RIDGE, a tool tailored to detect gene flow barriers across species pairs

Ewen Burban, M. I. Tenaillon, Sylvain Glémin

Received: 15 September 2023 | Revised: 19 January 2024 | Accepted: 5 February 2024

DOI: 10.1111/1755-0998.13944

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

RIDGE, a tool tailored to detect gene flow barriers across species pairs

Ewen Burban[1] | Maud I. Tenaillon[2] | Sylvain Glémin[1,3]

[1]University of Rennes, CNRS, ECOBIO-UMR 6553, Rennes, France

[2]University Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE-Le Moulon, Gif-sur-Yvette, France

[3]Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Correspondence
Maud I. Tenaillon, University Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE-Le Moulon, Gif-sur-Yvette, France.
Email: maud.tenaillon@inrae.fr

Sylvain Glémin, University of Rennes, CNRS, ECOBIO-UMR 6553, Rennes, France.
Email: sylvain.glemin@univ-rennes.fr

Funding information
Région Bretagne; Agence Nationale de la Recherche, Grant/Award Number: ANR-17-EUR-0007 and ANR-19-CE32-0009-02; Centre National de la Recherche Scientifique, Grant/Award Number: GDR 3765

Handling Editor: Michael M. Hansen

Abstract

Characterizing the processes underlying reproductive isolation between diverging lineages is central to understanding speciation. Here, we present RIDGE—Reproductive Isolation Detection using Genomic polymorphisms—a tool tailored for quantifying gene flow barrier proportion and identifying the relevant genomic regions. RIDGE relies on an Approximate Bayesian Computation with a model-averaging approach to accommodate diverse scenarios of lineage divergence. It captures heterogeneity in effective migration rate along the genome while accounting for variation in linked selection and recombination. The barrier detection test relies on numerous summary statistics to compute a Bayes factor, offering a robust statistical framework that facilitates cross-species comparisons. Simulations revealed RIDGE's efficiency in capturing signals of ongoing migration. Model averaging proved particularly valuable in scenarios of high model uncertainty where no migration or migration homogeneity can be wrongly assumed, typically for recent divergence times <0.1 $2N_e$ generations. Applying RIDGE to four published crow data sets, we first validated our tool by identifying a well-known large genomic region associated with mate choice patterns. Second, while we identified a significant overlap of outlier loci using RIDGE and traditional genomic scans, our results suggest that a substantial portion of previously identified outliers are likely false positives. Outlier detection relies on allele differentiation, relative measures of divergence and the count of shared polymorphisms and fixed differences. Our analyses also highlight the value of incorporating multiple summary statistics including our newly developed outlier ones that can be useful in challenging detection conditions.

KEYWORDS
approximate Bayesian computation, crows, gene flow barrier detection, hybrid zones, reproductive isolation, speciation

Maud I. Tenaillon and Sylvain Glémin equally contributed to this study.

1 | INTRODUCTION

The process of speciation involves a gradual and divergent evolution of populations, passing through conditions of semi-isolated species, named the 'grey zone of speciation' (De Queiroz, 2007; Roux et al., 2016), until complete genetic isolation is achieved, resulting in the formation of distinct species (Wu, 2001). Population divergence can occur through various scenarios, ranging from the complete absence of genetic exchanges, known as allopatric speciation (e.g. due to geographical barriers between populations), to almost unrestricted genetic exchanges in sympatric speciation. These extreme scenarios are not mutually exclusive, as genetic exchanges can reoccur after a period of allopatric divergence followed by secondary contacts (Schluter, 2001). Regardless of the scenario, the question of how reproductive isolation is established between divergent populations is central to understanding speciation. This involves comparing the proportion and identity of the relevant genomic regions across biological systems (Delmore et al., 2018; Fraïsse et al., 2021; Schluter, 2001).

Extensive exploration of the genomic bases of speciation have been conducted, in particular, in the case of ecological speciation where environmental disparities among populations drive both phenotypic divergence and reproductive isolation (Rundle & Nosil, 2005; Schluter, 2000; Shafer & Wolf, 2013). A recurrently observed pattern is that pre-mating reproductive isolation is facilitated by the physical linkage between genes that govern reproductive isolation and those responsible for divergent traits, which can potentially result from adaptation to contrasted environmental conditions. The gradual establishment of linkage disequilibrium between these genes can then lead to the progressive arrest of gene flow during the speciation process (Schluter & Rieseberg, 2022).

For example, in stickleback fish, divergent mate preferences have been mapped to the same set of genomic regions controlling body size, shape and ecological niche utilization (Bay et al., 2017). Another striking example concerns the genomic determinants of mate selection based on feather colour patterns in carrion and hooded crows (Metzler et al., 2021; Poelstra et al., 2014). Specifically, genes encoding feather pigmentation and genes responsible for perceiving colour patterns have been identified within the same 1.95 Mb region of chromosome 18. This region displays significant genetic differentiation between carrion and hooded crows. Similarly, in the neotropical butterflies *Heliconius cydno* and *H. melpomene*, assortative mating patterns correlate with a genomic region proximate to *optix*, a crucial locus influencing distinct wing colour patterns between these species (Merrill et al., 2019). Note that, inversions can help build linkage disequilibrium by generating large genomic regions of suppressed recombination, maintaining combinations of co-adapted alleles encoding ecologically relevant traits. For example, in three species of wild sunflowers, 37 large non-recombining haplotype blocks (1–100 Mbp in size) contribute to strong prezygotic isolation between ecotypes through multiple traits such as soil, climate and flowering characteristics (Todesco et al., 2020).

Another key genetic mechanism involved in speciation is the epistatic interaction between genes that produce deleterious phenotypes in hybridization, also known as Bateson–Dobzhansky–Muller Incompatibility (BDMI) (Gavrilets, 2003). Across *Arabidopsis thaliana* strains, epistatic interactions between alleles from two loci located on separate chromosomes, result in an autoimmune-like responses in F1 hybrids (Bomblies et al., 2007). A more recent example in vertebrates concerns the Swordtail fish species, *Xiphophorus birchmanni* and *X. malinche*, where interaction between two genes generates a malignant melanoma in hybrids associated with strong viability selection (Powell et al., 2020).

As population-wide genomic data increase, genome-scan approaches enable a more systematic search of the genetic factors behind reproductive isolation. One popular approach relies on the search for genomic islands of elevated differentiation compared with the genomic background, typically through F_{ST} scans (Wolf & Ellegren, 2017). However, it is now widely recognized that processes other than selection against gene flow can generate such islands. For example, selective sweeps and background selection against deleterious alleles both decrease genetic diversity at linked sites especially in low recombination regions (Charlesworth, 1993; Charlesworth & Jensen, 2021; Cruickshank & Hahn, 2014; Kaplan et al., 1989). Because gene flow barriers are more likely to occur in functional regions, they are also more affected by those forms of selection, further complicating the distinction of gene flow reduction (Ravinet et al., 2017). Demography, which affects the entirety of the genome, is also key to account for barrier detection because barrier loci are harder to identify when the time split is recent and/or the migration rate is low (Sakamoto & Innan, 2019). Yet, recent splits of partially isolated taxa are of paramount interest in speciation research as they allow access to the key determinants of reproductive isolation while avoiding the confusion with other differences accumulated since speciation (Tenaillon et al., 2023).

Linked selection (at least some forms of) can be approximated by a local reduction in effective population size (Cruickshank & Hahn, 2014; Ravinet et al., 2017; Sakamoto & Innan, 2019) and several methods have proposed to decouple its effect from the heterogeneity in effective migration rate to detect gene flow barrier on genomic polymorphism patterns (Fraïsse et al., 2021; Laetsch et al., 2023; Sethuraman et al., 2019; Sousa et al., 2013). These methods relax the assumption that all loci share the same demography. Some of them use likelihood methods to directly estimate and decouple the effects of differential introgression and demography across genomic loci (Laetsch et al., 2023; Sethuraman et al., 2019; Sousa et al., 2013). However, they make specific assumptions about demography. For example, gIMble simulates population divergence under isolation with migration (IM) only, thereby considering no variation in migration rate through time (Laetsch et al., 2023). DILS proposes a more flexible approximate Bayesian computation (ABC) approach. First, it infers the best demographic models among four models that include migration rate variation through time while accounting for genomic heterogeneity in effective population size N_e (to mimic linked selection) and in

effective migration m_e (to mimic gene flow barriers). Such account of genomic heterogeneity has been shown to enhance the quality of model inferences (Roux et al., 2014). Second, DILS infers the migration model at the locus scale—arrest of migration versus migration similar to the genome-wide level–conditioned on the chosen best model (Fraïsse et al., 2021). Although effective in detecting gene flow barrier, this reliance on an initial model choice restricts comparability among species pairs.

Overall, an adequate method to identify potential reproductive isolation barriers would require a cross-species comparative framework that takes genomic heterogeneity into account, while making analysis comparable despite differences in demographic histories. Here, we propose an innovative method to identify gene flow barrier loci satisfying these requirements and that also quantifies the confidence in locus detection. We used an ABC-based model averaging approach that accounts for different modalities of divergence between pairs of populations/taxons. We considered both heterogeneity in N_e along the genome, by modelling the mosaic effect of linked selection as in the DILS program (Fraïsse et al., 2021), and heterogeneity in recombination, by including an option for the user to provide a recombination map. In addition, we not only relied on a number of classic summary statistics but also incorporated new ones, related to outlier detection, which improved the inferences of barrier loci. Finally, the method provides Bayes factors associated with barrier detection, which facilitate cross-species comparisons.

2 | MATERIALS AND METHODS

2.1 | RIDGE pipeline

RIDGE utilizes ABC based on random forest (RF) to detect barrier loci between two diverging populations in the line of the framework proposed in DILS (Fraïsse et al., 2021). The observed data consist of a set of loci sequenced on several individuals of the two populations. The general principle of RIDGE is as follows: first, we simulate 14 demographic × genomic models to produce a *reference table*. This table serves to train one RF per parameter that generates corresponding estimate of each parameter in addition to providing weights for each model according to their fit to the target (observed) data set. Second, we construct a hypermodel where the posterior distribution of each parameter is obtained as the weighted average over the 14 models. Finally, we use this hypermodel to produce data sets for control loci (thereafter non-barrier) and barrier loci that have undergone no gene flow during divergence. Simulated data sets are employed to train a second RF model that subsequently calculates posterior probabilities and associated Bayes factors for categorizing each locus as barrier or non-barrier. RIDGE was executed using Snakemake (v7.7.0) with Singularity as the container manager. Data visualization was conducted using R v4.1.2 (R Core Team, 2021) and involved the utilization of the following packages: ggpubr (Kassambara, 2020), scales (Wickham, 2018), FactoMineR (Le et al., 2008), factoextra (Kassambara & Mundt, 2017) and latex2exp (Meschiari 2023).

2.2 | ABC summary statistics

ABC inferences rely on summary statistics that are computed either at the locus-level or across loci, that is, genome-wide distributions of summary statistics and correlations among loci, and either within- or between-populations. For a given observed data set, the number of loci used for construction of the hypermodel is set by the user. To reduce computation time for large data sets, a subset of loci can be randomly sampled to represent the whole genome (by default, we used 1000 loci).

For each locus, RIDGE computes the following within population statistics: the number of Single Nucleotide Polymorphisms—SNPs (S), π (Nei & Li, 1979), Watterson θ (Watterson, 1975), as well as Tajima's D (Tajima, 1989). As measures of population differentiation between populations, RIDGE computes F_{ST} (Bhatia et al., 2013; Hudson et al., 1992), the absolute (*Dxy*) and the net (*Da*) divergence (Nei & Li, 1979), the summary of the joint Site Frequency Spectrum (jSFS) (Wakeley & Hey, 1997) with *ss* (the proportion of shared polymorphisms between populations), *sf* (the proportion of fixed differences between populations), *sxA* and *sxB* (the proportion of exclusive polymorphisms to each population).

Across loci, RIDGE computes the mean, the median and the standard deviation for each summary statistic described above. In addition, RIDGE computes the Pearson correlation coefficient between *Dxy* and F_{ST} and between *Da* and F_{ST}. Regarding the jSFS, RIDGE determines the number of loci that contains both shared polymorphisms ($ss > 0$) and fixed differences ($sf > 0$) between populations, $ss^{+}sf^{+}$ and following the same rational $ss^{+}sf^{-}$, $ss^{-}sf^{+}$, $ss^{-}sf^{-}$. These statistics are commonly used in ABC to simplify the jSFS while keeping the most relevant information (e.g. in DILS, Fraïsse et al., 2021). To obtain better insights into the proportion of barriers, we introduced new statistics: the proportion of outlier loci, defined as the proportion of loci that exceeds certain thresholds for F_{ST}, *Dxy*, *sf*, *Da* and *ss* or falling below certain thresholds for π and θ. The thresholds are determined using Tukey's fences: $t_{min} = Q_{min} - 1.5 * (Q_{max} - Q_{min})$ and $t_{max} = Q_{max} + 1.5 * (Q_{max} - Q_{min})$, for the lower and upper thresholds respectively, where Q_{min} is the lowest and Q_{max} the highest quartiles (Tukey, 1977). All summary statistics are computed using the python packages scikit-allel (Miles et al., 2021) and numpy (Harris et al., 2020).

2.3 | Coalescence simulations

We simulated the evolution of neutral loci (1000 by default) under 14 demographic × genomic models using the scrm simulator (Staab et al., 2015), an efficient ms-like program (Hudson, 2002). We stored corresponding simulation parameters as well as all summary statistics in the *reference table*.

2.3.1 | Demographic models

RIDGE simulates the split of a single ancestral population of effective size N_a, into two daughter populations of size N_1 and N_2 at time T_{split}. Four different demographic models are considered as in DILS (Fraïsse et al., 2021) (Figure 1): (1) strict isolation with no migration (SI), (2) isolation with constant migration rate since T_{split} (IM), (3) secondary contact with no migration after the split until a secondary contact at time T_{SC} occurs (SC), and (4) ancestral migration with migration occurring initially and ceasing after time T_{AM} (AM). Migration rate m is assumed to be symmetrical between the two populations.

2.3.2 | Genomic models

In addition to modelling demography, RIDGE also incorporates heterogeneity in effective population size along the genome generated by linked selection, and heterogeneity in effective migration generated by selection against migrants at barrier loci. Thus, demographic models are combined with two effective population size modalities (homo-*N* vs. hetero-*N*) and with two migration rate modalities (homo-*m* vs. hetero-*m*), so that four genomic models are considered, except for the SI model where there is no migration and only two genomic models (homo-*N* and hetero-*N*). This gives 14 demographic x genomic models. For simplicity, genomic models are named using a combination of *1N* (homo-*N*), *2N* (hetero-*N*), *1m* (homo-*m*), *2m* (hetero-*m*). While in the *1N* modality all loci display the same effective population size genome-wide, heterogeneity of effective population size under *2N*, is modelled by a rescaled Beta distribution. Effective size at locus *i* is given by:

$$N_i = \overline{N} \cdot \left(\frac{\alpha + \beta}{\alpha} \right) \cdot B(\alpha, \beta) \quad (1)$$

where B(α,β) is a Beta distribution with parameter α and β and $\overline{N}$ is the mean effective population size across the genome. In other words, under *2N* and for a given locus, three independent values are sampled

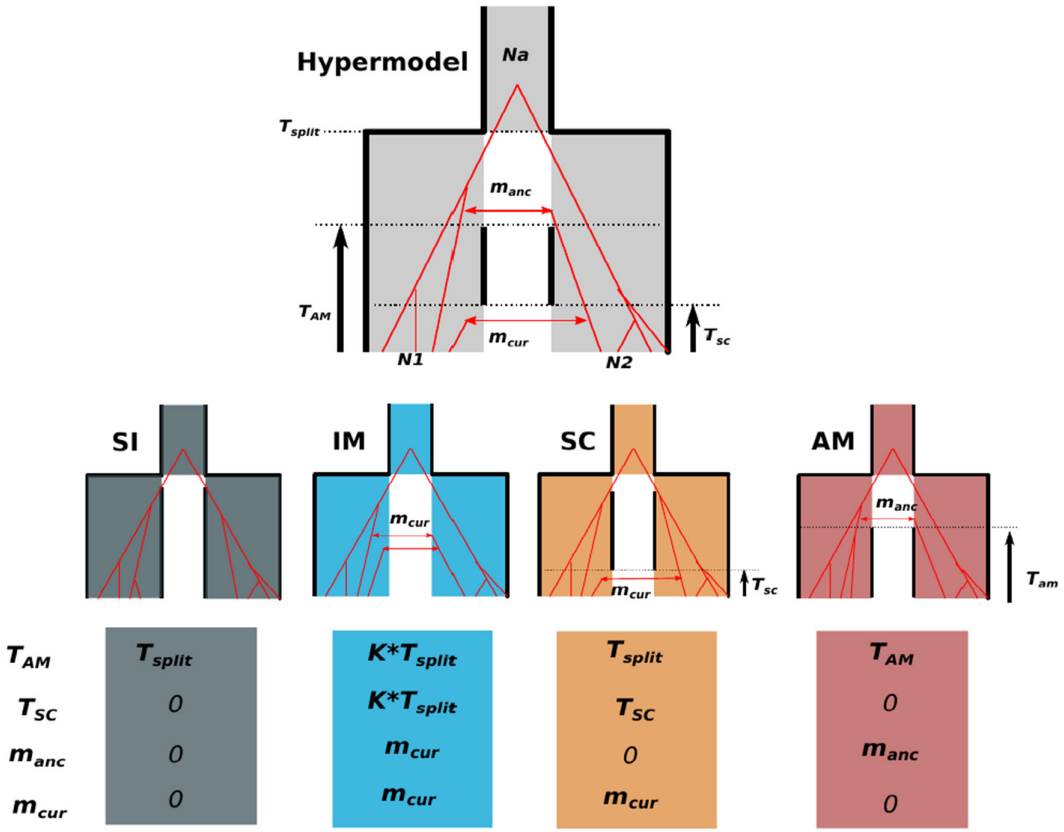

FIGURE 1 Demographic models implemented in RIDGE. The hypermodel combines all four demographic models considered: Strict Isolation (SI), Ancestral Migration (AM), Secondary contacts (SC) and Isolation-Migration (IM) plus genomic models. In the hypermodel, an ancestral population of effective size N_a split at T_{split} in two populations of effective size N_1 and N_2. At T_{AM} ancestral migration ceases, and it restarts at the time of secondary contact,T_{SC}. m_{anc} and m_{cur}denote the ancestral and current migration rates between populations respectively. To fit in the hypermodel, each of the four demographic models adopt specific values for four of the parameters as indicated below each graph. For example, under SI, T_{AM} is set to T_{split} as there is no ancestral migration, and T_{SC} is set to 0 as there is no secondary contact, and so are m_{anc} and m_{cur}. Note that under IM, in order to model uninterrupted gene flow, we considered $T_{AM} = T_{SC} = K * T_{split}$, where K is a random value drawn from a uniform distribution in [0,1]. These demographic models are then combined with four genomic models: homogenous or heterogenous N_e (*1N*, *2N*) and homogeneous and heterogenous m (*1m*, *2m*). For the SI model there are only two possible genomic models (*1N* or *2N*) because there is no migration. This yields 14 models in total.

from the same B(α,β) distribution albeit distinct $\overline{N}$ are used in equation 1 so that there is no covariation of the effective population size across populations. For migration (m), the genome-wide heterogeneity in effective migration is modelled by a Bernoulli distribution where a proportion Q of loci displays $m = 0$ and a proportion $1 - Q$ loci displays $m > 0$, m corresponding either to current migration (m_{cur}) or to ancestral migration (m_{anc}). Likewise, we referred to the proportion of barriers under current (Q_{cur}) and ancestral (Q_{anc}) migration. It is important to note that coalescent simulations use the scaled parameter $M=4N_em$, and that M (rather than m) is the standard way to report migration rate. Variable M across the genome can thus be due to variation in N_e alone, m alone or both. For example, $2N$ and $1m$ models, M is variable across the genome but its variation parallels variation in N_e. This approach differs from the one implemented in DILS, where N_e can be variable but M fixed, which implicitly implies that m is proportional to $1/N_e$ and can thus over-detect heterogeneity in m. Also note that under *2N2m* models, variations in N_e and m are assumed to be independent. RIDGE assumes that all loci are independent and experience a genome-wide homogeneous mutation rate (μ, set by the user) and recombination rate (r, set by the user) unless a recombination map is provided, in which case locus-specific recombination rates are given by the recombination map.

2.4 | Generation of the reference table

RIDGE explores 14 demographic×genomic models of divergence using a hypermodel that integrates them all. This model contains 12 parameters, eight demographic parameters ($N_a, N_1, N_2, T_{split}, T_{AM}, T_{SC}, m_{cur}, m_{anc}$) as described in Figure 1, and four genomic parameters ($\alpha, \beta, Q_{cur}, Q_{anc}$). Regarding the demographic parameters, population sizes (N_a, N_1, N_2) and times ($T_{split}, T_{AM}, T_{SC}$) are sampled in uniform distributions with boundaries specified by the user. Migration rates are drawn from a truncated log-uniform distribution, with the boundary also specified by the user. We used log-normal instead of uniform distributions as migration affects most statistics in a non-linear, multiplicative way. Preliminary simulations showed that it improved the performance of migration estimation. Note that, depending on the considered demographic model, some of the parameters are set to 0 (Table S1, Figure 1). For example, under SI, only four demographic parameters are estimated (Table S1). Regarding the genomic parameters, parameters of the Beta distribution and the Q parameter, are sampled in a uniform distribution where $\alpha, \beta \in [0, 10]$ and $Q_{anc}, Q_{cur} \in [0, Q_{max}]$. $Q_{max} \leq 1$ is the maximal proportion of the genome under gene flow barrier set by the user. RIDGE produces the *reference table* from a set of simulations with parameters sampled from these prior distributions.

2.5 | Point estimates and goodness of fit of posteriors

RIDGE utilizes the *reference table* for training a regression RF model (Raynal et al., 2019). This model produces point estimates for the predicted values of each parameter and assigns weights to simulations based on their proximity to the real data using the *regAbcrf* function. The weight for each simulation is calculated as the mean of the weights across all parameters. Subsequently, a set of simulations (and their corresponding parameter values) are subsampled in proportion of these average weights to represent a set of simulations that better match the observed data. This subsample of the *reference table* is referred to as the *posterior table*. Note that subsampling of parameters according to the averaged weights over simulations effectively account for the non-independence of parameters. We evaluated the goodness of fit of the posterior distributions using an enhanced version of the *gfit* function of the *abc* packages (Csilléry et al., 2012), which employs a goodness-of-fit statistics approach described in Lemaire et al. (2016) and summarized here. To assess the goodness of fit of the posterior G_{post}, we followed these steps: first, summary statistics (in both observed data set and posterior table) are normalized by their mean absolute deviation determined from the *posteriors table*. Then, we computed the Euclidean distance between each summary statistics computed from the observed data set and those computed from each η simulation contained in the *posterior table*. Together it form a vector of Euclidean distances $d_1 \ldots d_\eta$ on which we computed the average, denoted D_{post}. To derive the null distribution of G_{post}, we considered a data set randomly sampled in the *posterior table* as 'observed' and discarded from subsequent analyses. The remaining η−1 data sets of the *reference table* were used to compute D_{post}', the average Euclidean distance between the *posterior table* and the 'observed' data set. Repeated as such Z times, we obtained a vector of $D1_{post}' \ldots D_{post}^{Z}{}'$. Then we computed G_{post} as the proportion of values for which $D_{post}' > D_{post}$.

2.6 | Detection of barrier loci

Each set of parameters of the *posterior table* is used to generate two sets of individual-locus simulations, one set for non-barrier loci (m equals to the value of the *posterior table*) and one set for barrier loci (m set to 0), with two corresponding *per-locus reference tables*. The RF algorithm (*abcrf* package) was trained on these *per-locus reference tables* to predict the most probable status of each locus, either barrier (model x_1) or non-barrier (model x_2). Since there are only two models, the posterior probabilities satisfied: $P[x_1] = 1 - P[x_2]$ so that we were able to compute a Bayes Factor (BF) for each locus i, denoted as BF_i:

$$BF_i = E\left[\frac{1-Q}{Q}\right] \cdot \left(\frac{P[x_1]_i}{1-P[x_1]_i}\right) \quad (2)$$

Here, $E[]$ represents the average of $1 - \hat{Q}$ and $\hat{Q}$ over the posterior distribution obtained from the hypermodel. Q can be zero in the empirical distribution, so the ratio undefined. Instead of removing zero values that makes the BF highly stochastic from one simulation to another, we used the following approximation (based on the Taylor expansion of the expectation of a ratio of random variables):

$$BF_i = \left(\frac{E[1-Q]}{E[Q]} + \frac{V[Q]}{E[Q]^3}\right) \cdot \left(\frac{P[x_1]_i}{1-P[x_1]_i}\right) \quad (3)$$

2.7 | Evaluation of RIDGE performance on pseudo-observed data sets

We evaluated RIDGE performance on pseudo-observed data sets (i.e. simulated data sets considered as 'observed' data and compared with simulation outputs). As a first step, we evaluated the ability of RIDGE to correctly infer demographic x genomic models. We next used the pseudo-observed data sets to evaluate the accuracy of RIDGE in estimating the proportion of barrier loci, and detecting their locations throughout the genome. SI model where all loci should be detected as barriers was used as a positive control.

We simulated pseudo-observed data sets under the four demographic models and under both *2N2m* and *2N1m* genomic models (only *2N1m* for SI). For simplicity, we fixed $N_a = N_1 = N_2 = 50,000$ individuals. The time of the secondary contact (T_{SC}) was set to $0.2 \times T_{split}$ and the time of arrest of ancestral migration (T_{AM}) was set to $0.7 \times T_{split}$. We used a range of parameter values (Table S2) for divergence (from 1000 to 2 million generations, i.e., from 0.1 to 20 in $2N_e$ generation unit), for migration (mean $4N_em = 1$ and 10) and barrier loci proportion ($Q = 1\%$, 5% or 10%). We set the mutation rate to $\mu = 1.10^{-8}$ and the recombination rate to $r = 1.10^{-7}$ so that their ratio was 10. In total, we simulated 15,000 data sets using the scrm coalescent simulator (Staab et al., 2015). Each multilocus data set contained 1000 loci of 10 kb each, and we performed 100 replicates per scenario.

To evaluate the inference of demographic × genomic models, we calculated the goodness of fit of the estimated model and determined the contribution of each model to the estimation of posteriors obtained from pseudo data sets. Contributions were evaluated through four criteria: (i) the average weight of the simulated demographic (among the four) model called here the 'correct' model, (ii) the average weight of *2m* models, (iii) the average weight of *2N* models, and (iv) the average weight of models displaying current migration. We also compared the point estimates obtained from simulations with the input parameter values.

Next, we assessed our ability to detect barrier loci using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The ROC curve relates the false-positive rate (FPR) to the true-positive rate (TPR) and provides insights into the discriminant power of a method. The AUC of the ROC ranges from 0 to 1. An AUC of 0.5 indicates that FPR and TPR are equal irrespective of the threshold, which implies a random classification of loci into barrier and non-barrier loci while an AUC of 1 indicates perfect classification. Additionally, we computed the precision as the number of true positives (TP) divided by the sum of true positives and false positives (TP + FP).

2.8 | Application to experimental data on crow hybrid zones

To assess the performance of RIDGE on experimental data, we focused on two published data sets produced by Poelstra et al. (2014) and Vijay et al. (2016). All sequencing data from crows were extracted from the NCBI database under project number PRJNA192205 and the reference genome used to map them is GCF_000738735.1. In the first one, a comparison was made between 30 individuals of *Corvus corone* (carrion crows) populations from Spain and Germany, and 30 individuals of the *C. cornix* (hooded crows) population from Poland and Sweden. In the second one, three crow contact zones, among which two well-characterized hybrid zones, with similar divergent times around ~80,000 generations are described, from the most recently diverged pair *C. corone*–*C. cornix* (RX), to the most anciently diverged *C. cornix*–*C. orientalis* (XO) and *C. orientalis*–*C. pectoralis* (OP) pairs (Vijay et al., 2016). This data set consisted of 124 sequenced individuals. The number of individuals sampled varied for each pair (RX: 15–14 individuals; XO: 6–6 individuals; OP: 5–3 individuals).

All alignments were done on a reference genome (NCBI assembly: GCF_000738735.1) consisting of 1299 scaffolds resulted in the detection of 16,064,921 common SNPs with an average density of 15 SNPs per kilobase. Previous genome-wide scans across the three pairs identified a number of candidate loci potentially involved in population/species divergence (Vijay et al., 2016). Two metrics were employed in those scans: (i) a Z-transformed F_{ST} computed on 50 kb non-overlapping windows between population/species pairs and normalized by the local level of Z-transformed F_{ST} from allopatric pairs, denoted as F_{ST}', (ii) an unsupervised genome-wide recognition of local relationship pattern using Hidden Markov Model and a Self-Organizing Map (HMM-SOM) method implemented in Saguaro (Zamani et al., 2013) to identify local phylogenetic relationships based on matrices of pairwise distance measures, across each of the target hybrid zones.

Here, we applied RIDGE on 50 kb non-overlapping windows considering a mutation rate of 3.10^{-9} for both data sets as is Poelstra et al. (2014) and Vijay et al. (2016). We, therefore, focused on scaffolds longer than 50 kb, which accounted for 9% of the total scaffolds but represented 98% of the genome, corresponding to 20,975 windows. Prior bounds are given in Table S3, and were determined based on the observed data sets and results of analysis from Vijay et al. (2016). First, we compared Bayes factor outliers ($BF > 50$) from RIDGE results with outlier loci detected in (Poelstra et al., 2014) to assess the ability of RIDGE to correctly detect barrier loci. Secondarily, we analysed RIDGE results produced on three species pairs on a lager data set (Vijay et al., 2016) to understand how *BF* correlates with summary statistics and which summary statistics are able to discriminate outlier loci ($BF > 50$).

3 | RESULTS

3.1 | Demographic inferences

The RIDGE's ability to infer demographic parameters, measured by the goodness of fit of posteriors (G_{post}), far exceeded the rejection threshold of 5% and was stable across all models and conditions tested in pseudo-observed data sets (Figure 2 and Figure S1). However, the model's contribution to the estimation of the demographic and genomic parameters varied across conditions. The

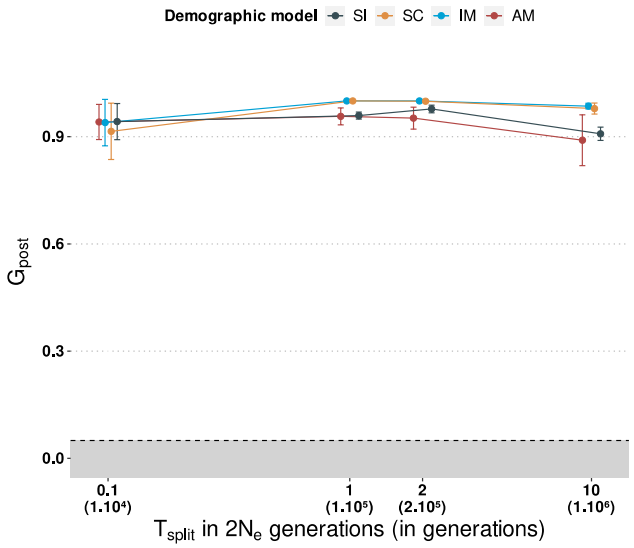

FIGURE 2 Evolution of the goodness of fit of the posteriors (G_{post}) as a function of time split, for four demographic models. The rejection threshold of 5% (under which an inferred model is discarded) is represented by the grey zone. Average values over 100 replicates with error bars (standard deviation) are presented. The data used in this figure were obtained from pseudo-observed data sets simulated under the *2N2m* model with migration set to $4N_em = 10$ and a proportion barrier $Q = 10\%$ (except for SI, no migration and no barrier).

percentage of simulations correctly attributed to the correct model increased with the time split (T_{split}), reaching over 51% for IM, 51% for SI, 60% for AM and up to 84% for SC (Figure 3). Consistently, we observed that the more recent the time split, the more balanced the contribution of different demographic models, and the greater the uncertainty surrounding the designation of a model (Figure 3 and Figure S2). For recent time splits, the choice of model is thus arbitrary, highlighting the increased utility of the model averaging approach under these conditions. Next, we investigated in greater details the consequences of model misspecification. We trained RIDGE using a *reference table* generated under IM *2N2m* and then applied it to pseudo-observed data created under both SC and AM *2N2m*, in addition to IM *2N2m* (the 'correct model') used as a control. Our results revealed a significant impact of model misspecification on G_{post} for $T_{split} = 10^6$ (Figure S3a). More importantly, the AUC fell below 0.5 and exhibited a sharp decrease for oldest T_{split} when AM model was chosen (Figure S3b). This underscores that, while IM and SC displayed similar outputs, opting for the AM model drastically increases the false positive rate.

The percentage of simulations correctly detecting the presence or absence of ongoing migration increased with T_{split} (97.6% and 98.4% at 10^6 generation for IM and SC against 5.3% for AM, Figure 3). Heterogeneous migration (2 m) was better captured under ongoing rather than ancestral migration but even under the most favourable conditions, ~25% of the simulations exhibited consistent patterns of homogeneous migration where barriers were undetectable (Figure 3). This once again emphasizes the enhanced value of employing the model averaging approach. The detection of the heterogeneity in population size (*2N*) varied little across T_{split} but tended to be more effectively detected under recent T_{split}, irrespective of the demographic model (Figure 3). Overall, these results indicated that while the correct demographic model was accurately inferred only under specific conditions, the occurrence of current migration was generally well captured.

We also examined the specific point estimates associated with each parameter. The accuracy of $\hat{T}_{split}$ estimation was only slightly affected by the proportion of barriers and migration rate, closely approximating the simulated value irrespective of the demographic model (Figure S4). Similar patterns were observed for $\hat{T}_{SC}$ and $\hat{T}_{AM}$ albeit T_{SC} tended to be slightly overestimated (Figure S5). As T_{split} increased, estimates of current population sizes $\hat{N}_1$ and $\hat{N}_2$ improved, approaching simulated values when T_{split} reached 1.10^5 generations (Figure S6). Estimates of past population size $\hat{N}_A$ is theoretically possible if $T_{MRCA} < 4N_e$ in each diverging population (with T_{MRCA} the coalescent time of the Most Recent Common Ancestor). When T_{split} is much greater than $4N_e$, most sequences are expected to coalesce before T_{split} so that less signal is available for $\hat{N}_A$ inference. In our case, $T_{MRCA} \approx 4N_e = 2.10^5$ generations, and $\hat{N}_A$ deteriorated beyond this value, converging towards the prior mean (Figure S6). Current migration estimates ($\hat{M}_{curr}$) were more reliable than ancestral migration ones ($\hat{M}_{anc}$). The proportion of barriers had minimal impact on $\hat{M}_{curr}$, under SC and IM models (Figure S7). Deeper T_{split} resulted in greater migration signal and therefore improved the accuracy of $\hat{M}_{curr}$ (Figures S7 and S8 left). In contrast, T_{split} had no clear effect on $\hat{M}_{anc}$ (Figures S8 and S9).

3.2 | Inferences of barrier proportion

The barrier proportion estimate, $\hat{Q}$, plays a crucial role in the computation of Bayes factors (Equation 2) and the detection of barrier loci. We obtained reliable estimates of the barrier proportion, $\hat{Q}$, when there was current migration (IM and SC models) and when T_{split} exceeded 1.10^5 generations (Figure 4 and Figure S10). For more recent T_{split} ($<0.2 \times 2N_e$ generations, approximately), $\hat{Q}$ was not properly estimated and converged to the prior mean, indicating that RIDGE lacks power to discriminate between barrier and non-barrier loci. Irrespective of the conditions, $\hat{Q}$ was unreliable under ancestral migration (AM model), except for both high migration rate and divergence time. Under the SI model, for which the proportion of barriers has no significance, the estimates corresponded to the prior mean. The *Q* parameter had a minimal impact on the effective migration rate as shown in Figure S8, reciprocally *M* had little impact on $\hat{Q}$ (Figure S10), so that $\hat{Q}$ was expected to exhibit a weak correlation with the genome-wide level of genetic differentiation/divergence between populations, as measured by statistics such as F_{ST}, *Da* and *Dxy*. We, therefore, introduced additional summary statistics based on the proportions of outliers for F_{ST}, *Da*, *Dxy*, *sf* and π. To assess the usefulness of these new statistics, we compared $\hat{Q}$ estimated with or without them. Overall, outlier statistics reduced estimation errors

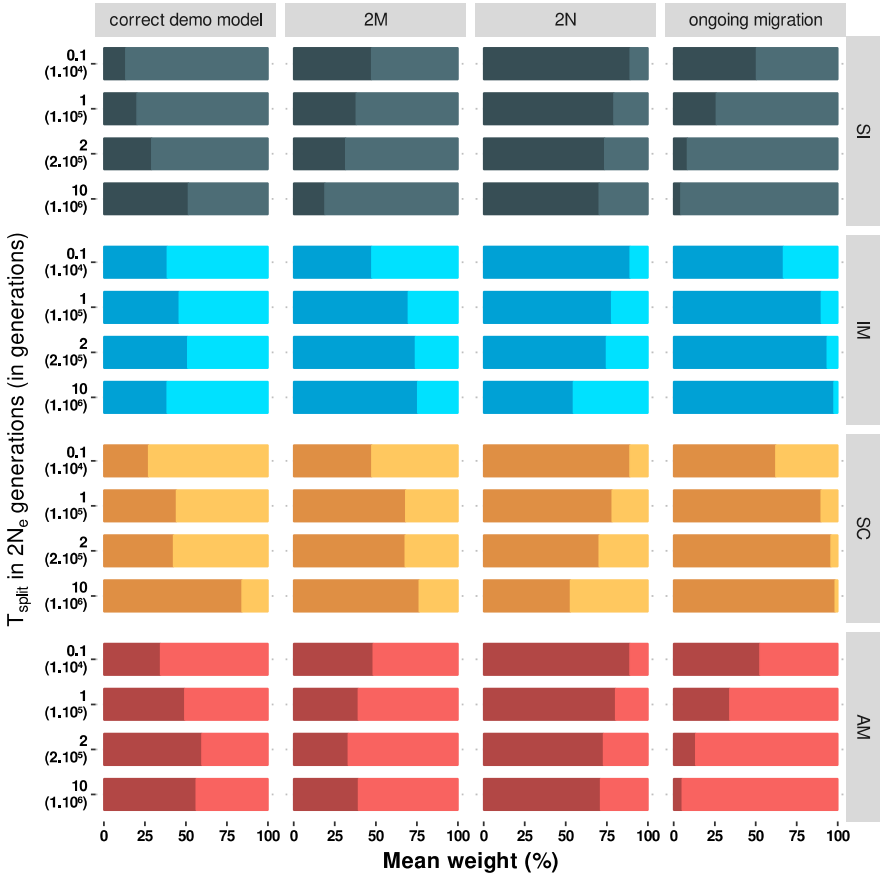

FIGURE 3 Demographic × genomic model weights in posteriors across time splits. Weight was measured by considering four criteria: (i) the average joint weight of the false and true demographic (among the fours) model—called here the 'correct' model—in posteriors, (ii) the average joint weight of *1m* and *2m* models, (iii) the average weight of *1N* and *2N* models, (iv) and the average weight of models displaying no ongoing (current) migration and ongoing migration. Proportion of accurate model predictions are shown in dark colours. As an example, for a time split of 10^6, an average weight of 0 for ongoing migration under the SI model signifies that across 100 replicates, simulations under ongoing migration represent 0% of the posteriors and so did not contribute to parameter estimation. All models were simulated under *2N2M*, and $4N_em_{curr}$ or $4N_em_{anc} = 1$.

by 8.4%. They were particularly effective in improving $\hat{Q}$ under challenging conditions for barrier proportion estimation, such as when migration was low ($M \leq 1$) and the proportion of barriers was small $Q \leq 1\%$ (Figure S11). The impact of outlier statistics varied across models and T_{split} values (Table S4). At $T_{split} = 1.10^4$, results remained difficult to interpret with variation in the signs of correlations. For $T_{split} > 1.10^4$, under the AM model, *Da* outliers positively correlated with $\hat{Q}$ (Pearson $R > 0.51$), while under the IM and SC models both *sf* and *ss* outliers exhibited a positive correlation with $\hat{Q}$ ($R > 0.88$). At $T_{split} = 1.10^6$, $\hat{Q}$ additionally correlated with *Dxy* for all models (Table S4).

3.3 | Detection of barrier loci

The parameter T_{split} plays a crucial role in detecting gene flow barriers. This is because the contrast between gene flow barriers and the rest of the genome increases with T_{split} as illustrated in Figure 5a. As T_{split} increased, the overlap between the space of summary statistics occupied by barrier and non-barrier loci decreased resulting in a more pronounced shift between the corresponding *BF* distributions (Figure 5a, b). A consistent signal was observed on posterior probability distributions where under IM, a single mode was detected for the most recent $T_{split} = 1.10^4$, while two modes corresponding to barrier and non-barrier loci emerged for older time splits (Figure S12). Note that, as expected, the SI model produces a single mode distribution irrespective of T_{split}, where all loci become barriers as T_{split} increases (Figure S12). To quantify the discriminant power of RIDGE, we used the area under the curve (AUC) of the receiver operating characteristic (ROC), as depicted in Figure 5c. When T_{split} was low, the AUC remained close to 0.5, indicating no power to detect barriers. This was confirmed by similar distributions of posterior probabilities under SI and IM for $T_{split} = 1.10^4$ (Figure S12). Our results on pseudo-observed data demonstrated that both the ability to detect barriers (measured by the AUC of the ROC) and the precision in barrier detection (measured by the PV/P ratio) increased with T_{split} (Figure 6). Moreover, barriers were more efficiently detected and at lower T_{split} under current (IM and SC models) than ancestral gene

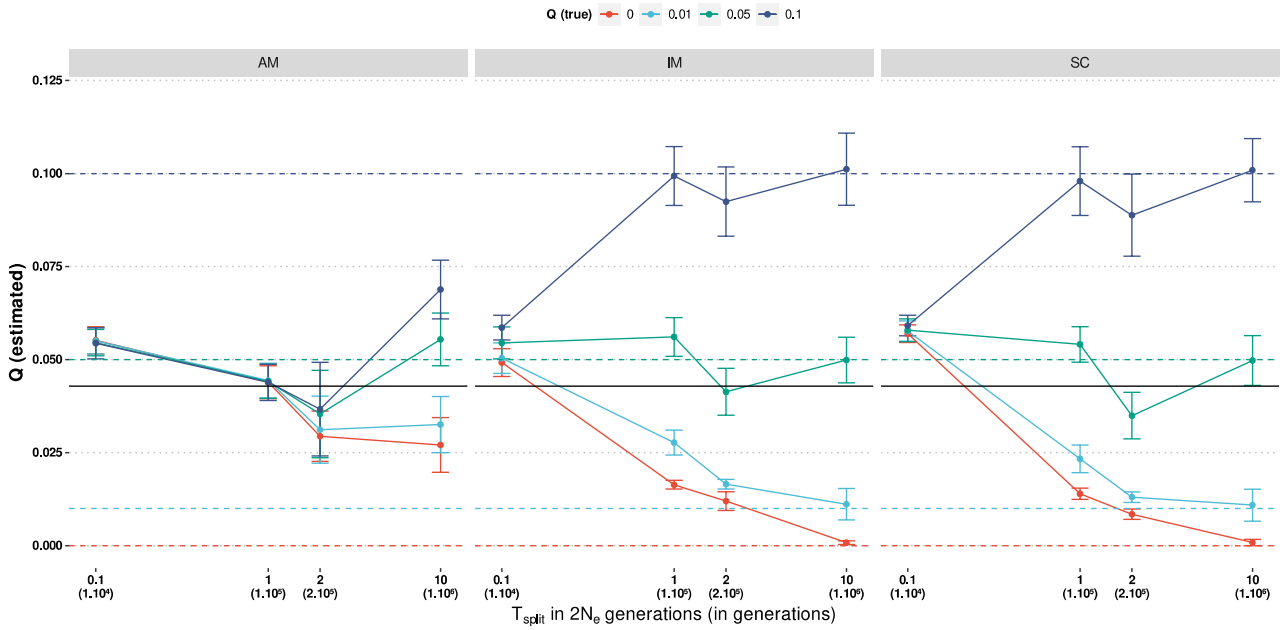

FIGURE 4 Barrier proportion estimates as a function of divergence time under three demographic models. In this figure, migration is set to $M = 10$ and the plain black line represents the priors mean. Each data point represents the average value over 100 replicates with standard deviation as error bars. Results overall conditions explored are represented in Figure S8.

flow (AM model) as shown in Figures S10 and S11. Noteworthy, the AUC never dropped below 0.5, indicating that RIDGE did not generate an excess of false positives (Figures S13 and S14).

3.4 | Detection of barrier loci on crow data sets

Poelstra et al. (2014) identified a highly divergent region on scaffold 78 and 60, which contained multiple genes identified through genomic scan, functional analysis and differential expression. These genes are involved in the melanogenesis pathway and visual perception. This region was thus considered by the author as a 'speciation island' allowing for the maintenance of phenotypic differences between crows based on colour phenotypes and colour-assortative mate choice.

We ran RIDGE on the same data set using the same window size as in Poelstra et al. (2014). Our analysis successfully fitted the observed data, with a goodness of fit indicated by $G_{post}=0.29$. The estimated value of $\hat{T}_{split}$ in $2N_e$ generation is $\hat{T}_{split}/2\hat{N}_e = 0.25$ (Table S5), corresponding to the favourable range for RIDGE to effectively detect gene flow barriers. The distribution of Bayes Factors (*BF*) was clearly bimodal with a distinct group of outliers ($BF > 50$), which accounted for 0.13% of the genome (Figure 7b). Interestingly, among these outlier loci, four genes (CACNG1, CACNG4, PRKCA and RSG9) were also found by Poelstra et al. (2014) and located on scaffold 78 (Figure 7c). The probability of detecting the same four genes just by chance was low ($p = 2.04\ 10^{-6}$). We next applied RIDGE on a genome-wide data set produced for three pairs of *Corvus* species that form hybrid zones (pair RX: *C. corone–C. cornix*; pair XO: *C. cornix–C. orientalis*; pair OP: *C. orientalis–C. pectoralis*) where current gene flow is detected (Vijay et al., 2016). For a single pair of crow species, the program took approximately 1,883,000 s of CPU runtime on four CPUs running at a minimum of 2.5 GHz. Therefore, in real-time, it took around 36 h for the whole data set on a cluster of 280 CPUs, which takes into account server latencies, job queues and CPU availability. The goodness of fit of the demographic parameters inferred by RIDGE was similar across all three pairs (RX: 0.33; XO: 0.21; OP: 0.26). The ratio of $\hat{T}_{split}/2\hat{N}_e$ was approximately 0.3 for all three pairs (RX: 0.28; XO: 0.27; OP: 0.31; Table S5), suggesting a comfort zone for RIDGE to detect gene flow barriers in all three data sets.

PCA analyses coloured by *BF* show a main group of outliers (characterized by elevated F_{ST} and/or *Da* and/or reduced level of diversity in all four pairs Figures 7a and 8 and Figure S15). Those signals were consistent with some theoretical expectations for gene flow barriers (i.e. increased *Da*, *sf*, F_{ST} and reduced *ss* and diversity), but almost no relationship with *Dxy*. In each pair, we identified a subset of loci with elevated Bayes factors ($BF > 50$) clearly separated from the genome-wide distribution (Figure 8c). These subsets detected on a per-locus basis (RX: 0.12%; XO: 0.02%; OP: 0.17%), represented smaller proportions than the expected proportion estimated in the general model, $\hat{Q}$ (RX: 4.9%; XO: 4.8%; OP: 4.7%) but still fell within the credibility intervals (Figure 8b and Table S5).

We found significant overlap between our outliers and those of Vijay et al. (2016) for the RX and OP pairs (69% and 28%, respectively, Figure 8a, b). For XO, we only detected four candidates, which makes the comparison difficult with Vijay et al. (2016) although using a less stringent $BF > 10$, the overlap was significant ($p = 0.007$). The *BF* revealed various correlation patterns among the three pairs, with F_{ST} and *Da* being consistently strongly positively correlated with

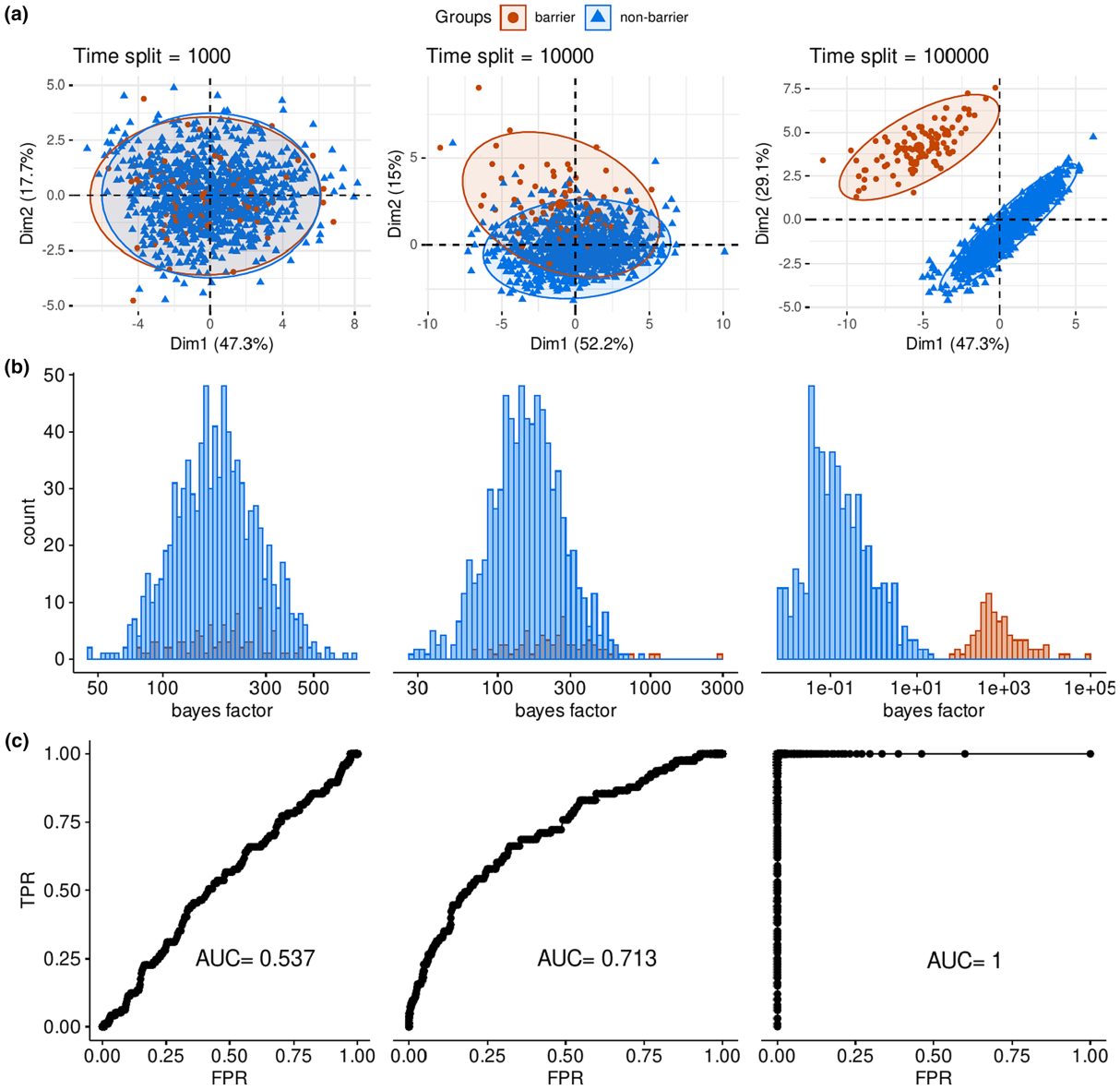

FIGURE 5 Impact of the divergence time on the overlap between barrier and non-barrier loci. Overlap revealed by a principal component analysis (PCA) computed on all 14 summary statistics (a), the log of the Bayes factor (BF) produced by RIDGE (b) and the area under the ROC curve (AUC) of the Bayes factor (c). The greater the AUC the higher the discriminant power is. A single pseudo-observed data set was used for each of the three values of T_{split}. Data sets were simulated under an IM *2N2m* model, with the following parameters: $4N_em = 10$ and $Q = 0.1$.

BF and *ss* being consistently negatively correlated with *BF* but to a lesser extent (Figure 9).

4 | DISCUSSION

A key goal of speciation research is to elucidate the genetic mechanisms behind reproductive isolation. Although diverging populations have been analysed in many studies, a challenging aspect remains the ability to capture the sequence of events that lead to the establishment of reproductive barriers. To answer this question, one approach is to compare populations that exhibit varying degrees of temporal and/or spatial divergence, including recently diverged ones. This requires the use of a comparative framework capable of detecting barriers to gene flow at both early and ancient stages across diverse biological systems, independently of their demographic history. In this context, we introduce RIDGE, a tool designed to facilitate this task.

4.1 | RIDGE offers a comparative framework where current migration is well captured

Currently, two methods explicitly model heterogeneity in the effective migration rate across the genome. Both tools utilize variations in effective population size to approximate selective effects along

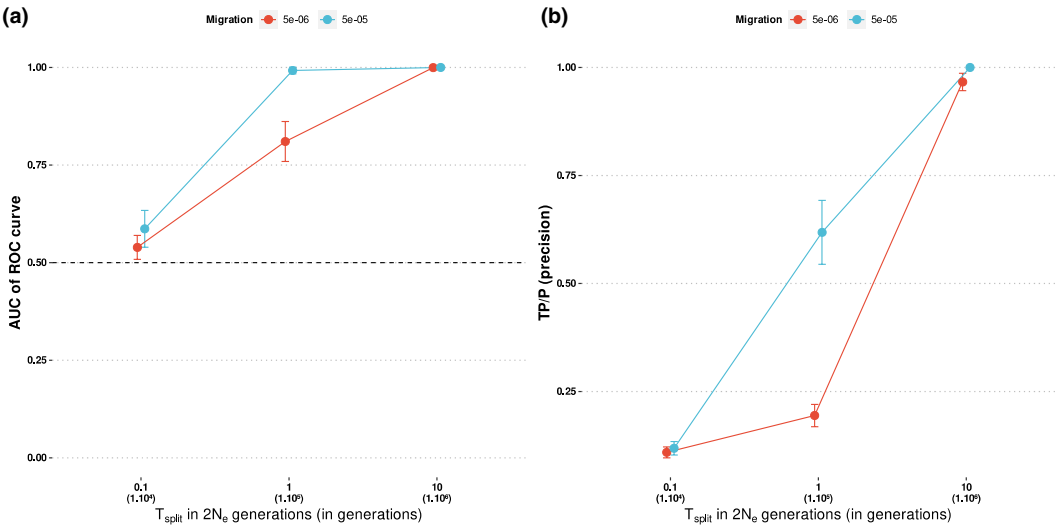

FIGURE 6 Ability and precision in the detection of barrier loci as a function of divergence time and migration. Ability is measured by the AUC of the ROC (a) and precision by TP/P (b). Considering a proportion of barrier $\hat{Q}$, barrier loci are those displaying a Bayes factor superior to the quantile at $1 - \hat{Q}$. Each data point represents the average value over 100 replicates with standard deviation as error bars. Simulations were performed under an IM *2N2m* model with $Q = 0.1$.

the genome. DILS (Fraïsse et al., 2021) uses an ABC framework under four demographic models of divergence (SI, IM, SC, AM) to assess alternative models of effective migration's homogeneity/heterogeneity and to provide corresponding genome-wide estimates. While not primarily designed to perform barrier detection, DILS can still offer valuable insights on potential barrier loci, conditioned on the selected demographic model (Fraïsse et al., 2021). There are, however, two main limits to this approach. Firstly, selecting a model can be rather arbitrary when two models explain the data equally well, which is often the case when divergence is shallow between populations (as shown in Fraïsse et al. (2021) and confirmed here, Figure 3 and Figure S2); and model misspecification can have strong consequences on the rate of false positives (Figure S3). Secondly, the use of potentially different demographic models complicates comparison across species pairs. gIMble (Laetsch et al., 2023) relies on composite likelihood to identify windows of unexpected level of effective migration along the genome. It first computes a general homogeneous model (homo-*N*, homo-*m*) and then fits a model for each window yielding local estimate of N_e and *m*. Then, it uses a parametric bootstrap approach to assess the statistical significance of a putative barrier. However, because it relies on likelihood computation, gIMBle is less flexible than ABC methods and can only handle the IM model, while secondary contacts may be rather frequent in nature (ex: Leroy et al., 2020; Roux et al., 2013; Vijay et al., 2016).

RIDGE builds on DILS, offering a high degree of model flexibility, while proposing a comparative framework. In order to do so, RIDGE employs a model averaging approach by assigning weights to each demographic x genomic model without directing the user's choice towards a single model. In addition, model averaging is also useful in reducing the uncertainty on parameter estimation when individual models present high variance (Dormann et al., 2018). Our results show that model averaging is especially relevant when data offers little discriminant power. For example, when T_{split} is low, the discriminatory power of summary statistics is reduced, resulting in similar assignation to all models (Figure 3). Opting for the best scenario under such conditions might be misleading. For example, at $T_{split} = 0.1 \times 2N_e$, when current migration is simulated (IM or SC models), it is detected in only ~60% of the cases (Figure 3), thus potentially leading to the selection of the SI or AM models, thereby impeding the estimation of gene flow barriers. In contrast, the model averaging approach always provides an estimate of the proportion of gene flow barrier with a credibility interval, which can be large and include 0 when the statistical power is low. RIDGE thus allows for formal comparison of any data sets despite differences in demographic history and/or statistical power.

In addition, compared to DILS, RIDGE makes another improvement in the way heterogeneity of migration is modelled. DILS models separately the heterogeneity in N_e and $M = 4N_e m$, which can lead to unrealistic scenarios where *m* is inversely proportional to N_e (when N_e is heterogeneous and *M* constant), which should inflate the detection of heterogeneity in migration rate. To illustrate it, we ran a modified version of RIDGE on the crow data sets where migration is modelled as in DILS (constant or variable *M* instead of *m*, independently of N_e, see Text S1). Employing the DILS-like version resulted in the detection of numerous additional putative barriers, some of which were challenging to interpret (e.g. high diversity and relatively low F_{st}). Moreover, the correlations between Bayes factors (*BF*) and summary statistics varied across data sets, lacking a clear interpretation for the RX and XO pairs.

A direct consequence of using a demographic×genomic hypermodel is that RIDGE is not intended for precise estimation of a demographic model and its underlying parameters but rather to address

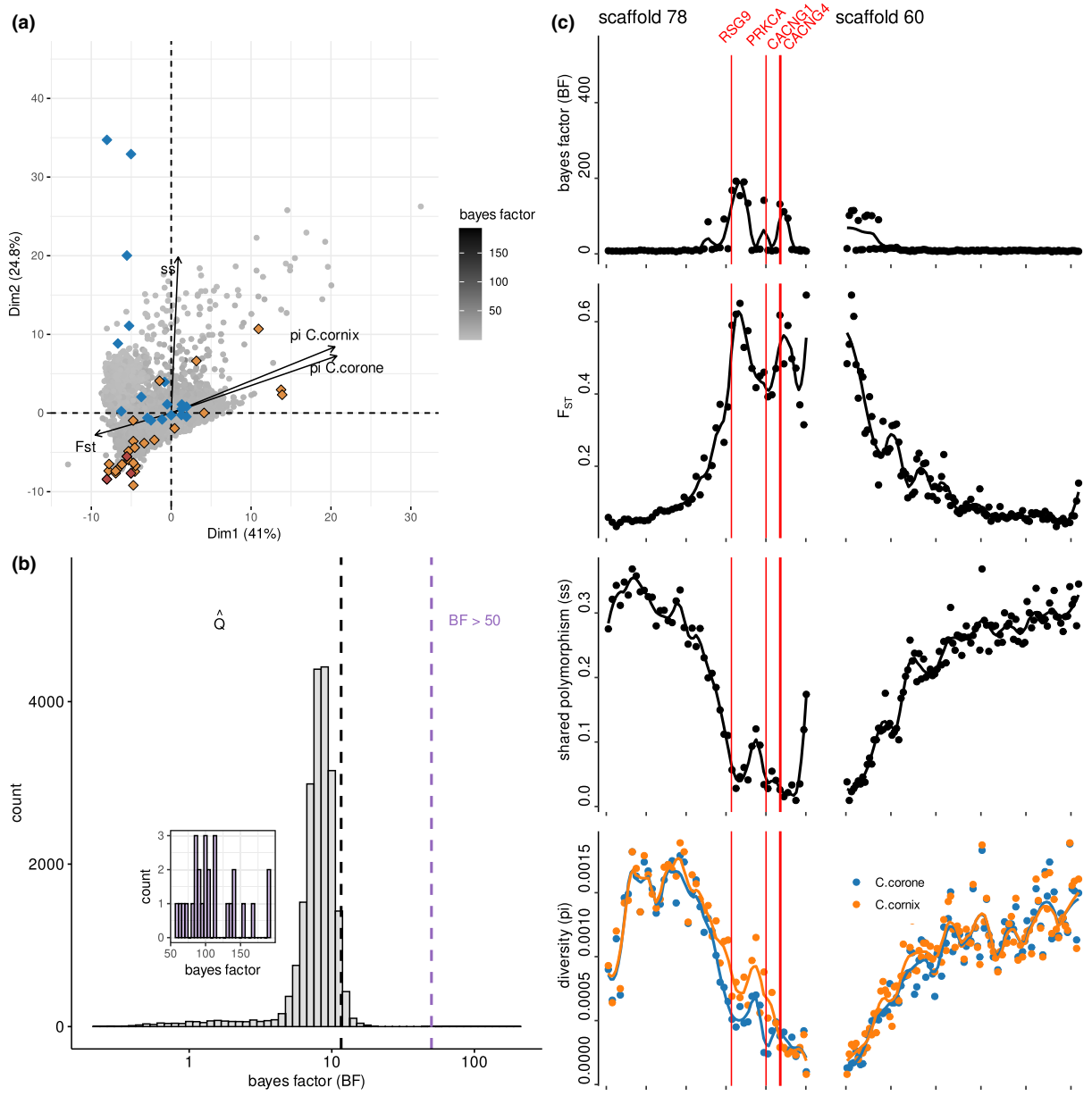

FIGURE 7 Results of the analysis conducted using RIDGE on the crow hybrid zone between carrion and hooded crows. PCA computed on summary statistics obtained from 50-kb windows along genomes with axes 1 and 2 (only 4 of 14 summary statistics are represented), where each datapoints (windows) are coloured according to the values of Bayes factors (a). Blue diamonds represent loci detected in Poelstra et al. (2014), yellow diamonds indicate loci detected by RIDGE that exceeded the population-specific Bayes factor threshold and red diamonds represent loci detected both in Poelstra et al. (2014) and RIDGE. Distribution of Bayes factors across the genome (b). Genomic landscape of scaffold 78 and 60 through Bayes factor, F_{ST}, shared polymorphism (*ss*) and diversity (π) (c). Data are from Poelstra et al. (2014).

demography as a confounding factor in the detection of gene flow barriers. High and stable values of goodness of fit across models and conditions indicate that we achieved this goal (Figure 2 and Figure S1) and more moderately for complex/real scenario as for crow data sets (Table S5) where the goodness of fit is lower (G_{post} ~0.9 for simulated data sets, G_{post} ~0.25 for crow data sets). However, as expected, the accuracy of parameter estimation largely depends on the divergence time (Figures S6–S9). Similar to DILS (Fraïsse et al., 2021), the correct model's contribution to parameter estimation and the detection of ongoing migration increases with divergence time (Figure 3). Overall, current migration is well captured, both in model weights and in parameter estimation (Figure 3, Figure S7).

This is well illustrated with the analysis of the crow data sets. After the ice cap had retreated in Europe around 10,000 years ago (~2000 crow generation), the ancestors of remnant carrion (*C. corone*) and hooded crow (*C. cornix*) populations met in a secondary contact in Central Europe, forming a narrow and stable hybrid zone (Knief et al., 2019; Metzler et al., 2021; Poelstra et al., 2014). Based on the sampling by Poelstra et al. (2014), which covers a wide geographical area away from the central European hybrid zone, RIDGE favoured

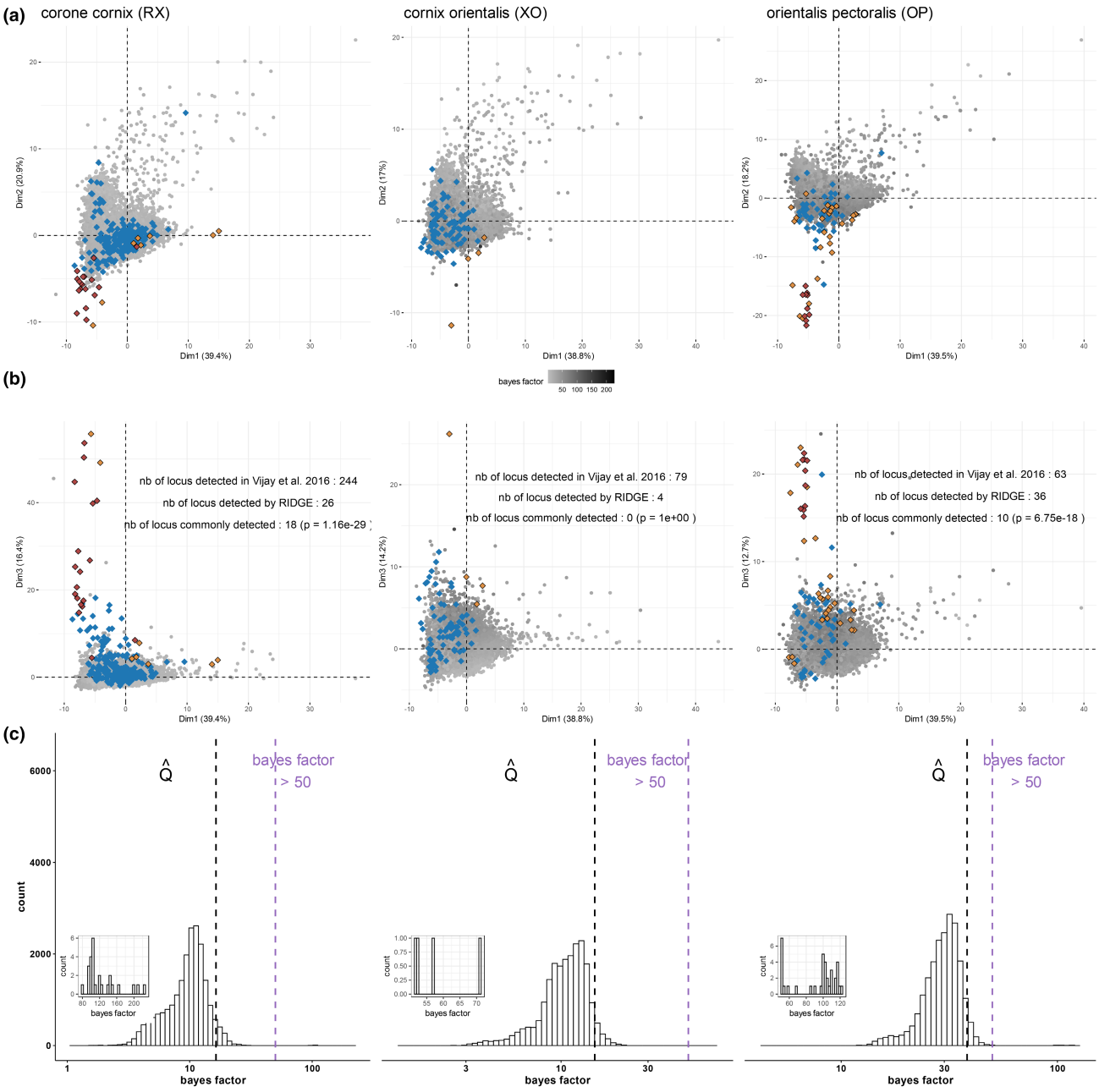

FIGURE 8 Barrier loci detection by RIDGE on three crow hybrid zones. PCA computed on summary statistics obtained from 50-kb windows along genomes with axes 1 and 2 (a) and 1 and 3 (b) displayed. Datapoints (windows) are coloured according to the values of Bayes factors. Blue diamonds represent loci detected in Vijay et al. (2016), yellow diamonds indicate loci detected by RIDGE that exceeded the population-specific Bayes factor threshold and red diamonds represent loci detected both in Vijay et al. (2016) and RIDGE. Distribution of Bayes factor values for each species pair (c). The histogram inside the figure shows the Bayes factor distribution of detected loci, which are the loci exceeding the population-specific Bayes factor threshold indicated by the violet dashed line. Black dashed lines indicate the Bayes factor threshold based on the estimated barrier proportion $\hat{Q}$. Data are from Vijay et al. (2016).

the correct scenario, especially the occurrence of ongoing migration (model weight for SC = 45% and IM = 44%) (Table S6). Similar results were obtained for the RX hybrid zone with IM at 43% and SC at 39%. Overall, in all four data sets the current status of migration has been correctly captured with ongoing migration accounting for the majority of the model weight (RX: 82%; XO: 84%; OP: 91%; Poelstra et al., 2014: 89%).

4.2 | Informative summary statistics are context-dependent

One drawback of the ABC approach is that parameter inference relies on summary statistics to capture the genomic signal. Historically, F_{st}, a measure of relative divergence, has been the most widely used statistic in genome scans (Wolf & Ellegren, 2017). To avoid

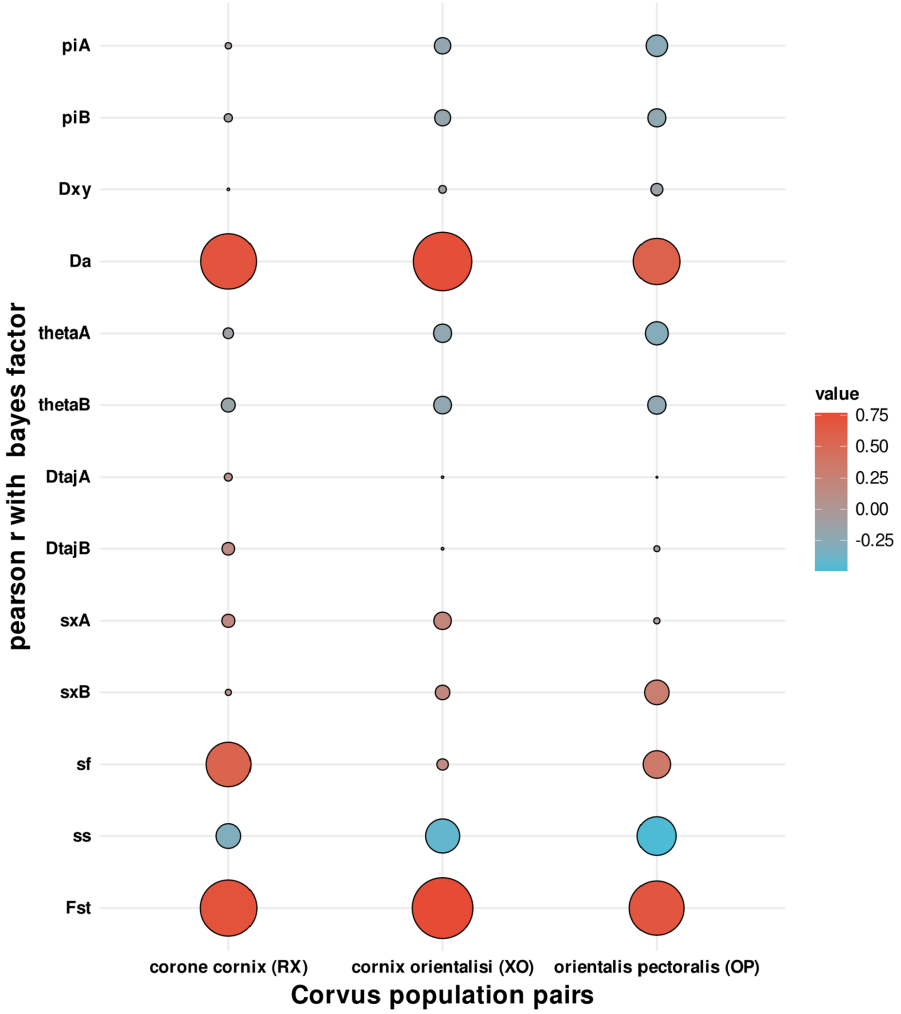

FIGURE 9 Pearson correlation between RIDGE Bayes factor and summary statistics used in the gene flow barrier detection for the three hybrid zones. Colours correspond to the values of correlations while circle size reflects the absolute values. Data are from Vijay et al. (2016).

the confounding effect of reduced diversity in either of the compared populations due to other causes than barrier to migration (Cruickshank & Hahn, 2014; Ravinet et al., 2017), it is now common practice to combine it to absolute measure of divergence (*Dxy*) to other related statistics such as net divergence (*Da*) or the number of fixed differences (*sf*) (Han et al., 2017; Hejase et al., 2020). Here, we devised a new set of summary statistics based on outlier detection, and proved them to be useful for estimating barrier proportions. The reasoning was that loci showing local increase in divergence (measured by F_{ST}, *Dxy*, *Da*, *sf*, *ss*) and decrease in diversity would generate outliers in the genome wide divergence and diversity distributions. Our results show that outlier statistics mostly contribute to $\hat{Q}$ under moderate gene flow ($M = 1$), and mainly for low level of barrier proportion ($Q < 0.1$) (Figure S11), where estimation of barrier proportion may be challenging.

Interestingly, the set of summary statistics that effectively capture the signal of barrier loci slightly differed among data sets, as illustrated with the three pairs of crows (Figure 9). For the three pairs, F_{ST} and *Da* strongly correlated with *BF* and contributed the most to barrier detection, in agreement with theoretical predictions (Cruickshank & Hahn, 2014). Quite unexpectedly, however, *Dxy* did not correlate with *BF* and did not contribute to barrier detection. A possible explanation is that, at low divergence, variations in *Dxy* mainly reflect local variations in N_e (as confirmed by the strong positive association with π in the PCA, Figure S15), while the main signal of variation in migration rate is already captured by *Da*. Other statistics also correlated with *BF* but at lower and variable levels in the three data sets, and, similarly outliers correlated differently to the PCA axes (Figure 8 and Figure S15). Differences in genomic signatures may reflect not only the difference in the environment in which incipient crow species evolved but also the difference in the geographical area covered by the hybrid zone (Vijay et al., 2016).

These examples illustrate that considering a few statistics in the detection of barrier loci can be misleading as signatures can be complex and context dependent. It thus advocates for the use of a more inclusive approach as implemented in the *BF* derived from the random-forest-based ABC approach of RIDGE. One contribution of the Random Forest (RF) is to reduce the curse of dimensionality (Bellman & Kalaba, 1959), which improves accuracy and computation time, RF also makes ABC a calibration-free problem by automating the inclusion of

summary statistics (Raynal et al., 2019). In return, a possible drawback is that RF results are less interpretable due to their complex nature. Indeed, even if the *abcrf* package provides a way to understand the contribution of variables to parameters estimations, it still remains difficult to interpret the RF decision for a specific locus.

4.3 | Detection of barrier loci using RIDGE

We validated the ability of RIDGE to detect gene flow barriers on empirical data sets from Poelstra et al. (2014) and Vijay et al. (2016). In particular, we clearly detected the large and well-established region of scaffold 78 on chromosome 18. It contains major loci that are involved in mate choice patterns between *C. corone* and *C. cornix* (RX) (Knief et al., 2019; Metzler et al., 2021; Poelstra et al., 2014). The study by Vijay et al. (2016) was conducted on three species pairs that had similar demographic histories. For all three pairs of populations, we identified a portion of loci exhibiting elevated *BF*. For the RX and OP pairs, we found less loci than previously detected by Vijay et al. (2016) but a significant overlap between the two set of genes. Using a rather stringent threshold of $BF > 50$, 69% (for RX) and 28% (for OP) of the loci that RIDGE detected were also identified by Vijay et al. (2016). For the three pairs, Vijay et al. (2016) detected (many) more loci than RIDGE. On average these additional loci, not detected by RIDGE, displayed low diversity without distinctive divergence patterns. This observation can be attributed to the confounding effect of the heterogeneity in N_e, not explicitly accounted for in Vijay et al. (2016) and which is a classic pitfall of F_{st} scan approaches (Cruickshank & Hahn, 2014). The fact that RIDGE detected only a limited number of loci displaying such a pattern implies that it effectively circumvents this problem. For the XO pair, its wide spatial range—three to seven times wider than the hybrid zone of RX pair—leads to a reduction in selection strength as documented in Vijay et al. (2016), and consequently, candidate regions in our results exhibit shallow divergence patterns (Figure 8). Furthermore, since low signal can increase noise in detection results, we did not detect any direct overlap between the candidate XO gene from Vijay et al. (2016) and our results. However, when examining the regions surrounding the candidate gene, we observed common regions such as the gene *LRP5*, which was consistently present in XO and OP pairs in Vijay and was consistently located at a distance of 50 kb from an outlier locus in our results.

4.4 | Benefits of RIDGE and guidelines for its use

RIDGE relies on an ABC approach that offers a lot of flexibility, enabling it to explore genomic heterogeneity and to incorporate customized summary statistics. We have also devised a method for generating multidimensional parameter estimates, extending beyond the initial single-parameter focus of *abcrf* (Raynal et al., 2019). This improvement enables RIDGE to deal effectively with parameter interdependencies and increase the precision of parameter estimations. Another improvement introduced by RIDGE is the incorporation of Bayes factors, facilitating result comparisons. In addition, RIDGE explicitly models variation in the migration rate, m rather than the population-scaled migration rate ($4N_em$) as in DILS (Fraïsse et al., 2021) which results is a much more stringent detection of barrier loci (Text S1). Our interpretation is that by fixing both N_e and $4N_em$ as in DILS, the heterogeneity of migration, m, tends to be too frequently inferred because it allows reconciling the observed patterns for different statistics.

One limitation of RIDGE is the need to define a priori the size of windows, an arbitrary choice that can pose problems in cross-species comparisons. One possible improvement would be to define window size based on the genetic instead of the physical distance when a genetic map is available. Alternatively, one could use criteria based on local topologies to segment the genome into windows, as implemented in Saguaro, which relies on a Hidden Markov Chain model coupled with unsupervised pattern recognition and classification algorithms (Zamani et al., 2013).

The simulated data sets we explored gave us guidelines for the conditions where RIDGE can provide useful and accurate results. We suggest to use data sets with SNP density higher than 0.1%, such as in crows and simulated data sets, where the SNP density was around 1%. We also advise to use a minimum of three samples per population. The goodness-of-fit statistics enables users to check the quality of inferences made. If $G_{post} < 5\%$, the user should verify the prior bounds. The guidelines for interpreting and thresholding *BF* depend on the user's goals. If RIDGE is used solely to discover new candidate genes involved in gene flow barriers for a specific population pair, we recommend using a customized threshold that optimally captures Bayes factor outliers. For the purpose of comparison, it is recommended to use a standard threshold for all data sets, for example, $BF > 50$ or 100, or to keep the number of outlier loci corresponding to the proportion of barriers estimated in the first step of RIDGE ($\hat{Q}$). In addition, it is also important to consider the whole distribution of *BF* (or posterior probability) to help interpreting the results. For example, under the SI model (with sufficient divergence) all loci or a large proportion of loci appear as barrier but the global distribution is unimodal in sharp contrast with an IM model with barriers, which presents a clear bimodal distribution (Figure S12).

Crucially, genomic data alone cannot provide conclusive evidence of barrier loci and so RIDGE results should be coupled with other analysis such as functional analysis (Ravinet et al., 2017). It is worth noting that window length (default set to 10 kb) can significantly affect the results of RIDGE. It should be determined according to the extent of linkage disequilibrium as well as the level of diversity, since it determines the amount of polymorphism and consequently affects the strength of the signal.

As is the case with all ABC approaches, the quality of the priors given by the user affects the results obtained using RIDGE. A T_{split} of $0.1 \times 2N_e$ generations (10,000 generations in our simulations) appears to be a lower bound for both demography (Figures 4 and 5) and barrier inferences (Figure 6), below which RIDGE fails to capture informative signals. RIDGE can detect gene flow barriers on both simulated

(Figure 6) and empirical data (Figure 7), starting at 0.1 × 2Ne generation, which represents a very low level of divergence. For context, DILS correctly inferred a gene flow barrier when $T_{split} > 0.5 \times 2N_e$ generations, while gIMble only demonstrated its effectiveness on one pair of *Heliconius* species that diverged 4.5 million generations ago, estimated to represent $0.49 \times 2N_e$ generations (Martin et al., 2015).

Comparative approaches have been useful in understanding the genomic basis involved in the process of reproductive isolation (e.g. Roux et al., 2016) and they will continue to play an important role in speciation research. By its flexibility and its comparative framework, RIDGE should become a useful tool to follow this direction.

AUTHOR CONTRIBUTIONS

Designed research: Maud Tenaillon, Sylvain Glémin; Performed research: Ewen Burban; Funding acquisition: Maud Tenaillon, Sylvain Glémin; Informatics tool development: Ewen Burban; Analysed data: Ewen Burban; Supervision: Sylvain Glémin, Maud Tenaillon; Wrote the paper—original draft: Ewen Burban; Wrote the paper—review and editing: Ewen Burban, Sylvain Glémin, Maud Tenaillon.

ACKNOWLEDGEMENTS

We thank Camille Roux for the help with the DILS code and Miguel de Navascués for advice in the use of the ABC-RF method. We also thank Thibault Leroy, Christelle Fraïsse, Yves Vigouroux, Maxime Bonhomme and Claire Mérot for their insightful discussions and valuable inputs during the course of the project. We thank Augustin Desprez, Harry Belcram, Clementine Tocco and Arthur Wojcik for helping to improve RIDGE by beta-testing it. We also thank Chyi Yin Gwee and Jochen Wolf for providing us with the pre-mapped VCF data set of crows. We are also extremely grateful to two anonymous reviewers that helped improving the manuscript. This work benefited from the computing resources provided by the GenOuest cluster, the Cornuta cluster and the IFB core cluster. This work was supported by the grant DomIsol overseen by the French National Research Agency (ANR-19-CE32-0009-02). GQE-Le Moulon benefits from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007) as well as from the Institut Diversité, Ecologie et Evolution du Vivant (IDEEV). E.B. was financed by a doctoral contract from DomIsol and from Région Bretagne through the Doctoral School EGAAL. In addition, E.B. benefited from a travel grant from GDR 3765 'Approche Interdisciplinaire de l'Évolution Moléculaire'.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Source codes to deploy RIDGE and user manual are freely available from GitHub: https://github.com/EwenBurban/RIDGE.git. This GitHub repository also includes a pipeline for simulating pseudo-observed data sets and an optimized pipeline for running RIDGE on thousands of pseudo-observed data sets. Raw vcf of crow data accompanied by population file, simulation data, results of RIDGE and scripts are available as supplementary files.

ORCID

Sylvain Glémin https://orcid.org/0000-0001-7260-4573

REFERENCES

Bay, R. A., Arnegard, M. E., Conte, G. L., Best, J., Bedford, N. L., McCann, S. R., Dubin, M. E., Chan, Y. F., Jones, F. C., Kingsley, D. M., Schluter, D., & Peichel, C. L. (2017). Genetic coupling of female mate choice with polygenic ecological divergence facilitates stickleback speciation. *Current Biology*, *27*(21), 3344–3349.e4. https://doi.org/10.1016/j.cub.2017.09.037

Bellman, R., & Kalaba, R. (1959). On adaptive control processes. *IRE Transactions on Automatic Control*, *4*(2), 1–9. https://doi.org/10.1109/TAC.1959.1104847

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., Dangl, J. L., & Weigel, D. (2007). Autoimmune response as a mechanism for a Dobzhansky-muller-type incompatibility syndrome in plants. *PLoS Biology*, *5*(9), e236. https://doi.org/10.1371/journal.pbio.0050236

Charlesworth, B. (1993). Directional selection and the evolution of sex and recombination. *Genetics Research*, *61*(3), 205–224. https://doi.org/10.1017/S0016672300031372

Charlesworth, B., & Jensen, J. D. (2021). Effects of selection at linked sites on patterns of genetic variability. *Annual Review of Ecology, Evolution, and Systematics*, *52*(1), 177–197. https://doi.org/10.1146/annurev-ecolsys-010621-044528

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. https://doi.org/10.1111/mec.12796

Csilléry, K., François, O., & Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, *56*(6), 879–886. https://doi.org/10.1080/10635150701701083

Delmore, K. E., Lugo Ramos, J. S., Van Doren, B. M., Lundberg, M., Bensch, S., Irwin, D. E., & Liedvogel, M. (2018). Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evolution Letters*, *2*(2), 76–87. https://doi.org/10.1002/evl3.46

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., … Hartig, F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504. https://doi.org/10.1002/ecm.1309

Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., Loire, É., Simon, A., Galtier, N., Duret, L., Bierne, N., Vekemans, X., & Roux, C. (2021). DILS: Demographic inferences with linked selection by using ABC. *Molecular Ecology Resources*, *21*, 2629–2644. https://doi.org/10.1111/1755-0998.13323

Gavrilets, S. (2003). Perspective: Models of speciation: What have we learned in 40 years? *Evolution*, *57*(10), 2197–2215. https://doi.org/10.1111/j.0014-3820.2003.tb00233.x

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, *27*(6), 1004–1015. https://doi.org/10.1101/gr.212522.116

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hejase, H. A., Salman-Minkov, A., Campagna, L., Hubisz, M. J., Lovette, I. J., Gronau, I., & Siepel, A. (2020). Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30554–30565. https://doi.org/10.1073/pnas.2015987117

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Kaplan, N. L., Hudson, R. R., & Langley, C. H. (1989). The "hitchhiking effect" revisited. *Genetics*, *123*(4), 887–899. https://doi.org/10.1093/genetics/123.4.887

Kassambara, A. (2020). Ggpubr: 'ggplot2' based publication ready plots. R Package Version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kassambara, A., & Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Knief, U., Bossu, C. M., Saino, N., Hansson, B., Poelstra, J., Vijay, N., Weissensteiner, M., & Wolf, J. B. W. (2019). Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone. *Nature Ecology & Evolution*, *3*(4), 570–576. https://doi.org/10.1038/s41559-019-0847-9

Laetsch, D. R., Bisschop, G., Martin, S. H., Aeschbacher, S., Setter, D., & Lohse, K. (2023). *Demographically explicit scans for barriers to gene flow using glMble (p. 2022.10.27.514110)*. bioRxiv. https://doi.org/10.1101/2022.10.27.514110

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://www.jstatsoft.org/v25/i01/

Lemaire, L., Jay, F., Lee, I.-H., Csilléry, K., & Blum, M. G. B. (2016). *Goodness-of-fit statistics for approximate Bayesian computation (arXiv:1601.04096)*. arXiv. https://doi.org/10.48550/arXiv.1601.04096

Leroy, T., Rougemont, Q., Dupouey, J.-L., Bodénès, C., Lalanne, C., Belser, C., Labadie, K., Le Provost, G., Aury, J.-M., Kremer, A., & Plomion, C. (2020). Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytologist*, *226*(4), 1183–1197. https://doi.org/10.1111/nph.16039

Martin, S. H., Eriksson, A., Kozak, K. M., Manica, A., & Jiggins, C. D. (2015). *Speciation in Heliconius Butterflies: Minimal Contact Followed by Millions of Generations of Hybridisation (p. 015800)*. bioRxiv. https://doi.org/10.1101/015800

Merrill, R. M., Rastas, P., Martin, S. H., Melo, M. C., Barker, S., Davey, J., McMillan, W. O., & Jiggins, C. D. (2019). Genetic dissection of assortative mating behavior. *PLoS Biology*, *17*(2), e2005902. https://doi.org/10.1371/journal.pbio.2005902

Metzler, D., Knief, U., Peñalba, J. V., & Wolf, J. B. W. (2021). Assortative mating and epistatic mating-trait architecture induce complex movement of the crow hybrid zone. *Evolution*, *75*(12), 3154–3174. https://doi.org/10.1111/evo.14386

Miles, A., pyup.io bot, Murillo, R., Ralph, P., Harding, N., Pisupati, R., Rae, S., & Millar, T. (2021). *cggh/scikit-allel: V1.3.3 [Computer software]. Zenodo*. https://doi.org/10.5281/zenodo.4759368

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273.

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M. G., & Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410–1414. https://doi.org/10.1126/science.1253226

Powell, D. L., García-Olazábal, M., Keegan, M., Reilly, P., Du, K., Díaz-Loyo, A. P., Banerjee, S., Blakkan, D., Reich, D., Andolfatto, P., Rosenthal, G. G., Schartl, M., & Schumer, M. (2020). Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science*, *368*(6492), 731–736. https://doi.org/10.1126/science.aba5216

R Core Team. (2021). *R: A language and environment for statistical computing (version 4.1.2)*. R Foundation for Statistical Computing. https://www.R-project.org/

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: A road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. https://doi.org/10.1111/jeb.13047

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728. https://doi.org/10.1093/bioinformatics/bty867

Roux, C., Fraïsse, C., Castric, V., Vekemans, X., Pogson, G. H., & Bierne, N. (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. *Journal of Evolutionary Biology*, *27*(8), 1662–1675. https://doi.org/10.1111/jeb.12425

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding light on the Grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, *14*(12), e2000234. https://doi.org/10.1371/journal.pbio.2000234

Roux, C., Tsagkogeorga, G., Bierne, N., & Galtier, N. (2013). Crossing the species barrier: Genomic hotspots of introgression between two highly divergent Ciona intestinalis species. *Molecular Biology and Evolution*, *30*(7), 1574–1587. https://doi.org/10.1093/molbev/mst066

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, *8*(3), 336–352. https://doi.org/10.1111/j.1461-0248.2004.00715.x

Sakamoto, T., & Innan, H. (2019). The evolutionary dynamics of a genetic barrier to gene flow: From the establishment to the emergence of a peak of divergence. *Genetics*, *212*(4), 1383–1398. https://doi.org/10.1534/genetics.119.302311

Schluter, D. (2000). *The ecology of adaptive radiation*. OUP Oxford.

Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution*, *16*(7), 372–380. https://doi.org/10.1016/S0169-5347(01)02198-X

Schluter, D., & Rieseberg, L. H. (2022). Three problems in the genetics of speciation by selection. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(30), e2122153119. https://doi.org/10.1073/pnas.2122153119

Sethuraman, A., Sousa, V., & Hey, J. (2019). *Model-based assessments of differential introgression and linked natural selection during divergence and speciation*. BioRxiv. https://doi.org/10.1101/786038

Shafer, A. B. A., & Wolf, J. B. W. (2013). Widespread evidence for incipient ecological speciation: A meta-analysis of isolation-by-ecology. *Ecology Letters*, *16*(7), 940–950. https://doi.org/10.1111/ele.12120

Sousa, V. C., Carneiro, M., Ferrand, N., & Hey, J. (2013). Identifying loci under selection against gene flow in isolation-with-migration Models. *Genetics*, *194*(1), 211–233. https://doi.org/10.1534/genetics.113.149211

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). Scrm: Efficiently simulating long sequences using the approximated coalescent with

recombination. *Bioinformatics*, *31*(10), 1680–1682. https://doi.org/10.1093/bioinformatics/btu861

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Tenaillon, M. I., Burban, E., Huynh, S., Wojcik, A., Thuillet, A.-C., Manicacci, D., Gérard, P. R., Alix, K., Belcram, H., Cornille, A., Brault, M., Stevens, R., Lagnel, J., Dogimont, C., Vigouroux, Y., & Glémin, S. (2023). Crop domestication as a step toward reproductive isolation. *American Journal of Botany*, *110*(7), e16173. https://doi.org/10.1002/ajb2.16173

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., Huang, K., Ostevik, K. L., Drummond, E. B. M., Imerovski, I., Lande, K., Pascual-Robles, M. A., Nanavati, M., Jahani, M., Cheung, W., Staton, S. E., Muños, S., Nielsen, R., Donovan, L. A., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), 602–607. https://doi.org/10.1038/s41586-020-2467-6

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co. http://archive.org/details/exploratorydataa00tuke_0

Vijay, N., Bossu, C. M., Poelstra, J. W., Weissensteiner, M. H., Suh, A., Kryukov, A. P., & Wolf, J. B. W. (2016). Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13195

Wakeley, J., & Hey, J. (1997). Estimating ancestral population parameters. *Genetics*, *145*(3), 847–855. https://doi.org/10.1093/genetics/145.3.847

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Wickham, H. (2018). Scales: Scale functions for visualization. R Package Version 1.1.1. https://CRAN.R-project.org/package=scales

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*(2), 87–100. https://doi.org/10.1038/nrg.2016.133

Wu, C.-I. (2001). The genic view of the process of speciation: Genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Zamani, N., Russell, P., Lantz, H., Hoeppner, M. P., Meadows, J. R., Vijay, N., Mauceli, E., di Palma, F., Lindblad-Toh, K., Jern, P., & Grabherr, M. G. (2013). Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics*, *14*(1), 347. https://doi.org/10.1186/1471-2164-14-347

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Burban, E., Tenaillon, M. I., & Glémin, S. (2024). RIDGE, a tool tailored to detect gene flow barriers across species pairs. *Molecular Ecology Resources*, *00*, e13944. https://doi.org/10.1111/1755-0998.13944