Genomic approach to detecting barriers to gene flow

Ewen Burban

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

THÈSE DE DOCTORAT DE

Rapporteurs avant soutenance :

Thibault Leroy	Chargé de recherche, INRAE (GenPhySE), Université de Toulouse
Christophe Lemaire	Enseignant-Chercheur, INRAE (IRHS), Université d'Angers

Composition du Jury :

Président :	Guillaume Evanno	Directeur de recherche, INRAE (DECOD), Université de Rennes
Examinateurs :		
	Violaine Llaurens	Directrice de recherche, (ISYEB) Museum National d'Histoire Naturelle
	Christelle Fraïsse	Chargée de recherche, (EEP) Université de Lille
Dir. de thèse :	Sylvain Glémin	Directeur de recherche, (ECOBIO), Université de Rennes
Co-dir. de thèse :	Maud Tenaillon	Directrice de recherche, (GQE-LeMoulon), Université Paris-Saclay

"All models are wrong, but some are useful". George E. P. Box

Awknowledgements

To begin with, this whole story wouldn't have happened without Maud saying to herself one day, "Well, we've got a candidate for our evolution internship who knows nothing about evolution and nothing about programming, but he's a pharmacist! let's take him! I'd like to thank you, Maud, for your confidence, which you've shown again and again for this thesis. Next, I'd like to thank you, Maud and Sylvain, for your guidance and your confidence in me. I'd like to thank you for giving me so much freedom in my thesis. And finally, thank you for allowing me to take time out from this thesis to write my pharmacy thesis. Thanks to Augustin and Clémentine who were great master's students to supervise, and also thanks to them for their work and contribution to my thesis. Thanks to Arthur for all his help with my work. Thanks to Augustin, Clémentine, Arthur and Harry for being RIDGE alpha-testers.

This thesis wasn't all work, it was also an opportunity to meet new people, quickly turning into friendships that I hope will last a long time. So thank you to Nathan, Morgane, Harriet, Thomas, Rémi, Léa, Solène, Sacha and Antoine (even if he doesn't have a PhD :D). You have all been wonderful companions during this thesis. I'd also like to thank my family for always supporting me and taking an interest in my work, despite the complexity of the subjects.

Finally, thank you Camille, my best friend, my companion, my PACS, the mother of my daughter. Thank you for putting up with me and supporting me during these three years. Without you, this thesis experience would have been less fun and probably more difficult. And then there was Alix. Thank you Alix, you were my precious companion (even if you slept during the process) during those many evenings of writing. Camille, Alix, you were my crutch when my mind was wavering under fatigue and laziness, my lights at the end of the tunnel, my two reasons to push myself.

Odi panem quid meliora. Which mean nothing but I think it finish well.

Enjoy your reading !

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Speciation: from concept to models and empirical evidences

1.1.1 The Species concept

In 1859, in his book "On the Origin of Species", Charles Darwin (Darwin 1859) concluded about species definition with this: *"In short, we shall have to treat species in the same manner as those naturalists treat genera, who admit that genera are merely artificial combinations made for convenience. This may not be a cheering prospect, but we shall at least be freed from the vain search for the undiscovered and undiscoverable essence of the term species"* (Darwin 1859). Nowadays, species is recognized as the fundamental unit of biology and is defined as a group of organisms that can successfully interbreed and produce fertile offspring; it follows from this definition that organisms that form two different species can not interbreed i.e. they are reproductively-isolated (Dobzhansky 1937; Mayr 1942). For non-specialist scientists and the general public, a simplified representation of natural variation made of discrete categories is appreciated (Galtier 2019). Such simplification is also useful in macroecology or conservation biology for the study of the effects of climate changes on biodiversity which relies on our ability to count species, monitor their diversity and track their evolution. However, delineating species is a difficult task as it means clustering living organisms into discontinuous categories (i.e species) despite the fact that they cover a continuous gradient of reproductive isolation (RI). Indeed, intermediate individuals are frequently observed, and in many analyses, the number and nature of categories depend on arbitrary thresholds or parameters such as the reproductive isolation threshold above which population should be considered as species (Galtier 2019) Furthermore, reproductive isolation is often hard or impossible to test (e.g between extinct species).

To address the challenges associated with delineating species, De Queiroz (2007) proposes to separate the species conceptualisation from the operational delineation of species. Using this framework, species concept is defined as mentioned previously and species delineation is based on multiple criteria (e.g ecological differentiation, genotypic cluster, etc...) depending

on their availability and operationality for the considered taxon. If a population of organisms satisfies all the criteria, then it can be considered as a species. Conversely, if it fails to meet the relevant criteria, it is not considered a species. He also introduces the concept of the "gray zone", confirmed empirically by extensive comparisons across species (Roux et al. 2016). The gray zone represents a transitory stage between the status of population and the status of species. In this zone, species/populations meet certain criteria but not all of them. This zone is particularly relevant to understand the evolutionary forces governing reproductive isolation and, over time, the process of speciation.

1.1.2 Forms of Reproductive isolation

As explained above, the notion of RI is central to understanding species definition and thus speciation. A recent survey of 231 researchers studying speciation indicates divergence of opinions on the definition of RI (Westram et al. 2022). From this survey, two primary perspectives on RI emerge: RI is either considered as a reduction in hybrid production and/or fitness or as a reduction of gene flow. Importantly, these two views of RI are not mutually exclusive: decreasing hybrid production or fitness diminishes gene flow. But the decision to adopt one definition over the other affects the empirical measurement of RI in practice. In this thesis, I considered RI from a genomic perspective, and therefore used RI as a quantitative measure of the effect of genetic differences on gene flow (Westram et al. 2022). Genetic variations that reduce gene flow include any genetic difference influencing organisms-level traits that limit gene flow between populations. Conceptually RI is characterized by its effect on the effective migration rate, m_e, which represents the rate of gene flow between populations. It is worth distinguishing m_e from migration rate m: m quantifies the exchange of individuals between populations without considering their effect on the evolution of the genetic composition of the population whereas m_e quantifies their genetic inflow. In other words, m_e is the migration rate that results in the same change in allele frequencies as observed in an ideal population where all migrants would contribute equally to gene flow (Barton and Bengtsson 1986). Selection for migrants can increase gene flow ($m_e/m > 1$, the effective migration rate is larger than the actual migration rate) as in the case of adaptive introgression. On the contrary, selection against migrants can decrease gene flow ($m_e/m < 1$, the effective migration rate is smaller than the actual migration rate) as in the case of a genetic barrier. The ratio m_e/m also called the gene flow factor is a measure of the penetrability of a genetic barrier (Bengtsson 1985), and its inverse defines the strength of a genetic barrier (Barton and Bengtsson 1986). In the following, I describe different categories and mechanisms of RI.

Prezygotic isolation

Prezygotic isolation encompasses a spectrum of mechanisms influencing the likelihood of zygote formation. Here, I describe the main types of prezygotic RI and illustrate them with a biological example.

A straightforward condition that prevents zygote formation is when the two potential parents cannot meet because they live in different habitats, which can result from local adaptation to distinct ecological conditions. For example, premating isolation between stickleback ecomorphs arose as a simple by-product of divergent natural selection on traits such as prey (Rundle et al. 2000). One ecomorph (Benthic) lives in the littoral zone of coastal lakes in Canada and feed on invertebrates, the other (Limnetic) feeds on plankton in open water of the same lakes. Both ecomorphs derived from marine stickleback fishes that invaded the lakes after the post-Pleistocene period. The ecological prezygotic reproductive isolation between Benthic and Limnetic ecomorphs independently occurred on multiple lakes leading to the emergence of both ecotypes in multiple lakes (Rundle et al. 2000).

Other than RI through habitat, temporal RI can limit interactions between two potential parents. This temporal isolation can occur because of differences in life-history traits, or even daily behaviors. For example, RI between two wild species of rice (*Oryza nivara* the daughter species derived from *O. rufipogon*) is linked to flowering time differences with *O. nivara* flowering around 80 days before *O. rufipogon*. Hence, despite no decline in hybrid viability and fertility, RI is maintained by temporal isolation (Xu et al. 2020).

In the context of populations coexisting in the same geographic area and breeding simultaneously, reproductive prezygotic isolation can still arise even among populations. Variations in mating preferences is one of the primary mechanisms that triggers RI in such conditions. Mating preferences rely on various traits such as body size but also visual, olfactory or auditory signals. During the last glaciation period, crow (*Corvus*) populations underwent geographical isolation, resulting in the accumulation of divergence and the emergence of incipient species exhibiting distinct feather color patterns. Even with repeated secondary contacts, RI has persisted. In the central European contact zone, a specific region in chromosome 18 experiences robust divergent selection and harbors genes implicated in the feather color pathway, as well as color pattern recognition (Poelstra et al. 2014; Vijay et al. 2016). Collectively, these genes influence mate choice creating mate preference towards individuals that present the same feather pattern - also called assortative mating, as confirmed by hybrid analysis (Knief et al. 2019; Metzler et al. 2021). In plant species that rely on pollinators to mate, pollinators can act as agents of divergent selection on floral traits. The evolution of these traits is triggered by variations in pollinator distribution, and plant adaptation to the locally most efficient/frequent ones. *Mimulus aurantiacus* (monkey flower) is a good example of such an RI based on pollinator preferences. In Southern California, two ecotypes of *M. aurantiacus* are parapatric, evolving in distinct environmental niches, with a contact zone facilitating migration between populations. The primary difference between these ecotypes lies in flower color. This color disparity is attributed to differences in the abundance of pollinators in the respective environments. In the eastern region, hummingbirds, more receptive to red flowers, are the pollinators of monkeyflowers. Conversely, in the western region, hawkmoths prefer yellow flowers. Incipient species of *M. aurantiacus* have evolved based on the composition of pollinators in their environment through a gene called *MaMyb2*, regulating the amount of anthocyanin in the petal and, consequently,

the red coloration of the petal (Streisfeld et al. 2013). Different variants of *MaMyb2* are closely associated with flower color, resulting in prezygotic isolation by attracting distinct pollinators and thereby reducing the likelihood of hybrid formation. It's important to note that the isolation based on flower color is only partial and no significant intrinsic postzygotic isolation has been reported (Sobel and Streisfeld 2015).

Finally physiological isolation can also cause structural or chemical barriers that keep species isolated from one another. For example, maize underwent domestication approximately 9000 years ago. Within certain regions of Mexico, coexistence of cultivated maize and annual teosintes, the wild relatives of maize, demonstrates close proximity in growth and synchronized flowering. Despite this proximity, the occurrence of hybrids between them is infrequent. Crucially, a pivotal gene for speciation, *Tcb1*, impedes the fertilization process of female teosinte plants by maize pollen through restriction of pollen tube growth within the teosinte pistils (Evans and Kermicle 2001; Lu et al. 2019). This cross-incompatibility is unilateral. That is, in reciprocal circumstances, teosinte pollen exhibits the capacity to fertilize maize, albeit with a slight disadvantage when under direct competition with maize pollen (Lu et al. 2019).

Postzygotic isolation

In the face of gene flow, RI may also occur through postzygotic isolation that diminishes the viability or fertility of the resulting hybrids. As an example, in the *Drosophila melanogaster* and *D. virilis* groups — two closely related species — divergence in gonadal proteins, particularly male-reproductive-tract proteins, is closely associated with sterility of F1 hybrid males (Coyne and Orr 1989). Continuing with *D. melanogaster* and *viridis*, Turissini et al. (2017) showed a reduction in ability to locate food in F1 hybrids, consequently reducing hybrid fitness. Reduction of the ability to locate food relies on the disruption of neural circuitry used to detect olfactory cues and is negatively correlated with parental divergence (Turissini et al. 2017). Finally, RI can be total and immediate through karyotype incompatibilities such as in the case of alterations of the ploidy level. This mechanism has played a critical role in plant diversification, as approximately 35% of species within vascular plant genera have undergone such polyploidization (Wood et al. 2009). Indeed, the doubling of the number of chromosomes in a single generation triggers an immediate reproductive isolation. Specifically, when a diploid individual (2n) mates with a recently polyploidized individual (2*2n = 4n), the gametes of the 4n individuals carry twice the chromosomal content of the 2n, often preventing the formation of a viable zygote.

Between pre- and post-zygotic mechanisms: the case of segregation distorters

Segregation distorters (SDs) are genomic elements that induce a distortion in Mendelian segregation, resulting in the preferential transmission of SD alleles in the progeny of an heterozygote, a phenomenon referred to as meiotic drive. As a result, SDs are overrepresented in viable gametes, ultimately leading to fixation of SDs in the population. Although selected for at the gametic level, such drivers can affect fertility, generating the selection of suppressors. In plants,

a classical example is the case of cytoplasmic male sterility (CMS) mutations that increase their transmission by favoring the female over the male transmission pathway, often balanced by nuclear mutations restoring male fertility (Postel and Touzet 2020). Geographically isolated populations, however, may evolve distinct pairs of segregation distorter/restorer, which can generate incompatibilities upon hybridization and constitute a source of reproductive isolation (Orr and Presgraves 2000; Postel and Touzet 2020). For instance, the allele S1 from African rice (*O. glaberrima*) causes selective abortion of male and female gametes carrying its allelic alternative from Asian rice (*O. sativa*) in a heterozygous individual, resulting in reproductive isolation (Koide et al. 2008). Teosinte mexicana (*Zea mays ssp. mexicana*), contains a complex genetic toxin-antidote system encompassing multiple loci (*Tpd1*, the Teosinte Pollen Drive,*Tdr1* the responder gene as well as *Dcl2* Dicer like-2 and *Tpd2*) that contributes to postzygotic hybrid incompatibility between maize and teosintes by aborting pollen in hybrids (Berube et al. 2023).

Intrinsic and extrinsic RI origin

Beyond the classification based on zygote formation, RI can be categorized by its origin, as either intrinsic or extrinsic. Intrinsic RI is independent of environmental influences, whereas extrinsic RI depends on the interplay between the genome and the environment (genotype-environment interaction). For example, in maize and teosinte, the rejection of maize pollen by teosinte females is unaffected by environmental conditions (Lu et al. 2019). In contrast, in closely related populations of the threespine stickleback fish (*Gasterosteus aculeatus*) living in either the Japanese or the Pacific ocean, divergence in body size, driven by adaptation to dissimilar salinity levels, predator presence, and food availability, influences in turn mate choice but also reduces hybrid fitness in these distinct environments, thus substantiating extrinsic RI (Kume et al. 2010).

1.1.3 Models for the emergence of reproductive isolation

In the previous section, I explained how reproductive isolation manifested. However, to understand speciation, it is essential to grasp the process through which this isolation develops. In other words, how an allele, as seen in *Drosophila*, can emerge and spread, even when it causes a significant reduction in feeding capabilities in heterozygous individuals (hybrids), thereby substantially decreasing their chances of reproducing. Several decades of scientific research have been dedicated to addressing this question of the establishment and maintenance of RI. Beyond comprehending the mechanism of the formation and preservation of this isolation, theoretical models also enable the formulation of testable hypotheses. These hypotheses then guide empirical research, providing a systematic approach to explore and understand the underlying mechanisms of evolution and speciation. In this section, I synthesize the main models of speciation that explain the establishment and maintenance of RI.

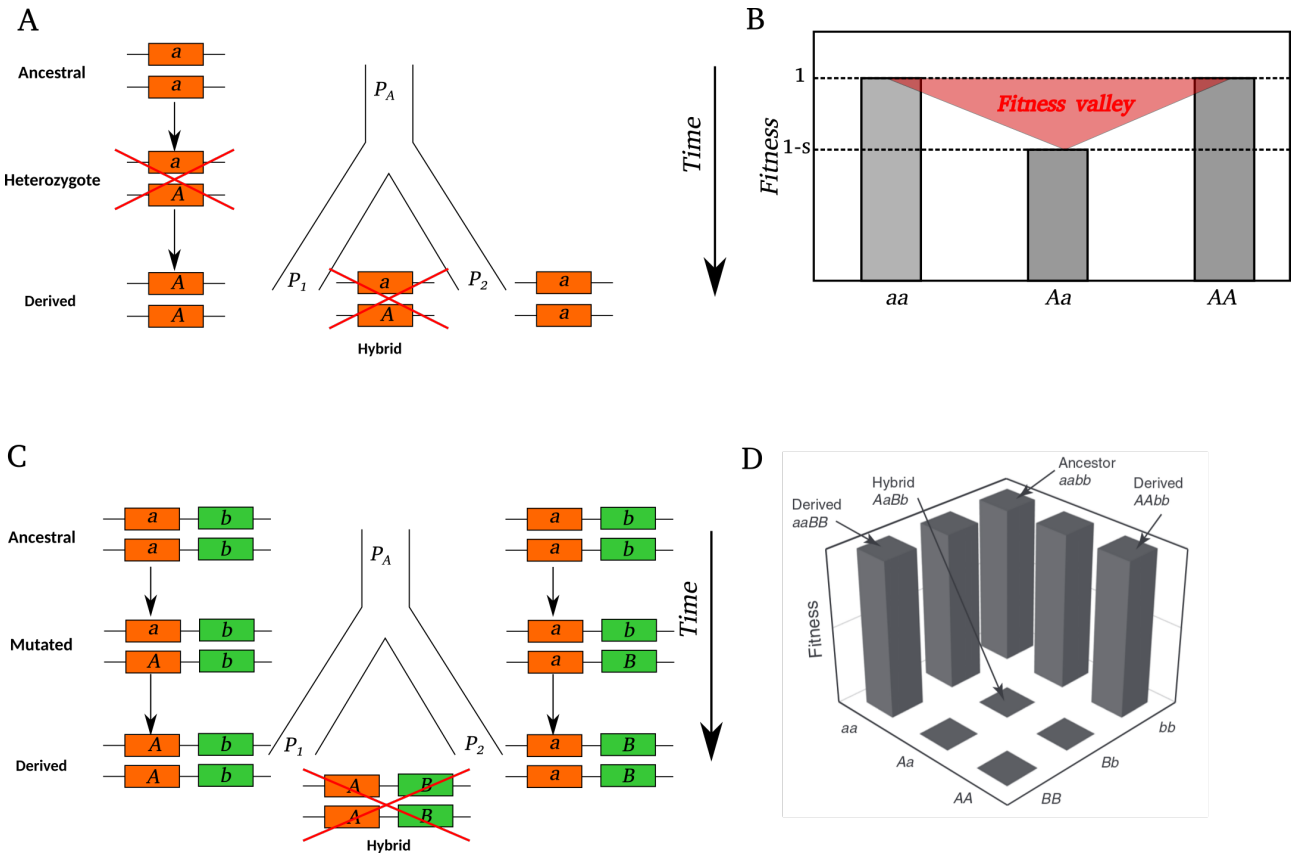

Figure 1.1: Representation of genotype evolution through divergence history and their respective fitness landscape for the underdominance model (A & B) and BDMI model (C & D). Red cross in A & C indicates genotype of reduced fitness (relatively to other genotypes). In C, only one of the fourth incompatible genotypes (as represented in D) is represented. Figure (D) comes from (Cutter 2012).

The problem of crossing fitness valleys

As a first basic model, let's consider two populations, denoted as P_1 and P_2, deriving from an ancestral common population P_A. Let us focus on a specific locus, where every individual in both populations, P_1 and P_2, exhibits the genotype **aa**. To introduce RI, we introduce an additional allele, **A**, such that homozygosity for **A** confers the same fitness as **aa** ($w_{AA} = w_{aa} = 1$), whereas heterozygosity (**Aa**) results in reduced fitness ($w_{Aa} = 1-s$), with s representing the selection coefficient. In such a model, called underdominance, reproduction between **AA** and **aa** individuals produces **Aa** heterozygotes with lower fitness, hence generating RI (see Figure 1.1 B). This reduction in fitness for heterozygotes creates what is commonly referred to as a "fitness valley". Assuming that one population fixes **AA** genotype and the other population fixes **aa** genotype, the RI would be maintained. The problem with this "paradoxical" situation is that it implies that one population should have evolved from the **aa** to the **AA** genotype (or the other way around), such that one population must have crossed the "fitness valley" (through the genotype **Aa**), which is unlikely under many conditions (except if drift or selfing are strong enough to overwhelm selection against heterozygotes, Gavrilets (2003); Charlesworth (1992)). It is noteworthy that the introduction of gene flow between populations P_1 and P_2 does not alter the scenario. Given that the genotype **aa** is ubiquitous in both populations, the effect of gene flow simply amplifies the quantity of alleles accessible within each population.

The BDMI model

Bateson, then later Dobzhansky and Muller each proposed an alternative model (the so-called BDMI for Bateson-Dobzhansky-Muller Incompatibility) that resolved the problem of crossing fitness valleys by considering a two-locus and two-alleles model (Orr 1996). Initially, every individual in populations P_1 and P_2 carries the ancestral genotype at both loci noted **aabb**. The derived alleles, **A** and **B**, are incompatible with each other, reducing fitness of individuals carrying both of them. Yet, **A** and **B** are not incompatible with the ancestral background. If one population fixes the **A** allele, which is possible as the **A** allele in a **bb** background has no negative fitness effect, and the other population fixes the **B** allele, crosses between individuals of these two evolved populations will put together the incompatible **A** and **B** alleles, generating low-fitness hybrids. So, the major result of this model is that RI can take place without requiring crossing any fitness valley. Note that, for simplicity, the BDMI model as presented here involves negative epistasis between two loci, but this model can be generalized to more than two loci. Since then, the BDMI theory has been confirmed multiple times through various empirical example. For example, in the case of *Arabidopsis thaliana*, two strains were identified, *Uk1* and *Uk3*, which, when hybridized, manifest a hybrid necrosis phenotype characterized by severely stunted growth and reduced seed production (see Figure 1.2A from Bomblies et al. (2007)). The authors identified two distinct loci, DM1 and DM2, situated on different chromosomes (see Figure 1.2B). Hybrid necrosis akin to that observed in F1 hybrids occurs only when both the DM1 allele from *Uk1* and the DM2 allele from *Uk3* are found in the same genotype. The authors validated this fact by introgressing incompatible alleles into another strain (*Cor-0*).

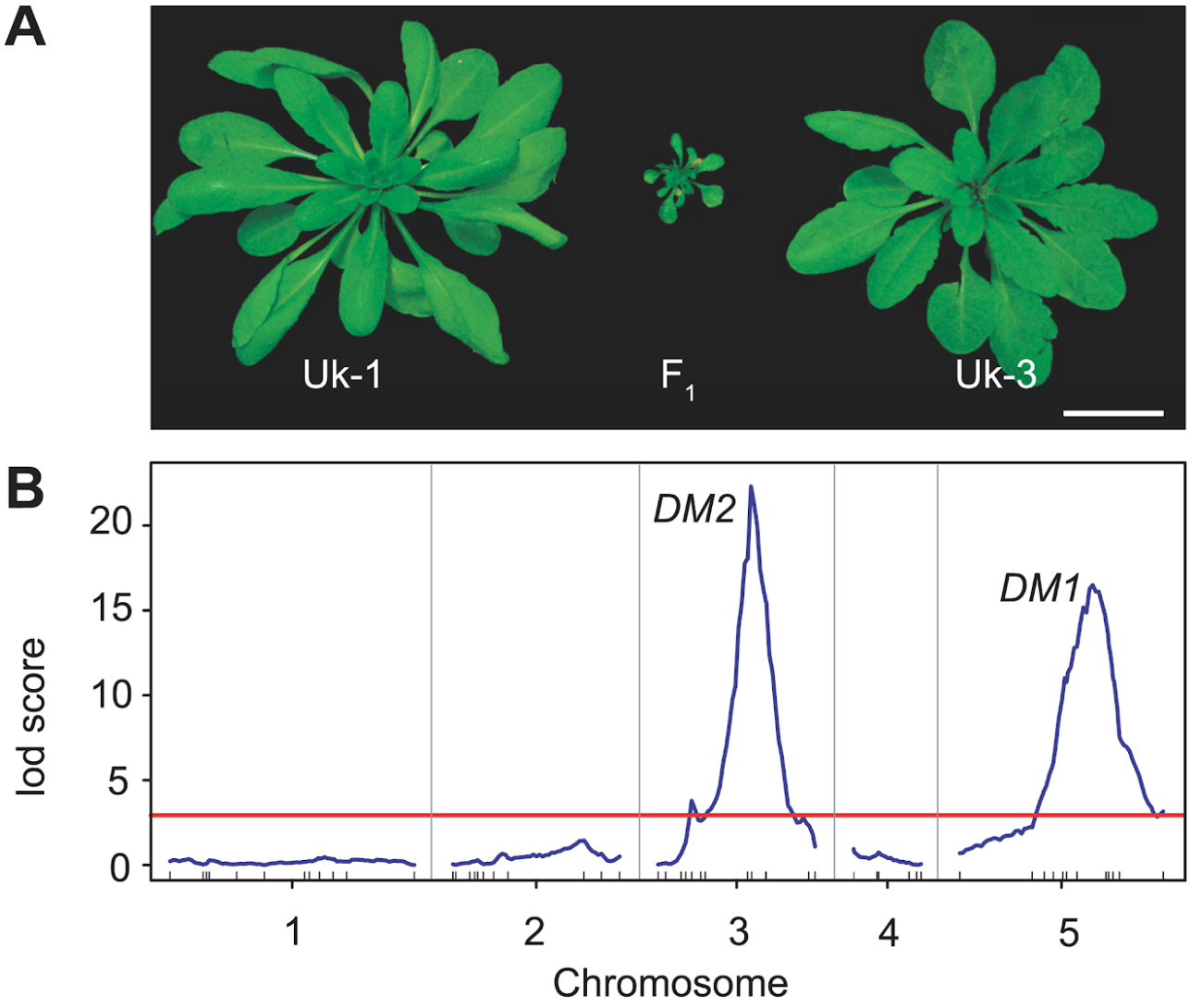

Figure 1.2: BDMI system in A. thaliana between *Uk-1* and *Uk-3* strains producing a hybrid necrosis phenotype A) Two regions (DM1& 2) where identified as associated with F1-like phenotype by doing a QTL mapping represented in B). Figures come from Bomblies et al. (2007).

This validation aligns with the hypothesis of BDMI, wherein epistatic interactions between alleles from two distinct loci, though non-harmful in their original genetic contexts, trigger hybrid necrosis involving an auto-immune response in F1 hybrids, here produced from distinct strains within a species.

Model of ecological speciation

In the initial model described, assuming one locus with two alleles, we neglected the influence of the environment on populations. Let's revisit this model by incorporating local selection and, consequently, local adaptation to the environment. In this refined model, we begin with a population of diploid individuals, each carrying the genotype **aa**, a genotype selected in their respective environment named $E1$. Subsequently, a portion of this population migrates to a different ecological niche, $E2$, where the allele **A** is advantageous and the allele **a** is disadvantageous). Over successive generations and due to mutation, the population in environment $E2$ becomes fixed for the allele **A**, resulting in the entire population carrying the genotype **AA**. Upon hybridization between populations, the hybrids possess the genotype **Aa**, which is not favored in either environment. Consequently, the hybrids experience reduced fitness in the parental environments. For example, under conditions intermediate between the two environments, the **Aa** genotype may have the highest fitness. Thus, local adaptation can initially trigger RI, which is called ecological speciation – where speciation is a by-product of selection against migrants in a heterogeneous environment context – but is not sufficient to generate full and permanent RI.

Model of speciation under gene flow

Gene flow plays a pivotal role in the evolution of RI, it tends to erode RI by reducing the genetic disparities between divergent lineages. Through BDMI, speciation can happen as a by-product of neutral divergence without the need of selection. However, under fully neutral conditions, even a tiny amount of gene flow, as encountered under parapatric scenarios, may prevent the establishment of BDMI or by collapsing previously established BDMI into a single genotype, and so suppressing RI (Lindtke and Buerkle 2015; Bank et al. 2012). Two main mechanisms can drive the evolution of BDMI despite gene flow: first, exogenous selection favoring local alleles at least at one locus is required, that is selection against migrants must act in addition to selection against hybrids. Then, genetic linkage between the two interacting loci facilitate the establishment and maintenance of BDMI (Bank et al. 2012).

Apart from BDMI, speciation under gene flow involves the same key mechanism to maintain RI. Indeed, for maintenance of RI by local adaptation, the strength of selection (s) must overcome the rate of effective migration (m_e), ensuring that selection maintains local adaptation (Yeaman and Otto 2011). Also, selection against migrants is more effective if incompatibility loci are tightly linked to local adaptation loci, so selection tends to select closely linked architecture, strongly enhancing RI (Flaxman et al. 2013). Ecological speciation of the wildflower *Mimulus guttatus* inhabiting regions near copper mines, is a striking case involving both

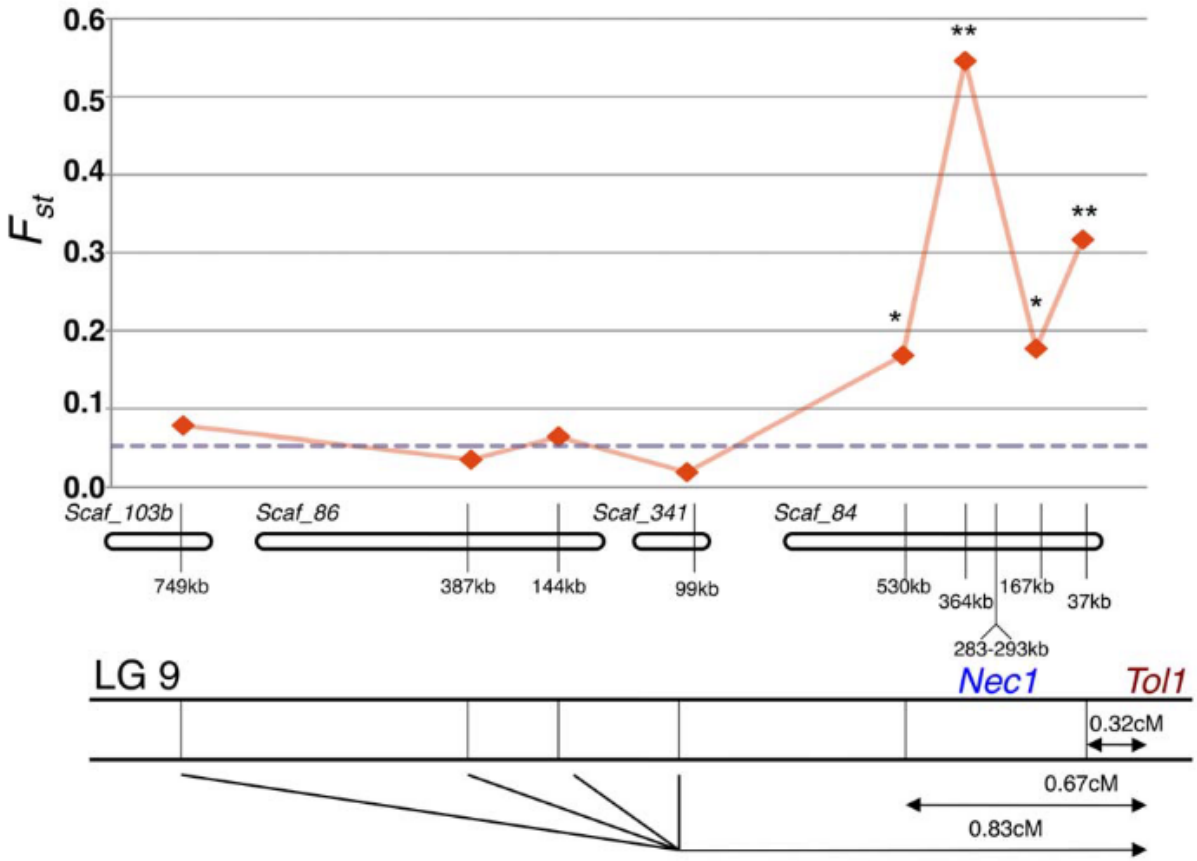

Figure 1.3: F_{ST} landscape at *Tol1* marker and linked marker *Nec1* along LG9 chromosome. Figure from Wright et al. (2013)

mechanisms previously mentioned. These mines offer a highly selective habitat due to poor nutrient availability and elevated heavy-metal levels, resulting in a highly oxidizing environment. *M. guttatus*, showing a remarkable copper tolerance, has colonized multiple copper mines (copper mine populations are called *copperopolis*). The genetic control of copper tolerance is associated with a locus known as *Tol1*. The copper-tolerant allele is nearly fixed in all *copperopolis* populations but occurs at a very low frequency in off-mine populations situated 40 km away from the copper mines. Hybridization between *copperopolis* and off-mine populations leads to a hybrid necrosis phenotype, signifying RI. Wright et al. (2013) established that *Tol1* experiences strong divergent selection, with the copper-tolerant allele being highly favored in the *copperopolis* population but disfavored in the off-mine population. While this constitutes a low-intensity extrinsic postzygotic barrier locus alone, it is insufficient for causing hybrid necrosis. Additionally, Wright et al. (2013) identified the closely linked locus *Nec1*, which, in conjunction with *Tol1*, induces hybrid necrosis between *copperopolis* and off-mine populations. The likely scenario is that with the invasion of the copper mine, wildflowers were subjected to high selection pressure leading to a localized selective hard sweep at *Tol1* loci. This, coupled with linked selection, resulted in the hitchhiking of the *Nec1* incompatible allele, forming a *copperopolis* haplotype that couple both a copper tolerance allele and an incompatible *Nec1* allele with off-mine variants (Figure 1.3), subsequently yielding a strong postzygotic barrier. This example emphasizes the importance of considering recombination, as it can transform a low-intensity barrier locus into a highly effective one. In some cases, tight linkage between loci in unecessary as a single "magic" locus may provide local adaptation and RI (Servedio et al. 2011). This is the case for the *MaMyb2* involved in RI between *Mimulus aurantiacus* ecotypes as detailed above in 1.1.2 (Streisfeld et al. 2013; Sobel and Streisfeld 2015).

1.1.4 How reproductive isolation evolves across genome and time

The previous example of *Mimulus* serves as an illustration of partial isolation, which results in a decrease in the number of hybrids, but hybridization can continue as long as hybrid fitness is not zero. To achieve full speciation, which means hybrids must be sterile or non viable, increased selection against migrants or recruitment of additional divergent loci may happen. How, where and when these new loci are recruited and how do they evolve? In this section, I describe the continuous process of the genomic accumulation of RI, with a distinction made between processes influenced by gene flow and those that are not.

Genomic patterns of RI

At the beginning of the speciation process, population divergence takes place at a small number of loci responsible for divergence. In the presence of gene flow, progress toward speciation is rapidly eroded. Under such conditions barrier loci must be selected at a strength counteracting migration. Establishment of haplotypes bearing locally adapted alleles and isolation loci (Schilling et al. 2018) confers a drastic advantage as it enables adaptive divergence and speciation even under elevated migration rates (Schluter and Rieseberg 2022). This may happen

either via linked selection where multiple genes may hitchhike around the initial barrier loci, or via selection of recombination suppressors. In both cases, clustering of genes translates into a broad genomic signal, referred to as "genomic islands of divergence". In neotropical butterflies of the *Heliconius* genus, assortative mating patterns correlate with a genomic region close to *Optix*, which is a crucial locus influencing their unique color patterns (Merrill et al. 2019). Likewise, for stickleback fish, divergent mate preferences have been identified to have originated from the same set of genomic regions that regulate body size, shape, and ecological niche utilization (Bay et al. 2017). Inversions contribute to the accumulation of divergent haplotypes by generating large genomic regions of suppressed recombination as in sunflowers where large haplotype blocks (1-100 Mbp in size) confer prezygotic isolation through traits involved in abiotic factors and life-history traits (Todesco et al. 2020).

Time courses of reproductive isolation

The gradual nature of speciation has been observed in multiple pairs of species showing a correlation between RI and genetic divergence (Coyne and Orr 1989; Presgraves 2002). One of the most comprehensive studies was conducted by Roux et al. (2016). They utilized genomic data and an Approximate Bayesian Computation framework to estimate the probability of ongoing gene flow ($P_{geneflow}$) for 61 independent pairs of diverging taxa across the animal kingdom. They found that at intermediate levels of divergence (0.5% to 2%), there is a noticeable decrease in $P_{geneflow}$ until speciation is completed ($P_{geneflow} = 0$). Those intermediate levels of divergence which correspond to the gray zone of speciation are characterized by patterns of genomic heterogeneity of gene flow where most genomic regions are permeable to gene flow while a few have established barriers. Furthermore, with the same approach a recent study demonstrated that RI (and so the gray zone) appears at divergence level lower in plants than in animals (Monnet et al. 2023).

From this observation comes the question: at what rate does RI builds up? In his theoretical work, Orr (1995) concluded that postzygotic RI is expected to increase at a faster-than-linear rate, i.e. as "snowball" process. This result is obtained when considering BDMI as the sole source of RI and assuming the cumulative effect of BDMI on RI where the number of potential incompatible combinations increases faster than the number loci involved in epistatic interactions. The "snowball" nature of RI process across time is still debated, with some studies that failed finding evidence (Presgraves 2002; Stelkens et al. 2010; Price and Bouvier 2002), and others supporting the theory as in *Drosophila* (Matute et al. 2010) and in *Solanum* (Moyle and Nakazato 2010) species.

1.2 The challenge of detecting barrier loci *in silico*

A central objective in speciation research is to understand the genetic and genomic mechanisms at work in the emergence and maintenance of RI. To do so, one needs to identify gene flow barrier loci and to compare the results from multiple pairs of diverging lineages to capture

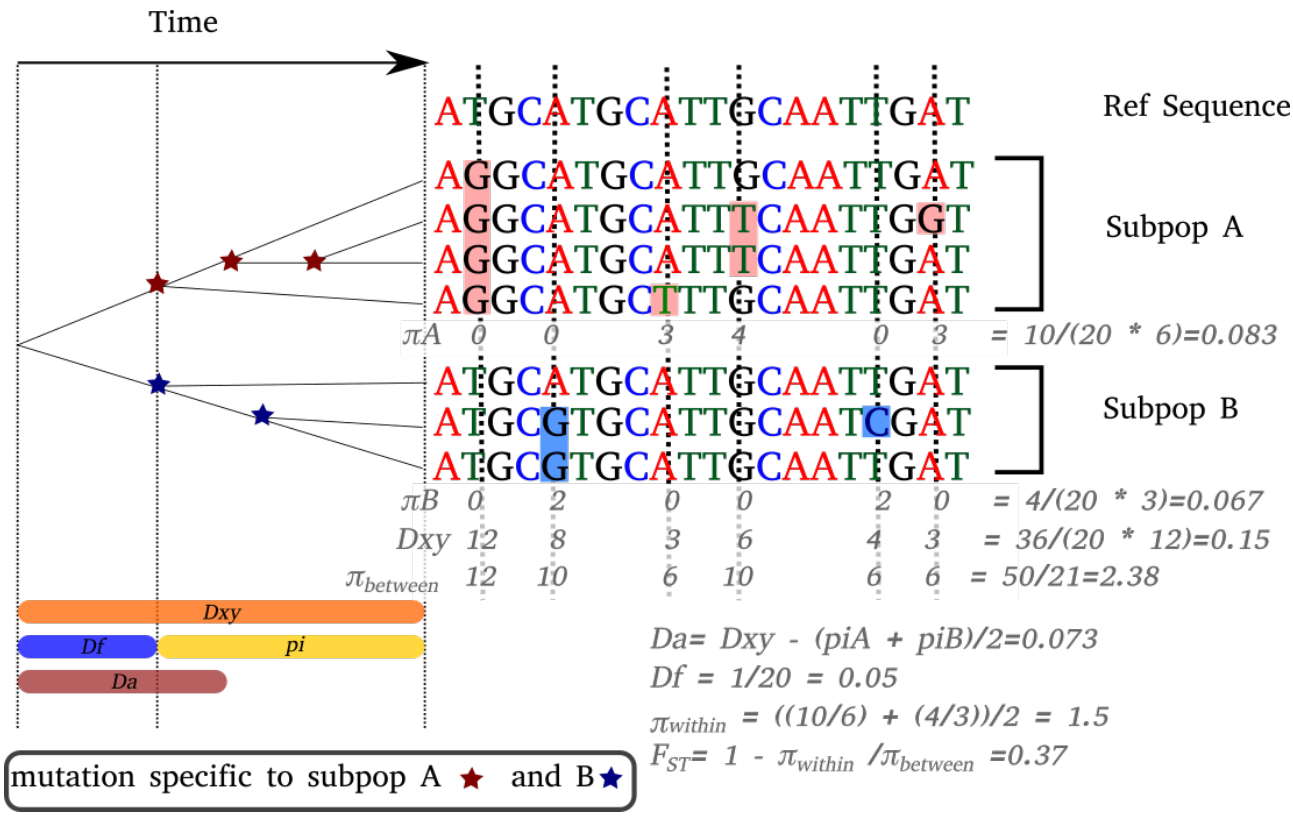

Figure 1.4: SNP data presented in the form of a gene tree (left) and sequence alignment on the sequence of reference (right) between two populations of 4 (subpop A) and 3 (subpop B) haplotypes. For each metric, the part of the divergence covered by the statistics is represented by bars (left) and the calculated values (right).

the sequential events that contribute to the establishment and maintenance of reproductive barriers. Traditionally, barrier detection relied on Quantitative Trait Locus (QTL) analyses and functional assessments. However, the advancement of genome-wide population genetics data has paved the way for genome-scan approaches that are cheaper and easier to undertake, allowing to study a much broader spectrum of populations/species pairs. To effectively investigate genetic determinants of RI through genome-wide population data, it is crucial to possess theoretical expectations regarding the genomic signatures of barrier loci, enabling their detection. In this section, I review: i) the theoretical expectations pertaining to the signatures of barrier loci under various conditions, ii) the confounding effects that can impact the detection of barrier loci, and iii) the main methods employed for their detection, along with the principles underpinning these methodologies.

1.2.1 Genomic signatures of barrier loci

Population genetics data

Population genetics uses genetic polymorphism data which comprises specific genomic locations where individuals within a population exhibit variations. Generating these data often involves re-sequencing the genomes of individuals from a population - but older datasets are generally restricted to exploration of a subset of genomic regions. Subsequently, each indi-

vidual's genome is mapped onto a reference genome (as illustrated in Figure 1.4). The discrepancies between any individual's genome and the reference are recorded as variants. One prominent form of variant frequently utilized is Single Nucleotide Polymorphisms (SNPs) corresponding to point mutations. SNPs are comparatively easier to identify than, for example, structural variants such as inversions or deletions, and they exhibit a higher density along the genomes than any other types of markers (ex: insertion, deletion, small inversion). They are therefore extremely informative to study polymorphism genomic landscapes.

Box genomic metrics

I present in this box how to compute genomic metrics from population genetics data using examples from Figure 1.4.

- **Diversity with π** (Tajima 1989): in Figure 1.4, at position 9 in the sequence, one individual from the subpopulation A, is mutated (A->T). By doing all possible comparisons between all haplotypes from subpopulation A (there are six possible comparisons), three differences are shown. So, at the SNP, position 9, the diversity is of $\pi_{SNP} = 3/6 = 0.5$. Repeating the process at each locus, the diversity of the locus is $\pi_{locus} = (3 + 4 + 3)/6 = 1.667$, and the average diversity per-site in this sequence is $\pi = (3 + 4 + 3)/(L * 6) = 10/(20 * 6) = 0.083$, with L the length of the sequence

- **Divergence with D_{xy}** (Nei and Li 1979a): in Figure 1.4, population A consists of four individuals and population B consists of three individuals, resulting in a total of 12 comparisons. At position 2, population A has completely fixed a mutation (T->G), leading to complete divergence at this locus. Consequently, the 12 comparisons yield a D_{xy} value of $12/12 = 1$ at position 2. The average D_{xy} of a sequence is the cumulative sum of differences averaged over the total number of comparisons (number of positions x number of comparisons).

- **Differentiation with F_{ST}** (Hudson et al. 1992): F_{ST} is computed based on two elements: π_{within} and $\pi_{between}$. First, π_{within} is the average sequence diversity (π_{locus}) within sub populations. So here in Figure 1.4, $\pi_{within} = (\pi_{locusA} + \pi_{locusB})/2 = ((10/6) + (4/3))/2 = 1.5$. Then we compute $\pi_{between}$, which is the diversity across all sequences in the entire population, which make 21 possible comparison, so $\pi_{between} = 50/21 = 2.38$ which make an $F_{ST} = 1 - (1.5/2.38) = 0.37$.

Measure of Genomic diversity

Molecular genetic variation depends on the product of mutation rate (μ) and effective population size (N_e), $\theta = 4 * N_e * \mu$ (for diploids). Empirically, expected molecular genetic variation (genetic diversity) can be estimated at the scale of a loci by Watterson's theta (θ_W) (Watterson 1975a) which is calculated as: $\theta_W = S/a$, where S is the the number of polymorphic sites among n sequence and a is equal to $a = \sum_{i=1}^{n-1} \frac{1}{i}$. Classically θ_W is expressed in per-site unit, so

$\theta_W = S/(a * L)$ where L is the length of the sequence. Alternatively, Nei's π, which measures the nucleotide diversity, is computed as the number of differences between all pairs of sequences, divided by the total number of pairs (Nei and Li 1979b). Nei's π is also usually expressed in per-site units. Both statistics are used to quantify the diversity inside a population, which can vary along chromosomes as a function of m_e but also of other factors (see below).

Measures of differentiation and divergence

Divergence and differentiation are two ways to quantify how much two populations differ. Differentiation measures the disparity in allele frequencies between the populations. Divergence measures the accumulation of genetic difference between populations since their split from a common ancestor. A quantitative measure of absolute divergence known D_{xy} or Nei's D (Nei and Li 1979b) is obtained by counting, at the scale of a SNP, the number of differences between population sequences divided by the total number of comparisons made (see Box Genomic metrics). By excluding all comparisons between sequences of the same population D_{xy} has the advantage of being independent from the population polymorphism in contrast with relative measure of divergence. However, it is affected by the level of ancestral polymorphism present before the split of the daughter populations and the substitution rate (Cruickshank and Hahn 2014). In order to subtract the former from D_{xy} which facilitates comparisons across biological models, the net divergence (D_a) uses the average of within-population diversity (π) found in the two daughter populations as a proxy of the ancestral polymorphism. is therefore a measure of divergence that aims at quantifying differences that have accumulated since divergence. Interestingly, Roux et al. (2016) found that D_a was the best predictor of the probability of ongoing migration between two populations/species, with a value of D_a around 0.01 characterizing the gray zone of speciation. F_{ST} quantifies the differentiation and so the disparity in allele frequencies between populations. It's computed as follows: $F_{ST} = 1 - \frac{H_S}{H_T}$, with H_S the average expected heterozygosity across subpopulations ($H = 2pq$ with p the frequence of the first allele and $q = 1 - p$, the frequency of the second allele assuming a bi-allel site). H_T is expected heterozygosity for the entire metapopulation. Another way to compute F_{ST} provided by Hudson et al. (1992), replaces the expected heterozygosity by nucleotide polymorphism as follows: $F_{ST} = 1 - \frac{\pi_{within}}{\pi_{between}}$, with π_{within} being the average diversity within populations and $\pi_{between}$, the expected nucleotide diversity of sequences sampled between two different populations. Note that, just like Da, F_{ST} is strongly affected by within-population diversity and therefore is a relative measure of differentiation. Another way to selectively capture divergence signals (excluding diversity) is Df, which accounts only for sites where the difference is fixed (Wakeley and Hey 1997).

Expected diversity, divergence and differentiation around barrier loci

Barrier loci are expected to generate RI, thereby diminishing gene flow at the locus level between populations. In the absence of gene flow, barrier loci do not exert a local influence on evolutionary divergence, as geographical isolation already hinders gene flow across the entire

genome. In the presence of gene flow, barrier loci contribute to shaping the differentiation and the divergence landscape by obstructing gene flow in their vicinity, causing genetic sequences to evolve separately between populations at these loci. Due to this independent evolution, independent mutations arise over time, and local adaptation may further favor specific alleles resulting in increased divergence and differentiation (Hejase et al. 2020; Sakamoto and Innan 2019). So, barrier loci are predicted to induce an escalation in net divergence and potentially an increase in D_{xy} (Figure 1.8). However, as D_{xy} also depends on local diversity, this increase may be masked by local variations in diversity. In the presence of gene flow, the diversity of each population can be enriched by migrants. Conversely, barrier loci, which do not experience the effects of gene flow, tend to exhibit lower diversity compared to the remainder of the genome (Figure 1.8). This reduction in diversity is often further accentuated by strong local selection. To summarize, the genetic landscape surrounding barrier loci that have existed for a "sufficiently long time" (the term is deliberately imprecise, as the time required depends on many conditions) is predicted to show higher levels of divergence (primarily net divergence) and differentiation compared to the rest of the genome.

Hence over time, the contrast between the genomic landscape at barrier loci and the rest of the genome intensifies, resulting in the emergence of regions often referred to as "islands of speciation." The metaphorical concept of islands regards the rest of the genome as the sea level. Depending on the metric used, they may also be refered to as "islands of differentiation" (typically for islands identified through F_{ST}) or "islands of divergence". As time increases, an island of divergence emerges (see Figure 1.5), producing the expected patterns for divergence and diversity landscapes.

1.2.2 Factors affecting barrier detection from genomic data

As presented above, barrier loci should leave detectable genomic signatures. Unfortunately, other processes can generate similar confounding signatures. Among factors affecting detection, we can distinguish two groups: one acting at the local level, which partially mimics the genomic pattern of barrier loci, and another at the whole-genome level, which decreases the difference between barrier loci and the rest of the genome reducing detection power.

Confounding factors and power detection

I described in the previous part the expected landscape around gene flow barrier through diversity, divergence and differentiation metrics. Taken separately, there is a large variety of processes that can produce a similar landscape than gene flow barrier and so false positives (Figure 1.6) during detection. F_{ST} measures allele frequency differences, which can be produced by barrier to gene flow but also selection. For example, balancing selection (Figure 1.6B or local adaptation are processes whereby positive selection drives an adaptive allele in a given population to high frequency which leads to an elevated F_{ST} although effective migration is not affected (Figure 1.6 C). Selection sweeps are an extreme form of such positive selection, producing a nearly fixation of a new arisen selected allele and consequently a reduction of

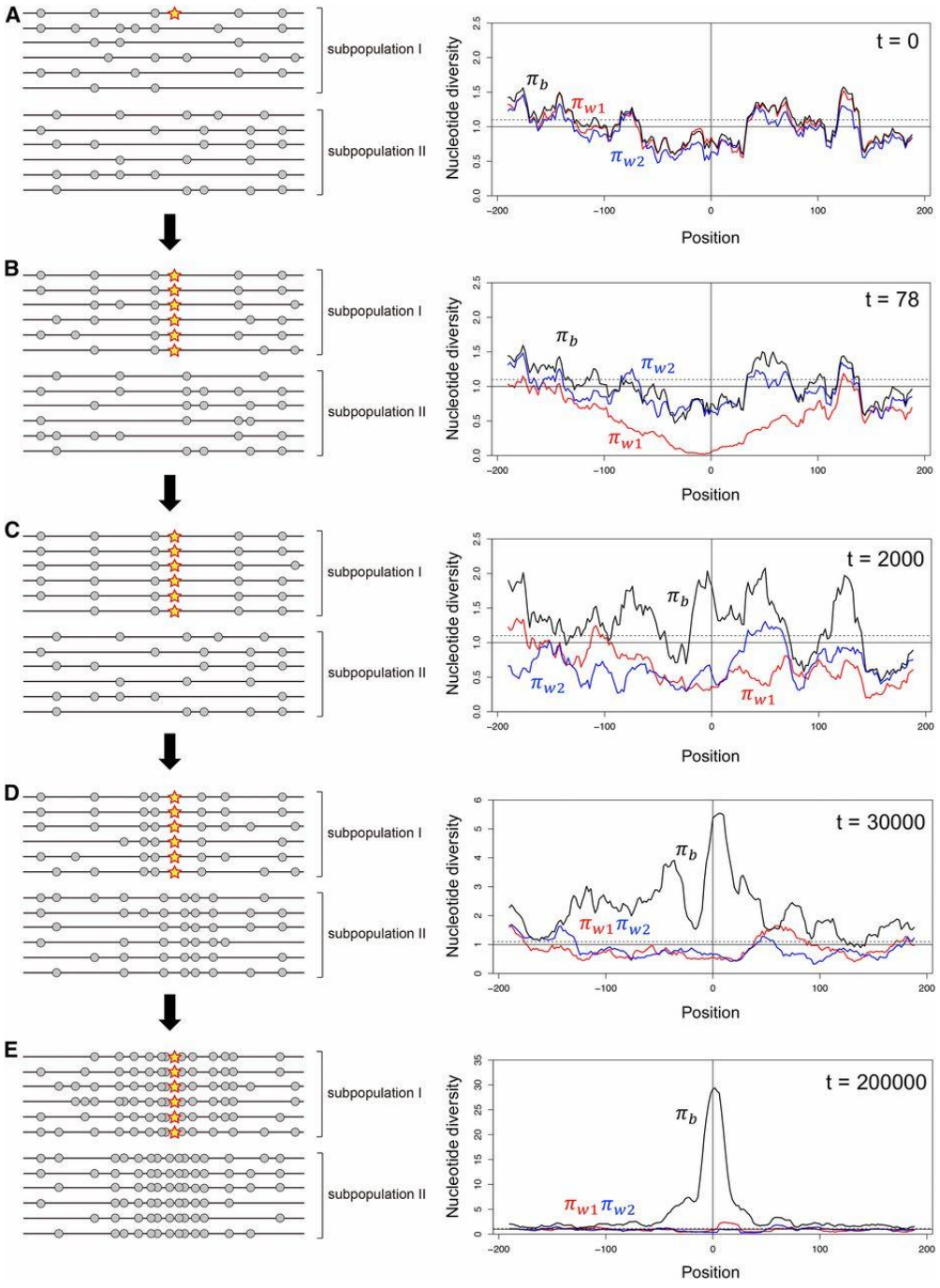

Figure 1.5: Evolution of a barrier locus in a simple two-populations model with high migration between them. (A) A locally adaptive de novo mutation arises in subpopulation I at position 0. A typical pattern of polymorphism is shown on the left. The star is the locally adaptive mutation, and gray circles are neutral polymorphism in the surrounding region. The right panel shows the spatial distributions of nucleotide diversity obtained via simulations. π_1 & π_2 describe the within-population local diversity, and π_b is another notation for D_{xy} which measures the divergence between populations 1 and 2. The simulations considers two populations with $2 * N_1 = 2 * N_2 = 2000$ between which symmetric migration is allowed at rate $4 * N_1 * m = 4 * N_2 * m = 5$. They assume selection intensity $s_1 = 0.2$; $s_2 = -0.2$ The entire simulated region is 400 kb assuming a population-recombination rate of $\rho = 0.001$ per site. From Sakamoto and Innan (2019)

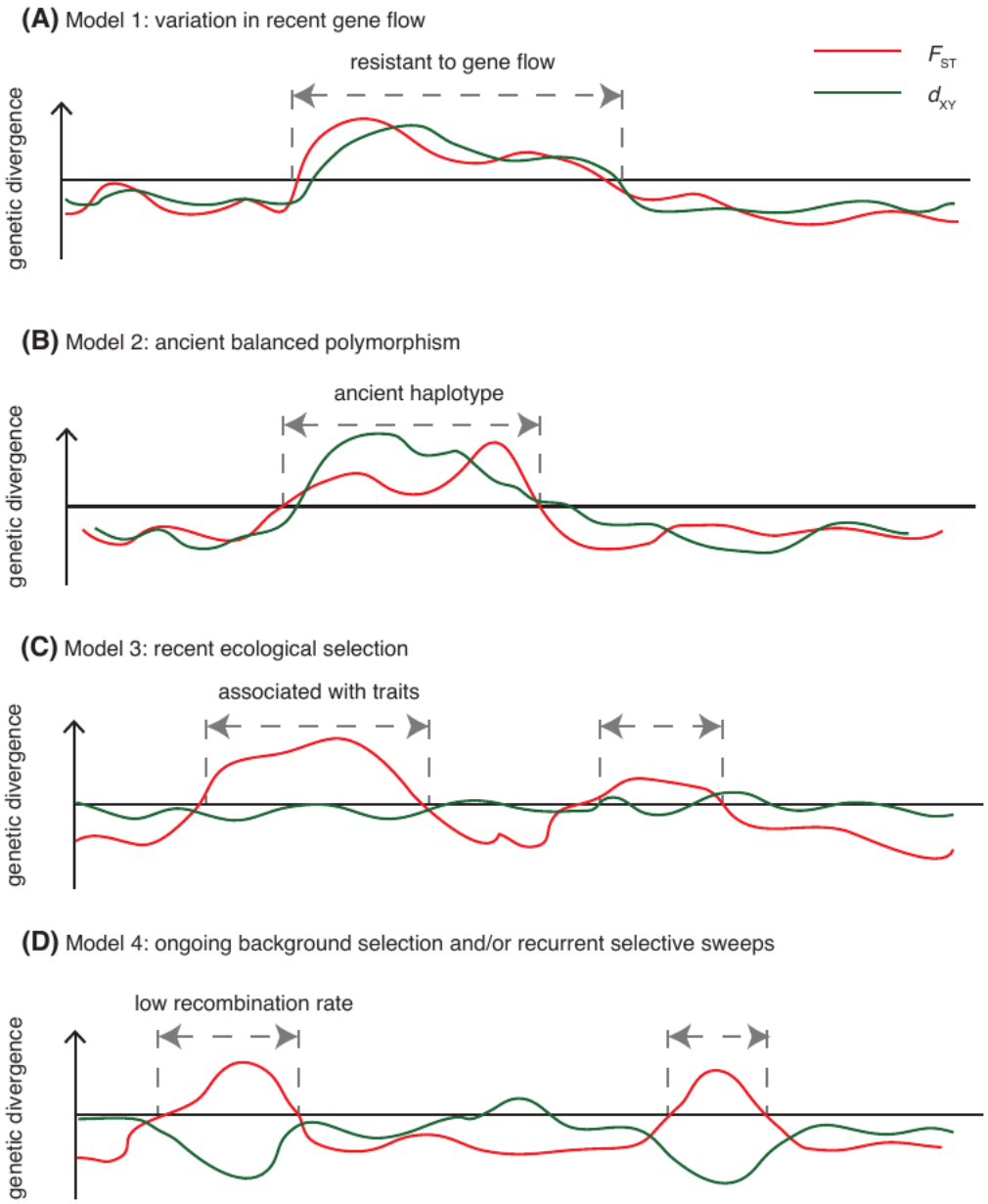

Figure 1.6: Expected patterns of genomic islands of divergence under different models. (A) Model 1: variation in recent gene flow. Gene flow is restricted at island regions, and homogenized the rest of the genome. (B) Model 2: ancient balanced polymorphism. Highly diverged haplotypes present before speciation form genomic islands due to lineage sorting. (C) Model 3: recent selection without gene flow. In allopatric speciation, selection forms the genomic island at adaptive loci. (D) Model 4: ongoing background selection and/or recurrent selective sweep. Recurrent selection accumulates divergence at genomic regions of low recombination. Recurrent selective sweep causes a similar pattern of divergence. Figure from Han et al. (2017)

the diversity at the considered locus as well as a marked increase in F_{ST} (B. Charlesworth 1998; Martin et al. 2015). Negative selection against deleterious alleles (also called background selection) cause similar patterns. In both positive and negative selection, low recombination regions enhance linked selection, that is selection at loci in linkage with the target locus creating extended genomic patterns of selection. In fact, the extent of the genome around loci affected by linked selection directly depends on the balance between local recombination rate and the strength of selection. (Figure 1.6 D). D_{xy} theoretically measures the absolute divergence with $E[D_{xy}] = 2T + 4N_e\mu$ (Nei and Li 1979b), with μ the mutation rate, N_e the effective population size of the ancestral population and T the time of divergence. If T is extremely low, $E[D_{xy}] \approx 4N_e\mu = \theta$. So using only D_{xy}, without considering θ, may lead to interpret regions as divergence outliers where they are just highly diverse, as for regions under balancing selection (Figure 1.6 B).

Because of these multiple confounding factors, it is highly recommended to combine several genomic statistics to mitigate misinterpretation. For example, (Han et al. 2017) propose to combine F_{ST} and D_{xy} to differentiate gene flow barriers from regions under positive or negative selection, as the latter affect differentiation without generating an excess of divergence (Fig 1.6 from Han et al. (2017). M. I. Tenaillon et al. (2023) provided a more complete view considering D_{xy}, F_{ST} and π. They illustrate that the anticipated genomic signature concerning F_{ST} for a barrier loci is identical to that for a selective sweep and a region affected by background selection. However, if D_{xy} and π are examined in addition to F_{ST} to investigate the genomic landscape, the genomic signatures of the three scenarios become distinguishable (Figure 1.8).

Factors limiting power detection

Detection of barrier loci relies on contrasting the genomic landscape of these loci (island level) with the remainder of the genome (sea level). Under certain conditions, the difference of level between "island" and "sea" is reduced, which drastically affects the detection power. This can happen for instance in the case of two populations that diverged in parapatry with very limited gene flow between them (e.g selfers, see below). Another crucial factor to consider is the time of population split, denoted as T_{split}. Assuming a constant gene flow through time, a gene flow barrier will accumulate more population-specific mutations (and so divergence) than the rest of the genome due to the homogenizing effect of migration. But at very low T_{split}, barriers primarily exhibit ancestral polymorphism, indistinguishable from alleles shared by migration in the rest of the genome. We expect the barrier signal to increase with T_{split} (Ravinet et al. 2017; Sakamoto and Innan 2019). The effective population size also significantly impacts this dynamic by increasing the amount of mutations and the time required to generate a divergence signal. To account for this, T_{split} is generally expressed in $2N_e$ units

The mating system exerts a profound influence on gene flow, recombination, selection and so on the genomic landscape of barrier loci (Burgarella and Glémin 2017). The mating system can vary from obligate outcrossing to full selfing where individuals reproduce with themselves. Selfing reduces effective population size due to correlation in gamete sampling and due to re-

current bottlenecks as they often experience extinction-recolonization. Selfing also drastically reduces effective recombination, as both haplotypes available for recombination are identical, precluding the formation of novel genetic combinations (reviewed in Burgarella & Glémin, 2017). Linked selection is thus expected to be more prevalent in selfing than outcrossing species. These processes jointly contribute to strongly reducing genetic diversity in selfing compared to outcrossing species. Finally, selfing also significantly diminishes effective gene flow between populations (Burgarella and Glémin 2017). Consequently, divergence and differentiation level between barrier and the rest of the genome in selfers is strongly reduced compared with outcrossers. Overall, this makes most genomic scan approaches, including detection of barrier loci but also selection, less powerful in selfing than in outcrossing species.

1.2.3 Current genomic-based methods to detect barrier loci

A variety of approaches propose to detect barriers to gene flow from genome-wide patterns exists. They can be categorized into two groups (Tenaillon and Tiffin, 2008): i) data-driven methods involve empirically constructing null distributions from one or multiple statistics obtained from genome scans and rely on arbitrary thresholds to detect outliers; ii) model-based methods involve inferring a demographic model (either beforehand or simultaneously) to establish a null model, followed by the identification of outliers corresponding to barrier loci based on this model. Demography is incorporated to mitigate confounding effects.

Data-driven methods: empirical genome scans

Genome scans consist of measuring genomic features through summary statistics across the whole genome usually computed from sliding-windows. The resulting distributions can then be used to define outliers - loci for which values exceed a certain threshold set by the user. The most commonly used summary statistics is the F_{ST} (78% of studies analyzed uses F_{ST} to detect gene flow barrier through genome scans) according to a survey done by Wolf & Ellergren (2017). F_{ST} scans have inherent limitations, notably in their ability to distinguish between local reduction in effective population size (N_e) and reduction in gene flow (m_e) (B. Charlesworth 1998; Cruickshank and Hahn 2014; Ravinet et al. 2017). That is why, local F_{ST} is sometimes normalized by the local level of F_{ST} from closely related species pairs to the target species (Vijay et al. 2016). Because barrier loci generate a complex signal of differentiation, divergence and diversity, some studies use a combination of summary statistics to perform a stringent detection and avoid as much as possible confounding effects (Hejase et al. 2020; Han et al. 2017). Yet, relying on a combination of signals encounters common limitations with F_{ST} genome-scans. Because thresholds are specific to a given dataset, the results are hardly comparable across studies and biological systems. In addition, the use of empirical threshold precludes any conclusion about the actual number of barriers (i.e. a 5% threshold will provide 5% of barrier loci/windows among all loci). For these reasons, model-based approaches are preferred but are also much more challenging to implement.

Model-based methods

Model-based methods aim at modeling scenarios of population divergence - from one ancestral population splitting in two daughter populations - as well as their demography and possible gene flow between them. Four scenarios of divergence are classically considered: Strict isolation (SI) where divergence occurs under complete allopatry; isolation with migration (IM) where divergence occurs under constant gene flow since the time of split (T_{split}); secondary Contact (SC) where a period of divergence without gene flow is followed by gene flow at a time T_{SC} until present; ancestral migratory model (AM) where an initial period of divergence with gene flow, ending at T_{AM}, is followed by complete isolation until present. These four models are coupled with estimates of population-specific demographic parameters, sometimes also encompassing events such as population expansion or population bottleneck. In this part I describe how these models are used under i) likelihood ii) approximate bayesian computation approach.

Methods based on likelihood:

Likelihood methods provide a statistical framework for estimating parameters describing the genetic data and for assessing the likelihood of the data under different hypotheses, including the presence or absence of barriers under a comprehensive divergence model. For a given divergence model M with a parameter set θ, the likelihood represents the probability (P) of the observed data given these parameters, denoted as $P_M(Data|\theta)$. So the likelihood approach involves maximizing the likelihood function to find the parameter values that make the observed data most probable. This can be achieved for simple demographic models as proposed in *gIMble* (Laetsch et al. 2023). *gIMble* operates under the hypothesis that divergence follows an IM (isolation with migration) model for parameter estimation. Rather than co-estimating all parameters of the IM model in each sliding-window, *gIMble* first infers the best-fitting global IM history, assuming single constant N_e and m_e across the genome. Subsequently, in a second step, local variations in N_e and m_e parameters are co-estimated in sliding windows. While likelihood approaches are extensively utilized, this process usually requires solving optimization problems, and the complexity of these problems can vary depending on the structure of the likelihood function and the number of parameters. An additional limitation is that, in models with a large number of parameters such as SC, the likelihood function may be too complex to solve analytically.

Method based of Approximate Bayesian Computation:

For complex models, likelihood functions can be impossible to implement and Approximate Bayesian Computation (ABC) methods offer an interesting alternative. The concept of ABC traces back to the rejection algorithm, a fundamental technique for generating samples from a probability distribution (Csilléry et al. 2010). The basic rejection algorithm entails simulating numerous datasets under a presumed evolutionary scenario. The parameters of this scenario are not deterministically chosen but are rather sampled from a prior probability distribution. Simulated data are next generated using parameter values and further condensed into sum-

mary statistics. Subsequently, the sampled prior values are accepted or rejected based on the disparity between the simulated summary statistics and the observed ones. The samples of accepted values approximate the full posterior distributions of parameters as they encapsulate the parameter values that best fit the data (Csilléry et al. 2010). Inference can also be achieved by replacing summary statistics through local linear or non-linear regression of simulated parameter values on simulated summary statistics (Beaumont et al. 2002). There are two pivotal factors that significantly impact ABC results: i) the quality of the models if the values of the parameters fall outside the range of the observed dataset, the resulting ABC outcomes may be of suboptimal quality; ii) the selection of appropriate summary statistics (Beaumont 2010). Recent advances, in particular the use of machine learning, have led to improvements in ABC methods, notably through the utilization of random forests as implemented in the abcrf package (Pudlo et al. 2016; Raynal et al. 2019). Random forests enable ABC to be a calibration-free problem – calibration refers to the process of choosing the most informative summary statistic and rejecting the misleading ones – and significantly reduce the number of simulations required to obtain robust estimates. Building upon *abcrf*, a tool named DILS has been developed to infer demography and detect barrier loci (Fraisse et al. 2021). DILS operates in two distinct steps. Initially, it infers the global demographic model from a pool of 14 models, considering not only classic demographic model parameters but also genome-wide heterogeneity in N_e (effective population size) and m_e (migration rate). Incorporating such heterogeneity has been demonstrated to enhance model inference (Roux et al. 2014). Subsequently, conditional on the best-fitting model, DILS infers the local best model at the locus scale, distinguishing between barrier and non-barrier loci models. However, its reliance on the first step of best-fitting model selection makes it challenging to compare results across species. The best-fitting model indeed conditions barrier detection and may vary across species.

1.3 Crop domestication as a step towards reproductive isolation

My PhD project was primarily motivated by the understanding of the mechanisms involved in the early set-up of barrier loci. The challenges here were double: I aimed to design a tool that efficiently detected barrier loci in highly challenging conditions (recent time splits, selfing species) but that also allowed cross-species comparisons to gain insights into general mechanisms. Within this framework, the biological models provided by crops and their wild relatives appeared to be particularly relevant, furnishing ideal empirical datasets to investigate questions related to RI. The upcoming section, co-authored and recently published in the American Journal of Botany, provides a concise explanation for this choice, outlining the anticipated benefits of these models in addressing key questions on RI while advocating for the development of new detection tools (M. I. Tenaillon et al. 2023).

1.3.1 Background

Speciation, Darwin's mystery of mysteries, is a continuous process that results in genomic divergence accompanied by the gradual increment of reproductive barriers between lineages. Since the beginning of research on the genetics of speciation, several questions have emerged such as: What are the genetic bases of incompatibilities? How many loci are necessary to prevent hybridization and how are they distributed along genomes? Can speciation occur despite gene flow and how common is ecological speciation? Early stages of divergence are key to understanding the ecology and genetics of speciation, and semi-isolated species where hybrids can still be produced are particularly relevant. Here we argue that the recent divergence between wild and domesticated lineages is an excellent model to capture the very-first steps of reproductive barriers formation, and will bring novel insights into the speciation process.

1.3.2 Why is domestication a good model to study speciation?

Domestication is the process of divergent selection between wild forms undergoing natural selection in their habitats, and domesticates evolving under combined natural and human-mediated selection. It has been increasingly recognized that evolution of domesticated species shows many similarities with evolution in the wild: it results primarily from changing environmental conditions and involves unconscious selection under a protracted process (Purugganan 2019) with selection intensities of the same magnitude or even smaller (Yang et al. 2019). Thus, domestication has been considered as a choice example to study adaptation. Here, we argue that it also offers an excellent opportunity to catch the very-first processes at work in ecological speciation, where adaptive divergence between nascent lineages triggers the onset of reproductive isolation (RI). Allele differentiation resulting from divergent selection can be measured by F_{ST}. F_{ST} between wild and domestic pairs range between 0.05 in sweet cherry and 0.51 in Tomato (Appendix S1, and references herein; see Supplemental Data with this article), which cover a wide range of divergence within a "grey zone of speciation" in which barriers to gene flow exist but are not complete (Roux et al. 2016). Interestingly, within this continuum, self-fertilizing taxa display greater genetic differentiation than outcrossers (Figure 1.7). As mating systems are predicted to affect the speciation process, domestication also offers the opportunity to address this question (Marie-Orleach et al. 2022). In contrast, life span seems to have no significant effect on divergence (Figure 1.7), although annuals and perennials experience contrasted domestication dynamics in many respects (Gaut et al. 2015).

The existence of reproductive barriers between wild and domesticated plants has been repeatedly documented. Despite the occurrence of wild-cultivated gene flow, the establishment of wild alleles into domesticated populations and reciprocally – introgressions – is rare (Ellstrand et al. 2013). Perhaps the best documented examples come from maize, where the introgression from the mexicana teosinte subspecies has contributed to highland adaptation of maize landraces (Calfee et al. 2021); and conversely, introgression from locally-adapted maize has contributed to teosinte adaptation in Europe (Le Corre et al. 2020). Interestingly, introgressions in

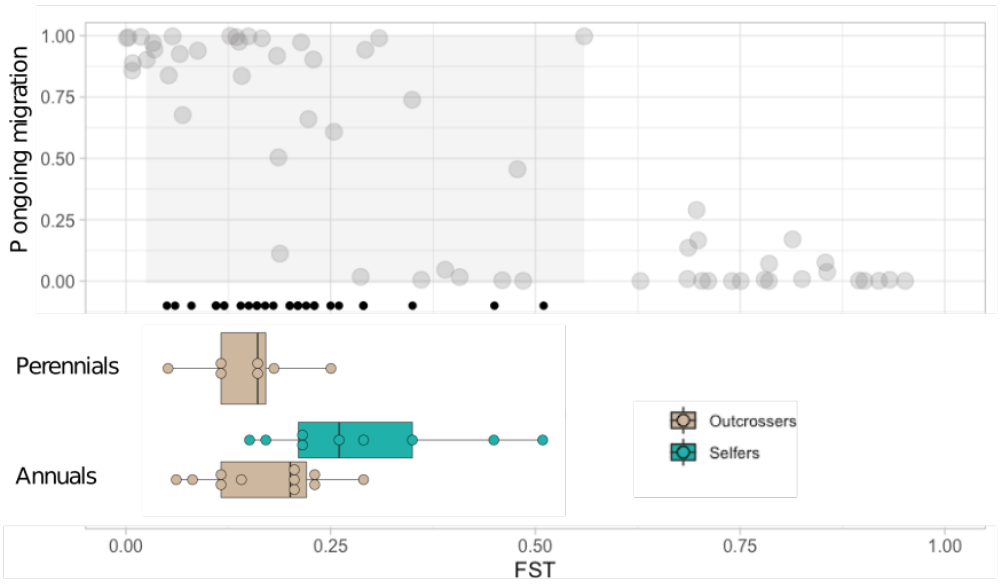

Figure 1.7: The grey zone of speciation as defined by Roux et al. (2016) encompasses the range of allele differentiation between wild and domestic forms across 27 plant species. Upper panel: Data illustrating the grey zone of speciation are taken from (Roux et al. 2016). F_{ST} values (x-axis) were computed for 61 pairs of animal populations/species across sequenced loci (natural divergence/speciation). The posterior probability of ongoing migration (y-axis) for a given pair reflects the capacity of demographic models that allow for ongoing exchange of migrants between diverging lineages to predict the observed data compared to models where gene flow has stopped. The light grey rectangle spans the range of F_{ST} values in which both currently isolated and currently connected pairs are found, and therefore defines the co-called grey zone of speciation. Lower panel: Black dots along the x-axis correspond to F_{ST} values obtained for 27 wild/domestic plant species. F_{ST} values for plant species were used to compute boxplots for annual (/biannual) species and perennial species. Boxplots are colored according to mating system.

the two directions are removed by selection around domestication genes (Le Corre et al. 2020; Calfee et al. 2021). This points to a prominent role of pre- and post-zygotic genetic barriers in the divergence of wild and domesticated lineages, and some genes involved in reproductive barriers have been identified such as the *Tcb1* locus in maize that governs pollen rejection by teosinte (Lu et al. 2019).

Whether partial isolation between wild and domesticated forms will ultimately result in full speciation is unknown. But clearly, partial RI does occur and has contributed to the maintenance of the distinct features between wild and domesticated forms, the so-called domestication syndrome. RI therefore stands as a major component of the domestication syndrome, but has been so far largely ignored (Dempewolf et al. 2012). It is even possible that reinforcement played a role in the establishment of the domestication syndrome, which involves the evolution of stronger RI due to the costs associated with producing low-fitness hybrids (Rushworth et al. 2022).

1.3.3 The genetic bases of reproductive isolation

The establishment of reproductive barriers can occur through various mechanisms. Selection leading to the fixation of advantageous alleles in different environments, resulting in local adaptation, can cause hybrid offspring to have lower fitness in parental environments, which strengthens isolation as populations adapt to differing conditions. This process may contribute to RI between wild and domesticated forms, and some crops may already be considered as independently evolving lineage once human-mediated cessation of gene flow is complete. Loci involved in such adaptation, those governing domestication traits, display a high degree of differentiation between wild and domesticated forms as well as a pattern of positive selection within forms compared with neutral loci (Figure 1.8). They contribute to limiting effective gene flow at nearby loci, leading to the progressive buildup of the so-called genomic islands of divergence (Wolf and Ellegren 2017).

RI may also be promoted by the buildup of intrinsic barriers from the differential fixation of alleles that are incompatible at two or more interacting loci (Bateson-Dobzhansky-Muller Incompatibilities – BDMIs). Such BDMIs can evolve as a by-product of local adaptation to contrasting environments or through non-adaptative processes (Wolf and Ellegren 2017). If selection favors distinct mutational steps at several loci in each population, deleterious side effect interactions may arise when brought together in hybrids. These interactions may in turn provoke detrimental symptoms and/or Transmission Ratio Distortions (TRDs) at F2 generation for recessive alleles, contributing to intrinsic post-zygotic isolation.

In domesticated forms, the accumulation of deleterious mutations through domestication bottlenecks and linked selection may have accelerated the evolution of BDMIs between wild and domesticated forms. The loci underlying BDMIs should display genomic fingerprints that can be similar to those left by selection for habitat adaptation in domestic or wild populations (Figure 2). However, in the absence of intra-form selection, we expect increased divergence between forms while the level of polymorphism is not affected by the cessation of gene flow

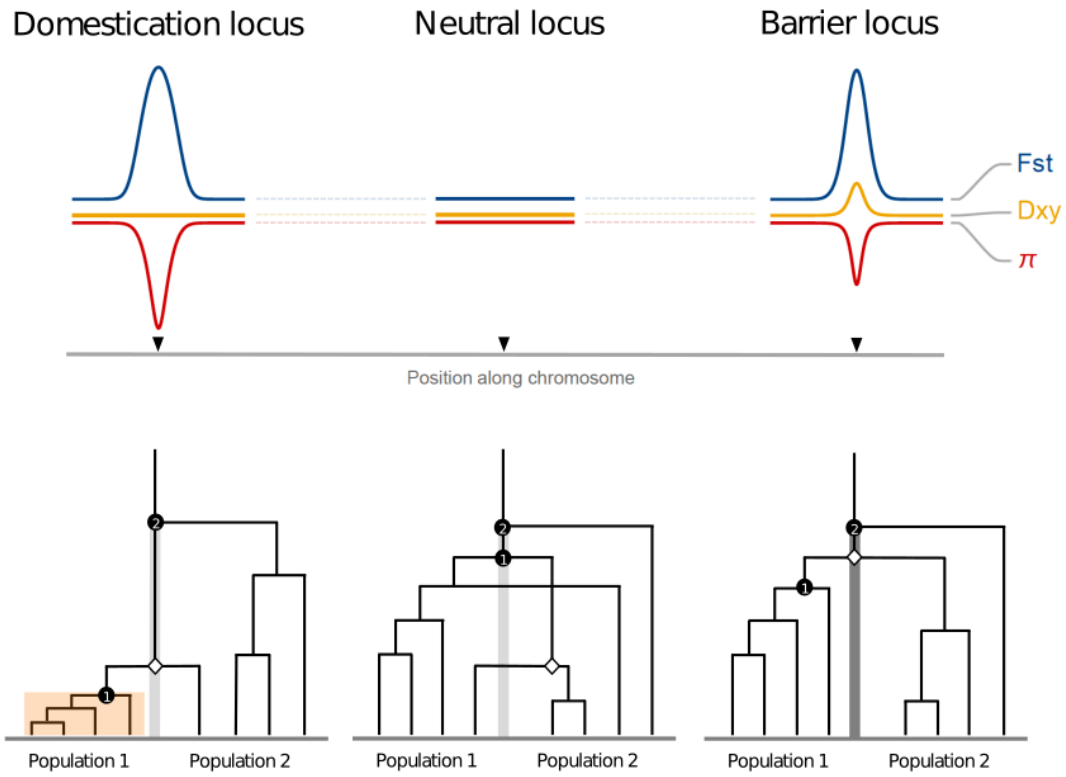

Figure 1.8: Theoretical expectations of summary statistics under divergence with gene flow in wild and domestic populations. Patterns of allele differentiation (F_{ST}), divergence (D_{xy}) and diversity within the domesticated populations (π) are displayed along the chromosome around three loci (black arrows) evolving under distinct scenarios: selective sweep at a locus involved in environmental adaptation and/or governing a domestication trait (domestication locus), neutrality (neutral locus), gene flow arrest at a locus that contributes to RI between populations (barrier locus). Representative genealogies of eight individuals from two divergent populations, a domestic population 1 and a wild population 2 are displayed (adapted from Hejase et al. (2020)). At a neutral locus and under continuous gene flow (light grey vertical bar), no allelic differentiation ($F_{ST} = 0$) is observed between populations that behave as a single population. Allelic differentiation ($F_{ST} > 0$) can be initiated either because the time to the most recent common ancestor – T_{MRCA}s are represented by black circles for population 1 and 2 – is reduced by a selective sweep (light orange rectangle) in one of the two populations, in this case the domestic population 1; or because the time to the first cross-coalescence between the populations (diamond) is increased by selection against gene flow (barrier locus, solid vertical bar). Note that in all graphs, the T_{MRCA} of the population 2 is also the T_{MRCA} of wild and domestic populations.

(Figure 1.8). This illustrates how the use of different statistics helps to clarify the mechanisms at work in RI (Cruickshank and Hahn 2014). There is of course a continuum of scenarios between those presented above: the barrier loci that limit gene flow between wild and domesticates can be directly targeted by selection within one of the two forms.

A particular kind of negative epistatic interaction can emerge as a by-product of coevolution between nuclear and cytoplasmic genomes. If different combinations have coevolved in divergent lineages, this may result in organelle dysfunction and hybrid breakdown when inter-lineage crosses occur (Burton et al. 2013). Such negative cytonuclear conflicts often result in asymmetrical reproductive barriers, which can be revealed by reciprocal crosses between wild and domesticated lineages as observed in Citrus (Wang and al 2022).

Finally, reproductive barriers may result from parental conflicts generating allelic dosage perturbations. If species have evolved contrasting levels of parental conflicts, it can translate into paternal or maternal excess of gene expression in a hybrid context (Florez-Rueda et al. 2016). As evidenced by transcriptomic comparisons of wild and domestic forms (e.g., common bean, Bellucci et al. (2014); tomato, Sauvage et al. (2017)) and simulations (Burban et al. 2022), domestication led to a profound reorchestration of coexpression networks, which can then cause disruptions in allelic dosage between wild and domestic forms resulting in fitness decline in wild x domesticated crosses.

1.3.4 Hypotheses testing & challenges

Do the number of generations since domestication correlate with hybrid defects? Does the mutation load depend on the domestication history and the strength of RI? Do "stronger" domestication syndromes and/or higher genome-wide neutral divergence and/or extent of islands of differentiation induce stronger isolation? Answering these questions will bring unique insights into the very-first steps of reproductive barriers formation, but detecting barrier genes is a daunting task. Divergent selection and BDMIs among loci create patterns of strong allelic differentiation relative to the genomic background (Figure 1.8) that together with linked loci, form genomic islands of differentiation. Their detection requires overcoming confounding effects such as local variation in recombination rates and effective population size (Wolf and Ellegren 2017). At the species level, parameters such as mating system, intensity of domestication and changes in effective population size (e.g., due to domestication bottlenecks) determine the extent of selection, genetic drift and linkage disequilibrium, and in turn the expected size and depth of islands of differentiation. Ultimately, interpretations of genomic differentiation patterns need to be guided by modelling in order to properly estimate the fraction of the genome recalcitrant to gene flow and identify the corresponding regions, which can then be combined to experimental results.

1.3.5 Conclusion

The alterations of habitats due to human activities are precious laboratories to explore the mechanisms involved in adaptive divergence and the initial phases of speciation (Thompson et al. 2018; Touchard et al. 2023). The establishment of RI between wild and domestic forms is a crucial aspect of domestication that has received little attention. In addition to providing basic information about the processes at work in the early stages of speciation, testing for cross-compatibility between cultivated plants and their wild relatives and detecting the underlying barrier loci are essential for overcoming them. Crops wild relatives have faced continuous environmental challenges in their natural environment and often exhibit greater genetic diversity than their domesticated relatives, so they are a valuable reservoir of adaptive alleles that transferred to crops could help mitigate their vulnerability.

1.4 Thesis objectives

Understanding the genetic mechanisms underlying reproductive isolation is a primary objective in speciation research. Analyzing diverging populations is a common approach, but capturing the sequence of events that lead to reproductive barriers remains challenging. One promising avenue involves comparing populations at varying levels of temporal and/or spatial divergence, including recently diverged ones. Achieving this necessitates a comparative framework capable of detecting gene flow barriers at different evolutionary stages across diverse biological systems, regardless of their demographic history. The introduced method, RIDGE (Reproductive Isolation Detection using Genomic Polymorphisms), aims to address this need. The first chapter of the thesis - accepted manuscript - consists of a description of RIDGE and an evaluation of its performance in detecting gene flow barriers both by simulations and on empirical datasets from crow species. The second chapter is more technical and provides a detailed manual for RIDGE while supporting the choices that have been made following extensive trials and improvements made to the pipeline - the manual and code are open source and available online at `https://github.com/EwenBurban/RIDGE.git` The third chapter presents contrasted application of RIDGE on domesticated crops, encompassing a range of biological contexts, mating systems, and divergence histories, leading to change in RIDGE performances in barrier detection. This chapter explains how to evaluate the results and eventually how to improve them.

Chapter 2

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

This chapter consists of our recently accepted article in Molecular Ecology Resources, which describes RIDGE and evaluates its performance at detecting gene flow barriers by simulation and on empirical datasets. See Appendix A for additional figures and tables.

Ewen Burban, Maud I. Tenaillon, Sylvain Glémin

2.1 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 datasets, 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, a substantial portion of previously identified outliers might be false positives. The utilization of RIDGE for outlier detection accommodates a diversity of demographic scenarios, and relies significantly 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 one that can be useful in challenging conditions.

Keywords: Speciation; Reproductive isolation; gene flow barrier detection; approximate bayesian computation; Hybrid zones; Crows.

2.2 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; Fraisse 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 and Nosil 2005; Schluter 2000; Shafer and 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 and 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 color patterns in carrion and hooded crows (Metzler et al. 2021; Poelstra et al. 2014). Specifically, genes encoding feather pigmentation and genes responsible for perceiving color 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 *melopomene*, assortative mating patterns correlate with a genomic region proximate to optix, a crucial locus influencing distinct wing color 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 **B**ateson-**D**obzhansky-**M**uller **I**ncompatibility (BDMI) (Gavrilets 2003). Across *Arabidopsis thaliana* strains, epistatic interactions between alleles from two loci located on separate chromosomes, resulted 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 and 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 (B. Charlesworth et al. 1993; Charlesworth and Jensen 2021; 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 and 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 (M. I. Tenaillon et al. 2023).

Linked selection (at least some forms of) can be approximated by a local reduction in effective population size (Cruickshank and Hahn 2014; Ravinet et al. 2017; Sakamoto and 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 (Fraisse 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 relying four demographic 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). This account of genomic heterogeneity has been shown to enhance the quality of model inferences (Roux et al. 2014). Second, the method infers the migration model at the locus scale – arrest of migration vs migration similar to the genome-wide level –, conditioned on the chosen model (Fraisse et al. 2021). Although effective in detecting gene flow barrier, this reliance on the 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 modeling the mosaic effect of linked selection as in the DILS program (Fraisse et al. 2021), and heterogeneity in recombination, by including an option for the user to provide a recombination map. In addition, we 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.

2.3 Material and Methods

2.3.1 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 (Fraisse 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 x 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) dataset. 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 datasets for control loci (thereafter non-barrier) and barrier loci that have undergone no gene flow during divergence. Simulated datasets 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 v 4.1.2 (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 and Mundt 2017) and latex2exp (Meschiari 2023).

ABC Summary statistics

ABC inferences rely on summary statistics that are computed either at the locus-level or across loci i.e. genome-wide distributions of summary statistics and correlations among loci, and either within- or between- populations. For a given observed dataset, the number of loci used for construction of the hypermodel is set by the user. To reduce computation time for large datasets, 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 and Li 1979b), Watterson θ (Watterson 1975b), 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 (D_{xy}) and the net (D_a) divergence (Nei and Li 1979b), the summary of the joint Site Frequency Spectrum (jSFS) (Wakeley and 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 D_{xy} and F_{ST} and between D_a and F_{ST}. Regarding specific jSFS status, 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, (Fraisse 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}, D_{xy}, sf, D_a and ss while falling below certain thresholds for π and θ. The thresholds are determined using Tukey's fences: $t_{min} = Q_1 - 1.5 * (Q_3 - Q_1)$ and $t_{max} = Q_3 + 1.5 * (Q_3 - Q_1)$, for the lower and upper thresholds respectively, where Q_1 is quantile at 25% and Q_3 the quantile at 75% (Tukey 1977). All summary statistics are computed using the python packages scikit-allel (Miles et al. 2021) and numpy (Harris et al. 2020).

Coalescence simulations

We simulated the evolution of neutral loci (1000 by default) under 14 demographic x 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*.

Demographic models RIDGE simulates the split of a single ancestral population of effective size N_a, in two daughter populations of size N_1 and N_2 at time T_{split}. Four different demographic models are considered as in DILS (Fraisse et al. 2021) (Figure 2.1: Demographic and genomic models): (1) strict isolation with no migration (SI), (2) isolation with constant migration rate since (IM), (3) secondary contact with no migration after the split until a secondary contact at time occurs (SC), and (4) ancestral migration with migration occurring initially and ceasing

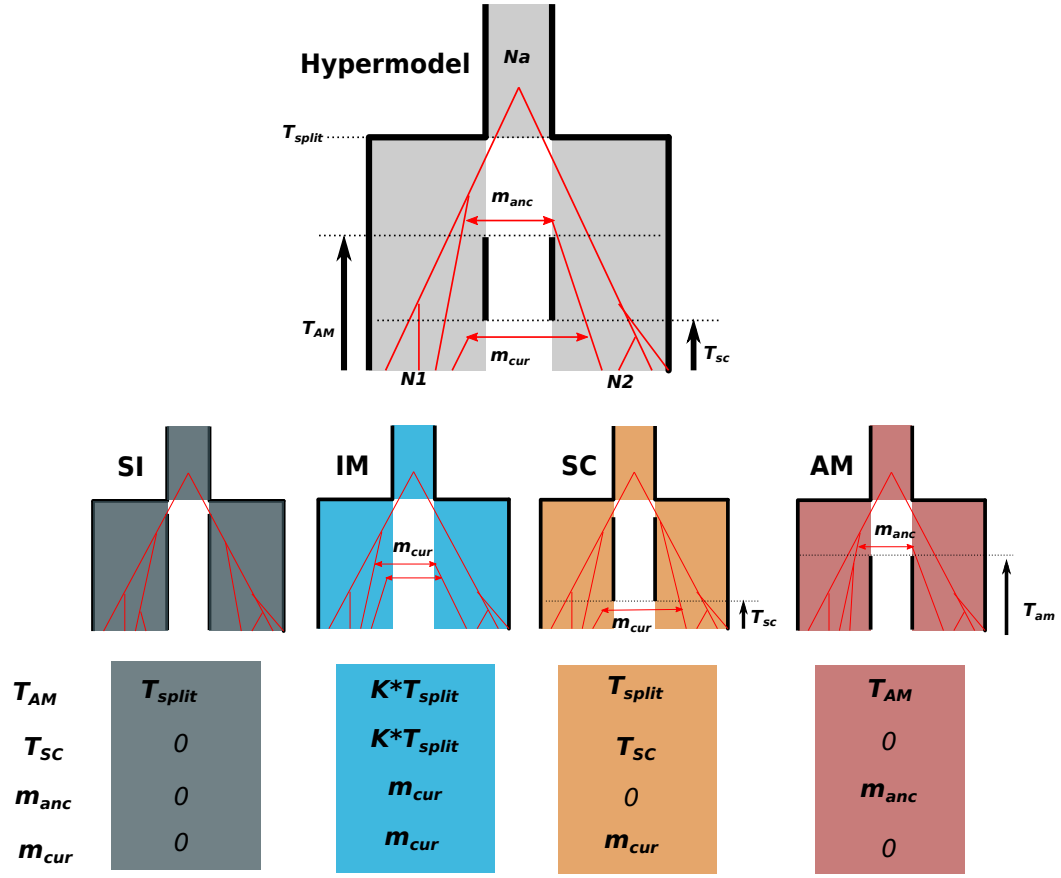

Figure 2.1: Demographic models implemented in RIDGE. The hypermodel combines all four demographic models considered: Strict Isolation (SI), Ancestral Migration (AM), Secondary contacts (SC) and Isolation-Migration (IM) plus genomic models. In the hypermodel, an ancestral population of effective size 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.

after time (AM). Migration rate m is assumed to be symmetrical between the two populations.

Genomic models In addition to modeling 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 modeled by a rescaled Beta distribution. Effective size at locus i is given by:

$$N_i = \bar{N}.\frac{\alpha+\beta}{\alpha}.B(\alpha,\beta) \tag{2.1}$$

where $B(\alpha,\beta)$ is a Beta distribution with parameter α and β and $\bar{N}$ is the mean effective population size across the genome. In other words, under $2N$ and for a given locus, three independent values are sampled from the same $B(\alpha,\beta)$ distribution albeit distinct $\bar{N}$ are used in equation 2.1 so that there is no covariation of the effective population size across populations. For migration (m), the genome-wide heterogeneity in effective migration is modeled by a Bernouilli distribution where a proportion Q of loci displays $m = 0$ and a proportion $1-Q$ loci displays $m > 0$, m designating either the current migration (m_{cur}) or the 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 $M = 4 * N_e * m$, and 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, in hetero-N and homo-m 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 $1/4N_e$ 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.

Generation of the *reference table*

RIDGE explores 14 demographic x 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 2.1, and four genomic parameters ($\alpha, \beta, Q_{anc}, Q_{cur}$).

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 A.1, Figure 2.1). For example, under SI, only four demographic parameters are estimated (Table A.1). 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} 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.

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 (Csillery 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 dataset 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 dataset and those computed from each η simulation contained in the posterior table. Together it form a vector of Euclidean distances $d_1 \dots d_\eta$ on which we computed the average, denoted D_{post}. To derive the null distribution of G_{post}, we considered a dataset randomly sampled in the posterior table as "observed" and discarded from subsequent analyzes. The remaining $\eta - 1$ datasets of the *reference table* were used to compute D'_{post}, the average Euclidean distance between the posterior table and the "observed" dataset. Repeated as such Z times, we obtained a vector of ${D^1_{post}}' \dots {D^Z_{post}}'$. Then we computed G_{post} as the proportion of values for which $D'_{post} > D_{post}$

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:

$$BF_i = E\left[\frac{1-Q}{Q}\right] . \left(\frac{P[X_1]}{1-P[X_1]}\right) \tag{2.2}$$

Here, E[] represent 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):

$$BF_i = \left(\frac{E[1-Q]}{E[Q]} + \frac{V[Q]}{E[Q]^3}\right) . \left(\frac{P[X_1]}{1-P[X_1]}\right) \tag{2.3}$$

2.3.2 Evaluation of RIDGE performance on pseudo-observed datasets

We evaluated RIDGE performance on pseudo-observed datasets (i.e., simulated datasets considered as "observed" data and compared with simulation outputs to validate the accuracy and reliability of the simulation models). As a first step, we evaluated the ability of RIDGE to correctly infer demographic x genomic models. We next used the pseudo-observed datasets 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 datasets under the four demographic models and under both $2N2m$ and $2N1m$ genomic models (only $2N$ for SI). For simplicity, we fixed $N_a = N_1 = N_2 = 50000$ individuals. The time of the secondary contact (T_{SC}) was set to $0.2 * T_{split}$ and the time of arrest of ancestral migration (T_{AM}) was set to $0.7 * T_{split}$. We used a range of parameter values (Table A.2) for divergence (from 1000 to 2 million generations, i.e., from 0.1 to 20 in $2N_e$ generation unit), for migration (mean $M = 4N_em = 1$ 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 datasets using the *scrm* coalescent simulator (Staab et al. 2015). Each multilocus dataset contained 1000 loci of 10kb each, and we performed 100 replicates per scenario. To evaluate the inference of demographic x 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).

2.3.3 Application to experimental data on crow hybrid zones

To assess the performance of RIDGE on experimental data, we focused on two published datasets produced by Poelstra et al. (2014) and Vijay et al. (2016). All sequencing data from crows were extracted from 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 dataset 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 Organising 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 datasets 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 A.3, and were determined based on the observed datasets 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 analyzed RIDGE results produced on three species pairs

on a lager dataset (Vijay et al. 2016) to understand how BF correlate with summary statistics and which summary statistics are able to discriminate outlier loci ($BF > 50$).

2.4 Results

2.4.1 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 datasets (Figure 2.2 & A.1). However, the model's contribution to the estimation of the demographic and genomic parameters varied across conditions. The 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 2.3). 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 2.3 and A.2). 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 $2N2m$ utilizing IM $2N2m$ (the "correct" model") as a control. Our results revealed a significant impact of model misspecification on G_{post} for $T_{split} = 1.10^6$ (Figure A.3A). More importantly, the AUC fell below 0.5 and exhibited a sharp decrease for oldest when AM model was chosen (Figure A.3B). 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 (97.6% and 98.4% at 10^6 generation for IM and SC against 5.3% for AM, Figure 2.3). Heterogeneous migration ($2m$) was better captured under ongoing rather than ancestral migration but even under the most favorable conditions, 25% of the simulations exhibited consistent patterns of homogeneous migration where barriers were undetectable (Figure 2.3). 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 2.3). 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 $\hat{T}_{AM}$ and $\hat{T}_{SC}$ albeit T_{SC} tended to be slightly overestimated (Figure A.5). As increased, estimates of current population sizes $\hat{N}_1$ and $\hat{N}_2$ improved,

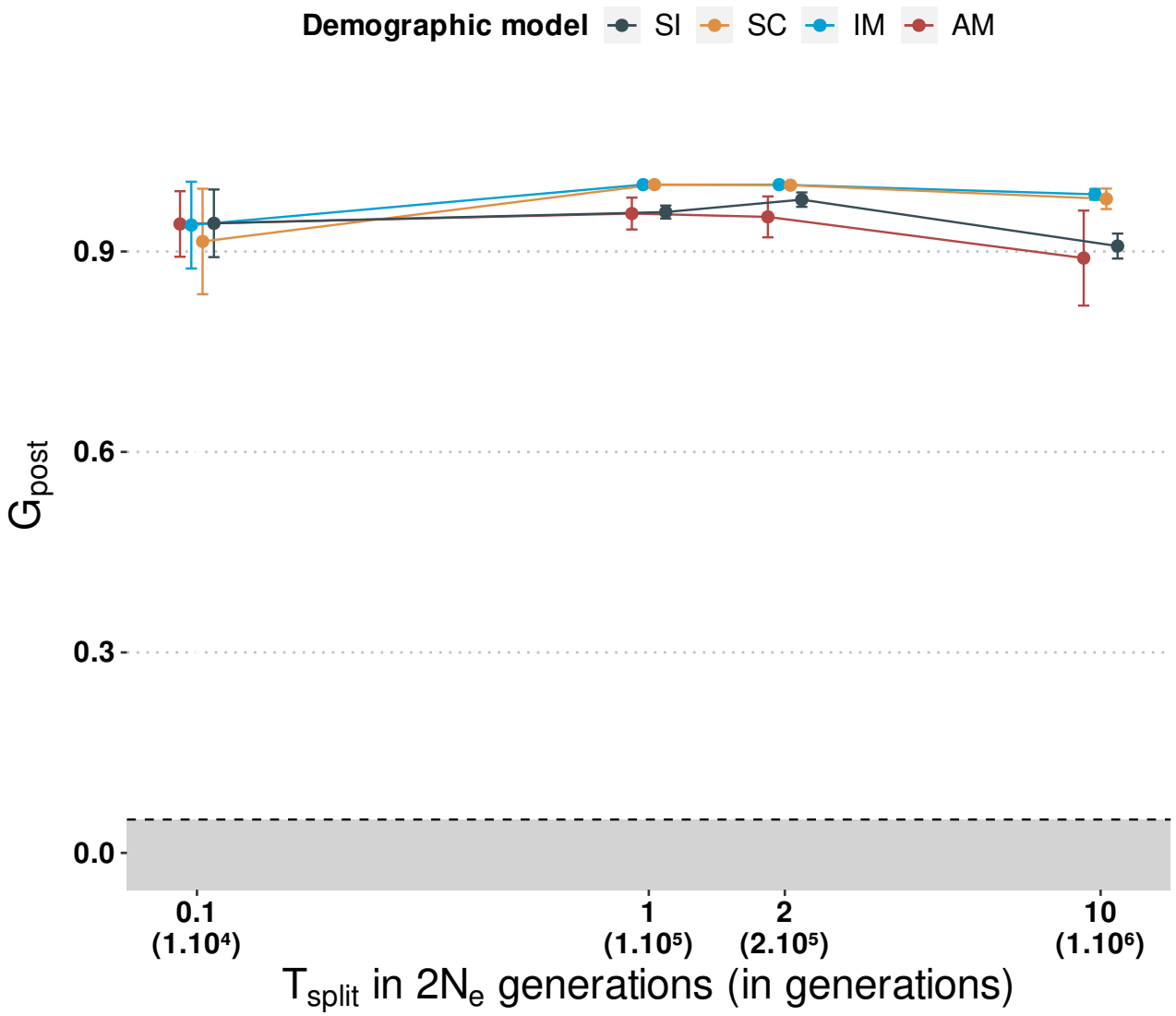

Figure 2.2: Evolution of the goodness-of-fit of the posteriors (G_{post}) as a function of time split, for four demographic models. The rejection threshold of 5% (under which an inferred model is discarded) is represented by the gray zone. Average values over 100 replicates with error bars (standard deviation) are presented. The data used in this figure were obtained from pseudo-observed datasets 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).

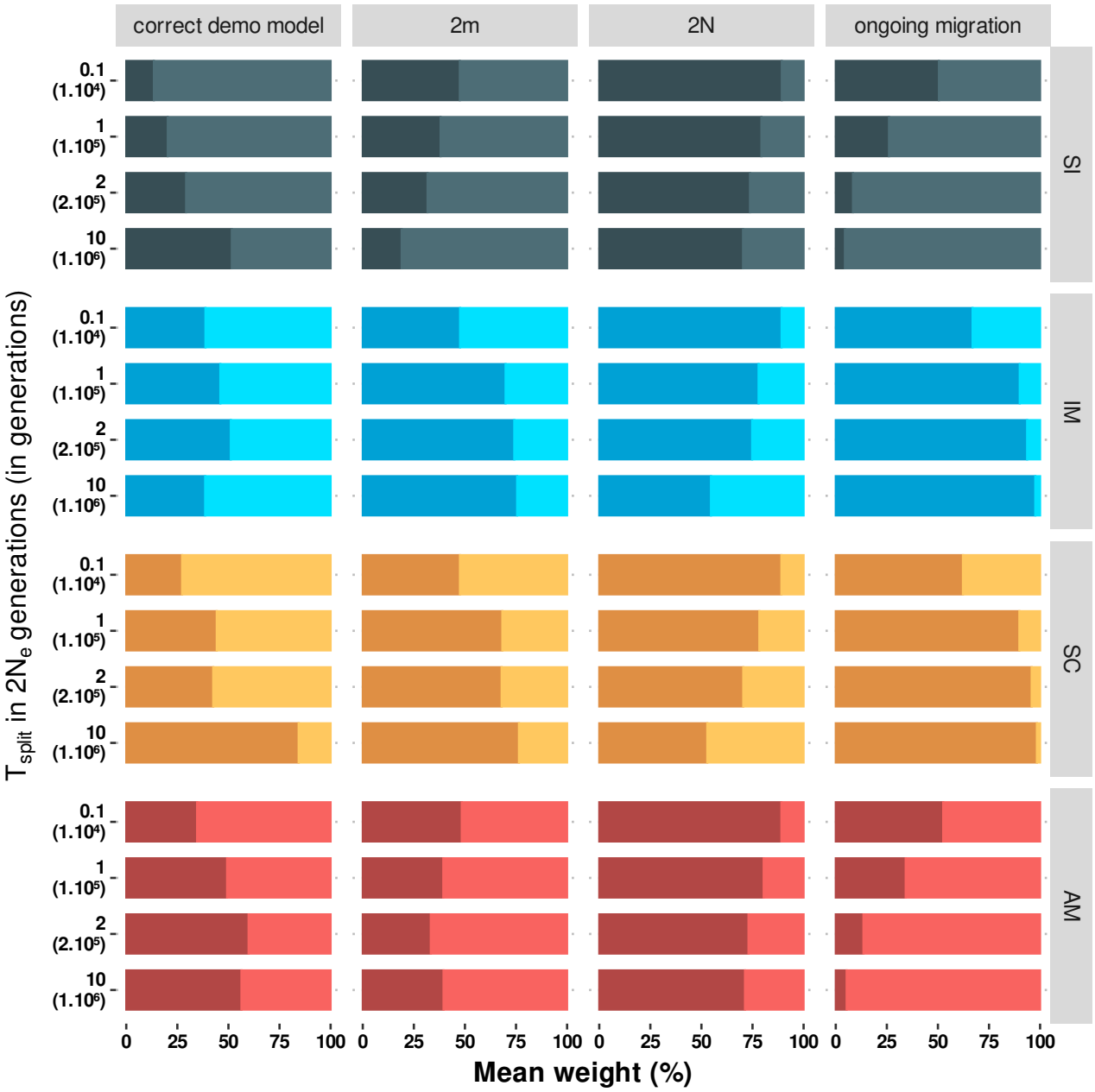

Figure 2.3: Demographic x genomic model weights in posteriors across time splits. Weight was measured by considering four criteria: i) the average joint weight of the false and true demographic (among the fours) model –called here the "correct" model– in posteriors, ii) the average joint weight of $1m$ 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 colors. 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 $4N_e m_{curr}$ or $4N_e m_{anc} = 1$.

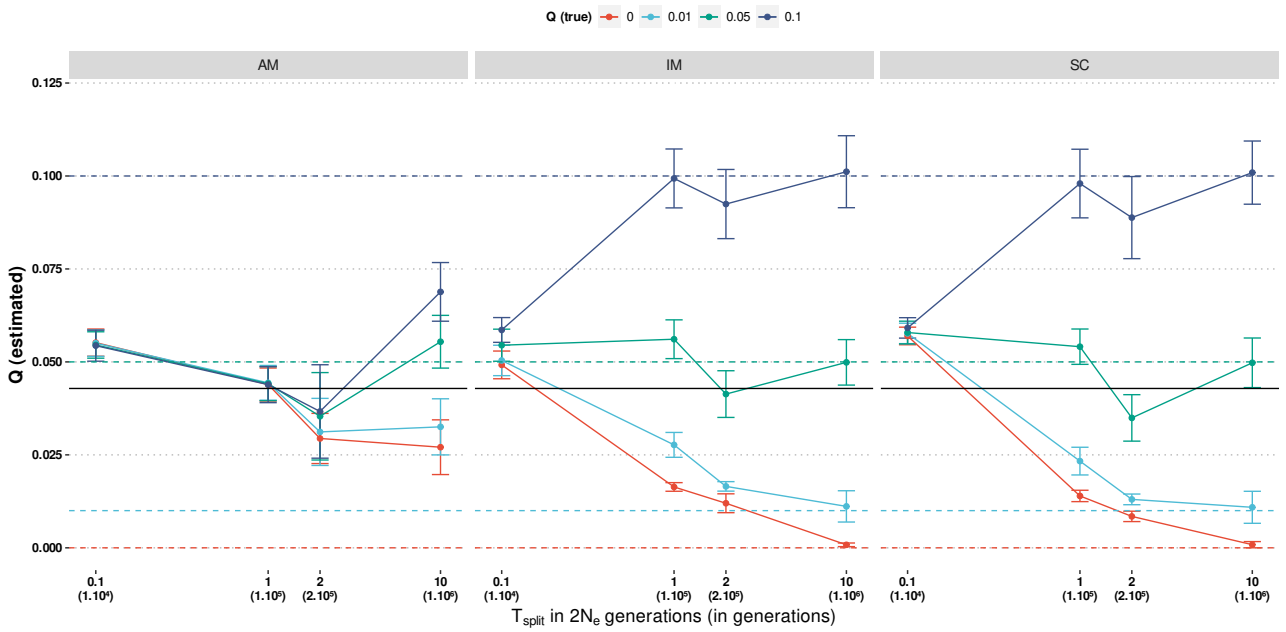

Figure 2.4: Barrier proportion estimates as a function of divergence time under three demographic models. In this figure, migration is set to $M = 10$ and the plain black line represents the priors mean. Each data point represents the average value over 100 replicates with standard deviation as error bars. Results overall conditions explored are represented in Figure A.8.

approaching simulated values when T_{split} reached 1.10^5 generations (Figure A.6). Estimates of past population size $\hat{N}_a$ is theoretically possible if $T_{MRCA} \leq 4N_e$ in each diverging population (with T_{MRCA} the coalescent time of the Most Recent Common Ancestor). When $T_{split} >> 4N_e$ most sequences are expected to coalesce before T_{MRCA} 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 A.6). Current migration estimates ($\hat{M}_{cur}$) were more reliable than ancestral migration ones ($\hat{M}_{anc}$). The proportion of barriers had minimal impact on $\hat{M}_{anc}$, under SC and IM models (Figure A.7). Deeper T_{split} resulted in greater migration signal and therefore improved the accuracy of $\hat{M}_{cur}$ (Figure A.7 & Figure A.8 left). In contrast, T_{split} had no clear effect on $\hat{M}_{anc}$ (Figure A.8 & A.9).

2.4.2 Inferences of barrier proportion

The barrier proportion estimate,$\hat{Q}$, plays a crucial role in the computation of Bayes factors (Eq 2.2) 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 2.4 & A.10). For more recent T_{split} ($< 0.2\ 2N_e$ 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 A.8 and, reciprocally M had little impact on $\hat{Q}$ (Figure S10), so that 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}, D_a, and D_{xy}. We therefore introduced additional summary statistics based on the proportions of outliers for F_{ST}, D_a, D_{xy}, sf, ss 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 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 A.11). The impact of outlier statistics varied across models and T_{split} values (Table A.4). 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 D_a 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 ($r > 0.88$). At $T_{split} = 1.10^6$, $\hat{Q}$ additionally correlated with D_{xy} for all models (Table A.4).

2.4.3 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 2.5A. As 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 2.5A & 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 A.12). Note that, as expected, the SI model produce a single mode distribution irrespective of T_{split} where all loci become barriers as T_{split} increases (Figure A.12). 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 2.5C. When 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 A.12). 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 2.6). Moreover, barriers were more efficiently detected and at lower T_{split} under current (IM and SC models) than ancestral gene flow (AM model) as shown in Figure A.10 & A.11. Noteworthy, the AUC never dropped below 0.5, indicating that RIDGE did not generate an excess of false positives (Figure A.13 & A.14).

2.4.4 Detection of barrier loci on crow datasets

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.

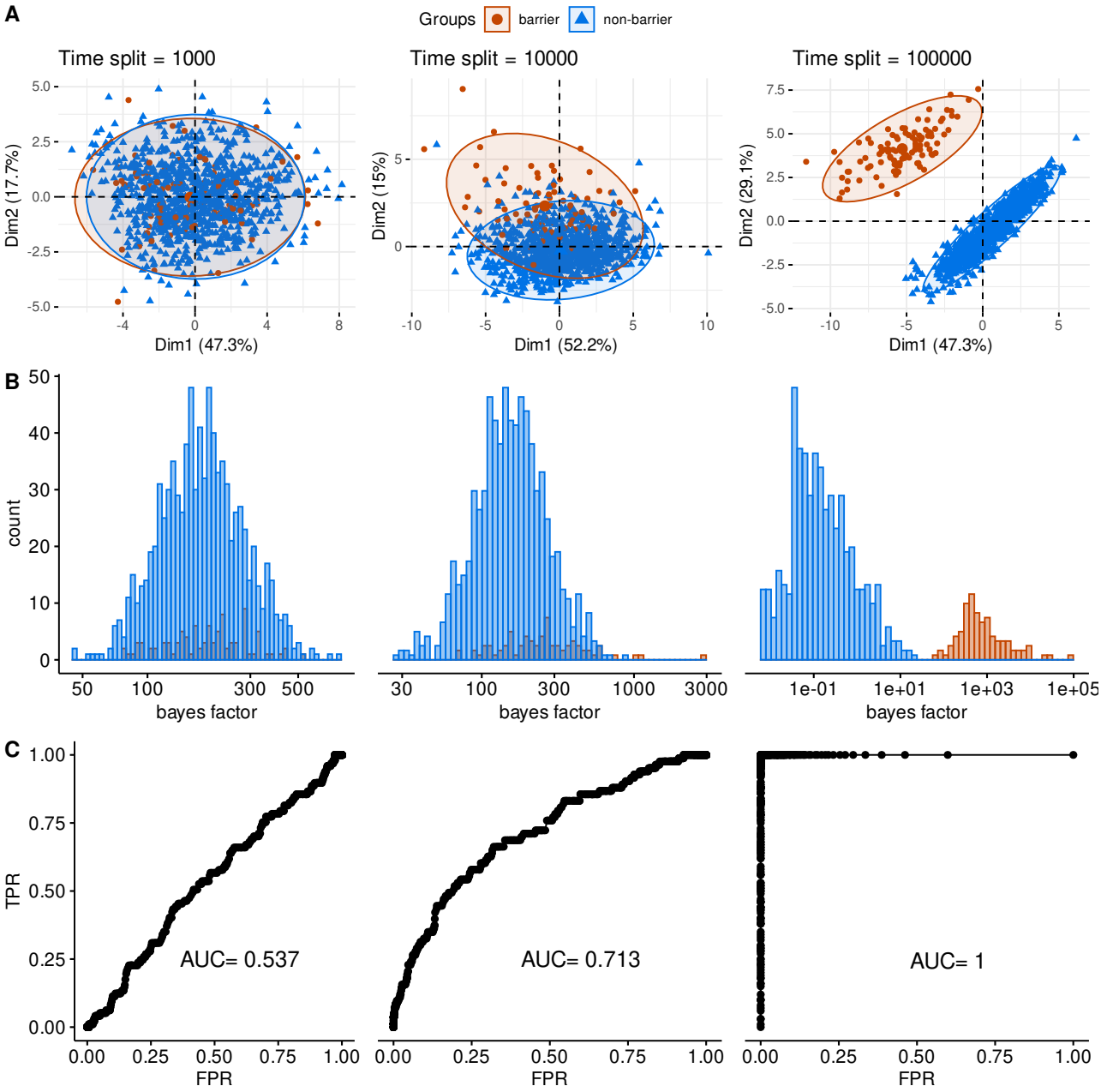

Figure 2.5: Impact of the divergence time on the overlap between barrier and non-barrier loci. Overlap revealed by a principal component analysis (PCA) computed on all 14 summary statistics (A), the log of the bayes factor (BF) produced by RIDGE (B) and the area under the ROC curve (AUC) of the bayes factor (C). The greater the AUC the higher the discriminant power is. A single pseudo-observed dataset was used for each of the three values of T_{split} Datasets were simulated under an IM $2N2m$ model, with the following parameters: $4Nem = 10$ and $Q = 10\%$.

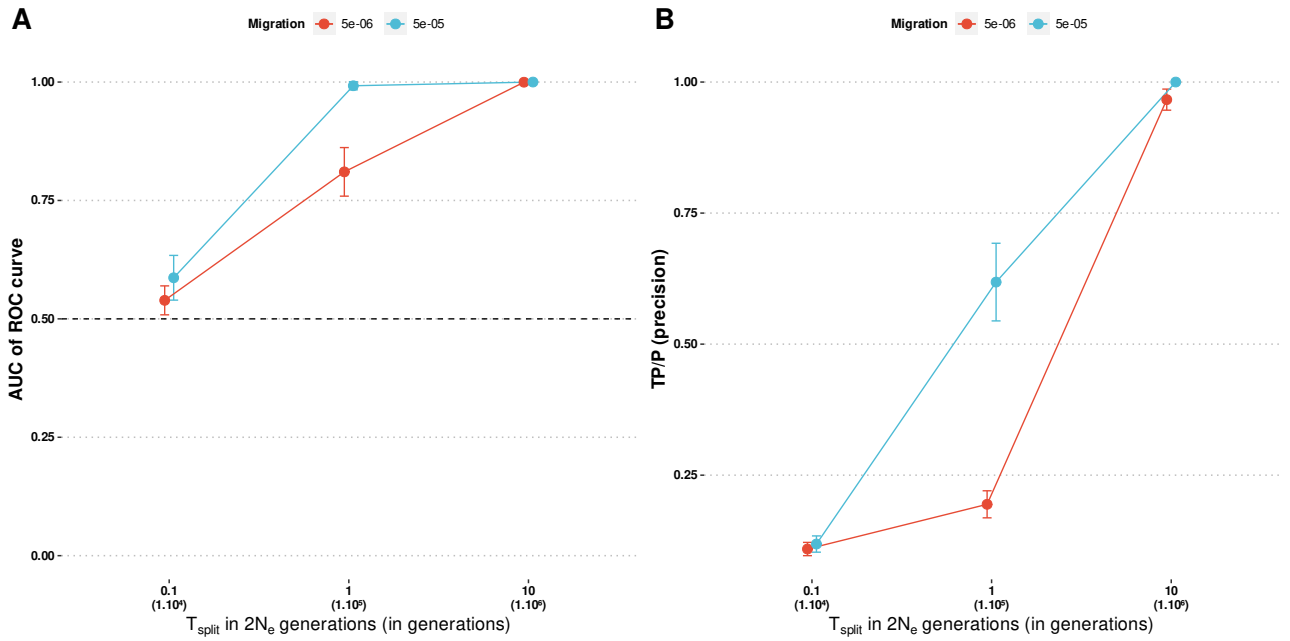

Figure 2.6: Ability and precision in the detection of barrier loci as a function of divergence time and migration. Ability is measured by the AUC of the ROC (A) and precision by TP/P (B). Loci are considered as barrier when theyr posterior probability of barrier model $P[X_1] > 0.5$. Each data point represents the average value over 100 replicates with standard deviation as error bars. Simulations were performed under an IM $2N2m$ model with $Q = 10\%$

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 color phenotypes and color-assortative mate choice. We ran RIDGE on the same dataset 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 A.5), indicating that we were within a favorable 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 2.7B). 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 2.7C). 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 dataset 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 seconds of CPU runtime on four CPUs running at a minimum of 2.5GHz. Therefore, in real-time, it took around 36 hours for the whole dataset 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 $\hat{T}_{split}/2\hat{N}_e = 0.25$ was approximately 0.3 for all three pairs (RX: 0.28; XO: 0.27; OP: 0.31; Table A.5), suggesting

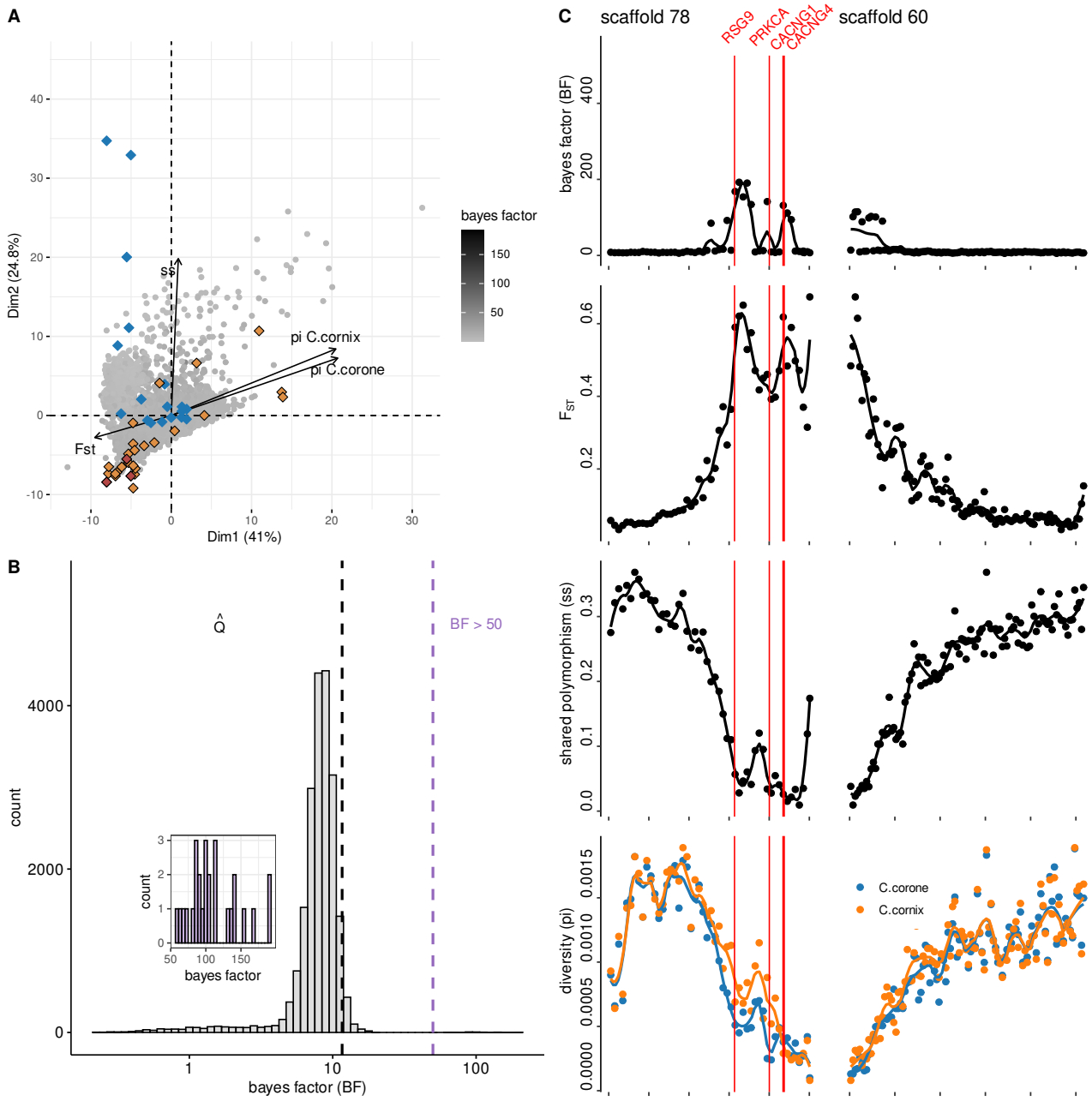

Figure 2.7: Results of the analysis conducted using RIDGE on the crow hybrid zone between carrion and hooded crows. PCA computed on summary statistics obtained from 50kb-windows along genomes with axes 1 and 2 (only 4 of 14 summary statistics are represented), where each datapoints (windows) are colored according to the values of Bayes factors (A). Blues 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 fromPoelstra et al. (2014)

a comfort zone for RIDGE to detect gene flow barriers in all three datasets.

PCA analyses colored by BF show a main group of outliers (characterized by elevated F_{ST} and/or D_a and/or reduced level of diversity in all four pairs Figure 2.7A & 2.8 & A.15). Those signals were consistent with some theoretical expectations for gene flow barriers (i.e., increased D_a, sf, F_{ST}, and reduced ss and diversity), but almost no relationship with D_{xy}. In each pair, we identified a subset of loci with elevated Bayes factors ($BF > 50$) clearly separated from the genome-wide distribution (Figure 2.8C). 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 2.8B & A.5).

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 2.8A & 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 D_a being consistently strongly positively correlated with BF and ss being consistently negatively correlated with BF but to a lesser extent (Figure 2.9).

2.5 Discussion

A key goal of speciation research is to elucidate the genetic mechanisms behind reproductive isolation. Although diverging populations have been analyzed 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.

2.5.1 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 the genome. DILS (Fraisse 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 provides corresponding genome-wide estimates. While not primarily designed to perform barrier detection, DILS can still provide valuable insights on potential barrier loci, conditioned on the selected demographic model (Fraisse et al. 2021). There are however two main limits to this approach. Firstly, selecting a model can be rather

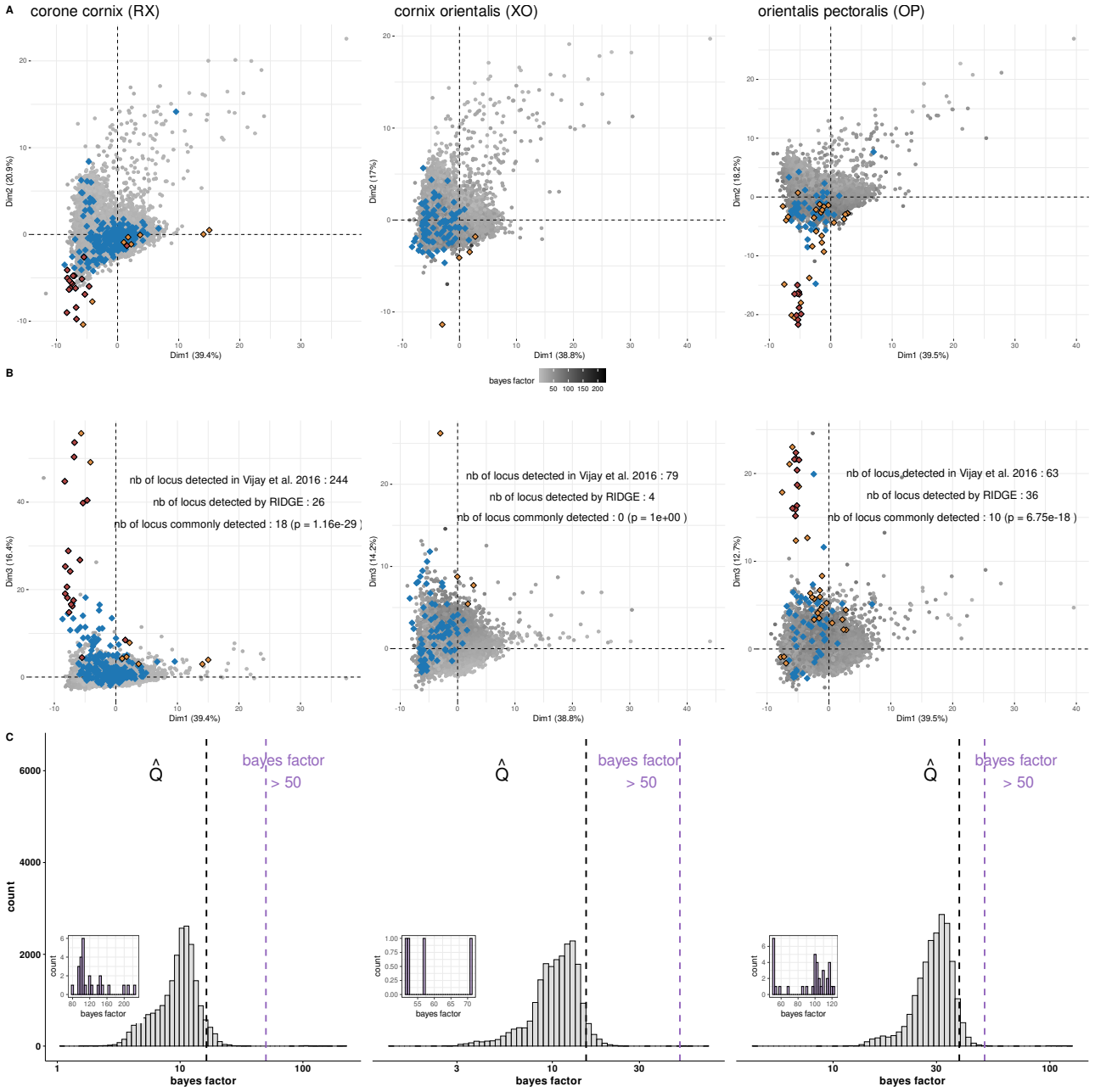

Figure 2.8: Barrier loci detection by RIDGE on three crow hybrid zones. PCA computed on summary statistics obtained from 50kb-windows along genomes with axes 1 and 2 (A) and 1 and 3 (B) displayed. Datapoints (windows) are colored according to the values of Bayes factors. Blue diamonds represent loci 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 line indicate the Bayes factor threshold based on the estimated barrier proportion $\hat{Q}$. Data are from Vijay et al. (2016).

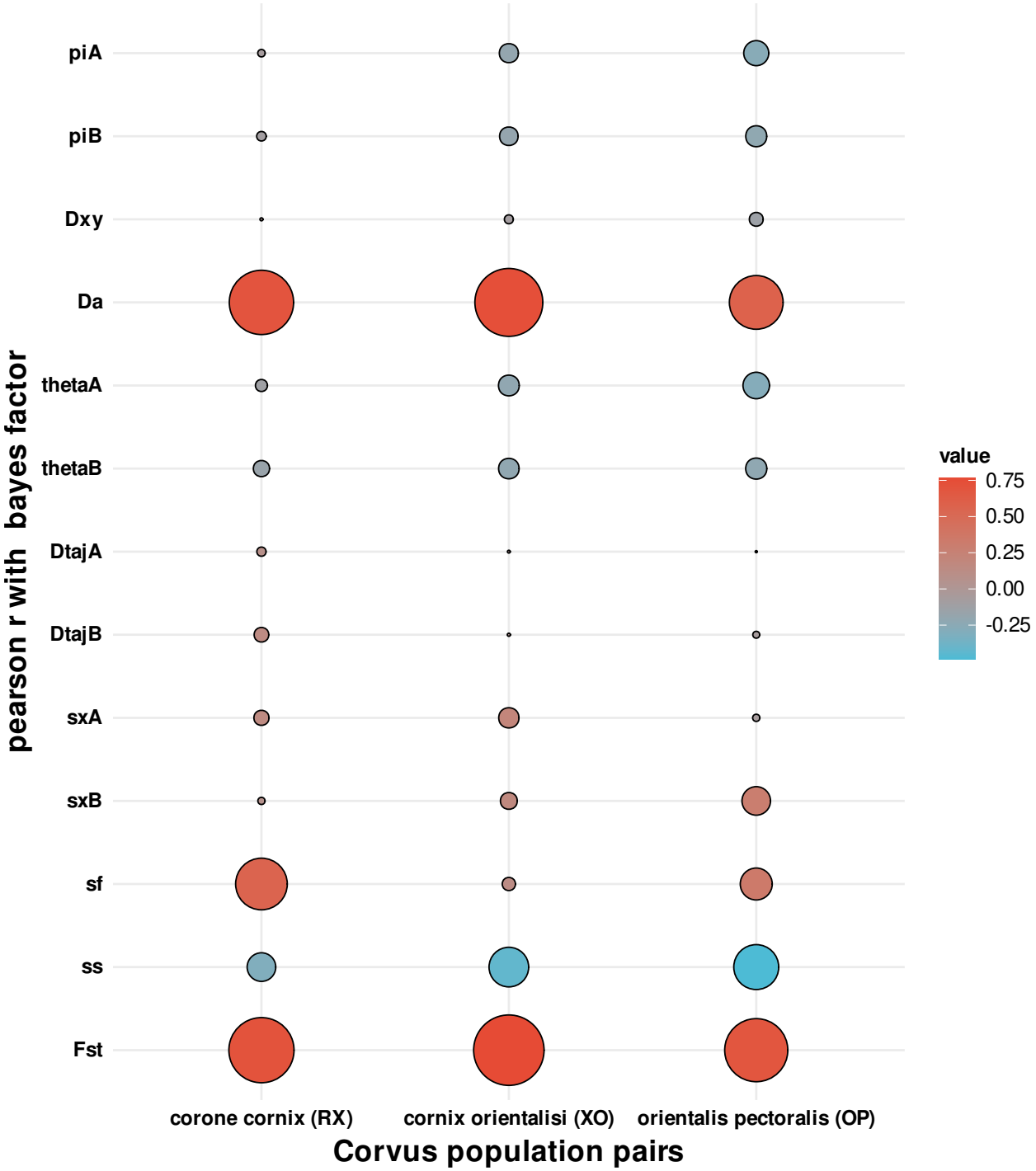

Figure 2.9: Pearson correlation between RIDGE Bayes factor and summary statistics used in the gene flow barrier detection for the three hybrid zones. Colors correspond to the values of correlations while circle size reflects the absolute values. Data are from Vijay et al. (2016).

arbitrary when two models explain the data equally well, which is often the case when divergence is shallow between populations (as shown in Fraisse et al. (2021) and confirmed here, Figure 2.3 and A.2); and model misspecification can have strong consequences on the rate of false positives (Figure A.3). 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. 2016; Camille 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 2.3). Opting for the best scenario under such conditions might be misleading. For example, at $T_{split} = 0.1 * 2N_e$, when current migration is simulated (IM or SC models), it is detected in only 60% of the cases (Figure 2.3), 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 datasets 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 modeled. DILS models separately the heterogeneity in N_e and $M = 4N_em$, 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 datasets where migration is modeled as in DILS (constant or variable M instead of m, independently of N_e). 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 datasets, lacking a clear interpretation for the RX and XO pairs.

A direct consequence of using a demographic x genomic hypermodel is that RIDGE is not intended for precise estimation of a demographic model and its underlying parameters but rather to address 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 2.2 & A.1) and more moderately for complex/real scenario as for crow datasets (Table A.5) where the goodness-of-fit is lower (G_{post} 0.9 for simulated datasets,G_{post} 0.25 for crow datasets). However, as expected, the accuracy of parameter estimation largely depends on the divergence time (Figure A.6-A.9). Similar to DILS (Fraisse et al. 2021), the correct model's contribution to parameter estimation and the detection of ongoing migration increases with divergence time (Figure 2.3). Overall, current migration is well captured, both in model weights and in parameter estimation (Figure 2.3, Figure A.7).

This is well illustrated with the analysis of the crow datasets. After the ice cap had retreated in Europe around 10,000 years ago ($\approx$ 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 geographic area away from the central European hybrid zone, RIDGE favored the correct scenario, especially the occurrence of ongoing migration (model weight for SC = 45% and IM=44%) (Table A.6). Similar results were obtained for the RX hybrid zone with IM at 43% and SC at 39%. Overall, in all four datasets 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%).

2.5.2 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 and Ellegren 2017). To avoid the confounding effect of reduced diversity in either of the compared populations due to other causes than barrier to migration (Cruickshank and Hahn 2014; Ravinet et al. 2017), it is now common practice to combine it to absolute measure of divergence (D_{xy}) to other related statistics such as net divergence (D_a) 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}, D_{xy}, D_a, 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 A.11) 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 datasets, as illustrated with the three pairs of crows (Figure 2.9). For the three pairs, F_{ST} and D_a strongly correlated with BF and contributed the most to barrier detection, in agreement with theoretical predictions (Cruickshank and Hahn 2014). Quite unexpectedly, however, D_{xy} did correlate with BF and did not contribute to barrier detection.

A possible explanation is that, at low divergence, variations in D_{xy} mainly reflect local variations in N_e (as confirmed by the strong positive association with π in the PCA, Figure A.15) while the main signal of variation in migration rate is already captured by D_a. Other statistics also correlated with BF but at lower and variable levels in the three datasets, and, similarly outliers correlated differently to the PCA axes (Figure 2.8 and A.15). These difference in genomic signatures may reflect 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 and 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.

2.5.3 Detection of barrier loci using RIDGE

We validated the ability of RIDGE to detect gene flow barriers on empirical datasets 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 and 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 2.8). 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.

2.5.4 Benefits of RIDGE and Guidelines for it uses

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 (Fraisse et al. 2021) which results is a much more stringent detection of barrier loci. 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 windows 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 datasets we explored gave us guidelines for the conditions where RIDGE can provide useful and accurate results. We suggest to use datasets with SNP density higher than 0.1%, such as in crows and simulated datasets, 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 datasets, 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 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 A.12).

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 $2N_e$ generations (10,000 generations in our simulations) appears to be a lower bound for both demography (Figure 2.4 & 2.5) and barrier inferences (Figure 2.6), below which RIDGE fails to capture informative signals. RIDGE can detect gene flow barriers on both simulated (Figure 2.6) and empirical data (Figure 2.7), starting at 0.1 $2N_e$ generation, which represents a very low level of divergence. For context, DILS correctly inferred a gene flow barrier when $T_{split} > 0.5$ $2N_e$ 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 $2N_e$ 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.

2.6 Acknowledgement

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, Clemetine 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 dataset 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”.

Chapter 3

Approach method

This chapter outlines the technical and conceptual enhancements integrated into DILS to transform it into RIDGE. Technical intricacies of RIDGE are then discussed, providing a comprehensive description of each input file within the pipeline, explaining script functions, and examining resulting files. As an illustration, I applied RIDGE to the example dataset provided with the code, explaining the step-by-step procedure.

3.1 From DILS to RIDGE

As mentioned before, RIDGE was developed using DILS (Fraisse et al. 2021) as base code. At the beginning of the project, RIDGE was supposed to be a slightly modified version of DILS devoted to the detection of barriers. Modification after modification, the whole code has been rewritten to incorporate all necessary features and to allow robust and reliable barrier detection. In this part, I present the major changes that I added to RIDGE during the development, the conceptual reasons that motivated these changes and their effects on the results. In the end, I present how RIDGE works.

3.1.1 Reducing the simulation time

Simulation time represents roughly 80% to 90% of the runtime of RIDGE. So even if runtime is not the first priority in my work, having a pipeline that takes a reasonably low amount of time to run allowed me to test a wider range of conditions and features. DILS uses a derived version of the program *ms* (Hudson 2002), called *msnsam* (Ross-Ibarra et al. 2008), which allows the use of a vector of values rather than a single value for a parameter, generating multiple simulations with different parameter values. This property of *msnsam* allows DILS to simulate a dataset with genomic heterogeneity without calling the program for each individual locus in the dataset, reducing the complexity of the code. The runtime performances of *msnsam* are the same as *ms*. One major limitation of *ms* and *msnsam* is that they become very slow when the ratio of ρ/θ becomes high. For context, one single run of RIDGE implies simulating 14 scenarios * 10 000 replicates * 1 000 loci, which makes 141 million independent loci. In the literature, there are two candidate coalescent simulation programs that can replace *ms* using the same syntax and

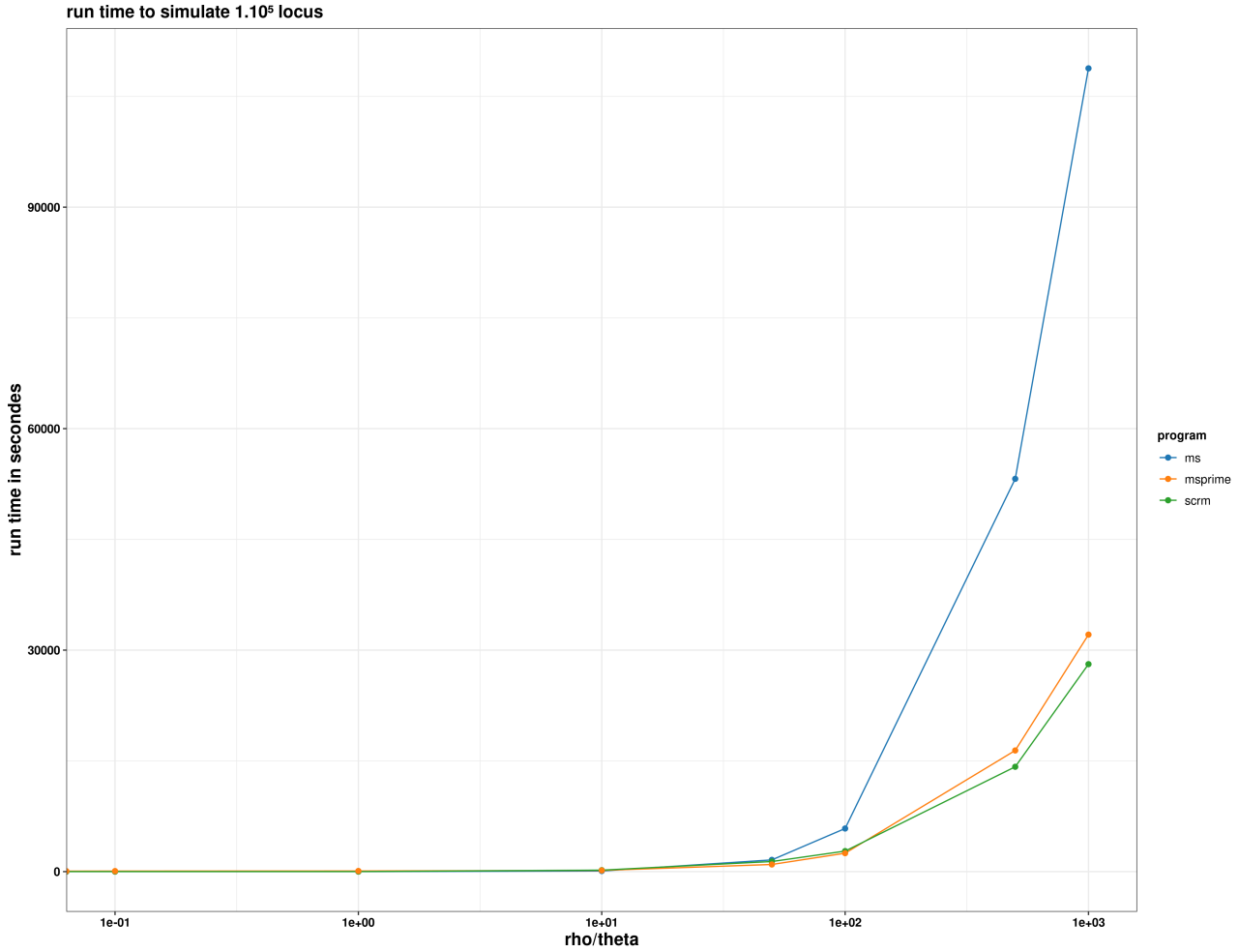

Figure 3.1: Run time in seconds to simulate 1.10^5 locus of 1kb with a fixed value of $\theta = 20$. Only 2 haploid individuals per locus are generated. Benchmark runs were done with a processor at 2.5 GHz. Command used for *scrm* is `scrm 2 100000 -t 20 -r {set value} 1000 -l 100r` ; command used for *ms* is `ms 2 100000 -t 20 -r {set value} 1000` and *msprime* command used is `mspms 2 100000 -t 20 -recombination-rate {set value} -length 1000`, using following recombination rate value: 0, 2, 20, 200, 1000, 2000, 10000, 20000.

provide better performance: *scrm* (Staab et al. 2015) and *msprime* (Baumdicker et al. 2022). *scrm* is a program that aims specifically to simulate faster using approximate coalescence for recombination. *msprime* is an efficient coalescent simulator using tree sequence data structure.

In real data, the ratio ρ/θ can vary a lot and go up to 1000. In such conditions, simulation times of *msnsam* and *ms* become very high. To illustrate this problem, I simulated 2 haploid individuals for 1.10^5 loci of 1kb with a fixed theta value of 20 and varying values of ρ/θ, ranging from 0 to 1000. For simulations, I used *ms* program, *msprime*, and *scrm* with the recombination approximation parameter set to "100r" (this is the recommended value by the authors to gain in runtime without generating too much deviation from *ms*). The results show that at low values of ρ/θ ($<$50), the difference between among programs is negligible, but when ρ/θ increases, *scrm* and *msprime* are 2 to 4 times faster than *ms* (Figure 3.1). For example, assuming a value of $\rho/\theta = 500$, to run a single simulation, *scrm* takes 0.14 s and *ms* takes 0.532 s for loci of 1kb size. Under the same condition, if running RIDGE on 70 cores, it would take about 3 days with *scrm* or *msprime*, whereas with *ms* it would take about 12 days.

The challenge was to implement genomic heterogeneity using *scrm* or *msprime*. Because both programs do not offer the possibility to input a vector of n locus at once, we have to call the program for each locus. But there is a crucial difference between *scrm* and *msprime*. *msprime* is encoded in Python whereas *scrm* is in C++. This difference is noticeable when we call the program multiple times. In the previous example, I only called the program once and asked it to generate 1.10^5 locus under the same conditions. To illustrate the difference, I called *scrm* and *msprime* 100 times to generate the sequence of one locus for two haploid individuals each time with a θ of 20. The run time for 100 loci of *msprime* is 23,846 s, whereas it's 0.189 s for *scrm*. In conclusion, for the case of RIDGE, I chose to use *scrm* with the recombination approximation parameter set to "100r".

3.1.2 The log uniform distribution of migration parameter in priors

During different tests, it appeared that our results delivered poorl estimates "low" values of migration ($M < 1$) (Figure 3.2 method uniform), an observation also made with DILS (see Figure 5E from Fraisse et al. (2021)). Because we set a uniform prior covering three orders of magnitude for $M \in [0.1; 50]$, random draws values predominantly cover values between $M = 10$ and $M = 50$. Hence, the random forest algorithm was primarly trained under conditions where $M < 10$ was rare. The effect of migration on F_{ST} (Fig 3.3) is nonlinear, which argues for the use of a higher prior density toward small rather than elevated values of M. To address this problem, a log-uniform distribution is used (see Fig 3.2) in RIDGE.

3.1.3 Model averaging & joint parameter estimates powered by random forest

To perform parameter estimates, DILS chooses the model that best fits the data and then estimates the parameters of the best model. The detection of gene flow barriers is contingent upon

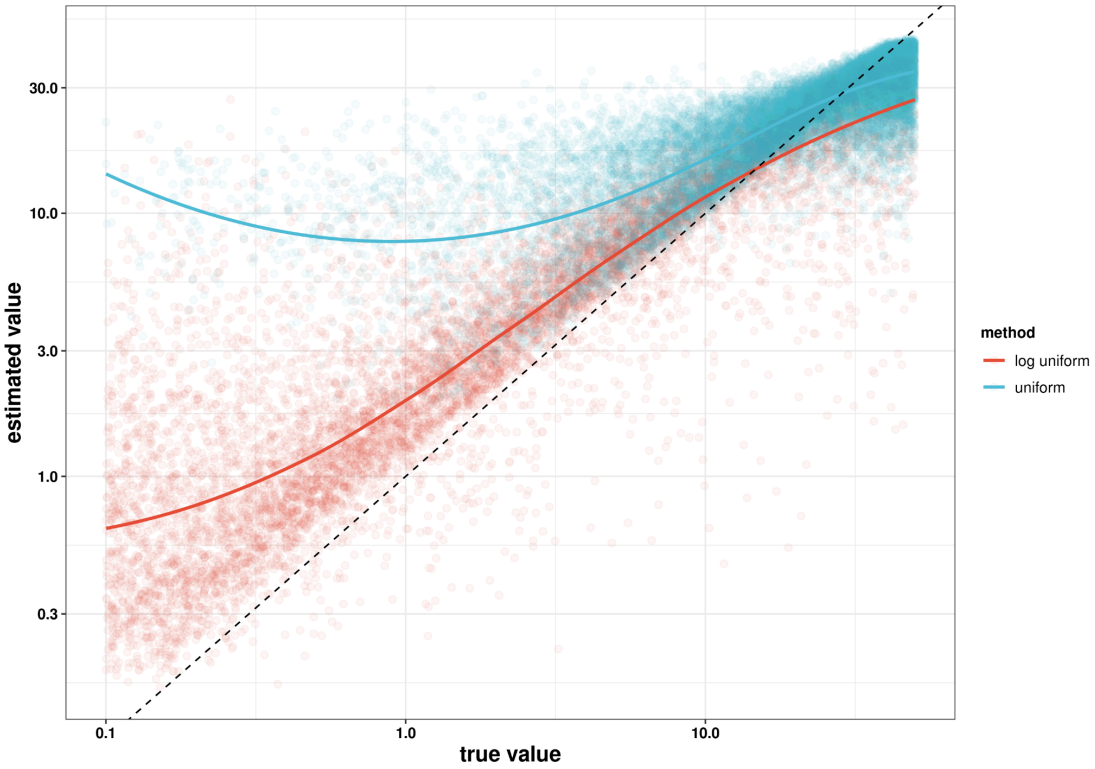

Figure 3.2: Performance of migrant rate M ($M = 4 * N_e * m$) parameter estimate, using a uniform or a log uniform distribution to generate priors. Plain lines represent the loess (locally estimated scatterplot smoothing). The dashed line represents $x = y$. Under each condition 10 000 pseudo-observed dataset under IM_2N_2m model are simulated using random parameter values distributed between bounds $M \in [0.1; 50]$. Parameter estimation was done with the "predict" function on *regAbcrf* (from *abcrf* R packages) using 1000 trees. X and Y axes are in log scale.

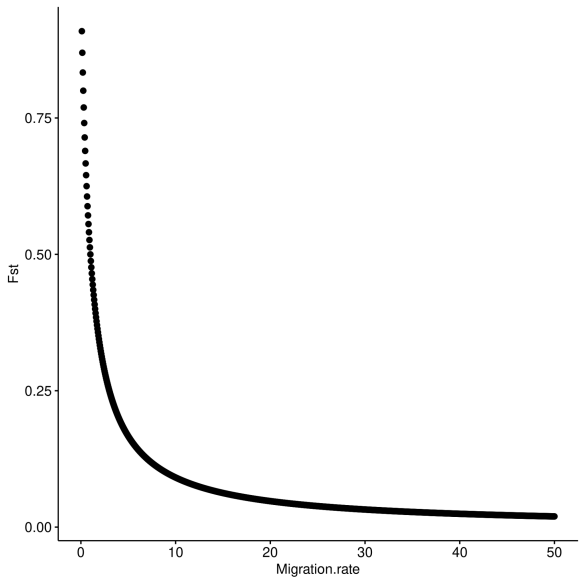

Figure 3.3: Evolution of F_{ST} with the migration rate, assuming that $E[F_{ST}] = 1/(1 + M)$

this step, as the search for barriers occurs only when the best-fitting model includes migration heterogeneity. This is a major limitation for comparative studies. RIDGE instead uses model

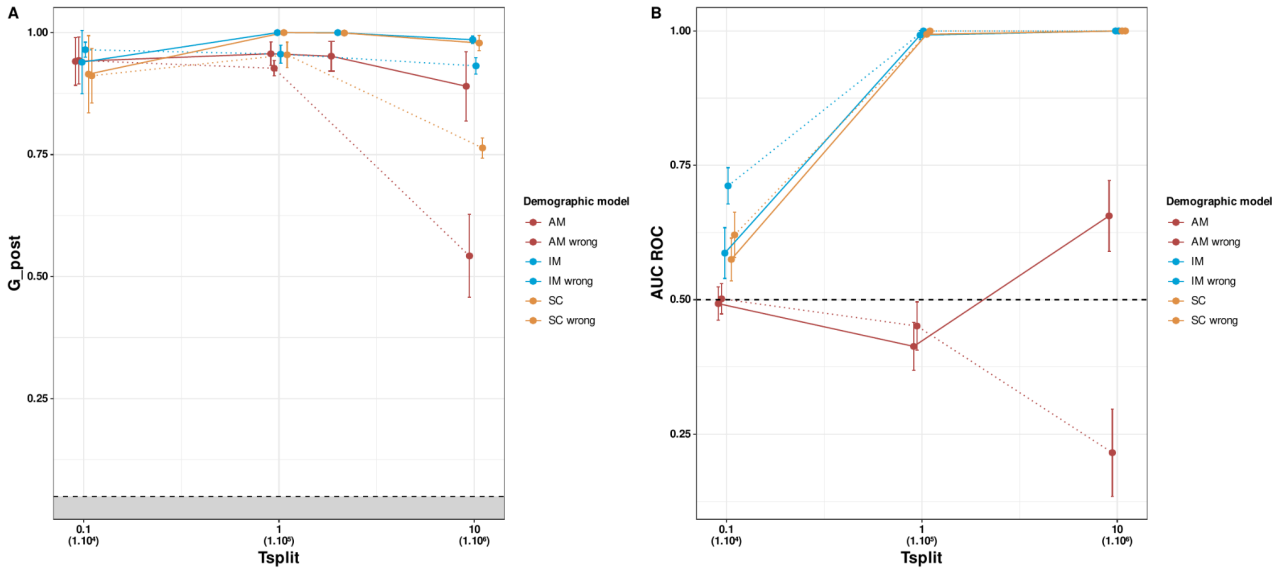

Figure 3.4: Goodness of fit of posterior (A) and AUC of ROC curve of barrier detection (B), generated using model averaging (plain lines) or generated forcing RIDGE to allocate a model weight of 100% to IM_2N_2m model (dotted lines, called "wrong" in legend) under each T_{split} condition. Dataset was simulated with $M = 10$ and $Q = 0.1$.

averaging to avoid the best-model choice step. DILS estimatse demographic parameters separately using a machine learning algorithm (*regAbcrf* from *abcrf* R package ; Pudlo et al. (2016), and *nnet* R package ; Venables and Ripley (2002)), but parameters are highly correlated and accounting for these correlations may greatly improve parameter estimations. To solve this problem, RIDGE estimates jointly all parameters. These two improvements are detailed below.

Model averaging

Classic model averaging methods relie on models' weight. To obtain model weight, our first idea was to use the proportion of trees votes for each model as a proxy to model weight. But using tree votes, mean implicitly that each tree has the same weight, which is wrong. Indeed, in a random forest, some trees are less informative than others, as they are more or less able to classify data. Giving the same weight to each tree adds noise to the signal and reduces discriminant power. Instead of inferring model weight, RIDGE uses all simulated datasets to build a hypermodel that accounts for all possible scenarios, allowing the RF to directly estimate hypermodel parameters, and implicitly weighting models. The model weight can be quantified as the proportion of simulations from a specific model within the posteriors. The significance of model averaging was demonstrated by comparing the results obtained through averaging with those obtained by forcing RIDGE to allocate 100% of model weight into a single model (IM_2N_2m). For small T_{split} values, model averaging is relevant because it is difficult to choose the correct model (see Figure A.2). For large T_{split} values, model averaging prevents a reduction in goodness of fit and barrier detection power due to errors in model selection. The goodness of fit of posterior was reduced, and the barrier detection ability for the dataset simulated under SC and AM demographic models at high split time ($T_{split} > 1T/2N_e$) was also

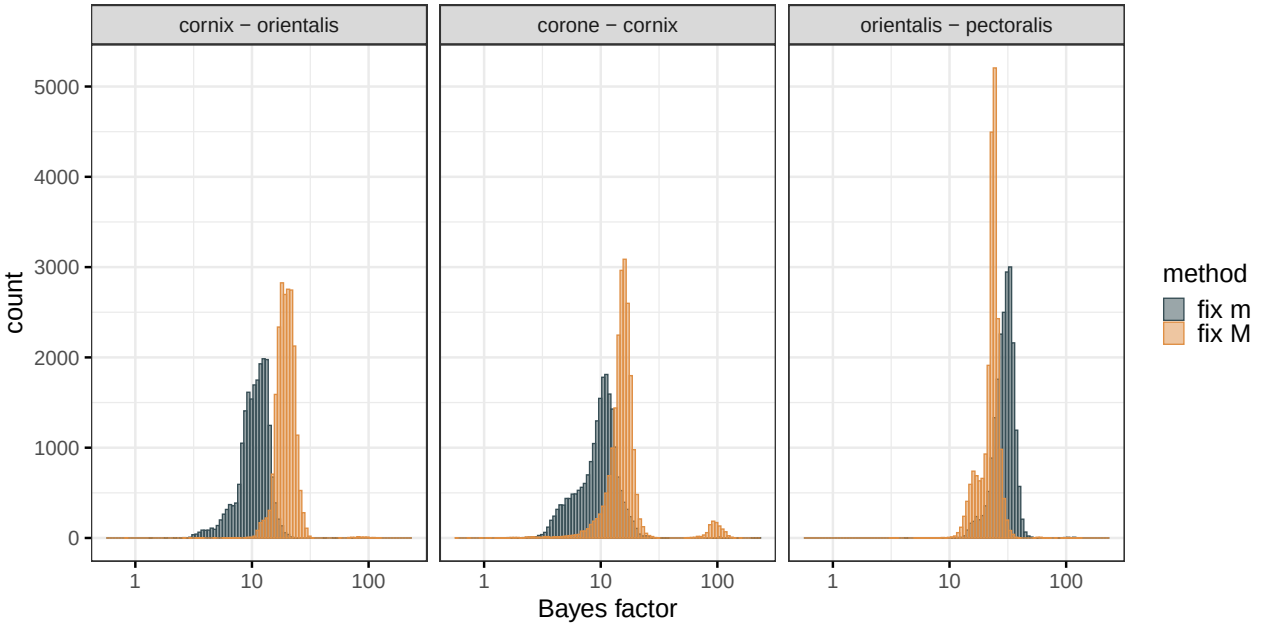

Figure 3.5: Distribution of Bayes factors simulating migration using M (DILS like method) or m.

reduced (Figure 3.4).

Joint parameter estimates

Parameters in demographic models are extremely correlated. For example, to estimate T_{split} from divergence D, where $D = 2 * T_{split} + 4N_e\mu$, one must take the population size N_e into account (assuming a constant mutation rate, μ). Ideally, one would like to infer the correlation between parameters in the observed dataset and then, from parameter specific posterior distribution, to create a joint distribution in which one could sample posterior parameter values. In theory the *CovRegAbcrf* (Raynal et al. 2019) should allow us to estimate covariance between parameters in a dataset, but our test was not satisfying. So RIDGE uses a different method based on the weights attributed by the random forest. When the *regAbcrf* (Raynal et al. 2019) function performs a prediction, it attributes a weight to each simulated dataset of the reference table, and the prediction of a given parameter is the weighted mean of the reference table parameter values. The higher the weight value, the greater the contribution of the simulated parameter values to the parameter estimation RIDGE computes the average weight across all parameters - joint weights - for each dataset from the reference table, and uses it to sub-sample parameters sets in the reference table to generate posterior parameters. RIDGE therefore implicitly accounts for the non-independence of parameters.

3.1.4 Migration rate versus effective number of migrants

There are two common ways to use migration rate: (i) the number of effective migrants at each generation, denoted as M, where $M = 4 * N_e * m$, and (ii) the migration rate, m. DILS (Fraisse et al. 2021) simulates migration by modulating $M = 4 * N_e * m$ rather than m, so

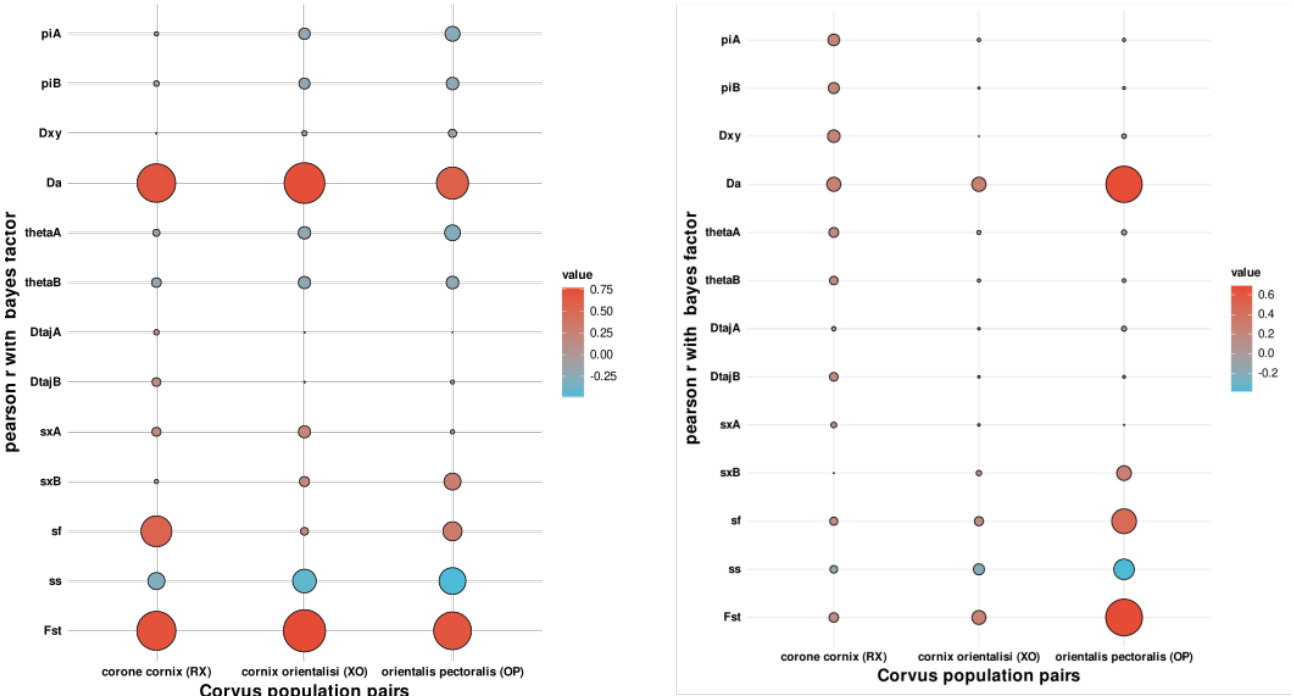

Figure 3.6: Comparison of correlations of BF with the summary statistics of crow dataset presented in Vijay et al. (2016) simulating migration using m (left) or M (DILS-like method, right). Circle size represents the absolute value of correlation and coloration the value of correlation.

that m is automatically adjusted along the genome to achieve a given M value according to N_e. Hence, in the case of hetero-N_e and homo-m, variation in N_e along the genome therefore automatically translates into variation in M and we expect hetero-N_e hetero-M models to be often wrongly infered. This causes a bias toward barrier detection. In RIDGE, we decided to simulate migration through m, which appears more biologically realistic. M can thus vary along the genome because of variations in N_e, in m, or both. As we expected, this change reduced the number of barrier loci detected in some cases. For example, taking the '*corone - cornix*' pair (RX) examined in the crow dataset (see 2.4), the number of loci exhibiting a BF > 50 was reduced from 958 to 26 loci (see Figure 3.5). Examination of the correlation between summary statistics and Bayes factors, using either m or M to model migration, reveal results more consistent with our expectations with the former. In particular, the summary statistics that are the most correlated with Bayes factors are those linked to divergence, such as D_{xy}, Da, ss, and sf. In contrast, the use of M displays a pattern with contrasted correlations not easily interpretable and, in some cases, with no summary statistics involved in barrier detection (see Figure 3.6). Further data analyzes reveal that the loci identified as barriers using M (referred as "M") in the "corone - cornix" dataset exhibit a pattern similar to the rest of the genome, unlike loci that are consistently identified as barriers using both methods (referred as "m", see Figure 3.7). This means that the "M" loci are likely false positives resulting from the M-based approach. Therefore, we chose to implement the simulations of migration using m rather than M.

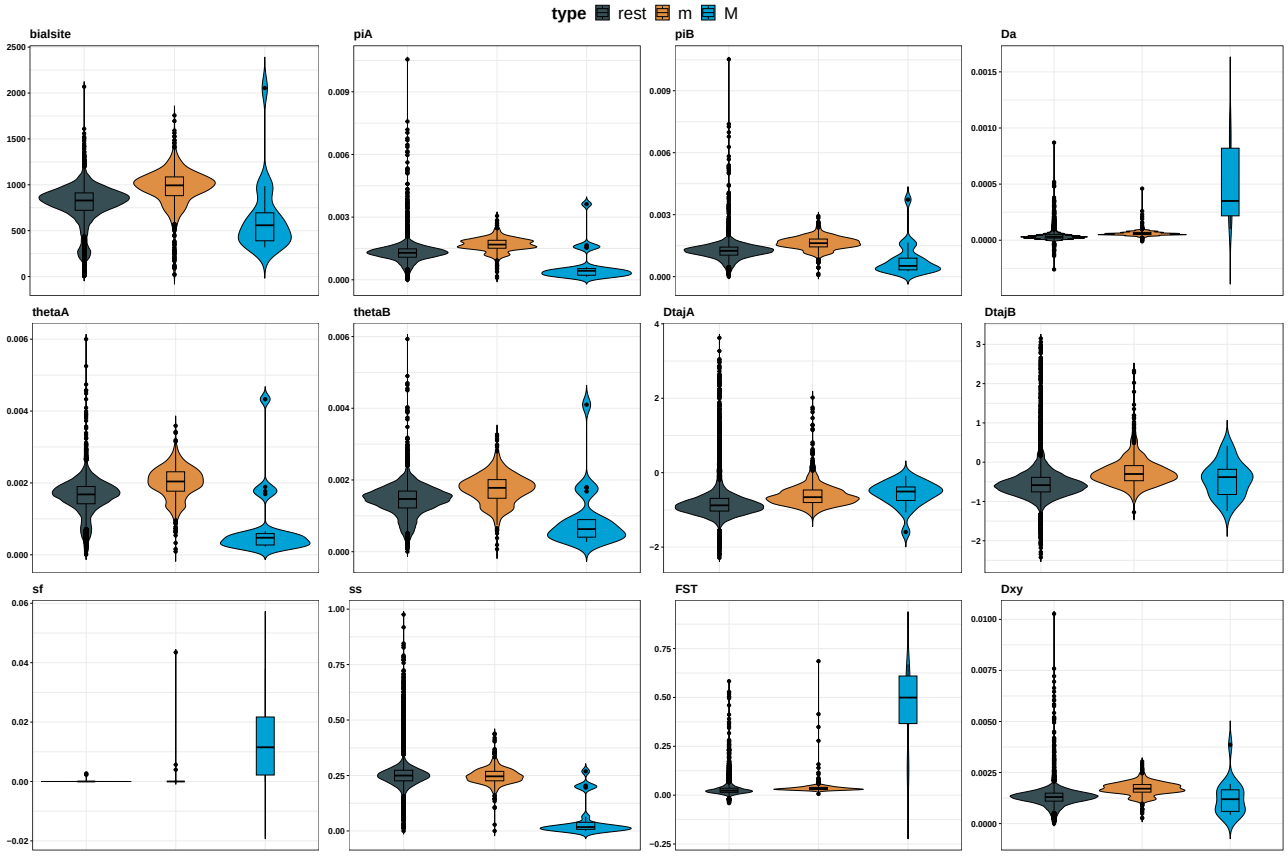

Figure 3.7: summary statistics distribution for loci from "*corone* - *cornix*" (RX) dataset (Vijay et al. 2016), constantly detected as barrier with both methods (BF>50) called "*m*", loci that were previously detected as barrier with the *M*-based approach but not detected as barrier with the *m*-based approach referred to as "*M*" and the "rest" of the genome.

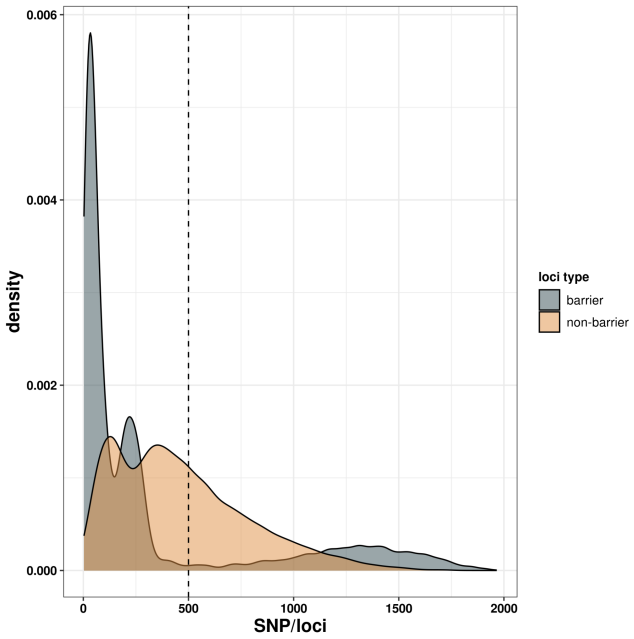

Figure 3.8: Distribution of the number of per-window SNPs for barrier and non-barrier loci in the maize dataset detected using the "hetero θ" option. The dashed line distinguishes low and high SNP density barrier loci

3.1.5 Taking heterogeneity of data quality or mutation rate into account

DILS assumes a uniform mutation rate and recombination rate across the entire genome. Initially, RIDGE only took into account the heterogeneity of the recombination landscape. However, in certain analyses, such as the one presented in the next chapter using a maize dataset (refer to chapter 4 for more detailed results), it was observed that 24% of the genome was categorized as barriers. Upon closer examination of the results, it became apparent that among the loci classified as barriers, the majority were devoid of polymorphisms. This could be attributed either as a mutation cold spot or inadequate coverage of the region linked to mapping issues. Since the RF is trained based on anticipated diversity derived from simulated loci, which inadequately represents regions with a limited number of polymorphisms, it struggles to handle this data scarcity effectively. Consequently, the existence of only a few polymorphisms, some being unique to a particular population, is perceived as a barrier, when, in fact, these regions should be recognized as missing data. Rather than opting for the complex task of identifying and excluding such regions from analyses, we chose to enhance the training of the RF by relying on the distribution of θ_W which is directly linked to S, the number of polymorphic sites. This "hetero θ" option notably improved the quality of barrier detection. It should be particularly relevant for low-depth data and/or complex genomes where mapping can be challenging and/or genomes of poor-quality assembly. In the case of *corone* species, we found no clear impact of this option because the genome was assembled in scaffolds with an average coverage of 12.5X and we chose to retain only the scaffold exceeding 50kb.

In the maize example (described in chapter 4), in contrast, there was a marked difference

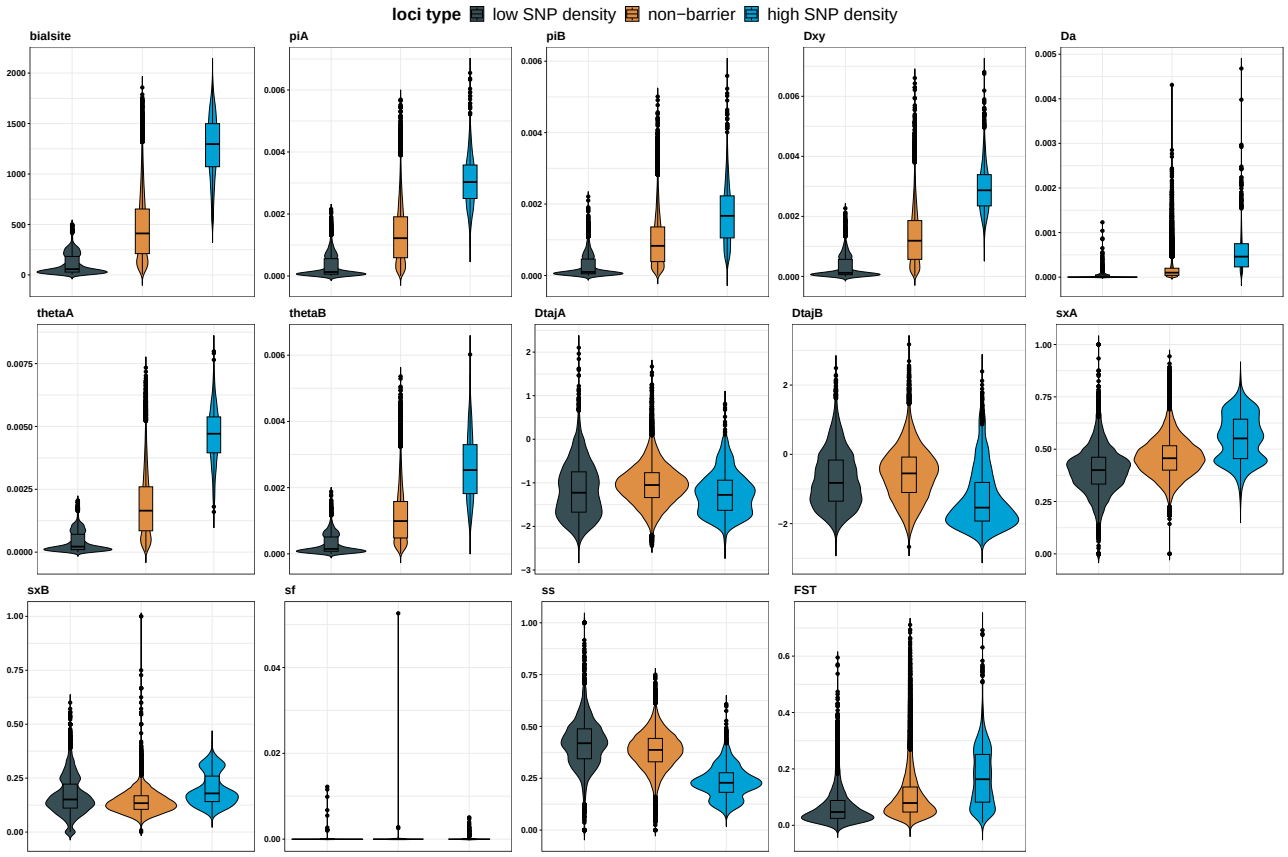

Figure 3.9: The distribution of summary statistics in maize dataset, based on three groups: barrier loci of low-density SNPs ($\leq$ 500), high-density SNPs ($>$ 500) and non-barrier loci. Results were obtained without the "hetero θ" option.

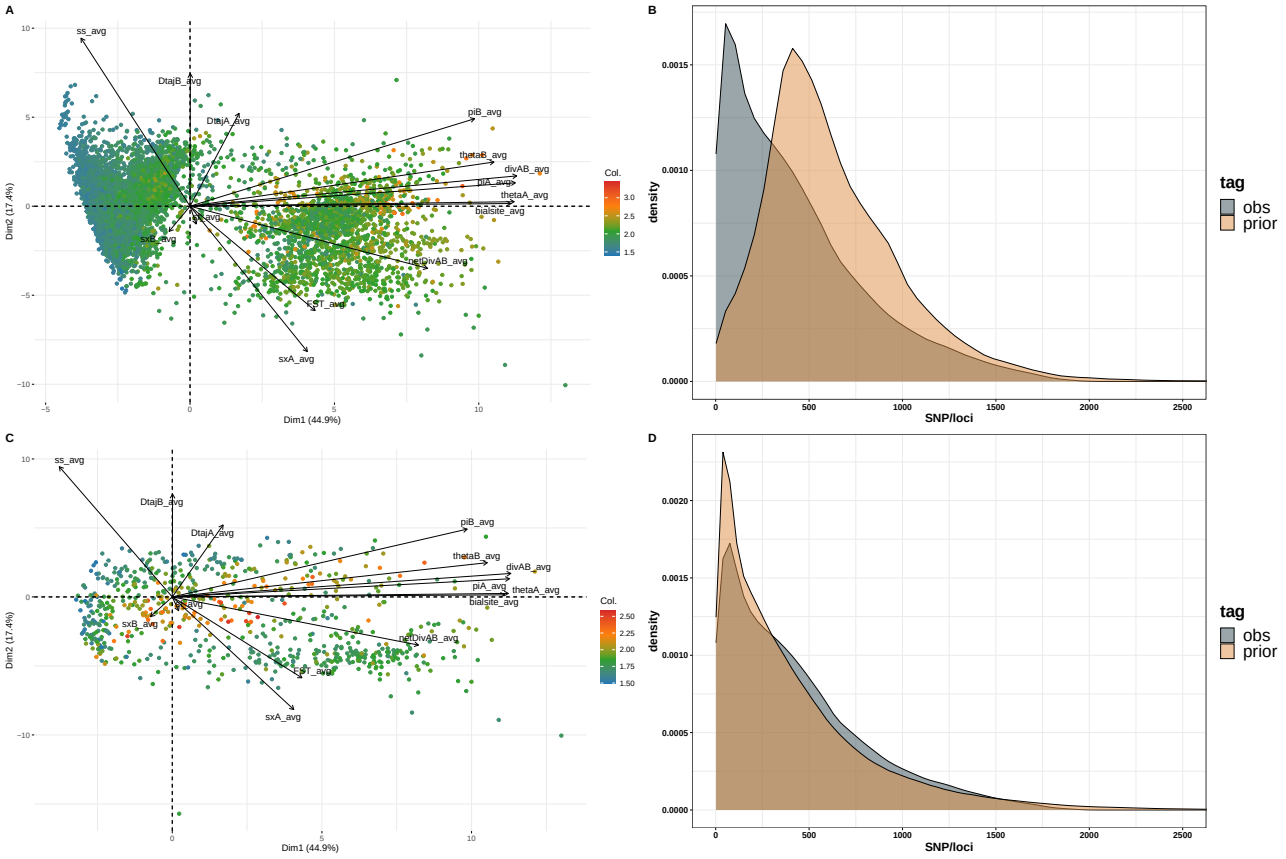

Figure 3.10: PCA of all loci considered as barrier in maize dataset without (A) and with (C) "hetero θ" option in simulations. The color scale relates to log of Bayes factor. Distribution of the density of per-window SNPs between observed dataset (maize) and simulated dataset without (B) and with (D) "hetero θ" option in simulations.

when considering expected diversity and observed distribution of θ_W (referred to as "hetero θ"). Without relying on *heteroθ* to train the RF, 24% of the genome was deemed a barrier (where a locus was considered as barrier with a post.prob>0.5). Using PCA, the loci identified as barriers were visually divided into two groups with a threshold of 500 SNPs per window (see Figure 3.10A & 3.8): one group encompassing regions with low per-window SNP density with the majority peaking below 100 SNPs and one above 500 SNPs (see Figure 3.8). The former exhibited patterns of depleted levels of diversity for both populations A (teosinte) and B (maize), and no sign of divergence and differentiation between populations, signals yet expected for barrier loci which markedly contrast with high-density SNP windows (Figure 3.9). Note that the levels of diversity, differentiation and divergence of this first group fall below the level observed for non-barrier loci (Figure 3.9). In contrast, the group containing SNP density above 500 displayed levels of diversity, differentiation and divergence significantly higher than the non-barrier loci. With the inclusion of "hetero θ", we observe a drastic decrease of the number of barrier loci representing 1.7% of the genome (instead of 24%). Noteworthy, barrier loci exhibited patterns of summary statistics consistent with "high-density loci" in Figure 3.9. This pattern included a global increase in divergence and differentiation albeit a diversity of profiles (Figure 3.10C).

3.2 User manual

This manual is intended to guide users through the installation and use of RIDGE by describing each script function and the resulting files in detail. The following notations help clarifying the nature of the different elements in the pipeline in this section :

- `Gray highlighted` text stands for input parameters of RIDGE that are used in config.yaml file (see Input files section).
- ***Italic bold*** text stands for files or programs needed in RIDGE.
- ***`Italic bold highlighted`*** text stands for RIDGE scripts files.

3.2.1 RIDGE v1

RIDGE takes as input a vcf file containing sequences of individuals from two populations, accompanied by accessory files providing complementary information. From this, RIDGE first uses Approximate Bayesian Computation (ABC) to infer demographic data by simulating 14 demographic x genomic models to produce a reference table. This table serves to train a random forest (RF) that generates weights and parameter estimates for each model according to their fit to the target (observed) dataset. Second, RIDGE constructs a hypermodel where the posterior distribution of each parameter is obtained as the weighted average over the 14 models. Finally, it uses this hypermodel to simulate one set of control loci (thereafter non-barrier) and one set of barrier loci that have undergone no gene flow during divergence. Simulated datasets

generated for barrier and non-barrier loci are used to train a second RF that generates posterior probabilities and associated Bayes factors for each locus to belong to the barrier or non-barrier category.

3.2.2 Installation

Requirements

RIDGE uses the Snakemake workflow management system as well as Singularity containers. First ensure that Singularity and Snakemake (v 7.7.0) are installed on the machine where RIDGE will run. If it's not the case, contact the admin system in case of a cluster installation; otherwise, follow the installation instructions: `https://snakemake.readthedocs.io/en/stable/getting_started/installation.html` (pip installation is recommended)

Get the code

Download the code (v1.0, which is the version use in Burban et al. (2023)) with the following command

```
git clone -b v1.0 https://github.com/EwenBurban/RIDGE.git
cd RIDGE
```

Install containers

After completing this step, you will be provided with a list of .sif files in the container folder, including python.sif, R.sif, R_visual.sif and scrm_py.sif.

cluster installation To install RIDGE on a cluster, create a free account on https://sylabs.io and then go to https://cloud.sylabs.io/tokens to create an authentication token. Afterwards, input the command below and paste your token.

```
singularity remote login
```

Go into the RIDGE folder and launch container creation using the following command.

```
cd <path to RIDGE> bash cluster_configure.sh
```

This process installs all required programs and their dependencies within the container folder.

local installation If you install RIDGE on a local machine, simply execute the following command:

```
bash configure.sh
```

Set-up config folder

The configuration folder must contain at least one file (***config.sh***) that can tailor the behavior of RIDGE to fit your specific installation. As it is not included in the git clone, you will need to create the folder first. The content of this folder is called as the launch of RIDGE by ***RIDGE.sh***.

```
cd <path to RIDGE>
mkdir config
cd config
touch launch_param.sh
```

launch_param.sh This file regulates the number of jobs that snakemake will attempt to execute simultaneously and monitors snakemake. So open and then edit it using the following instruction and example:

Example of file :

```
module load snakemake singularity
mode='cluster'
ntask_load=140
ID=ridge_project
```

- The beginning of the file involves calling Snakemake and Singularity, which is essential in case these programs are not available by default in your working environment – often the case while using clusters. The command to call them is highly dependent on your installation, so do not take into account the command used in the example file. If you place Snakemake and Singularity within a Conda environment, this is the location to invoke the Conda environment 'conda activate <name of your env>.
- **mode** option defines the behavior of Snakemake. If the mode is set to 'cluster', then RIDGE will initiate jobs using SLURM. Afterwards, the ***cluster.json*** file must be completed (further information regarding this file is provided below). For 'local' mode, Snakemake will start jobs automatically. This mode is recommended for local installations or clusters that do not employ SLURM job managers.
- **ntask_load** option allows you to define the number of jobs that Snakemake will try to launch.
- **ID** option is specific to mode='cluster', as it defines the name of the user who starts the job. This only applies to the SLURM structure, and is commonly your user name. If mode is set to 'local', this line can be removed.

cluster.json This document outlines the resources available to each job and enables you to fine-tune and optimize RIDGE for your specific cluster. Further information regarding cluster configuration can be found in the cluster configuration section of `https://snakemake.readthedocs.io/en/stable/snakefiles/configuration.html`. An example is also available at template/cluster.json. Input files To launch RIDGE, you need to provide at least four files in your work folder (i.e the folder where RIDGE will work and generate output). So, your work folder must follow the following configuration before any launch:

work_dir/
├─ vcf_file
├─ contig_data.txt
├─ popfile.csv
├─ config.yaml
├─ rec_rate_map (optional

3.2.3 Input file

Vcf file

In the actual version, RIDGE only takes as genomic polymorphism data a vcf file (vcf file format >= 4.0). RIDGE can manage haploid and diploid data. The vcf file must contain only biallelic sites. See `https://en.wikipedia.org/wiki/Variant_Call_Format` for detailed information on the file format

Config.yaml file

The config file contains all the data to start RIDGE. Note, that in this file, the priors used in the ABC process are defined, and so, an incorrect specification of the hyperpriors can drastically affect the performances and results of RIDGE The fields of the config file are the followings:

- `config_yaml`: the name of the ***config.yaml*** file that you are actually filling. (Note that you do not need to give the absolute path, but only the filename, otherwise it will stop)
- `vcf_file`: the name of vcf file (only filename expected)
- `contig_data`: the name of the contig data file (only filename expected)
- `rec_rate_map`: the name of the recombination map (only filename expected or NA if no map)
- `popfile`: the name of the popfile (only filename expected)
- `nameA` and `nameB`: name of one of the two populations. The names must be the same as used in popfile

- `container_path`: the absolute path to the container folder, which contain all the Singularity container, and so all programs
- `ploidy`: the level of ploidy of the dataset. 1 is for haploid, 2 is for diploid
- `ligthMode`: Activate the ligthMode, which is a fastest but less precise version of RIDGE
- `work_dir`: absolute path to the work folder
- `nLoci`: number of loci sampled, used to avoid unnecessary computational time. If nLoci is set to -1, all the genome will be used in the process (it may slow down the process by 10 to 100 times, depending on the size of the dataset). Note that a total number of loci around 1000 loci is a good trade-off between genome representation and computation time limitations
- `window_size`: size of each locus in bp Choose the value according to the SNP density in your data.
- `homo_rec`: If True, recombination rate is considered homogeneous along the genome, if False,it uses the recombination map provided with `rec_rate_map`
- `homo_rec_rate`: recombination rate value along the genome (used only if `homo_rec`=True)
- `mu`: mutation rate (assumed constant along the genome) per pb. Note that mu is needed for prior bound suggestion even if hetero_theta=True.
- `Nref`: Population size of reference (in number of individuals) used to rescale all values in coalescent units.
- `N_min` & `N_max`: minimum and maximum population size (in number of individuals). It is highly recommended to set the value on the basis of the real diversity value in the vcf rather than expected value from the literature (to have a good estimation you can launch RIDGE in scan mode and follow suggestion from ***prior_bound_suggestion.txt***).
- `M_min` & `M_max`: minimum and maximum migration rate (in $4 * Nref * m$ unit). By default, `M_min`=0.1 and `M_max`=50
- `Tsplit_min` & `Tsplit_max`: minimum and maximum time of split of the ancestral population in generations !!! Note that it is highly recommended to set the value on the basis of the real data in the vcf rather than expected value from the literature (to have a good estimation you can launch RIDGE in scan mode and follow suggestions from ***prior_bound_suggestion.txt***).
- `Pbarrier_max`: maximum proportion of the genome under barrier to gene flow. By default, `Pbarrier_max`=0.2 (i.e. 20% of loci are considered as barriers).

- `hetero_theta`: Activate/Deactivate (True/False) hetero θ option. If True, RIDGE ignores `mu` to set an expected level of diversity and rather uses θ computed on the observed number of SNPs using Watterson θ estimator: $\theta_W = S/a$.

Example of file:

```
M_max: 50
M_min: 0.1
N_max: 200000
N_min: 10000
Nref: 50000
Tsplit_max: 20000
Tsplit_min: 1000
Pbarrier_max: 0.2
config_yaml: config.yaml
container_path: /home/RIDGE/container
contig_data: contig_data.txt
lightMode: False
mu: 1e-8
nameA: wild
nameB: dom
popfile: popfile.csv
rec_rate_map: rho_map.txt
work_dir: /home/example_ridge
window_size: 50000
ploidy: 1
vcf_file: vcf_file.vcf
homo_rec: False
homo_rec_rate: NA
hetero_theta: True
nLoci: 1000
```

Popfile

This file lists the individuals from each population in csv format (with ',' as separator). The name of each population must be in the header. The popfile must contain at least two populations (so two columns) and each list must be of the same length as the others. If the populations are not of equal length, you can fill the missing individuals with NA.

Example of file:

```
wild,dom
```

```
W1,Dx1
W3,Dx10
W5,NA
W7,Dx11
```

Contig data file

A file containing the length of each chromosome/contig and their associated names and order.

- **contig_name**: is the name of the chromosome/contig in the vcf file
- **contig_length**: the length in bp of the chromosome/contig
- **index**: the index in the order of contigs

Example of file:

```
contig_name contig_length index
Chr1 43270923 1
Chr2 35937250 2
Chr3 36413819 3
Chr4 35502694 4
Chr5 29958434 5
Chr6 31248787 6
Chr7 29697621 7
Chr8 28443022 8
Chr9 23012720 9
```

Recombination rate data

RIDGE uses either a recombination map or a constant recombination rate to work. If you choose to use a constant recombination rate, you must set `homo_rec` to True in the ***config.yaml*** file and fill the field `homo_rec_rate` with the mean recombination rate estimated for your dataset. Otherwise, you need to fill the field `rec_rate_map` with the name of your recombination map and place it in the work folder. Note that the recombination rate r must be the number of recombination rate per bases and per generation and the recombination map uses the tabulation as a separator.

- **chr**: the index of the contig (see contig file)
- **start** and **stop**: the beginning and ending in bp of the window
- **r**: the recombination rate inside the window in bp

Example of file :

```
chr start end r
9 21800000.0 21900000.0 7.170443918444081e-07
9 21900000.0 22000000.0 6.771961140602021e-07
9 22000000.0 22100000.0 6.44356192372138e-07
9 22100000.0 22200000.0 6.08745943319314e-07
9 22200000.0 22300000.0 5.709059200375581e-07
```

3.2.4 Usage

Gather all necessary information

Before any launch you must fill and provide all mandatory input files. See section 3.2.3 for details.

Scan launch

To correctly fill your config files you will need a measure of the diversity and divergence. Under "scan" mode, you are not obliged to set up values for the prior bounds (`N_min, N_max, N_ref, Tsplit_min, Tsplit_max, M_min, M_max, Pbarrier_max`), but for "all" mode – which run RIDGE analysis – they are mandatory. You can choose to use your own prior bouds or to use on estimation done by RIDGE. If you choose to use your own prior, see section 3.2.6 to validate the quality of you prior, otherwise read the following part to generate prior suggestion. For `M_min ; M_max` and `Pbarrier` default values are suggested (`M_min`=0.1, `M_max`=50 and `Pbarrier_max`=0.2). Next you have to launch RIDGE in scan mode with the following command

```
bash <path to RIDGE>/RIDGE.sh <work_dir>/config.yaml scan
```

RIDGE will generate a file called ***prior_bound_suggestion.txt***. It is highly recommended to define prior bounds based on this file, as the values are calculated from the raw data. The procedure for validating prior bounds is explained in section 3.2.6.

Complete launch

Once `N_min, N_max, N_ref, Tsplit_min, Tsplit_max, M_min, M_max, Pbarrier_max` are correctly set, relaunch RIDGE, but this time with the whole process, without forgetting to delete ***modelComp/*** folder:

```
rm -r <work_dir>/modelComp <work_dir>/gof_prior.txt <work_dir>/QC_plot
bash <path to RIDGE>/RIDGE.sh <work_dir>/config.yaml all
```

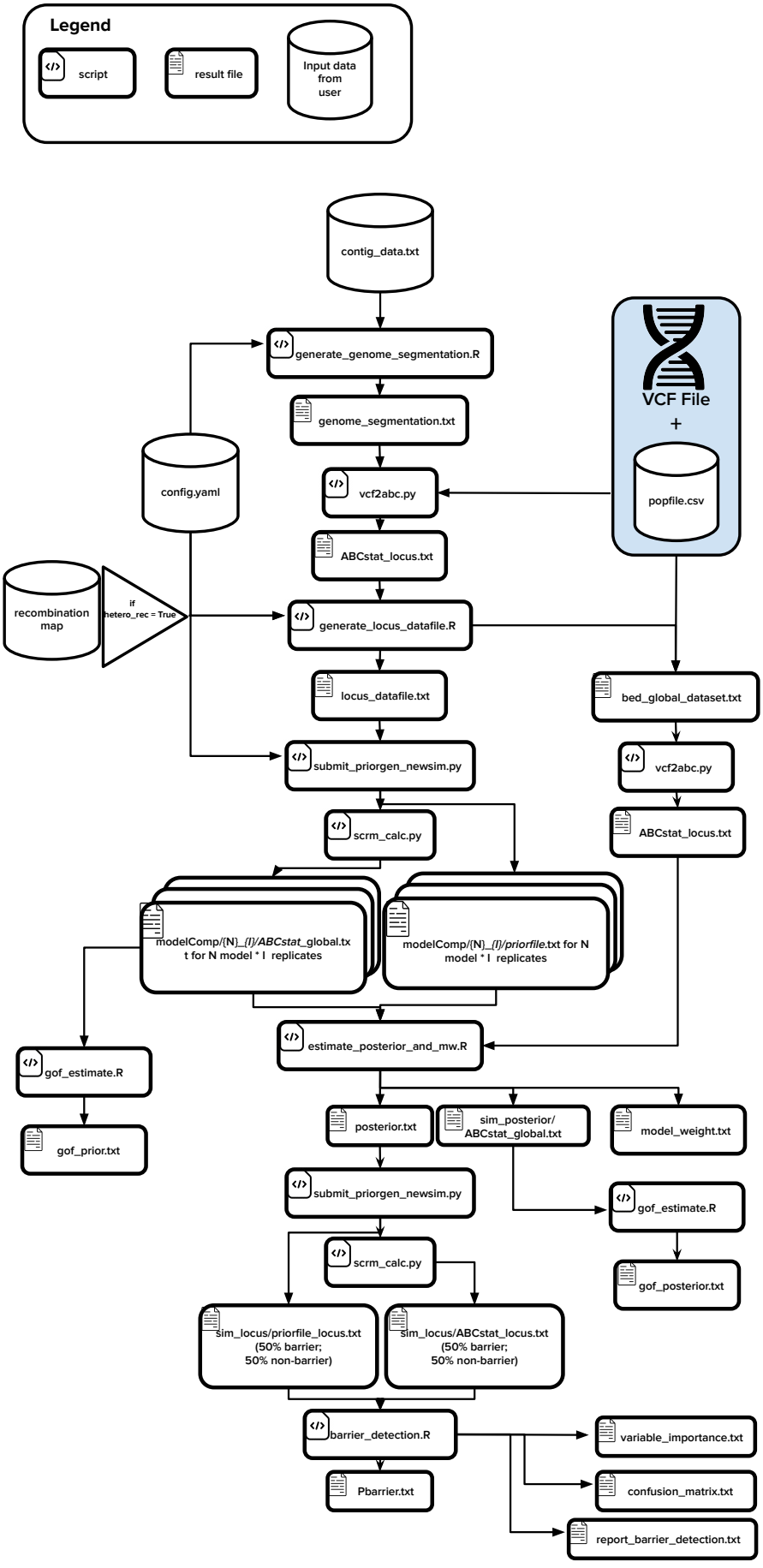

Figure 3.11: Graphical representation of RIDGE pipeline

3.2.5 RIDGE Pipeline

RIDGE relies on ABC for detecting gene flow barriers by incorporating the effect of demography. Firstly, RIDGE infers the parameters of a hyper demographic x genomic model; secondly, it infers the probability of each locus to be a barrier. Each step of the pipeline is detailed below in the order of execution and represented in Fig 3.11.

Genome splitting and scanning

Initially, RIDGE takes the chromosome size information from the ***contig_data.txt*** file and combines it with the `window_size` parameter provided in ***config.yaml***. The script ***generate_genome_segmentation.R*** then divides the genome into non-overlapping windows of `window_size` base pairs. The genome segmentation is stored in the file ***genome_segmentation.txt***. This file is then used by ***vcf2abc.py*** to determine the boundaries of each locus, accompanied by ***popfile.csv*** which indicates the sample composition of each population, to calculate summary statistics on the vcf file. The computation method used for summary statistics is given in chapter 2. The summary statistics for each locus are stored in the file ***ABCstat_locus.txt***.

Locus data gathering, subsampling and average summary statistics

The script ***generate_locus_datafile.R*** creates the ***locus_datafile*** file containing the data needed to simulate loci in further steps. The data needed to simulate a locus are:

- The number of haploid samples from each population, for which we multiply the number of samples from the population by the ploidy level. To get this information, ***generate_locus_datafile.R*** uses ***popfile.csv*** combined with population names (`nameA` and `nameB`) and `ploidy` level from the ***config.yaml*** file.
- The total number of haploid samples, which is the sum of both population sizes multiplied by their `ploidy` level.
- The recombination rate within the locus is calculated as follows: $\rho = 4 * Nref * r * L$, where `Nref` is the reference population size, L is the `window_size` defined in ***config.yaml***, and r is the expected number of recombinations on the locus per generation and per individual. There are two ways to provide r to RIDGE. The first is to use a fixed value across the genome by setting `homo_rec` to 'True' and specifying the value of r through the `homo_rec_rate` option. The second is to provide a recombination map by setting `homo_rec` to 'False' and declaring the path file to the recombination map in `rec_rate_map`.
- θ represents the expected diversity. RIDGE offers two ways to compute it: 1) Using a fixed mutation rate across the genome (μ), by setting `hetero_theta` to 'False' and specifying the `mu` value in the ***config.yaml*** file and computing $\theta = 4 * Nref * \mu * L$. 2)

Obtaining it from the observed amount of SNPs using $\theta_W = S/a$ (see 1.2.1 for details on measuring genomic diversity). This option allows for the consideration of heterogeneity in mutation rates across the genome and/or variations in data coverage (refer to section 3.1.5 for the rationale behind this option).

For each locus, we obtain the necessary information for simulations. Then, we randomly select `nLoci` loci across the genome without replacement. This reduces computation time and avoids unnecessary simulations for demographic model inference, which accounts for 80% of the RIDGE running time. For optimal results, we recommend using `nLoci`=1000. Locus data is stored in the ***locus_datafile***, and the genomic boundaries of the `nLoci` loci are stored in ***bed_global_dataset.txt***. The program ***vcf2abc.py*** uses ***bed_global_dataset.txt***, ***vcf file***, and ***popfile.csv*** to compute the average summary statistics across `nLoci` loci stored in ***ABCstat_global.txt*** (see 2.3.1 for summary statistics computation method).

Generating priors and simulated datasets

The script ***submit_priorgen_newsim.py*** generates prior simulation parameters using the method explained in 2.3.1. The parameters are stored in ***modelComp/{N}_{I}/priorfile.txt*** (where N is the model name and I is the number of replicates) and in a *scrm* compatible command format. After generating the prior parameters, simulations are run from them using *scrm* (Staab et al. 2015). The results of these simulations are then piped directly into ***scrm_calc.py***, which transforms the simulation results into summary statistics using the same method explained in 2.3.1. The results of ***scrm_calc.py*** are stored in ***modelComp/{N}_{I}/ABCstat_global.txt***.

Inferring average demographic history

The modelComp folder contains the "reference table" – a table containing the parameter for simulations and the summary statistics – on which a RF is trained to infer each parameter. Each model among the 14 demographic x genomic models, has a different number of parameters. So at first ***estimate_posterior_and_mw.R***, put each model under the "hypermodel" – a model that uses all parameters, all models combined – by filling missing parameters (see Table A.1). Then, for each of the 12 parameters of the hypermodel, a regression RF (*regAbcrf* from *abcrf* R packages Raynal et al. (2019)) is trained. It predicts from summary statistics of the observed dataset (stored in ***ABCstat_global.txt***), the parameters values. Joint posteriors and model weights are generated following the procedure explained in 2.3.1 and stored in ***posterior.txt*** and ***model_weight.txt***. The summary statistics of ***posterior.txt*** are stored in ***sim_posterior/ABCstat_global.txt***.

Estimating the goodness of fit of prior and posterior parameters

The goodness of fit of the posterior distributions (from ***sim_posterior/ABCstat_global.txt***) and prior (from ***modelComp/{N}_{I}/ABCstat_global.txt***) are evaluated using an en-

hanced version of the *gfit* function of the abc packages (Csillery et al. 2012) coded in ***gof_estimate.R***, which uses a goodness-of-fit statistical approach described in 2.3.1. The results of the posterior and prior goodness of fit are stored in ***gof_posterior.txt*** and ***gof_prior.txt***, respectively.

Simulating locus reference table and Detecting barrier

The ***posterior.txt*** file, containing posterior parameters, undergoes transformation into simulation parameters, with half of simulation migration set to $m = 0$ and the second half $m > 0$ using the ***submit_priorgen_locus.py*** script and is subsequently stored in ***sim_locus/priorfile_locus.txt***. In contrast to the previous simulation step, where only average information across `nLoci` were retained, this iteration preserves information at the locus scale. Similar to the ***submit_priorgen_newsim.py*** procedure, simulation parameters are formatted in *scrm* command style and executed using *scrm*. The resulting simulations are then immediately transformed into summary statistics through the utilization of ***scrm_calc.py***. This time, summary statistics at the locus level are stored in ***sim_locus/ABCstat_locus.txt***. Combining ***sim_locus/priorfile_locus.txt*** and ***sim_locus/ABCstat_locus.txt*** forms the "locus scale reference table", on which the RF algorithm *abcrf* (from the *abcrf* package Raynal et al. (2019)) is trained to classify barriers ($m = 0$) and non-barriers ($m > 0$). Subsequently, using summary statistics for each locus from ***ABCstat_locus.txt***, the RF classifies each locus into the barrier or non-barrier class. From the RF outcomes, Bayes factors are computed (refer to 2.3.1 for details on bayes factor computation). The Bayes factors and corresponding summary statistics for each locus are stored in ***Pbarrier.txt***. Variable importance for each variable, following the Random Forest model, is stored in ***variable_importance_barrier.txt***. Additionally, the confusion matrix, as defined in the Random Forest context, is stored in ***confusion_matrix_barrier.txt***. Finally, the average barrier proportion and resulting barrier/non-barrier ratios are stored in ***barrier_proportion_and_ratio.txt***.

3.2.6 Example of usage of RIDGE and recommendations

The RIDGE code comes with a small test dataset in the ***example/*** folder. In this part, I describe step-by-step the process of using RIDGE on a dataset and how to interpret RIDGE's outputs. The dataset is a subset of the published dataset in Poelstra et al. (2014) and Vijay et al. (2016), stored at NCBI under PRJNA192205, containing 9 out of 1300 scaffolds, where SNPs were called on the following reference genome GCF_000738735.1 (available at NCBI). This subsample focuses on the 9 scaffolds where genes of interest were found in Poelstra et al. (2014). The individuals used are the same as in Poelstra et al. (2014). The example folder contains :

- An archive name test_dataset.vcf.tar.gz
- A preset contig_data.txt

- A preset popfile.csv
- A list of the genes of interest detected in Poelstra et al. (2014) and their positions

Setup

Prepare files : First, go in the ***example*** folder and decompress the vcf file:

```
cd example
tar -xvzf test_dataset.vcf.tar.gz
```

Then, create the file ***config.yaml***, by taking the template from template folder:

```
cp ../template/RIDGE_template.yaml config.yaml
```

Adapt the content from config.yaml to your installation (follow instruction from 3.2.3). To correctly fill your config files you will need a measure of the diversity and divergence. At first, you are not obliged to set up values for the prior bounds (`N_min, N_max, N_ref, Tsplit_min, Tsplit_max, M_min, M_max, Pbarrier_max`), but in the end they are mandatory. Here, leave all parameters empty exept for `M_min, M_max` and `Pbarrier` to fill with the following value : `M_min`=0.1, `M_max`=50 and `Pbarrier_max`=0.2.

Scan genome & setup prior bounds

Launch RIDGE in scan mode,

```
bash <path to RIDGE>/RIDGE.sh <work_dir>/config.yaml scan
```

The genome will be scanned and summary statistics will be generated for each window. RIDGE also suggest prior bounds based on statistics (available in ***prior_bound_suggestion.txt***). It is advisable to utilize these values as they are scaled to the `mu` value provided in the config.yaml file. In the event of any changes to the `mu` values, rerun or re-adjust other values.

Test prior bounds

!!! Inference power of RIDGE depends on the quality of priors. So, pay attention to prior bounds. In the following steps, it explains how to measure the quality of prior bounds !!!

To test prior bounds, launch RIDGE in test mode using the following command:

```
bash <path to RIDGE>/RIDGE.sh <work_dir>/config.yaml test
```

This way, only 1% of the simulation in ***modelComp*** folder are launched, then the goodness of fit of prior is evaluated in ***gof_prior.txt*** and ***QC_plot/QC_prior_acp.pdf*** and ***QC_plot/QC_prior_density.pdf***. To demonstrate the significance of selecting a suitable prior bound, I executed the RIDGE in test mode with a "good" and a "bad" choice of priors.

parameter	"Good" prior bounds	"Bad" prior bounds
N_min	42500	140500
N_max	147500	300500
Tsplit_min	5000	100000
Tsplit_max	120000	120000
M_min	0.1	0.1
M_max	50	50
Pbarrier_max	0.2	0.2

Table 3.1: Values of prior bound for "good" (generated based on ***prior_bound_suggestion.txt***) and "bad" (value choosen arbitrary) prior bounds

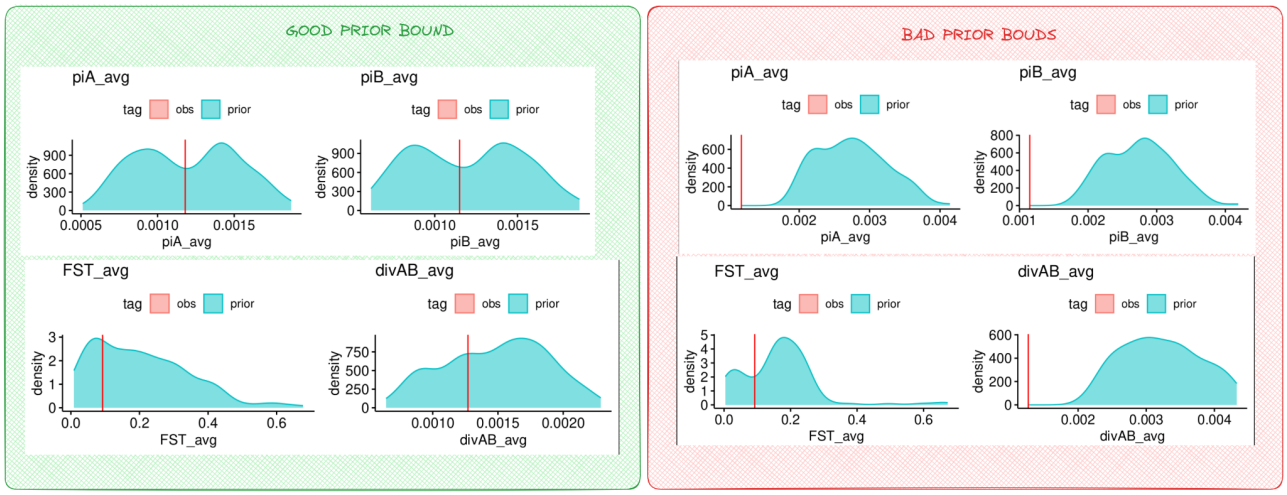

Figure 3.12: Distribution of Summary Statistics from Simulation Generated Within Prior Bounds. "Obs" represents the observed dataset value for the summary statistic, while "Prior" denotes the distribution of summary statistics generated from simulations within the specified prior bounds. The "Good" choice of prior is deemed favorable, as the observed value is contained within 95% CI values of the prior distribution. Conversely, the "Bad" choice shows observed values falling in the distribution tail or outside the bounds of the prior distribution.

For the "good" choice, I used the value suggested in the ***prior_bound_suggestion.txt*** file, while for the "bad" choice, I chose a version biased towards higher values.

In the "bad" prior bound choice, the observed values fall outside the distribution range for piA_avg, piB_avg, and divAB_avg statistics (respectively the π of population A and B and the D_{xy} between both populations). This shows that the prior bounds do not include the true observed value, contrary to "good" prior bound choice (see Figure 3.12). It is necessary to evaluate the quality of prior bounds by examining the prior density in ***QC_plot/QC_prior_density.pdf*** before running the entire RIDGE. Additionally, users may evaluate the quality through PCA on summary statistics (accessible at ***QC_plot/QC_prior_acp.pdf***), but this method is less sensitive and less informative. With this visual approach, the prior must include the observations to validate the prior. Finally, the goodness of fit (GOF) test can be used, but it is only reliable if the threshold is set at 5%, indicating a "bad" prior estimation. If the observations demonstrate a poor quality of prior bounds, follow these steps:

```
cd <work_dir>
rm -r modelComp/ QC_plot/ gof_prior.txt
```

And then test relaunch RIDGE until you achieve satisfactory prior bounds. It is recommended to use large prior bounds as there is little cost associated with doing so. !!! For N and Tsplit, avoid having more than two orders of magnitude between your min and max bounds. !!!

Analyze outputs

To launch RIDGE use the following command :

```
bash <path to RIDGE>/RIDGE.sh <work_dir>/config.yaml all
```

Depending on the `lightMode` option, the runtime is around 470000 s for `lightMode`=True and 1883000 s for `lightMode`=False (assuming that you run RIDGE on 4 cores running at least at 2.5GHz) for a dataset with a ρ/θ=20 with `nLoci`=1000 and `window_size`. Run-time can be multiply by 3-4 for high ρ/θ (e.g.ρ/θ=500).

Check the quality of demographic infereneces

To evaluate the quality of demographic inferences, there are two levels of verification:

- Agnostic Level: At this level, the assessment involves confirming that the posterior aligns more closely with the observed dataset than the prior.
- Documented Level: This level involves checking whether the parameter values are consistent with established knowledge. The second level is optional and relies on the user's familiarity with the dataset.

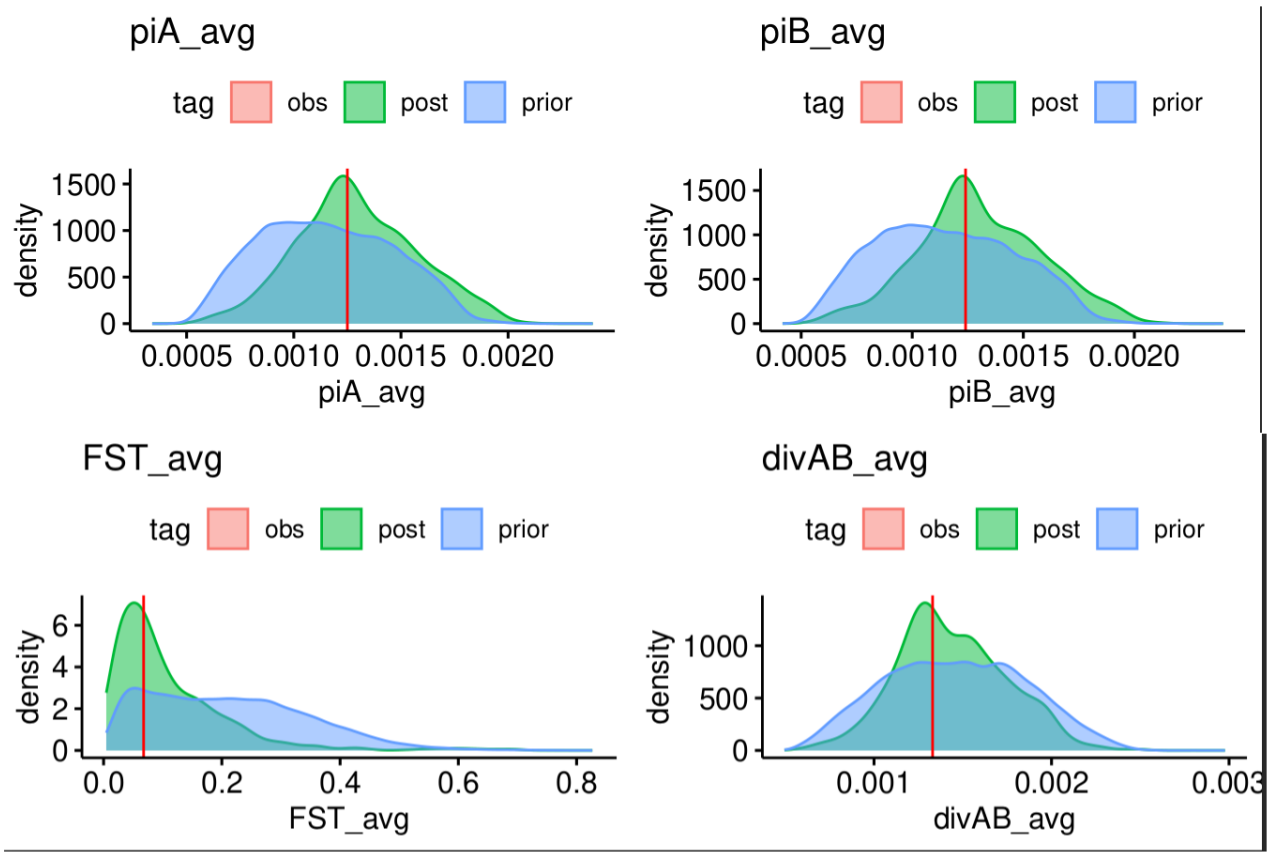

Figure 3.13: Distribution of Summary Statistics for Prior and Posterior Simulations. "Obs" represents the observed dataset value for the summary statistic, "Prior" denotes the distribution of summary statistics generated from simulations based on the prior, and "Post" illustrates the distribution of summary statistics from simulations of the posterior.

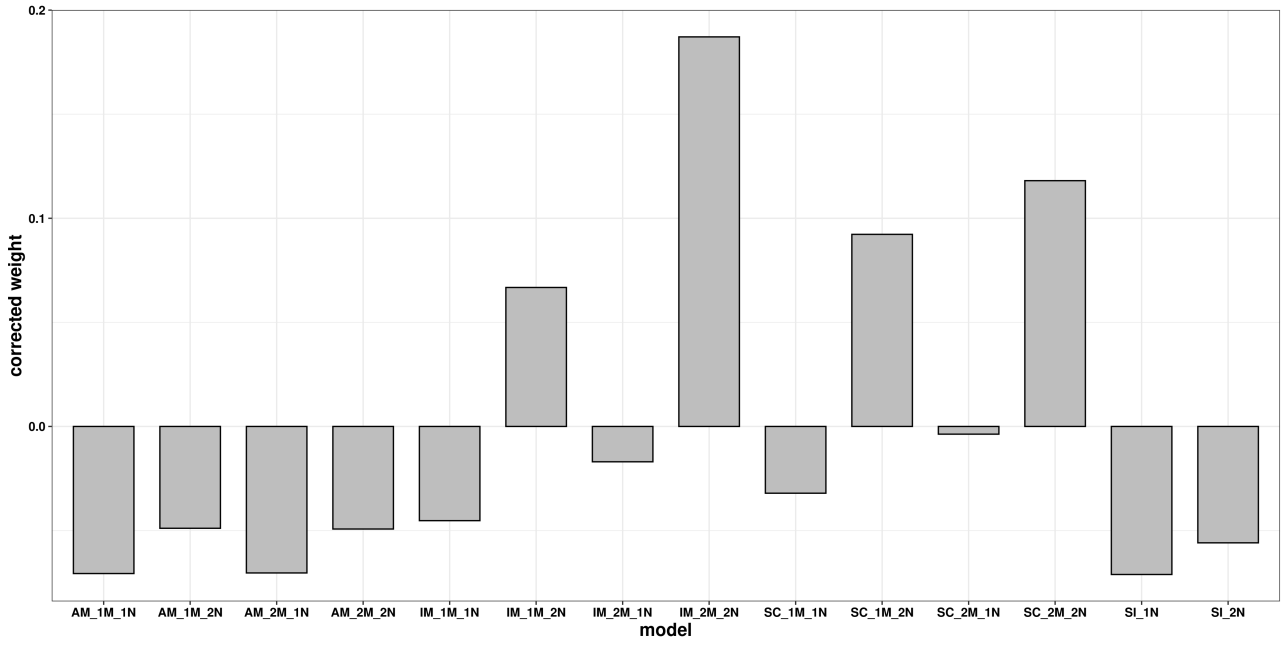

Figure 3.14: Model weight of each 14 model corrected by the uniform distribution model weight

For the first level of checking, we use visual observation of posterior distributions (available in ***QC_plot/QC_posterior_density.pdf*** and ***QC_plot/QC_posterior_acp.pdf***) and evolution of goodness of fit between ***gof_prior.txt*** and ***gof_posterior.txt***. We expect posterior distributions to be more centered on the obs dataset and the distribution less wide than for priors (as observable Fig 3.13). Furthermore, the goodness of fit should increase between prior and posterior. In our example the goodness of fit slightly increased from 0.26 (for prior) to 0.31 (for posterior).

The second level of verification relies on pre-existing knowledge. For this dataset, T_{split} is anticipated to be approximately 80,000 generations (Poelstra et al. 2014; Vijay et al. 2016), with an effective population size ranging between 100,000 and 300,000 individuals. The demographic model is expected to align with a secondary contact model. Results, available at ***model_weight.txt*** and ***visual_model_weight.pdf***, indicate a predominant contribution of the IM_2M_2N and SC_2M_2N models to the estimation of demographic history, which is consistent with existing data. Indeed, both models (IM & SC) involve ongoing migration and exhibit heterogeneity in both migration (2M) and effective population size (2N)(see Figure 3.14). The estimated value (available at ***posterior.txt*** and ***visual_posterior.pdf***) of T_{split} closely matches the expected value (average $T_{split} = 73414$ generations), and the population size falls within the anticipated interval ($\approx 100,000$ for N1, N2, and Na)(see Figure 3.15). The migration rate is higher for current migration (M_{cur}) than ancestral migration (M_{anc}), suggesting an increase in migration over time, even if the estimated model is not optimal.

Detecting barrier to gene flow

In this section, I distinguish between 1) detecting barriers to gene flow for a specific dataset without a comparative framework and 2) detecting barriers to gene flow across multiple datasets with the goal of comparing results between datasets. This difference is crucial as it affects how

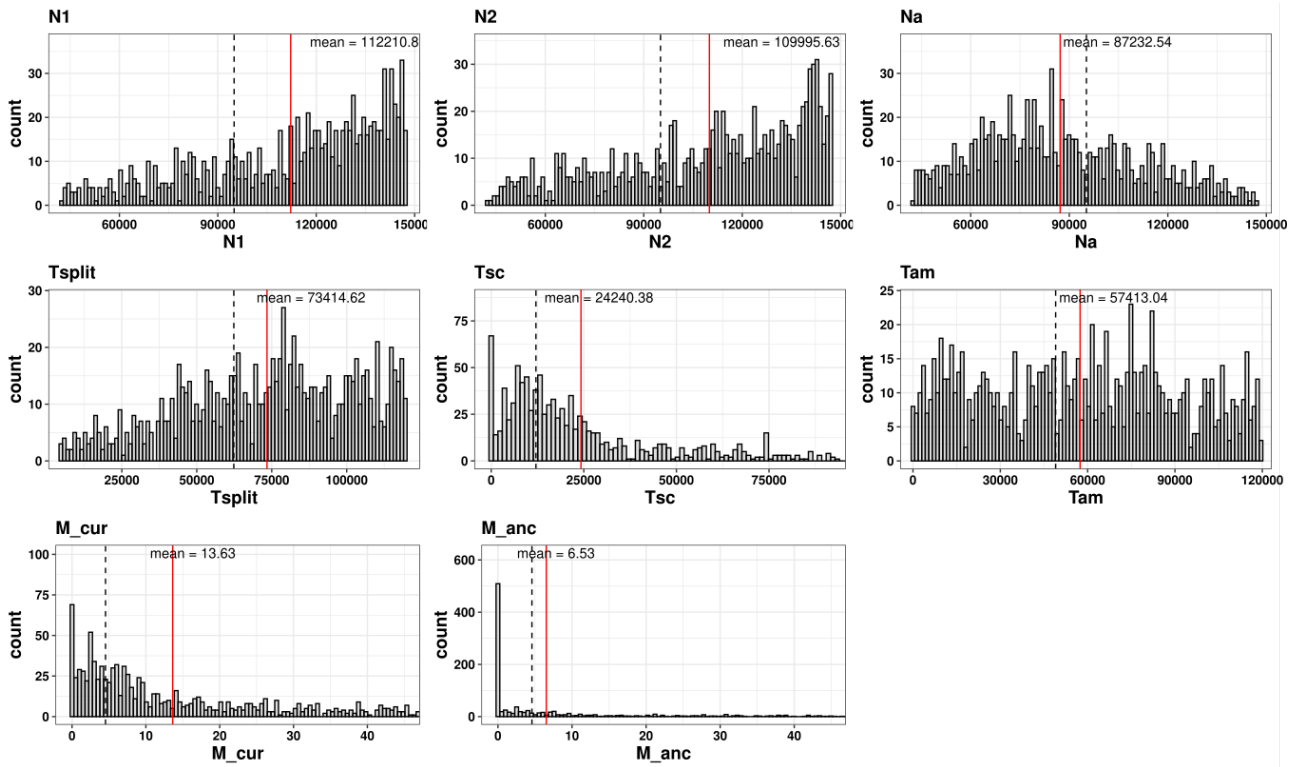

Figure 3.15: Distribution of parameter posterior values for N_1, N_2, N_a, T_{split}, T_{SC}, T_{AM}, M_{cur}, and M_{anc}. Dashed lines represent the mean value of priors, and the red line represents the mean value of posterior.

Allocation\true status	barrier	non-barrier	class.error
barrier	95695	1464	0.0150
non-barrier	1285	101407	0.0125

Table 3.2: Confusion matrix produced for the example case

the user should utilize the different outputs.

Quality of barrier detection

Test of barrier detection quality relies on two elements: i) the confusion matrix of the random forest and ii) the distribution of variable importance of the random forest. In our example, the obtained confusion matrix (available in ***confusion_matrix_barrier.txt***) shows the distribution of allocation of simulated loci used to train the random forest. It demonstrates the ability of the RF to accurately classify the simulated loci. Here, 95695 loci simulated as barriers have been classified as barrier and 1464 as non-barrier, corresponding to an error of 1.5% (see Table 3.2). The higher the class error, the lower the confidence in the results. A high degree of class error (class.error=0.5) indicates that the Random Forest (RF) is unable to distinguish between barrier and non-barrier loci.

Furthermore, the variable importance, available at ***variable_importance_barrier.txt***, provides insights into which variables enable the RF to distinguish between barrier and non-barrier classes. Variable importance is calculated by considering how much each feature contributes to differentiate barrier from non-barrier classes across all the trees. Higher importance

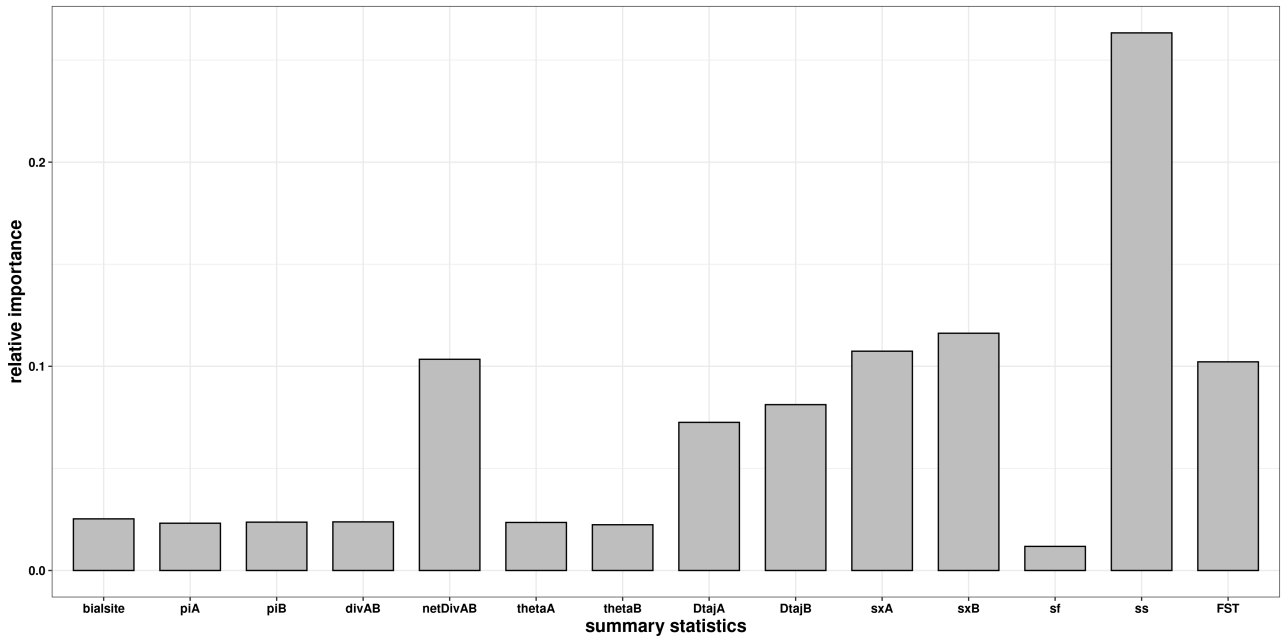

Figure 3.16: Relative importance associated with each summary statistic during the random forest building for the example case.

values indicate more influential features. The expectation was that some summary statistics such as ss, F_{ST}, netDivAB ($=D_a$), and sf would be important for barrier barrier detection. Our results are consistant with this prediction (see Fig 3.16). A pattern where all variables have the same importance reduces confidence in the results, as the RF is then unable to prioritize specific summary statistics to detect barrier loci.

Detecting barrier for a focal dataset

RIDGE provides a posterior probability (post.prob) for each locus to quantify the probability of it being a barrier, which is accessible through ***Pbarrier.txt***. Loci with a post.prob > 0.5 have a high chance of being a barrier. By nature post.prob are not comparable from one dataset to another, as they are conditionated to demographic model. So, if there is only one dataset post.prob could be used. We suggest setting a threshold of at least $post.prob > 0.5$, and even stricter if necessary. In the example presented in Figure 3.17, a threshold of 0.5 effectively distinguishes the two observable groups of loci in the distribution.

Detecting barrier across multiple dataset

RIDGE provides, for each locus, a posterior probability and the Bayes Factor (BF) (refer to section 2.3.1 for details). By default, a BF is interpreted as follows: BF=10 means that the locus has 10 times more chance to be a barrier than a non-barrier, given Q. A BF=100 indicates a very high probability that the locus is a barrier. Compared to post.prob, BF is comparable from one dataset to another and allow the user to compare multiple datasets. To compare multiple datasets, the user must choose a threshold in BF value that is applied consistently across all datasets. A BF value of 100 is a relevant starting point, but it may not be suitable for all datasets. Indeed, depending on the datasets, there might be significant variations in the

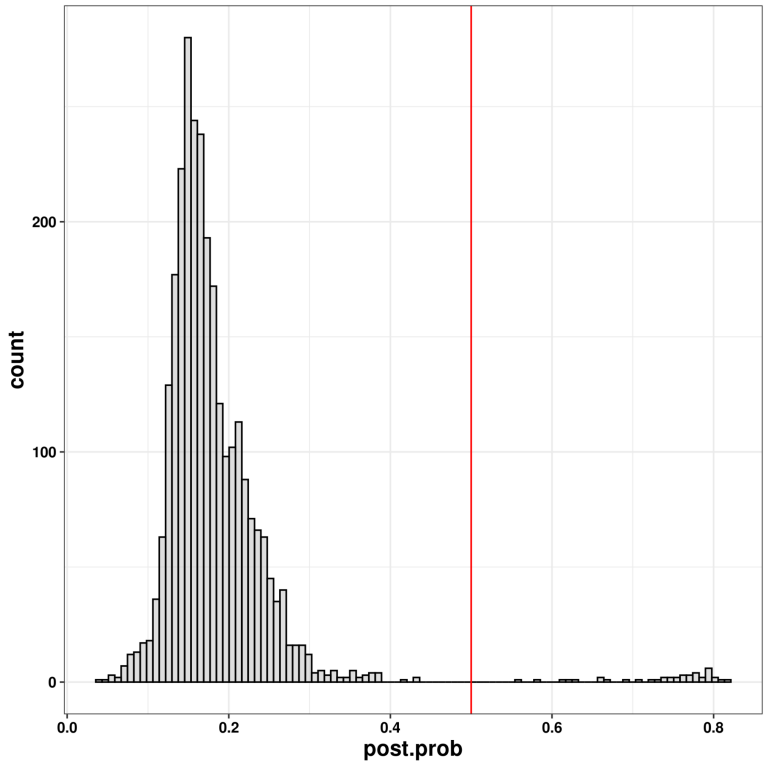

Figure 3.17: Distribution of posterior probability of barriers across the genome of the example data. The threshold for post.prob is set to 0.5 with a red line.

scales of BF (e.g., one dataset may range BF values between 0 and 100, while another ranges between 0 and 3000). This variability can occur when a dataset has a low T_{split}, causing BF to be biased toward high values. In such cases, we recommend using the "BF_approxQ" column in ***Pbarrier.txt***, as the approximation is more reliable under these circumstances (see 2.3.1 for details).

Chapter 4

Application on empirical dataset

Chapter 2 demonstrates the performance of RIDGE through simulated and crow datasets that diverged approximately 90,000 years ago. This chapter tests RIDGE under low divergence times by applying it to domesticated systems, including maize and foxtail millet that underwent recent human-mediated selection. Both systems diverged from their respective most direct wild relative ("ancestor") around 9000 years ago. However, they have different mating systems; maize is an outcrosser while foxtail millet is a selfer. This chapter aims at examining RIDGE's ability to detect barriers even at very low divergence times and provide insights to the following questions: can we compare the results between these two different biological systems? What does RIDGE detect as a barrier? What is the effect of the mating system on RIDGE results? What additional information would be relevant to improve our understanding of the results? Preliminary analyses of this work were conducted during the M2 internship of Augustin Desprez in 2022 (for foxtail millet) and M1 internship of Clementine Tocco also in 2022 (for maize), both of whom I co-supervised with Maud Tenaillon. I worked more recently in close collaboration with a PhD student, Arthur Wojcik, who was recruited on the project to continue my work and apply RIDGE to multiple wild/domestic systems. I contributed to train him on implementing RIDGE to various datasets. See Appendix B for additional figures and tables.

4.1 Data generation

For both systems, 20 wild and 20 cultivated individuals were sampled and their genome sequenced with Illumina. Stella Huynh of IRD Montpellier conducted data cleaning, alignment, and SNP detection analyses using the following pipeline (note that steps are the same for both species): First, reads were cleaned using *cutadapt* (Martin 2011) separately for R1 and R2 (respectively "forward" and "reverse" reads), and filtered to keep only pairs of reads. Second, mapping was performed on the reference genome using *bwa-mem2 v2.2.1* with default parameters (Li et al. 2009). Finally, SNP calling was done using *GATK4* (McKenna et al. 2010). The *GATK HaplotypeCaller*, *genomicsDBImport*, and *GenotypeGVCFs* tools were applied to each chromosome individually. Subsequently, SNP filtering was performed using *GATK4 VariantFiltration* and *SelectVariants*, also on a per-chromosome basis. During this step, we applied

the following filters at each SNP: a coverage depth (DP) above 2.5 times the number of individuals for selfing species, and 5 times the number of individuals for outcrossers; a quality (QUAL) above 30 for selfers and 60 for outcrossers. Finally, all chromosomes were combined using *bcftools concat* (Danecek et al. 2021) and filtered based on the depth of coverage per site and per individual. For allogamous individuals, the depth of coverage must be superior to 5X, while for autogamous individuals, it must be superior to 3X. Additionally, all SNPs with more than 10% missing data across all individuals were filtered out. In addition, we phased the obtained data using *Shapit2* (Delaneau et al. 2008) and, for selfing species, we haploidized data using a custom script. Then, I extracted a single haplotype per individual for future comparison between allogamous and autogamous species. It is worth noting that diversity can be up to twice as high in a heterozygous individual as in a homozygous one, leading to a strong comparison bias between cross-pollinating species, such as maize, and self-pollinating species, such as millet. However, for a species that is over 90% homozygous, such as millet (P. Huang et al. 2014; Jia et al. 2013), this phasing/haploidization step has little impact on the results.

4.2 Maize

4.2.1 History of maize and demographic context

Maize is an emblematic plant of Mesoamerican culture and is the species that has historically served as a model for domestication studies. Molecular and archaeological data indicate that maize was domesticated from an annual grass of the teosinte subspecies *Zea mays ssp. parviglumis* (thereafter parviglumis) around 9,000 years ago (Matsuoka et al; 2002). The cradle of its domestication is thought to be located in south western Mexico, in the plains of the fertile Balsas river basin. Unlike most species, cultivated maize (*Zea mays ssp. mays*) is quite distinct from parviglumis, and its origin has long remained a mystery. Several key characters differentiate the two forms (Beadle, 19395): (1) teosinte has long, elongated branches terminating in a male inflorescence; in maize, the lateral branches are very short and terminate in female inflorescences (cobs); (2) teosinte has numerous tillers, whereas most maize has a single tiller; (3) maize cobs have more rows and grains than teosinte cobs; (4) teosinte cups have abscission layers at their base, enabling the kernels to separate at maturity and disperse, whereas in maize, abscission layers are absent and the cob retains its integrity at maturity; (5) in teosinte, the kernels are covered by glumes that harden at maturity, making the kernel difficult to access; in maize, the kernels are naked and therefore tender at maturity (J. Doebley 1992).

At the genomic level, domestication has led to a reduction in the overall diversity of domesticated forms compared with wild forms. This reduction, estimated at nearly 40% in maize (Wright et al. 2005), is due on the one hand to a demographic effect, since domesticated forms have been selected from a reduced number of wild individuals (bottleneck), and, on the other hand, to intense selection that led to the fixation of "domesticated" alleles, also reducing diversity at neighboring neutral loci through genetic hitchhiking. To date, 11 domestication genes have been described in maize (Table B.1). In addition to these genes, many other regions have

been detected. In fact, around 2-4% of the coding genome is thought to have contributed to domestication (S. I. Wright et al. 2005). Maize has a fairly large genome, of 2.3 gigabases divided into 20 chromosomes (2n = 20) (Schnable et al. 2009). It is a highly repetitive genome of which around 85% consists of transposable elements and genome size can vary from line to line (Diez et al. 2013). The Zea mays species complex comprises another annual subspecies than parviglumis, a teosinte called *Zea mays ssp. mexicana* (therefafter mexicana). Both grow in Mexico, but in different conditions of temperature, humidity and altitude. Mexicana grows mainly at high altitude (1600-2700 m) in the relatively dry regions of central Mexico. Parviglumis, in contrast is adapted to the warmer, wetter low to medium altitudes of southwestern Mexico (below 1,800 m)(M. B. Hufford et al. 2012). The genetic differentiation of mexicana from the lowland parviglumis was influenced by its adaptation to low temperatures and soils with low phosphorus content in highland environment (Aguirre-Liguori et al. 2019a).

Gene flow between Zea subspecies occurs naturally, due to their geographical proximity. There is ample evidence suggesting that the adaptation of maize from low altitudes to high plateaus was facilitated by introgression of alleles belonging to the mexicana subspecies (Hufford et al. 2013). Gene flow also exists between parviglumis and mexicana (Aguirre-Liguori et al. 2019a). Finally, gene flow from cultivated maize to the European teosinte of mexicana ancestry has contributed to the adaptation of the latter, facilitating its establishment as a weed. European teosinte has indeed acquired an early-flowering allele from maize as well as herbicide resistance allele (Le Corre et al. 2020). So, although domesticated maize is the result of domestication of individuals from the parviglumis subspecies, it is also the product of intense gene flow from mexicana (Yang et al. 2023).

These introgressions are sometimes beneficial (adaptation to the altitude of cultivated maize), but can also be counter-selected (hybrid depression). Unlike parviglumis, experiments involving hand-crossing reveal that mexicana and maize demonstrate genetically rooted cross-incompatibility (Kermicle 1997; Baltazar et al. 2005). In line with those results, natural hybridization rate between maize and parviglumis was estimated at 100% - based on a single parviglumis population, while for mexicana, it was «1%. The low rate of natural hybridization with mexicana can also be partially explained by short silks, short-lived pollen and earlier flowering in mexicana compared to maize (Baltazar et al. 2005; Rodriguez et al. 2006).

4.2.2 Maize genes involve in RI

Three distinct post-meiotic and pre-zygotic loci, *Ga1*, *Ga2* and *Tcb1*, have all been identified as barriers between maize and mexicana (M. M. S. Evans and J. L. Kermicle 2001; Chen et al. 2022; Wang et al. 2022). Those are characterized as unilateral cross incompatibilities as they prevent mexicana to be pollinated by maize while maize can be freely pollinated by mexicana. All of these loci consist of both a female and a male determinant, associated with different genes. The female determinant is expressed in silks and inhibits pollen germination. The male determinant is expressed in pollen and is able to overcome the inhibition of the female determinant. The *Ga1* locus encompasses two genes responsible for regulating cross-incompatibility. ZmPme3 codes

for a pectin methylesterase (PME) expressed in silks (Moran et al. 2017; Wang et al. 2022; Zhang et al. 2023) that hinders pollen tube growth, hindering pollination by maize varieties lacking a functional version of the second gene at the *Ga1* locus. The second gene, named *ZmGa1P* (Zhang et al. 2018), also encodes a PME. Expressed in pollen, this gene enables pollen carrying it to overcome the cross-pollination barrier imposed by *ZmPme3*. In addition to the initially reported single *ZmGa1P* gene, several extra tandem repeated sequences and full-length duplicates of *ZmGa1P* form the male function (Wang et al. 2022; Zhang et al. 2023). Three haplotypes of the *Ga1* locus have been delineated based on the functionality of these two genes. For example, Ga1-S carries functional *ZmPme3* and *ZmGa1P*, while ga1 carries neither. *Ga1-M* possesses a functional *ZmGa1P* but lacks a functional *ZmPme3* (Lu et al. 2020). Consequently, three natural haplotypes can be identified. The S haplotype possesses both active male and female determinant alleles, the M haplotype only has the male determinant allele, and the wild haplotype lacks both active determinant alleles. S haplotypes can pollinate all other haplotypes but can only be pollinated by other S or M haplotypes. M haplotypes are capable of pollinating all haplotypes, but they are also susceptible to pollination by the wild haplotype. The wild haplotype can be pollinated by all others but cannot pollinate the S haplotype. Consequently, the M haplotype exists as an intermediate haplotype between a complete barrier and the absence of a barrier.

Note that S haplotypes are also present in some populations of parviglumis. Maize predominantly comprises wild haplotypes, with only a limited presence of M haplotypes in sympatric areas (M. M. S. Evans and J. L. Kermicle 2001; Chen et al. 2022; Wang et al. 2022) but also of S as in popcorn varieties (Bapat et al. 2023).

Two other unidirectional cross-incompatibility systems known as *Ga2* and *Tcb1* are functionally analogous but incompatible with Ga1 and are located at different genetic loci. The *Ga2* locus has been mapped to chromosome 5 (Chen et al. 2022). The *Tcb1* locus is located approximately 44 cM away from the *Ga1* locus (M. M. S. Evans and J. L. Kermicle 2001). Y. Lu et al. (2019) described the female function gene of the *Tcb1* locus, *Tcb1-f*, which encodes a PME protein slightly differing from *ZmPme3*, and the male function is also a PME gene (Zhang et al. 2023).

Another barrier genetic system between maize and mexicana, named TPD for Teosinte Pollen Drive, has been described, has a post-meiotic post-zygotic barrier affecting hybrids fitness (Berube et al. 2023). TPD is a male meiotic driver that manipulates transmission using a poison-antidote system. Although it operates in a similar fashion to a Bateson-Dobzhansky-Muller interactions, it is a special case as backcrossing hybrids into the teosinte background results in significant to complete pollen abortion in the progeny no matter the number of generations. This barrier locus also acts as a one-way barrier, affecting only progeny in the teosinte background and allowing gene flow from teosinte to maize, but not the other way.

The TPD system relies on the interaction between three loci: *Tpd1*, *dcl2T*, and *Tpd2*. *Tpd1* produces abundant 22nt long hp-siRNAs, which disrupt the functioning of *Tdr1*, a lipase necessary for pollen grain development encoding locus. *Tpd2* and *dcl2T* both act as partial

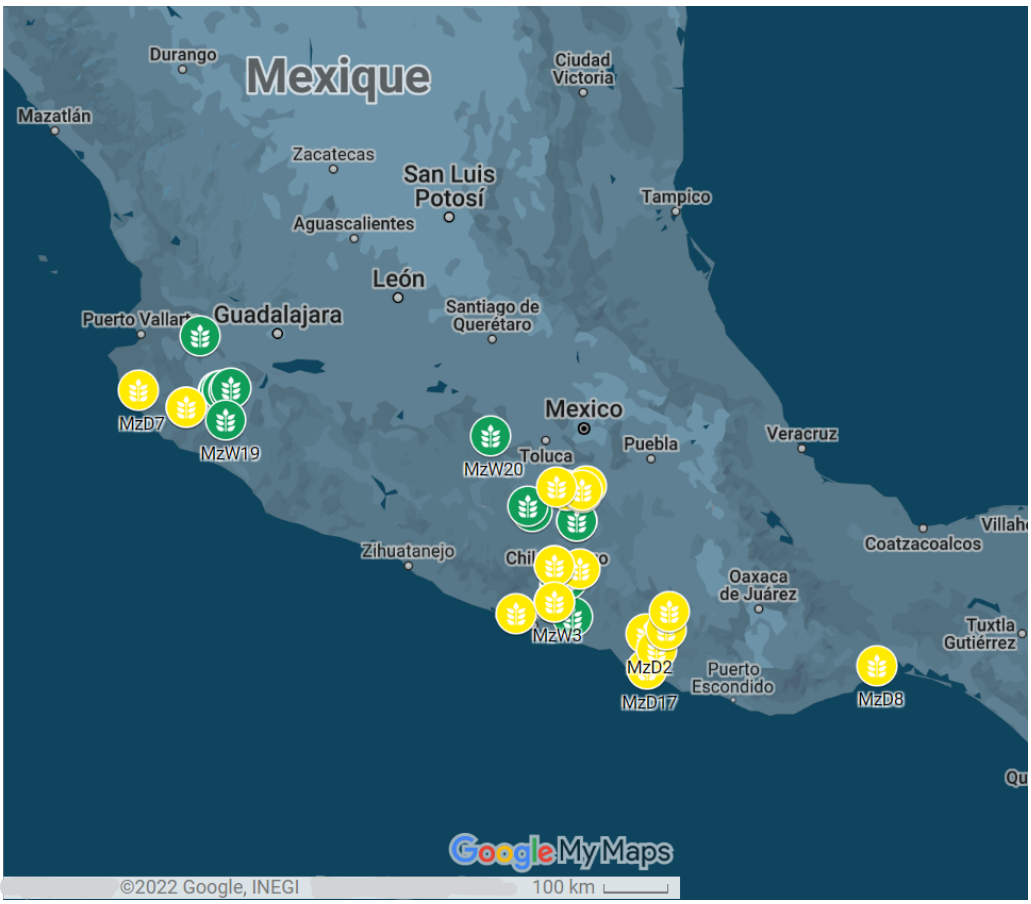

Figure 4.1: geographical origins of maize (yellow) and teosinte (green) varieties studied (source: GoogleMyMaps)

antidotes, repressing secondary processing of siRNAs and restoring viability and fertility. An original aspect of this system is that while *Tpd1* and *dcl2T* are both on chromosome 5, *Tpd2* is located on chromosome 6, resulting in the abortion of 3/4 of tetrad in hybrids.

4.2.3 Genomic material

We used a sample of 20 wild individuals of parviglumis subspecies. These individuals were selected from 19 different populations of the same genetic group (G1, Aguirre-Liguori et al. (2019)). Two individuals were obtained from the Paso Morelos population, which exhibits a relatively high level of diversity but also some inbreeding (with F_{IS} value close to 0.2). The 20 domesticated individuals of lowland maize were selected from 20 traditional varieties grown at altitudes below 1800m to ensure genetic proximity to parviglumis. The sampling strategy aimed to minimize intra-form genetic structure and distance from the center of domestication (Figure 4.1). By studying traditional varieties, we avoided as much as possible modern varietal selection, which often diversifies due to the diversity of uses and food preferences, and instead focus primarily on the selection of domestication traits.

Sequence alignment was performed on the maize B73 reference sequence (Zm-B73-REFERENCE-NAM-5.0 Hufford et al. (2021). The sequencing depth for our data was between 10X and 15X for all individuals. In total, 266 278 466 SNP were detected and only 18 359 214 SNP passed

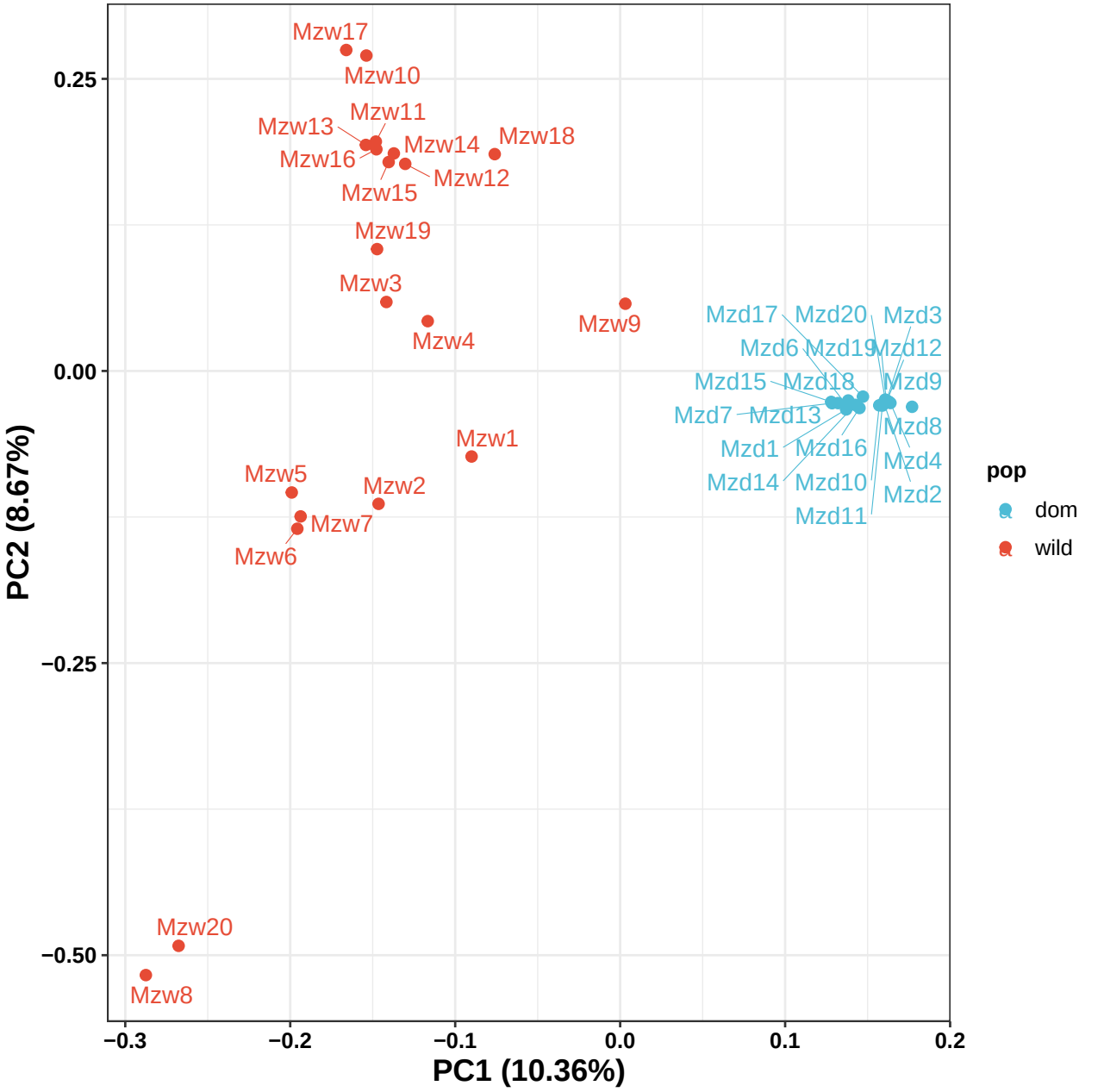

Figure 4.2: PCA performed on all filtered SNPs in maize dataset

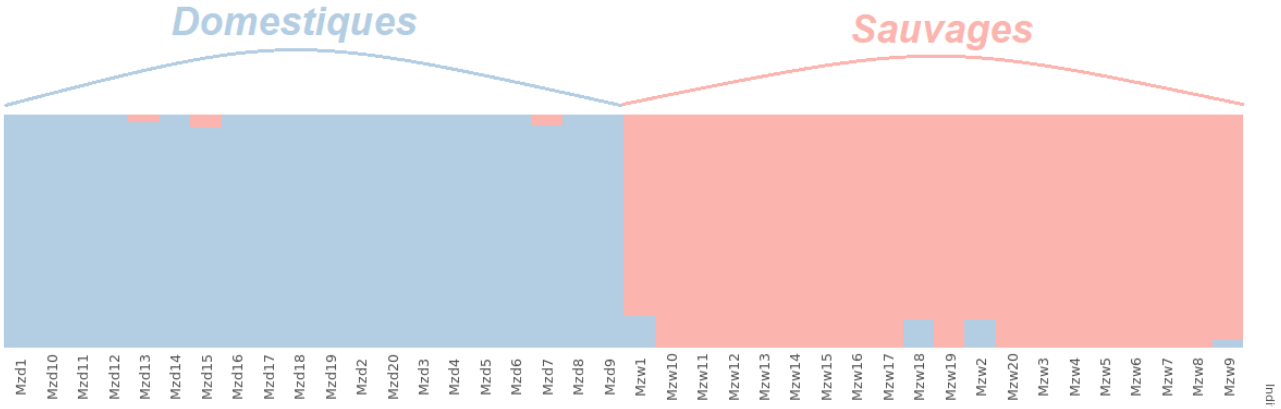

Figure 4.3: Genetic structure for 2 groups (K=2). This grouping received the lowest cross-validation (cv) (best K value).

all filters (which represent 6.89% of all detected SNPs). Note there is a difference in raw diversity between our rawdata and and data from Beissinger et al. (2016). This difference might be due to sampling differences as well as SNP calling filter strength (Table B.2). Principal component analysis (PCA) of the SNP set demonstrates a clear differentiation between the wild and domestic forms on axis 1, which explains 10.36% of the variation (see Figure 4.2), but little structure within wild genetic pools (MzW8 & 20 are genetically isolated from the rest of wild individuals) and no structure in the domestic genetic pools As expected, the domesticated individuals are much more clustered on the PCA than the wild individuals (Figure 4.2). which indicated greater genetic diversity in teosintes than in maize. An analysis using the *Admixture* software resulted in an optimal categorization of two genetic groups (K=2). The individuals are grouped by their wild or domestic form. Figure 4.3 shows that the observed structuring is mainly due to domestication.

4.2.4 Application of RIDGE on maize dataset

By applying RIDGE to our maize/parviglumis dataset, we expected the models with recent gene flow (in particular the IM model) to receive the highest weight and we also expected to detect few genetic barriers. As haplotype S, that is responsible for incompatibilities with maize, is rarer in parviglumis than in mexicana (Wang et al. 2022), we were unsure to detect *Tcb1/Ga1* and *Ga2* in our data. Finally, the maize dataset provided an opportunity to test whether RIDGE would detect domestication genes (see Table B.1), which underwent strong selective sweep in the domestic form, as barriers. It is indeed unclear whether such genes that have most likely acted as postzygotic barriers where hybrids are counterselected in the two environments - the field and the natural environment - are detected since the expected signal differs at least partly from barrier loci which do not necessarily display intra-form selection (M. I. Tenaillon et al. 2023). We set prior bounds based on real data and literature information as follows: N_e [1,000; 150,000], M [0.1; 50], T_{split} [1,000; 50,000], Pbarrier_max=0.2, and Nref = 75,000 using a window size of 50kb. We first ran RIDGE, taking into account heterogeneity of diversity (with 'hetero θ' set to True) and heterogeneity of recombination (with 'homo_rec' set to False). The recombination map for maize cM was created using the genetic map of the B73 maize model (Brazier and Glémin 2022).

Demographic inferences

As expected, the most favored demographic model is the IM model, which incorporates hetero-m and hetero-Ne, and the models under ongoing migration represent 90%.Note that the IM model that includes heterogeneity in migration (2M) concentrates 37% of the model weights – IM_2M_2N=16% and IM_2M_1N=21% –, which is consistent with the existence of barriers between maize and parviglumis. The posterior summary statistics fit the data well (G_{post} = 0.567). Hence, the goodness of fit for each summary statistic shows that the posterior distributions are more concentrated around the observed dataset than the prior (Figure B.3). The ratio of the population size are close to previous estimates of the diversity loss of 40%

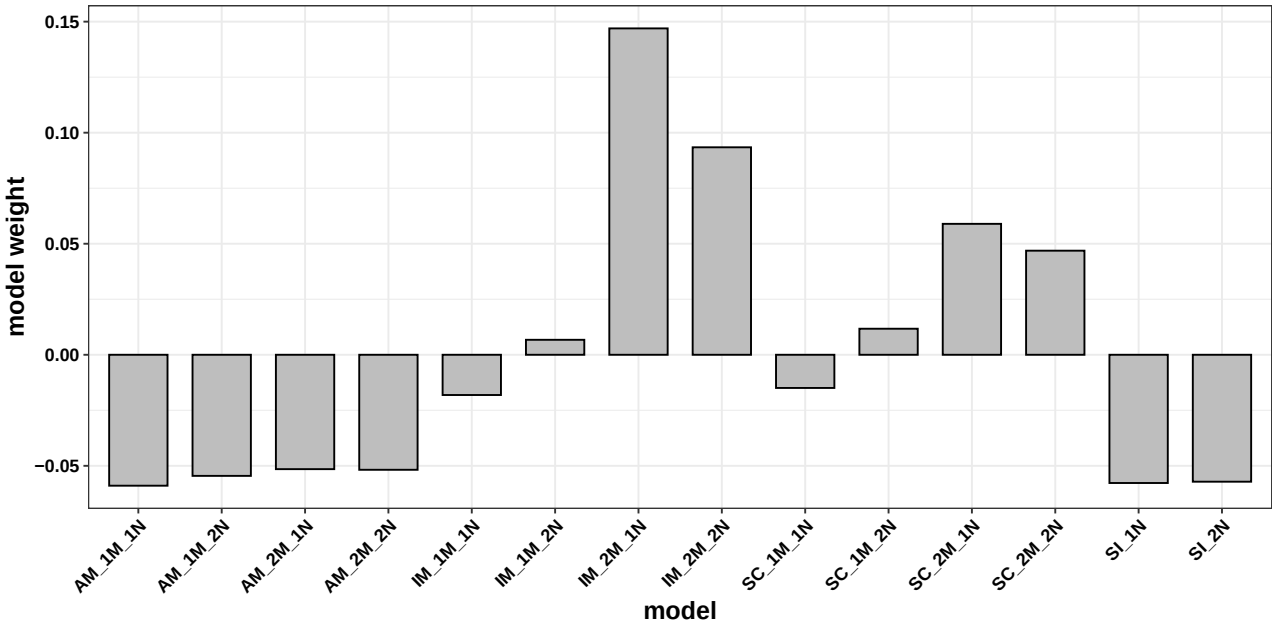

Figure 4.4: Distribution of model weight in maize. Model weight here are calculated as the difference between model weight estimated and an uniform distribution of model weigth.

associated with the domestication bottleneck (S. I. Wright et al. 2005). The point estimate for N_e in teosintes is 88,844 [7904; 144342] while that for maize is 52,648 [6642 ; 134544] (see Figure B.4). The estimated T_{split} is 32,902 [8073 ; 48582] generations (Figure B.4), which although two large for domestication given the archeological records (Piperno et al. 2009), is within the same order of magnitude.

Barrier detection

Posterior probability (post.prob) of barrier model quantify the probability of a loci fit the barrier model, knowing the error rate of the RF. Bayes factor is a rescaled post.prob, by taking into account the expected abundances provided through Q from the estimated demographic model. The distribution of posterior probability and Bayes factor are bimodal. The threshold of post.prob=0.5 (corresponding here to BF=30) divides the distribution into two modes, representing 98.3% and 1.7% of the genome, respectively (see Figure 4.5) for an estimated barrier proportion of Q=7% [0% ; 18%] in posterior (Figure B.4). The estimated false discovery rate by the RF is 12%. When visualizing the barrier signature (Figure B.2 BF>30) based on summary statistics using PCA, it does not appear to specifically correlate with one axis. However, the barrier loci signature appears to be homogeneous as they are grouped in the same place across the first three axes of the PCA (Fig B.2 BF>30). Loci identified as barriers using a BF>30 exhibit an increase in divergence, which is notably more pronounced for D_a than D_{xy}, as well as for F_{ST}. Loci also display strong decrease in ss and an increase in sf compared to the rest of the genome (Figure 4.6). Not all summary statistics contribute equally to barrier detection. The summary statistics that contribute the most to barrier detection, in decreasing order, are ss, F_{ST}, sXB, sxA, and D_a(see Figure 4.7). Among the barriers detected with the

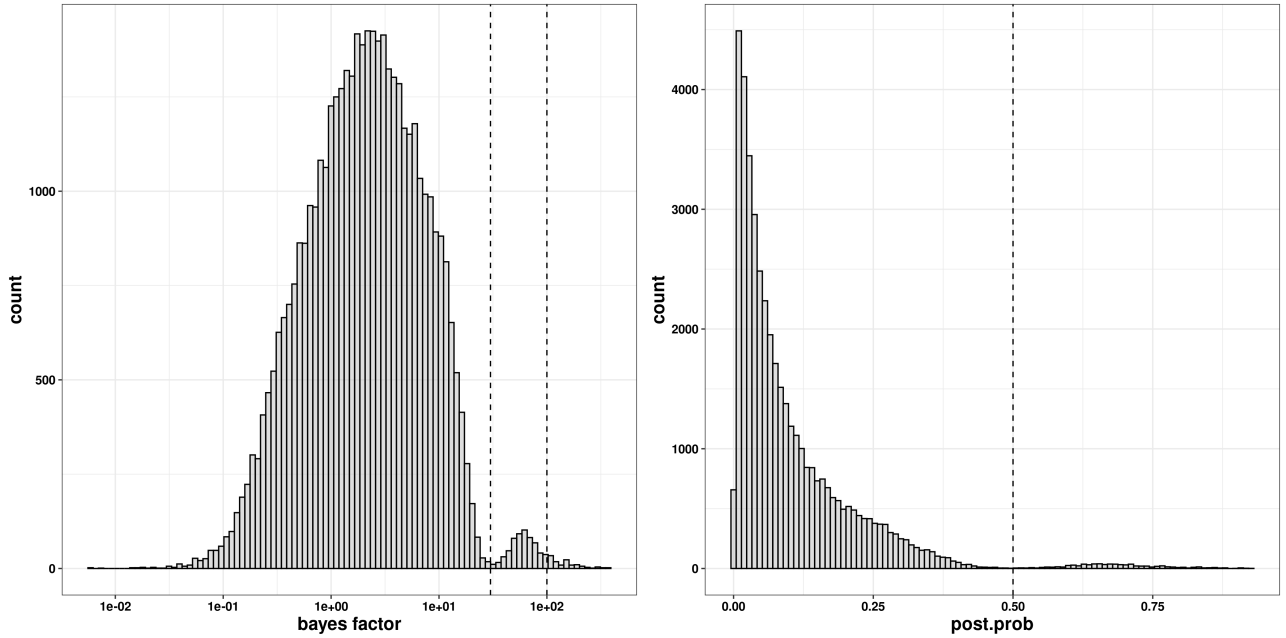

Figure 4.5: Distribution of Bayes factor (left) and barrier model posterior probability (post.prob)(right) for the maize dataset. For BF, BF=30 and BF=100 are represented by the dashed line, and for post.prob, post.prob=0.5 is also represented.

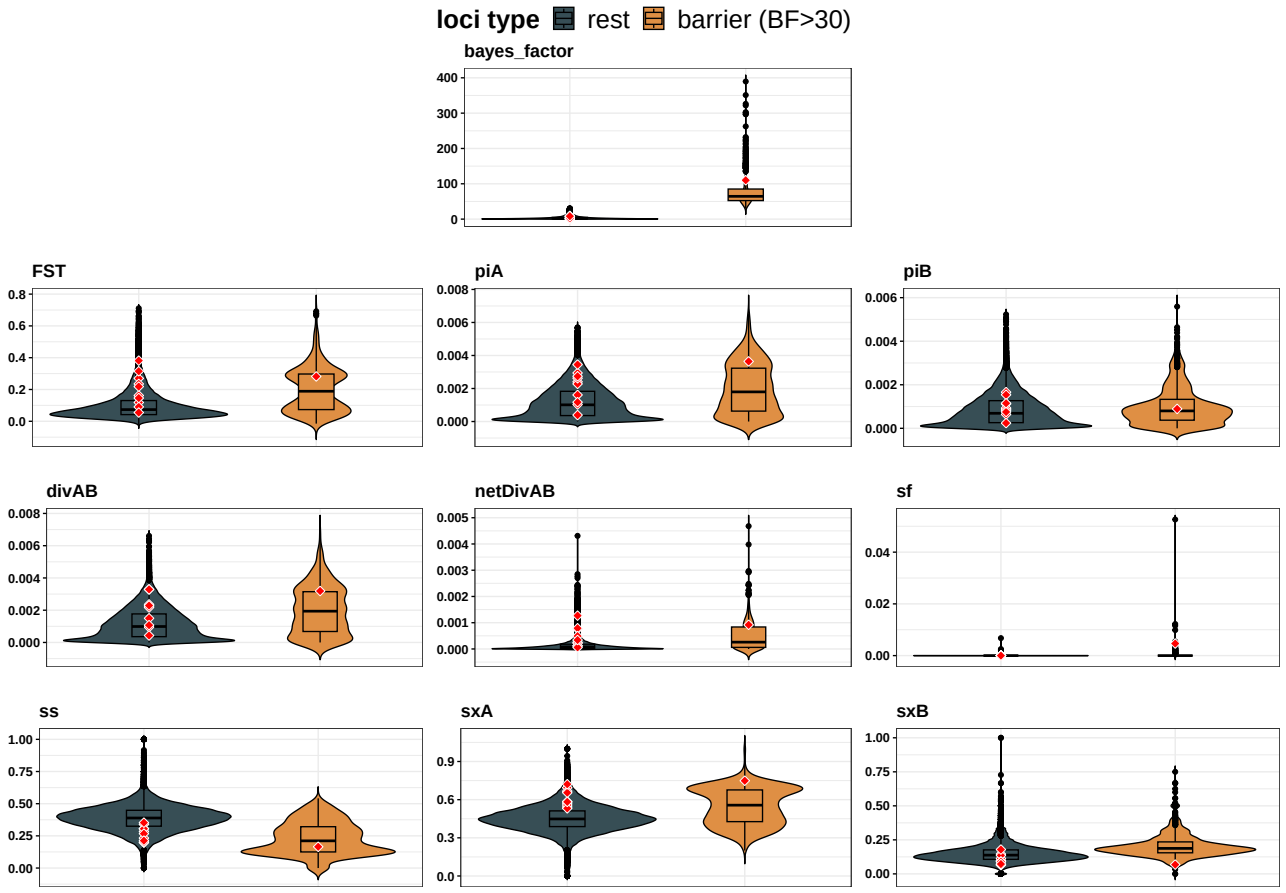

Figure 4.6: Distribution of summary statistics values for maize loci as a function of their type. Domestication genes from Table B.1 are represented by red diamonds. Barrier loci correspond to posterior probability > 0.5 and so a BF>30. A complete version of the plot is available in the annexe at Figure B.1.

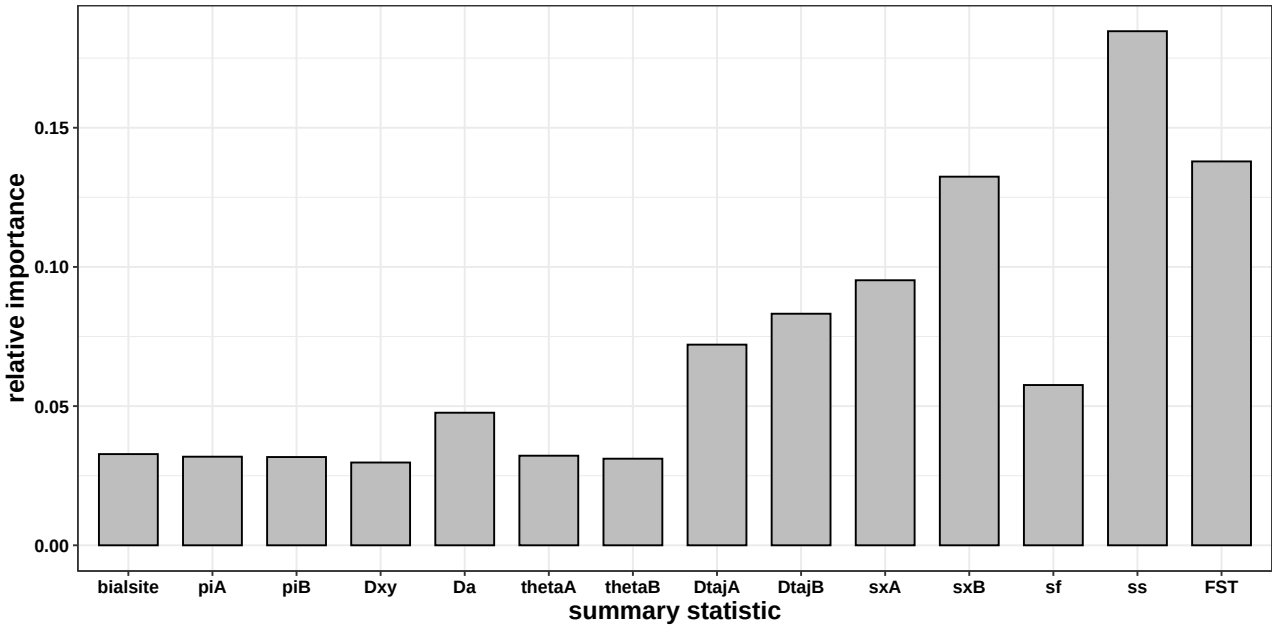

Figure 4.7: Relative contribution of each summary statistic to barrier detection

BF>30 threshold, Ga2 (BF=277) and a domestication gene named *Bt2* (BF=110) were identified (Figure 4.8). We also tested for the presence of gene barrier in flowering genes (listed in M. Tenaillon et al. (2019)), as flowering genes are potential candidates for RI through temporal isolation. We did not find any flowering gene exceeding BF=30 (the average value in flowering gene regions is around 4)

Are domestication genes detected as barriers to gene flow?

During the process of domestication, domestication loci are strongly selected, leaving a selective sweep footprint that can be detected. Our results show that a single locus, known in the literature as a domestication gene named *Bt2* in maize, is detected as a barrier (with a BF=110.03) among all loci known in maize (see Table B.1). Domestication genes exhibit a genomic pattern that shows an increase in differentiation, a specific reduction of diversity in maize population, and a slight increase in divergence. This pattern differs from barrier loci mainly on ss and sf (Figure 4.6), resulting in a lower Bayes factor (except *Bt2*), the average BF is 4.34 for domestication loci) than for barrier loci. Hence our results indicate that RIDGE is capable of differentiating barrier from domestication loci.

4.3 Foxtail millet

4.3.1 Domestication of the foxtail millet

Foxtail millet (*Setaria italica*) was first domesticated in China from *Setaria viridis* and then became a cultivated grain throughout Eurasia (Diao and Jia 2017). The oldest foxtail millet grains found to date were recovered from the Donghulin site in Beijing and date from 11,000 to

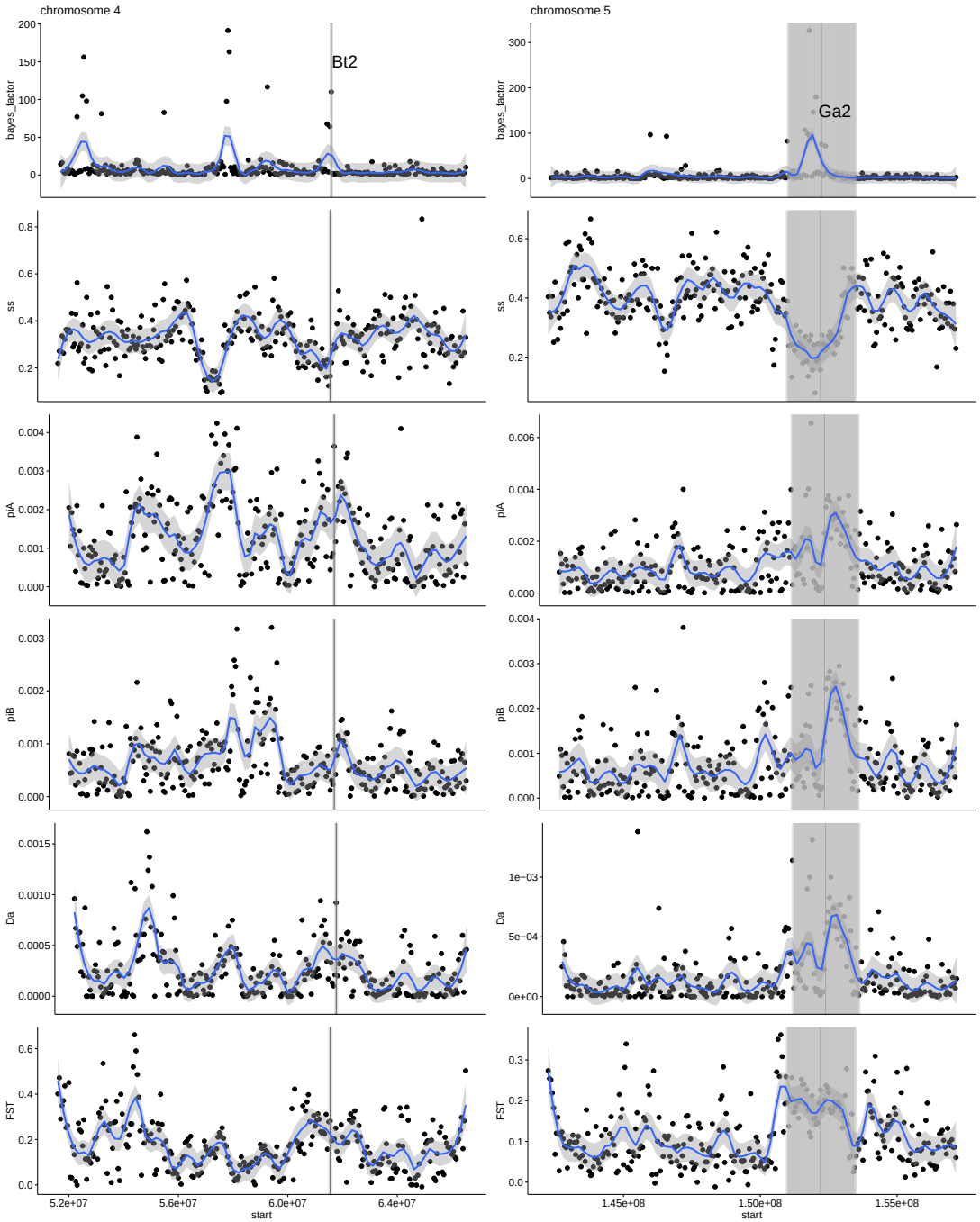

Figure 4.8: Genomic landscape of chromosome 4 around the *Bt2* locus and chromosome 5 around the *Ga2* locus (both marked with gray line) for BF, *ss*, π for teosinte (piA) and maize (piB), D_a (netDivAB) and F_{ST}. The blue line represents the loess regression with standard deviation in grey.

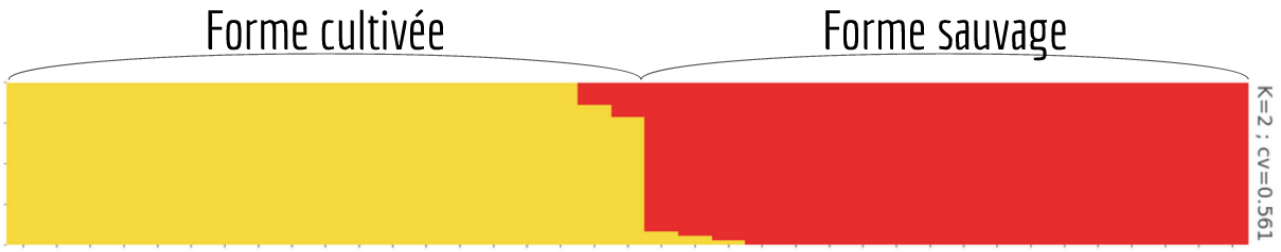

Figure 4.9: Genetic structure for 2 groups (K=2). This result has the lowest cross-validation (cv) (better clustering).

9,000 years ago (Jia et al. 2013). Grain has also been found at the Zhangmatun site in Shandong province, dating from 9,000 to 8,500 years ago (W. Wu et al. 2014). To this day, foxtail millet remains a major crop in the arid and semi-arid regions of China and India. At the molecular level, foxtail millet has experienced a loss of approximately 55% of its wild genetic diversity (wild $\theta = 0.0059$, domestic $\theta = 0.0027$) (Wang et al. 2010). This reduction in diversity is attributed, in part, to a population bottleneck with an intensity parameter $k = T_b/N_b = 0.6095$ (T_b the duration of the bottleneck and N_b the effective population size during the bottleneck) that occurred at the onset of the domestication event (Wang et al. 2010).

Its genome is small (490 Mb), making it an interesting genomic model among the Poaceae. The self-fertilization rate of *Setaria viridis*, has been estimated at 96% on the basis of SNP data and 90% on the basis of microsatellite data, compared with 98% for the domestic form based on microsatellite data (Jia et al. 2013; P. Huang et al. 2014).

As with other species, the introgression of domestic alleles into wild forms has contributed to the emergence of weeds with morphological similarities to foxtail millet. Hence the trait conferring great height in the domestic form was transferred to the wild form, which likely conferred a competitive advantage in natural environments, producing an invasive weed, the giant green foxtail (*S. viridis var. major*) frequently found around cultivated plots (Pohl 1951; Pohl 1966; Rominger 1962; Darmency et al. 1987a). Foxtail millet thus offers an example of a wild-weed-cultivated complex (Rao et al. 1987), where partial introgression between cultivated and wild forms may be advantageous for wild forms (Darmency et al. 1987a). In addition to the domestication traits common among the Poaceae, wild foxtail millet displays chloroplastic resistances to herbicides that do not originate from the domestic compartment. These resistances have been exploited in recent millet breeding. The atrazine-resistant F1 cross-platform hybrids tested by Darmency et al. (2006) show lower yields than the cultivated form. However, the yield loss can be offset by effective weed control in the presence of weeds. According to Darmency et al. (2017), this is the only known use of wild millet diversity for breeding cultivated millets. These results show that reproductive barriers do not prevent gene flow between wild and domesticated forms in millet, but do lead to hybrid depression.

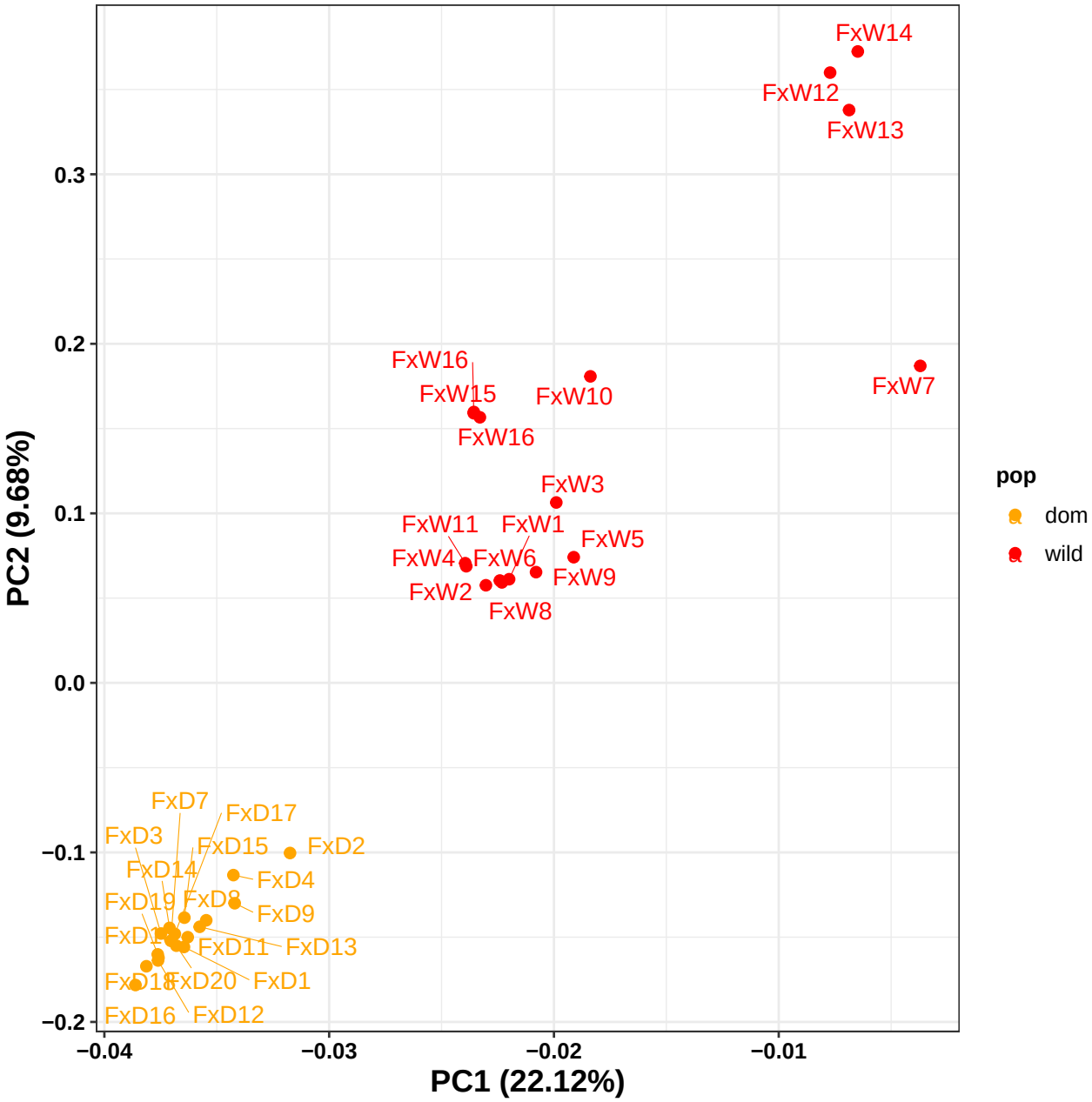

Figure 4.10: PCA performed on all filtered SNPs from foxtail millet dataset

4.3.2 Genomic material

A sample of 14 wild individuals (*Setaria viridis*) from 14 populations and 18 domestic individuals (*Setaria italica*) from 18 traditional varieties was taken. Sampling was designed to minimize intra-form genetic structure and distance from the center of domestication. The aim was to avoid confounding demographic effects. Just like for maize, by studying traditional varieties, we aimed at avoiding the counfounding effects of modern breeding.

Sequences were aligned to the foxtail millet reference sequence published in J. Wang et al. (2021) and treated using the method pipeline of Stella Hyung (see 4.1). The target sequencing depth for our data was 10X, and it ended up being between 5X and 15X, depending on the genotype. In results, from the initial 34 404 192 SNPs, 9 368 787 SNPs passed the filter (which represented 27.2% of the initially called SNPs). Principal component analysis (PCA) of the SNP set shows a clear differentiation between the wild and domestic form on axis 1 and 2 combined, which explains respectively 22.12% and 9.69%of the variation (Figure 4.10). Note that both axes rescaled population structure within our wild sample with FxW12, 13, 14 and 7 forming an independent group. Analysis using Admixture software (Alexander et al. 2009) resulted in an optimal categorization for two genetic groups (K=2). Individuals are grouped by wild or domestic form. As in maize, this result (Figure 4.9) is consistent with a structuring mainly explained by domestication.

4.3.3 Application of RIDGE on foxtail millet dataset

As for maize, there is evidence of repeated gene flow between *S. italica* and *S. viridis*, making the IM model the most likely. The primary distinction between maize and foxtail millet is their mating system. This may result in less contrast between barrier and non-barrier due to reduced gene flow, and more false positives as selfing breaks numerous assumptions made by coalescent simulation, such as panmictic reproduction and absence of linkage disequilibrium between loci. We established prior bounds based on real data and literature information. N_e [4,250; 152,000], M [0.1; 50], T_{split} [1,000; 80,000], Pbarrier_max=0.2, and Nref = 63,500, with a window size of 50kb. We ran RIDGE while taking into account heterogeneity (with 'hetero θ' set to True) and recombination (with 'homo_rec' set to False). The foxtail recombination map used in this study was obtained from Brazier and Glémin (2022).

Demographic infereneces

As for maize, the goodness of fit increased between prior and posterior ($G_{prior} = 0.147$;$G_{post} = 0.481$), and the posterior distribution of the summary statistic fits the observed data set better than the prior (Figure B.8). The dominant models receiving most weight are the IM and SC models with heterogeneous migration rates and the N_e (IM_2M_2N: 23% and SC_2M_2N: 15%)(see Figure 4.11), which together represent 38% of model weight. Furthermore, the IM and SC models represent 90% of the model weight, advocating for the presence of ongoing migration. This is in agrement with studies demostrating introgression and partial barriers

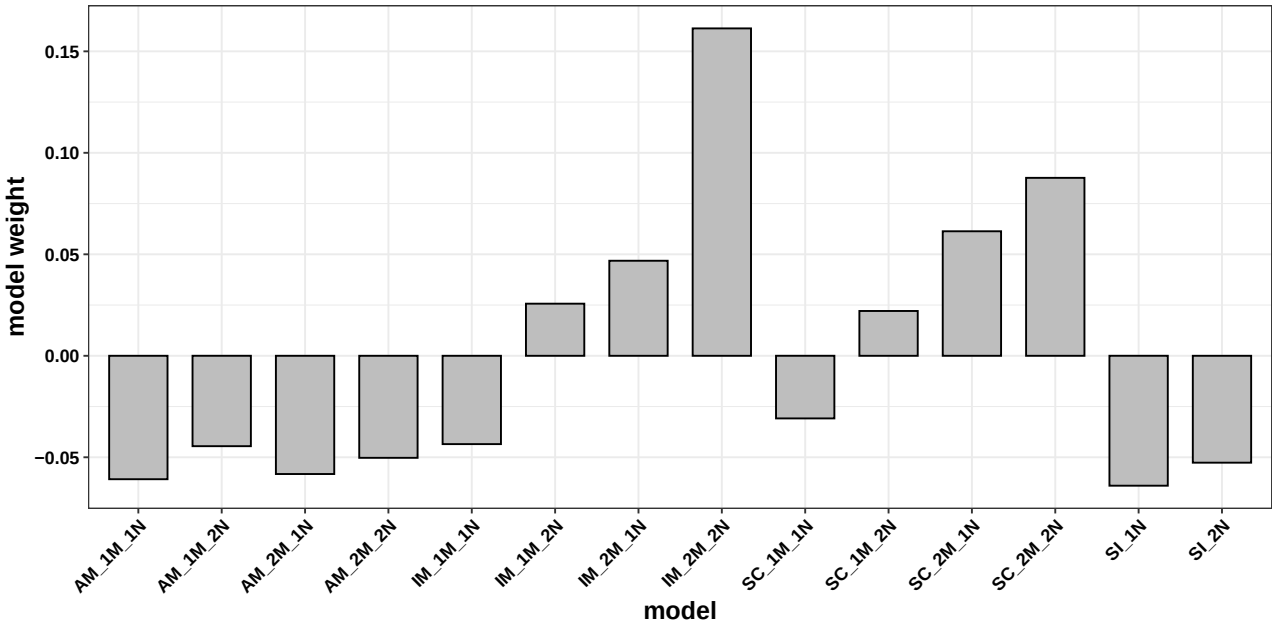

Figure 4.11: Distribution of model weight in foxtail millet. Model weight here are calculated as the difference between model weight estimated and an uniform distribution of model weigth.

between wild and cultivated forms (Pohl 1951; Pohl 1966; Rominger 1962; Darmency et al. 1987a).Based on the parameter, the estimated loss of diversity is 26%. The effective population size is estimated to be 70,709 individuals for the wild gene pool and 52,050 individuals for the domesticated gene pool. These values are lower than the observed loss of 55% in Wang et al. (2010) (see Figure B.5).

Barrier detection

As for maize, the posterior distribution and thus the BF distribution follows a bimodal distribution, which is split in half at post.prob=0.5, corresponding to BF=16 (Fig 4.12).The estimated false discovery rate by RF is 7.2%, which is of the same order as for maize (12%). In contrast to maize, foxtail millet has a higher percentage of its genome classified as a barrier (15.5% compared to 1.7% of maize using post.prob=0.5 threshold), for an estimated barrier proportion of 8% in posteriors (Fig 4.12). When using a common BF threshold of 30, which is the threshold used for maize, the barrier proportion decreased to 10%. In terms of megabases (Mb), both maize and foxtail millet have a similar amount of genome detected as a barrier when using a BF threshold of 30 (44.2Mb for maize and 49Mb for foxtail millet). To visualize barriers, we conducted a PCA analysis and found that at BF>16, two groups are observable (Figure B.6). This suggests that the barrier detected (BF>16) may have two different genomic signatures. Notably, at BF>30, the barrier forms one homogeneous group. Such loci exhibit an increase in divergence (more pronounced for D_a than D_{xy}) and in differentiation summary statistics. They also show a strong decrease in ss and an increase in sf compared to the rest of the genome (Figure 4.14). Additionally, the distribution of summary statistic importance is similar to that of maize, except that sf plays a more significant role in barrier detection than in maize (Figure

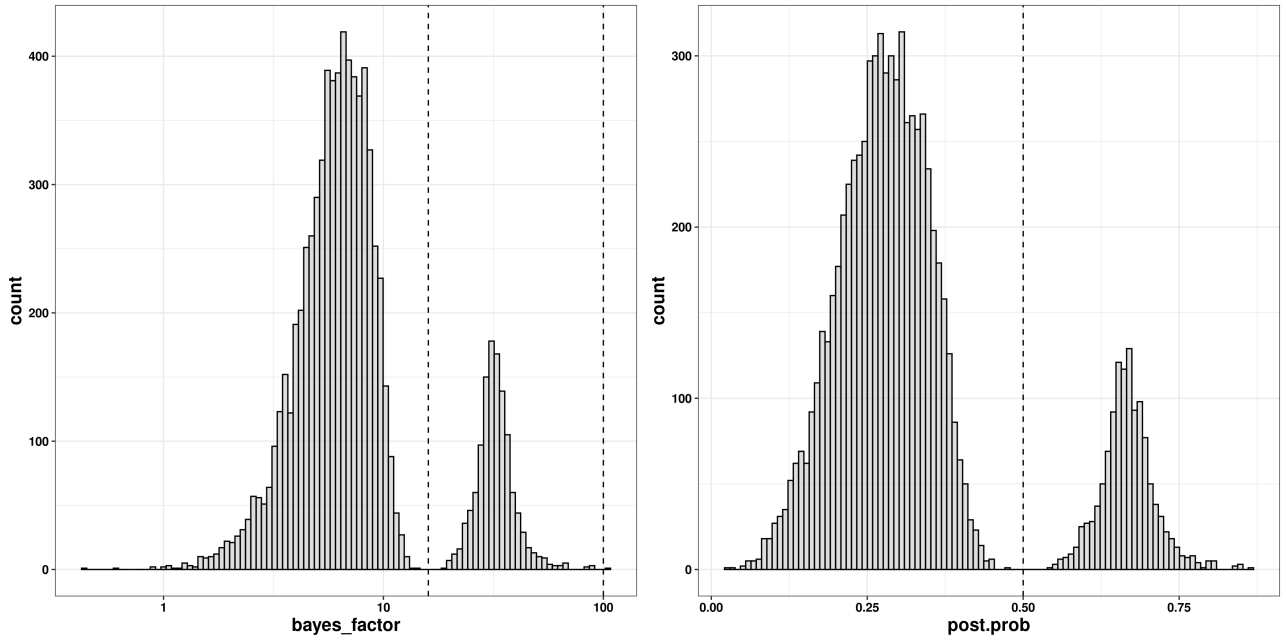

Figure 4.12: Distribution of Bayes factor (left) and barrier model posterior probability (post.prob)(right) for the foxtail millet dataset. For BF, BF=16 and BF=100 are represented by the dashed line, and for post.prob, post.prob=0.5 is also represented.

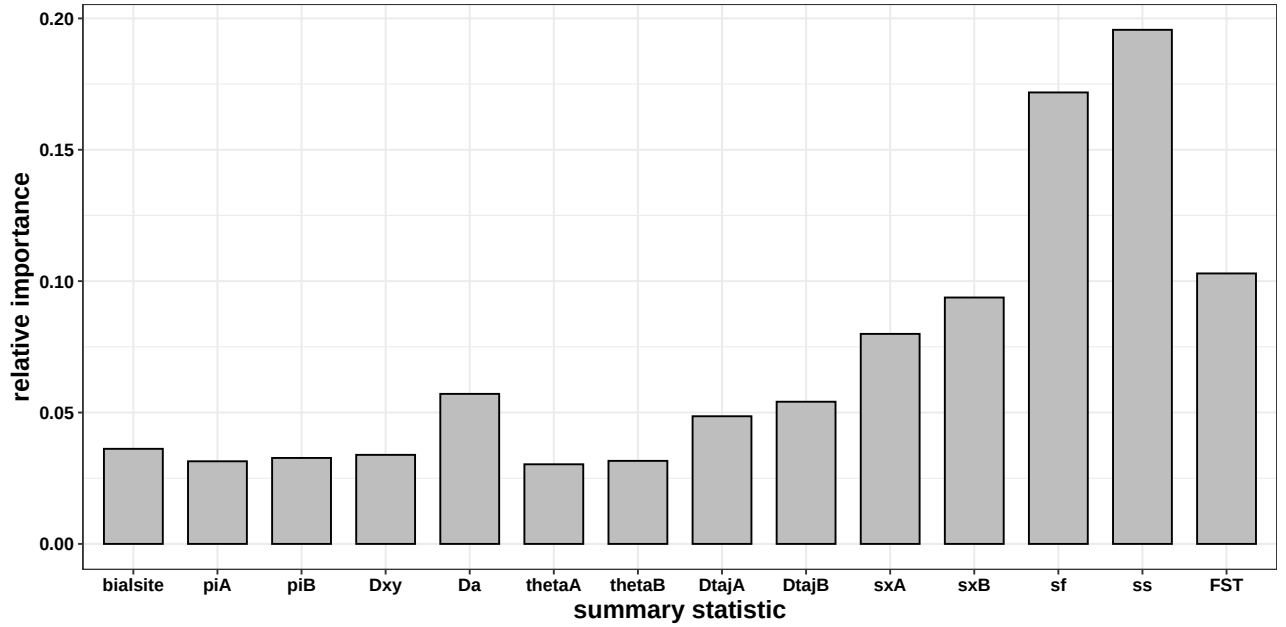

Figure 4.13: relative importance associated with each statistics during the random forest building for the foxtail millet dataset.

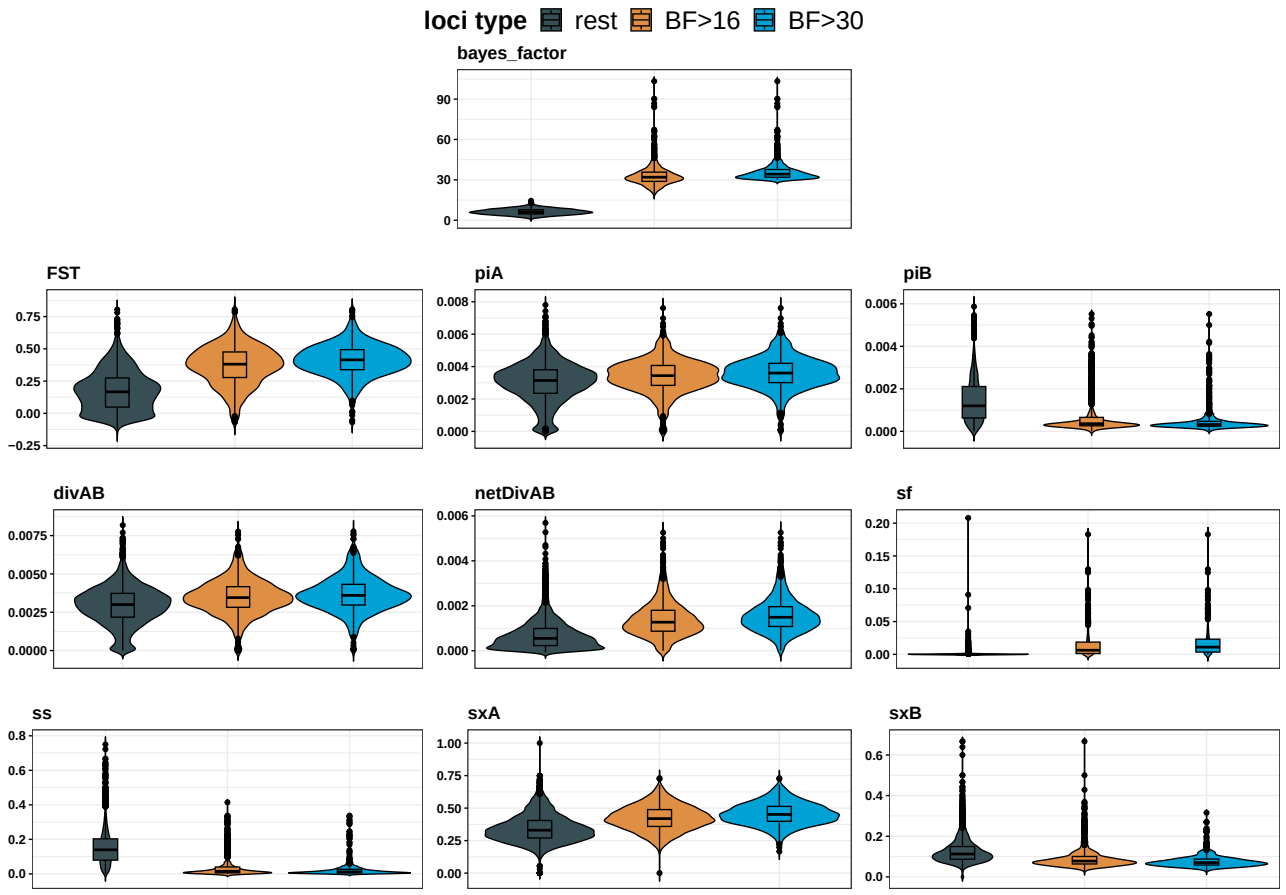

Figure 4.14: Distribution of summary statistics values for foxtail millet loci in function of their BF level. Complete figure is showed at Fig B.7.

4.13 & 4.7).

Effect of mating system on RIDGE results

As previously stated, comparing foxtail millet and maize provides an opportunity to observe how the mating system can affect results. Both species diverged from their respective wild relatives 9000 years ago and belong to the Poaceae family. We observe a higher proportion of the genome involved in reproductive isolation in foxtail millet than in maize, but it represents the same amount of megabases. The genomic signature of the barrier is nearly the same, except for sf, which was more involved in barrier detection in foxtail millet. The stronger contribution of sf may be due to reduced gene flow and/or N_e resulting from selfing, which induces more fixed differences (sf). As fixed differences are more abundant, sf is more sensitive for barrier detection. In maize, the average sf is 0, while in foxtail millet, sf was 0.002 (see Fig 4.15F). Similarly, we observed a higher mean of D_a, D_xy, and F_{ST}. Finally, the landscape in terms of Bayes factor and posterior probability is flatter in foxtail millet than in maize (see Figure 4.15A&B). This indicates that in foxtail millet, barriers are more challenging to differentiate from non-barriers than in maize. Interestingly, the landscape of F_{ST} and D_a around the barrier is more elevated from the sea level for foxtail millet than for maize.

4.4 Discussion

Both datasets are primarily genetically structured by the differentiation between wild and domestic gene pools. However, it appears that the wild gene poll in foxtail millet still exhibits genetic structure despite efforts to avoid it during genetic sampling. In future work, re-analyzing foxtail millet dataset after removing the outliers hrom the wild gene pool should be done to test if our current observations are affected by wild group population structure.

The dataset, for maize and foxtail millet, because of their recent divergence from their wild relatives, presents more challenging conditions than the crow dataset. Barrier detection using a BF=30 threshold showed nearly identical genomic patterns for barriers in both cases, with the same amount of Mb detected as a barrier (44.2Mb for maize and 49Mb for foxtail millet), despite the maize genome being five times larger than the foxtail millet genome.

Interestingly, all loci known to be barriers between maize and teosinte are from the maize-mexicana interaction. Maize is known to hybridize easily with parviglumis, which may result in fewer RI loci. Our wild population consists solely of parviglumis individuals. Nevertheless, we observe a large region of RI at the position of *Ga2* (Figure 4.8), indicating that even between maize and parviglumis, *Ga2* acts as a barrier. As for the other gene, it exhibits a BF level of around 10 for *Ga1/Tcb1*. Furthermore, one domestication gene, *Bt2*, has been detected as a barrier (with a BF=110), showing that domestication can play a role in reproductive isolation, even though it is an exception among the 11 well-described domestication genes in maize that express a very low BF (BF 4). Domestication is a form of strong local adaptation that could indirectly lead to reproductive isolation (RI). Our results suggest that, with the exception of

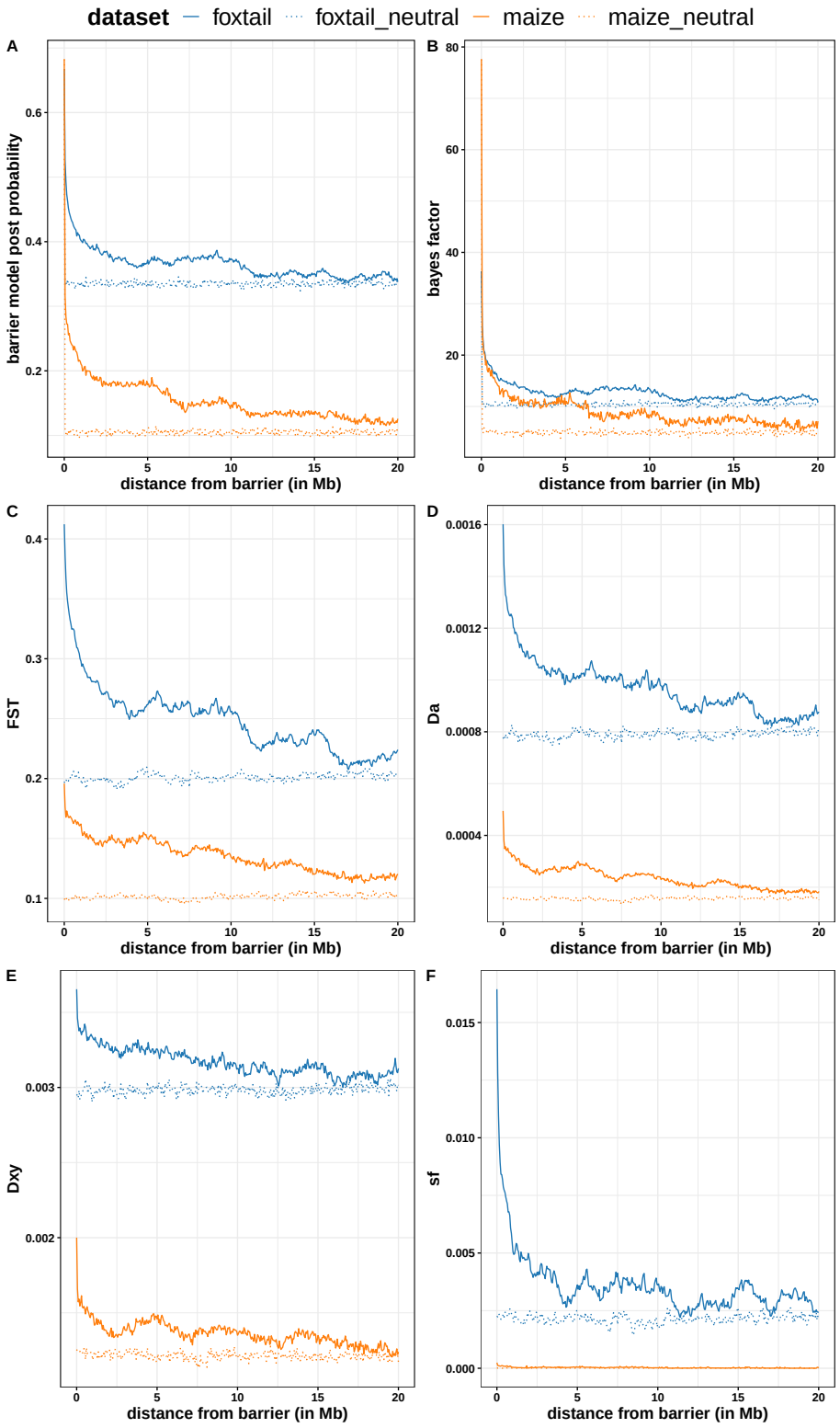

Figure 4.15: Landscape of barrier model posterior probability (A) and bayes factor (B) F_{ST} (C), D_a (D), D_{xy} (E) and D_f or s_f (F) around loci detected as barrier (BF>30) for maize and foxtail millet dataset compared to a randomized version of each dataset (called neutral). To generate a neutral dataset, values of posterior probability are resampled without replacement.

one gene, domestication genes are not involved in RI. We also tested for the presence of gene barriers in flowering genes as they are potential candidates for reproductive isolation through temporal isolation. Our results did not find any flowering genes exceeding BF=30 (the average value in flowering gene regions is the same as the average mean of the genome).

Overall, selfing appears to affect the results when examining demographic inferences and genomic patterns of barriers. For instance, barriers detected in foxtail millet with a Bayes Factor of 16 represent 15%, which is significantly higher than in maize. Additionally, these barriers are separated into two groups using PCA (Figure B.6). This result could potentially be explained by lower gene flow and higher linkage disequilibrium in selfing generating more gene flow barrier, or by the fact that selfing breaks many assumptions made during simulations, consequently leading to potential false positives. The contrast between island and sea level in Bayes Factor is lower for foxtail millet than for maize (Figure 4.15). This could be due to the global reduction in gene flow induced by selfing coupled with a reduction in effective recombination. However, if this were the case,the BF distribution should not show a clear bimodality and the BF landscape should be flatter. Alternatively, the lower contrast could be explained by a lack of statistical power. The results are more difficult to interpret than those for maize due to the effect of selfing on both biological and statistical aspects.

Chapter 5

General Conclusion & Perspectives

The primary objective of this thesis was to develop a method for detecting barriers to gene flow that could be applied across diverse biological systems. I built on an existing tool, DILS (Fraïsse and al 2021), and modified and extended it to enable comparative analysis across multiple contexts. Firstly, I improved the method by implementing a model averaging instead of a best-model approach. This modification enabled RIDGE to accurately estimate demographic models across a broad range of conditions, both in simulated and empirical datasets, making comparisons possible among datasets even when the best models differ. Secondly, I refined the estimation of parameters related to the proportion of barriers to gene flow within genomes, through the introduction of new metrics of outliers of divergence, differentiation, and diversity. In particular, it improved estimation under challenging conditions such as low divergence time and/or low migration rates.

Finally, RIDGE is able to detect barrier on simulated dataset even at very low time of divergence ($T_{split} < 0.2N_e$), but also in real dataset as seen for crows (for recent divergence time) and maize (very recent) for which we successfully detected well identified barrier loci from literature (*RSG9*, *LRP5*, *PRKCA* and *CACNG1&4* for crows and *Ga2* for maize). As observed in the case of foxtail millet, selfing mating system seems to reduces RIDGE's ability to distinguish barriers.

5.1 What have we learnt with RIDGE and where should we go?

In all tests conducted with RIDGE, it appears that ss, F_{ST}, and D_a are the three main contributors to barrier detection. However, the contribution of each varies depending on demographic history and mating system. Additionally, sf may also be involved (as in the case of foxtail millet) for the same reasons. Interestingly, across all empirical datasets tested, net divergence (D_a) emerged as a more informative metric than D_{xy}. Unlike the expected pattern of increasing D_{xy} with decreasing gene flow (as shown in Figure 1.8), D_{xy} followed diversity trends at low divergence times, making D_{xy} unable to discriminate barriers from the rest. In contrast, D_a accurately measured divergence. This observation contrasts with the point of view of T.

Cruickshank and M. Hahn (2014), who strongly advocate for the use of D_{xy} rather than F_{ST}. At lower levels of divergence, such as those observed in crows, maize, and foxtail millet, D_{xy} is primarily influenced by the internal diversity of the population. However, at higher T_{split} values, this is not the case and D_{xy} should contribute more to barrier detection for more anciently diverged systems. Therefore, there is probably no universal pattern for gene flow barriers as it depends on multiple factors, which make crucial the use of multiple summary statistics.

The genomic pattern used to detect barriers may be influenced by some simulation choices, as illustrated with the heterogeneity in migration. The migrant rate ($M = 4 * N_e * m$) can be affected by local reductions in diversity (N_e) and in gene flow (m_e). Modeling heterogeneity of migration through M rather than m (as in Fraïsse and al (2021)), resulted in false positive detection, as some loci with a strong reduction in N_e (which induces a reduction in M) were also detected as barrier (refer to chapter 2). Another questionable aspect of the simulations, which is common to DILS and RIDGE is the method of simulating linked selection at the locus level (hetero-N models). For a given locus, three independent values are sampled from the same beta distribution (albeit with distinct mean) so that there is no covariation in effective population size across populations. Alternatively, N_a values could be sampled and then rescale to give values for N_1 and N_2 using N_a/N_1 and N_a/N_2 ratio ensuring complete correlation between population size at a given loci. The reality is probably somewhere in between these two solutions. Selective sweeps may occur independently across the genome in different populations making local N_e uncorrelated but for regions with low recombination, local population sizes may be highly correlated. More realistic models of linked selection could be investigated in the future.

The sliding window approach with fixed size is also questionable when it comes to linked selection. Since loci are simulated completely independently, it assumes that there is no linked selection between windows. This assumption may be more challenging in autogamous systems, where the extent of linked LD is much larger due to inefficient recombination (Burgarella and Glémin 2017). One way to deal with the sliding window size problem is to allow the window size to vary across the genome, without using an arbitrary value. For example, one could segment the genome based on the local phylogeny (Zamani et al. 2013). This will generate windows with a homogeneous signal, capturing the local signal and thus avoiding the problem of varying LD across the genome. Also, in the early stages of speciation, the signal may be concentrated in small regions, potentially smaller than the window size. Conversely, in later stages, RI is expected to generate larger patterns, larger than the window size. In both cases, the RI signal is affected by the window size, and a method such as the one suggested above could improve barrier detection during the early and late stages of speciation.

5.2 Perspectives of RIDGE usages for speciation research

RIDGE successfully detects intrinsic reproductive isolation without distinguishing between pre- and postzygotic barriers, at low and high time of divergence. Its ability to produce comparable

results across different contexts, at low as at higher divergence degree, paves the way for gathering new insights for various questions, as discussed below.

5.2.1 What are the genomic patterns of reproductive isolation during speciation?

During the early stages of divergence, reproductive isolation may be limited to a small number of genes (Wu 2001). As reproductive isolation becomes complete, it is expected to progress from a genetic mosaic pattern to genome-wide divergence (Feder et al. 2012; Wu 2001). The recombination landscape around loci responsible for reproductive isolation, particularly for recombination suppressors such as inversions, has been extensively described. Genes that contribute to pre- and postzygotic isolation tend to map to inversions that distinguish species of sunflowers (Todesco et al. 2020), stickleback fish (Bay et al. 2017), and *Heliconius* (Merrill et al. 2019). To detect reproductive isolation in a comparative context, RIDGE could be used on multiple species pairs encompassing a large spectrum of divergence to address the following question: is there a correlation between local barrier richness and recombination? Are regions of inversion enriched in gene flow barrier? Also, does barrier region size increase with divergence and linkage disequilibrium as presented in Wu (2001) and Feder et al. (2012)? Using the same comparative approach we could measure the evolution of the barrier proportion with divergence, and test if the barrier proportion correlates with RI (measured for example through crossing experiment and hybrid phenotyping).

5.2.2 Testing the snowball theory

In 1995, Orr published a paper aimed at describing the evolution of hybrid incompatibilities over time. The study demonstrated that when using BDM incompatibilities, the number of incompatibilities grows faster than linearly with time. Specifically, Orr and Turelli (2001) demonstrated that it grows as the square of the divergence time. Several attempts have been made to detect the snowball effect by measuring the evolution of RI between lineages at different times of divergence. However, these attempts have failed to find evidence of the snowball effect and instead found a linear increase (Presgraves 2002; Stelkens et al. 2010; Price and Bouvier 2002). Nevertheless, testing for the snowball effect requires information on the number of BDMIs contributing to reproductive isolation, rather than the effect on RI. Moyle and Nakazato (2010) and Matute et al. (2010) successfully demonstrated a snowball effect in *Solanum* and *Drosophila* species, respectively, using this approach. However, this method is expensive as it requires producing numerous hybrids between multiple lineages. Consequently, testing the snowball theory on a broader range of species is extremely time-consuming and costly. In this context, RIDGE may offer a way to test the theory using only genomic data. This way we could count the number of loci involved in BDMI and see if they accumulate faster than the linear rate. However, the limitation of this approach is that RIDGE is unable to distinguish between loci involved in hybrid incompatibility among loci considered as gene flow barriers.

5.2.3 What is the nature of speciation genes?

Speciation genes are genes that actively contribute to RI. They can be associated with any form of pre- or post-zygotic isolation. One of the aims of speciation research is to understand the nature of speciation genes or speciation gene networks. For example, do all bird species that rely on feather color patterns for mate choices have genes involved in feather color pattern and color recognition, like *RSG9* and *LRP5* in crows (Poelstra et al. 2014; Vijay et al. 2016)? Do plants that rely on pollination to reproduce speciate through genes that affect flower and flowering time, as seen in *Mimulus aurantiacus* with *MaMyb2* (Streisfeld et al. 2013)? To answer this question, functional analysis is mandatory to identify the degree of contribution of a gene to reproductive isolation. However, RIDGE offers a way to detect candidate regions that act as gene flow barriers, with a quantification of their probability to be a barrier, allowing further functional study to focus only on a subset of genes, reducing the cost and the time needed for this type of study.

Bibliography

Aguirre-Liguori, Brandon Gaut, Juan Pablo Jaramillo-Correa, et al. (2019a). "Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (Zea mays parviglumis and Zea mays mexicana)." In: *Molecular ecology* 28.11, pp. 2814–2830.

Aguirre-Liguori, Santiago Ramirez-Barahona, Peter Tiffin, et al. (2019b). "Climate change is predicted to disrupt patterns of local adaptation in wild and cultivated maize." In: *Proceedings of the Royal Society B* 286.1906, p. 20190486.

Alexander, John Novembre, and Kenneth Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals." In: *Genome research* 19.9, pp. 1655–1664.

Baltazar et al. (2005). "Pollination between maize and teosinte: an important determinant of gene flow in Mexico." In: *Theoretical and Applied Genetics* 110, pp. 519–526.

Bank, Claudia, Reinhard Bürger, and Joachim Hermisson (July 2012). "The Limits to Parapatric Speciation: Dobzhansky–Muller Incompatibilities in a Continent–Island Model." In: *Genetics* 191.3, pp. 845–863.

Bapat, Amruta R et al. (2023). "The Ga1 locus of the genus Zea is associated with novel genome structures derived from multiple, independent nonhomologous recombination events." In: *G3: Genes, Genomes, Genetics* 13.11, jkad196.

Barton, Nick and Bengtsson (Dec. 1986). "The barrier to genetic exchange between hybridising populations." In: *Heredity* 57.3. Number: 3 Publisher: Nature Publishing Group, pp. 357–376.

Baumdicker, Franz et al. (2022). "Efficient ancestry and mutation simulation with msprime 1.0." In: *Genetics* 220.3, iyab229.

Bay, Rachael A. et al. (Nov. 2017). "Genetic Coupling of Female Mate Choice with Polygenic Ecological Divergence Facilitates Stickleback Speciation." In: *Current Biology* 27.21, 3344–3349.e4.

Beaumont (Dec. 1, 2010). "Approximate Bayesian Computation in Evolution and Ecology." In: *Annual Review of Ecology, Evolution, and Systematics* 41.1, pp. 379–406.

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

Beissinger et al. (June 13, 2016). "Recent demography drives changes in linked selection across the maize genome." In: *Nature Plants* 2.7, pp. 1–7.

Bellman, R. and R. Kalaba (Nov. 1959). "On adaptive control processes." In: *IRE Transactions on Automatic Control* 4.2, pp. 1–9.

Bellucci, Elisa et al. (2014). "Decreased nucleotide and expression diversity and modified co-expression patterns characterize domestication in the common bean." In: *Plant Cell* 26, pp. 1901–1912.

Bengtsson (1985). *The flow of genes through a genetic barrier in Evolution Essays in Honor of John Maynard Smith, eds Greenwood JJ, Harvey PH, Slatkin M., editors.*

Berube, Benjamin et al. (July 13, 2023). *Teosinte Pollen Drive guides maize domestication and evolution by RNAi.* Pages: 2023.07.12.548689 Section: New Results.

Bhatia, Gaurav et al. (Jan. 9, 2013). "Estimating and interpreting FST: The impact of rare variants." In: *Genome Research* 23.9. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1514–1521.

Bomblies, Kirsten et al. (Sept. 4, 2007). "Autoimmune Response as a Mechanism for a Dobzhansky-Muller-Type Incompatibility Syndrome in Plants." In: *PLoS Biology* 5.9. Ed. by James C Carrington, e236.

Brazier, Thomas and Sylvain Glémin (Aug. 30, 2022). "Diversity and determinants of recombination landscapes in flowering plants." In: *PLOS Genetics* 18.8, e1010141.

Burban, Ewen, Maud I Tenaillon, and Arnaud Le Rouzic (2022). "Gene network simulations provide testable predictions for the molecular domestication syndrome." In: *Genetics* 220.2, iyab214.

Burban, Ewen, Maud Irene Tenaillon, and Sylvain Glemin (Sept. 17, 2023). *RIDGE, a tool tailored to detect gene flow barriers across species pairs.* Pages: 2023.09.16.558049 Section: New Results.

Burgarella, Concetta and Sylvain Glémin (2017). "Population genetics and genome evolution of selfing species." In: *John Wiley & Sons, Ltd. doi* 10.9780470015902, a0026804.

Burton, Ronald S, Ricardo J Pereira, and Felipe S Barreto (2013). "Cytonuclear genomic interactions and hybrid breakdown." In: *Annual Review of Ecology, Evolution, and Systematics.* Vol. 44, pp. 281–302.

Calfee et al. (2021). "Selective sorting of ancestral introgression in maize and teosinte along an elevational cline." In: *PLOS Genetics* 17, e1009810.

Charlesworth (July 1992). "Evolutionary Rates in Partially Self-Fertilizing Species." In: *The American Naturalist* 140.1, pp. 126–148.

Charlesworth, B (May 1, 1998). "Measures of divergence between populations and the effect of forces that reduce variability." In: *Molecular Biology and Evolution* 15.5, pp. 538–543.

Charlesworth, B, M T Morgan, and D Charlesworth (Aug. 1, 1993). "The effect of deleterious mutations on neutral molecular variation." In: *Genetics* 134.4, pp. 1289–1303.

Charlesworth and Jensen (2021). "Effects of Selection at Linked Sites on Patterns of Genetic Variability." In: *Annual Review of Ecology, Evolution, and Systematics* 52.1, pp. 177–197.

Chen et al. (2022). "A pair of non-Mendelian genes at the Ga2 locus confer unilateral cross-incompatibility in maize." In: *Nature Communications* 13.1.

Coyne and Orr (1989). "Patterns of Speciation in Drosophila." In: *Evolution* 43.2, pp. 362–381.

Cruickshank and Hahn (July 2014). "Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow." In: *Molecular Ecology* 23.13, pp. 3133–3157.

Cruickshank, Travis and Matthew Hahn (2014). "Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow." In: *Molecular ecology* 23.13, pp. 3133–3157.

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

Csilléry et al. (July 2010). "Approximate Bayesian Computation (ABC) in practice." In: *Trends in Ecology & Evolution* 25.7, pp. 410–418.

Cutter, Asher D. (Apr. 1, 2012). "The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities." In: *Trends in Ecology & Evolution* 27.4, pp. 209–218.

Danecek, Petr et al. (2021). "Twelve years of SAMtools and BCFtools." In: *Gigascience* 10.2, giab008.

Darmency, Christian Ouin, and Jacqueline Pernes (1987a). "Breeding foxtail millet (Setaria italica) for quantitative traits after interspecific hybridization and polyploidization." In: *Genome* 29, pp. 453–456.

Darmency, XiaoXia Tian, and Christophe Delye (2006). "Molecular evidence of biased inheritance of trifluralin herbicide resistance in foxtail millet." In: *Journal of Heredity* 97.3, pp. 254–258.

Darmency, Tao Wang, and Christophe Delye (2017). "Herbicide Resistance in Setaria." In: *Genetics and Genomics of Setaria*. Springer, pp. 283–296.

Darwin, Charles (1859). "On the origins of species by means of natural selection." In: *London: Murray* 247, p. 1859.

De Queiroz, Kevin (Dec. 1, 2007). "Species Concepts and Species Delimitation." In: *Systematic Biology* 56.6, pp. 879–886.

Delaneau, Olivier, Cédric Coulonges, and Jean-François Zagury (Dec. 16, 2008). "Shape-IT: new rapid and accurate algorithm for haplotype inference." In: *BMC bioinformatics* 9, p. 540.

Delmore, Kira E. et al. (Apr. 1, 2018). "Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds." In: *Evolution Letters* 2.2, pp. 76–87.

Dempewolf, Hannes et al. (2012). "Reproductive isolation during domestication." In: *Plant Cell* 24, pp. 2710–2717.

Diao, Xianmin and Guanqing Jia (2017). "Origins and domestication of foxtail millet." In: *Genetics and Genomics of Setaria*. Springer, pp. 61–72.

Diez, C. M., B. S. Gaut, E. Meca, et al. (2013). "Genome size variation in wild and cultivated maize along altitudinal gradients." In: *New Phytol* 199, pp. 264–276.

Dobzhansky (July 1937). "Genetic Nature of Species Differences." In: *The American Naturalist* 71.735, pp. 404–420.

Doebley, J. (1992). "Mapping the genes that made maize." In: *Trends Genet* 8.9, pp. 302–307.

Dormann, Carsten F. et al. (Nov. 2018). "Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference." In: *Ecological Monographs* 88.4, pp. 485–504.

Ellstrand, Norman C et al. (2013). "Introgression of crop alleles into wild or weedy populations." In: *Annual Review of Ecology, Evolution, and Systematics*. Vol. 44, pp. 325–345.

Evans and Kermicle (Aug. 1, 2001). "Teosinte crossing barrier1, a locus governing hybridization of teosinte with maize." In: *Theoretical and Applied Genetics* 103.2, pp. 259–265.

Evans, M. M. S. and J. L. Kermicle (2001). "Teosinte crossing barrier1, a locus governing hybridization of teosinte with maize." In: *Theoretical and Applied Genetics* 103.2, pp. 259–265.

Feder, Jeffrey L. et al. (Feb. 5, 2012). "Establishment of new mutations under divergence and genome hitchhiking." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587, pp. 461–474.

Flaxman, Samuel M, Jeffrey L Feder, and Patrik Nosil (2013). "Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow." In: *Evolution* 67.9, pp. 2577–2591.

Florez-Rueda, Ana M. et al. (Nov. 1, 2016). "Genomic Imprinting in the Endosperm Is Systematically Perturbed in Abortive Hybrid Tomato Seeds." In: *Molecular Biology and Evolution* 33.11, pp. 2935–2946.

Fraisse, Christelle et al. (Jan. 15, 2021). "DILS: Demographic Inferences with Linked Selection by using ABC." In: *Molecular Ecology Resources*, pp. 1755–0998.13323.

Fraïsse and et al (2021). "DILS: Demographic inferences with linked selection by using ABC." In: *Molecular Ecology Resources*.

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

Gaut, Concepción M Díez, and Peter L Morrell (2015). "Genomics and the Contrasting Dynamics of Annual and Perennial Domestication." In: *Trends in Genetics* 31, pp. 709–719.

Gavrilets, Sergey (Oct. 2003). "PERSPECTIVE: MODELS OF SPECIATION: WHAT HAVE WE LEARNED IN 40 YEARS?" In: *Evolution* 57.10, pp. 2197–2215.

Han, Fan et al. (Jan. 6, 2017). "Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches." In: *Genome Research* 27.6, pp. 1004–1015.

Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy." In: *Nature* 585.7825. Number: 7825 Publisher: Nature Publishing Group, pp. 357–362.

Hejase, Hussein A. et al. (Dec. 2020). "Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps." In: *Proceedings of the National Academy of Sciences* 117.48, pp. 30554–30565.

Huang, Pu et al. (2014). "Population genetics of Setaria viridis, a new model system." In: *Molecular Ecology* 20, pp. 4912–4925.

Hudson (Feb. 1, 2002). "Generating samples under a Wright–Fisher neutral model of genetic variation." In: *Bioinformatics* 18.2, pp. 337–338.

Hudson, Slatkin, and Maddison (Oct. 1, 1992). "Estimation of levels of gene flow from DNA sequence data." In: *Genetics* 132.2, pp. 583–589.

Hufford, Matthew, Xun Xu, Joost van Heerwaarden, et al. (2012). "Comparative population genomics of maize domestication and improvement." In: *Nature genetics* 44.7, pp. 808–811.

Hufford, Matthew B et al. (2012). "Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight." In: *PLoS One* 7.11, e47659.

Hufford, Xun Xu, Joost Van Heerwaarden, et al. (2021). "De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes." In: *Heredity* 126, pp. 365–376.

Hufford et al. (May 9, 2013). "The Genomic Signature of Crop-Wild Introgression in Maize." In: *PLOS Genetics* 9.5, e1003477.

Jia, Guanqing et al. (2013). "A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (Setaria italica)." In: *Nature genetics* 45.8, pp. 957–961.

Kaplan, N L, R R Hudson, and C H Langley (Dec. 1, 1989). "The "hitchhiking effect" revisited." In: *Genetics* 123.4, pp. 887–899.

Kassambara, Alboukadel (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. Version 0.4.0.

Kassambara, Alboukadel and Fabian Mundt (2017). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Version 1.0.7.

Kermicle (1997). "Cross compatibility within the genus Zea." In: *Gene Flow Among Maize Landraces, Improved Maize Varieties, and Teosinte: Implications for Transgenic Maize. JA Serratos, MC Willcox, F. Castillo (eds). CIMMYT, México, D. F. pp*, pp. 40–43.

Knief, Ulrich et al. (Apr. 2019). "Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone." In: *Nature Ecology & Evolution* 3.4. Number: 4 Publisher: Nature Publishing Group, pp. 570–576.

Koide, Yohei et al. (2008). "Sex-independent transmission ratio distortion system responsible for reproductive barriers between Asian and African rice species." In: *New Phytologist* 179.3, pp. 888–900.

Kume, M. et al. (2010). "Ecological divergence and habitat isolation between two migratory forms of Japanese threespine stickleback (Gasterosteus aculeatus)." In: *Journal of Evolutionary Biology* 23.7, pp. 1436–1446.

Laetsch, Dominik R. et al. (June 6, 2023). *Demographically explicit scans for barriers to gene flow using gIMble*. Pages: 2022.10.27.514110 Section: New Results.

Le, S., J. Josse, and F. Husson (2008). "FactoMineR: An R Package for Multivariate Analysis." In: *Journal of Statistical Software* 25.1, pp. 1–18.

Le Corre, Valerie et al. (2020). "Adaptive introgression from maize has facilitated the establishment of a teosinte as a noxious weed in Europe." In: *Proc. Proceedings of the National Academy of Sciences* 117, pp. 25618–25627.

Lemaire, Louisiane et al. (Jan. 15, 2016). *Goodness-of-fit statistics for approximate Bayesian computation*. arXiv: `1601.04096[stat]`.

Leroy, Thibault et al. (2020). "Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers." In: *New Phytologist* 226.4, pp. 1183–1197.

Li et al. (2009). "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics* 25.16, pp. 2078–2079.

Lindtke, Dorothea and C. Alex Buerkle (Aug. 2015). "The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact: THE GENETIC ARCHITECTURE OF HYBRID INCOMPATIBILITIES." In: *Evolution* 69.8, pp. 1987–2004.

Lu, Moran, Srilakshmi Makkenna, et al. (2020). "Insights into the molecular control of cross-incompatibility in Zea mays." In: *Plant reproduction* 33.3, pp. 117–128.

Lu, Yongxian, Samuel A. Hokkin, Jerry L. Kermicle, et al. (2019). "A pistil-expressed pectin methylesterase confers cross-incompatibility between strains of Zea mays." In: *Nature communications* 10.1, pp. 1–7.

Lu et al. (May 24, 2019). "A pistil-expressed pectin methylesterase confers cross-incompatibility between strains of Zea mays." In: *Nature Communications* 10.1, p. 2304.

Marie-Orleach, Lucas, Christian Brochmann, and Sylvain Glémin (2022). "Mating system and speciation I: Accumulation of genetic incompatibilities in allopatry." In: *PLOS Genetics* 18, e1010353.

Martin (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." In: *EMBnet.journal* 17.1, pp. 10–12.

Martin, John W. Davey, and Chris D. Jiggins (Jan. 1, 2015). "Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci." In: *Molecular Biology and Evolution* 32.1, pp. 244–257.

Matute, Daniel R. et al. (Sept. 17, 2010). "A Test of the Snowball Theory for the Rate of Evolution of Hybrid Incompatibilities." In: *Science*.

Mayr, E (1942). "Systematics and the origin of species–Columbia Univ." In: *Press, New York*, pp. 99–107.

McKenna, A., M. Hanna, E. Banks, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." In: *Genome Res* 20, pp. 1297–1303.

Merrill, Richard M. et al. (Feb. 7, 2019). "Genetic dissection of assortative mating behavior." In: *PLOS Biology* 17.2, e2005902.

Meschiari, Stefano (2023). *latex2exp: Use LaTeX Expressions in Plots*. Available from: `https://github.com/stefano-meschiari/latex2exp`.

Metzler, Dirk et al. (Dec. 1, 2021). "Assortative mating and epistatic mating-trait architecture induce complex movement of the crow hybrid zone." In: *Evolution* 75.12, pp. 3154–3174.

Miles, Alistair et al. (May 14, 2021). *cggh/scikit-allel: v1.3.3*.

Monnet, François et al. (Oct. 20, 2023). *Rapid establishment of species barriers in plants compared to animals*. Pages: 2023.10.16.562535 Section: New Results.

Moran et al. (2017). "A pectin methylesterase ZmPme3 is expressed in Gametophyte factor1-s (Ga1-s) silks and maps to that locus in maize (Zea mays L.)" In: *Frontiers in plant science* 8, p. 1926.

Moyle, Leonie C and Takuya Nakazato (2010). "Hybrid incompatibility "snowballs" between Solanum species." In: *Science* 329.5998, pp. 1521–1523.

Nei and Li (1979a). "Mathematical model for studying genetic variation in terms of restriction endonucleases." In: *Proceedings of the National Academy of Sciences* 76, pp. 5269–5273.

— (Oct. 1979b). "Mathematical model for studying genetic variation in terms of restriction endonucleases." In: *Proceedings of the National Academy of Sciences of the United States of America* 76.10, pp. 5269–5273.

Orr (1996). "Dobzhansky, Bateson, and the genetics of speciation." In: *Genetics* 144.4, p. 1331.

Orr and Presgraves (2000). "Speciation by postzygotic isolation: forces, genes and molecules." In: *BioEssays* 22.12, pp. 1085–1094.

Orr and Turelli (2001). "The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities." In: *Evolution* 55.6, pp. 1085–1094.

Piperno et al. (2009). "Starch grain and phytolith evidence for early ninth millennium BP maize from the central Balsas River valley." In: *Proc Natl Acad Sci USA* 106.13, pp. 5019–24.

Poelstra, J. W. et al. (June 20, 2014). "The genomic landscape underlying phenotypic integrity in the face of gene flow in crows." In: *Science* 344.6190, pp. 1410–1414.

Pohl, R. W. (1951). "The genus Setaria in Iowa." In: *IA St. Jour. Sci.* 25, pp. 501–508.

— (1966). "The grasses of Iowa." In: *Iowa St. Jour. Sci.* 40, pp. 341–373.

Postel, Zoé and Pascal Touzet (2020). "Cytonuclear Genetic Incompatibilities in Plant Speciation." In: *Plants* 9.4.

Powell, Daniel L. et al. (May 15, 2020). "Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish." In: *Science* 368.6492, pp. 731–736.

Presgraves (2002). "Patterns of Postzygotic Isolation in Lepidoptera." In: *Evolution* 56.6, pp. 1168–1183.

Price, Trevor D and Michelle M Bouvier (2002). "The evolution of F1 postzygotic incompatibilities in birds." In: *Evolution* 56.10, pp. 2083–2089.

Pudlo, Pierre et al. (Mar. 15, 2016). "Reliable ABC model choice via random forests." In: *Bioinformatics* 32.6, pp. 859–866.

Purugganan, Michael D (2019). "Evolutionary Insights into the Nature of Plant Domestication." In: *Current Biology* 29, R705–R714.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Version 4.1.2. R Foundation for Statistical Computing. Vienna, Austria.

Rao, P. K. E. et al. (1987). "Intraspecific variation and systematics of cultivated Setaria italica, foxtail millet (Poaceae)." In: *Econ. Bot.* 41, pp. 108–116.

Ravinet, M. et al. (Aug. 2017). "Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow." In: *Journal of Evolutionary Biology* 30.8, pp. 1450–1477.

Raynal, Louis et al. (May 15, 2019). "ABC random forests for Bayesian parameter inference." In: *Bioinformatics* 35.10, pp. 1720–1728.

Rodriguez et al. (2006). "Characterization of floral morphology and synchrony among Zea species in Mexico." In: *Maydica* 51, pp. 383–398.

Rominger, J. M. (1962). "Taxonomy of Setaria (Gramineae) in North America." In: *Illinois Biol. Monogr.* 29.

Ross-Ibarra, Jeffrey et al. (June 11, 2008). "Patterns of Polymorphism and Demographic History in Natural Populations of Arabidopsis lyrata." In: *PLOS ONE* 3.6, e2411.

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

Roux, Camille et al. (July 1, 2013). "Crossing the Species Barrier: Genomic Hotspots of Introgression between Two Highly Divergent Ciona intestinalis Species." In: *Molecular Biology and Evolution* 30.7, pp. 1574–1587.

Roux et al. (Dec. 27, 2016). "Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence." In: *PLOS Biology* 14.12. Ed. by Craig Moritz, e2000234.

Rundle and Nosil (2005). "Ecological speciation." In: *Ecology Letters* 8.3, pp. 336–352.

Rundle et al. (Jan. 14, 2000). "Natural Selection and Parallel Speciation in Sympatric Sticklebacks." In: *Science* 287.5451, pp. 306–308.

Rushworth, Catherine A et al. (2022). "Conflict over fertilization underlies the transient evolution of reinforcement." In: *PLOS Biology* 20, e3001814.

Sakamoto, Takahiro and Hideki Innan (Aug. 2019). "The Evolutionary Dynamics of a Genetic Barrier to Gene Flow: From the Establishment to the Emergence of a Peak of Divergence." In: *Genetics* 212.4, pp. 1383–1398.

Sauvage, Christopher et al. (2017). "Domestication rewired gene expression and nucleotide diversity patterns in tomato." In: *Plant Journal* 91, pp. 631–645.

Schilling, Martin et al. (May 24, 2018). "Transitions from Single- to Multi-Locus Processes during Speciation with Gene Flow." In: *Genes* 9.6, p. 274.

Schluter, Dolph (Aug. 31, 2000). *The Ecology of Adaptive Radiation*. OUP Oxford. 302 pp.

— (July 2001). "Ecology and the origin of species." In: *Trends in Ecology & Evolution* 16.7, pp. 372–380.

Schluter, Dolph and Loren H. Rieseberg (July 26, 2022). "Three problems in the genetics of speciation by selection." In: *Proceedings of the National Academy of Sciences* 119.30, e2122153119.

Schnable, Patrick et al. (2009). "The B73 maize genome: complexity, diversity, and dynamics." In: *science* 326.5956, pp. 1112–1115.

Servedio, Maria R. et al. (Aug. 2011). "Magic traits in speciation: 'magic' but not rare?" In: *Trends in Ecology & Evolution* 26.8, pp. 389–397.

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

Shafer and Wolf (2013). "Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology." In: *Ecology Letters* 16.7, pp. 940–950.

Sobel and Streisfeld (Feb. 1, 2015). "Strong premating reproductive isolation drives incipient speciation in Mimulus aurantiacus." In: *Evolution* 69.2, pp. 447–461.

Sousa, Vitor C et al. (May 1, 2013). "Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models." In: *Genetics* 194.1, pp. 211–233.

Staab, Paul R. et al. (May 15, 2015). "scrm: efficiently simulating long sequences using the approximated coalescent with recombination." In: *Bioinformatics* 31.10, pp. 1680–1682.

Stelkens, Rike B, Kyle A Young, and Ole Seehausen (2010). "The accumulation of reproductive incompatibilities in African cichlid fish." In: *Evolution* 64.3, pp. 617–633.

Streisfeld, Young, and Sobel (Mar. 21, 2013). "Divergent Selection Drives Genetic Differentiation in an R2R3-MYB Transcription Factor That Contributes to Incipient Speciation in Mimulus aurantiacus." In: *PLOS Genetics* 9.3, e1003385.

Studer, Anthony J and John F Doebley (2011). "Do Large Effect QTL Fractionate? A Case Study at the Maize Domestication QTL teosinte branched1." In: *Genetics* 188.3, pp. 673–681.

Tajima, F. (Nov. 1989). "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." In: *Genetics* 123.3, pp. 585–595.

Tenaillon, Maud et al. (2019). "Transcriptomic response to divergent selection for flowering times reveals convergence and key players of the underlying gene regulatory network." In: *Peer Community in Evolutionary Biology*. eprint: `hal-02345286`.

Tenaillon, Maud I. et al. (2023). "Crop domestication as a step toward reproductive isolation." In: *American Journal of Botany* 110.7, e16173.

Thompson, Kim A, Loren H Rieseberg, and Dolph Schluter (2018). "Speciation and the City." In: *Trends in Ecology & Evolution* 33, pp. 815–826.

Todesco, Marco et al. (Aug. 2020). "Massive haplotypes underlie ecotypic differentiation in sunflowers." In: *Nature* 584.7822. Number: 7822 Publisher: Nature Publishing Group, pp. 602–607.

Touchard, Florence et al. (2023). "Urban rendezvous along the seashore: Ports as Darwinian field labs for studying marine evolution in the Anthropocene." In: *Evolutionary Applications* 16, pp. 560–579.

Tukey, John Wilder (1977). *Exploratory data analysis*. In collab. with Internet Archive. Reading, Mass. : Addison-Wesley Pub. Co. 714 pp.

Turissini, David A. et al. (Apr. 1, 2017). "The ability of Drosophila hybrids to locate food declines with parental divergence." In: *Evolution* 71.4, pp. 960–973.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. New York: Springer.

Vijay, Nagarjun et al. (Oct. 31, 2016). "Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex." In: *Nature Communications* 7.1. Number: 1 Publisher: Nature Publishing Group, p. 13195.

Wakeley, John and Jody Hey (Mar. 1, 1997). "Estimating Ancestral Population Parameters." In: *Genetics* 145.3, pp. 847–855.

Wang and et al (2022). "Pan-mitogenomics reveals the genetic basis of cytonuclear conflicts in citrus hybridization, domestication, and diversification." In: *Proceedings of the National Academy of Sciences* 119, e2206076119.

Wang, Jie et al. (2021). "De novo genome assembly of a foxtail millet cultivar Huagu11 uncovered the genetic difference to the cultivar Yugu1, and the genetic mechanism of imazethapyr tolerance." In: *BMC Plant Biology* 21.1, pp. 1–11.

Wang et al. (2010). "Population genetics of foxtail millet and its wild ancestor." In: *BMC Genetics* 11.1, p. 90.

Wang et al. (2022). "Three types of genes underlying the Gametophyte factor1 locus cause unilateral cross incompatibility in maize." In: *Nature Communications* 13.1.

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

— (Apr. 1, 1975b). "On the number of segregating sites in genetical models without recombination." In: *Theoretical Population Biology* 7.2, pp. 256–276.

Westram, Anja M. et al. (2022). "What is reproductive isolation?" In: *Journal of Evolutionary Biology* 35.9, pp. 1143–1164.

Wickham, Hadley (2018). *scales: Scale Functions for Visualization*. Version 1.1.1.

Wolf and Hans Ellegren (2017). "Making sense of genomic islands of differentiation in light of speciation." In: *Nature Reviews Genetics* 18, pp. 87–100.

Wood, Troy E. et al. (Aug. 18, 2009). "The frequency of polyploid speciation in vascular plants." In: *Proceedings of the National Academy of Sciences* 106.33, pp. 13875–13879.

Wright, Stephen I., Irie Vroh Bi, Steve G. Schroeder, et al. (2005). "The effects of artificial selection on the maize genome." In: *Science* 308.5726, pp. 1310–1314.

Wright et al. (May 27, 2005). "The Effects of Artificial Selection on the Maize Genome." In: *Science* 308.5726, pp. 1310–1314.

Wright et al. (Feb. 26, 2013). "Indirect Evolution of Hybrid Lethality Due to Linkage with Selected Locus in Mimulus guttatus." In: *PLOS Biology* 11.2, e1001497.

Wu (Dec. 20, 2001). "The genic view of the process of speciation: Genic view of the process of speciation." In: *Journal of Evolutionary Biology* 14.6, pp. 851–865.

Wu, WenWan et al. (2014). "The early Holocene archaeobotanical record from the Zhangmatun site situated at the northern edge of the Shandong Highlands, China." In: *Quaternary International* 348, pp. 183–193.

Xu, Xun et al. (Nov. 1, 2020). "Divergence in flowering time is a major component contributing to reproductive isolation between two wild rice species (Oryza rufipogon and O. nivara)." In: *Science China Life Sciences* 63.11, pp. 1714–1724.

Yang et al. (2019). "The genetic architecture of teosinte catalyzed and constrained maize domestication." In: *Proceedings of the National Academy of Sciences* 116, pp. 5643–5652.

Yang et al. (2023). "Two teosintes made modern maize." In: *Science* 382.6674, eadg8940.

Yeaman, Sam and Sarah P. Otto (July 2011). "ESTABLISHMENT AND MAINTENANCE OF ADAPTIVE GENETIC DIVERGENCE UNDER MIGRATION, SELECTION, AND DRIFT." In: *Evolution* 65.7, pp. 2123–2129.

Zamani, Neda et al. (May 24, 2013). "Unsupervised genome-wide recognition of local relationship patterns." In: *BMC Genomics* 14.1, p. 347.

Zhang et al. (2018). "A PECTIN METHYLESTERASE gene at the maize Ga1 locus confers male function in unilateral cross-incompatibility." In: *Nature communications* 9.1.

Zhang et al. (2023). "A pollen expressed PME gene at Tcb1 locus confers maize unilateral cross-incompatibility." In: *Plant Biotechnology Journal* 21.3, p. 454.

Appendix A

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

Table A.1: Demographic parameters used under four demographic models (SI: Strict Isolation, IM: Isolation Migration, SC: Secondary Contact, AM: Ancestral Migration) and four Genomic model (1M, 2M, 1N, 2N). Parameters are either estimated (empty field) or fixed to a value defined as indicated - either 0 or the value of another parameter. Note that for a single simulation, K value is drawn in $U \in [0,1]$ only once, so it means that $T_{SC} = T_{AM} = K * T_{split}$

	AM				IM				SC				SI	
	1M1N	1M2N	2M1N	2M2N	1M1N	1M2N	2M1N	2M2N	1M1N	1M2N	2M1N	2M2N	1N	2N
T_{split}														
T_{AM}					$K^* T_{split}$	$K^* T_{split}$	$K^* T_{split}$	$K^* T_{split}$	T_{split}	T_{split}	T_{split}	T_{split}	T_{split}	T_{split}
T_{SC}	0	0	0	0	$K^* T_{split}$	$K^* T_{split}$	$K^* T_{split}$	$K^* T_{split}$					0	0
Na														
$N1$														
$N2$														
M_{cur}	0	0	0	0									0	0
M_{anc}					M_{cur}	M_{cur}	M_{cur}	M_{cur}	0	0	0	0	0	0
α	1.10^4		1.10^4		1.10^4		1.10^4		1e4		1.10^4		1.10^4	1.10^4
β	1.10^4		1.10^4		1.10^4		1.10^4		1.10^4		1.10^4		1.10^4	1.10^4
Q_{anc}	0	0			0	0	Q_{cur}	Q_{cur}	0	0	0	0	0	0
Q_{cur}	0	0	0	0	0	0			0	0			0	0

Table A.2: Parameter values used in the simulations of pseudo-observed datasets. Note that for the strict isolation model, only T_{split} varies. The number of loci in a pseudo-observed dataset is 1000 loci of 10 kb each. The mutation rate was set to 1.10^{-8} and the recombination rate to 1.10^{-7} event/generation/bp. Populations size $N_1 = N_2 = N_a = 5.10^4$ individuals. From each daughter's populations, 20 haploid samples are produced. Each condition is repeated 100 times. To run RIDGE on each pseudo-observed dataset, prior were defined as follows: T_{split} and N_e prior distribution is bounded by one order below and above the true value (e.g, for 1.10^5 the distribution is bounded between 1.10^4 and 1.10^6). The M prior distribution is bounded between 0.1 and 50 4N.m, and the Q prior distribution is bounded between 0 and 0.2.

Parameter	Parameter value
T_{split}	$1.10^4, 1.10^5, 2.10^5, 1.10^6, 2.10^6$
m_{cur} and m_{anc}	1, 10
Q_{cur} and Q_{anc}	0.01, 0.05, 0.1

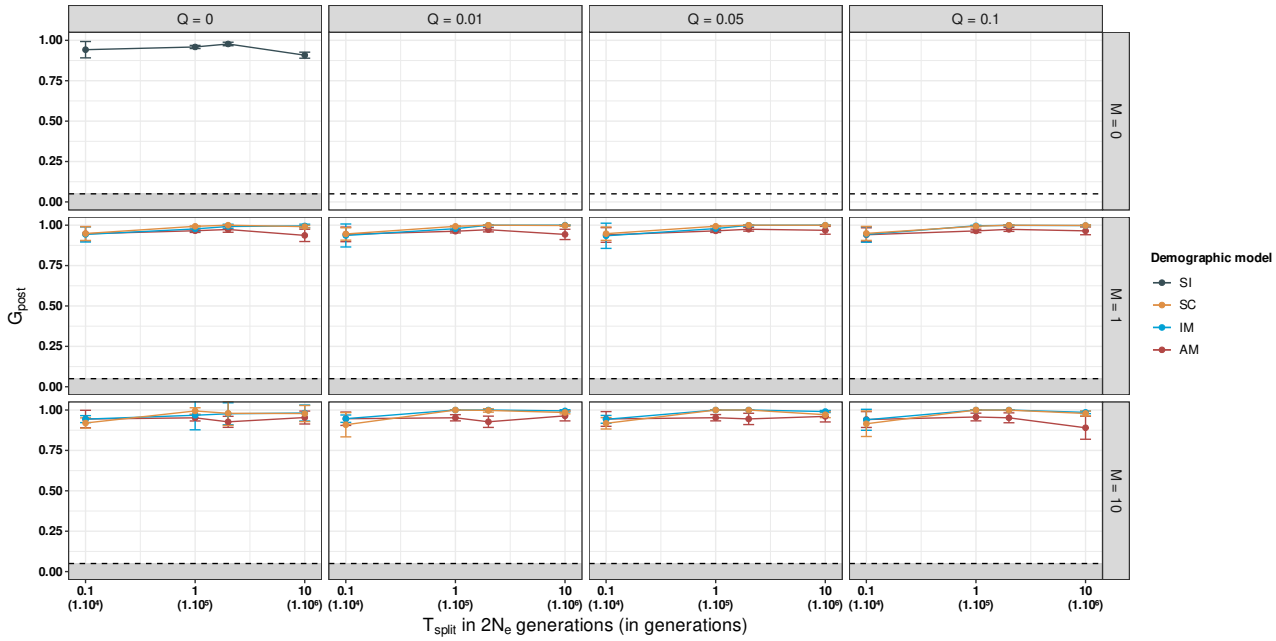

Figure A.1: Evolution of the goodness-of-fit of the posteriors as a function of T_{split}, migration (M) and barrier proportion (Q), for four demographic models. The gray zone represents the rejection zone, in which inferred models are discarded. Average values over 100 replicates with error bars (standard deviation) are presented. Pseudo-observed datasets were simulated under $2N2m$ and $2N1m$ models.

Table A.3: Prior bound used to run RIDGE over all crow population pairs

Parameter	Parameter bound (min-max)
T_{split}	10 000 - 150 000
N_e	30 000 - 250 000
M	0.1 - 50
Q	0 - 0.2

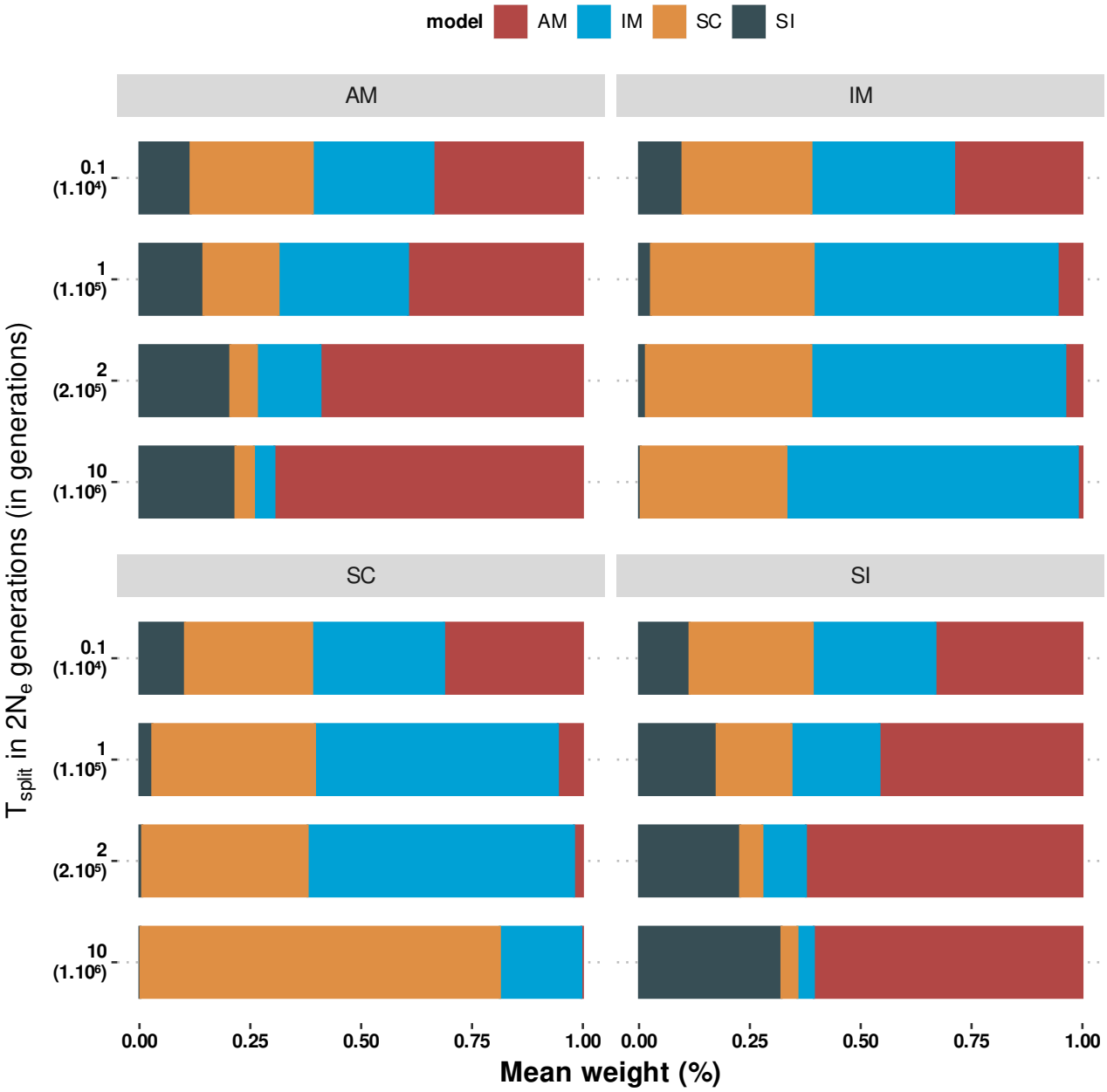

Figure A.2: Demographic model weights in posteriors across time splits. The simulated demographic model is indicated above each plot in grey and the proportion of model predictions are shown in colors. All models were simulated under $2N2m$ and under $2N$ for SI.

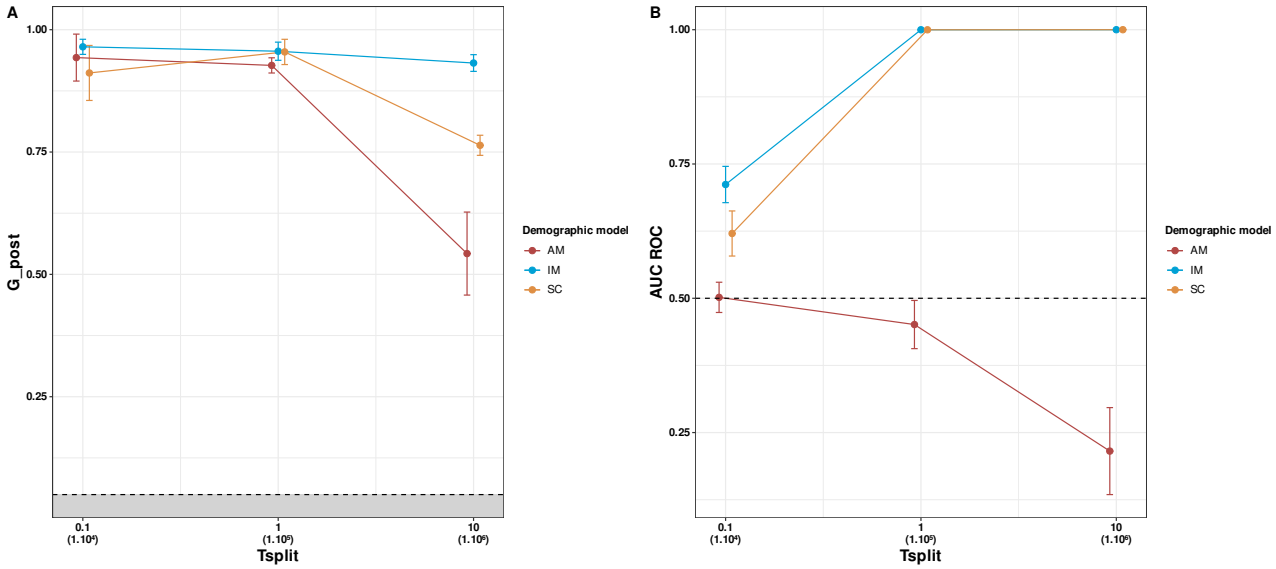

Figure A.3: Effect of model misspecification on the evolution of the goodness-of-fit of the posteriors (A) and the as a function of T_{split} and the discriminant power measured through the AUC of ROC (B). RIDGE were trained on a *reference table* only composed of IM $2N2m$ simulated dataset and applied on model under $2N2m$ with $M = 10$ and $Q = 0.1$ Average values over 100 replicates with error bars (standard deviation) are presented.

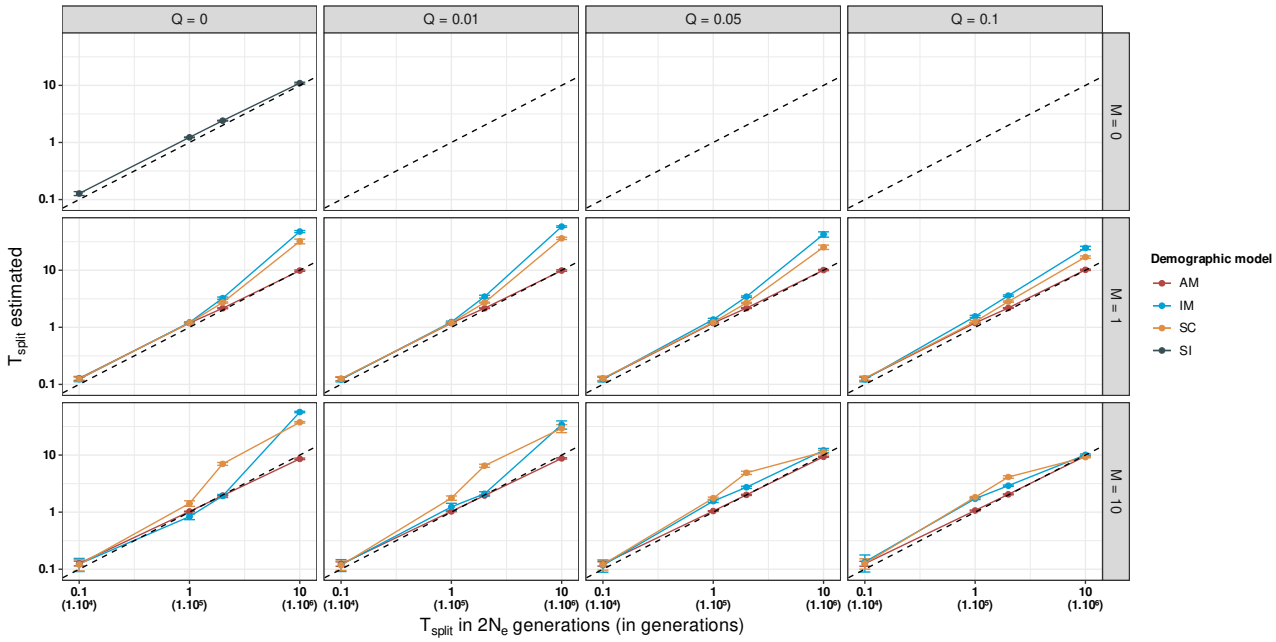

Figure A.4: Estimated Time split ($\hat{T}_{split}$) as a function of simulated time (T_{split}) split, migration (M) and barrier proportion (Q) for four demographic models. Average values over 100 replicates with error bars (standard deviation) are presented. Pseudo-observed data were simulated under $2N2m$ and $2N1m$. The dashed line represents the reference (simulated=estimated).

Table A.4: Pearson correlation (r) between estimated proportion of barrier Q and outlier statistics under three demographic models with different Time of split. Simulations were ran under $2N2m$ model with $M = 10$ and $Q = 0$. Values of r >0.5 are shown in bold, NA indicates that correlation could not be computed.

Model	Tsplit	r between Q and outlier statistic					
		F_{ST}	*Dxy*	*Da*	*sf*	*ss*	π
AM	1.10^4	-0.40	-0.03	**-0.55**	-0.04	-0.08	0.03
	1.10^5	-0.33	0.30	**0.51**	-0.30	0.27	-0.14
	2.10^5	0.19	0.35	**0.96**	0.02	NA	0.05
	1.10^6	0.007	**0.51**	**0.92**	-0.06	NA	0.04
IM	1.10^4	NA	0.11	NA	-0.14	-0.16	0.08
	1.10^5	NA	0.02	NA	**0.96**	**0.94**	0.1
	2.10^5	NA	0.21	NA	**0.96**	**0.95**	0.13
	1.10^6	NA	**0.99**	NA	**0.99**	**0.90**	-0.11
SC	1.10^4	0.28	-0.03	0.20	-0.01	-0.41	-0.12
	1.10^5	NA	0.09	NA	**0.95**	**0.96**	-0.04
	2.10^5	NA	0.01	NA	**0.97**	**0.96**	0.031
	1.10^6	NA	**0.990**	NA	**0.98**	**0.88**	-0.43

Table A.5: Estimated demographic and genomic parameters for each pair of crow species from Vijay et al. (2016) and Poelstra et al. (2014) (=Poelstra comp). For each parameter, the mean is presented with a credibility interval [5%;95%]. Note that time are expressed in crow generations, migration in $4N_em$ units and population size in number of individuals.

	RX	XO	OP	Poelstra comp
$\hat{N}_1$	156626.63 [41274.11 ; 243898.74]	147427.24 [39831.97 ; 239393.74]	155749.25 [39642.31 ; 244688.48]	154688.21 [49016.26 ; 242075.48]
$\hat{N}_2$	145483.9 [36191.17 ; 241586.99]	159308.69 [42520.75 ; 243157.84]	140602.76 [36838.31 ; 239681.69]	155425.86 [47325.64 ; 239872.01]
$\hat{N}_A$	117120.17 [40547.57 ; 224625.69]	118832.69 [36789.34 ; 228009.74]	113690.48 [38945.11 ; 230523.18]	111402.79 [40332.77 ; 213925.65]
$\hat{M}_{cur}$	17.19 [0 ; 44.85]	13.7 [0 ; 43.16]	13.83 [0 ; 44.18]	13.18 [0 ; 41.64]
$\hat{M}_{anc}$	7.83 [0 ; 39.95]	7.21 [0 ; 37.21]	7.57 [0 ; 38.4]	7.04 [0 ; 35.1]
$\hat{\alpha}$	1743.85 [0.92 ; 10000]	1853.61 [0.24 ; 10000]	1753.49 [0.2 ; 10000]	1423.73 [0.7 ; 10000]
$\hat{\beta}$	1744.19 [0.75 ; 10000]	1854.24 [0.73 ; 10000]	1753.88 [0.72 ; 10000]	1424.66 [1.05 ; 10000]
$\hat{T}_{SC}$	25492.74 [0 ; 94052.16]	25492.74 [0 ; 94052.16]	26978.14 [0 ; 98910.78]	22265.8 [0 ; 77347.69]
$\hat{T}_{AM}$	66990.1 [6827.93 ; 143187.47]	66911.39 [8244.63 ; 139332.62]	69877.64 [6698.72 ; 142416.09]	59263.08 [4917.05 ; 139704.26]
$\hat{T}_{split}$	85051.2 [14839.87 ; 145398.07]	85587.62 [15740.83 ; 146227.89]	91977.35 [16403.09 ; 146742.26]	78540.5 [16975.93 ; 144101.38]
$\hat{Q}$	0.05 [0 ; 0.18]	0.05 [0 ; 0.18]	0.05 [0 ; 0.17]	0.05 [0 ; 0.18]

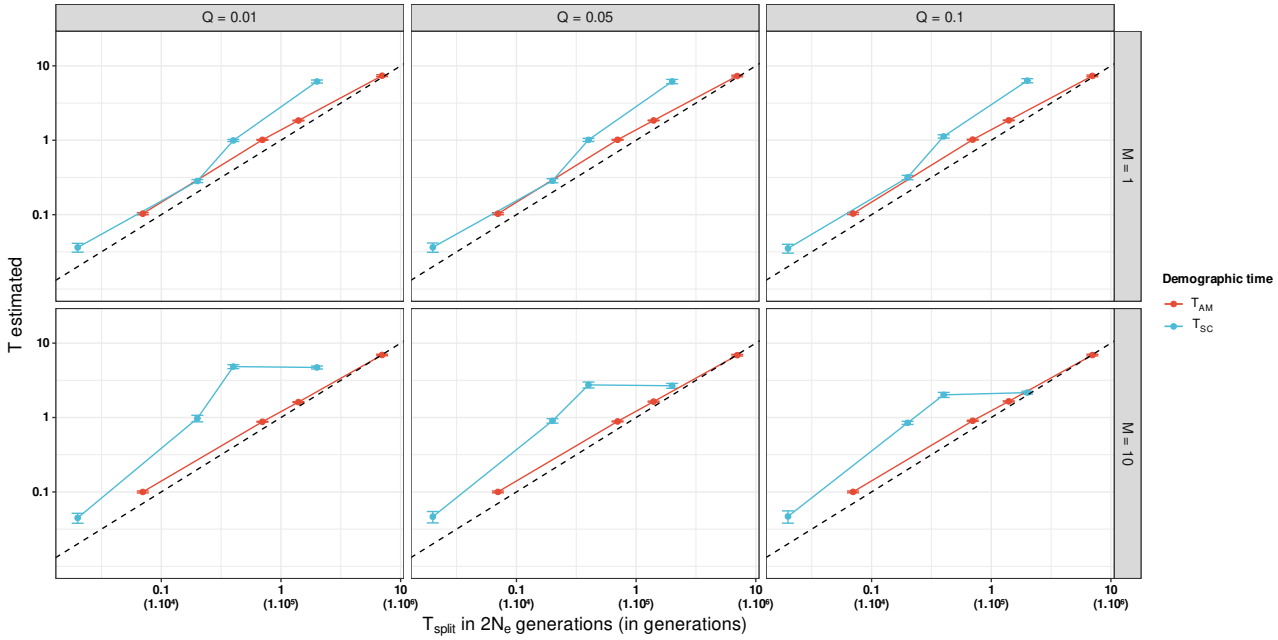

Figure A.5: Estimated Time of secondary contact ($\hat{T}_{SC}$) and time of last migratory contact ($\hat{T}_{AM}$) as a function of simulated time split (T_{split}), migration (M) and barrier proportion (Q) for respectively SC and AM demographic model. Average values over 100 replicates with error bars (standard deviation) are presented. Pseudo-observed data were simulated under $2N2m$. The dashed line represents the reference (simulated=estimated).

Demographic model weight	AM	IM	SC	SI
RX	0.13	0.43	0.39	0.04
OP	0.07	0.45	0.46	0.03
XO	0.11	0.45	0.39	0.04
Poelstra comp	0.08	0.44	0.45	0.03

Table A.6: Weight of each demographic model in posteriors for each pair of crow species from Vijay et al. (2016) and Poelstra et al. (2014) (=Poelstra comp).

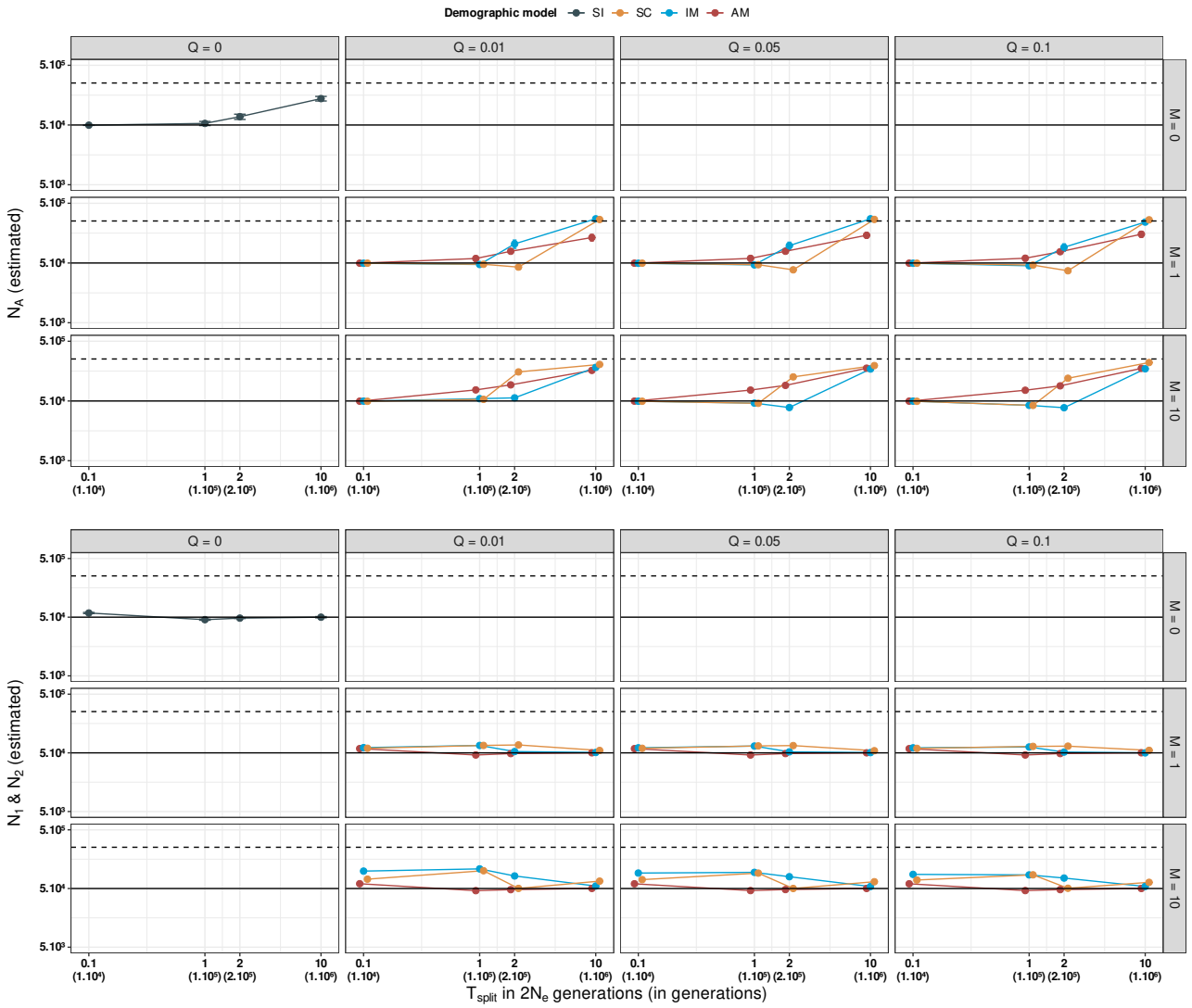

Figure A.6: Estimated population size (past with $\hat{N}_A$ and current with $\hat{N}_1$ and $\hat{N}_2$) under fours demographics models. Average values over 100 replicates with error bars (standard deviation) are presented. The plain line represents the value used in the simulation (N_e=50 000), and the dashed line represents the mean value of priors (N_e=252 500). Simulated data were obtained under $2N2m$.

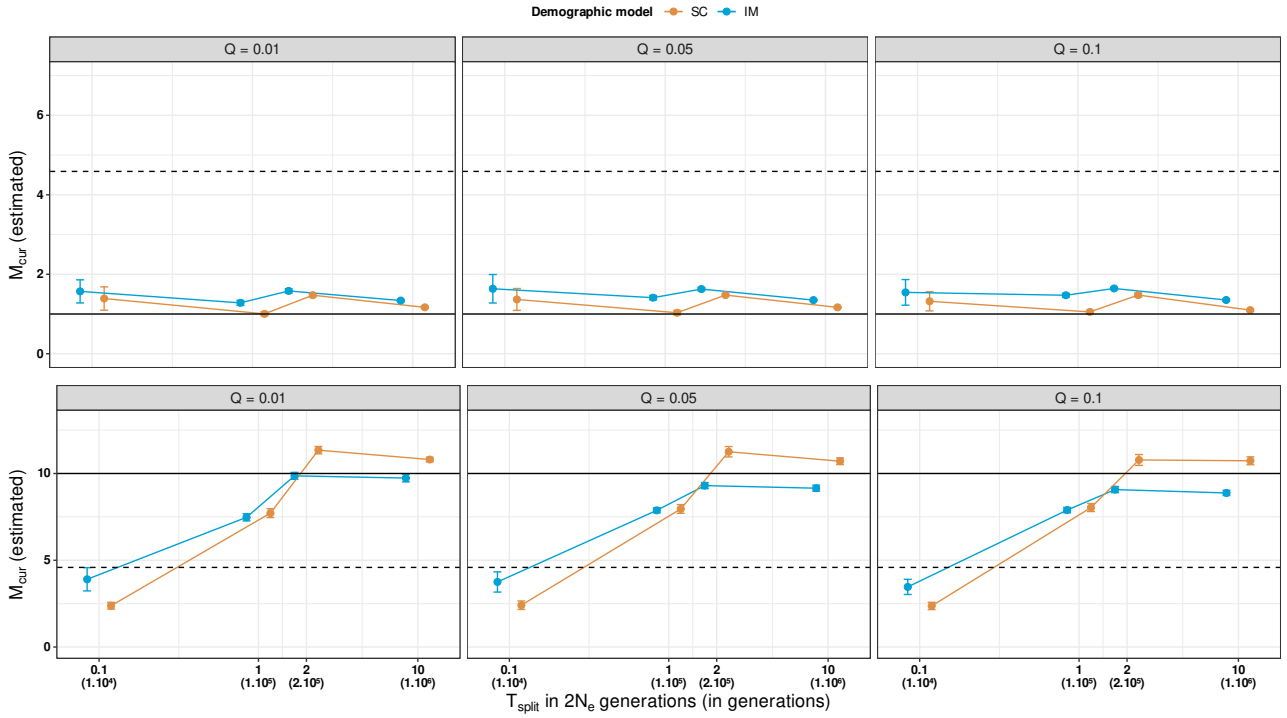

Figure A.7: Current migration rate ($\hat{M}_{cur}$) estimation accuracy under IM and SC models under $2N2m$. Average values over 100 replicates with error bars (standard deviation) are presented. The plain black line represents the true value ($M_{cur} = 1$ and 10) used to generate the pseudo-observed datasets and the dashed line represents the mean of priors $M_{cur} = 4.58$

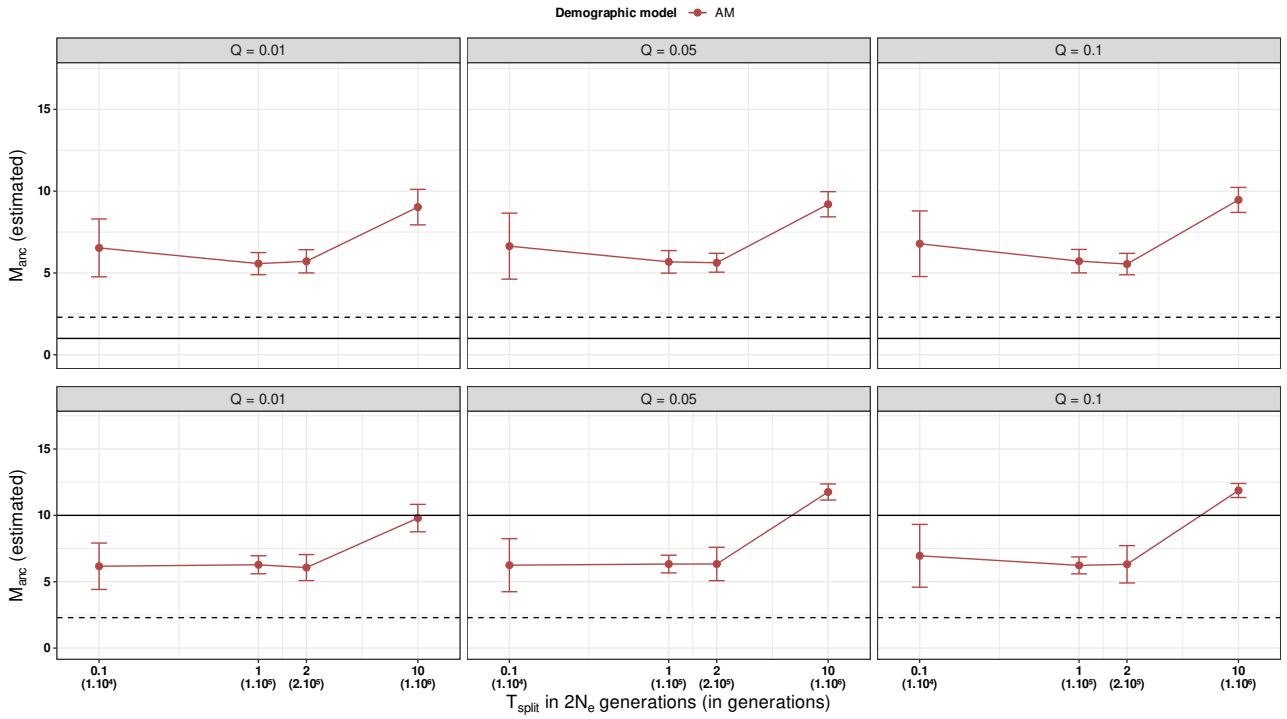

Figure A.8: Ancestral migration rate ($\hat{M}_{anc}$) estimation accuracy for AM model under $2N2m$.Average values over 100 replicates with error bars (standard deviation) are presented. The plain black line represents the true value ($M_{anc} = 1$ and 10) used to generate the pseudo-observed datasets and the dashed line represents the mean of priors $M_{anc} = 2.29$.

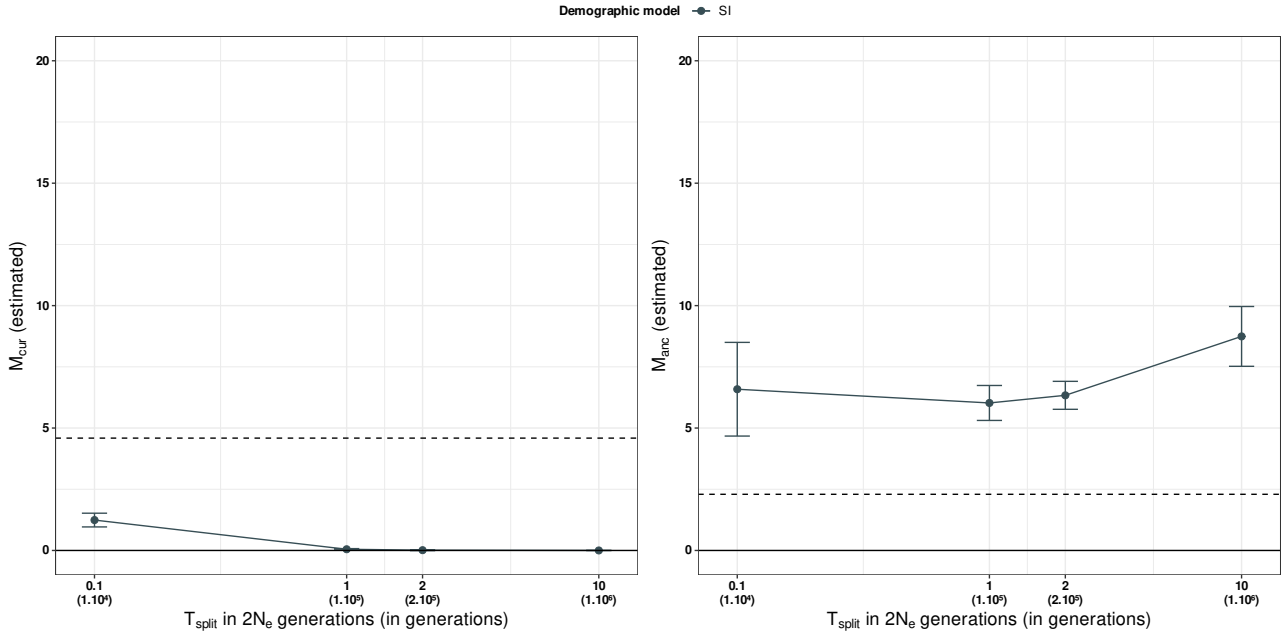

Figure A.9: Current (left) and Ancestral (right) migration rate estimation accuracy under SI model. Average values over 100 replicates with error bars (standard deviation) are presented. The plain black line represents the true value ($M_{cur} = M_{anc} = 0$) used to generate the pseudo-observed datasets and the dashed line represents the mean of priors $M_{cur} = 4.58$ (left) and $M_{anc} = 2.29$ (right).

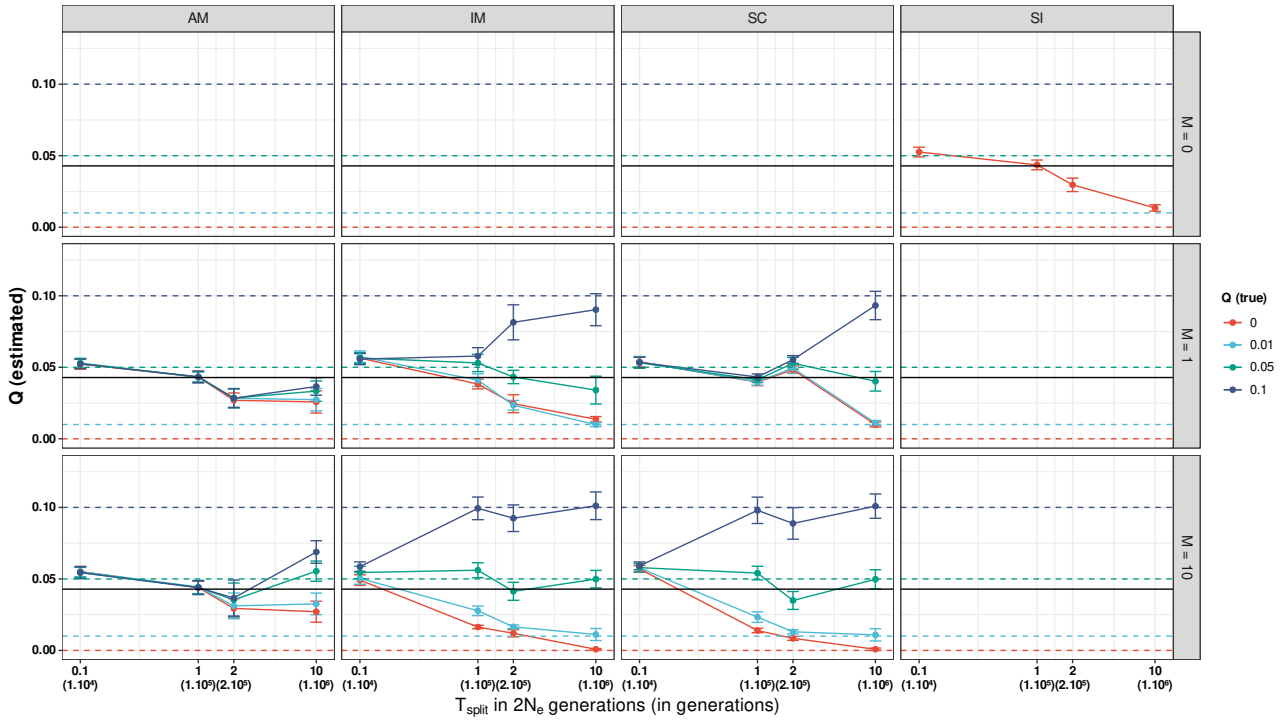

Figure A.10: Barrier proportion estimates ($\hat{Q}$) as a function of divergence time under four demographic models. Average values over 100 replicates with error bars (standard deviation) are presented and the plain black line represents the mean of priors $Q = 4.2\%$ Dashed lines represent references values corresponding to each barrier proportion conditions.

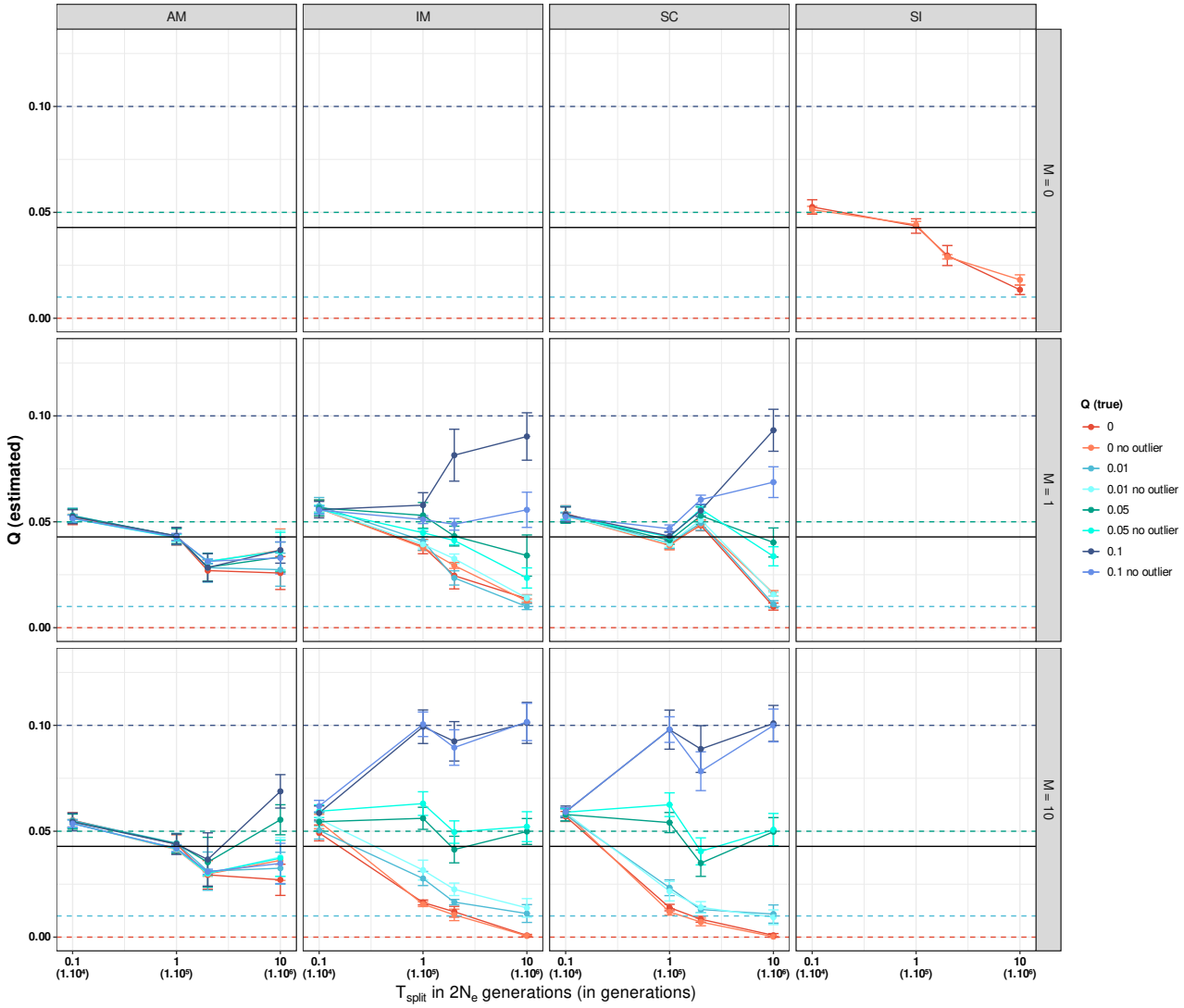

Figure A.11: Comparison between barrier proportion estimates with or without outlier summary statistics (no outlier) as a function of divergence time under the fours demographic models. Average values over 100 replicates with error bars (standard deviation) are presented and the plain black line represents the mean of priors $Q = 4.2\%$ Dashed lines represent the initial value of barrier proportion used in pseudo-observed dataset.

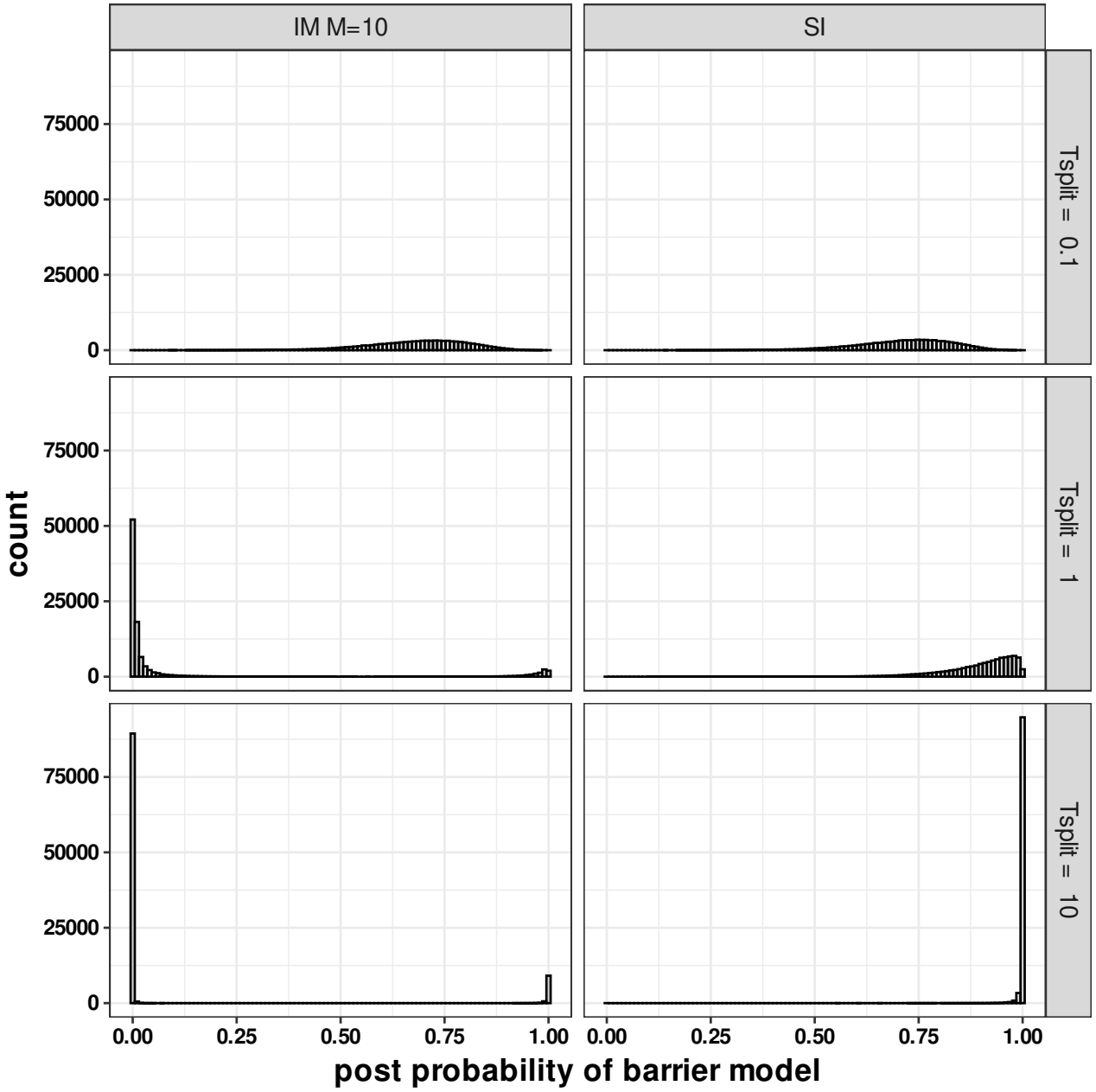

Figure A.12: Cumulative distribution of posterior probability of the barrier model under IM model with $2N2m$ and SI with 2N model in function of divergence time. For each condition, the cumulative distribution of 100 dataset of 1000 loci is represented.

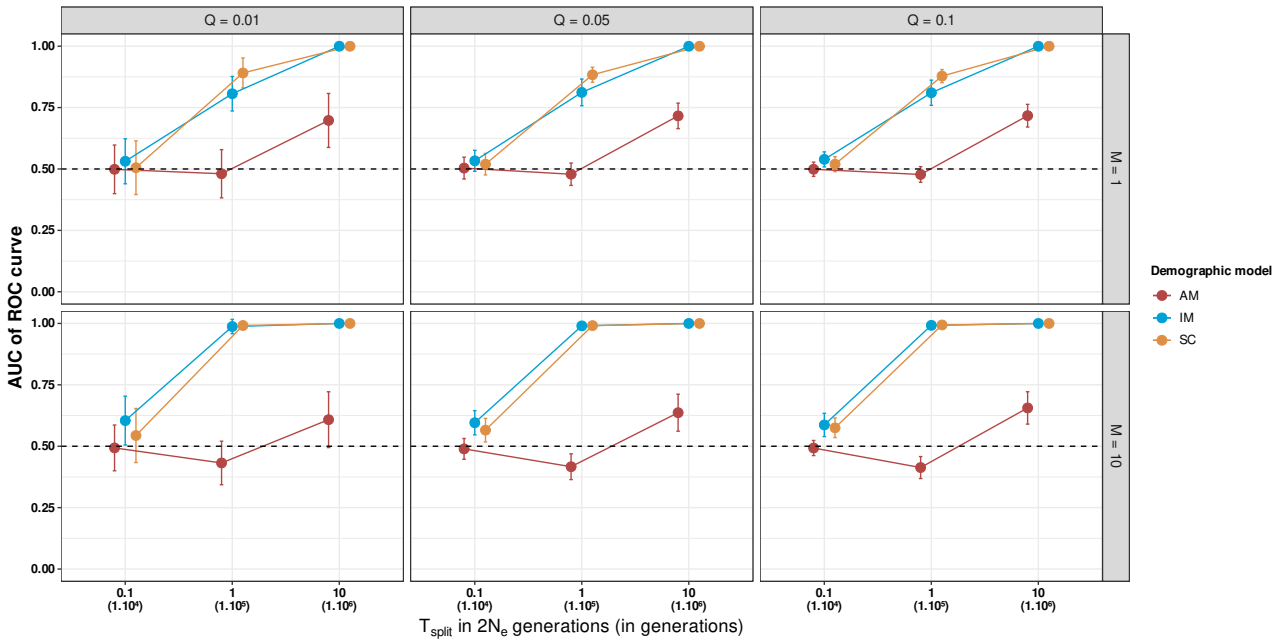

Figure A.13: Discriminant power measured through the AUC of ROC as a function of divergence time T_{split}, migration M, demographic model and the proportion of barrier Q. The AUC relates the False Positive Rate (FPR) to the True Positive Rate (TPR), the greater the AUC the higher the discriminant power. Average values over 100 replicates with error bars (standard deviation) are presented. The dashed line represents the $AUC = 0.5$ threshold, above which signal is captured. Pseudo-observed data were simulated under $2N2m$.

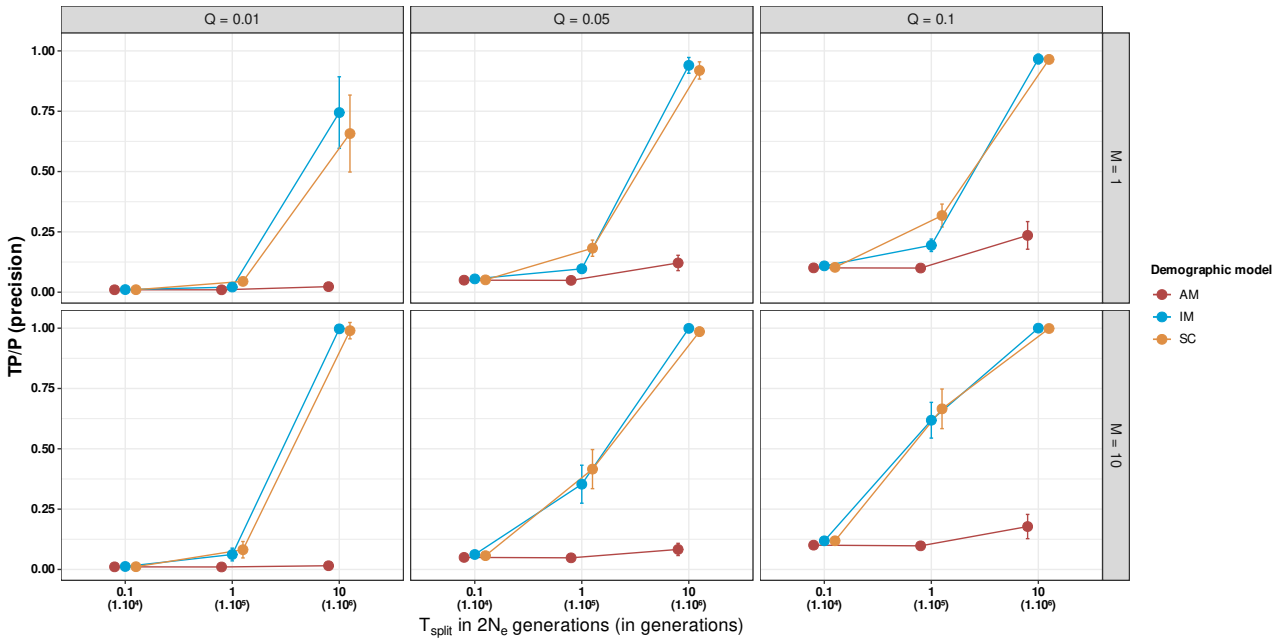

Figure A.14: Precision of barrier loci identification as a function of divergence time T_{split}, migration M, model and the proportion of barrier Q. Precision, which is the ratio of the number of true positives (TP) divided by the number of detected loci (P) – which is true positives plus the number of false positives – are shown with average values over 100 replicates with error bars (standard deviation) are presented Detected loci are loci that exhibit a barrier model posterior probability superior to 0.5Pseudo-observed data were simulated under $2N2m$.

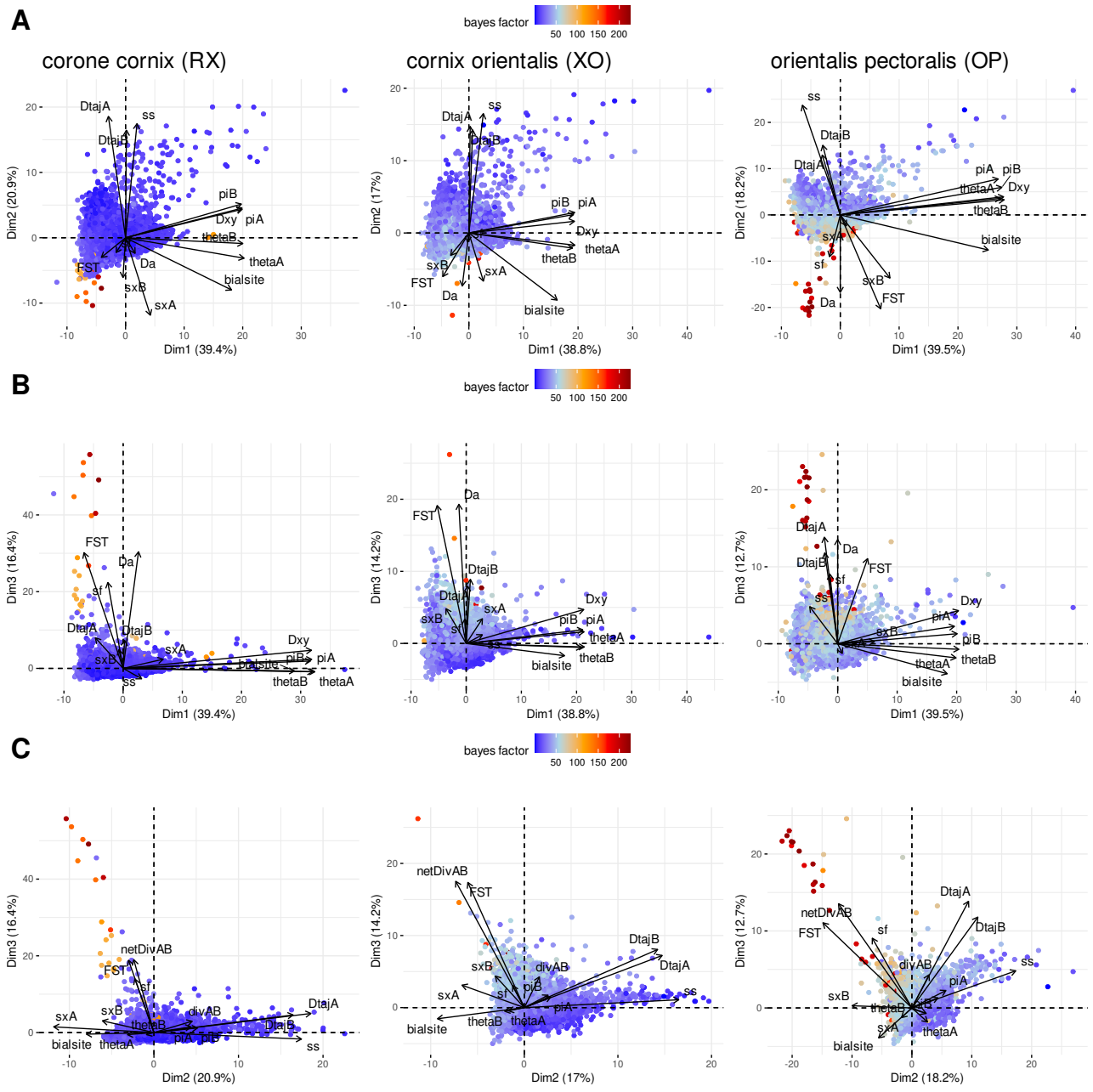

Figure A.15: PCA computed on summary statistics obtained from 50kb-windows along genomes with axes 1 and 2 (A) and 1 and 3 (B) and 2 and 3 (C) displayed. Datapoints (windows) are colored according to the values of Bayes factors. Data are from Vijay et al. (2016)

Appendix B

Appendix of application on empirical dataset

Table B.1: List of domestication genes (dom) and genes involve in RI known in maize, with their names and position.

Function	name	chr	Start pos	End pos	reference
dom	Tb1	1	272330844	272332595	Studer and J. F. Doebley (2011)
dom	Tga1	4	46648115	46652648	Wang et al, 2015
dom	Zagl1	1	4932462	4948248	Wills et al, 2018
dom	ZmSh1-5.1/5.2	5	16842725	16849077	Lin et al, 2012
dom	Gt1	1	23433586	23435122	Wills et al, 2018
dom	Ba1	3	188190176	188191189	Gallavotti et al, 2004
dom	Bt2	4	61578520	61584597	Whitt et al, 2002
dom	Ra1	7	114958643	114959358	Simon and Vollbrecht, 2010
dom	Su1	4	43429928	43438749	Whitt et al, 2002
dom	Zfl2	2	13161522	13164487	Bomblies et al, 2006
dom	ZmSh1-1	1	230589293	230594150	Lin et al, 2012
RI	Ga2	5	151000000	153500000	Chen et al., 2022
RI	Ga1/Tcb1	4	8530000	10230000	Evans & Kermicle, 2001; Lu et al.2019

Table B.2: Distribution of summary statistics values compared to observed values in maize (B)/teosinte (A) data, with IC representing quantile at 5% and 95% of the distribution. In addition, value from literature are provided.

	obs	literature
SNP/window	4.3e+02 [2.4e+01 ; 1.2e+03]	
πA	1.2e-03 [5.0e-05 ; 3.2e-03]	0.0115 (Beissinger et al. 2016)
πB	8.6e-04 [4.0e-05 ; 2.3e-03]	0.00691 (Beissinger et al. 2016)
D_{xy}	1.2e-03 [5.0e-05 ; 3.1e-03]	
D_a	1.5e-04 [0.0e+00 ; 5.8e-04]	
Dtaj A	-1.1e+00 [-1.9e+00 ; -2.2e-01]	
Dtaj B	-6.3e-01 [-1.9e+00 ; 6.7e-01]	
F_{ST}	1.0e-01 [1.5e-02 ; 2.8e-01]	0.11 (M. Hufford et al. 2012)

Table B.3: Distribution of prior and posterior summary statistics values compared to observed values in *S. italica* (B)/*S. viridis* (A) data, with IC representing quantile at 5% and 95% of the distribution. In addition, values from the literature are provided.

	obs	literature
SNP/window	1.2e+03 [2.1e+02 ; 1.7e+03]	
πA	3.1e-03 [5.9e-04 ; 4.9e-03]	0.0059 (Wang et al. 2010)
πB	1.3e-03 [1.7e-04 ; 3.2e-03]	0.0027 (Wang et al. 2010)
D_{xy}	3.0e-03 [5.8e-04 ; 4.9e-03]	
D_a	8.0e-04 [2.0e-05 ; 2.1e-03]	
Dtaj A	-7.2e-01 [-1.5e+00 ; 4.9e-02]	
Dtaj B	-1.1e+00 [-2.5e+00 ; 9.1e-01]	
F_{ST}	2.0e-01 [-4.1e-02 ; 5.0e-01]	0.15 (Wang et al. 2010)

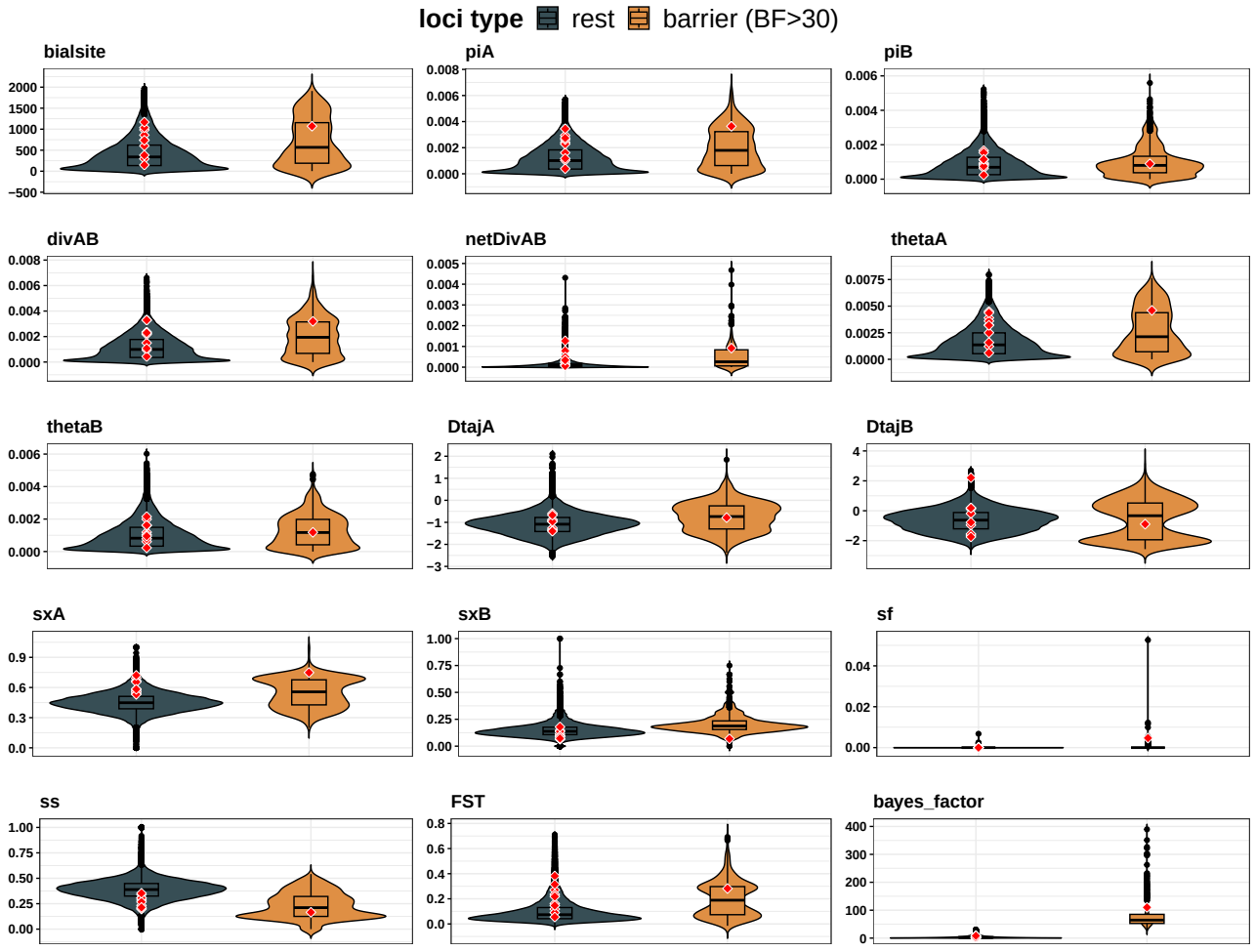

Figure B.1: Distribution of summary statistics values for maize loci in function of their loci type. Domestication genes from Table B.1 are represented by red diamonds. Barrier loci are loci showing a barrier model posterior probability > 0.5 and so a BF$>$30.

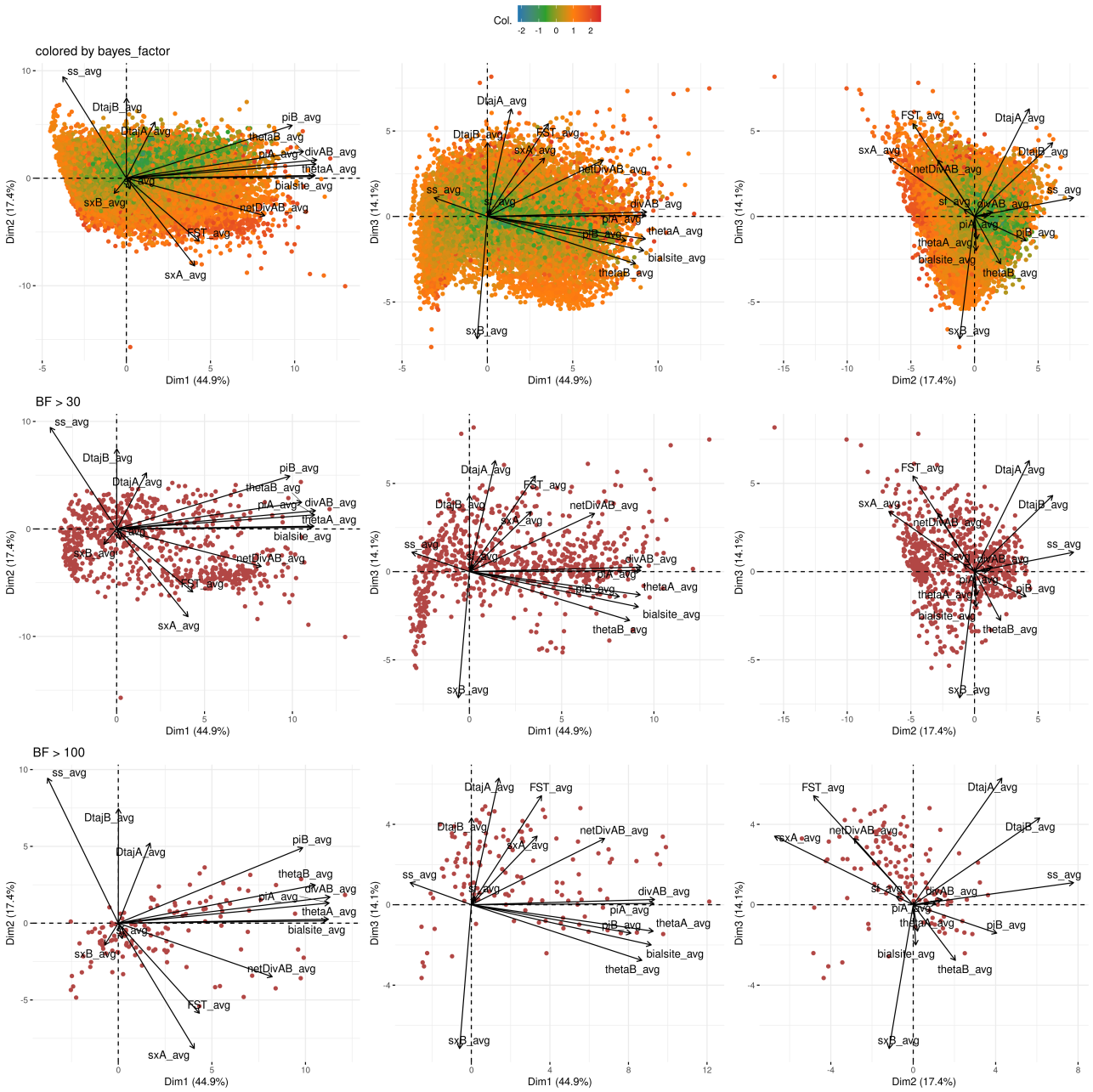

Figure B.2: PCA computed from summary statistics of maize loci colored by log of their Bayes factor. All loci (first row),with BF>30 (middle row) and BF>100 (last row). Three axes are represented and together explain 76.4% of the variance.

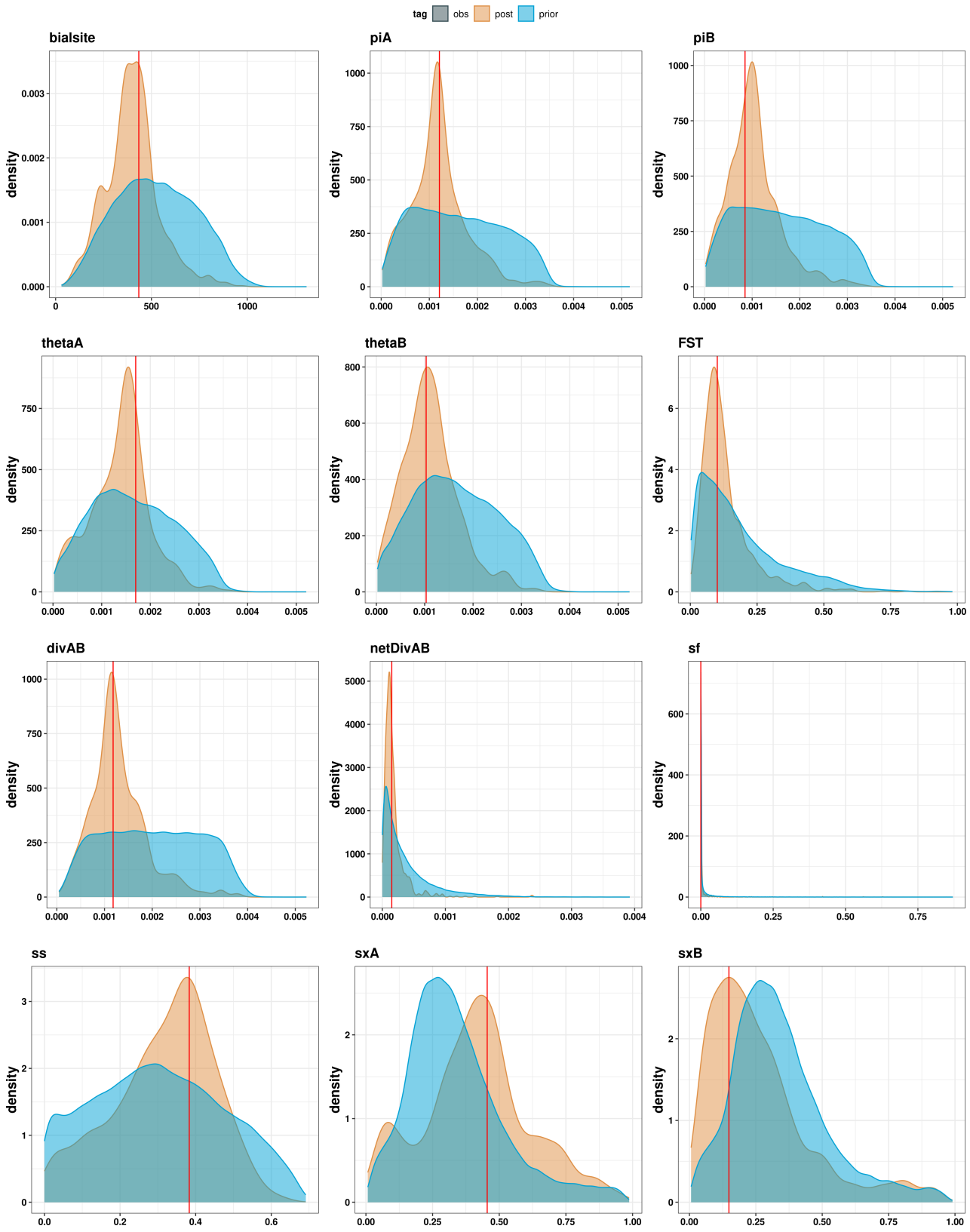

Figure B.3: Distribution of prior and posterior summary statistics values compared to observed values in maize (B)/teosinte (A) data (red line).

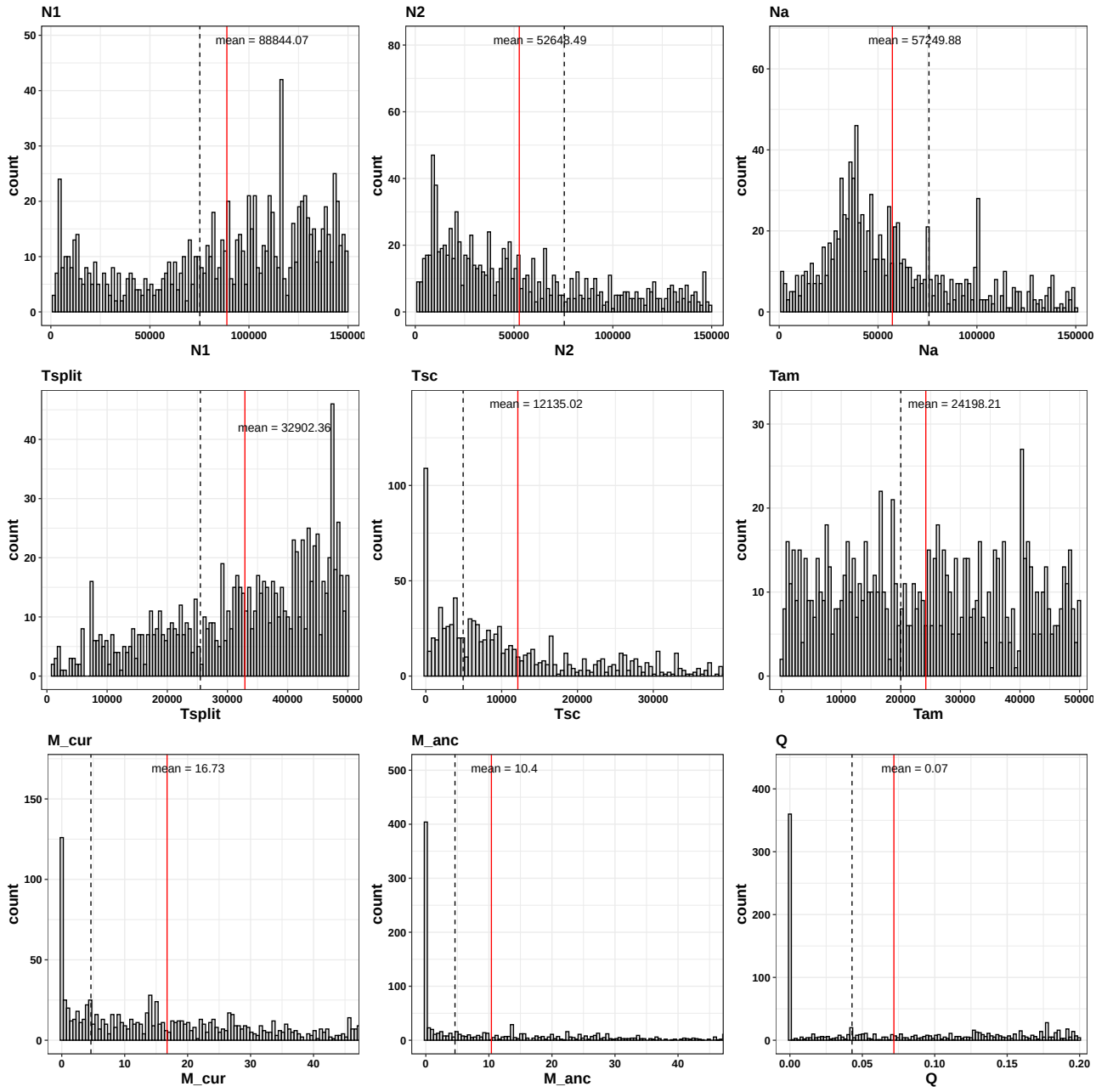

Figure B.4: Distribution of parameter posterior values for N_1, N_2, N_a, T_{split}, M_{cur}, and M_{anc} and Q. Dashed lines represent the mean value of priors, and the red line represents the mean value of posterior.

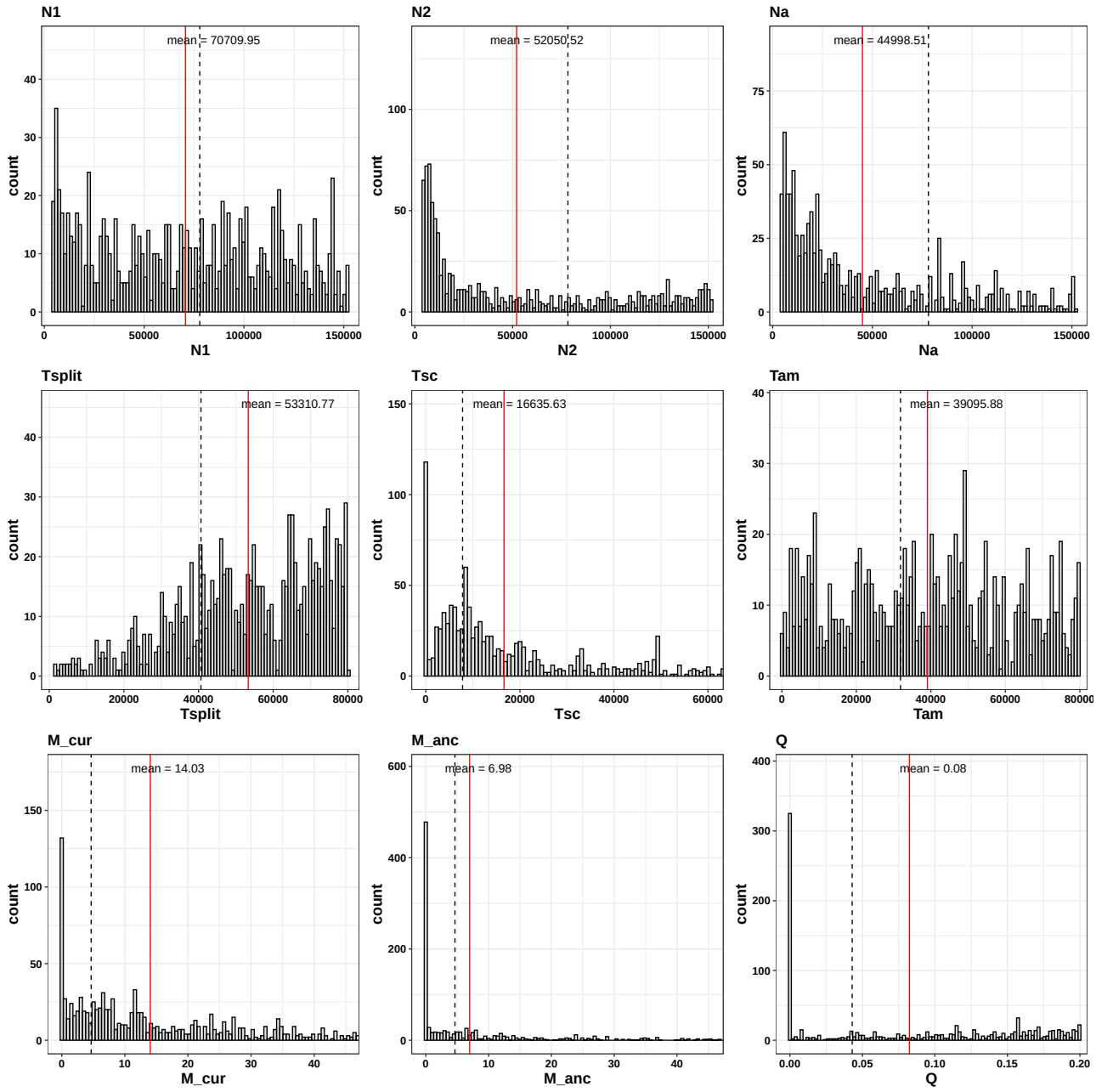

Figure B.5: Distribution of parameter posterior values for N_1, N_2, N_a, T_{split}, M_{cur}, and M_{anc} and Q. Dashed lines represent the mean value of priors, and the red line represents the mean value of posterior.

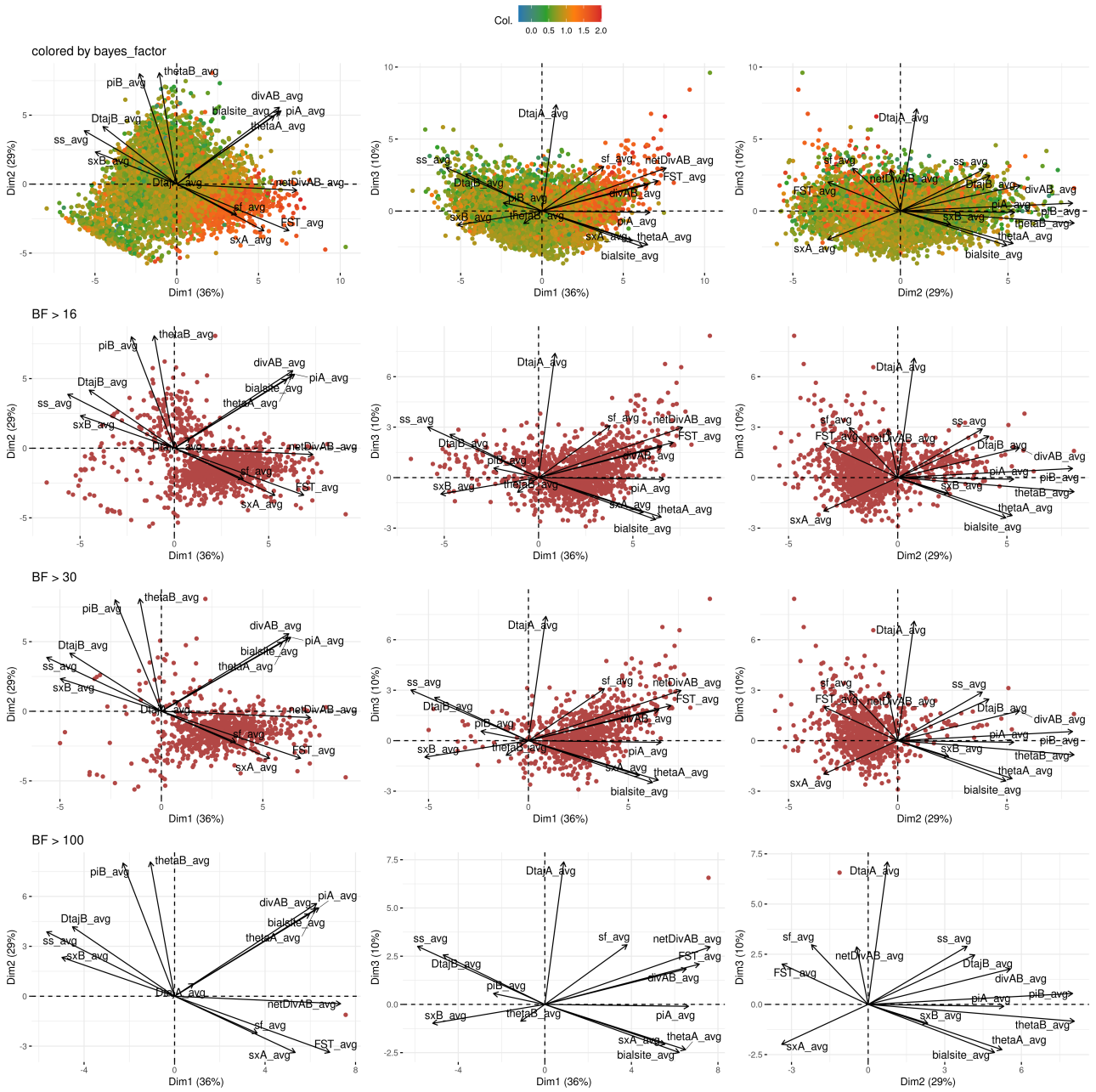

Figure B.6: PCA computed from summary statistics of foxtail millet loci colored by log of their Bayes factor. All loci (first row),with BF>30 (middle row) and BF>100 (last row). Three axes are represented and together explain 76.5% of the variance.

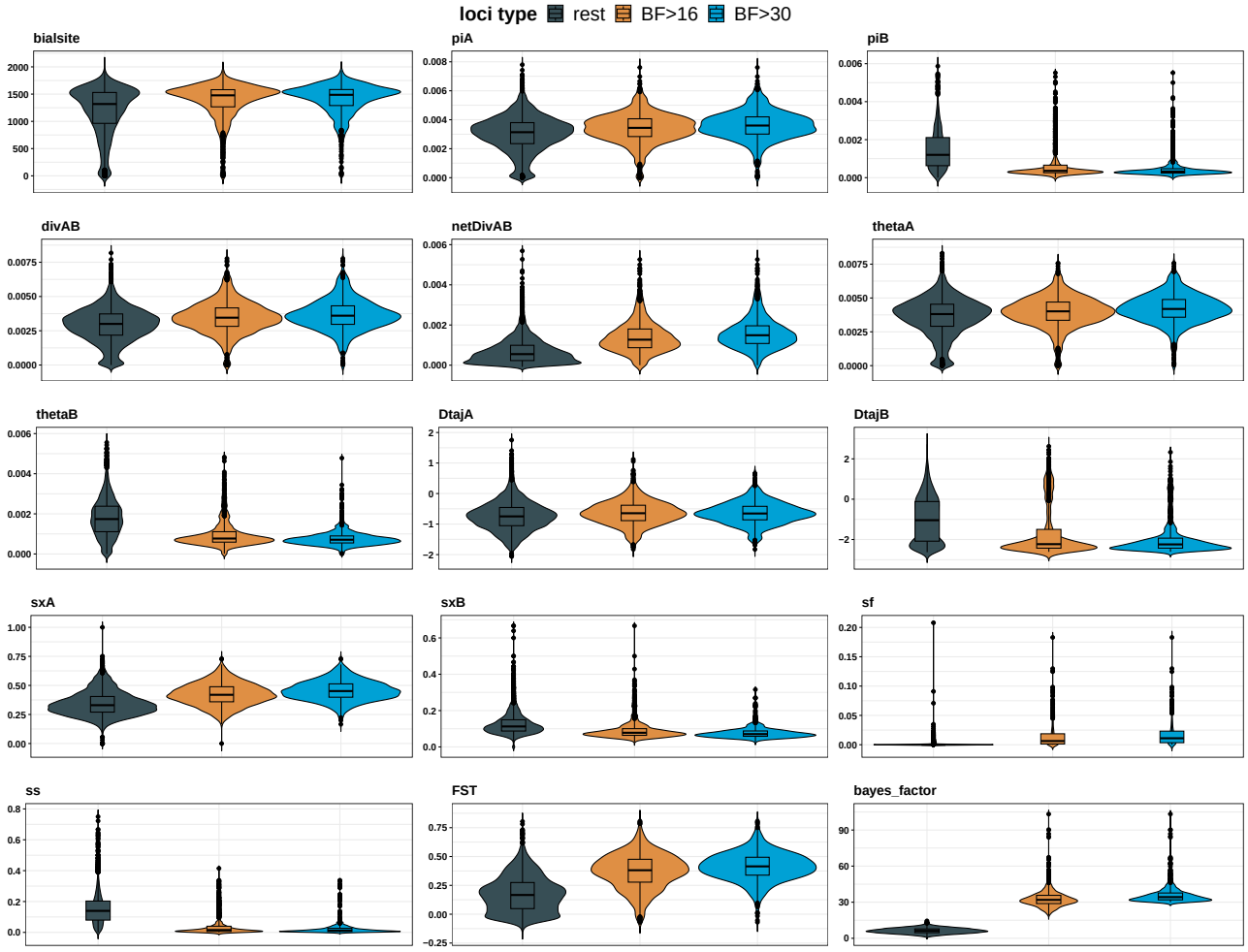

Figure B.7: Distribution of summary statistics values for foxtail millet loci in function of their BF level.

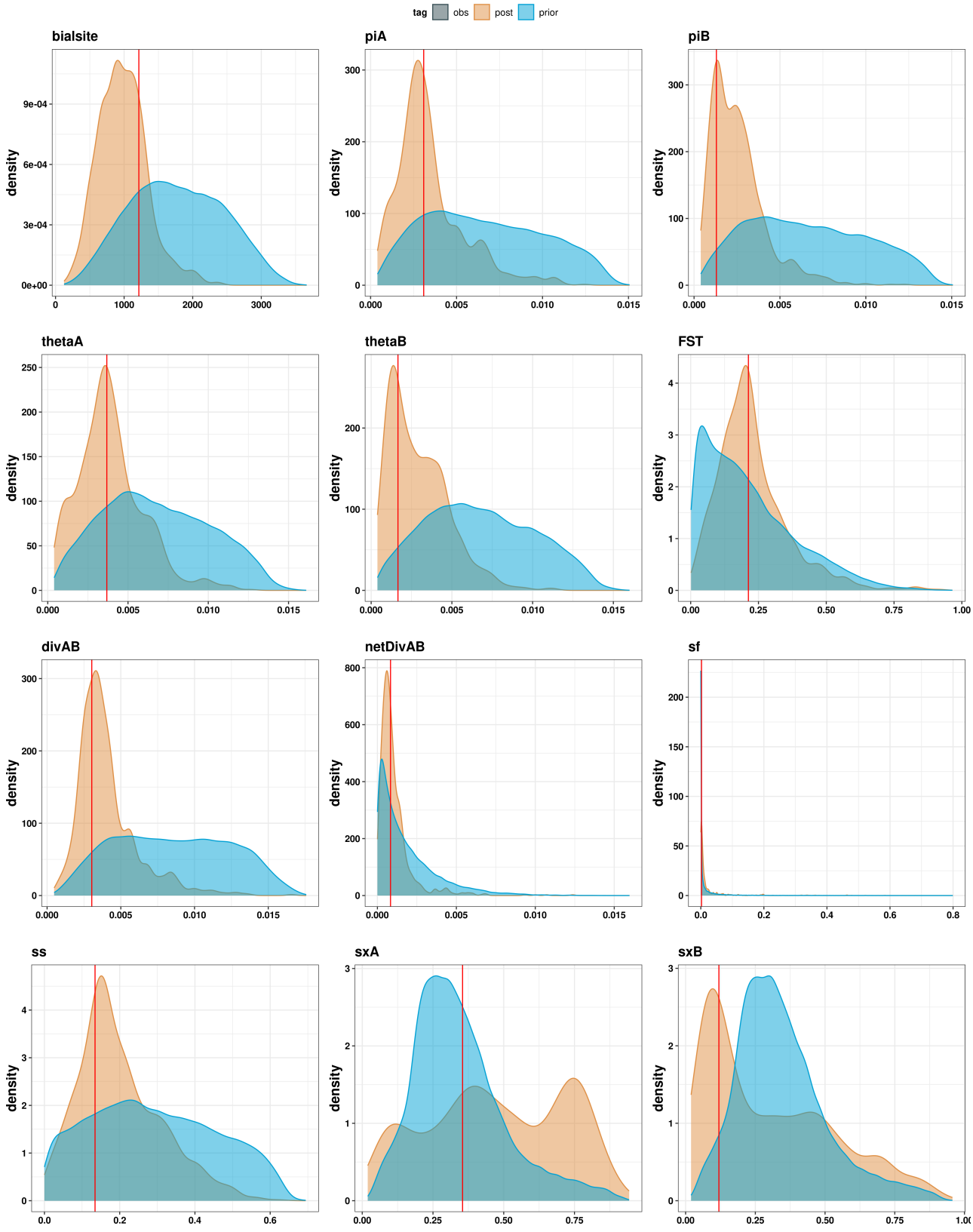

Figure B.8: Distribution of prior and posterior summary statistics values compared to observed values in *S.italica* (B)/*S.viridis* (A) data (red line).

Appendix C

Extended abstract (in french)

Les processus qui sous-tendent l'isolement reproductif entre des lignées divergentes sont essentiels pour comprendre la spéciation. Au fur et à mesure de leur évolution, les populations développent progressivement un isolement reproductif (IR). L'IR est défini comme une réduction de la production d'hybrides et/ou de l'aptitude à la reproduction ou comme une réduction du flux génétique. Dans cette thèse, j'ai étudié l'IR d'un point de vue génomique et j'ai donc utilisé l'IR comme mesure quantitative de l'effet des divergences génétiques sur le flux de gènes (Westram et al. 2022). Les mécanismes d'IR sont classiquement classés en prézygotiques et post-zygotiques. L'isolement prézygotique englobe un éventail de mécanismes influençant la probabilité de formation d'un zygote, comme l'isolement écologique, qui se produit lorsque les deux parents potentiels ne peuvent pas se rencontrer parce qu'ils vivent dans des habitats différents, ce qui peut résulter d'une adaptation locale à des conditions écologiques distinctes. L'isolement post-zygotique diminue la viabilité ou la fertilité des hybrides qui en résultent. Par exemple, dans les groupes Drosophila melanogaster et D. virilis - deux espèces étroitement apparentées - la divergence des protéines gonadiques, en particulier des protéines du tractus reproducteur mâle, est étroitement associée à la stérilité des mâles hybrides F1 (Coyne et Orr 1989).

En dehors de ces deux types, les distordeurs de ségrégation (SD) sont des éléments génomiques qui induisent une distorsion de la ségrégation mendélienne, entraînant la transmission préférentielle des allèles SD dans la descendance d'un hétérozygote, un phénomène appelé entraînement méiotique. En conséquence, les SD sont surreprésentés dans les gamètes viables, ce qui conduit finalement à la fixation des SD dans la population. Au-delà de la classification basée sur la formation du zygote, l'IR peut être classé en fonction de son origine, soit intrinsèque, soit extrinsèque. intrinsèque ou extrinsèque. L'IR intrinsèque est indépendant des influences environnementales, tandis que l'IR extrinsèque dépend de l'interaction entre les deux. l'IR extrinsèque dépend de l'interaction entre le génome et l'environnement (interaction génotype-environnement). environnement).

Pour comprendre la spéciation, il est essentiel de saisir le processus par lequel cet isolement se développe. En d'autres termes, comment un allèle, comme chez la drosophile, peut émerger et se propager, même s'il entraîne une réduction significative des capacités alimentaires

des individus hétérozygotes (hybrides), diminuant ainsi substantiellement leurs chances de se reproduire. Plusieurs décennies de recherche scientifique ont été consacrées à la question de l'établissement et du maintien de l'IR. Au-delà de la compréhension du mécanisme de formation et de préservation de cet isolement, les modèles théoriques permettent également de formuler des hypothèses testables. Ces hypothèses guident ensuite la recherche empirique, fournissant une approche systématique pour explorer et comprendre les mécanismes sous-jacents de l'évolution et de la spéciation. Bateson, puis Dobzhansky et Muller ont chacun proposé un modèle alternatif (appelé BDMI pour Bateson-Dobzhansky-Muller Incompatibility) qui résout le problème du franchissement des vallées d'aptitude (il s'agit de la situation paradoxale où la réduction de l'aptitude des hétérozygotes est nécessaire pour établir l'IR, mais en même temps, la réduction de l'aptitude des hétérozygotes empêche la population de diverger car les hybrides ne pourraient pas survivre) en considérant un modèle à deux locus et à deux allèles (Orr 1996). Il utilise une interaction d'épistasie négative à deux locus pour permettre l'émergence de l'IR sans croiser la vallée de l'aptitude. Un autre modèle courant utilise l'adaptation locale entre deux niches/environnements pour déclencher l'IR (appelé spéciation écologique) et/ou maintenir l'IR en sélectionnant des hybrides qui ne s'adaptent à aucun environnement.

La spéciation est un processus dynamique qui s'étend dans le temps et dans le génome. Au début du processus de spéciation, la divergence des populations se produit au niveau d'un petit nombre de loci responsables de la divergence. de loci responsables de la divergence. En présence d'un flux génétique, les progrès vers la spéciation sont rapidement érodés. Dans ces conditions, les loci barrières doivent être sélectionnés à une force qui contrecarre la migration. L'établissement d'haplotypes portant des allèles adaptés localement et des loci d'isolement (Schilling et al. 2018) confère un avantage considérable, car il permet la divergence adaptative et la spéciation même en présence de taux de migration élevés (Schluter et Rieseberg 2022). Cela peut se produire soit par le biais de la sélection liée, où plusieurs gènes peuvent faire de l'auto-stop autour des loci de la barrière initiale, soit par le biais de la sélection des suppresseurs de recombinaison. Dans les deux cas, le regroupement des gènes se traduit par un large signal génomique, appelé « îlots génomiques de divergence ».

La nature graduelle de la spéciation a été observée dans de nombreuses paires d'espèces montrant une corrélation entre l'IR et la divergence génétique (Coyne et Orr 1989 ; Presgraves 2002 ; Roux 2016). De cette observation découle la question suivante : à quelle vitesse l'IR s'accumule-t-il ? Dans ses travaux théoriques théorique, Orr (1995) a conclu que l'IR postzygotique devrait augmenter à un rythme plus rapide que la ligne, c'est-à-dire comme un processus « boule de neige ». La nature « boule de neige » du processus d'IR dans le temps est toujours débattue, certaines études n'ayant pas apporté de preuves de financement (Presgraves 2002). n'ont pas apporté de preuves de financement (Presgraves 2002 ; Stelkens et al. 2010 ; Price et Bouvier 2002), tandis que d'autres d'autres soutiennent la théorie comme chez la drosophile (Matute et al. 2010) et chez les espèces Solanum (Moyle et Nakazato 2010).

Un objectif central de la recherche sur la spéciation est de comprendre les mécanismes génétiques et génomiques à l'œuvre dans l'émergence et le maintien de l'IR. Pour ce faire, il faut

identifier les locus de barrière de flux génétique et comparer les résultats de plusieurs paires de lignées divergentes afin de saisir les événements séquentiels qui contribuent à l'établissement et au maintien des barrières reproductives. Traditionnellement, la détection des barrières repose sur des analyses de locus de traits quantitatifs (QTL) et des évaluations fonctionnelles. Toutefois, les progrès des données génétiques sur l'ensemble des populations ont ouvert la voie à des approches de balayage du génome moins coûteuses et plus faciles à mettre en œuvre, qui permettent d'étudier un spectre beaucoup plus large de populations/paires d'espèces. Pour étudier efficacement les déterminants génétiques de l'IR à l'aide de données de population à l'échelle du génome, il est essentiel de disposer d'attentes théoriques concernant les signatures génomiques des loci barrières, ce qui permet de les détecter.

On s'attend à ce que les loci barrières génèrent des RI, diminuant ainsi le flux de gènes au niveau du locus entre les populations. En l'absence de flux génétique, les loci barrières n'exercent pas d'influence locale sur la divergence évolutive, car l'isolement géographique entrave déjà le flux génétique sur l'ensemble du génome. En présence d'un flux de gènes, les loci barrières contribuent à façonner le paysage de la différenciation et de la divergence en faisant obstacle au flux de gènes dans leur voisinage, ce qui entraîne une évolution distincte des séquences génétiques entre les populations au niveau de ces loci. En raison de cette évolution indépendante, des mutations indépendantes apparaissent au fil du temps, et l'adaptation locale peut favoriser davantage des allèles spécifiques, ce qui entraîne une divergence et une différenciation accrues (Hejase et al. 2020 ; Sakamoto et Innan 2019). On prévoit donc que les loci barrières induisent une escalade de la divergence nette et potentiellement une augmentation de Dxy . Cependant, comme la Dxy dépend également de la diversité locale, cette augmentation peut être masquée par des variations locales de la diversité. En présence d'un flux génétique, la diversité de chaque population peut être enrichie par les migrants. Inversement, les loci barrières, qui ne subissent pas les effets du flux génétique, ont tendance à présenter une diversité plus faible que le reste du génome.

Malheureusement, d'autres processus peuvent générer des signatures confondantes similaires. Parmi les facteurs affectant la détection, nous pouvons distinguer deux groupes : l'un agissant au niveau local, qui imite partiellement le modèle génomique des loci de barrière (comme la réduction locale de la taille de la population), et l'autre au niveau du génome entier (comme un temps de divergence récent), qui diminue la différence entre les loci de barrière et le reste du génome, réduisant ainsi la puissance de détection. Il existe une variété d'approches proposant de détecter les barrières au flux génétique à partir des modèles génomiques à l'échelle du génome. Elles peuvent être classées en deux groupes (Tenaillon et Tiffin, 2008) : i) les méthodes basées sur les données impliquent la construction empirique de distributions nulles à partir d'une ou plusieurs statistiques obtenues à partir de scans génomiques et reposent sur des seuils arbitraires pour détecter les valeurs aberrantes ; ii) les méthodes basées sur des modèles impliquent l'inférence d'un modèle démographique (soit au préalable, soit simultanément) pour établir un modèle nul, suivi de l'identification des valeurs aberrantes correspondant aux loci de barrière en fonction de ce modèle. La démographie est incorporée pour atténuer les effets

confondants.

Comprendre les mécanismes génétiques sous-jacents à l'isolement reproductif est un objectif principal de la recherche sur la spéciation. L'analyse des populations divergentes est une approche courante, mais capturer la séquence des événements qui mènent aux barrières reproductives reste un défi. Une avenue prometteuse consiste à comparer des populations à différents niveaux de divergence temporelle et/ou spatiale, y compris celles récemment divergentes. Pour y parvenir, il est nécessaire de disposer d'un cadre comparatif capable de détecter les barrières au flux génétique à différents stades évolutifs à travers divers systèmes biologiques, indépendamment de leur histoire démographique. La méthode introduite, RIDGE (Détection de l'Isolement Reproductif utilisant les Polymorphismes Génomiques), vise à répondre à ce besoin.

RIDGE prend en entrée un fichier vcf contenant les séquences d'individus provenant de deux populations, accompagné de fichiers accessoires fournissant des informations complémentaires. À partir de cela, RIDGE utilise d'abord la méthode de l'Approximate Bayesian Computation (ABC) pour inférer des données démographiques en simulant 14 modèles démographiques x génomiques afin de produire une table de référence. Cette table sert à entraîner une forêt aléatoire (Random Forest, RF) qui génère des poids et des estimations de paramètres pour chaque modèle en fonction de leur adéquation avec l'ensemble de données cible (observé).

Ensuite, RIDGE construit un hypermodèle où la distribution postérieure de chaque paramètre est obtenue comme la moyenne pondérée sur les 14 modèles. Enfin, il utilise cet hypermodèle pour simuler un ensemble de loci de contrôle (ci-après non-barrière) et un ensemble de loci de barrière qui n'ont subi aucun flux génétique pendant la divergence. Les ensembles de données simulés pour les loci de barrière et non-barrière sont utilisés pour entraîner une seconde forêt aléatoire qui génère des probabilités postérieures et des facteurs de Bayes associés pour chaque locus afin de déterminer s'il appartient à la catégorie des loci de barrière ou de non-barrière.

RIDGE repose sur une approche ABC qui offre une grande flexibilité, lui permettant d'explorer l'hétérogénéité génomique et d'incorporer des statistiques récapitulatives personnalisées. Nous avons également conçu une méthode pour générer des estimations de paramètres multidimensionnelles, dépassant le focus initial sur un seul paramètre de abcrf (Raynal et al. 2019). Cette amélioration permet à RIDGE de gérer efficacement les interdépendances entre les paramètres et d'augmenter la précision des estimations de paramètres.

Une autre amélioration introduite par RIDGE est l'incorporation des facteurs de Bayes, facilitant la comparaison des résultats. De plus, RIDGE modélise explicitement la variation du taux de migration, m, plutôt que le taux de migration à l'échelle de la population (4Ne m) comme dans DILS (Fraisse et al. 2021), ce qui aboutit à une détection beaucoup plus stricte des loci de barrière. Nous interprétons que, en fixant à la fois Ne et 4Ne m comme dans DILS, l'hétérogénéité de la migration, m, tend à être trop fréquemment inférée car cela permet de concilier les modèles observés pour différentes statistiques.

Une limitation de RIDGE est la nécessité de définir a priori la taille des fenêtres, un choix arbitraire qui peut poser des problèmes dans les comparaisons entre espèces. Une amélioration possible serait de définir la taille des fenêtres en fonction de la distance génétique plutôt que

de la distance physique lorsqu'une carte génétique est disponible. Alternativement, on pourrait utiliser des critères basés sur des topologies locales pour segmenter le génome en fenêtres, comme implémenté dans Saguaro, qui repose sur un modèle de chaîne de Markov cachée couplé à des algorithmes de reconnaissance et de classification de motifs non supervisés (Zamani et al. 2013). Nous avons testé RIDGE d'abord sur un ensemble de données simulées pour mesurer les performances de cet outil sous divers scénarios. Ensuite, nous l'avons appliqué à un ensemble de données empiriques, en commençant par un ensemble de données de corbeaux (provenant de Poelstra 2014 et Vijay 2016) pour lequel une barrière est bien décrite et la biologie sous-jacente bien connue.

Les ensembles de données simulées que nous avons explorés nous ont fourni des lignes directrices pour les conditions où RIDGE peut fournir des résultats utiles et précis. Nous suggérons d'utiliser des ensembles de données avec une densité de SNP supérieure à 0,1 %, comme dans les corbeaux et les ensembles de données simulées, où la densité de SNP était d'environ 1 %. Nous conseillons également d'utiliser un minimum de trois échantillons par population. Les statistiques de bonté d'ajustement permettent aux utilisateurs de vérifier la qualité des inférences effectuées. Si Gpost < 5 %, l'utilisateur doit vérifier les bornes des prioris. Les lignes directrices pour interpréter et définir les seuils des facteurs de Bayes (BF) dépendent des objectifs de l'utilisateur.

Si RIDGE est utilisé uniquement pour découvrir de nouveaux gènes candidats impliqués dans les barrières au flux génétique pour une paire de populations spécifique, nous recommandons d'utiliser un seuil personnalisé qui capture de manière optimale les valeurs aberrantes du facteur de Bayes. À des fins de comparaison, il est recommandé d'utiliser un facteur de Bayes standard BF > 50 ou 100, ou de conserver le nombre de loci aberrants correspondant à la proportion de barrières estimée dans la première étape de RIDGE ($\hat{Q}$). Il est également important de considérer la distribution globale du facteur de Bayes (ou de la probabilité postérieure) pour aider à interpréter les résultats. Par exemple, sous le modèle SI (avec une divergence suffisante), tous les loci ou une grande proportion de loci apparaissent comme des barrières, mais la distribution globale est unimodale, ce qui contraste nettement avec un modèle IM avec des barrières, qui présente une distribution clairement bimodale (Figure A.12).

Il est crucial de noter que les données génomiques seules ne peuvent fournir des preuves concluantes des loci de barrière et que les résultats de RIDGE doivent être couplés à d'autres analyses telles que l'analyse fonctionnelle (Ravinet et al. 2017). Il est important de noter que la longueur de la fenêtre (par défaut réglée à 10 kb) peut affecter de manière significative les résultats de RIDGE. Elle doit être déterminée en fonction de l'étendue du déséquilibre de liaison ainsi que du niveau de diversité, car elle détermine la quantité de polymorphisme et affecte donc la force du signal.

Comme pour toutes les approches ABC, la qualité des priors fournis par l'utilisateur affecte les résultats obtenus avec RIDGE. Un Tsplit de 0.1 2Ne générations (10 000 générations dans nos simulations) semble être une limite inférieure pour à la fois les inférences démographiques (Figures 2.4 et 2.5) et les inférences de barrière (Figure 2.6), en dessous de laquelle RIDGE ne

parvient pas à capturer des signaux informatifs. RIDGE peut détecter des barrières au flux génétique sur des données simulées (Figure 2.6) et empiriques (Figure 2.7), à partir de 0.1 2Ne génération, ce qui représente un très faible niveau de divergence. Pour contextualiser, DILS a correctement inféré une barrière au flux génétique lorsque Tsplit $>$ 0.5 2Ne générations, tandis que gIMble n'a démontré son efficacité que sur une paire d'espèces d'Heliconius qui ont divergé il y a 4.5 millions de générations, estimé à représenter 0.49 2Ne générations (Martin et al. 2015).

Les approches comparatives ont été utiles pour comprendre les bases génomiques impliquées dans le processus d'isolement reproductif (par exemple, Roux et al. (2016)) et continueront de jouer un rôle important dans la recherche sur la spéciation. Par sa flexibilité et son cadre comparatif, RIDGE devrait devenir un outil utile pour suivre cette direction. Enfin, j'ai testé RIDGE sous des temps de divergence faibles en l'appliquant à des systèmes domestiqués, y compris le maïs et le millet , qui ont subi une sélection récente médiée par l'homme. Les deux systèmes ont divergé de leur parent sauvage le plus direct ("ancêtre") il y a environ 9000 ans. Cependant, ils ont des systèmes de reproduction différents ; le maïs est allogame tandis que le millet est autogame.

Les deux ensembles de données sont principalement structurés génétiquement par la différenciation entre les pools géniques sauvages et domestiques. Cependant, il semble que le pool génique sauvage du millet présente encore une structure génétique malgré les efforts pour l'éviter lors de l'échantillonnage génétique. Dans les travaux futurs, il faudra ré-analyser l'ensemble de données du millet après avoir éliminé les valeurs aberrantes du pool génique sauvage afin de vérifier si nos observations actuelles sont affectées par la structure de la population sauvage.

Les ensembles de données pour le maïs et le millet , en raison de leur divergence récente de leurs parents sauvages, présentent des conditions plus difficiles que l'ensemble de données des corbeaux. La détection de barrières en utilisant un seuil BF = 30 a montré des modèles génomiques presque identiques pour les barrières dans les deux cas, avec la même quantité de Mb détectée comme barrière (44,2 Mb pour le maïs et 49 Mb pour le millet), bien que le génome du maïs soit cinq fois plus grand que celui du millet .

Fait intéressant, tous les loci connus comme barrières entre le maïs et la téosinte proviennent de l'interaction maïs-mexicana. Le maïs est connu pour s'hybrider facilement avec parviglumis, ce qui peut entraîner moins de loci d'isolement reproductif (IR). Notre population sauvage se compose uniquement d'individus parviglumis. Néanmoins, nous observons une grande région d'IR à la position de Ga2 (Figure 4.8), indiquant que même entre le maïs et parviglumis, Ga2 agit comme une barrière. Quant à l'autre gène, il présente un niveau de BF d'environ 10 pour Ga1/Tcb1. De plus, un gène de domestication, Bt2, a été détecté comme une barrière (avec un BF = 110), montrant que la domestication peut jouer un rôle dans l'isolement reproductif, bien que ce soit une exception parmi les 11 gènes de domestication bien décrits chez le maïs qui expriment un BF très faible (BF 4). La domestication est une forme d'adaptation locale forte qui pourrait indirectement mener à l'isolement reproductif (IR). Nos résultats suggèrent qu'à l'exception d'un gène, les gènes de domestication ne sont pas impliqués dans l'IR. Nous avons également testé la présence de barrières génétiques dans les gènes de floraison, car ils sont des

candidats potentiels pour l'isolement reproductif par isolation temporelle. Nos résultats n'ont trouvé aucun gène de floraison dépassant BF = 30 (la valeur moyenne dans les régions des gènes de floraison est la même que la moyenne du génome).

Dans l'ensemble, l'autofécondation semble affecter les résultats lors de l'examen des inférences démographiques et des modèles génomiques de barrières. Par exemple, les barrières détectées dans le millet avec un facteur de Bayes de 16 représentent 15 %, ce qui est nettement plus élevé que dans le maïs. De plus, ces barrières sont séparées en deux groupes en utilisant l'ACP (Figure B.6). Ce résultat pourrait potentiellement s'expliquer par un flux génétique plus faible et un déséquilibre de liaison plus élevé dans l'autofécondation générant plus de barrières au flux génétique, ou par le fait que l'autofécondation viole de nombreuses hypothèses faites lors des simulations, entraînant ainsi des faux positifs potentiels. Le contraste entre l'île et le niveau de la mer dans le facteur de Bayes est plus faible pour le millet que pour le maïs (Figure 4.15). Cela pourrait être dû à la réduction globale du flux génétique induite par l'autofécondation couplée à une réduction de la recombinaison effective. Cependant, si tel était le cas, la distribution du BF ne devrait pas montrer une bimodalité claire et le paysage du BF devrait être plus plat. Alternativement, le contraste plus faible pourrait s'expliquer par un manque de puissance statistique. Les résultats sont plus difficiles à interpréter que ceux du maïs en raison de l'effet de l'autofécondation sur les aspects biologiques et statistiques.

L'objectif principal de cette thèse était de développer une méthode pour détecter les barrières au flux génétique pouvant être appliquée à divers systèmes biologiques. Je me suis basé sur un outil existant, DILS (Fraïsse et al. 2021), et je l'ai modifié et étendu pour permettre une analyse comparative dans plusieurs contextes. Tout d'abord, j'ai amélioré la méthode en implémentant une moyenne des modèles au lieu d'une approche par le meilleur modèle. Cette modification a permis à RIDGE d'estimer avec précision les modèles démographiques dans une large gamme de conditions, à la fois sur des ensembles de données simulées et empiriques, rendant possibles les comparaisons entre ensembles de données même lorsque les meilleurs modèles diffèrent. Ensuite, j'ai affiné l'estimation des paramètres liés à la proportion de barrières au flux génétique dans les génomes, en introduisant de nouvelles métriques pour les valeurs aberrantes de divergence, de différenciation et de diversité. En particulier, cela a amélioré l'estimation dans des conditions difficiles telles que des temps de divergence faibles et/ou des taux de migration faibles.

Enfin, RIDGE est capable de détecter des barrières sur des ensembles de données simulées même à des temps de divergence très faibles (Tsplit $<$ 0.2Ne), mais aussi sur des ensembles de données réels comme observé pour les corbeaux (pour un temps de divergence récent) et le maïs (très récent) pour lesquels nous avons réussi à détecter des loci de barrière bien identifiés dans la littérature (RSG9, LRP5, PRKCA et CACNG1&4 pour les corbeaux et Ga2 pour le maïs). Comme observé dans le cas du millet , le système de reproduction par autofécondation semble réduire la capacité de RIDGE à distinguer les barrières.

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

Université de Rennes

Titre : Approche génomique de la détection des barrières au flux de gènes

Mots clés : spéciation, barrières au flux de gènes, ABC, machine learning

Résumé : La caractérisation des mécanismes qui sous-tendent l'isolement reproductif entre des lignées divergentes est essentielle pour comprendre le processus de spéciation. Au cours de leur évolution, les populations développent progressivement un isolement reproductif (IR) en passant par des étapes intermédiaires, souvent appelées "zone grise de la spéciation". L'établissement de l'IR se manifeste par l'apparition de régions génomiques qui agissent comme des barrières réduisant le flux de gènes local par rapport au reste du génome. Les approches de génomique des populations impliquent donc l'identification de locus avec des signatures spécifiques, différentes du reste du génome. Cependant, d'autres processus peuvent créer des signatures similaires, ce qui fait de la détection des barrières une tâche difficile. Dans ma thèse, j'ai développé un nouvel outil, RIDGE - Reproductive Isolation Detection using Genomic Polymorphisms – un nouvel outil libre et portable adapté en particulier aux approches comparatives. RIDGE utilise une approche ABC (Approximate Bayesian Computation) et de “model averaging” basée sur des “random forest” pour prendre en compte divers scénarios de divergence entre lignées. Il prend en compte l'hétérogénéité du taux de migration, de la sélection en liaison et de la recombinaison le long du génome, estimant la proportion de barrières et effectuant des tests par locus pour détecter les barrières au flux génique. Des simulations et des analyses de jeux de données publiés sur des paires d'espèces de corbeaux indiquent que RIDGE est efficace pour détecter la migration en cours et identifier les locus barrières, même pour des temps de divergence récents. De plus, la contribution des statistiques résumées varie en fonction du jeux de données, ce qui met en évidence la complexité des signaux génomiques des barrières et l’intérêt de combiner plusieurs statistiques résumées. Par la suite, j'ai appliqué RIDGE à des paires de populations sauvages/domestiques : le maïs (allogame) et le millet (autogame), les deux ayant été domestiquées il y a environ 9 000 ans. Des flux de gènes entre les formes ont été documentés dans ces deux systèmes. Les modèles avec migration continue au cours du temps et hétérogène le long du génome sont clairement ressortis comme dominants. RIDGE a également démontré sa capacité à distinguer les locus barrière des locus de domestication (qui ont subi des balayages sélectifs au sein des formes domestiques). Les perspectives de ce travail comprennent l'application de RIDGE à de multiples paires population/espèce englobant un large spectre de divergence afin de déterminer les bases génomiques de l’IR au cours de la spéciation, de tester la théorie de «l’effet boule de neige” formulée par Orr en 1995 ou de déterminer la nature des gènes de spéciation.

Title : Genomic Approach to Detecting Barriers to Gene Flow

Keywords : speciation, gene flow barrier, ABC, machine learning

Abstract : Characterizing the mechanisms that underlie reproductive isolation between diverging lineages is central in understanding the speciation process. As populations evolve, they gradually develop reproductive isolation (RI) by passing through intermediate steps, often referred to as the "gray zone of speciation". This isolation is marked by the emergence of genomic regions acting as barriers to local gene flow, distinct from the rest of the genome. Detecting these barrier loci involves identifying outlier loci with specific signatures. However, other processes can create similar patterns, which challenges barrier loci detection. In my thesis, I developed a new tool, RIDGE - Reproductive Isolation Detection using Genomic Polymorphisms, a novel free and portable tool tailored for this purpose in a comparative framework. RIDGE utilizes an Approximate Bayesian Computation model-averaging approach based on a random forest to accommodate diverse scenarios of lineage divergence. It considers heterogeneity in migration rate, linked selection, and recombination, estimates barrier proportion and conducts locus-scale tests for gene flow barriers. Simulations and analyses of published datasets in crow species pairs demonstrate RIDGE's efficacy in detecting ongoing migration and identifying barrier loci, even for recent divergence times. Furthermore, the contribution of summary statistics varies depending on the dataset, highlighting the complexity of gene flow barrier genomic signals and the interest of combining several statistics. Subsequently, I applied RIDGE to wild/domestic pairs in maize (an outcrosser), and foxtail millet (a selfer), both domesticated around 9,000 years ago. Gene flow between forms has been reported in these two systems. Consistently, models with ongoing migration and heterogeneity in migration rate were clearly dominant over other models. RIDGE also demonstrated its ability to distinguish between barrier loci and domestication loci (that experienced selective sweeps within the domestic forms). The perspectives of this work include applying RIDGE to multiple population/species pairs encompassing a large spectrum of divergence to determine the genomic pattern of RI during speciation, to test the snowball theory formulated by Orr in 1995 or to determine the nature of speciation genes.