DILS : Demographic Inferences with Linked Selection by using ABC

Christelle Fraïsse[1], Iva Popovic[2], Clément Mazoyer[3], Jonathan Romiguier[4], Étienne Loire[5], Alexis Simon[4], Nicolas Galtier[4], Laurent Duret[6], Nicolas Bierne[4], Xavier Vekemans[3], Camille Roux[3]

[1] *Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria*
[2] *School of Biological Sciences, University of Queensland, St Lucia, Qld, Australia*
[3] *Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France*
[4] *ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France*
[5] *Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR, ASTRE, Montpellier, France*
[6] *Laboratoire de Biologie et Biométrie Évolutive CNRS UMR 5558, Université Claude Bernard Lyon 1, Lyon, France*

✉: christelle.fraisse@ist.ac.at; camille.roux@univ-lille.fr

ABSTRACT

We present DILS, a deployable statistical analysis platform for conducting demographic inferences with linked selection from population genomic data using an Approximate Bayesian Computation framework. DILS takes as input single-population or two-population datasets and performs three types of analyses in a hierarchical manner, identifying: 1) the best demographic model to study the importance of gene flow and population size change on the genetic patterns of polymorphism and divergence, 2) the best genomic model to determine whether the effective size Ne and migration rate $N.m$ are heterogeneously distributed along the genome and 3) loci in genomic regions most associated with barriers to gene flow. Also available via a web interface, an objective of DILS is to facilitate collaborative research in speciation genomics. Here, we show the performance and limitations of DILS by using simulations, and finally apply the method to published data on a divergence continuum composed by 28 pairs of $Mytilus$ mussel populations/species.

Contents

1 Introduction **2**

2 Results **3**
2.1 Comparisons of demographic and genomic models 4
2.1.1 Single population models . . . 4
2.1.2 Two population models 5
2.1.3 Detection of barriers to gene flow 7
2.2 Parameter estimates 8
2.2.1 Single-population models . . . 8
2.2.2 Two-population models 8
2.3 Illustration of DILS with RNA-seq data from 28 pairs of *Mytilus* populations . 9

3 Discussion **10**
3.1 Performances of DILS 10
3.2 Collaborative research 11
3.3 Non-detailed features 11
3.4 Prospects for the future 11

4 Materials and methods **12**
4.1 ABC 12
4.1.1 Summary statistics 12
4.1.2 Simulations 12
4.1.3 Model comparisons 12
4.1.4 Parameter estimates 13
4.1.5 Locus-specific model comparison 13
4.2 Pseudo-observed datasets 13
4.3 Analysis of the *Mytilus* dataset 13

5 Acknowledgements **14**

6 Resources **14**

7 Supplementary material **16**

1 Introduction

Population genomic data along with efficient computational methods are becoming increasingly available, paving the way to broad-scale application of model-based inferences for understanding signatures of evolutionary processes (Lohse, 2017). Neutral processes such as divergence, gene flow and changes in population size all shape patterns of genomic variation; and so demographic models attempting to reconstruct the past history of single populations or closely-related species can also serve as null models in genome scans for selection. Considering a single species, model-based inferences are especially suitable in domesticated crops for disentangling the effect of population size changes from selection on agronomic traits (Gaut et al., 2018). Two-population models allow to tackle issues on speciation genomics, where this approach provides direct testing of distinct modes of speciation (Sousa and Hey, 2013), with at the two extremes a model of allopatric speciation that occurs in complete isolation and a model where speciation is opposed by continuous gene flow. This is critical to build-up a unifying picture of the genic view of speciation by quantifying the reduction in gene exchange between lineages as a function of their molecular divergence (Roux et al., 2016; Peñalba, Joseph, and Moritz, 2019); and identify *in silico* genomic regions harboring speciation genes, given that their barrier effects can only be detected in the presence of ongoing gene flow (Roux et al., 2013). At a broader scale, model-based inferences can be applied to community ecology to infer, for example, the assembly history of trophically linked species (Bunnefeld et al., 2018).

Various methods have been proposed to extract such information from population genomic data. Site frequency spectrum (SFS)-based methods compute or approximate the likelihood of the allele frequency distribution from a demographic model using either the diffusion approximation (Gutenkunst et al., 2009), the moment closure (Jouganous et al., 2017) or the coalescent (Excoffier et al., 2013). While these methods are fast, they ignore linkage information which is informative about past demography (Terhorst and Song, 2015). Therefore, other methods rely on the block-wise SFS, i.e. the SFS of short non-recombining blocks of sequences (Lohse, Harrison, and Barton, 2011). That way the genealogical information contained within each block is combined along the genome. Other multi-locus methods can jointly infer recombination and demography, therefore capturing longer range linkage disequilibrium, but they are still restricted to simple demographic histories excluding migration (Terhorst, Kamm, and Song, 2017). Still, the flexibility of simulation-based approximate Bayesian computation (ABC) enables including recombination within unlinked blocks in multi-locus inference of complex (and hopefully more realistic) evolutionary scenarios (Beaumont, Zhang, and Balding, 2002). Although more computationally expensive, the analysis of thousands of loci results in high-precision parameter estimation for most demographic scenarios (Robinson et al., 2014; Smith and Flaxman, 2020).

In this paper, we present an ABC framework (DILS) building upon and extending current statistical machinery (Pudlo et al., 2015; Roux et al., 2016). Our method is flexible both in terms of the evolutionary scenarios that can be accommodated (allowing changes in population size over time, linked selection and implementing various models of migration), and type of data (SFSs and/or multi-locus sequences). A major improvement compared to most existing methods is decoupling the effect of linked selection and neutral history by relaxing the assumption that all loci share the same demography (see Sethuraman, Sousa, and Hey, 2019; Sousa et al., 2013 and Lohse pers. comm. for similar ideas). We model variation in the rate of drift among loci to account for linked selection effects due to background selection (i.e. purifying selection) and adaptive sweeps in low-recombining and gene-dense regions. And by explicitly modelling variation in migration rates among loci in two-population models, we can capture the effect of selection against migrants at neutral markers linked to species barriers, and so analyse further these candidate genomic regions for reproductive isolation (Roux et al., 2013).

DILS is deployable on every machine, but also offers an online platform for configuring demographic inferences based on genomic data of thousands of loci, performing them and visualizing the returned output. These advances are made possible by progress in simulator performance (Hudson, 2002), reduction in the number of simulations required to train prediction algorithms (Pudlo et al., 2015) and development of computer clusters and tools facilitating parallelism (Köster and Rahmann, 2012). Following other user-friendly ABC programs, DILS aims to ease the use of high-performance tools for non-experts in methodology (Cornuet et al., 2008; Cornuet et al., 2014). Importantly, as there is a limit to how much information can be extracted from genomic data, DILS also implements rigorous quality controls. Therefore, not only does the user receive 1) the best-supported model among those proposed (figure 1), 2) an estimate of the demographic parameters describing this model and 3) a locus-specific test to identify barriers to gene flow (when relevant); the user will also get feedback on whether the best model is relevant and to which extent the estimates are able to reproduce the observed data.

A long-term aim of DILS is to facilitate collaborative research in speciation genomics. The degree of reproductive isolation appears to follow a quasi-shared molecular clock among animals, depending on the level

of net genomic divergence between lineages (Roux et al., 2016). However, for the same level of divergence, two opposite situations coexist in the so-called grey zone of speciation with, on the one hand, semi-isolated pairs capable of exchanging genetic material and, on the other hand, pairs of species that are fully reproductively isolated. Many hypotheses have been advanced to explain such a reproductive barrier contrast within the same range of molecular divergence, including differences related to life history traits (internal versus external fertilization), ecology (marine versus terrestrial organisms), reproductive systems (e.g. in plants: self-incompatibility versus self-fertilization), genome size and recombination landscape, functional redundancies in genomes, etc... Speciation is such a multi-factorial process that it seems impossible for a single research group to study these different components.

Consequently, our aim is to include in DILS a collaborative science option, allowing to feed in real time the relationship between molecular divergence and genetic isolation between lineages. Available as a choice, the sharing of the inferences made in DILS, associated with the expertise that users have about their biological model, will contribute to a long-term collaborative study aiming to better understand the speciation process. This objective is illustrated here with the analysis of 28 new pairs of mussel populations whose transcriptomes were recently published, revealing ongoing gene flow for levels of divergence greater than 2%.

In this study, we have four objectives: 1) providing a flexible and powerful demographic inference method with linked selection to analyse genome-scaled dataset in single and two-population models; 2) presenting a user-friendly tool that implements this approach and paves the way for collaborative science; 3) testing the performance and limitations of the method by using simulations; 4) and applying it to an empirical dataset of $Mytilus$ mussels.

2 Results

In the current version of DILS, evolutionary scenarios can be investigated for sampling schemes involving one or two populations. For both types of analysis, the first step in DILS is to compare the demographic models described in figure 1. With a single population, DILS will examine the changes in size over time. With two populations, such variations in population size are also implemented, but DILS will additionally compare alternative temporal patterns of introgression.

An innovative feature of DILS is to include linked selection, either through the effect of background selection (and selective sweeps) that modulates the effective population size along the genome, or through the effect of selection against migrants that reduces locally the effective introgression rate in genomic regions locked to gene flow. Therefore, all demographic models exist under two alternative genomic models regarding the effective population size (homo-*Ne versus* hetero-*Ne*), and, in models with migration, the introgression rate (homo-*N.m versus* hetero-*N.m*), depending on whether these parameters are homogeneous or heterogeneous among loci.

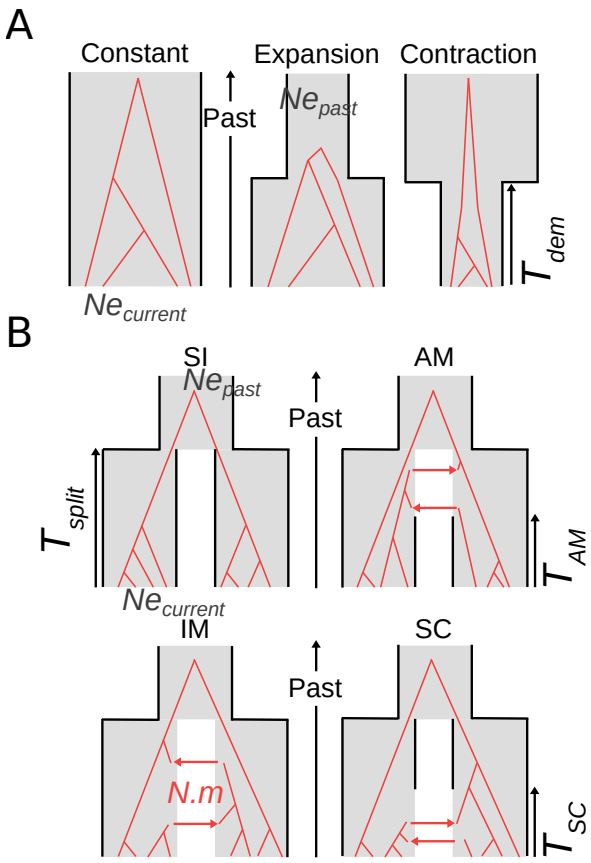

Figure 1. Demographic models currently implemented in DILS
A: compared single-population models. Demographic changes occurring T_{dem} generations ago are modeled backwards in time by sudden transitions from Ne_{current} to Ne_{past}, either for expansions or contractions.
B: compared two-population models. The Strict Isolation (SI) and Ancient Migration (AM) models are characterized by an absence of ongoing migration. Conversely, the Isolation with Migration (IM) and Secondary Contact (SC) models describe two populations that are currently connected by introgression events at rate $N.m$. The two-population models shown here are of constant size, but DILS optionally incorporates alternative versions of the same four models where effective size can change independently in both daughter populations between the present time and T_{split}. This is a relevant addition given the influence of over-time size changes on demographic inferences in speciation scenarios (Momigliano, Florin, and Merilä, 2020)

2.1 Comparisons of demographic and genomic models

In this section, we present DILS performance to compare demo-genomic models involving one (section 2.1.1) or two (section 2.1.2) populations. These evaluations were performed by analyzing pseudo-observed datasets simulated under specified models, in order to assess the efficacy of our approach to correctly support the true model. In both analyses, a given demographic model corresponds to the set of all its genomic submodels. All model comparisons are performed using random forest algorithms (Pudlo et al., 2015; Fraimout et al., 2017).

2.1.1 Single population models

For studies where a single population is sampled, three demographic models are compared describing either $i)$ a Constant population size Ne_{current}, $ii)$ recent demographic Expansion or $iii)$ Contraction. Demographic changes are assumed to be instantaneous, with a transition from Ne_{past} to Ne_{current} occurring T_{dem} generations ago (figure 1-A).

For a given dataset, we first estimate the best-fitting demographic model among those depicted in figure 1 by carrying out $10,000$ simulations under each genomic alternative sub-model (homogeneous *versus* heterogeneous *Ne*). These simulations produce reference tables, *i.e.*, a set of simulated summary statistics used to train a random forest algorithm to predict which of the proposed models best explains observed data. Thus, in the comparison Expansion *versus* Constant *versus* Contraction (algorithm 1), each model was simulated $10,000$ times under the homogeneous *Ne* model, and $10,000$ times under the heterogeneous *Ne* model.

To test the performance of DILS in model comparisons (among demographic models and among genomic models), we simulated $30,000$ pseudo-observed datasets of 100 loci by using random combinations of parameters. Throughout the manuscript, pseudo-observed datasets are used to evaluate the performance of the random forest and do not contribute to its training. These simulated datasets are equally distributed between the demographic models ($10,000$ for each of the three Expansion/Constant/Contraction models) and between the genomic models ($5,000$ for each of the two homo/hetero *Ne* alternative genomic models for a given demographic model). Then, for each of these pseudo-observed datasets, we apply step 1 and step 2 of the algorithm 1 in order to obtain for each model *M*, the proportion of pseudo-observed datasets that is *i*) correctly and strongly captured by the random forest approach, *ii*)falsely and strongly captured and *iii*) ambiguous, i.e, associated to an insufficiently high posterior probability (figure 2).

For a given dataset, the model comparison will provide two pieces of information: *i*) What is the best model among all those arbitrarily proposed in the analysis? *ii*) With which posterior probability is the best model supported? It is from the latter probability that we determine whether an inference is strong or ambiguous. For this we use a probability threshold beyond which an inference is considered strong. This threshold is determined recursively on the basis of the false positive rate which decreases monotonically by increasing the value of the threshold. From datasets randomly simulated under different models, we establish a threshold value such that the false positive rate is less than or equal to 1%. With this approach, the false-positive rate remains consistently low, but the relative proportions of true positives *versus* ambiguous cases vary according to the power of the ABC to discriminate among an arbitrary set of models. Thus, among the $30,000$ pseudo-observed datasets simulated under the Expansion, Constant and Contraction models, if a threshold is applied that keeps the error rate below 1%, the proportions that are correctly supported by our approach are 86%, 89% and 99% respectively, while the proportions that are ambiguous are 13%, 10% and 0.2% (figure 2).

The second step of the hierarchical comparison, which classifies models with genomic variation in effective size (hetero) or without variation (homo), is evaluated using the same procedure as step 1 (figure 2). For this purpose, the proportion of ambiguous, correct and false inferences are measured among $5,000$

Data: A single fasta containing all genes sequenced in all individuals sampled in the studied population
Result: Posterior probabilities for the best $1)$ demographic and $2)$ genomic models

- **Data cleaning:**
  forall *genes i making the dataset* **do**
  .discard from the alignment of gene *i* the individuals with too many *N*s;
  .discard gene *i* if there are not enough retained individuals;
  .discard gene *i* if it doesn't contain enough positions without *N*
  end
- **Reference simulations to train the random-forest:** $10,000$ multilocus datasets under each combination of [demographic models] x [genomic models];
- **Model comparisons:**
  Random Forest (RF) comparisons of:
  step 1. Expansion [homo + hetero] *versus* Constant [homo + hetero] *versus* Contraction [homo + hetero];
  step 2. best demographic model supported in the previous step with homogeneous *Ne versus* heterogeneous *Ne*;

Algorithm 1: Single-population hierarchical model comparisons

pseudo-observed datasets simulated for each of the 6 combinations [demographic models] x [genomic models].

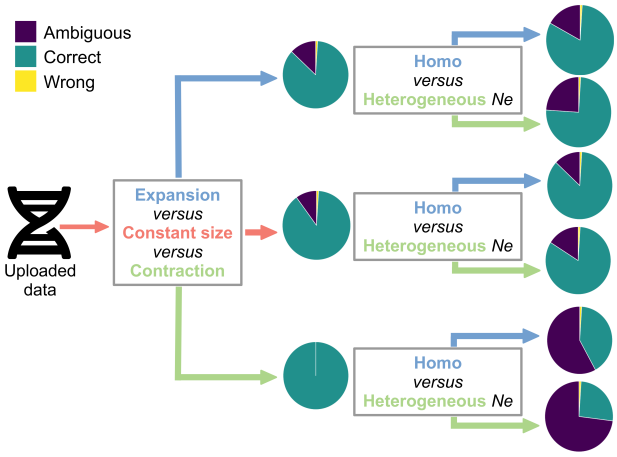

Figure 2. Performance of DILS for hierarchical comparison of single-population models
The required data is a single fasta file containing the sequences obtained for different genes, in different individuals. DILS will first perform a comparison of the three demographic models (Expansion $versus$ Constant $versus$ Contraction). In a second step, DILS will compare two genomic models (homogeneous $versus$ heterogeneous genomic distribution of Ne) for the best supported demographic model.
The pie charts designate for each model the proportion of simulations performed under the corresponding model that is strongly and correctly captured (correct: blue), strongly and incorrectly captured (wrong: yellow) and without strong statistical support for any of the studied models (ambiguous: purple).
The performance of DILS for each comparison was based on $10,000$ pseudo-observed datasets for each of the Expansion/Constant/Contraction demographic models. Each of these 10,000 simulated datasets are evenly distributed between the two genomic models, homo and hetero *Ne*. The parameters used for the simulated datasets are randomly drawn from uniform laws, with Ne in [1-1,000,000] individuals and T_{dem} in [1-2,000,000] generations.
Each simulated dataset consists of 100 loci of length $1,000$ nucleotides.

For the Expansion and Constant demographic models, the correct recapture rates of homo and hetero *Ne* genomic models range from 75% (Expansion + hetero *Ne*; 24% of ambiguity) to 86% (Constant + homo *Ne*; 13% of ambiguity). Finally, while recovering the Contraction demographic model is a very robust analysis with 99% of inferences that are both correct and associated with a high posterior probability, it is more complicated to distinguish "Contraction + homo *Ne*" from "Contraction + hetero *Ne*". About 41% of the pseudo-observed datasets simulated in the "Contraction + homo *Ne*" model are correctly captured by the random forest, and only 26% for the "Contraction + hetero *Ne*" model. The occurrence of a recent bottleneck tends to reduce the genomic variance of *Ne* to levels that generate apparent homogeneity.

2.1.2 Two population models

The two-population models are grouped into two supermodels: with current isolation (Strict Isolation + Ancient Migration; see figure 1) and ongoing migration (Isolation Migration + Secondary Contact). The first step of the hierarchical comparisons performed by DILS therefore aims to determine which supermodel best explains the data observed in the two sampled populations (see algorithm 2 and figure 3).

Data: A single fasta containing all genes sequenced in all individuals sampled in the two studied populations
Result: Posterior probabilities for $1)$ ongoing migration, $2)$ [SI, AM] (in case of current isolation) or [IM, SC] (in case of ongoing migration), $3)$ *Ne* and *N.m* (in case of ongoing migration)

- **data cleaning:**
  forall *genes i* **do**
  forall *population/species j* **do**
  .discard from the alignment of gene *i* the individuals with too many *N*s;
  .discard gene *i* if there are not enough retained individuals in population *j*;
  .discard gene *i* if it doesn't contain enough positions without an aligned null allele *N*;
  end
  end
- **Reference simulations to train the random-forest:** $20,000$ multilocus datasets under each combination of [demographic models] x [genomic models];
- **Model comparisons:**;
  Random Forest (RF) comparisons of:;
  step 1. isolation ([all SI + all AM]) *versus* migration ([all IM + all SC]);
  step 2. if isolation **then** [all SI] *versus* [all AM] **else** [all IM] *versus* [all SC];
  step 3. if isolation **then** [homo-*Ne* SI + homo-*Ne* AM] *versus* [hetero-*Ne* SI + hetero-*Ne* AM] **else** [homo-*Ne* IM + homo-*Ne* SC] *versus* [hetero-*Ne* IM + hetero-*Ne* SC];
  step 4. if migration **then** [homo-*N.m* IM + homo-*N.m* SC] *versus* [hetero-*N.m* IM + hetero-*N.m* SC];

Algorithm 2: Two-population hierarchical model comparisons

This is achieved by labeling as "isolation" all reference simulations performed under the two SI models

(with homo-*Ne* or hetero-*Ne*) and the four AM models (homo-*Ne* or hetero-*Ne*, and homo-*N.m* or hetero-*N.m*). All other models are labeled as "migration" supermodel.

To test the power of DILS to correctly recapture an isolation or migration model, the same simulation-based evaluation as in section 2.1.1 is performed here. We evaluated the performance of DILS for $60,000$ pseudo-observed datasets simulated under the "isolation" supermodel ($10,000$ for each combination of [SI; AM], [homo-*Ne*; hetero-*Ne*] and [homo-*N.m*; hetero-*N.m*] for the AM model only) and $80,000$ under the "migration" supermodel ($10,000$ for each combination of [IM; SC], [homo-*Ne*; hetero-*Ne*] and [homo-*N.m*; hetero-*N.m*]). As shown in figure 3, 95% of the datasets simulated under the supermodel "isolation" with random combinations of parameters from large priors are correctly recaptured by the random forest approach with a high probability (4% ambiguity and 1% error if we apply a posterior probability threshold of 0.84; table S1)). Similarly, 98% of the pseudo observed datasets under the "migration" supermodel are strongly recaptured (with 1% of ambiguity and 1% of error for a threshold of 0.665). Models with migration are globally more efficiently recaptured by DILS, relying on a lower threshold probability to be robustly supported.

The second step of the hierarchical model comparisons is to classify the demographic models described in figure 1-B, within the supermodel that was best supported in the previous step (figure 3-B).

The results of the performance analyses first show that a pseudo-observed dataset simulated under an SI model (homo-*Ne* and hetero-*Ne*) is very unlikely to be strongly supported in an SI *versus* AM comparison. Out of $20,000$ simulations, only 1% are correctly recaptured by DILS (98% ambiguity and 1% error for a threshold of 0.845). The AM model is more robustly supported than SI (56%), but the $10,000$ inferences made under each of the AM sub-models lead to weak support (43% ambiguity, 1% error for a threshold of 0.705).

On the contrary, the two models making the "migration" supermodel (IM and SC) are more efficiently distinguished by DILS. The $40,000$ pseudo-observed datasets randomly simulated under the IM model are captured at 79% with a high probability in the IM *versus* SC comparison (20% ambiguity, 1% error for a threshold of 0.885). Similarly, 70% of the $40,000$ pseudo-observed datasets from the SC model are correctly recaptured by DILS (29% ambiguity, 1% error for a threshold of 0.915).

We then evaluate DILS performance for discriminating among alternative models of genomic distribution for the *Ne* (figure 3-B; table S2) and *N.m* (figure 3-C).

Concerning the effective population size, DILS systematically recaptures the homogeneous model more easily than the heterogeneous model for each of the four demographic models tested. The most complicated model to recapture is the genomic heterogeneity of *Ne* in an SC model ($\approx 64\%$ true positives), while homo-*Ne* under an SI model is the most straightforward.

Concerning the genomic model of introgression, the main parameter determining the quality of inference is the relative duration of gene flow versus speciation time T_{split} (figure 1). This leads to a higher robustness for the IM model (probabilities of correctly supporting homo-*N.m* and hetero-*N.m* of $\approx 86\%$ and $\approx 76\%$ respectively) compared to the SC model ($\approx 71\%$ and $\approx 60\%$).

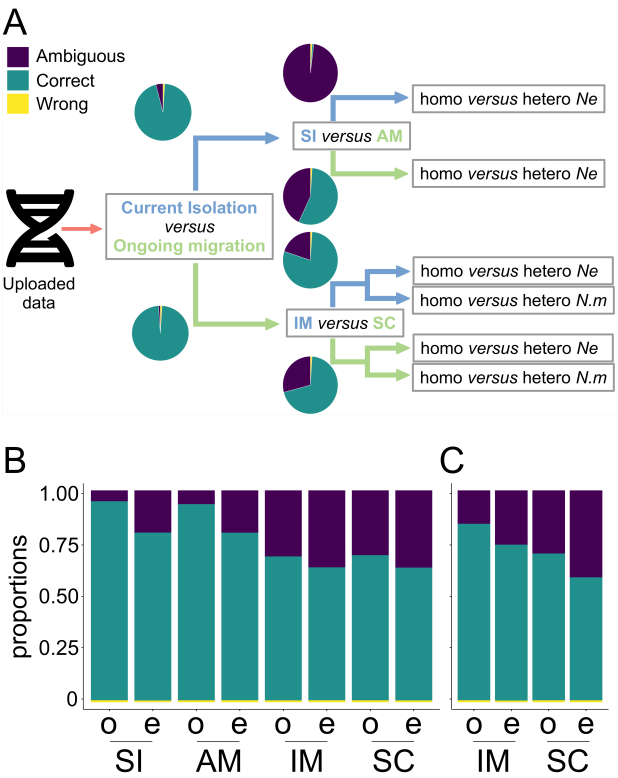

Figure 3. Performance of DILS for hierarchical comparison of two-population models
Two-population analyses are performed in three steps
1. Testing the general demographic models "Current isolation" versus "Ongoing migration" (panel **A**)
2 Testing the demographic sub-models (panel **A**):
2.a if the best model is 'current isolation', then DILS tests SI versus AM
2.b if the best model is 'ongoing migration', then DILS tests IM versus SC
3. Testing the genomic models for variation of:
3.a effective population size, Ne (panel **B**)
3.b migration rate, $N.m$ (panel **C**; only for ongoing migration models)
The letters 'o' and 'e' in panels B and C indicate simulations performed under genomic h**o**mogeneity and h**e**terogeneity models, respectively
The pie charts designate for each model the proportion of simulations performed under the corresponding model that is strongly and correctly captured (correct: blue), strongly and incorrectly captured (wrong: yellow) and without strong statistical support for any of the studied models (ambiguous: purple).

2.1.3 Detection of barriers to gene flow

One additional purpose of DILS is to identify genomic regions which are associated to a local reduction in the introgression rate *N.m*. This analysis will only be carried out by DILS if the observed dataset is better explained by 1) a demographic model with ongoing migration (IM or SC) and 2) a genomic model with gene flow heterogeneity (hetero-*N.m*). To achieve this purpose, DILS will infer the parameters under the model that best explains the data. Then, for each gene, DILS performs a comparison between two models that differ only for the parameter *N.m*: 1) the *migration* model corresponds to the whole set of parameters estimated under the best supported model; 2) the *isolation* model corresponds to the previous model whose *N.m* has been set to zero, because a barrier gene impedes gene flow locally along the chromosome. Therefore, such a gene should be supported by the isolation model if the barrier effect is strong. This approach therefore seeks to approximate a continuous variable, *N.m*, by a dichotomous choice of model: region with a local-isolation *versus* local-migration.

In order to evaluate the performance of DILS, the locus-specific model comparison was applied for locus simulated under an IM model with different values of *N.m* in $[0, 1]$ (figure 4). A value of zero means no exchange during the divergence process from one population to another. A value of 1 means that there is one immigrant individual on average every generation. We simulated genes $10,000$ times for different combinations of *N.m* and T_{split} under an IM model. Then, for each simulated dataset, we applied the locus-specific model comparison to finally record for each locus which model is the best between local-migration and local-isolation. Ideally, we aim that DILS considers 100% of the simulations with $N.m = 0$ as local-isolation, and 100% of the simulations with $N.m = 1$ as local-migration. Values of *N.m* greater than 1 were not explored because the comparison between "high migration" and "very high migration" is not relevant here.

As shown in figure 4, in the case of two populations of $100,000$ individuals separated only $5,000$ generations ago ($T_{\text{split}} = 0.05$), DILS will support gene flow for $\approx 50\%$ of the genes that have a *N.m* migration equals to zero (figure 4). For $N.m = 1$, the proportion of genes inferred as local-isolation is of similar magnitude, indicating that DILS is not at all designed to detect barriers to gene flow in the genomes of populations that have separated very recently. Our recommendation, therefore, is to disregard the results of DILS if the studied populations are extremely recent. However, as soon as barrier regions have enough time to differentiate ($T_{\text{split}} \geq 0.5$; figure 4), then $\approx 100\%$ of loci with $N.m = 0$ are correctly inferred as local-isolation, and only few loci with $N.m = 1$ are incorrectly supported by the model of local-isolation. The performance of DILS therefore depends directly on the true history of the studied populations/species, not on the amount of data. The ideal case for identifying which genes in the genome are linked to barriers occurs when the patterns of polymorphism and divergence at such genes differ greatly from the rest of the genomic background (figure 4). An ideal demographic scenario for identifying barriers with DILS would be :

1. a divergence that is old enough to allow the neutral regions linked to barriers to have at least one position with two variants that are differentially fixed between the two populations/species.
2. a migration rate in the genomic background high enough to counteract the effect of differentiation in non-barrier regions.

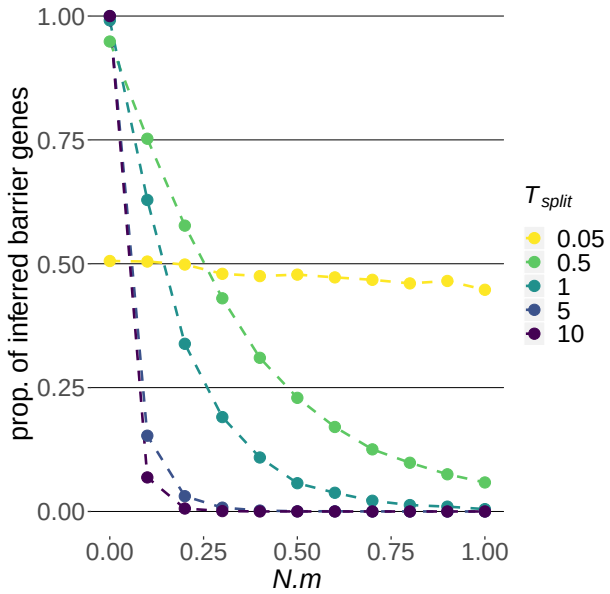

Figure 4. Detection of barriers to gene flow
x-axis : 11 explored values of the locus-specific $N.m$ migration rate under an IM model.
y-axis : proportion of simulations supported by DILS as being linked to a barrier to gene flow.
The colors designate five different divergence times of the IM model (T_{split}, figure 1). The unit time is in *Ne* generations where *Ne* is the number of haploid individuals making up the population. If $Ne = 100,000$ individuals, then $T_{\text{split}} = 5$ means a divergence time of $500,000$ generations under the IM model. If *Ne* is the number of diploids, then T_{split} must be multiplied by two to find the same relationship.
Each combination of T_{split} and *N.m* was independently simulated $10,000$ times and analyzed by DILS to get the proportion of model-assignation for a given combination of parameters.
The estimated points are connected by dotted lines for visibility.

2.2 Parameter estimates

In this section, we describe performance tests for estimating the parameters of different demographic models. The same procedure was applied for single and two-population models: first, simulating pseudo-observed datasets ($10,000$ for the three single-population models, $2,000$ for the 14 two-population models) and then ABC estimation of the parameters to test DILS ability to recapture the parameter values used. We only detail here the results obtained for the demographic parameters, *i.e.*, those describing the mean effective population size (Ne_{current} and Ne_{past}), time of split (T_{split}), the date for the cessation of gene flow (T_{AM}), the age of the secondary contact (T_{SC}) and the migration rates ($N.m$).

2.2.1 Single-population models

The current effective population size is by far the parameter that is most accurately recaptured, especially in a constant and homogeneous model with a mean-squared error (MSE) close to zero (MSE≈ 0.005; figure 5-A; table S2). The introduction of a recent demographic change reduces the quality of inferences for Ne_{current}, more for the expansion model than for the contraction model. The inference of ancestral size conducted for 4 x $10,000$ pseudo-observed datasets shows globally a low error rate on the raw values of Ne_{past} with a $MSE_{\text{max}} \approx 0.09$ (table S2). Errors depend very closely on the relative values between Ne_{past} and Ne_{current} as shown on figure 5-B with the estimates of the ratio $Ne_{\text{past}}/Ne_{\text{current}}$. Hence, while the estimate of a Ne_{past} is reliable for a change in size by a factor of 10, it becomes less accurate as the contrast with Ne_{current} is increasingly sharp. In a similar manner, the quality of inferences of the age of demographic change T_{dem} is highly dependent on its relative value with Ne_{current} (figure 5-C where the age of the demographic events expressed as the ratio $T_{\text{dem}}/Ne_{\text{current}}$). Any change more recent than $0.1Ne_{\text{current}}$ generations ago will be dated with poor precision. Conversely, the age of events older than $0.1Ne_{\text{current}}$ appears more accurately recaptured by our ABC approach.

2.2.2 Two-population models

To estimate the accuracy of DILS in recapturing model parameters, we followed the same procedure based on the analysis of simulated datasets as in the previous paragraph. Note that the two-population models comprise two additional parameters that affect patterns of divergence: the time of split (T_{split}) and the migration rate ($N.m$).

The error rate in the estimation of the parameter Ne_{current} is of the same order of magnitude as in models with a single population (figure 6-A; table S3). However, the imprecision increases with ongoing migration and tends to underestimate Ne_{current}. Estimates of

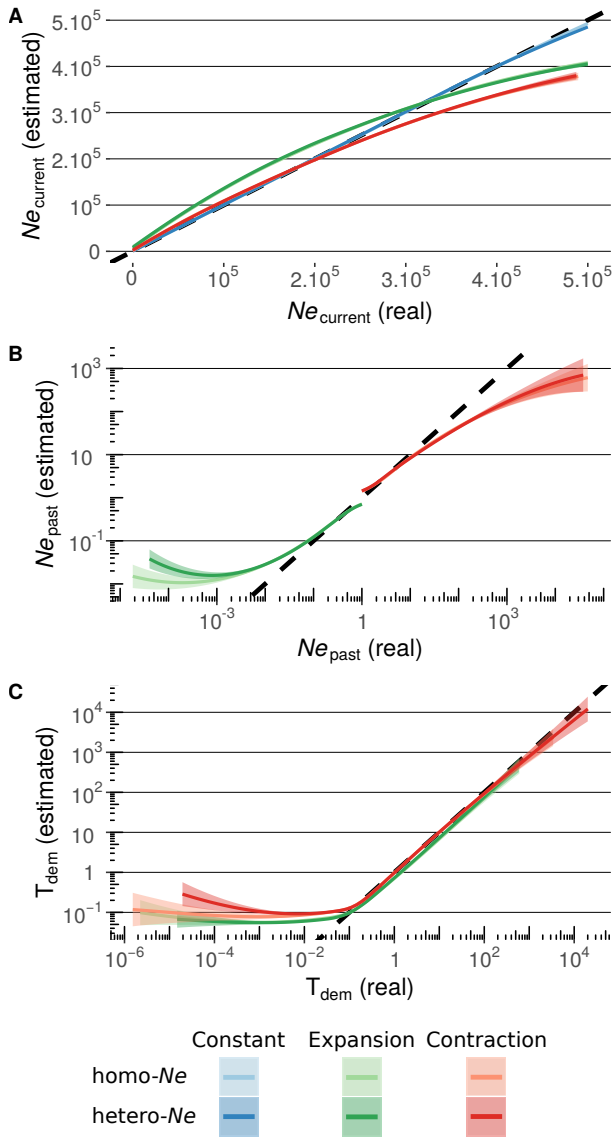

Figure 5. Parameter estimation for single-population models
$10,000$ pseudo-observed datasets are simulated by taking random parameter values (x-axis) under the 6 models. These parameters are estimated using DILS (y-axis). The lines represent the loess (locally estimated scatterplot smoothing) regressions between exact and estimated parameter values for each of the six models. The fields represent the 99% confidence interval and the dotted line represents $x = y$.
Estimation of the effective size of the current population Ne_{current} (**A**), of the ancestral population Ne_{past} (**B**) and the time of demographic changes T_{dem} (**C**). Ne_{past} and T_{dem} are both expressed here in terms of Ne_{current} individuals.

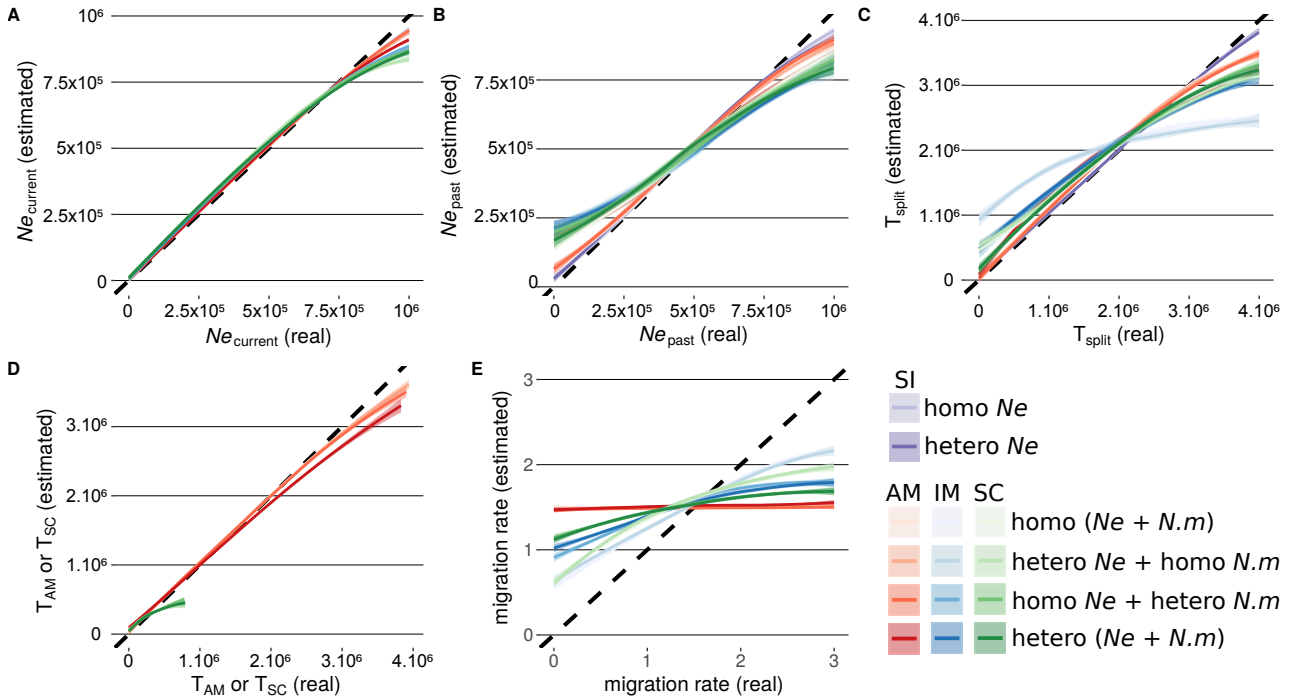

Figure 6. Parameter estimation for two-population models
$2,000$ pseudo-observed datasets are simulated by taking random parameter values under the 14 models and analyzed using the same procedure as to produce the figure 5, but for
the effective size of the current population Ne_{current} (**A**), the ancestral population size Ne_{past} (**B**), the time of split T_{split} (**C**, in generations), the times of demographic transitions T_{AM} and T_{SC} (**D**) and the migration rate *N.m* (**E**).

Ne_{current} are thus more accurate for the SI model, than for the AM model, and the worst for IM and/or SC. This negative effect of ongoing migration on the accuracy of parameter estimation is more pronounced for the ancestral population size Ne_{past} (figure 6-B). Hence, the ongoing migration implemented in the IM and SC models will lead to the overestimation of very low Ne_{past} values and underestimation of large Ne_{past} values.

The precision of the estimate of T_{split} for a given model is of the same order of magnitude as for the ancestral size, with the exception of an accentuated imprecision of T_{split} in the IM model when the migration is homogeneous along the genome (figure 6-C; table S3).

The AM and SC models both have an additional parameter describing the time of the demographic transition between two periods (with and without migration). In the AM model, T_{AM} describes the number of generations during which the two current populations remain genetically isolated after a period of ancestral migration. Conversely, in the SC model, T_{SC} describes the number of generations where the two current populations are connected by gene flow during a secondary contact occurring after a past period of isolation. For the AM model, T_{AM} is better estimated than T_{split} unlike T_{SC} under the SC model (figure 6-D; table S3).

Finally, the performance of DILS to estimate the migration rate $N.m$ (expressed in number of individuals immigrating per generation) is reported on the figure 6-E. The poor estimation accuracy for $N.m$ contrasts sharply with the reliable inferences obtained when comparing the 'ongoing migration' *versus* 'current isolation' supermodels (paragraph 2.1.2). Indeed, it is straightforward to discriminate between these two categories of supermodels while an accurate estimate of the migration rate is more challenging to obtain (figure 6-E). We were unable to reach a reliable measure of $N.m$ for the AM model. More accurate inferences are obtained for both the IM and SC models. Hence, accuracy is reported to increase for genomic models where $N.m$ is homogeneous (table S3).

2.3 Illustration of DILS with RNA-seq data from 28 pairs of *Mytilus* populations

We now illustrate the potential of DILS to contribute to the study of speciation among 8 populations of a complex of four *Mytilus* mussels species (figure 7-A; Bierne et al., 2003; Popovic et al., 2019). The relationship between molecular divergence and genetic isolation was established over 28 pairs of *Mytilus* populations using DILS. The aim was to identify which pairs of populations, characterized by different levels of molecular divergence (net synonymous divergence between 0.003% and 6.705%), are inferred to be cur-

rently connected by ongoing gene flow (figure 7-B). This large scale analysis within the same genus was made possible by the use of a large RNA-seq dataset recently published by Popovic et al., 2019 in 44 individuals of *Mytilus* from 8 populations (one *M. edulis*, five *M. galloprovincialis*, one *M. planulatus* and one *M. trossulus*; figure 7-A).

Out of 28 pairs of *Mytilus* populations that have been tested for ongoing gene flow, 9 pairs receive support for models with current isolation, while models with migration suggest a better fit to the patterns of polymorphism and divergence observed in the remaining 19 pairs (figure 7-B). Within the group composed of *M. galloprovincialis* and *M. planulatus*, the 15 possible pairs are characterized by levels of divergence ranging from $\approx 0.003\%$ (Crique - Herceq) to $\approx 0.673\%$ (Primel - Spring). All of them are supported by models with ongoing gene flow. Our ABC analysis provides support for gene flow for a single interspecific pair with high divergence level $\approx 6\%$ (Darling - Lighthouse). This pair is the only one to be genetically connected by migration among the 8 pairs in our analysis that have a level of net synonymous divergence higher than 2%, most probably due to gene flow between West Atlantic *trossulus* populations and *edulis* ones. Such gene flow also contrasts with an analysis realized for 61 pairs of animal species that had been studied without *a priori* on their speciation history, and for which a threshold of $\approx 2\%$ had emerged above which all interspecific pairs were currently isolated (figure 7-B; Roux et al., 2016).

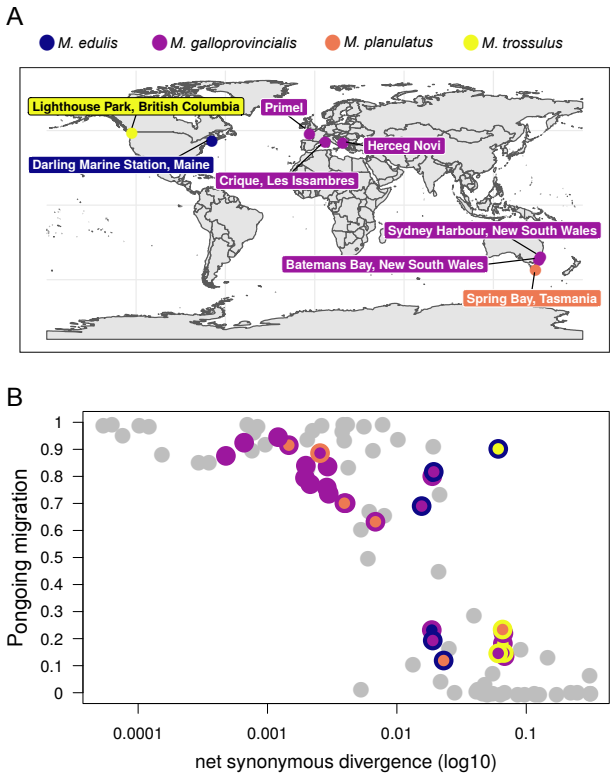

Figure 7. Application of DILS to a *Mytilus* RNA-Seq dataset
A: Transcriptomes were obtained from Popovic et al., 2019 from 44 individuals sampled from 4 labelled species (throughout 8 localities), providing 28 possible pairs of *Mytilus* localities analysed to test for a genetic connection through migration events (SRP218536; `https://cutt.ly/ErrDuwj`). A median number of $1,407$ coding genes was used to perform demographic inferences after filtering the data ($min = 144$ genes; $max = 2,899$ genes; depending on the pair of *Mytilus* considered).
B: The x-axis shows the net divergence measured at the synonymous positions of the sequenced genes.
The y-axis shows the probability provided by DILS of models with ongoing migration (IM + SC).
Grey dots correspond to the 61 pairs of populations/semi-isolated species/species studied in Roux et al., 2016. The coloured dots correspond to the 28 pairs of newly analyzed *Mytilus*. The colours refer to the labelled species from which the partners of each designated pair originate.

3 Discussion

Our statistical analysis platform, DILS, goes beyond simple summary statistics by explicitly testing evolutionary scenarios in a model-based inference framework. In this sense, it is important for users to be mindful that all biological interpretations must be made in light of the models discussed here. This approach is especially time-wise in speciation genomics as comparative studies of closely-related species are accumulating (e.g. in butterflies: Cong, Zhang, and Grishin, 2019; Edelman et al., 2019; Martin et al., 2019; in birds: Peñalba, Joseph, and Moritz, 2019; in fishes: Malinsky et al., 2018; in plants: Stankowski et al., 2019); and so there is a strong demand for efficient and powerful inference tools. Applied to genomic data (SFSs and/or multi-locus sequences), DILS will first identify the best demographic model to test for changes in effective size and migration rate over time, then it will identify the best genomic model to test for genome-wide heterogeneity in these parameters, and finally it will identify loci most associated with genomic regions locked to gene flow.

3.1 Performances of DILS

For single-population models, DILS is highly efficient at distinguishing the three demographic models (Expansion *versus* Constant *versus* Contraction). It fairly discriminates among the two genomic models (homo-*Ne versus* hetero-*Ne*) for the Expansion and Constant models, but has too much ambiguity to distinguish them in a Contraction model. Current population sizes (Ne_{current}) are accurately estimated under all three

demographic models, as well as the time of size change (T_{dem}) provided that it is not too recent. However, the past population size (Ne_{past}) is increasingly overestimated in an expansion model (respectively, underestimated in a contraction model) when the contrast with Ne_{current} is increasingly sharp.

We also found that in two-population models, DILS very accurately discriminates between the two supermodels classically tested in speciation ("current isolation" *versus* "ongoing migration"), and it discriminates reasonably well among models with ongoing migration (IM *versus* SC), but quite poorly among models without (SI *versus* AM). Within each demographic model, the two *Ne*-genomic models are fairly discernible; and the same is true for the two *N.m*-genomic models (homo-*N.m versus* hetero-*N.m*) in scenarios of ongoing migration. Parameters are reasonably well estimated in all models (i.e. population sizes Ne_{current} and Ne_{past}, and times T_{dem}, T_{AM} and T_{SC}); except for the migration rate ($N.m$). It is poorly estimated in ongoing migration models, and cannot be evaluated at all when migration happened in the past (AM).

Therefore it is critical for users to be aware of the limits of DILS; especially when it does accurately discriminate among models and estimate parameter values, and when it does not. For example, the best scenarios for identifying barriers to gene flow is when the genetic signal for these genes strongly contrasts with the rest of the genome, i.e. when speciation time is long enough to build-up divergence at barrier regions, and migration rate is high enough to homogenize the genomic background between species. In general, DILS fails to make accurate inferences when divergence or changes in population size have occurred very recently.

3.2 Collaborative research

DILS was designed with the objective to facilitate collaborative research in speciation. One major question in the field is to understand how fast reproductive isolation builds-up with divergence between lineages, and so how fast introgression decreases along a continuum of molecular divergence. This relationship has been investigated in 61 pairs of animals (Roux et al. 2016) only providing a partial picture. Here, we extended this work by analyzing genomic data of 28 species/populations of $Mytilus$ mussels. Within this specific clade, we found a pattern of non-linear decrease of migration probability with the neutral molecular divergence, similar to what was observed in Roux et al. (2016). However, we also documented ongoing migration between two highly divergent mussel species, hence pushing the grey zone of speciation threshold beyond 2% of net synonymous divergence, maybe due to the outstanding life history traits of mussels (i.e. broadcast spawning, high-dispersal larvae, large effective population sizes and living in a highly connected marine environment).

DILS offers the possibility for the users to participate to this enterprise, and record where their biological model falls within this global speciation picture. Such a global picture of transition from gene flow to no gene flow is necessary for the central problem of species delineation (Hey and Pinho, 2012). Although a universal criterion for delineating species seems impossible, as exemplified by the mussel dataset, the idea of defining a grey zone by taxonomic system is promising (Galtier, 2019). Thus, our collaborative approach option included in DILS will allow in the future to establish a relationship between molecular divergence and genetic isolation for different taxa, *i.e.* vertebrates, terrestrial plants, algae, etc ..., and thus will provide delimitation rules by system.

3.3 Non-detailed features

The raw data can be easily visualized with DILS as a site frequency spectrum and summary statistics across loci. DILS produces comprehensive results for each inference step: (1) the global model comparison to estimate the best demo-genomic model; (2) the locus-specific model comparison to identify barrier loci; and (3) the estimation of parameter values for the best model. To help users interpreting these results, DILS produces a series of goodness-of-fit tests to the data. These tests are performed by simulating under the best model each population genetic statistic calculated in section 4.1.1 (genomic mean and variance of π, θ, F_{ST}, etc.), as well as for each bin of the SFS (or jSFS for two-population models). In addition to an individual test for each summary statistic, a test is also performed from statistics transformed by a PCA following Cornuet et al., 2008; Cornuet et al., 2014. DILS also provides values for each locus of: 1) each summary statistic; 2) the approximated recombination rate calculated based on the four-gamete rule (Galtier et al., 2017; Hudson and Kaplan, 1985); and 3) the posterior probability of being genetically linked to a barrier to gene flow (for two-population models only). These results are outputted as interactive graphics in the application.

Our method is implemented in a user-friendly platform allowing the configuration of the ABC analysis via a graphical interface, its execution and the visualization of the results. Detailed information for how to use DILS is provided in the manual. The released version of DILS is currently hosted by the CC LBBE/PRABI, but to ensure full reproducibility and portability on any server, DILS is packaged in a singularity container freely available at `https://github.com/popgenomics/DILS_web`. The complete analysis of a dataset (model comparison + parameter estimates + locus-specific tests + goodness-of-fit tests) on the host server takes 2h30 on average.

3.4 Prospects for the future

With the improvement of computational methods, it is now possible to simulate entire chromosomes under the

full ancestral process of coalescence and recombination (Kelleher, Etheridge, and McVean, 2016). Combining this type of coalescent simulators with haplotype-based statistics in our ABC framework would be very promising to improve estimates of the timing and extent of gene-flow after secondary contact (Harris and Nielsen, 2013). The architecture of DILS has been designed to easily add simulators other than *ms* and its modified versions (Hudson, 2002). Thus, it would be readily achievable to use forward-in-time simulations including direct selection (Haller and Messer, 2019), and therefore making inferences for any selective scheme while taking into account the demographic history of the sample, without changing the pipeline upstream or downstream of the simulator.

4 Materials and methods

4.1 ABC

4.1.1 Summary statistics

Since ABC is a category of inferential method based on the comparison between statistics summarizing simulated and observed datasets, we first describe here the statistics computed in our framework.

We assume that users are interested in carrying out inferences from multilocus datasets. For single population models, DILS calculates for each locus: *i*) the pairwise nucleotide diversity (π) (Tajima, 1983); *ii*) Watterson's θ (Watterson, 1975) and *iii*) Tajima's *D* (Tajima, 1989). In addition to these three statistics, the site-frequency spectrum (SFS; Fischer, 1930; Wright, 1931; Wright, 1938) is also used to summarize the data, *i.e*, the number of single-nucleotide polymorphism (SNP) where the derived allele is present in $[2, ..., nseq - 1]$ copies in the studied population/species, where $nseq$ represents the number of copies sampled for a given locus. If the SFS is folded by the absence of an outgroup, then the SFS will be described by the number of SNPs where the minor allele is present in $[2, ..., nseq/2 - 1]$ copies. Finally, for single population models, multilocus inferences are based on 6 multilocus summary statistics which are the means and standard deviations of π, θ and Tajima's D, to which we add $[nseq - 2]$ individual statistics corresponding to the SFSs summed over all loci. In absence of an outgroup, the SFS will be represented by $[nseq/2 - 2]$ individual statistics.

For models with two populations/species, π, θ and Tajima's D statistics are also calculated for each of the two samples. These are supplemented with statistics approximating the joint SFS (jSFS; Ramos-Onsins et al., 2004): 1) the fraction of sites showing a fixed difference between the populations/species (Sf), 2) the fraction of sites showing an exclusive polymorphism to a given population/species (Sx_{A} and Sx_{B}) and 3) the fraction of sites with a polymorphism shared between the population/species (Ss). Statistics describing the divergence between the two populations/species are also calculated, including the raw (D_{xy}; Nei, 1987) and the net (D_{a}; Nei and Li, 1979) divergence between the population/species, and their relative genetic differentiation measured by F_{ST} (Wright, 1943). Finally, all bins in the jSFS (except singletons) can optionally be used as an additional vector of summary statistics. If the jSFS is unfolded, then this vector has a length of [($nseq_{\mathrm{A}}$+1) * ($nseq_{\mathrm{B}}$+1) - 4] available statistics (minus 4 to remove the two bins corresponding to singletons and the two bins corresponding to the fixation of the derived or ancestral allele in both samples), and a length of [($nseq_{\mathrm{A}} + 1/2$ * $nseq_{\mathrm{B}} + 1/2$ - 3] if the jSFS is folded. The use of the jSFS as a vector of summary statistics is an option that the user can switch on or off, to avoid cases where the jSFS is composed by a large number of bins.

4.1.2 Simulations

In the current version of DILS, all simulations are performed using the msnsam coalescent simulator (Hudson, 2002; Ross-Ibarra et al., 2008). Each simulated multilocus dataset takes properties from the observed datasets (same number of genes, lengths and sample size). Since the summary statistics used to perform the ABC inferences are averages and standard deviations measured over all the surveyed genes, then for model comparisons and parameter estimations we randomly subsample $1,000$ genes if more loci are present in the total dataset. The purpose of this subsampling is to avoid unnecessarily long simulation times because the values of statistics for a given locus will not impact the used summary statistics over $1,000$ loci.

If an outgroup is specified by the user, then it will be used for each locus (or contig, or gene) to correct its mutation rate μ_i to $\widehat{\mu}.\frac{div_i}{\widehat{div}}$ where $\widehat{\mu}$ is the neutral mutation rate assumed by the user, div_i is the raw divergence between the focal population/species and the outgroup measured at a given locus i, and $\widehat{div}$ is the average raw divergence between the focal population/species and the outgroup measured over all loci. The other implication of using an outgroup will be to orientate the mutations and consequently to unfold the jSFS. Finally, the loci are assumed to be genetically independent, and a $\frac{\widehat{\rho}}{\widehat{\theta}}$ ratio value has to be specified by the user where $\widehat{\rho}$ is the average population recombination rate $4.Ne.r$ (r in number of recombination events per generation and per nucleotide).

4.1.3 Model comparisons

Here, when used alone, the term model means a given combination between a demographic and a genomic model. All comparisons are performed by using the *abcrf* function of the eponymous R package (Pudlo et al., 2015). The comparison is a two-step process.

First, grow the random forest with the *abcrf* function. This requires one reference table per model for

the training. The reference table of each model is produced by $10,000$ multilocus simulations whose parameters correspond to random combinations sampled from priors. They are composed of one row per multilocus simulation and one column for each summary statistic described in section 4.1.1. When categories of models are compared following the hierarchical approaches (figures 2 and 3), the reference tables of the different models in the same category are merged together. For instance, in the comparison between Current isolation and Ongoing migration (figure 3), $60,000$ multilocus simulations are used for the training of the super-model Current isolation, and $80,000$ multilocus simulations for the training of Ongoing migration. Each forest is made up of $1,000$ grown decision trees regardless of the comparison made throughout the hierarchical approach.

The last step is the prediction of the best model among those proposed by passing the observed data through the trained random forest.

DILS reports the model supported by the largest number of decision trees in the random forest and its associated posterior probability.

4.1.4 Parameter estimates

Two strategies are applied simultaneously to estimate the parameters describing the best-supported model among those compared:

1) a joint estimation of the set of parameters using a rejection/regression method (Csilléry, François, and Blum, 2012). Estimation is based on the $5,000$ multilocus simulations producing the statistics closest to the observed dataset among $1,000,000$ simulations. We then correct for imperfect matches between observed and retained values of statistics. The parameter values of the selected simulations are weighted by their Euclidean distance and corrected according to a non-linear regression method using a neural network. 10 trained neural networks with 10 hidden networks are used in the regression.

2) an individual estimation of each parameter by constructing a random forest of $1,000$ trees per parameter (Raynal et al., 2019).

The results from both approaches are returned to the user. There is no evidence to further support a method over the other in terms of estimation accuracy. Within the framework of the models currently compared in DILS, both approaches produce similar estimates when tested on pseudo-observed datasets. However, joint parameter estimation has the advantages of including parameter co-variations as well as providing a probability density. This is achieved at a computational cost that is ≈ 100 times greater regarding the number of multilocus simulations, since $10,000$ are required for a parameter estimation using random forest versus $1,000,000$ when using the rejection/regression algorithm.

4.1.5 Locus-specific model comparison

To identify barriers to gene flow among a set of sequenced DNA fragments (genes for instance), we adopt the same procedure as in *Ciona intestinalis* (Roux et al., 2013) and *Mytilus* (Roux et al., 2014) but by replacing the neural network with a random forest to divide the computational cost by 100. This step is performed by DILS only if 1) observed data better fits models with ongoing migration (IM or SC; figure 1) and 2) genomic models of *N.m.* variation explain the data better than homogeneous models. We first estimate the parameters of the best model from the multilocus dataset. Based on this estimation, two models are compared at each locus: 1) local-migration: the multilocus estimated model with the non-zero migration rate estimated over the whole genome. 2) local-isolation: the multilocus estimated model with a migration rate set to zero. A random forest of $1,000$ trees is then trained to recognize combinations of summary statistics specific to each of the two evaluated models. This forest allows to return for each sequenced DNA fragment the locus-specific model that best explains the statistics observed, and its posterior probability.

4.2 Pseudo-observed datasets

In this study we distinguish two types of simulations. Simulations carried out to build reference tables used to train a random forest, or from which a small proportion will be sub-sampled in a rejection/regression algorithm based on the Euclidean distance with the observed data. Then a second type of simulations produces pseudo-observed datasets. These are not used for training, but to evaluate the inferential power of the ABC approach, and test whether it can recapture the parameters used to simulate the pseudo-observed datasets. To assess the reliability of model comparisons and parameter estimates, for single and two population models, we simulate pseudo-observed datasets consisting of 100 loci, of length equal to $1,000$ nucleotides, sampled from 10 diploid individuals in each population/species and a mutation rate of 5.10^{-8} mutations/nucleotide/generation. These datasets are simulated according to demographic histories using random combinations of parameters from the priors.

4.3 Analysis of the *Mytilus* dataset

We downloaded the raw RNA-seq data deposited to the NCBI sequence read archive (BioProject ID: PRJNA560413; https://cutt.ly/OtQN1Y0) by Popovic et al., 2019. The raw data consist in a total of $\approx 145Gb$ from the sequenced transcriptomes of 47 mussel individuals. Three individuals from the *M. californianus* species were removed as they do not belong to the *M. edulis* complex. The reference transcriptome used for the mapping is made up of $16,151$ CDS, for a total length of $\approx 23Mb$. The reference was indexed using

bowtie2 (version 2.2.6 Langmead and Salzberg, 2012). For each individual, reads were aligned to the reference with bowtie2, and cleaned using samtools with a mapping quality threshold of 20 (version 1.3.1; Li et al., 2009). Individual genotypes were called using reads2snp (Tsagkogeorga, Cahais, and Galtier, 2012) at positions covered by at least 8 reads. We then ran DILS for each of the 28 possible pairs of localities by tolerating up to 20% of missing data, rejecting genes with less that 100 codons without a missing data, and by keeping 6 copies per genes within each population/species. Simulations were conducted by exploring uniform priors for effective population sizes between 0 and $500,000$ diploid individuals, times of different demographic events (split, secondary contact, arrest of migration) between 0 and $1,750,000$ generations. Presentation of the results were carried out with R (Wickham et al., 2019; Chang et al., 2019; Sievert, 2018; R Core Team, 2020).

5 Acknowledgements

We are largely in debt to 'Pôle Rhône-Alpes de Bioinformatique' who kindly offered to host DILS, in particular Bruno Spataro and Stephane Delmotte. We would also like to acknowledge Cynthia Riginos for the approach to make public a large amount of produced data, and the associated metadata facilitating their processing. This project benefited greatly from discussions with Arthur Weyna and Pierre Barry, and the redaction from LaTeX tips by Julien Fraïsse.

6 Resources

Source codes to deploy DILS can be freely used from GitHub `https://github.com/popgenomics/DILS_web`. This GitHub repository also contains the *Mytilus* sequences analyzed (subdirectory 'example'). The corresponding raw data ara available under the accession number SRP218536 (`https://cutt.ly/ErrDuwj`).

Bibliography

Beaumont, Mark A, Wenyang Zhang, and David J Balding (2002). "Approximate Bayesian computation in population genetics." In: *Genetics* 162.4, pp. 2025–2035.

Bierne et al. (2003). "Introgression patterns in the mosaic hybrid zone between Mytilus edulis and M. galloprovincialis." In: *Mol. Ecol.* 12.2, pp. 447–461.

Bunnefeld, Lynsey et al. (2018). "Whole-genome data reveal the complex history of a diverse ecological community". In: *Proceedings of the National Academy of Sciences* 115.28, E6507–E6515.

Chang, Winston et al. (2019). *shiny: Web Application Framework for R*. R package version 1.4.0. URL: `https://CRAN.R-project.org/package=shiny`.

Cong, Qian, Jing Zhang, and Nick Grishin (2019). "Genomic determinants of speciation". In: *bioRxiv*, p. 837666.

Cornuet, Jean-Marie et al. (2008). "Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation". In: *Bioinformatics* 24.23, pp. 2713–2719.

Cornuet, Jean-Marie et al. (2014). "DIYABC v2. 0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data". In: *Bioinformatics* 30.8, pp. 1187–1189.

Csilléry, Katalin, Olivier François, and Michael G B Blum (June 2012). "abc: an R package for approximate Bayesian computation (ABC)". In: *Methods Ecol. Evol.* 3.3, pp. 475–479.

Edelman, Nathaniel B et al. (2019). "Genomic architecture and introgression shape a butterfly radiation". In: *Science* 366.6465, pp. 594–599.

Excoffier, Laurent et al. (2013). "Robust demographic inference from genomic and SNP data". In: *PLoS genetics* 9.10.

Fischer, RA (1930). *The genetical theory of natural selection. Clarendon*.

Fraimout, Antoine et al. (2017). "Deciphering the routes of invasion of Drosophila suzukii by means of ABC random forest". In: *Molecular biology and evolution* 34.4, pp. 980–996.

Galtier, Nicolas (2019). "Delineating species in the speciation continuum: A proposal". In: *Evolutionary applications* 12.4, pp. 657–663.

Galtier, Nicolas et al. (2017). "Codon usage bias in animals: disentangling the effects of natural selection, effective population size and GC-biased gene conversion". en.

Gaut, Brandon S et al. (2018). "Demography and its effects on genomic variation in crop domestication". In: *Nature plants* 4.8, pp. 512–520.

Gutenkunst, Ryan N et al. (Oct. 2009). "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data." In: *PLoS Genet.* 5.10. Ed. by Gil McVean, e1000695.

Haller, Benjamin C and Philipp W Messer (2019). "SLiM 3: forward genetic simulations beyond the Wright–Fisher model". In: *Molecular biology and evolution* 36.3, pp. 632–637.

Harris, Kelley and Rasmus Nielsen (June 2013). "Inferring demographic history from a spectrum of shared haplotype lengths." In: *PLoS Genet.* 9.6. Ed. by Jeffery D Jensen, e1003521.

Hey, Jody and Catarina Pinho (2012). "Population genetics and objectivity in species diagnosis." In: *Evolution* 66.5, pp. 1413–1429.

Hudson, Richard R (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation." In: *Bioinformatics* 18.2, pp. 337–338.

Hudson, Richard R and Norman L Kaplan (1985). "Statistical properties of the number of recombination events in the history of a sample of DNA sequences". In: *Genetics* 111.1, pp. 147–164.

Jouganous, Julien et al. (2017). "Inferring the joint demographic history of multiple populations: beyond the diffusion approximation". In: *Genetics* 206.3, pp. 1549–1567.

Kelleher, Jerome, Alison M Etheridge, and Gilean McVean (2016). "Efficient coalescent simulation and genealogical analysis for large sample sizes". In: *PLoS computational biology* 12.5.

Köster, Johannes and Sven Rahmann (2012). "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19, pp. 2520–2522.

Langmead, Ben and Steven L Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4, p. 357.

Li, Heng et al. (2009). "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079.

Lohse, Konrad (2017). "Come on feel the noise—from metaphors to null models". In: *J. Evol. Biol* 30, pp. 1506–1508.

Lohse, Konrad, Richard J Harrison, and Nicholas H Barton (2011). "A general method for calculating likelihoods under the coalescent process". In: *Genetics* 189.3, pp. 977–987.

Malinsky, Milan et al. (2018). "Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow". In: *Nature ecology & evolution* 2.12, pp. 1940–1955.

Martin, Simon H et al. (2019). "Recombination rate variation shapes barriers to introgression across butterfly genomes". In: *PLoS Biology* 17.2, e2006288.

Momigliano, Paolo, Ann-Britt Florin, and Juha Merilä (2020). "Biases in demographic modelling affect our understanding of the process of speciation". In: *bioRxiv*.

Nei, M (1987). "Molecular Evolutionary Genetics Columbia University Press New York 512". In:

Nei, M and W H Li (1979). "Mathematical model for studying genetic variation in terms of restriction endonucleases". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 76.10, pp. 5269–5273.

Peñalba, Joshua V, Leo Joseph, and Craig Moritz (2019). "Current geography masks dynamic history of gene flow during speciation in northern Australian birds". In: *Molecular ecology* 28.3, pp. 630–643.

Popovic, Iva et al. (2019). "Twin introductions by independent invader mussel lineages are both associated with recent admixture with a native congener in Australia". In: *Evolutionary Applications*.

Pudlo, Pierre et al. (2015). "Reliable ABC model choice via random forests". In: *Bioinformatics* 32.6, pp. 859–866.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramos-Onsins, Sebastián E et al. (Jan. 2004). "Multilocus analysis of variation and speciation in the closely related species Arabidopsis halleri and A. lyrata." In: *Genetics* 166.1, pp. 373–388.

Raynal, Louis et al. (2019). "ABC random forests for Bayesian parameter inference". In: *Bioinformatics* 35.10, pp. 1720–1728.

Robinson, John D et al. (2014). "ABC inference of multi-population divergence with admixture from unphased population genomic data". In: *Molecular ecology* 23.18, pp. 4458–4471.

Ross-Ibarra, Jeffrey et al. (2008). "Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata". In: *PloS one* 3.6.

Roux, C et al. (2014). "Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a Mytilus hybrid zone." In: *J. Evol. Biol.* 27.8, pp. 1662–1675.

Roux, Camille et al. (2013). "Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species." In: *Mol. Biol. Evol.* 30.7, pp. 1574–1587.

Roux, Camille et al. (2016). "Shedding light on the grey zone of speciation along a continuum of genomic divergence". In: *PLoS biology* 14.12, e2000234.

Sethuraman, Arun, Vitor Sousa, and Jody Hey (2019). "Model-based assessments of differential introgression and linked natural selection during divergence and speciation". In: *bioRxiv*, p. 786038.

Sievert, Carson (2018). *plotly for R*. URL: https://plotly-r.com.

Smith, Chris CR and Samuel M Flaxman (2020). "Leveraging whole genome sequencing data for demographic inference with approximate Bayesian computation". In: *Molecular ecology resources*.

Sousa, Vitor and Jody Hey (2013). "Understanding the origin of species with genome-scale data: modelling gene flow". In: *Nature Reviews Genetics* 14.6, pp. 404–414.

Sousa, Vitor C et al. (2013). "Identifying loci under selection against gene flow in isolation-with-migration models". en. In: *Genetics* 194.1, pp. 211–233.

Stankowski, Sean et al. (2019). "Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers". In: *PLoS biology* 17.7, e3000391.

Tajima, F (1983). "Evolutionary relationship of DNA sequences in finite populations." In: *Genetics* 105.2, pp. 437–460.

– (1989). "The effect of change in population size on DNA polymorphism." In: *Genetics* 123.3, pp. 597–601.

Terhorst, Jonathan, John A Kamm, and Yun S Song (2017). "Robust and scalable inference of population

history from hundreds of unphased whole genomes". In: *Nature genetics* 49.2, p. 303.

Terhorst, Jonathan and Yun S Song (2015). "Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum". In: *Proceedings of the National Academy of Sciences* 112.25, pp. 7677–7682.

Tsagkogeorga, Georgia, Vincent Cahais, and Nicolas Galtier (2012). "The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis." In: *Genome Biol. Evol.* 4.8, pp. 740–749.

Watterson, G A (1975). "On the number of segregating sites in genetical models without recombination." In: *Theor. Popul. Biol.* 7.2, pp. 256–276.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Wright, Sewall (1931). "Evolution in Mendelian populations". In: *Genetics* 16.2, p. 97.

– (1938). "The distribution of gene frequencies under irreversible mutation". In: *Proceedings of the National Academy of Sciences of the United States of America* 24.7, p. 253.

– (1943). "Isolation by distance". In: *Genetics* 28.2, p. 114.

7 Supplementary material

Supplementary Table 1. Performance of DILS model comparisons

comparison	target	correct	ambiguous	wrong	threshold
Single-population models					
Expansion *versus* Constant *versus* Contraction	Expansion	0.8627	0.1275	0.0098	0.919
	Constant	0.891	0.099	0.01	0.856
	Contraction	0.9937	0.0024	0.0039	0.501
Expansion homo-Ne *versus* hetero-Ne	homo-Ne	0.8248	0.1652	0.01	0.83
	hetero-Ne	0.7492	0.2408	0.01	0.92
Constant homo-Ne *versus* hetero-Ne	homo-Ne	0.8562	0.134	0.0098	0.835
	hetero-Ne	0.8336	0.1564	0.01	0.89
Contraction homo-Ne *versus* hetero-Ne	homo-Ne	0.4074	0.5828	0.0098	0.906
	hetero-Ne	0.2634	0.7266	0.01	0.901
Two-populations models					
migration *versus* isolation	migration	0.97922	0.01078	0.01	0.665
	isolation	0.95053	0.03968	0.00978	0.84
SI *versus* AM	SI	0.01465	0.9754	0.00995	0.845
	AM	0.81482	0.18518	0	0.705
IM *versus* SC	IM	0.791	0.19908	0.00992	0.885
	SC	0.69958	0.29042	0.01	0.915
SI homo-Ne *versus* SI hetero-Ne	homo-Ne	0.9688	0.0213	0.0099	0.8
	hetero-Ne	0.8173	0.1741	0.0086	0.96
AM homo-Ne *versus* AM hetero-Ne	homo-Ne	0.95495	0.0355	0.00955	0.82
	hetero-Ne	0.8156	0.17445	0.00995	0.95
IM homo-Ne *versus* IM hetero-Ne	homo-Ne	0.7006	0.29005	0.00935	0.855
	hetero-Ne	0.6469	0.3433	0.0098	0.93
SC homo-Ne *versus* SC hetero-Ne	homo-Ne	0.7063	0.284	0.0097	0.85
	hetero-Ne	0.64645	0.34445	0.0091	0.925
IM homo-N.m *versus* IM hetero-N.m	homo-N.m	0.85915	0.13105	0.0098	0.845
	hetero-N.m	0.759	0.23265	0.00835	0.95
SC homo-N.m *versus* SC hetero-N.m	homo-N.m	0.7144	0.27585	0.00975	0.855
	hetero-N.m	0.59875	0.39165	0.0096	0.915

For each model, $10,000$ sets of pseudo-observed data were analysed. These datasets were simulated by taking random combinations of parameters from large prior distributions. The table reports for each model comparison the proportions among these simulations that lead to correct, ambiguous or wrong inferences according to a threshold set to keep the rate of wrong inferences below 1%.

Supplementary Table 2. Mean-squared error in parameter estimations for single-population models

demographic model	genomic model Ne	Ne (current)	Ne (past)	α	β	T_{dem}
Constant	homo	0.00523				
	hetero	0.01178		0.39888	0.92462	
Expansion	homo	0.2009	0.00922			0.74168
	hetero	0.20754	0.01369	0.46283	0.94578	0.74924
Contraction	homo	0.12258	0.08716			0.88229
	hetero	0.12563	0.09423	0.52025	0.95338	0.90631

Estimation errors are calculated on $N = 10,000$ simulated datasets for each model, by using random combination of parameter values. The reported values are measured as follows

$$\frac{1}{N * variance(\theta_i)} \sum_{i=1}^{N} (\widehat{\theta_i} - \theta_i)^2,$$

where $\widehat{\theta_i}$ and θ_i represent the estimated and the true parameter values respectively

Supplementary Table 3. Mean-squared error in parameter estimations for two-population models

demographic model	genomic model $N.m$	genomic model Ne	Ne (current)	Ne (past)	T_{split}	T_{AM}	T_{SC}	$N.m$	number of barriers	α	β
AM	homo	homo	0.02114	0.09633	0.08024	0.02503		1.01278			
		hetero	0.02832	0.10649	0.09502	0.02797		1.0092		0.39736	0.96193
	hetero	homo	0.02134	0.09238	0.08401	0.04687		1.02054	1.01096		
		hetero	0.03793	0.21534	0.22518	0.07039		1.00063	0.94219	0.46799	0.97759
IM	homo	homo	0.08547	0.30056	0.60409			0.37401			
		hetero	0.09235	0.35789	0.62885			0.40637		0.452	0.97489
	hetero	homo	0.05707	0.31596	0.27658			0.6166	0.52701		
		hetero	0.07003	0.29829	0.27734			0.65794	0.55112	0.55623	0.97603
SC	homo	homo	0.08764	0.23384	0.19008		0.33885	0.50572			
		hetero	0.09052	0.26734	0.21924		0.35622	0.50668		0.46972	0.9829
	hetero	homo	0.07016	0.25806	0.14854		0.30055	0.78375	0.49972		
		hetero	0.07937	0.26373	0.15774		0.32815	0.77411	0.5055	0.49075	0.9738
SI		homo	0.0183	0.04164	0.01158						
		hetero	0.02918	0.05768	0.01458					0.38479	0.95003

Estimation errors are calculated on $N = 2,000$ simulated datasets for each model, by using random combination of parameter values. The reported values are measured as follows

$$\frac{1}{N * variance(\theta_i)} \sum_{i=1}^{N} (\widehat{\theta_i} - \theta_i)^2,$$

where $\widehat{\theta_i}$ and θ_i represent the estimated and the true parameter values respectively