Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence

Camille Roux, Christelle Fraïsse, Jonathan Romiguier, Yoann Anciaux, Nicolas Galtier, Nicolas Bierne

PLOS | BIOLOGY

RESEARCH ARTICLE

Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence

Camille Roux[1,2,3]*, Christelle Fraïsse[1,2,4], Jonathan Romiguier[1,2,3], Yoann Anciaux[1,2], Nicolas Galtier[1,2], Nicolas Bierne[1,2]

1 Université Montpellier, Montpellier, France, **2** CNRS Institut des Sciences de l'Évolution, CNRS-UM-IRD-EPHE, Montpellier, France, **3** Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, **4** Institute of Science and Technology, Klosterneuburg, Austria

* camille.roux.1983@gmail.com

OPEN ACCESS

Citation: 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 Biol 14(12): e2000234. doi:10.1371/journal.pbio.2000234

Academic Editor: Craig Moritz, Australian National University, Australia

Received: June 6, 2016

Accepted: November 21, 2016

Published: December 27, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: European Research Council (ERC) https://erc.europa.eu/ (grant number ERC grant 232971). PopPhyl project. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. French National Research Agency (ANR) http://www.agence-nationale-recherche.fr/en/project-based-funding-to-advance-french-research/ (grant number ANR-12-BSV7- 0011). HYSEA project. The funder had no role in study

Abstract

Speciation results from the progressive accumulation of mutations that decrease the probability of mating between parental populations or reduce the fitness of hybrids—the so-called species barriers. The speciation genomic literature, however, is mainly a collection of case studies, each with its own approach and specificities, such that a global view of the gradual process of evolution from one to two species is currently lacking. Of primary importance is the prevalence of gene flow between diverging entities, which is central in most species concepts and has been widely discussed in recent years. Here, we explore the continuum of speciation thanks to a comparative analysis of genomic data from 61 pairs of populations/species of animals with variable levels of divergence. Gene flow between diverging gene pools is assessed under an approximate Bayesian computation (ABC) framework. We show that the intermediate "grey zone" of speciation, in which taxonomy is often controversial, spans from 0.5% to 2% of net synonymous divergence, irrespective of species life history traits or ecology. Thanks to appropriate modeling of among-locus variation in genetic drift and introgression rate, we clarify the status of the majority of ambiguous cases and uncover a number of cryptic species. Our analysis also reveals the high incidence in animals of semi-isolated species (when some but not all loci are affected by barriers to gene flow) and highlights the intrinsic difficulty, both statistical and conceptual, of delineating species in the grey zone of speciation.

Author Summary

Isolated populations accumulate genetic differences across their genomes as they diverge, whereas gene flow between populations counteracts divergence and tends to restore genetic homogeneity. Speciation proceeds by the accumulation at specific loci of mutations that reduce the fitness of hybrids, therefore preventing gene flow—the so-called species barriers. Importantly, species barriers are expected to act locally within the genome, leading to the prediction of a mosaic pattern of genetic differentiation between populations at intermediate levels of divergence—the genic view of speciation. At the same time, linked selection also contributes to speed up differentiation in low-recombining and

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: ABC, approximate Bayesian computation; AM, ancient migration; IM, isolation with migration; PAN, panmixia; SC, secondary contact; SI, strict isolation.

gene-dense regions. We used a modelling approach that accounts for both sources of genomic heterogeneity and explored a wide continuum of genomic divergence made by 61 pairs of species/populations in animals. Our analysis provides a unifying picture of the relationship between molecular divergence and ability to exchange genes. We show that the "grey zone" of speciation—the intermediate state in which species definition is controversial—spans from 0.5% to 2% of molecular divergence, with these thresholds being independent of species life history traits and ecology. Semi-isolated species, between which alleles can be exchanged at some but not all loci, are numerous, with the earliest species barriers being detected at divergences as low as 0.075%. These results have important implications regarding taxonomy, conservation biology, and the management of biodiversity.

Introduction

An important issue in evolutionary biology is understanding how the continuous-time process of speciation can lead to discrete entities—species. There is usually no ambiguity about species delineation when distant lineages are compared. The continuous nature of the divergence process, however, causes endless debates about the species status of closely related lineages [1]. A number of definitions of species have thus been introduced over the 20th century, each of them using its own criteria—morphological, ecological, phylogenetic, biological, evolutionary, or genotypic. A major problem is that distinct markers do not diverge in time at the same rate [2]. For instance, in some taxa, morphological differences evolve faster than the expression of hybrid fitness depression, which in turn typically establishes long before genome-wide reciprocal monophyly [3]. In other groups, morphology is almost unchanged between lineages that show high levels of molecular divergence [4]. The erratic behavior and evolution of the various criteria is such that in a wide range of between-lineage divergence—named the grey zone of the speciation continuum—distinct species concepts do not converge to the same conclusions regarding species delineation [2].

Besides taxonomic aspects, the grey zone has raised an intense controversy regarding the genetic mechanisms involved in the formation of species [5–7]. Of particular importance is the question of gene flow between diverging lineages. How isolated must two gene pools be for speciation to begin? How long does gene flow persist as lineages diverge? Is speciation a gradual process of gene flow interruption or a succession of periods of isolation and periods of contact? These questions are not only central in the speciation literature but also relevant to the debate about species delineation, with the ability of individuals to exchange genes being at the heart of the biological concept of species.

As genomic data have become easier and less expensive to obtain, sophisticated computational approaches have been developed to perform historical inferences in speciation genomics (i.e., estimate the time of ancestral separation in two gene pools, changes in effective population size over evolutionary time, and the history of gene flow between the considered lineages [8–10]). Simulation-based approximate Bayesian computation (ABC) methods are particularly flexible and have recently attracted an increased attention in speciation genomics. One strength of ABC approaches is their ability to deal with complex, hopefully realistic models of speciation and test for the presence or absence of ongoing introgression between sister lineages. This is achieved by simulating molecular data under alternative models of speciation with or without current introgression and choosing among models based on their relative posterior probabilities [11].

Migration tends to homogenize allele content and frequency between diverging populations. This homogenizing effect, however, is often expected to only affect a fraction of the genome. This is because the effective migration rate is impeded in regions containing loci involved in assortative mating, hybrid fitness depression, or other mechanisms of isolation—the so-called genetic barriers [12]. Consequently, gene flow is best identified by models explicitly accounting for among-locus heterogeneity in introgression rates, as demonstrated by a number of recent studies [13–16]. When homogeneous introgression rate across the genome is assumed, distant lineages that have accumulated a large number of genetic barriers can be inferred as currently isolated, whereas they actually exchange alleles at a minority of loci unlinked to barriers [14]. On the other hand, neglecting heterogeneity in introgression rates between closely related lineages can result in a failure to identify some regions of the genome that are already evolving independently [16,17]. Heterogeneous introgression models therefore appear necessary according to the genic view of speciation [18].

Importantly, introgression rates alone do not govern local patterns of genetic differentiation [19]. Linked selective processes, such as hitchhiking effects [20] or background selection [21], are expected to affect the landscape of population differentiation by lowering polymorphism levels at particular loci, especially in low-recombining or gene-dense genomic regions. Neglecting this confounding effect tends to inflate the proportion of false positives in statistical tests of ongoing gene flow [19] and to mislead inferences [22,23]. Linked directional selection is expected to locally increase the stochasticity of allele frequency evolution, a process sometimes coined genetic draft [24]. Its effect can therefore be modeled by assuming that the effective population size, N_e, which determines the strength of genetic drift, varies among loci [25].

Multilocus analyses of the process of population divergence have been achieved in various groups of animals [26,27] and plants [28–30] for which genome-wide data are available, revealing a diversity of patterns. These case studies, however, are limited in number and have taken different approaches, such that we still lack a unifying picture of the prevalence of gene flow during early divergence between gene pools. Here, we gathered a dataset of 61 pairs of populations/species of animals occupying a wide continuum of divergence level. Species were selected in order to sample the phylogenetic and ecological diversity of animals [31], irrespective of any aspect related to population structure or speciation. We investigated the effects of genomic divergence between populations on patterns of gene flow, paying attention to the ability of ABC methods to distinguish between competing models and the influence of model assumptions.

Results

Simulations: ABC as a Powerful Approach to Test for Current Introgression

Five demographic models differing by the history of gene flow between two diverging populations were considered (Fig 1), namely strict isolation (SI), ancient migration (AM), isolation with migration (IM), secondary contact (SC), and panmixia (PAN). The latter three models involve ongoing gene flow between the two populations, whereas the former two do not. The five demographic models were subdivided into different genomic submodels that reflect alternative assumptions about the genomic distribution of indirect selective effects on the effective population size (homoN if homogeneous or heteroN if heterogeneous) and on the migration rate (homoM if homogeneous or heteroM if heterogeneous). Heterogeneous effective population size was considered in all the models, while heterogeneous migration rate was considered in models with gene flow (IM, AM, and SC). The SI and PAN models were divided into two

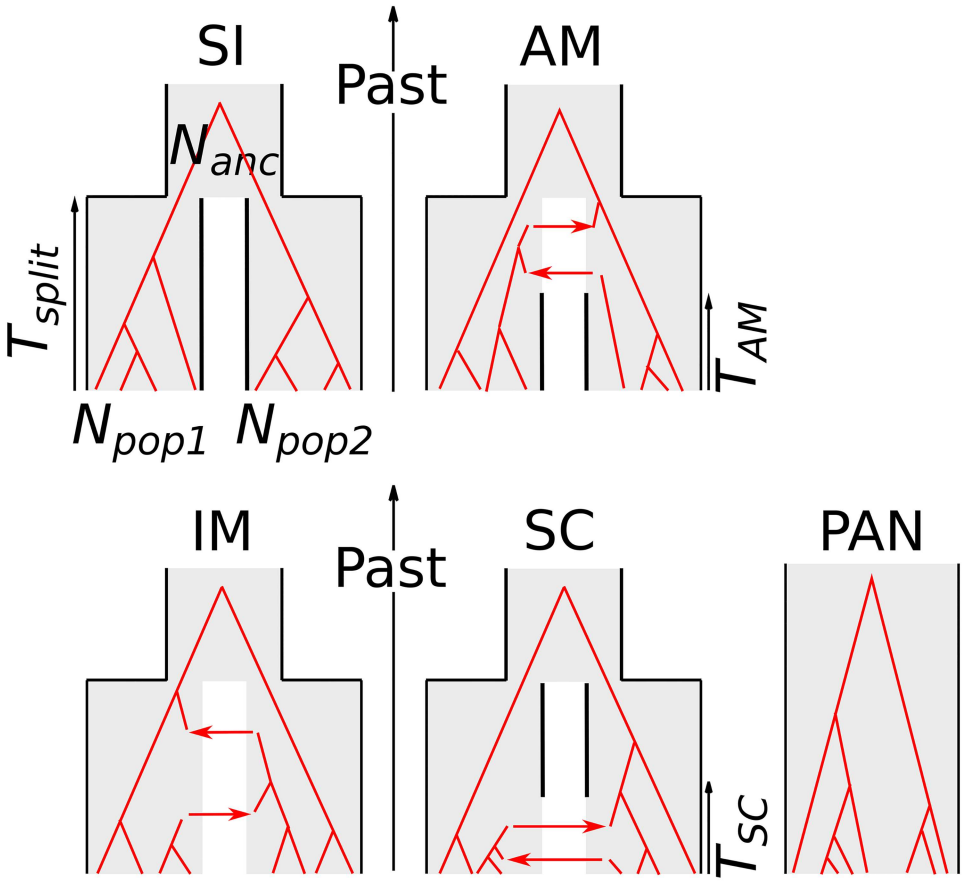

Fig 1. Compared alternative models of speciation. SI = strict isolation: subdivision of an ancestral diploid panmictic population (of size N_{anc}) in two diploid populations (of constant sizes N_{pop1} and N_{pop2}) at time T_{split}. AM = ancestral migration: the two newly formed populations continue to exchange alleles until time T_{AM}. IM = isolation with migration: the two daughter populations continuously exchange alleles until present time. SC = secondary contact: the daughter populations first evolve in isolation (forward in time), then experience a secondary contact and start exchanging alleles at time T_{SC}. PAN: panmictic model. All individuals are sampled from the same panmictic population. Red phylogenies represent possible gene trees under each alternative model.

doi:10.1371/journal.pbio.2000234.g001

submodels (homoN and heteroN), and the AM, IM, and SC models were divided into four submodels (homoN_homoM, homoN_heteroM, heteroN_homoM, and heteroN_heteroM).

The dominant assumption in published demographic inferences is the homoN submodel, in which it is assumed that most of the genetic variation in the genome is unaffected (or equally affected) by selection at linked sites. Here, homoN was simulated using a single value of effective population size shared by all loci across the genome, but the effective population size differed among populations. The heteroN submodel accounts for local genomic effects of directional selection (background selection, selective sweeps) by considering a variable effective population size among loci, here assumed to follow a rescaled beta distribution. The homoM submodel assumes that all loci share the same probability to receive alleles from the sister population (i.e., posits the absence of species barriers or of adaptively introgressed loci). Alternatively, the heteroM submodel accounts for the existence of local barriers to gene flow, of variable strengths, and of variable levels of genetic linkage to the sampled loci. HeteroM was here simulated by assuming that the effective introgression rate is beta distributed across the genome, thus intending to account for the combined effects of selection, recombination, and

gene density. In principle, one could explicitly include information on local recombination rates and gene density, but no such data was available in the species analyzed here.

We explicitly tested the hypothesis of current gene flow by comparing the relative posterior probabilities of 16 models for 61 pairs of species distributed along a continuum of molecular divergence. In the ABC framework, the posterior probability of a model corresponds to its relative ability to theoretically produce datasets similar to the observed dataset, compared to a set of alternative models. Before analyzing datasets from the 61 pairs of animal species, we first assessed the power of the adopted ABC approach to correctly distinguish between models involving current isolation (SI + AM) versus ongoing migration (IM + SC + PAN). This was achieved by randomly simulating 116,000 datasets distributed over the 16 compared models and applying our ABC inference method to each of them. Specifically, we investigated which model had the highest posterior probability and assessed significance by estimating the associated robustness—the probability to correctly support a model given its posterior probability. A robustness greater than 0.95 can be interpreted as a *p*-value below 0.05 [32]. The analysis of simulated datasets allowed us to empirically measure a threshold value of 0.6419 for the posterior probability $P_{migration}$ (= $P_{IM} + P_{SC} + P_{PAN}$), above which the robustness to support ongoing migration is greater than 0.95. Similarly, a posterior probability $P_{migration}$ below 0.1304 implied a statistical support for the current isolation model with a robustness greater than 0.95.

Among the 58,000 simulated datasets in which current gene flow was assumed (IM, SC, and PAN; Fig 2A), 99.462% were true positives ($P_{migration} > P_{isolation}$ and robustness ≥ 0.95), 0.129% were false positives ($P_{migration} < P_{isolation}$ and robustness ≥ 0.95), and 0.409% were ambiguous cases for which ABC did not provide any robust conclusion (robustness < 0.95). Among the 58,000 simulated datasets in which current isolation was assumed (SI and AM; Fig 2B), 99.649% were true positives ($P_{isolation} > P_{migration}$ and robustness ≥ 0.95), 0.002% were false positives ($P_{isolation} < P_{migration}$ and robustness ≥ 0.95), and 0.34% were ambiguous cases (robustness < 0.95). When current gene flow was assumed, the rates of false positive and ambiguity were very low at every level of population divergence. When current isolation was assumed, a higher rate of ambiguity, but no elevation of the rate of false inference, was observed at low levels of divergence ($D_a < 0.01$, Fig 2D). This contrasts with the recent suggestion that the full-likelihood method developed in the IMa2 software [33] might be biased towards supporting current gene flow when isolation is recent [19,34]—our approach appears to be immune from this bias. To specifically address this point, we repeated the exact same simulations as in [34] and confirmed that our ABC approach has a reduced power (i.e., more ambiguous cases with robustness <0.95) when the split is recent but still a very low rate of false positive in these conditions (see S1 text).

In addition, the robustness of the ABC inference was only weakly dependent on the sample size when the number of loci was greater than 100: similar results were obtained when we simulated samples of size 2, 3, 25, or 50 diploid individuals (S1 Fig). Finally, and importantly, simulations showed that ABC is not accurate enough to discriminate between the IM and SC models. Datasets simulated under SC were assigned to SC with high confidence only when the period of isolation before secondary contact represents at least a proportion of about 60% of the total divergence time (S2A Fig). When shorter periods of isolation were simulated, the method either assigned the datasets to IM or did not provide an elevated posterior probability to any demographic model (S2B Fig).

Dataset: Molecular Divergence and Population Differentiation in 61 Taxa

The posterior probability of ongoing gene flow was estimated in 61 pairs of species/populations of animals (S1 Data) showing variable levels of molecular divergence (S1 Data). Fifty

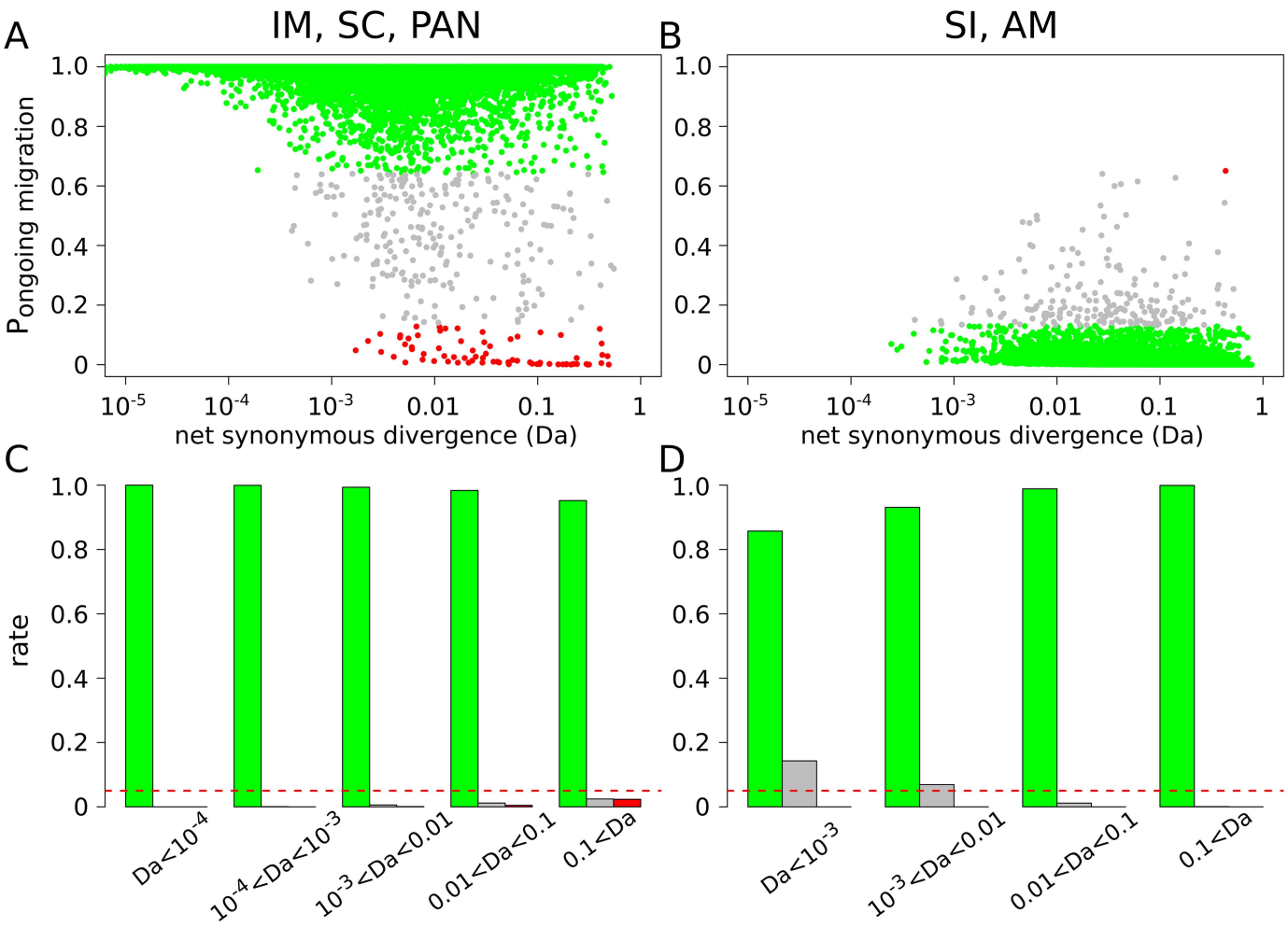

Fig 2. ABC analysis of randomly simulated datasets. Posterior probability $P_{migration}$ to support ongoing migration was estimated for a total of 116,000 simulated datasets across 16 models. A. $P_{migration}$ as a function of the net synonymous divergence D_a. Dots represent datasets simulated under the IM, SC, and PAN models. The colors show datasets for which gene flow is correctly supported (green) or wrongly rejected (red). Grey dots represent datasets for which the robustness of the ABC analysis is <0.95. B. $P_{migration}$ as a function of the net synonymous divergence D_a. Dots represent datasets simulated under the SI or AM models. The colors show datasets for which gene flow is correctly rejected (green; robustness $\geq$ 0.95) or wrongly supported (red; robustness $\geq$ 0.95). C. Proportion of true positives (green), false positives (red), and ambiguous analyses (grey) for different ranges of D_a across IM, SC, and PAN datasets. Horizontal red line shows 5%. D. Proportion of true positives (green), false positives (red), and ambiguous analyses (grey) for different ranges of D_a across SI and AM datasets.

doi:10.1371/journal.pbio.2000234.g002

pairs were taken from a recent transcriptome-based population genomic study [31], with two individuals per population/species being analyzed here. The datasets for the other 11 species pairs were downloaded from the NCBI (S1 Data). They correspond to sequences from published studies using either ABC, Ima [33], or MIMAR [35], for which 3 to 78 diploid individuals were analyzed.

We computed various measures of molecular divergence between species/populations: namely, D_a, the relative average divergence, corrected for within-species diversity [36]; D_{xy}, the absolute average divergence; and F_{ST}, a classical measure of population differentiation. In our dataset, D_a ranged from 5.10^{-5} (French versus Danish populations of *Ostrea edulis*) to 0.309 (*Crepidula fornicata* versus *C. plana*) and F_{ST} from 0 (between *Anas crecca shemya* and *A. crecca attu*) to 0.95 (between *Camponotus ligniperdus* and *C. aethiops*, S3 Fig). As expected, D_a was strongly correlated to F_{ST} and less well to the absolute divergence D_{xy} (S3B Fig). The

across-loci variance in F_{ST} was minimal for low and high values of D_a (S3B Fig), which reflects an F_{ST} homogeneously low at early stages of divergence, homogeneously high at late stages of divergence, and heterogeneous among genes at intermediate levels of D_a (S3 Fig).

Statistical Analysis: Assessment of Ongoing Gene Flow

For each of the 61 studied pairs of populations/species, we focused on synonymous positions and investigated the prevalence of ongoing gene flow by estimating the posterior probabilities of 16 different models under ABC. These 16 models represent the combinations of 5 demographic models (SI, AM, IM, SC, and panmixia) and four assumptions regarding the genomic heterogeneity in introgression (for AM, IM, and SC only) and drift rates (for all models; see above and Material and Methods). The posterior probability $P_{migration}$ that the two populations currently exchange migrants was estimated by summing the contributions of the PAN, IM, and SC models (Fig 1) and plotted against measures of molecular divergence (Fig 3). D_a, which can be understood as the per-site amount of neutral derived mutations being fixed in the different lineages, provided the best relationship (Fig 3). Results with other measures of divergence and with the estimated age of the split (T_{split} parameter under the IM model) are also shown (S4–S7 Figs).

Over the continuum of divergence, the 22 pairs with D_a lower than 0.5% received a support for ongoing gene flow with a robustness $\geq$0.95 (Fig 3). The first identified semipermeable

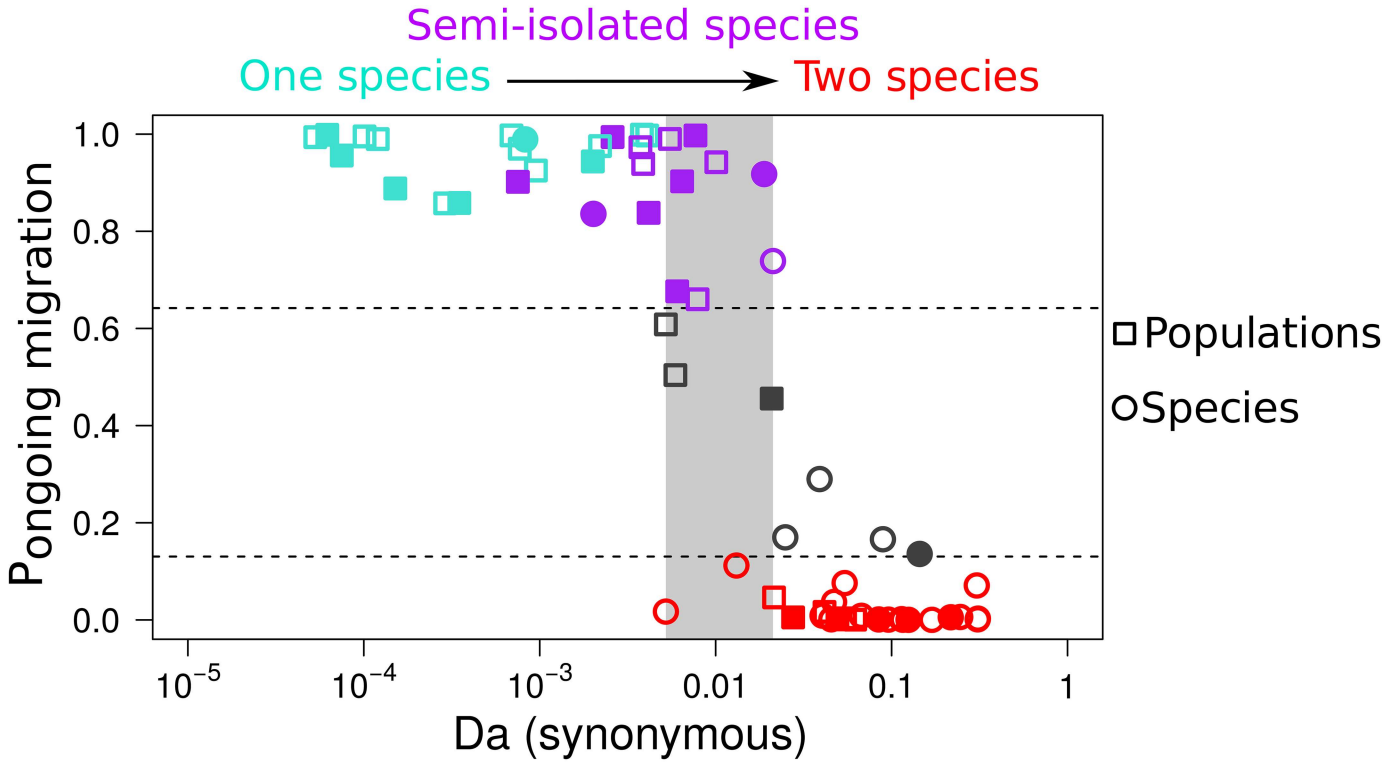

Fig 3. Probability of ongoing gene flow along a continuum of molecular divergence. Each dot is for one observed pair of populations/species. *x*-axis: net molecular divergence D_a measured at synonymous positions (log10 scale) and averaged across sequenced loci. *y*-axis: relative posterior probability of ongoing gene flow (i.e., SC, IM, and PAN models) estimated by ABC. Red dots: pairs with a strong support for current isolation. Grey dots: pairs with no strong statistical support for any demographic model (robustness <0.95). Blue dots: pairs with strong statistical support for genome-homogeneous ongoing gene flow. Purple dots: pairs with strong statistical support for genome-heterogeneous ongoing gene flow. Filled symbols: pairs with a strong support for genome-heterogeneous N_e. Open symbols: genome-homogeneous N_e. The light grey rectangle spans the range of net synonymous divergence in which both currently isolated and currently connected pairs are found (see S1 Data).

doi:10.1371/journal.pbio.2000234.g003

barrier to gene flow was detected at $D_a \approx 0.075\%$, a pair of *Malurus* (fairywren) species [37] for which ABC strongly supports heterogeneity in M. When the net divergence was between 0.5% and 2%, inferences about gene flow were variable and sometimes uncertain. In this grey zone, gene flow was strongly supported in 7 pairs, always with a strong support for genomic heterogeneity in introgression rates. Still, in the grey zone, ABC did not distinguish between isolation and introgression in 3 pairs of species and provided strong support for isolation in 2 other pairs. Finally, among the 27 most divergent pairs of species where D_a was greater than 2%, we found 23 pairs with a strong support for current isolation and 4 ambiguous pairs (Fig 3).

We investigated the impact of assumptions about genomic heterogeneity in N_e and M on the detection of current introgression (S4–S9 Figs). When both parameters were allowed to vary among loci, pairs of populations with D_a exceeding 0.1% and showing strong statistical support for ongoing migration tended to obtain support for genomic heterogeneity in introgression rates. But when constant introgression rate was assumed (homoM_heteroN and homoM_homoN models), the importance of gene flow became underestimated in several divergent pairs of species, consistent with previous reports (e.g. [15]). When we compared models assuming homogeneous versus heterogeneous effective population size across loci, we found that the former tended to overestimate the prevalence of ongoing gene flow (S8 Fig), again in line with published analyses [19]. Analyses assuming homogeneous N_e and M in many cases failed to support either isolation or migration, producing the highest number of ambiguous pairs (S8 Fig). The detected genomic heterogeneity in gene flow increased with D_a until 2% of divergence. Finally, across the whole continuum, there was no significant effect of the divergence on the probability of supporting genomic heterogeneity in effective population size in our dataset.

No Effect of Habitat, Geography, Phylogeny, or Life History Traits

We investigated the influence of a number of ecological, geographical, phylogenetic, and life history variables on the posterior probability of ongoing gene flow. This was achieved under the heteroM_heteroN model using data from [31]. We detected no significant effect of species longevity or log-transformed propagule size (size of the developmental stage that leaves the mother and disperses) on the log-transformed probability of ongoing gene flow. In the same vein, marine organisms ($n = 25$) did not exhibit a higher propensity for ongoing gene flow than terrestrial ones ($n = 36$; r^2 below 0.01%). The log-transformed probability of ongoing gene flow was significantly higher (p-value = 0.002, $r^2 = 0.14$) in vertebrates ($n = 20$) than in invertebrates ($n = 41$), but the effect disappeared when the level of divergence was controlled for (net synonymous divergence <0.04: 17 vertebrate pairs, 22 invertebrate pairs, $p = 0.32$, $r^2 = 0.03$). This effect only reflects the paucity of pairs of vertebrate population/species with a high divergence in our dataset. Finally, we tested whether the current geographic distribution of species coincides with the establishment of genetic structure in our data by distinguishing pairs in which the two considered species/populations occur on the same versus distinct continents or oceans. We did not find any significant effect of this variable on the estimated values of $P_{\text{migration}}$ in either of the three divergence zones: $D_a < 0.5\%$, t test = –0.015269, df (degrees of freedom for the t-statistic) = 18.522, p-value = 0.988; $0.5\% < D_a < 2\%$, t test = –0.74229, df = 7.1996, p-value = 0.4814; $2\% < D_a$, t test = 0.35512, df = 22.426, p-value = 0.7258.

Ongoing Gene Flow and Taxonomic Status

Finally, we verified whether our inferences confirmed or contradicted the current taxonomy (S1 table). Our dataset comprises 26 pairs of recognized species and 35 pairs of populations (or subspecies) sharing a common binomen. Twenty-one pairs of recognized species belonged to

the high-divergence zone ($D_a > 0.02$). Of these, 16 were inferred to be currently isolated, 4 produced ambiguous results and 1 pair, *Eunicella cavolinii* versus *E. verrucosa* (gorgonian), was found to be connected by heterogeneous gene flow. Among the 5 remaining recognized pairs of species (with $D_a < 0.02$), 2 were inferred as being fully isolated and 3 were inferred to be connected species: 2 pairs of semi-isolated species with heterogeneous gene flow (*Mytilus galloprovincialis* versus *M. edulis* and *Macaca mulatta* versus *M. fascicularis*) and the *Gorilla gorilla* versus *G. beringei* pair, which was found to be connected by homogeneous gene flow. Of the 35 pairs of recognized populations from the same species, 6 with $D_a > 0.02$ were inferred to be isolated cryptic species. Genetic isolation has been previously suspected between northern and southern populations of *Pectinaria koreni* (trumpet worms) [38], between the blue and purple morphs of *Cystodytes dellechiajei* (colonial ascidians) [39], and between the L1 and L2 lineages of *Allolobophora chlorotica* (earthworms) [40], but genetic isolation is here newly revealed between Morrocan and European populations of *Melitaea cinxia* (Glanville fritillary), between Spanish and French populations of *A. chlorotica* L2, and between Mediterranean and tropical populations of *Culex pipiens*.

Discussion

We performed a comparative speciation genomics analysis in 61 pairs of populations/species from various phyla of animals. Our ABC analysis, which takes into account the confounding effect of linked selection heterogeneity, provides a first global picture of the prevalence of gene flow between diverging gene pools during the transition from one to two species.

Accounting for Among-Locus Heterogeneity in Drift and Migration Rate

Inferring the history of divergence and gene flow, which determines the rate of accumulation of species barriers, is of prime importance to understand the process of speciation [17]. This can be achieved by various methods, among which ABC approaches have proven particularly flexible and helpful to compare alternative evolutionary models. Our analysis of simulated datasets illustrates that ABC methods have the power to effectively discriminate recent introgression versus current isolation based on datasets of several hundreds of loci and a few individuals per species—typical of population genomic studies. Comparisons of alternative demographic models, however, can be strongly impacted by assumptions regarding the genomic distribution of effective population size (N_e) and introgression rate (M). Heterogeneities in N_e and M are common in natural populations as a result of selective processes applying either globally (background selection [19,41,42]) or specifically against migrants (genetic barriers [12,43]).

Following [13], we here introduced a framework in which each of the two effects, or both, can be readily accounted for. In our analysis, the number of pairs of populations/species for which ambiguous conclusions were reached was maximal when genomic heterogeneities of both migration and drift were neglected. Incorporating within-genome variation in N_e tended to enhance the support for models with current isolation, as previously suggested [19]. The heteroN model makes a difference regarding inference of current gene flow between the highly divergent *Ciona intestinalis* and *C. robusta* species (see below). Conversely, incorporating heterogeneity in M doubled the number of pairs for which ongoing gene flow was supported when compared to analyses with homogenous M, in which most of these pairs exhibited ambiguous results. Our study therefore underlines the importance of accounting for genomic heterogeneities for both N_e and M when comparing alternative models of speciation [14,15,19] and calls for prudence regarding the conclusions to be drawn from the analysis of a single pair. However, it is important to recall here that the action of natural selection on its molecular

target and neighborhood is more complex than a simple reduction in N_e. Our modeling of genomic heterogeneity in drift and selection by a beta distribution of N_e throughout the genome is an approximation which cannot replace an explicit modeling of these processes. In our modeling, we assumed that a given locus *i* is independently affected by drift and selection in all of the simulated populations including the ancestral one. Our choice was motivated by the generality of this model. An alternative approach to model genomic heterogeneity in N_e can be to assume that background selection is the main process shaping genomic landscapes of diversity. This can be approximated by assuming that a locus *i* is equally affected by drift and selection in all populations instead of assuming independent effects as in our study.

Among models assuming ongoing gene flow, our ABC analysis of simulated and empirical data often failed to discriminate between the isolation with migration and secondary contact models. These two models yield similar signatures in genetic data, such that only relatively recent secondary contacts following long periods of interrupted gene flow can be detected with high confidence (S2D Fig) [44]. Similarly, among models excluding ongoing gene flow, distinguishing between strict isolation and ancient migration was not possible in a substantial number of cases. These are challenges for future methodological research in the field, with important implications regarding the debate about the requirement of geographic isolation to complete speciation [7,45]. Only two diploid individuals per population/species were used in this analysis for the sake of comparability between datasets (in many populations, no more than two individuals were available) and because of computational limitation. However, our evaluation of the effect of sample size on ABC-based demographic inference suggested that two diploid individuals per population were largely sufficient to capture the main signal when more than 100 loci are available (S1 Fig).

Prevalent Gene Flow between Slightly Diverged Gene Pools

Although ABC analyses of particular pairs of populations can be affected by the choice of model of genomic heterogeneity, the overall relationship between net molecular divergence and detected ongoing gene flow was qualitatively similar among analyses. Pairs of populations diverging by less than 0.5% were found to currently exchange migrants. This includes populations that form a single panmictic gene pool and pairs of diverging populations/species connected by gene flow. The low-divergence area contains pairs of populations showing conspicuous morphological differences, such as eastern versus western gorilla or the *cuniculus* and *algirus* subspecies of rabbit (*Oryctolagus cuniculus*).

No pair of populations in this range of divergence was supported to be genetically isolated or yielded ambiguous results. Simulations indicate that our ABC approach is not expected to yield false inference of gene flow in recently isolated populations, contrary to what was suggested with the full-likelihood approach of IMa2 [34]. The main risk is rather a false inference of isolation despite gene flow (Fig 2), which can be explained by the fact that the SI model is less parameterized than models assuming gene flow (IM and SC). ABC had a low false positive rate even when we simulated very recent splits, as has been done in previous papers [19,34]. This is probably because in strict isolation, shared polymorphisms are quickly sorted into private polymorphisms and fixed differences after population split, such that D_a can hardly be very small in the absence of gene flow [46]. Our analysis therefore identifies $D_a < 0.5\%$ as a good synthetic proxy to attest for the existence of gene flow. Other measures of divergence, although producing a qualitatively similar pattern, did not predict the existence of current gene flow as nicely as D_a did.

Pairs in the low range of divergence must correspond to populations that did not accumulate sufficiently strong and numerous genetic barriers, such that gene flow currently occurs at

important rates. The detection of significantly heterogeneous introgression rates in a number of low-diverged pairs ($D_a < 0.5\%$) demonstrates the ability of our ABC approach to detect semipermeable barriers quite efficiently at early stages of speciation and supports the rapid evolution of Dobzhansky–Muller incompatibilities [47,48]. A majority of the pairs from the low-divergence area, however, did not yield any evidence for among-locus heterogeneity of introgression rate. Some might correspond to effectively isolated backgrounds that are missed by our method by lack of power when the signal of heterogeneity is too tenuous. It is quite plausible, however, that some pairs of populations/species in the low-divergence zone have differentially fixed mutations with major effects on hybrid fitness, whereas others do not because of mutational stochasticity and/or across-taxa differences in the genetic architecture of barriers—i.e., simple (two locus) versus complex incompatibilities and strength of associated selective effects [49].

Suppressed Gene Flow at High Sequence Divergence

At the other end of the continuum, it appears that above a divergence of a few percent, barriers are strong enough to completely suppress gene flow: almost all pairs of species with $D_a > 2\%$ were found to have reached reproductive isolation with strong support. This might result from impaired homologous recombination because of improper pairing of dissimilar homologous chromosomes at meiosis, which would reduce the fecundity of hybrids [50,51]. Of note, the upper threshold for reproductive isolation ($D_a = 2\%$, $D_{x\,y} = 5.5\%$) is of the order of magnitude of the maximal level of within-species genetic diversity reported in animals [31,52], somewhat consistent with the hypothesis of a physical constraint imposed by sequence divergence on the ability to reproduce sexually. Alternatively, the 2% figure may represent a threshold above which Dobzhansky–Muller incompatibilities are normally in sufficient number and strength to suppress introgression. The two hypotheses are not mutually exclusive but pertain to distinctive processes of genetic isolation; the former would be maximally expressed during F1 hybrid meiosis, while the latter would affect recombined, mosaic individuals carrying alleles from the two gene pools at a homozygous state.

In the high-divergence area, no instance of among-locus heterogeneous migration was detected, indicating that introgression is blocked across the whole genome in these pairs of species. A number of highly divergent species pairs yielded support for among-locus heterogeneous N_e, suggesting that the same regions of the genome are under strong background selection in the two diverging entities—presumably regions of reduced recombination and/or high density in functional elements. Neglecting the genomic heterogeneity in N_e can lead to false inference of gene flow. For instance, allowing genomic heterogeneity in M but not in N_e led to strong statistical support for a secondary contact between the highly divergent *Ciona intestinalis* (formerly *C. intestinalis B*) and *C. robusta* (formerly *C. intestinalis* A) species (S4 and S5 Figs), consistent with [14], but accounting for heterogeneity in both M and N_e resulted in an ambiguous result without a sufficiently strong support for any models. The among-locus variance in differentiation between these two species, which was interpreted as mainly reflecting introgression at a few loci in [14], is shown here to possibly be the result of a more complex situation that our models failed to capture.

Intermediate Divergence Levels: The Grey Zone of Speciation

The area of intermediate divergence from 0.5% to 2% of net synonymous divergence unveils the grey zone of the speciation continuum. In this grey zone, isolated pairs of populations/species coexist with pairs connected by migration, and the latter are mainly composed of semi-isolated genetic backgrounds, the situation under which taxonomic conundrums flourish. Cases

of ambiguous conclusions about the demographic history also tended to be found in this intermediate zone, perhaps reflecting instances of complex divergence models that are not well predicted by our demographic models. Researchers should be ready to face problems regarding demographic inference—and therefore parameter estimation—when conducting a project of speciation genomics falling in the grey zone. Accounting for genomic heterogeneity of introgression and drift rates appears to be crucial for detecting current gene flow in this range of divergence (S4–S7 Figs). For instance, the mussel species *M. galloprovincialis* versus *M. edulis* and the gorgonian species *Eunicella cavolinii* versus *E. verrucosa* are the two most divergent pairs for which ongoing introgression was detected, but this only appeared when the genomic variation in *M* was accounted for—the homoM_homoN and homoM_heteroN models yielded ambiguous conclusions about these pairs of species, in which the existence of semipermeable barriers has previously been demonstrated [53,54].

Our analysis revealed significant among-locus heterogeneous migration in as many as thirteen pairs of populations/species (Fig 3). This illustrates the commonness of semipermeable genomes at intermediate levels of speciation, when some, but not all, genomic regions are affected by barriers to gene flow. Besides mussels and gorgonians, heterogeneous gene flow was newly detected between American and European populations of *Armadillidium vulgare* (wood lice) and *Artemia franciscana* (brine shrimp), between Atlantic and Mediterranean populations of *Sepia officinalis* (cuttlefish), and between the closely related *Eudyptes chrysolophus moseleyi* versus *E. c. filholi* (penguins) and *Macaca mulatta* versus *M. fascicularis* (macaques)—in addition to the previously documented mouse [55], rabbit [56], and fairywren [57] cases.

The grey zone, finally, includes populations between which unsuspected genetic isolation was here revealed, such as the Moroccan versus European populations of *Melitaea cinxia* (Glanville fritillary) and the Spanish versus French populations of *A. chlorotica* L2 (earthworm), which according to our analysis correspond to cryptic species. Our genome-wide approach and proper modeling of heterogeneous processes therefore clarified the status of a number of pairs from the grey zone, emphasizing the variety of situations and the conceptual difficulty with species delineation in this range of divergence.

Implications for Speciation and Conservation Research

Our dataset is composed of a large variety of taxa with deep phylogenetic relationships and diverse life history traits. In principle, the propensity to evolve prezygotic barriers might differ between groups of organisms (e.g., broadcast spawners versus copulating species [58]). We did not detect any significant effect of species biological and/or ecological features or taxonomy on the observed pattern. Highly polymorphic broadcast spawners and low-diversity large vertebrates with strong parental investment were equally likely to undergo current gene flow for a given divergence level. Whether the pace of accumulation of genetic barriers, the so-called speciation clock, varies among taxonomic group is a major challenge in speciation research and requires the dissection of the temporal establishment of barriers in many different taxa [59,60]. State-of-the-art ABC methods offer the opportunity to investigate the genome-wide effect of barriers to gene flow in natural populations but cannot provide answers about how and why barriers have evolved. However, our report of a strong and general relationship between molecular divergence and genetic isolation across a wide diversity of animals suggests that, at the genome level, speciation operates in a more or less similar fashion in distinct taxa, irrespective of biological and ecological particularities.

Interestingly, we did not detect any significant effect of geographic range overlap. This result may appear as unexpected at first sight because one expects gene flow to be dependent on geography. One explanation could be that we used a too crude measure of range overlap.

Alternatively, this result could support the idea that in many taxa, the observed genetic structure was established in the past in a geographic context different from the current one and only recently reshuffled by recent migration and/or colonization processes [61]. According to this hypothesis, genetic subdivision could have little to do with contemporary connectivity.

The width of the grey zone indicates that a number of existing taxonomic debates regarding species definition and delineation are difficult by nature and unlikely to be resolved through the analysis of a limited number of loci. Most of the molecular ecology literature, however, is based on datasets consisting of mitochondrial DNA and rarely more than a dozen microsatellite loci. The time when genome-wide data will be available in most species of interest is approaching, though not yet reached. Since then, we have to accept that knowledge about the existence of gene flow between diverged entities could not be settled from genetic data alone in a substantial fraction of taxa. In addition, our study highlights the commonness of semi-isolated entities, between which gene flow can be demonstrated but only concerns a fraction of loci, further challenging the species concept. We should therefore be prepared to make decisions regarding conservation and management of biodiversity in absence of well-defined species boundaries.

Materials and Methods

All of the informatic codes, data and command lines used to produce the analysis are openly available online in the following GitHub repository: https://github.com/popgenomics/popPhylABC.

Taxon Sampling

A total of 61 pairs of populations/species of animals were analyzed (S1 Data). These include 10 pairs taken from the speciation literature and 51 pairs newly created here based on a recently published RNAseq dataset [31], which includes 96 species of animals from 31 distinct families and eight phyla, and 1 to 11 individuals per species. Twenty-nine of the newly created pairs corresponded to distinct populations within a named species. Populations were here defined based on a combination of geographic, ecotypic, and genetic criteria: we contrasted groups of individuals (i) living in allopatry and/or differing in terms of their ecology and (ii) clustering as distinct lineages in a neighbor-joining analysis of genetic distances between individuals. The 2 most covered individuals per population were selected for ABC analysis. In 4 species, 3 distinct populations were identified, in which case the three possible pairwise comparisons were performed. Results were qualitatively unchanged when we kept a single pair per species. Twenty-two of the newly created pairs consisted of individuals from 2 distinct named species that belonged to the same family. Again, the 2 most covered individuals per species were selected for analysis. In the case of species in which several populations had been identified, we chose to sample 2 individuals from the same population for between-species comparison. When more than 2 species from the same family were available, we selected a single pair based on a combination of sequencing coverage and genetic distance criteria, with comparisons between closely related species being favored. Raw and final datasets are available from the PopPhyl website (http://kimura.univ-montp2.fr/PopPhyl/). Sample sizes, number of loci, and source of data are listed in S1 Data.

Transcriptome Assembly, Read Mapping, and Coding Sequence Prediction

For the 51 recently obtained pairs, Illumina reads were mapped to predicted cDNAs (contigs) with the BWA program [62]. Contigs with a per-individual average coverage below ×2.5 were

discarded. Open reading frames (ORFs) were predicted with the Trinity package [63]. Contigs carrying no ORF longer than 200 bp were discarded. In contigs including ORFs longer than 200 bp, 5′ and 3′ flanking noncoding sequences were deleted, thus producing predicted coding sequences that are hereafter referred to as loci.

Calling Single Nucleotide Polymorphisms (SNPs) and Genotypes

At each position of each locus and for each individual, diploid genotypes were called using the reads2snps program [64]. This method first estimates the sequencing error rate in the maximum-likelihood framework, calculates the posterior probability of each possible genotype, and retains genotypes supported at >95% if ten reads per position and per individual were detected. Possible hidden paralogs (duplicated genes) were filtered using a likelihood ratio test based on explicit modeling of paralogy. For our demographic inferences, only synonymous positions were retained. Synonymous length and positions were then computed for each loci using polydNdS [65].

Summary Statistics

For all of the 61 pairs of populations/species, we calculated an array of 31 statistics widely used for demographic inferences [32,35,66,67]: the average and standard variation over loci for (1) the number of biallelic positions; (2) the number of fixed differences between the two gene pools; (3) the number of polymorphic sites specific to each gene pool; (4) the number of polymorphic sites existing in both gene pools; (5) Wald and Wolfowitz statistics [68]; (6) Tajima's pi [69]; (7) Watterson's theta [70]; Tajima's *D* for each gene pool [71]; (8) the gross divergence between the two gene pools (D_{xy}); (9) the net divergence between the two gene pools (D_a); (10) F_{ST} measured by $1\text{-}p_W/p_T$, where p_W is the average allelic diversity based on the two gene pools and p_T is the total allelic diversity over the two gene pools; and (11) the Pearson's R correlation coefficient in *p* calculated between the two gene pools. Observed values of summary statistics are summarized for each species in S2 Data.

Demographic Models

Five distinct demographic models were considered: PAN, SI, AM, IM, and SC. (Fig 1). The PAN model assumes that the two investigated gene pools are sampled from a single panmictic population of size N_e sampled in the uniform prior [0–5,000,000] individuals. The SI model describes the subdivision of an ancestral panmictic population of size N_{anc} in two isolated gene pools of sizes $N_{pop\text{-}1}$ and $N_{pop\text{-}2}$. The two sister gene pools then evolve in absence of gene flow. Under the IM model, the two sister gene pools that split T_{split} (sampled in the uniform prior [0–10,000,000]) generations ago continuously exchange alleles as they diverge. Under the AM model, gene flow occurs between T_{split} and a more recent T_{AM} date sampled from the uniform prior [0–T_{split}], after which the two gene pools evolve in strict isolation. The SC model assumes an early divergence in strict isolation followed by a period of gene flow that started T_{SC} generations ago with T_{SC} sampled from the uniform prior [0–T_{split}].

Heterogeneity in Introgression and Effective Population Size

We assumed that the effects of selection on linked sites can be described in terms of heterogeneous effective population size (putatively affecting all demographic models) and/or migration rate (only affecting the IM, AM, and SC models). In the homoM setting, one gene flow parameter ($M = N.m$) is randomly sampled from a uniform prior distribution for each direction. M_1 is the direction from gene pool 2 to gene pool 1 and M_2 is the direction from gene pool 1 to

gene pool 2. All loci share the same M_1 and M_2 values, but M_1 and M_2 are independently sampled. In the heteroM setting, a specific migration rate is attributed per locus and per direction of migration. Thus, for each direction, a hyperprior is first randomly designed as a beta distribution. A value of $M_{1,i}$ and $M_{2,i}$ is then drawn for each loci i from the two hyperpriors. In the homoN setting, the effective population sizes N_{anc} (ancestral population), $N_{pop\text{-}1}$ (gene pool 1) and $N_{pop\text{-}2}$ (gene pool 2) are independent but shared by all loci. In the heteroN setting, heterogeneity in effective population size is independently modeled for the three populations (ancestor, gene pool 1, and gene pool 2). For each population, a proportion a of loci is assumed to evolve neutrally and share a common value for N_{anc}, $N_{pop\text{-}1}$, or $N_{pop\text{-}2}$, a being sampled from the uniform prior [0–1]. The remaining loci, in proportion 1-a, are assumed to be affected by natural selection at linked loci. They are assigned independent values of N, which are sampled from beta distributions defined on the intervals [0–N_{anc}], [0–$N_{pop\text{-}1}$], and [0–$N_{pop\text{-}2}$]. In this setting, a and N_e differ between the three populations but are sampled from distributions sharing the same shape parameters.

Approximate Bayesian Computation

The combination of demographic models and genomic settings resulted in a total of 16 distinct models, namely the homoN and heteroN versions of PAN and SI and the homoM_homoN, homoM_heteroN, heteroM_homoN, heteroM_heteroN versions of IM, AM, and SC. Model fit assessment and parameter estimation were performed under the ABC framework. Under each model, 3,000,000 multilocus simulations were conducted using the coalescent simulator msnsam, a modified version of ms allowing variation across loci of the number of sampled individuals [66,72]. For each of the 61 pairs of populations/species, the posterior probability of each model was estimated using a feed-forward neural network implementing a nonlinear multivariate regression by considering the model itself as an additional parameter to be inferred under the ABC framework using the R package "abc" [73]. The 10,000 replicate simulations (out of 16 x 3,000,000) falling nearest to the observed values of summary statistics were selected, and these were weighted by an Epanechnikov kernel that peaks when $S_{obs} = S_{sim}$. Computations were performed using 50 trained neural networks and 10 hidden networks in the regression. The posterior probability of each model was obtained by averaging over ten replicated ABC analyses.

Robustness

Among a set of compared models, ABC returns a best-supported model M and its posterior probability P_M. The returned model is validated when P_M is above an arbitrary threshold X, corresponding to the posterior probability above which the statistical support for a model is considered as being significant. The robustness of the inference—i.e., the probability to correctly support model M if true—obviously depends on X. To assess the reliability of our approach, we randomly simulated 116,000 pseudo-observed datasets (PODs) distributed over the 16 compared models. Simulations were independent of the 3,000,000 x 16 reference simulations used for model comparisons in our main analysis, but their parameters share the same boundaries.

For each simulated POD, we estimated the posterior probabilities P_i of the 16 compared models through ABC. The probability of correctly supporting M given X was calculated as: $P(P_M > X \mid M) \,/\, [\Sigma_1^m P(P_M > X \mid i)]$, where $P(P_M > X \mid i)$ is the probability that a dataset simulated under m will be supported by ABC as being M with a posterior probability above X [32]. This is the proportion (among simulated datasets inferred by ABC to correspond to M) of those actually generated under M.

For the "ongoing gene flow" versus "current isolation" model comparison, we empirically measured that robustness to support gene flow starts to be above 0.95 if $P_{migration} \geq 0.6419$ and the robustness to support isolation is above 0.95 if $P_{migration} \leq 0.1304$. For datasets with $P_{migration}$ between 0.1304 and 0.6419, we did not attribute a best model but treated them as "ambiguous cases."

Supporting Information

S1 Fig. Effects of the number of sampled individuals on robustness of model comparisons when 100 loci are investigated. Analyses were made by simulating four different datasets:

A-B: 100 loci sampled in two diploid individuals in each daughter species.

C-D: 100 loci sampled in three diploid individuals in each daughter species.

E-F: 100 loci sampled in 25 diploid individuals in each daughter species.

G-H: 100 loci sampled in 50 diploid individuals in each daughter species.

Panels on the left border show the distributions of *P*(current isolation | current isolation) (white bars) and *P*(current introgression | current introgression) (grey bars) measured after ABC analysis of 20,000 PODs simulated under each models. Panels on the right border show the distributions of *P*(SI | SI) (black lines), *P*(AM | AM) (red lines), *P*(IM | IM) (blue lines) and *P*(SC | SC) (green bars) measured after ABC analysis of 20,000 PODs simulated under each models.
(TIF)

S2 Fig. Effect of parameter combinations on the correct support of the SC model. A. Two-dimensional space of parameters of the SC model showing simulations leading to a correct support of SC (i.e $P(SC \mid SC) > 0.8$). X-axis represents the time since the ancestral split. Y-axis represents the relative time the two daughter species remained isolated before the secondary contact. Colors represent the density in simulations with $P(SC \mid SC) > 0.8$. B. Two-dimensional space of parameters of the SC model showing simulations leading to the absence of a robust conclusion using ABC. Colors represent the density in simulations with *P*(NA | SC).
(TIF)

S3 Fig. Relation between synonymous divergence and genetic differentiation. Each grey dot represents a pair of species/populations. Lepus (Spanish and Portuguese populations of *Lepus granatensis*), Eunicella (*Eunicella cavolinii* and *E. verrucosa*) and Crepidula (*Crepidula fornicata* and *Bostrycapulus aculeatus*) indicate representative pairs of poorly, intermediately and highly divergent species/populations. Effect of divergence on across-loci variance in F_{ST}. Genomic distribution of F_{ST} for the Lepus, Eunicella and Crepidula datasets (see S1 Data).
(TIF)

S4 Fig. Relation between net synonymous divergence D_a and probability of ongoing gene flow. Net synonymous divergence is the average proportion of differences at synonymous positions between individuals sampled in the two compared species due to mutations occurring after the ancestral split. The "hetero M + *Ne*" analysis was made by assuming genomic variation for both *M* and *Ne*. The "hetero M" analysis solely takes into account genomic variation in introgression rates over the whole genome. The "hetero *Ne*" analysis solely takes into account genomic variation in *Ne*. The "homo M + *Ne*" analysis considers one value of *M* and one value of *Ne* shared by the whole genome. Red arrows indicate pairs of species inferred as ambiguous in heteroM (robustness < 0.95), heteroNe and homoM_homoN analysis but not in heteroM_heteroN (robustness ≥ 0.95). Green arrows indicate pairs of species with different

and unambiguous inferences (robustness $\geq$ 0.95) made in heteroM, heteroNe and homoM_homoN when compared to heteroM_heteroN (see S1 Data).
(TIF)

S5 Fig. Relation between gross synonymous divergence D_{xy} and probability of ongoing gene flow. Gross synonymous divergence is the average proportion of differences at synonymous positions between individuals sampled in the two compared species, including differences present in the ancestral species. The "hetero M + *Ne*" analysis was made by assuming genomic variation for both *M* and *Ne*. The "hetero M" analysis solely takes into account genomic variation in introgression rates over the whole genome. The "hetero *Ne*" analysis solely takes into account genomic variation in *Ne*. The "homo M + *Ne*" analysis considers one value of *M* and one value of *Ne* shared by the whole genome. Red arrows indicate pairs of species inferred as ambiguous in heteroM (robustness $<$ 0.95), heteroNe and homoM_homoN analysis but not in heteroM_heteroN (robustness $\geq$ 0.95). Green arrows indicate pairs of species with different and unambiguous inferences (robustness $\geq$ 0.95) made in heteroM, heteroNe and homoM_homoN when compared to heteroM_heteroN (see S1 Data).
(TIF)

S6 Fig. Relation between F_{ST} and probability of ongoing gene flow. The "hetero M + *Ne*" analysis was made by assuming genomic variation for both *M* and *Ne*. The "hetero M" analysis solely takes into account genomic variation in introgression rates over the whole genome. The "hetero *Ne*" analysis solely takes into account genomic variation in *Ne*. The "homo M + *Ne*" analysis considers one value of *M* and one value of *Ne* shared by the whole genome. Red arrows indicate pairs of species inferred as ambiguous in heteroM (robustness $<$ 0.95), heteroNe and homoM_homoN analysis but not in heteroM_heteroN (robustness $\geq$ 0.95). Green arrows indicate pairs of species with different and unambiguous inferences (robustness $\geq$ 0.95) made in heteroM, heteroNe and homoM_homoN when compared to heteroM_heteroN (see S1 Data).
(TIF)

S7 Fig. Relation between the estimated T_{split} under the IM model and probability of ongoing gene flow. The "hetero M + *Ne*" analysis was made by assuming genomic variation for both *M* and *Ne*. The "hetero M" analysis solely takes into account genomic variation in introgression rates over the whole genome. The "hetero *Ne*" analysis solely takes into account genomic variation in *Ne*. The "homo M + *Ne*" analysis considers one value of *M* and one value of *Ne* shared by the whole genome. Red arrows indicate pairs of species inferred as ambiguous in heteroM (robustness $<$ 0.95), heteroNe and homoM_homoN analysis but not in heteroM_heteroN (robustness $\geq$ 0.95). Green arrows indicate pairs of species with different and unambiguous inferences (robustness $\geq$ 0.95) made in heteroM, heteroNe and homoM_homoN when compared to heteroM_heteroN.
(TIF)

S8 Fig. Number of pair of species supporting current isolation, current introgression, or ambiguity in model choice. A pair of species is associated to "current isolation" if the sum of posterior probabilities P(SI) + P(AM) is associated to a robustness $\geq$ 0.95. A pair of species is associated to "current introgression" if the sum of posterior probabilities P(SC) + P(IM) is associated to a robustness $\geq$ 0.95. The ambiguous status is attributed to a pair of species when "current isolation" and "current introgression" are not strongly supported. The "homo M + N" analysis was made by assuming an unique genomic introgression rate and an unique *Ne* over the whole genome. The "hetero M" analysis takes into account genomic variation in introgression rates over the whole genome. The "hetero N" analysis takes into account

genomic variation in *Ne*. The "hetero M + N" analysis takes into account genomic variation in introgression rates and in *Ne* (see S1 Data).
(TIF)

S9 Fig. Number of pair of species showing evidences for SI, AM, IM, SC, PAN, or ambiguity in model choice for three distinct ABC analyses. A pair of species is associated to SI or AM if its relative posterior probability is greater than 0.8696. A pair of species is associated to IM, SC or PAN tf its relative posterior probability is greater than 0.6419. The "homo M + N" analysis was made by assuming an unique genomic introgression rate and an unique *Ne* over the whole genome. The "hetero M" analysis takes into account genomic variation in introgression rates over the whole genome. The "hetero N" analysis takes into account genomic variation in *Ne*. The "hetero M + N" analysis takes into account genomic variation in introgression rates and in *Ne* (see S1 Data).
(TIF)

S10 Fig. Estimating α, the proportion of loci that introgress, under the IM model. 2,000 pseudo-observed datasets (PODs) were simulated under the IM model with heterogeneity in introgression rates. We estimated the parameters of this model by using the ABC approach described in the 'Materials and Methods' section. α is the proportion of the genome crossing the species barrier at a rate $N.m > 0$. x-axis: values of α used to produce the PODs; y-axis: values of α estimated by ABC from the simulated PODs. Solid line represents f(x) = x. Dotted lines represent f(x) = 2.x and f(x) = x/2 respectively. Estimated values of α for the observed pairs of population/species as a function of their net synonymous divergence.
(TIF)

S11 Fig. Estimating *N.m*, the effective migration rate, under the IM model. 2,000 pseudo-observed datasets (PODs) were simulated under the IM model with heterogeneity in introgression rates. A. x-axis: values of *N.m* used to produce the PODs; y-axis: values of *N.m* estimated by ABC from the simulated PODs. Solid line represents f(x) = x. Dotted lines represent f(x) = 2.x and f(x) = x/2 respectively. B. Estimated values of *N.m* for the observed pairs of population/species as a function of their net synonymous divergence.
(TIF)

S12 Fig. Estimating *N*, the effective population size of daughter populations, under the IM model. 2,000 pseudo-observed datasets (PODs) were simulated under the IM model with heterogeneity in introgression rates. A. x-axis: values of N used to produce the PODs; y-axis: current values of N estimated by ABC for all PODs. Solid line represents f(x) = x. Dotted lines represent f(x) = 2.x and f(x) = x/2 respectively. B. Estimated values of N for the observed pairs of population/species as a function of their net synonymous divergence.
(TIF)

S13 Fig. Estimating *Nanc*, the effective size of the ancestral population, under the IM model. 2,000 pseudo-observed datasets (PODs) were simulated under the IM model with heterogeneity in introgression rates. A. x-axis: values of *Nanc* used to produce the PODs; y-axis: estimated values of *Nanc* for all PODs. Solid line represents f(x) = x. Dotted lines represent f(x) = 2.x and f(x) = x/2 respectively. B. Estimated values of *Nanc* for the observed pairs of population/species as a function of their net synonymous divergence.
(TIF)

S14 Fig. Estimating T_{split}, the time of ancestral subdivision, under the IM model. 2,000 pseudo-observed datasets (PODs) were simulated under the IM model with heterogeneity in introgression rates. T_{split} is expressed in million of generations since the ancestral separation.

A. x-axis: values of T_{split} used to produce the PODs; y-axis: estimated values of T_{split} for all PODs. Solid line represents f(x) = x. Dotted lines represent f(x) = 2.x and f(x) = x/2 respectively. B. Estimated values of T_{split} for the observed pairs of population/species as a function of their net synonymous divergence.
(TIF)

S1 Table. Number of populations and species inferred to be isolated or connected by ABC.
(ODS)

S1 Text. Simulation study to test the robustness of ABC in face of recent times of divergence.
(PDF)

S1 Data. Accessions of surveyed individuals, geographic locations and summary statistics.
(XLSX)

Acknowledgments

We thank Aude Darracq, Vincent Castric, Pierre-Alexandre Gagnaire, Xavier Vekemans, and John Welch for insightful discussions. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics and the ISEM computing cluster at the platform Montpellier Bioinformatique et Biodiversité.

Author Contributions

Conceptualization: Camille Roux, Nicolas Galtier, Nicolas Bierne.

Data curation: Christelle Fraïsse, Jonathan Romiguier, Yoann Anciaux, Nicolas Galtier, Nicolas Bierne.

Formal analysis: Camille Roux, Nicolas Galtier, Nicolas Bierne.

Funding acquisition: Nicolas Galtier, Nicolas Bierne.

Investigation: Camille Roux, Christelle Fraïsse, Jonathan Romiguier, Yoann Anciaux, Nicolas Galtier, Nicolas Bierne.

Methodology: Camille Roux, Nicolas Galtier, Nicolas Bierne.

Resources: Jonathan Romiguier, Yoann Anciaux, Nicolas Galtier, Nicolas Bierne.

Software: Camille Roux, Nicolas Galtier.

Supervision: Nicolas Galtier, Nicolas Bierne.

Validation: Nicolas Galtier, Nicolas Bierne.

Writing – original draft: Camille Roux, Nicolas Galtier, Nicolas Bierne.

Writing – review & editing: Camille Roux, Christelle Fraïsse, Jonathan Romiguier, Nicolas Galtier, Nicolas Bierne.

References

1. Coyne JA, Orr HA. Speciation. Sunderland: Sinauer Associates Inc.; 2004.
2. De Queiroz K. Species concepts and species delimitation. Syst Biol. 2007; 56: 879–886. doi: 10.1080/10635150701701083 PMID: 18027281

3. Dettman JR, Sirjusingh C, Kohn LM, Anderson JB. Incipient speciation by divergent adaptation and antagonistic epistasis in yeast. Nature. 2007; 447: 585–588. doi: 10.1038/nature05856 PMID: 17538619
4. Amato A, Kooistra WHCF, Ghiron JHL, Mann DG, Pröschold T, Montresor M. Reproductive isolation among sympatric cryptic species in marine diatoms. Protist. 2007; 158: 193–207. doi: 10.1016/j.protis.2006.10.001 PMID: 17145201
5. Mayr E. Animal species and evolution. Cambridge: Harvard University Press; 1963.
6. Gavrilets S. Fitness landscapes and the origin of species. Princeton: Princeton University Press; 2004.
7. Bolnick DI, Fitzpatrick BM. Sympatric Speciation: Models and Empirical Evidence. Annu Rev Ecol Evol Syst. 2007; 38: 459–487.
8. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013; 9: e1003905. doi: 10.1371/journal.pgen.1003905 PMID: 24204310
9. McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Mol Phylogenet Evol. 2012; 62: 397–406. doi: 10.1016/j.ympev.2011.10.012 PMID: 22063264
10. Emerson BC, Paradis E, Thébaud C. Revealing the demographic histories of species using DNA sequences. Trends Ecol Evol. 2001; 16: 707–716.
11. Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol. 2010; 25: 410–418. doi: 10.1016/j.tree.2010.04.001 PMID: 20488578
12. Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity. The Genetical Society of Great Britain; 1986; 57: 357–376.
13. Sousa VMC, Carneiro M, Ferrand N, Hey J. Identifying Loci Under Selection Against Gene Flow in Isolation with Migration Models. Genetics. 2013; 211–233.
14. Roux C, Tsagkogeorga G, Bierne N, Galtier N. Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. Mol Biol Evol. 2013; 30: 1574–1587. doi: 10.1093/molbev/mst066 PMID: 23564941
15. Roux C, Fraïsse C, Castric V, Vekemans X, Pogson GH, Bierne N. Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a Mytilus hybrid zone. J Evol Biol. 2014; 27: 1662–1675. doi: 10.1111/jeb.12425 PMID: 24913446
16. Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun. 2014; 5: 5770. doi: 10.1038/ncomms6770 PMID: 25534655
17. Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet. 2013; 14, 404–414 doi: 10.1038/nrg3446 PMID: 23657479
18. Wu C-I. The genic view of the process of speciation. J Evol Biol. 2001; 14: 851–865.
19. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014; 23: 3133–3157. doi: 10.1111/mec.12796 PMID: 24845075
20. Maynard-Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23: 23–35. PMID: 4407212
21. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134: 1289–1303. PMID: 8375663
22. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 2016; 25:135–41. doi: 10.1111/mec.13390 PMID: 26394805
23. Schrider DR, Shanku AG, Kern AD. Effects of Linked Selective Sweeps on Demographic Inference and Model Selection. Genetics. 2016. E-pub ahead of print.
24. Gillespie JH. Is the population size of a species relevant to its evolution? Evolution. 2001; 55: 2161–2169. PMID: 11794777
25. Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009; 10: 195–205. doi: 10.1038/nrg2526 PMID: 19204717
26. Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. PLoS Genet. 2013; 9: e1003942. doi: 10.1371/journal.pgen.1003942 PMID: 24244198
27. Warmuth V, Eriksson A, Bower MA, Barker G, Barrett E, Hanks BK, et al. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. Proc Natl Acad Sci U S A. 2012; 109: 8202–8206. doi: 10.1073/pnas.1111122109 PMID: 22566639

28. Roux C, Pannell JR. Inferring the mode of origin of polyploid species from next-generation sequence data. Mol Ecol. 2015; 24: 1047–1059. doi: 10.1111/mec.13078 PMID: 25585898

29. Ross-Ibarra J, Tenaillon M, Gaut BS. Historical divergence and gene flow in the genus Zea. Genetics. 2009; 181: 1399–1413. doi: 10.1534/genetics.108.097238 PMID: 19153259

30. Diez CM, Trujillo I, Martinez-Urdiroz N, Barranco D, Rallo L, Marfil P, et al. Olive domestication and diversification in the Mediterranean Basin. New Phytol. 2015; 206: 436–447. doi: 10.1111/nph.13181 PMID: 25420413

31. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature. 2014; 515: 261–263. doi: 10.1038/nature13685 PMID: 25141177

32. Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, et al. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A. 2007; 104: 17614–17619. doi: 10.1073/pnas.0708280104 PMID: 17978179

33. Hey J, Nielsen R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A. 2007; 104: 2785–2790. doi: 10.1073/pnas.0611164104 PMID: 17301231

34. Hey J, Jody H, Yujin C, Arun S. On the occurrence of false positives in tests of migration under an isolation-with-migration model. Mol Ecol. 2015; 24: 5078–5083. doi: 10.1111/mec.13381 PMID: 26456794

35. Becquet C, Przeworski M. A new approach to estimate parameters of speciation models with application to apes. Genome Res. 2007; 17: 1505–1519. doi: 10.1101/gr.6409707 PMID: 17712021

36. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979; 76: 5269–5273. PMID: 291943

37. Lee JY, Edwards SV. Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (Malurus melanocephalus). Evolution. 2008; 62: 3117–3134. doi: 10.1111/j.1558-5646.2008.00543.x PMID: 19087188

38. Jolly MT, Viard F, Gentil F, Thiébaut E, Jollivet D. Comparative phylogeography of two coastal polychaete tubeworms in the Northeast Atlantic supports shared history and vicariant events. Mol Ecol. 2006; 15: 1841–1855. doi: 10.1111/j.1365-294X.2006.02910.x PMID: 16689902

39. López-Legentil S, Turon X. Population genetics, phylogeography and speciation of Cystodytes (Ascidiacea) in the western Mediterranean Sea. Biol J Linn Soc Lond. 2006; 88: 203–214.

40. Dupont L, Grésille Y, Richard B, Decaëns T, Mathieu J. Dispersal constraints and fine-scale spatial genetic structure in two earthworm species. Biol J Linn Soc Lond. 2015; 114: 335–347.

41. Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol. 1998; 15: 538–543. PMID: 9580982

42. Noor MAF, Bennett SM. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity. 2009; 103: 439–444. doi: 10.1038/hdy.2009.151 PMID: 19920849

43. Harrison RG. Hybrid zones and the evolutionary process. Oxford University Press: Oxford; 1993. p. 364.

44. Alcala N, Vuilleumier S. Turnover and accumulation of genetic diversity across large time-scale cycles of isolation and connection of populations. Proc Biol Sci. 2014; 281: 20141369. doi: 10.1098/rspb.2014.1369 PMID: 25253456

45. Barton NH. What role does natural selection play in speciation? Philos Trans R Soc Lond B Biol Sci. 2010; 365: 1825–1840. doi: 10.1098/rstb.2010.0001 PMID: 20439284

46. Wakeley J, Hey J. Estimating ancestral population parameters. Genetics. 1997; 145: 847–855. PMID: 9055093

47. Coyne JA, Orr HA. Patterns of Speciation in Drosophila. Evolution. 1989; 43: 362–381.

48. Fraïsse C, Gunnarsson PA, Roze D, Bierne N, Welch JJ. The genetics of speciation: Insights from Fisher's geometric model. Evolution. 2016; 70: 1450–1464. doi: 10.1111/evo.12968 PMID: 27252049

49. Fraïsse C, Elderfield JAD, Welch JJ. The genetics of speciation: are complex incompatibilities easier to evolve? J Evol Biol. 2014; 27: 688–699. doi: 10.1111/jeb.12339 PMID: 24581268

50. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007; 315: 476–480. doi: 10.1126/science.1127573 PMID: 17255503

51. Opperman R, Emmanuel E, Levy AA. The effect of sequence divergence on recombination between direct repeats in Arabidopsis. Genetics. 2004; 168: 2207–2215. doi: 10.1534/genetics.104.032896 PMID: 15611187

52. Cutter AD, Jovelin R, Dey A. Molecular hyperdiversity and evolution in very large populations. Mol Ecol. 2013; 22: 2074–2095. doi: 10.1111/mec.12281 PMID: 23506466

53. Bierne, Borsa, Daguin, Jollivet, Viard, Bonhomme, et al. Introgression patterns in the mosaic hybrid zone between Mytilus edulis and M. galloprovincialis. Mol Ecol. 2003; 12: 447–461. PMID: 12535095

54. Bierne N, Bonhomme F, David P. Habitat preference and the marine-speciation paradox. Proc Biol Sci. 2003; 270: 1399–1406. doi: 10.1098/rspb.2003.2404 PMID: 12965032

55. Janousek V, Václav J, Liuyang W, Ken L, Petra D, Vyskocilova MM, et al. Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between Mus musculus musculus and M. m. domesticus. Mol Ecol. 2012; 21: 3032–3047. doi: 10.1111/j.1365-294X.2012.05583.x PMID: 22582810

56. Carneiro M, Blanco-Aguiar JA, Villafuerte R, Ferrand N, Nachman MW. Speciation in the European rabbit (Oryctolagus cuniculus): islands of differentiation on the X chromosome and autosomes. Evolution. 2010; 64: 3443–3460. doi: 10.1111/j.1558-5646.2010.01092.x PMID: 20666840

57. Baldassarre DT, White TA, Karubian J, Webster MS. Genomic and morphological analysis of a semi-permeable avian hybrid zone suggests asymmetrical introgression of a sexual signal. Evolution. 2014; 68: 2644–2657. doi: 10.1111/evo.12457 PMID: 24889818

58. Nydam ML, Harrison RG. Introgression despite substantial divergence in a broadcast spawning marine invertebrate. Evolution. 2011; 65: 429–442. doi: 10.1111/j.1558-5646.2010.01153.x PMID: 21044056

59. Matute DR, Butler IA, Turissini DA, Coyne JA. A test of the snowball theory for the rate of evolution of hybrid incompatibilities. Science. 2010; 329: 1518–1521. doi: 10.1126/science.1193440 PMID: 20847270

60. Moyle LC, Nakazato T. Hybrid incompatibility "snowballs" between Solanum species. Science. 2010; 329: 1521–1523. doi: 10.1126/science.1193063 PMID: 20847271

61. Hewitt GM. Quaternary phylogeography: the roots of hybrid zones. Genetica. 2011; 139: 617–638. doi: 10.1007/s10709-011-9547-3 PMID: 21234647

62. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754–1760. doi: 10.1093/bioinformatics/btp324 PMID: 19451168

63. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29: 644–652. doi: 10.1038/nbt.1883 PMID: 21572440

64. Tsagkogeorga G, Turon X, Galtier N, Douzery EJP, Delsuc F. Accelerated evolutionary rate of housekeeping genes in tunicates. J Mol Evol. 2010; 71: 153–167. doi: 10.1007/s00239-010-9372-9 PMID: 20697701

65. Thornton K. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 2003; 19: 2325–2327. PMID: 14630667

66. Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, et al. Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata. PLoS ONE. 2008; 3: e2411. doi: 10.1371/journal.pone.0002411 PMID: 18545707

67. Roux C, Castric V, Pauwels M, Wright SI, Saumitou-Laprade P, Vekemans X. Does speciation between Arabidopsis halleri and Arabidopsis lyrata coincide with major changes in a molecular target of adaptation? PLoS ONE. 2011; 6: e26872. doi: 10.1371/journal.pone.0026872 PMID: 22069475

68. Navascués M, Legrand D, Campagne C, Cariou M-L, Depaulis F. Distinguishing migration from isolation using genes with intragenic recombination: detecting introgression in the Drosophila simulans species complex. BMC Evol Biol. 2014; 14: 89. doi: 10.1186/1471-2148-14-89 PMID: 24762206

69. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983; 105: 437–460. PMID: 6628982

70. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975; 7: 256–276. PMID: 1145509

71. Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989; 123: 597–601. PMID: 2599369

72. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18: 337–338. PMID: 11847089

73. Csilléry K, François O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012; 3: 475–479.