habitat preference and the marine-speciation paradox. *Proc. Biol. Sci.* **270**: 1399–1406.

Bierne, N., Bonhomme, F., Boudry, P., Szulkin, M. & David, P. 2006. Fitness landscapes support the dominance theory of post-zygotic isolation in the mussels *Mytilus edulis* and *M. galloprovincialis*. *Proc. Biol. Sci.* **273**: 1253–1260.

Bierne, N., Tanguy, A., Faure, M., Faure, B., David, E., Boutet, I. *et al.* 2007. Mark-recapture cloning: a straightforward and cost-effective cloning method for population genetics of single-copy nuclear DNA sequences in diploids. *Mol. Ecol. Notes* **7**: 562–566.

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. *Curr. Zool.* **59**: 72–86.

Billard, E., Serrão, E., Pearson, G., Destombe, C. & Valero, M. 2010. *Fucus vesiculosus* and *spiralis* species complex: a nested model of local adaptation at the shore level. *Mar. Ecol. Prog. Ser.* **405**: 163–174.

Blum, M.G.B. & François, O. 2009. Non-linear regression models for Approximate Bayesian Computation. *Stat. Comput.* **20**: 63–73.

Boon, E., Faure, M.F. & Bierne, N. 2009. The flow of antimicrobial peptide genes through a genetic barrier between *Mytilus edulis* and *M. galloprovincialis*. *J. Mol. Evol.* **68**: 461–474.

Butlin, R. 1987. Speciation by reinforcement. *Trends Ecol. Evol.* **2**: 8–13.

Charlesworth, B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**: 195–205.

Charlesworth, B., Nordborg, M. & Charlesworth, D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70**: 155–174.

Cornuet, J.-M., Santos, F., Beaumont, M.A., Robert, C.P., Marin, J.-M., Balding, D.J. *et al.* 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**: 2713–2719.

Csilléry, K., François, O. & Blum, M.G. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**: 475–479.

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T. *et al.* 2012. The genomic landscape of species divergence in *Ficedula flycatchers*. *Nature* **491**: 756–760.

Endler, J.A. 1977. Geographic variation, speciation, and clines. *Monogr. Popul. Biol.* **10**: 1–246.

Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L. *et al.* 2007. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* **104**: 17614–17619.

Faure, B., Bierne, N., Tanguy, A., Bonhomme, F. & Jollivet, D. 2007. Evidence for a slightly deleterious effect of intron polymorphisms at the EF1alpha gene in the deep-sea hydrothermal vent bivalve Bathymodiolus. *Gene* **406**: 99–107.

Faure, M.F., David, P., Bonhomme, F. & Bierne, N. 2008. Genetic hitchhiking in a subdivided population of *Mytilus edulis*. *BMC Evol. Biol.* **8**: 164.

Feder, J.L. & Nosil, P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution (N.Y.)* **64**: 1729–1747.

Fraisse, C., Roux, C., Welch, J.J. & Bierne, N. 2014. Gene-flow in a mosaic hybrid zone: is local introgression adaptive? *Genetics* doi: 10.1534/genetics.114.161380.

Gardner, J.P.A. & Skibinski, D.O.F. 1988. Historical and size-dependent genetic variation in hybrid mussel populations. *Heredity (Edinb).* **61**: 93–105.

Geraldes, A., Carneiro, M., Delibes-Mateos, M., Villafuerte, R., Nachman, M.W. & Ferrand, N. 2008. Reduced introgression of the Y chromosome between subspecies of the European rabbit (*Oryctolagus cuniculus*) in the Iberian Peninsula. *Mol. Ecol.* **17**: 4489–4499.

Gosset, C.C. & Bierne, N. 2013. Differential introgression from a sister species explains high F(ST) outlier loci within a mussel species. *J. Evol. Biol.* **26**: 14–26.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**: e1000695.

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M. & Excoffier, L. 2005. Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**: 409–417.

Harrison, R.G. 1993. *Hybrid Zones and the Evolutionary Process*. Oxford University Press, New York.

Hartigan, J.A. & Hartigan, P.M. 1985. The dip test of unimodality. *Ann. Stat.* **13**: 70–84. Institute of Mathematical Statistics.

Hey, J. 2006. Recent advances in assessing gene flow between diverging populations and species. *Curr. Opin. Genet. Dev.* **16**: 592–596.

Hey, J. & Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.

Hey, J. & Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. USA* **104**: 2785–2790.

Hilbish, T., Carson, E., Plante, J., Weaver, L. & Gilg, M. 2002. Distribution of *Mytilus edulis, M. galloprovincialis,* and their hybrids in open-coast populations of mussels in southwestern England. *Mar. Biol.* **140**: 137–142.

Hilbish, T.J., Lima, F.P., Brannock, P.M., Fly, E.K., Rognstad, R.L. & Wethey, D.S. 2012. Change and stasis in marine hybrid zones in response to climate warming. *J. Biogeogr.* **39**: 676–687.

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

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. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**: 1735–1747.

Liu, Z.-L., Zhang, D., Wang, X.-Q., Ma, X.-F. & Wang, X.-R. 2003. Intragenomic and interspecific 5S rDNA sequence variation in five Asian pines. *Am. J. Bot.* **90**: 17–24.

Matute, D.R., Butler, I.A., Turissini, D.A. & Coyne, J.A. 2010. A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science* **329**: 1518–1521.

McVean, G., Awadalla, P. & Fearnhead, P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.

Moyle, L.C. & Nakazato, T. 2010. Hybrid incompatibility "snowballs" between Solanum species. *Science* **329**: 1521–1523.

Nachman, M.W. & Payseur, B.A. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**: 409–421.

Navarro, A. & Barton, N.H. 2003. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution (N.Y.)* **57**: 447–459.

Nielsen, R. & Wakeley, J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**: 885–896.

Nosil, P. 2008. Speciation with gene flow could be common. *Mol. Ecol.* **17**: 2103–2106.

Nosil, P. 2012. *Ecological Speciation*. Oxford University Press, Oxford and New York.

Nosil, P. & Feder, J.L. 2012. Widespread yet heterogeneous genomic divergence. *Mol. Ecol.* **21**: 2829–2832.

Ort, B.S. & Pogson, G.H. 2007. Molecular population genetics of the male and female mitochondrial DNA molecules of the California sea mussel, *Mytilus californianus*. *Genetics* **177**: 1087–1099.

Palumbi, S.R. 1992. Marine speciation on a small planet. *Trends Ecol. Evol.* **7**: 114–118.

Pardo-Diaz, C., Salazar, C., Baxter, S.W., Merot, C., Figueired-o-Ready, W., Joron, M. *et al.* 2012. Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS Genet.* **8**: e1002752.

Pereyra, R.T., Bergström, L., Kautsky, L. & Johannesson, K. 2009. Rapid speciation in a newly opened postglacial marine environment, the Baltic Sea. *BMC Evol. Biol.* **9**: 70.

Pialek, J. & Barton, N.H. 1997. The spread of an advantageous allele across a barrier: the effects of random drift and selection against heterozygotes. *Genetics* **145**: 493–504.

Pinho, C. & Hey, J. 2010. Divergence with gene flow: models and data. *Annu. Rev. Ecol. Evol. Syst.* **41**: 215–230. Annual Reviews.

Ross-Ibarra, J., Wright, S.I., Foxe, J.P., Kawabe, A., De-Rose-Wilson, L., Gos, G. *et al.* 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* **3**: e2411.

Ross-Ibarra, J., Tenaillon, M. & Gaut, B.S. 2009. Historical divergence and gene flow in the genus Zea. *Genetics* **181**: 1399–1413.

Roux, C., Castric, V., Pauwels, M., Wright, S.I., Saumitou-Laprade, P. & Vekemans, X. 2011. Does speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation? *PLoS ONE* **6**: e26872.

Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V. & Vekemans, X. 2013a. Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A. lyrata*. *Mol. Biol. Evol.* **30**: 435–447.

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

Schluter, D. & Rambaut, A. 1996. Ecological speciation in postglacial fishes [and discussion]. *Philos. Trans. R. Soc. B Biol. Sci.* **351**: 807–814.

Secor, C.L., Day, A.J. & Hilbish, T.J. 2001. Factors influencing differential mortality within a marine mussel (Mytilus spp.) hybrid population in southwestern England: reproductive effort and parasitism. *Mar. Biol.* **138**: 731–739.

Skibinski, D.O.F., Beardmore, J.A. & Cross, T.F. 1983. Aspects of the population genetics of Mytilus (Mytilidae; Mollusca) in the British Isles. *Biol. J. Linn. Soc.* **19**: 137–183.

Smadja, C.M. & Butlin, R.K. 2011. A framework for comparing processes of speciation in the presence of gene flow. *Mol. Ecol.* **20**: 5123–5140.

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H. *et al.* 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr. Biol.* **21**: 1296–1301.

Sousa, V.C., Grelaud, A. & Hey, J. 2011. On the nonidentifiability of migration time estimates in isolation with migration models. *Mol. Ecol.* **20**: 3956–3962.

Sousa, V.C., Carneiro, M., Ferrand, N. & Hey, J. 2013. Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics* **194**: 211–233.

Strasburg, J.L. & Rieseberg, L.H. 2010. How robust are "isolation with migration" analyses to violations of the IM model? A simulation study. *Mol. Biol. Evol.* **27**: 297–310.

Strasburg, J.L. & Rieseberg, L.H. 2011. Interpreting the estimated timing of migration events between hybridizing species. *Mol. Ecol.* **20**: 2353–2366.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.

Tajima, F. 1989. The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.

Tavaré, S., Balding, D.J., Griffiths, R.C. & Donnelly, P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145**: 505–518.

Wakeley, J. & Hey, J. 1997. Estimating ancestral population parameters. *Genetics* **145**: 847–855.

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256–276.

Whitney, K.D., Randell, R.A. & Rieseberg, L.H. 2010. Adaptive introgression of abiotic tolerance traits in the sunflower Helianthus annuus. *New Phytol.* **187**: 230–239.

Wu, C.-I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* **14**: 851–865.

Yang, Z. 1997. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* **14**: 105–108.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Distribution of pairwise interspecific molecular divergence among loci.

Figure S2 Empirical relationship between the relative posterior probability of hetero- or homo-alternative model for the three models with migration and the associated probability to support the correct model.

Figure S3 Empirical distributions of the estimated relative probabilities of the SC model when the SI (red line), the IM (blue line), the AM (green line) and the SC (black line) models are the true models.

Figure S4 Parameter estimates of the best-supported model of speciation SC-hetero.

Table S1 Primers used to amplify investigated loci.

Table S2 Estimates of the recombination rate ρ (=4.N.r) using LDhat.

Table S3 Goodness-of-fit test to assess whether the parameter values drawn from the posterior distribution estimated by our ABC approach fit the data.

Received 27 February 2014; revised 25 April 2014; accepted 25 April 2014