Université de Lille

GHENT UNIVERSITY

Evo-Eco-Paléo

Speciation dynamics: contrasts between plants and animals

Dynamiques de speciation: contrastes entre plantes et animaux

Thèse de doctorat de l'Université de Lille préparée à l'EEP
en cotutelle avec l'Université de Gand

École doctorale n°104 Science de la Matière, du Rayonnement et de l'Environnement (SMRE)

Thèse soutenue à Villeneuve d'Ascq, le 20 Décembre 2023, par

François Monnet

pour obtenir le grade de Docteur en **Science: Bioinformatics** (Ghent university) et
Biologie de l'environnement, des organismes, des populations, écologie (Université de Lille)

Composition du Jury :

Maud Tenaillon Directrice de Recherche CNRS, Université Paris-Saclay	Rapportrice
Tatiana Giraud Directrice de Recherche CNRS, Université Paris-Saclay	Rapportrice
Myriam Valero Directrice de Recherche CNRS, Sorbonne-Université	Présidente
Zhen Li professeur assistant, Ghent University	Examinateur
Nicolas Bierne Directeur de Recherche CNRS, Université de Montpellier	Examinateur
Xavier Vekemans Professeur, Université de Lille	Co-directeur de thèse
Yves Van de Peer Professeur, Ghent University	Co-directeur de thèse
Camille Roux Chargé de Recherche CNRS, Université de Lille	Co-encadrant de thèse

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

I-SITE UNIVERSITÉ LILLE NORD-EUROPE

EDSMRE

Thèse de doctorat

Contents

List of figures

List of tables

List of code snippets

Abstract

Speciation, the process leading to the emergence of reproductively isolated species through the accumulation of genetic reproductive barriers, has been a subject of study since *the origin of species* and remains an active topic of research. One primary goal of these studies is to elucidate which microevolutionary processes shape the dynamics of speciation. In this thesis, we introduced a novel comparative approach aimed at disentangling the effect of several speciation-related factors. This approach is illustrated by an investigation tackling an historical assumption: the supposed faster speciation of animals in contrast to plants. When comparing the dynamics of speciation between plants and animals, we observed that complete reproductive isolation occurred, on average, at a lower level of divergence for plants. We further analysed the dynamics of speciation in plants using linear modelling but did not find any significant effects for the two factors tested: selfing rate and life form. Overall, these results highlight the potential of our novel comparative approach to conduct easy, rapid and flexible comparisons of speciation dynamics in future research.

Introduction

The diversity of living organisms on the broadest scale is the result of a major evolutionary process: speciation. Throughout the 20th century, this process was represented as a succession of dichotomous splits of evolutionary lineages, notably through phylogenetic trees. While such a dichotomous vision respects the evolutionary relationships between living genera, its relevance gradually diminishes when the level of divergence of the lineages being compared also decreases. Whether there is little ambiguity in distinguishing the highest groups in the hierarchy of divergence levels (e.g. Archaea *vs* Bacteria *vs* Protists *vs* Plants *vs* Fungi *vs* Animals), it becomes more complicated to assign a phylogenetic status between two recently separated gene pools (Hahn and Nakhleh, 2016). Speciation is therefore a continuous process over time that gradually leads to the formation of discrete entities called species. This discretisation is the end product of successive accumulations in the genomes of mutations that have deleterious effects on the reproductive success of hybrids. Studying speciation therefore means trying to understand the evolution of these barriers (DMI for Dobzhansky and Muller incompatibilities), which mainly involve the following questions:

- What is a barrier between two species?
- How do barriers accumulate in the genomes of two species?
- What is the dynamic of barrier accumulation: snowball or not?
- In what historical context did the divergence of a pair of species take place?

What is a barrier between two species?

When Darwin proposed his theory of evolution (Darwin, 1859), the main mechanism responsible for temporal changes was natural selection. For a given environment, the different phenotypic states that an inheritable biological trait can assume have different effects on the contribution of that trait to the fitness of individuals. Consequently, the phenotypic optimum of a trait is likely to vary spatially in a heterogeneous environment. In this conceptual framework, two gene pools derived from an ancestral pool and placed in contrasting environments will see their phenotypes diverge to reach their respective optima. Individuals with intermediate phenotypes, such as hybrids, will thus be selected against.

Although "*On the Origin of Species by Means of Natural Selection*" deals with the evolution of differences between populations, the process of speciation is not directly addressed. It was not until the 1930s that we began to identify which of the factors that differentiate two species are involved in reducing the fitness of hybrid individuals. In order to apply a QTL (quantitative trait locus) approach to the search for the determinants of barriers between species, Dobzhansky (T. H. Dobzhansky, 1936) crossed *Drosophila pseudoobscura* and *D. persimilis*, each of whose chromosomes can be easily traced by the morphological markers they carry. In F1 hybrids, the males are sterile and the females fertile. By backcrossing fertile females into one of the parental genomic backgrounds and then measuring the reproductive value of the introgressed individuals, Dobzhansky was able to produce the first study to show that certain chromosomal regions contribute more to the sterility of hybrids than other regions, thus demonstrating the genomic nature of the barriers between species. Barrier loci were initially seen as loci where the parental species had *AA* and *aa* genotypes respectively, and whose selective values were higher than those of heterozygous *Aa* hybrids. But how can we explain the divergence observed at such a locus? Since the evolutionary path linking *AA* and *aa* necessarily passes through a valley of low selective value (Fig. 1), natural selection prevents the evolution of monolocus barriers.

The solution proposed by Dobzhansky was that the decrease in the reproductive value of hybrids compared with parents is the result of negative epistatic interactions involving at least two loci (Fig. 1). An ancestral population with a two-loci genotype aabb can evolve towards the

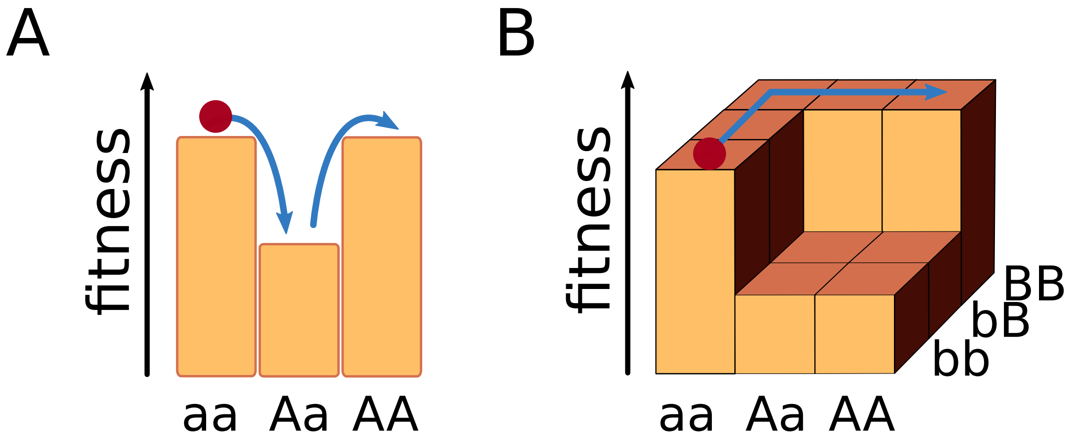


Figure 1: **Fitness landscapes of species barrier.**

Fitness of possible genotypes in the case of barriers involving either : (A) a single locus or (B) two interacting loci.

AABB genotype by following a ridge of high selective value (Fig. 1). Despite the deleterious effect of the interaction between the A and b alleles, it is possible to evolve initially from *aabb* to *aaBB*, then to fix the *A* allele in order to provide *AABB*. Hybrids with *AaBb* genotypes will have a lower reproductive value than the *aabb* and *AABB* parents (Fig. 1). This “Dobzhansky-Muller” model thus explains how two divergent populations can be separated by a fitness valley without having to cross it. Because the number of interacting loci can be greater than two, the Dobzhansky-Muller model is not restricted to simple epistasis and allows complex epistasis between more than two factors (Cabot et al., 1994).

How do barriers accumulate in the genomes of two species?

The “Dobzhansky-Muller”” model shows that barriers result from negative epistatic relationships that are expressed in a hybrid of two lineages that have already diverged. But this strictly verbal model makes no link between the molecular divergence of the lineages and the strength of postzygotic reproductive isolation. An initial series of experiments, mainly in *Drosophila* (but also in other lineages (Presgraves, 2002; Price and Bouvier, 2002; Sasa et al., 1998), sought to quantify the strength of postzygotic reproductive isolation by artificially generating hybrids from parental lines with different levels of divergence (Coyne and Orr, 1989, 1997, 1998). These experiments showed that there was a linear increase in the overall strength of isolation with molecular divergence. However, while the overall degree of postzygotic isolation measured is the end product of all the barrier loci, it says nothing about the number of these incompatibilities in the genomes, nor about the distribution of the effects of the different barriers, nor about the interactions between barriers. Nor do these results reveal the link between molecular divergence and the number of incompatibilities expressed in the hybrids, making it impossible to assert that the accumulation of barriers is also linear with molecular divergence. A second series of experiments therefore sought to unravel the genetic architecture of reproductive isolation using Dobzhansky's methodology, i.e. by measuring the selective value of individuals in which it had been possible to introgress chromosome segments from a more molecularly divergent line (Coyne and Orr, 1998; Elena and Lenski, 2001; N. A. Johnson, 2000; Ungerer et al., 2003). These random introgressions of different chromosomal fragments into a different genomic background revealed two main patterns of barrier evolution. The first pattern observed is that the barriers are predominantly asymmetric. The introgression of a chromosomal fragment from the genome of species A to that of species B will not have the same effect on the selective value of the hybrid as the opposite introgression. This pattern is interpreted as being the direct result of the evolution of barriers. In a two-locus DMI evolutionary model (called simple DMI), if the *A* and *b* alleles are incompatible, then the symmetrical incompatibility between a and B would block all possible evolutionary trajectories linking the ancestral *aabb* genotype to the derived *AABB* genotype, thus preventing the observation of *AABB* individuals in the wild. To exist, a DMI must allow at least one evolutionary trajectory linking *aabb* to *AABB*, implying an asymmetric barrier in the case of a two-locus DMI (Fig. 2).

The second observed pattern is that simple DMIs are less frequent than complex DMIs involving interactions between more than two loci (Fraïsse et al., 2014; Welch, 2004). In some cases, individual introgressions of *a*, *b* or *c* alleles into an *ABC* genome are not sufficient to significantly affect selective value, requiring at least introgressions at two loci to express a DMI (Cabot et al., 1994; Davis et al., 1994; Fraïsse et al., 2014; Orr and Turelli, 2001; Palopoli and Wu, 1994; Perez and Wu, 1995). A case of complex incompatibility has been well identified in *Drosophila pseudoobscura* where a specific combination of alleles at four loci reduces male

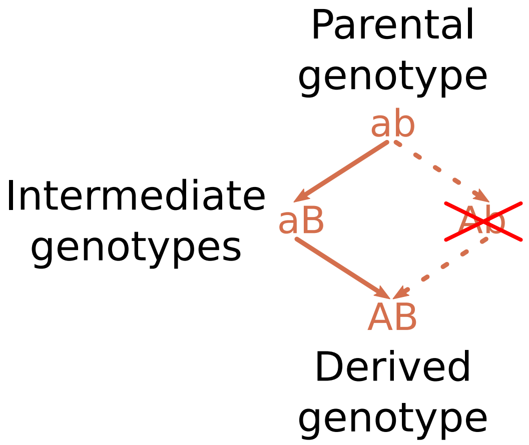


Figure 2: **Possible evolutionary paths linking the parental to derived genotypes: one locus model of DMI.**

Alleles *A* and *b* are incompatible in this example, whereas *a* and *B* are compatible: asymmetric barrier. A strong symmetrical barrier would prevent the existence of an *AB* genotype in nature, and therefore divergence at these loci, because all evolutionary paths leading to a DMI would be blocked.

fertility more strongly than the sum of the individual effects of each of these four loci (Orr and Irving, 2001).

One hypothesis to explain the preponderance in the data of cases of complex incompatibilities is that they are easier to evolve, blocking fewer evolutionary trajectories. In the previous example of a simple DMI, 50% of evolutionary trajectories are blocked by the negative epistasis between A and b (Fig. 2). In the example where the barriers are expressed by combining *A*, *B*, *c* and *d* in the same genomic background, only one-sixth of the 24 possible trajectories linking *abcd* and *ABCD* are blocked (Fig. 3).

An alternative but not exclusive hypothesis that could explain the preponderance of complex DMIs in crossing experiments is the importance of 1) the robustness and 2) the functional redundancies of metabolic cascades. The robustness of a metabolic cascade can be partly characterised by the degree of branching of the steps leading to its final product. Thus, for the same level of proteins involved in a cascade, a linear cascade (where each intermediate step requires the success of the previous one, Fig. 4-A) is more exposed to the deleterious effects of a DMI than a branched cascade (Fig. 4-B), the latter requiring on average more mutations than

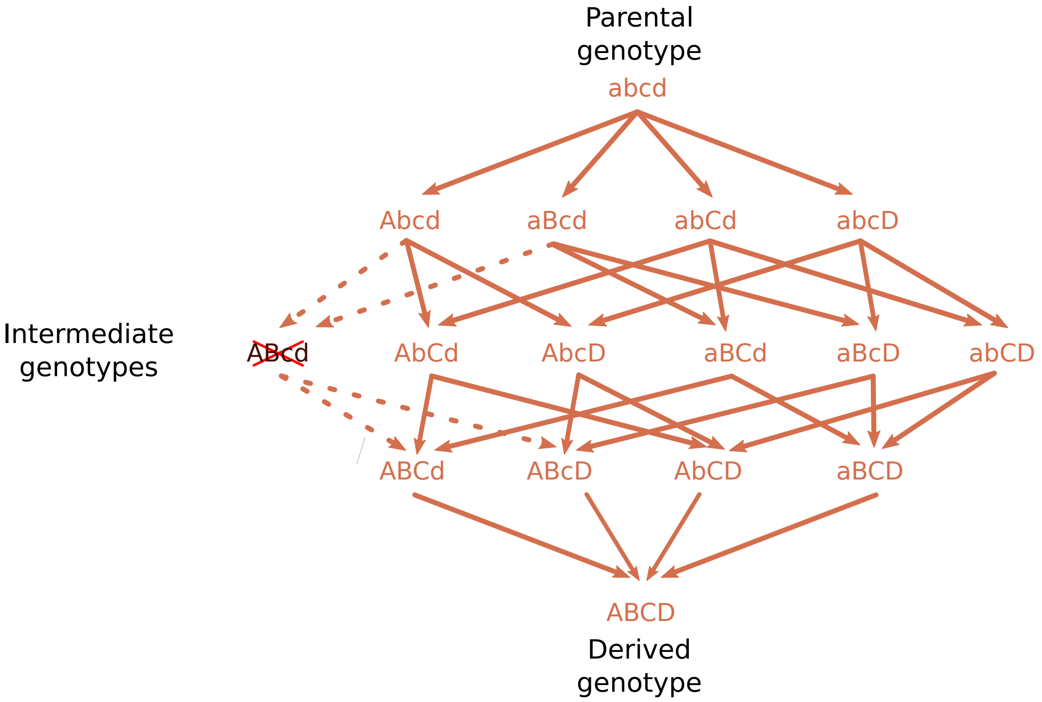


Figure 3: **Possible evolutionary paths linking the parental to derived genotypes: four loci model of DMI.**

24 possible evolutionary paths are linking the tetra-locus genotypes *abcd* and *ABCD*. The negative epistasis between alleles *A*, *B*, *c* and *d* blocks four of these evolutionary paths, making this complex DMI at 4 loci easier to evolve than simpler DMIs (Fig. 2).

the former to lead to a deletion of function.

While the Dobzhansky-Muller model describes what a barrier might be at the genetic level, it says nothing about the process of substitutions leading to incompatible alleles. Some DMI accumulation models make direct assumptions about the pattern of intra-lineage substitutions leading to incompatibilities. In these models, the substitutions responsible for DMIs occur successively through the action of epistatic positive selection within each isolated population (Kondrashov et al., 2002). This epistasis can be explained by the fact that the first substitution of an amino acid (called the “precursor”) in a lineage will condition the possible substitutions affecting a second site. The epistatically selected substitution at the second site can then determine the direction of selection acting at a third site, and so on throughout the divergence process. In this way, each substituted amino acid conditions the fate of subsequent mutations, with poten-

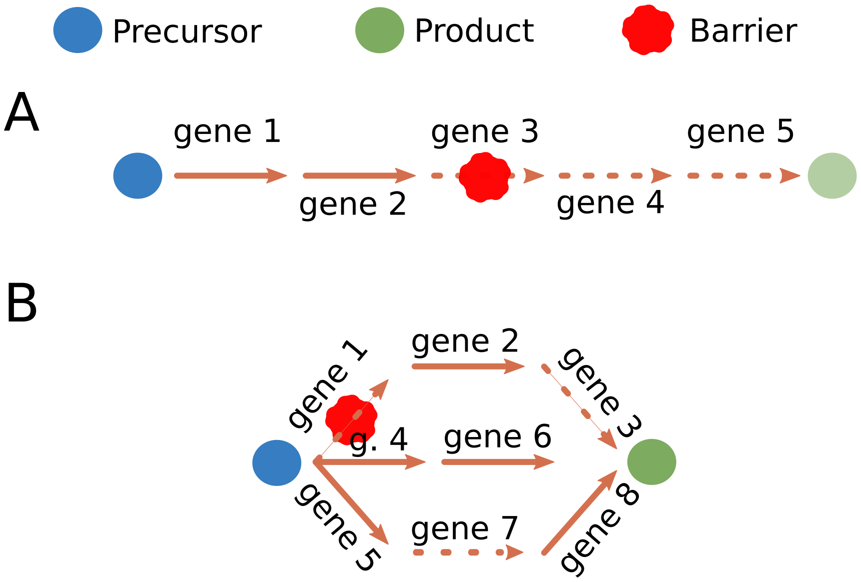


Figure 4: **Alternative geometries of metabolic cascades involving the same number of proteins or protein complexes.**

A: Linear cascade. A step negatively impacted by a DMI will short-circuit the synthesis of the final product.
B: Branched cascade. A greater number of mutations are expected to block all the pathways leading to synthesis of the final product.

tially significant consequences for the accumulation dynamics of DMIs. According to this model, the introgression of a fixed amino acid under the effect of positive epistasis towards a genome where the precursor is absent forms a DMI (Kondrashov et al., 2002). This DMI would be all the stronger as the positions in the adaptive landscapes of the two lineages become further apart in response to successive substitutions.

A concrete example of this co-evolution process are the multi-protein complexes involving numerous interactions at the interfaces of the peptides produced. Data measuring the robustness of multi-protein complexes in hybrids of highly divergent Saccharomyces lineages (*S. cerevisiae* and *S. kudriavzevii*) tested this hypothesis of accumulation of DMIs dependent on co-evolutionary processes (Leducq et al., 2013), and showed that these complexes are in fact robust to divergence. However, this type of study did not compare the effect of divergence on the robustness of several protein complexes involving different numbers of proteins.

What is the dynamic of barrier accumulation: snowball or not?

During a process of divergence, geographically isolated lineages will progressively accumulate substitutions, some of which will be involved in more or less complex DMIs for reasons described above. Certain hypotheses have been put forward to discuss the rate of accumulation of DMIs in genomes. Using the original Dobzhansky-Muller model and making certain assumptions, Orr (1995) formalises the theoretical number of DMIs as a function of the molecular divergence between two diverging lineages. For k substitutions accumulated between the two genomes, there would theoretically be $\frac{k.(k-1)}{2}$ interactions at two loci in an F1 hybrid that do not exist within the parental lines. This number of interactions can be approximated by $\frac{k^2}{2}$ for high values of k. Of these $\frac{k^2}{2}$ interactions in a hybrid, a proportion p will produce detectable deleterious effects. Thus, the expected number of DMIs at two loci in the genomes can be approximated by a square function of time: $E[I] = p.\frac{k^2}{2}$, where p is the probability that an interaction will reduce the selective value of the hybrid. More generally, in a case where DMIs are due to interactions between n loci, there would simply be $E[I_n] = p.\frac{k!}{n!(k-n)!}$ barriers to gene flow between the genomes. This acceleration in the accumulation of the number of barriers in the genomes during divergence has been called the 'snowball effect' by Orr (1995). While in this model the selective value depends directly on the number of allelic combinations at different loci in a hybrid, other models predict a linear increase in the number of DMIs over time. If the selective value is determined by the value at a trait according to Fisher's geometric model, then the snowball effect is not expected. Recent analyses have attempted to analyse the number of QTLs involved in sterility for pairs of species with different levels of divergence (Matute et al., 2010; Moyle and Nakazato, 2010). By statistically testing the linearity of the relationship, these authors propose to detect a non-linearity in the accumulation of DMIs over time (Fig. 5).

However, these studies have been heavily criticised for not taking into account the demographic history of the species in question (Städler et al., 2012). Part of the divergence measured between two current lineages comes from the polymorphism that was present in the ancestral population of these two lineages. DMIs do not segregate sufficiently within a population to con-

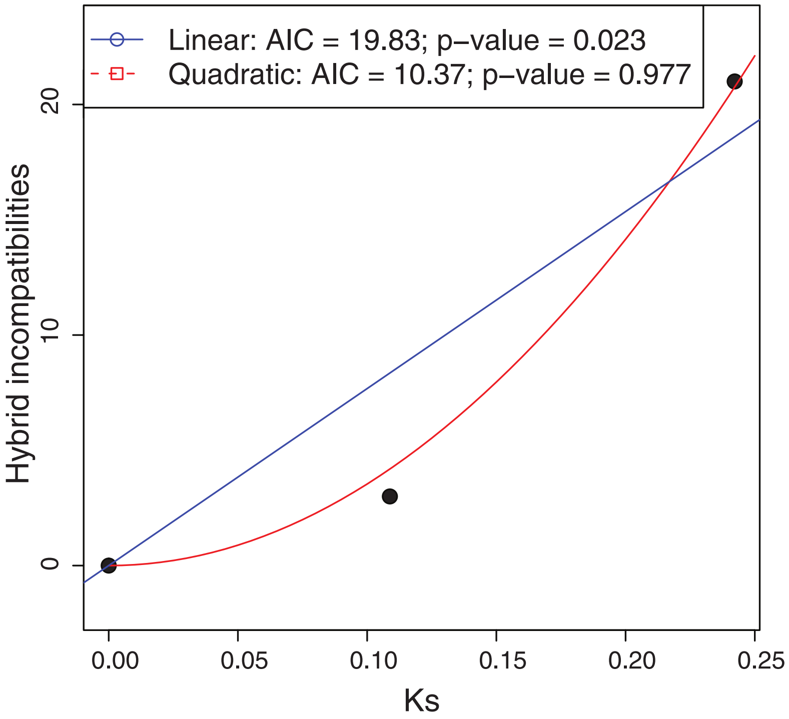


Figure 5: **Dynamic of accumulation of incompatibilities.**

Number of hybrid incompatibilities identified in *Drosophila melanogaster* × *D. santomea* and *D. melanogaster* × *D. simulans* hybrids from Matute et al. (2010). Due to the lack of points to test a quadratic versus linear relationship, the y-intercept was added as a pseudo-observed point. The authors use *Ks*, the raw synonymous divergence that neglects the effects of ancestral polymorphism containing no barriers. By subtracting this ancestral polymorphism, it becomes more difficult to reject a linear relationship (Städler et al., 2012).

tribute to its level of polymorphism, and therefore ultimately to the divergence between related lineages. A correct measure of divergence for testing the snowball effect must take account only of substitutions that have specifically occurred in sister lines, which has not yet been achieved.

This hypothesis of a snowball effect in speciation is also contradicted by other comparative genomics studies, but which focus on an even earlier evolutionary period (Kern and Kondrashov, 2004; Kondrashov et al., 2002). Kondrashov et al. (2002) looked at the proportion of mutations that are deleterious in humans but present in the wild in distantly related species. Exploring a continuum of non-synonymous divergence ranging from 5% to 50%, these authors found that an amino acid introgressed from another species has a probability of around 10% of expressing a DMI in humans, irrespective of the level of divergence (Fig. 6).

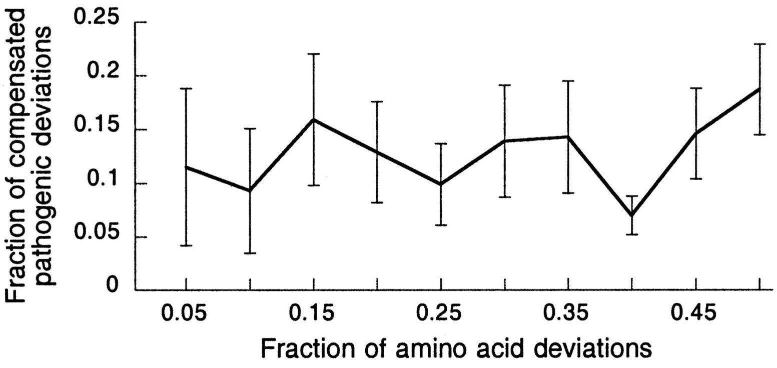


Figure 6: **Proportion of amino acid inducing DMI.**

Proportion of incompatible amino acids in a human genomic background but present in the wild in distantly related species, represented as a function of non-synonymous distance (Kern and Kondrashov, 2004; Kondrashov et al., 2002).

In what historical context did the divergence of a pair of species take place?

The processes described above concern the evolution and accumulation of barriers in geographically isolated lineages. With the reduction in costs associated with obtaining nucleotide sequences at high throughput, numerous studies have attempted to describe the effects of barrier genes on the genomic patterns of differentiation between populations in the presence of gene flow. By measuring the degree of genetic differentiation along genomes, it is theoretically possible to detect loci involved in reproductive isolation. Between two closely related species, the degree of differentiation is the result of global processes at the genomic level (such as the demographic history of each species or their reproductive system) and local selective effects at the level of the loci concerned (Lewontin and Krakauer, 1973). Thus, genes linked to isolation between two species should theoretically have higher levels of differentiation than the rest of the genome (Fig. 7). However, studies that have compared the levels of differentiation measured for several thousand markers in the genomes have shown that the highly differentiated genomic regions (also called 'genomic islands of differentiation') are too large to allow precise detection of the genes directly involved in reproductive isolation (Ellegren et al., 2012; Harr, 2006; Hohenlohe et al., 2012; Nadeau et al., 2012; Turner and Hahn, 2010; White et al., 2010; H. M. Wood et al., 2008) and that these islands can be explained by other evolutionary processes (Bierne et al., 2011; Cruickshank and Hahn, 2014).

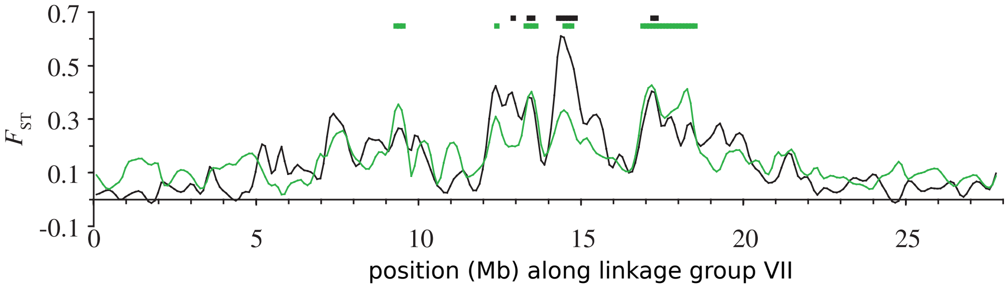


Figure 7: **Example of genomic islands.**

Genetic differentiation along linkage group VII in two pairs of sticklebacks measured by F_{ST}. Significantly more differentiated regions, containing candidate genes for reproductive isolation, are indicated above the curves by coloured bars, and form islands with widths greater than 100 kilobases. Figure from Hohenlohe et al. (2012)

The width of these islands of differentiation is strongly linked to the effects of recombination, whose rates themselves vary along the chromosomes (Navarro and Barton, 2003; Noor and Bennett, 2009). For example, it has been shown in the human and chicken genomes that there is a positive relationship between local recombination rate and genetic diversity along the chromosomes, but also a negative relationship between recombination rate and genetic divergence (Keinan and Reich, 2010; Mugal et al., 2013). Thus, regions with low recombination are both less polymorphic and more differentiated than regions with higher recombination rates. Despite this, some studies attempt to explain the size of islands of differentiation directly by the speciation process (Feder et al., 2012). According to the latter authors, at the start of the speciation process, the islets are loci linked to the local adaptations of each of the sister species, which continue to exchange gene flow. The rest of the genome would thus be weakly differentiated by the effect of migration, while a higher level of differentiation would be maintained at these loci by the effect of counter-selection of maladapted alleles from one species when they are placed in the environment of the other species. In this model, the progressive accumulation of adaptation genes would be accompanied by an increase in linkage disequilibrium between the selected alleles, leading to a decrease in the rate of recombination on a wider genomic scale, sometimes of several hundred kilobases (Via, 2012). The main problem with this model is that it attempts to verbally describe a speciation process by assuming that the demographic history of the species studied is known: the separation of an ancestral population into two daughter lineages that continue to exchange alleles through migration. However, the patterns described can be more easily achieved with an equally verbal demographic history scenario (Barton, 1979; Barton and Bengtsson, 1986; Endler, 1977). When the ancestral population splits into two geographically isolated populations, the two lineages that diverge in allopatry accumulate substitutions by drift as well as by selective effects. According to the Dobzhansky-Muller model, a proportion p of these k substitutions will be involved in reducing the selective value of the hybrids formed during an episode of secondary contact (Figure Demographic models.). Thus, the F_{ST} that was high throughout the genome before secondary contact will tend towards low values during secondary contact, with the exception of the regions genetically linked to the *p.k* barrier loci.

It therefore seems natural that a good understanding of the speciation process for a pair of species first requires correct inference of its demographic history, before proposing mechanisms linked to the evolution of the barriers observed. Recent progress has been made in this field with the appearance of tools for statistically evaluating the models described in figure 8 (Beaumont et al., 2002; Csilléry et al., 2010; Fagundes et al., 2007; Nielsen and Wakeley, 2001; Pritchard et al., 1999; Tavaré et al., 1997). However, two levels of complexity can lead to biases in demographic inferences. The first of these is that the selective effects that occurred independently in the two sister lineages after the ancestral separation will generate genomic heterogeneity in effective size, directly translated into heterogeneity in genetic differentiation, even for allopatric species Cruickshank and Hahn (2014). Current verbal models but also demographic inference tools (Hey and Nielsen, 2007) interpret this heterogeneity in differentiation (between two allopatric populations) as heterogeneity in introgression rate due to barrier loci. The second challenge is that introgression rates can actually vary across genomes and generate genomic heterogeneity in differentiation independently of variations in effective size (Roux et al., 2014; Roux et al., 2013). Only recently has it been possible to jointly infer the demographic scenario (Fig. 8) and potential genomic heterogeneities of effective size and introgression rate (Roux et al., 2016).

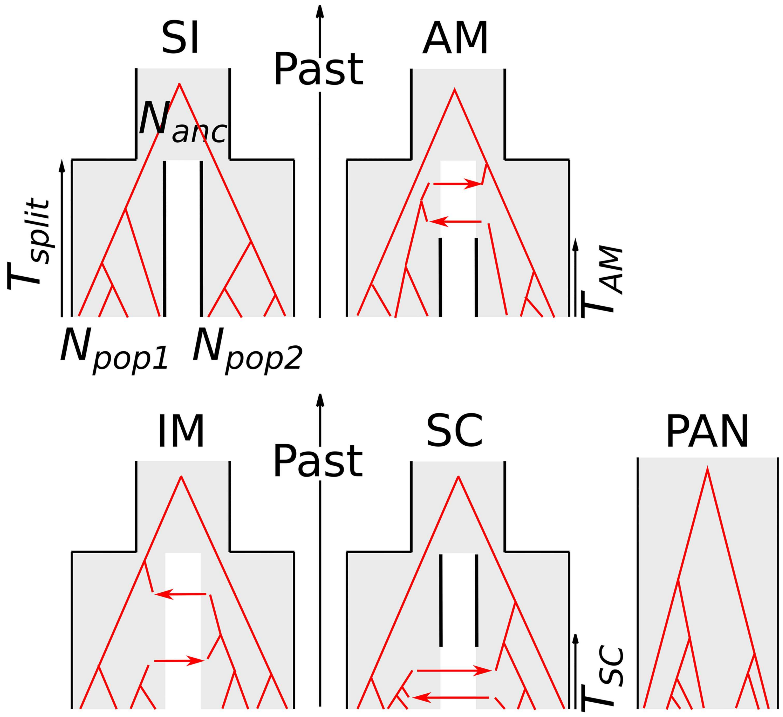


Figure 8: **Demographic models.**

Alternative speciation models from Roux et al. (2016). The SI model (for “Strict Isolation”) describes an ancestral population of size N_{anc}, panmictic, which subdivides into two daughter populations of constant size, and reproductively isolated from each other. The AM (for “Ancient Migration”) model is based on the SI model, with gene flow between the sister populations restricted to the first few generations after separation from the ancestral population. The IM model (for “Isolation with Migration”) incorporates continuous gene flow between the two populations from the time of separation to the present day. The SC model (for “Secondary Contact”) is based on the SI model, but where the daughter populations exchange alleles again during a secondary contact.

How do we study the dynamics of speciation ?

The completion of the speciation process can span millions of generations (Etienne et al., 2014), making it infeasible to observe the continuous accumulation of reproductive barriers, and thus RI, for a given speciation event. Therefore, we need to rely on ’snapshots’ of RI at different time points to reconstruct a comprehensive understanding of the dynamics of speciation (Stankowski and Ravinet, 2021). This can be achieved by analysing the RI of pairs of populations from speciation events of different ages, the latter usually approximated as the proportion of neutral genetic divergence. The estimation of the RI, however, is complex and multiple empirical methods have been developed to approximate it (reviewed in Westram et al., 2022). One of the most direct approaches are to perform crossing and transplanting experiments. Those experiments are designed to measure the fitness of hybrids and backcrossed progenies over several generations in controlled laboratory settings (or within natural environments), or the fitness of migrants and their descendants implanted in foreign populations. Different methods can be employed to calculate the RI based on various proxy of the hybrids' fitness, such as the rate of heterospecific matings or the proportion of hybrids seeds (for a review, see Sobel and Chen, 2014), according to the type of reproductive barrier of interest (e.g. mating preference, flowering time...). Westram et al. (2022) identify several possible concerns with these approaches. For instance, most of the crossing studies only account for the fitness of the first generation of hybrids. This can be illustrated with the study conducted by Coyne and Orr (1989) where RI was estimated by assessing the ratio of heterospecific matings to homospecific matings ($1 = \frac{freq(heterospecific\ matings)}{freq(homospecific\ matings)}$). Not only this approach provides no insights into the impact of post-zygotic barriers, but the reduction in fitness in the subsequent generations is not captured in the estimated RI. Another issue can be encountered in crossing studies while trying to avoid early generation hybrids, with samples being collected from locations far removed from an unimodal hybrid zone. These distant samplings introduce the risk of exposing incompatibilities (DMIs) that are not expressed in natural conditions. In the most simplest case of DMI, a derived allele *A* is favoured at a locus *A* and a derived allele *B* is favoured at a locus *B*. Alleles *A* and *B* are incompatible but they can rise in frequency in two different populations, as long as they are not exposed to each other (Fig. 9.A). This DMI should act as a reproductive barrier, as the low fitness of hybrids prevents gene flow between the two populations. However, the clines of the *A* and *B* alleles should repel each other on either side of the hybrid zone due to the counter-selection against *AB* individuals. Because of this repulsion effect, alleles *a* and *b* should make up the majority of the genotypes encountered near the hybrid zone, resolving the reproductive barrier by allowing a continuum of viable genotypes across the global cline (Westram et al., 2022; e.g. Hatfield et al., 1992; Virdee and Hewitt, 1994). By crossing individuals from samples distant from the hybrid zone, researchers might wrongly conclude the existence of reproductive barriers from incompatibilities that are not expressed in natural conditions (Fig. 9.B). Nonetheless, transplanting and crossing experiments remain useful to study the dynamic of speciation as they can provide measure of the RI at the genome scale. The combination of estimates of RI from different crossing experiments, associated with their respective level of divergence, provide a way to describe the dynamics of speciation (i.e. the accumulation of RI with time) (Stankowski and Ravinet, 2021).

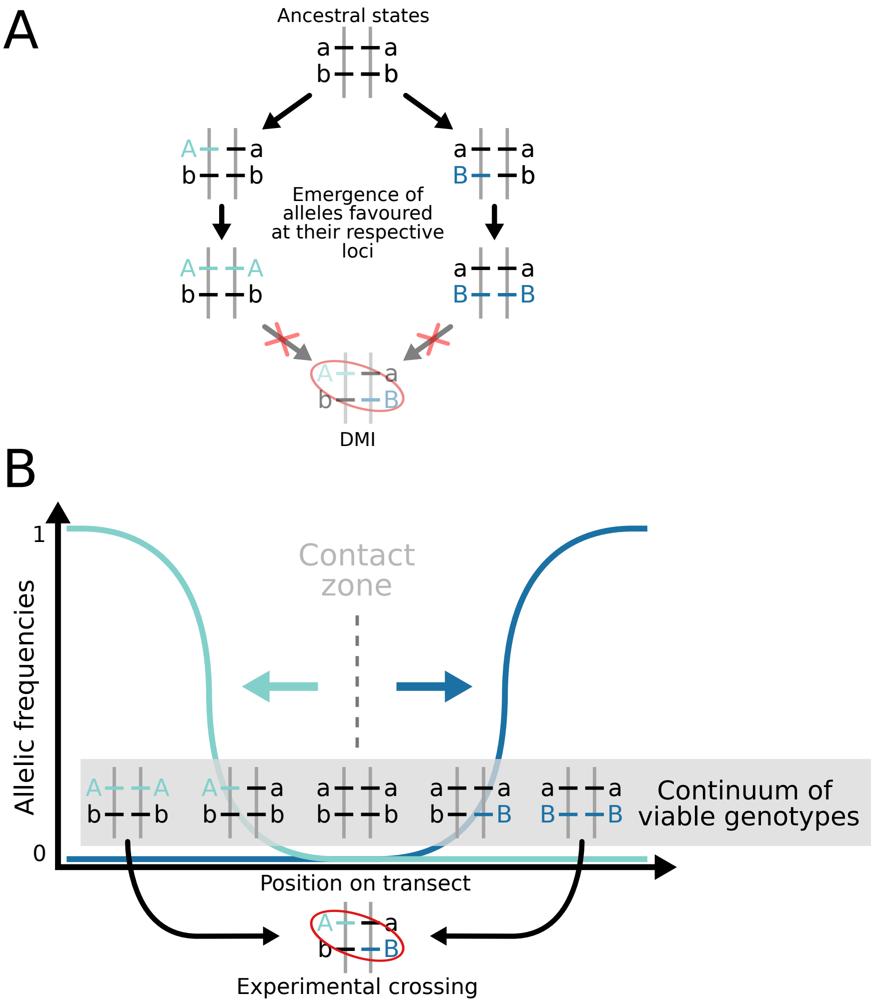


Figure 9: **Repel between DMIs.**

(A) Example of the emergence of a two loci DMI. Two incompatible alleles (A and B) from two loci appear and rise in frequency in allopatry. Hybrids between the two lineages suffer from a reduction of fitness as they carry both A and B incompatible alleles.
(B) Illustration of an effect of repel between two incompatible alleles (DMI). The y-axis represents the allelic frequencies, and the x-axis the geographical distance in function of the contact zone between the two demes (grey dotted line). As the genotype aabb is favoured near the contact zone, the A and B alleles repel each other from the contact zone, creating a continuum of viable genotypes. Experimental crosses of samples from the extreme end of the geographical distribution will generate non-viable hybrids carrying DMIs that are rarely expressed under natural conditions.

The use of hybrid zones in the investigation of the dynamic of speciation is not limited to crossing experiments. The effect of reproductive barriers, and thus the overall level of RI, can also be approximated through allele frequency gradients along the clines of hybrid zones. This approach requires a substantial effort of sampling, as the samples must be in sufficient number to cover most of the gradient of allele frequencies of multiple loci along the cline of interest. Multilocus clines usually have a steep slope of allele frequency at the centre of the hybrid zone, and long tails on each lateral gradient (Fig. 10). To improve the fitting of the data, the cline gradient of allele frequencies is usually divided into three parts (left, step and right gradient) and fit to a three-part 'stepped' cline model (Szymura and Barton, 1986, 1991). With this model, the strength of reproductive barriers *B* (or at least their effect on linked neutral loci) can be estimated through the equation $B = \frac{\Delta_p}{p'}$ (see Fig. 10), where Δ_p represents the central step (i.e. the difference of allele frequencies between the two extremes of the step gradient) and p' the slope of a lateral gradient. Note that both lateral gradients can be used to estimate *B*, thus a difference between the two values of *B* may indicate an asymmetry of gene flow.

The strength of barrier *B* obtained with this method must however be interpreted with caution, as physical features of the environment might mimic the effect of reproductive barriers although being completely independent of the genetic. Disentangling the effect of the environment on dispersal and habitat possibilities from the effect of reproductive barriers (i.e. RI of genetic origin) is not simple. The mapping of the population density on the different habitat variables should unveil the presence of physical constraint (Barton and Hewitt, 1985; Hewitt, 1988), although the density might also be decreased by genetic barriers in case of strong reproductive barriers (Barton, 1980). The genetic origin of the reproductive barrier observed might be furthermore established by the addition of other independent transects analysis or by direct studies of the occurring dispersal (Barton et al., 1993). Studies of hybrid zones and their allele frequency gradients can provide estimates of RI, making them valuable for studying the dynamics of speciation (once associated with levels of divergence). However, estimating RI at the genome-wide level may be challenging due to the difficulties in fitting clines (Westram et al., 2022), which reduces the adequacy of this method for studying the dynamics of speciation.

The application of allelic frequencies in investigating speciation dynamics extends beyond the examination of hybrid clines. As previously mentioned, gene flow between two populations

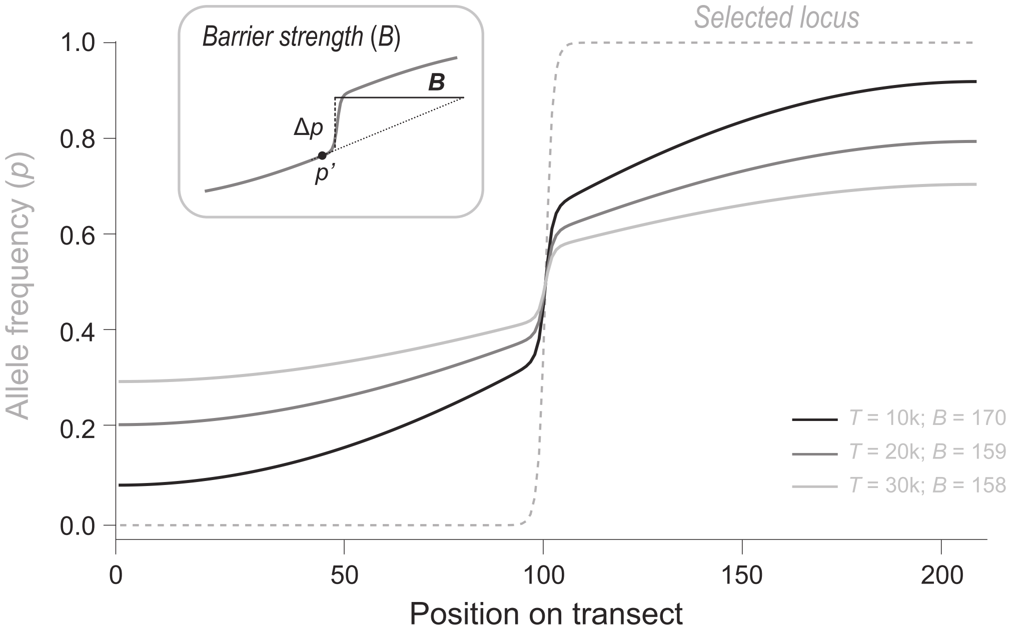


Figure 10: **Barrier strength estimated from allele frequency gradients.**

From Westram et al. (2022). Figure of the gradients of allele frequencies. The dotted line represents the frequency of the reproductive barrier (a selected locus), and the line with the different shade of grey are the allele frequencies at neutral loci (*r* = 0.01) at different times (*T* generations after the emergence of the barrier). The strength of the barrier (*B*) is measured with the ratio $\Delta p/p'$, with Δp the central step and p' the slope of a flanking gradient. Note that the value of *B* will decrease after the apparition of the reproductive barrier but rapidly stabilise in a stable hybrid zone (Nagylaki, 1976).

tends to homogenise allelic frequencies, producing low genetic differentiation throughout most of the genome, whereas high levels of genetic structure are expected in genomic regions linked to reproductive barriers. This effect of structuration is notably captured by statistics such as the F_{ST}. By measuring F_{ST} in windows along the genome, a method usually referred to as genome scan, it becomes possible to detect 'genomic islands of divergence' (i.e. segments of the genome with F_{ST} values that stand out of the overall distribution). This approach is convenient, as whole genome sequencing becomes more accessible with time, but is limited by the imprecision on the width of these islands (Ellegren et al., 2012; Harr, 2006; Hohenlohe et al., 2012; Nadeau et al., 2012; Turner and Hahn, 2010; White et al., 2010; H. M. Wood et al., 2008) and by the plurality of processes independent of reproductive barriers that can influence a statistic depending on the level of genetic diversity such as the F_{ST} (Cruickshank and Hahn, 2014; Lohse, 2017)[1]. The addition of other statistics not dependent on the nucleotide diversity such as the f_d (a statistic based on genome scan and a phylogenetic approach (Martin et al., 2015)) might help circumvent this limitation and provide a more genuine view of the reproductive barriers and level of RI.

Phylogenetic studies aim at uncovering the relationship among lineages by building bifurcating trees among related species. Moreover, it also allows inferences of previous occurrences of introgression (Hibbins and Hahn, 2022). This can be illustrated with a simple case with three species of interest (*S1*, *S2* and *S3*) and a fourth (*Ext*) only used to root the bifurcating tree (i.e. to determine the ancestral state of each polymorphic position, Fig. 11).

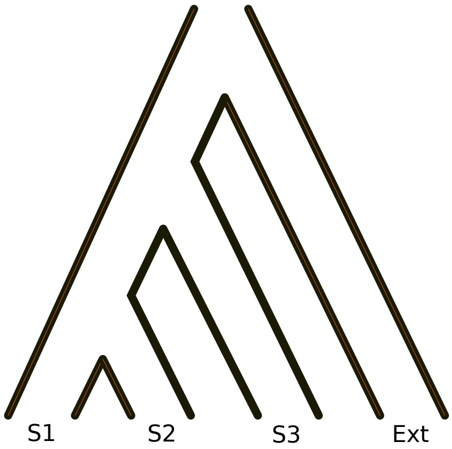


Figure 11: **Species phylogeny.**

Example of a phylogeny with 3 focal species plus one external group. The phylogeny of the species is based on the consensus phylogeny of all the loci.

Although this species tree, (((*S1*,*S2*),*S3*),*Ext*), is based on global information from all loci, the topology (branch length) of each individual locus' tree (referred to as gene trees) can vary from the overall species tree. Except errors, these incongruities are the result of two possible processes: *Incomplete Lineage Sorting* (*ILS*) of the ancestral polymorphism, or introgression

[1] Although, note that the extent of the influence of independent process is still debated, see for example Matthey-Doret and Whitlock (2019) for a discussion on the effect of background selection on locus-to-locus variation in F_{ST}.

events. *ILS* is observed when at least two lineages fail to coalesce before the previous event of speciation (Fig. 12.b, c and d). As *ILS* can mimic the genomic signature of introgression, we require a null model with only the effect of *ILS* to test for the presence of past introgression. The simple observation of congruent trees (Fig. 12.a and b) does not inform us much about introgression as scenarios with and without *ILS* can produce congruent trees, whereas incongruent trees (Fig. 12.c and d) are necessarily the result of *ILS* (or introgression).

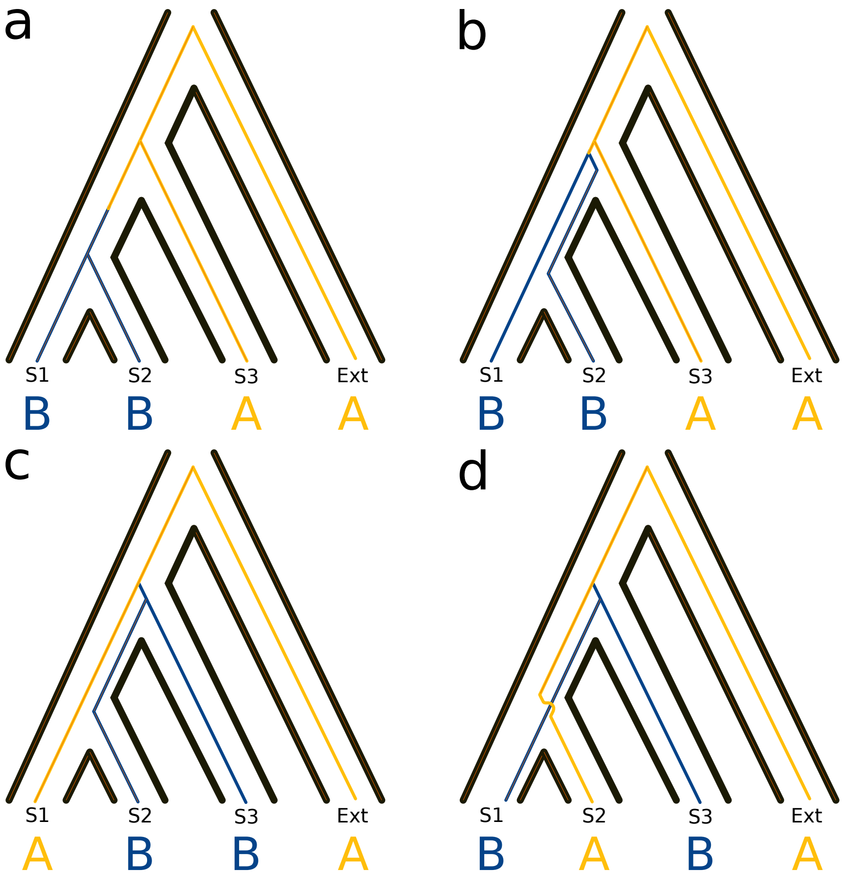


Figure 12: **Genetic tree topologies.**

Example of gene phylogenies. The black lines represent the species phylogeny, and the coloured lines represent the gene phylogeny of the allele *A* (yellow) and the allele *B* (blue). The gene and species phylogenies are congruent in the the two upper phylogenies (a and b), but the B phylogeny present an Incomplete Lineage Sorting (*ILS*) with a coalescence of the allele *B* (yellow) further in the past than the speciation event. The gene and species tree are incongruent in the two bottom phylogenies (c and d), and also present ILS.

The neutral multispecies coalescent model (MSC, Hudson, 1983; Pamilo and Nei, 1988; Tajima, 1983) establishes that a concordant tree topology (Fig. 12.a and b) can be produced by lineage sorting (Fig. 12.a) with a probability of $1 - e^{-\tau}$, and by *ILS* (Fig. 12.b) with a probability of $\frac{e^(-\tau)}{3}$, with $\tau = 2N$ generations. Furthermore, the probability of observing the two discordant gene tree topologies (Fig. 12.c and d) is equal to $\frac{e^(-\tau)}{3}$ each. This model inform us on the equiprobability of observing ABBA or BABA patterns ($\frac{e^(-\tau)}{3}$), providing a null model to test for the presence of introgression. Given the absence of introgression events, the frequencies of ABBA and BABA, due to *ILS* only, should be similar. In contrast, introgression between *S1* and *S3* or between *S2* and *S3* should increase the difference in observed frequencies between the two patterns. This difference is usually approximated as the Patterson's *D* statistic (or 'ABBA-BABA test', (Durand et al., 2011; Green et al., 2010)):

$$D = \frac{ABBA - BABA}{ABBA + BABA}$$

The absence of gene flow is inferred when $D \approx 0$, whereas $D \neq 0$ suggests gene flow occurred between *S2* and *S3* if *D* is positive, or between *S1* and *S3*, if *D* is negative. A first limitation of this method is associated with the fact that gene flow between S1 and S2 cannot be inferred, as it would only result in an increase in their internal branch length without changing the topology. Furthermore, the Patterson's *D* is not appropriate for scenarios of introgression between *S3* and the two lineages *S1* and *S2*, since this would lower the asymmetry of ABBA and BABA and reduce the value of *D*. Additionally, the strength of the approach is influenced by the direction, age, and degree of introgression, but it does not directly inform about their extent (Durand et al., 2011; Martin et al., 2015; Zheng and Janke, 2018). Derivatives of Patterson's *D* (usually referred as *D* statistics) have been proposed to extend the inferences to the direction of gene flow, the admixture proportion, and even, to a certain extent, the timing and rate of introgression (but see Dagilis et al., 2022, table S1). *D* statistics are now commonly used in speciation studies, but the meta-analysis of those studies is difficult as they often differ in their study effort, reporting standards, and methodology (Dagilis et al., 2022). *D* statistics are efficient approaches to infer the presence or absence of gene flow from whole genome data, but less reliable estimators of the admixture proportion (Dagilis et al., 2022), timing or direction of introgression events (Hibbins and Hahn, 2022). *D* statistics are efficient approaches to infer the presence or absence of gene flow from whole genome data, but less reliable estimators of the admixture proportion (Dagilis et al., 2022), timing or direction of introgression events (Hibbins and Hahn, 2022). This method's utility is limited in the context of studying speciation dynamics because it does not assess the age and extent of the most recent introgression event. Therefore, it is not suited to estimate the actual or recent RI necessary for depicting speciation dynamics.

Finally, inferences of demographic history from population genomic data (previously discussed in this introduction) can be used to estimate levels of RI along a continuum of divergence. Several methodologies have been developed for this purpose, most of them being grounded in the analysis of Site Frequency Spectrum (SFS). Noteworthy examples encompass $\delta a\delta i$ (Gutenkunst et al., 2010, relies on diffusion approximations), *Moments* (Jouganous et al., 2017, relies on the moment closure) or *fastsimcoal2* (Excoffier et al., 2021, relies on coalescence theory). In addition to the SFS, other methods can also take advantage of the Approximate Bayesian Computation framework (ABC, Beaumont et al., 2002). For example, *DILS* (Fraïsse et al., 2021), an ABC framework which presents the particularity to accommodate for varying levels of drift among loci induced by background selection. Furthermore, *DILS* implements variation in migration rates among loci, accounting for the effect of reproductive barriers linked to neutral markers (Roux et al., 2013). Overall, these methods all share the core concept of comparing statistical summaries of empirical data with statistical summaries predicted by demographic models.

Objectives of the thesis.

The methods presented in this introduction provide different and complementary elements of answer to elucidate the speciation continuum. However, their explanatory potential might remain too limited to allow a future thorough elucidation of the speciation continuum, and requests for the development of new complementary approaches (notably relying on large scale taxa comparison) are often found in the contemporary scientific literature (e.g. Baack et al., 2015; Dagilis et al., 2022; Payseur and Rieseberg, 2016; Stankowski and Ravinet, 2021). In this thesis, we present a new approach to test the effects of a given factor on the dynamic of speciation. We apply this method to test a historical assumption on the faster speciation of the animals in contrast with the plants, and to test for the effect of the selfing rate and the life form on the dynamics of speciation of the plants.

The approach we propose here is based on the comparison of multiple pairs of taxa for which RI has been estimated using population genomics approaches. Its idea is to explicitly test whether a given property, shared by a group of organisms, will have an impact on speciation dynamics. Speciation dynamics are investigated similarly as in Roux et al. (2016) or Dagilis et al. (2022), that is, by looking at the increase in RI as a function of time, based on samples from pairs of species with varying levels of divergence. The RI is approximated here by the genetic connection between populations/species in natura. Time is expressed here by molecular divergence. In this way, we look at the number of mutations required throughout the genome to interrupt gene flow within a comparative group. It is important to specify that by 'number of mutations', we mean neutral mutations, but that we have no precise idea of the genetic architecture of the underlying reproductive isolation. This approach can be summarised into four steps:

1. The identification of two sets of populations/species pairs with variable levels of divergence, each set sharing a property whose effect on speciation dynamics is to be tested. In this thesis, I will compare plants versus animals, autogamous versus allogamous plant species, and herb versus tree species but other comparisons can be made (e.g. other diploids vs polyploids , free living versus parasites...).
2. The collection of sequencing data for each taxa. The sequences obtained for a pair of populations/species makes it possible to describe, in natural populations, the patterns of polymorphism and divergence that result from demographic history.
3. The approximation of RI from demographic inferences for each pair of species. Here, we approximate RI as the ABC support for models with current isolation versus alternative scenarios with ongoing migration.
4. For each group (plants, animals, etc...), we model the relationship between divergence and IR by fitting a sigmoid using categorical regression. These different models can then be compared to test for a difference in speciation dynamics using a log-likelihood ratio test

(Fig. 13).

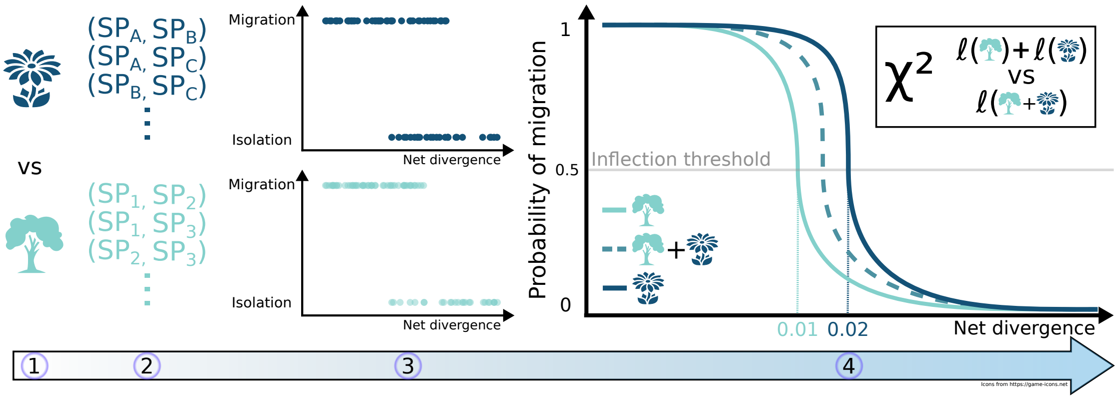


Figure 13: **Illustration of the approach.**

Summarised description of the approach presented in this thesis. ① Definition of groups for which speciation dynamics will be compared. Each group is represented by multiple pairs of species for which molecular data have been obtained. These groups are constructed to compare kingdoms at large phylogenetic scales (e.g., plants vs. animals) or to compare the effect of a biological trait (e.g., autogamous vs. allogamous). ② Once the taxa are selected, NGS data must be collected for a sufficient number of pairs of populations/species to cover the speciation continuum for both groups. ③ Analysis of demographic inferences is conducted for the retained pairs. Current gene flow is approximated from the most likely demographic model and plotted as a function of the net divergence (*da*). ④ The results of RI are used to fit a GLM (binomial) to model the dynamics of speciation by sigmoids. The sigmoid's inflection points are used to evaluate the average time for complete RI. Finally, the significance of the difference between the two sigmoids is tested with a log-likelihood ratio test.

The proposed approach requires that the investigated pairs of species cover the entire speciation continuum, starting from low level of divergence (i.e. conspecific populations) to high level of divergence (i.e. distinct species). Once the two taxa dataset is formed, the level of RI for each pair of species can be estimated. As previously discussed, the level of RI can be approximated with a variety of measures (e.g. fitness of hybrids, tree incongruity, demographic inferences...). In the approach proposed in this thesis, the approximation of RI is based on an ABC framework where the presence or absence of recent gene flow between two genetic clusters is inferred through likelihood comparisons of standard demographic models. Thus, the RI is simplified to a binary variable describing the recent/actual genetic connectivity of the two lineages. The RI is considered null between two populations if the most likely demographic scenario inferred is a model with ongoing migration, either a model of Isolation with migration (IM) or a model of Secondary Contact (SC). This binary proxy of RI has the benefit of avoiding some of the difficulties of calculating quantitative RI estimates (Westram et al., 2022), although being an oversimplification of the speciation continuum as RI is likely to emerge progressively with the accumulation of reproductive barriers (Wu, 2001), with the exception of 'instantaneous' speciation events such as those associated with polyploidy ((Coyne and Orr, 2004) or putative 'instant' homoploid hybrid speciation processes (Lamichhaney et al., 2018). After assigning a level of RI and estimating a level of divergence for each pair of species, the dynamics of speciation of both taxa can be represented as the evolution of RI along the continuum of divergence (Fig. 13.3). To compare the two dynamics, we propose a representation of these dynamics through linear regression, employing a generalised linear model (GLM) with a binomial link function, with the inflection point serving as a proxy for the average divergence at which complete RI is established. Then the difference between the two sigmoids is tested with a log-likelihood ratio test.

In the first chapter of this thesis, building upon animal data of Roux et al. (2016) and plant data obtained with open science, we introduce this novel approach by comparing the dynamics of speciation of the two kingdoms, testing the historical assumption that plants undergo speciation at a slower rate than animals. The comparison of the animal and plant taxa is illustrative of the approach but is also informative, as these two taxa show marked differences in speciation-related factors. In the second chapter of the thesis, we investigate the effect of two of those factors, the selfing rate and the life form, using the same plant dataset as in chapter I.

Table 1: **Glossary of key terms.**

Divergence continuum:	A gradient of genetic divergence between pairs of genetic clusters.
Gene flow:	The possibility of exchange of genetic information (not limited to genes) between two populations. Can be used to describe the proportion of genetic information actually exchanged (e.g. a low or high level of gene flow).
Genetic cluster:	Groups of individuals that form distinct clusters when grouped in function of the divergence among individuals.
Hybridization:	The production of hybrids between two semi-isolated species or species. Can sometimes be used improperly as a synonym of introgression although it does not necessarily imply introgression of genes in the foreign lineage.
Introgression:	The insertion of foreign genetic information in a population (or semi-isolated species).
Migration rate:	The proportion of individuals in a population that originate from another population.
Populations:	Groups of individuals from a species for which reproductive barrier does not prevent introgression. The expression 'semi-isolated species' is preferred with the accumulation of locus linked to reproductive barriers.
Reproductive barrier:	Allele or combination of alleles at two or more loci that reduce the fitness of hybrids or prevent their formation.
Semi-isolated species:	Populations that experience a reduction of gene flow at specific loci linked to reproductive barriers. They differ from population because of the heterogeneity of the migration rate along their genome, and from species as part of their genome are still able to be introgressed. Note that even neutral loci unlinked to reproductive barriers (i.e. with a recombination rate equal to 0.5) should be impacted by reproductive barriers (Westram et al., 2022).
Speciation continuum:	A continuum of RI (Stankowski and Ravinet, 2021), the absence of RI (conspecific populations) to a complete and durable RI (distinct species).
Species:	In this thesis, while the term 'species' is formally defined as a lineage with complete and durable reproductive isolation[2], its application exhibits considerable flexibility. It occasionally pertains to populations, semi-isolated species, or species. Therefore, opting for the term 'genetic cluster' as a more precise substitute was considered, but the retention of 'species' was favoured to maintain clarity for a broader readership.

[2] Defining the concept of species has been and remains a subject of much debates in the biologist community (e.g. Bolnick et al., 2023; Galtier, 2019; Mallet and Mullen, 2022; Stankowski and Ravinet, 2021; Westram et al., 2022. The main difficulty stemming from the dissonance between the discrete nature of our definitions and the continuous nature of speciation. I would like to add a personal reflection: To underscore the obvious, a definition is nothing more than a tool. Its purpose is to encapsulate a meaning we require under a common term. However, this necessity may not be the same for individuals seeking to define the same term. In the case of the concept of species, it is possible that a conservation biologist may not have identical needs to those of a population biologist or a palaeontologist. The first may require defining the species based on criteria that facilitate straightforward communication with decision-makers as well as ensuring optimal biodiversity preservation. The second may not be troubled by the impossibility of resolving every population/species dichotomy but may need a definition that gives meaning to the study of the speciation phenomenon. Finally, the third may need a species concept capable of manipulating populations from different time periods belonging to the same lineage and/or whose capacity to produce fertile hybrids remains unknown. Yet, it is not rare to read proposals for a definition to rule them all. Perhaps it would be more relevant to accept the idea of a plurality of species definitions? We might benefit from having several tailored tools rather than a single one incapable of satisfying all needs. Accepting a plurality of definitions would require authors to define the meaning they attribute to the term 'species' in each of their articles. However, adding a glossary to ensure the understanding of the terms used in an article is, at worst, a low-cost effort and, at best, beneficial for fostering better comprehension among researchers.

Chapter I

The results of this chapter have been submitted for publication. The main text of the manuscript submitted is attached as an appendix (Sup. article).

Introduction

In a nutshell, speciation is the accumulation of reproductive barriers between populations, leading to a complete reproductive isolation (RI) (Wu, 2001). Although all speciation events share this outcome, they differ in their dynamics, as they do not accumulate reproductive barriers following a universal molecular clock (Stankowski and Ravinet, 2021). As a result, populations' levels of divergence are not perfect predictors of their RI. The emergence, strength and persistence of reproductive barriers, the genetic core of RI, are shaped by multiple traits and environmental conditions varying at heterogenous rates over time (Coyne and Orr, 2004; Smadja and Butlin, 2011). Although being a crucial goal of evolutionary biology, elucidating the effects of these factors on the dynamic of speciation is not trivial as it faces major difficulties. First, the speciation, from the first reproductive barrier to the complete and irreversible RI, is a process that can extend up to millions of generations (Etienne et al., 2014). For example, by relating the fertility of first-generation hybrids reported on cross-breeding studies to estimates of the divergence times of species, Levin (2012) reports that hybrid sterility in herbaceous species is completed after on average four to five million years of divergence. Moreover, speciation is not necessarily straightforward. RI between two lineages can fluctuate in intensity through time, and sometimes RI can completely collapse (populations merge) (T. Dobzhansky, 1958; Seehausen et al., 1997; Taylor et al., 2005; Xiong and Mallet, 2022). Another difficulty relies on the diversity of effects that traits can have on speciation. A single trait can have antagonistic effects on the accumulation of species barriers. For instance, selfing can either promote (e.g. B. Charlesworth, 1992; Marie-Orleach et al., 2022; Wright et al., 2013) or impede (Gavrilets, 2004) the fixation of barriers, depending on the genetic determination of the barriers (see chapter II). Additionally, different traits involved in RI can be intercorrelated (Anderson et al., 2023). This is the case for example for dioecy and wind pollination. On one hand, dioecy is characterised by obligate outcrossing, which can enhance the potential for sexual selection and subsequently raise the accumulation of reproductive incompatibilities (Parker and Partridge, 1998). On the other hand, wind pollination may diminish reproductive isolation by enhancing the range of pollen dispersal (Loveless and Hamrick, 1984). Although the underlying causality is not well understood (D. Charlesworth, 1993), the observation of dioecy and wind pollination is often correlated (Bawa, 1980; Chazdon et al., 2003; Renner and Ricklefs, 1995). The challenge is therefore to unravel a process with a lifespan that exceeds the duration of scientific observation (at least for eukaryotes) and with a dynamic shaped by a rich combination of factors. This challenge requires different approaches (Stankowski and Ravinet, 2021): theoretical work provides models of the dynamics of speciation, for example the 'snowball model' that describe the dynamics of accumulation of intrinsic postzygotic barriers (Lynch and Real, 1994; Orr, 1995; Orr and Turelli, 2001) whereas experimental work allows testing for the processes driving and constraining genomic divergence (Nosil, 2012). Although essential to thoroughly describe how the reproductive barriers act on RI, those experiments are limited to a few generations and do not describe the dynamics of speciation but only 'instant snapshot' of it. In contrast, comparative studies can describe dynamics of speciation by estimating the RI of populations and comparing them among different groups of population's pairs. Comparing RI between groups of populations that differ in factors related to speciation might help to progressively characterise the effects of reproductive barriers on RI (Stankowski and Ravinet, 2021).

In a notable example of this approach, Roux et al. (2016) used sequencing data of 61 pairs of animal populations to conduct ABC demographic inferences. Pairs were then distributed in two groups in function of their most likely demographic history, i.e. pairs with ongoing migration versus isolated pairs, and plotted in function of net divergence (*da*). The results revealed an emergence of reproductive barriers around 0.075% of net divergence, a global isolation from 2% d_a and a greyzone of speciation between 0.5% and 2% of net divergence, where either migration or isolation could be inferred for different species pairs with similar levels of divergence, providing a first picture of an animal speciation continuum based on a binary approximation of the RI. In the continuity of this work, we introduce in this chapter a new comparative approach based on the ABC framework. To present this approach, we investigated a historical assumption opposing botanists and zoologists.

Hybridization has been historically mainly studied in plants. Botanists thought that hybridization was a common event, and that hybridization had an important role in evolution (Anderson and Stebbins, 1954; Stebbins, 1959), where zoologists claimed that hybridization had an insignificant effect on wild populations (Mayr, 1963) and used case of hybridization to study how species are isolated from each other rather than a supply of variation on which selection can operate (Taylor and Larson, 2019). Botanists suggested that hybridisation was an important driver of evolution, providing a large amount of variation between divergent lineages on which selection could act to produce a population adapted to a new environment/niche (Anderson and Stebbins, 1954) by providing new gene combinations (Stebbins, 1959). This view is notably supported by the work of botanists which showed the possibility for species to hybridise and introgress (Anderson, 1948; Anderson and Hubricht, 1938). On the contrary, as recalled by Dowling and Secor (1997), Grant (1971, p. 161) wrote that "several generations of zoologists have concluded that hybridization does not play an important role in animal evolution'. An opinion shared by Mayr (1963, p. 133) for which "the total weight of the available evidence contradicts the assumption that hybridization plays a major evolutionary role among higher animals". For Mayr, the limited number of animal hybrids observed could be explained because F1 and backcross progenies have a very low fitness (caused by unbalanced chromosome section). The idea that animal species hybridise (and by extension possibly introgress) less than plants is ancient and can be found in articles throughout the last decades (e.g. Dagilis et al., 2022; T. Dobzhansky, 1951; Gottlieb, 1984; P. R. Grant and B. R. Grant, 1992; Mallet, 2005; Mayr, 1963; Payseur and Rieseberg, 2016; Stebbins, 1959). Dobzhansky, in the third edition of his book Genetics and the Origin of Species (T. Dobzhansky, 1951, p. 300), discussed different reasons for the difference of hybrid observation frequencies between animals and plants. Among those arguments, which often come up in discussions between biologists, is suggested the idea that animals have more complex tissues and organ systems, which is thought to be correlated with a higher entanglement of genes. This entanglement would reduce the possibility of adaptive recombinants and increase the strength of reproductive barriers, an argument also advanced by Gottlieb (1984): 'the open, less integrative, and plastic patterns of plant morphogenesis are more permissive to large-effect genetic changes than those of animals'. In addition, clonal or asexual reproduction, which is more often encountered in plant studies, lowers the effective population size (due to longer lifespan and more clonal individuals), increasing the strength of the reproductive barriers (T. Dobzhansky, 1951, p. 300). In the same vein, Stebbins (1959) argued that animals are more complex in their adaptation to the environment. The range of viable phenotypes is then reduced (in comparison to plants), and so are the possibility of introgression, which would explain why more plant hybrids are observed. Noteworthy, some authors noted that the different expectations on the frequency of hybridization for plants and for animals may, at least in part, be due to a larger interest by botanists on the subject (T. Dobzhansky, 1951; Dowling and Secor, 1997; Whitney et al., 2010).

With the expansion of the next-generation sequencing (NGS) technology, the amount of sequenced genetic information has drastically increased during the last decade and with it the quantity of freely available data (Katz et al., 2022). This benefits scientific approaches that rely on important quantities of data as large datasets can be obtained with reasonable time and money investment. Based on a compilation of large datasets from about 27 published studies, a quantity of data that would have been difficult to reach with the resources of a single thesis, the present study aimed at (i) uncovering the dynamic of speciation in plants, (ii) compare this dynamic with the one observed for animals (Roux et al., 2016) with a new comparative approach. Based on arguments of historical scientific literature and on more recent results (Dagilis et al., 2022; Mallet, 2005), one might expect an earlier evolution of complete RI for animals in contrast with plants, rather than similar dynamics of speciation between plants and animals (null hypothesis) or faster speciation for plants (second alternative hypothesis).

Method

Data acquisition

The dataset was built only with sequencing data from other studies. Animal data come from Roux et al. (2016) were plant data come from the NCBI databank and its search engine (Sayers et al., 2022), the DDBJ search engine (Fukuda et al., 2021), Google scholar and direct data sharing (populations of Silene nutans , Muyle et al., 2021). Data directly obtained from NCBI accessions were searched with the following criteria :

- plant data
- raw data from RNA, WGS or RAD sequencing
- available on NCBI (Sayers et al., 2022) or from direct scientific collaboration
- from diploid species
- from sampled collected in the wild (no multi-generation crops, no manipulated crops)
- with at least two differentiated populations from a same genus
- with at least two individuals per population
- with sample location (at least the world region)

Biological models

Based on those criteria, data were collected for 25 genera from 27 independent sources (26 BioProjects and one direct data sharing) (Tab. S1). Of those 25 genera, 9 were issued from RNA sequencing, 13 from RAD sequencing and the 3 remaining from whole genome sequencing. The samples come mostly from the north hemisphere (USA, Europe and China), with only 5 genera sampled in South-America, south Asia and Oceania (Fig. S1). Most of the genera (17) are spread within the eudicot clade, the rest belong to monocots (5), magnoliids (1), as well as two genera outside of flowering plants: gymnosperm (1) and lycophyte (1) (Fig. 14). From the 131 species (all genera included), 516 pairs of congeneric species could be formed. As an exception, pairs of species were formed between the genera *Howea* and *Linospadix* because of the relatively low divergence between the two genera. The genus *Nepenthes* had almost as many pairs of species as the rest of the genera (233 *Nepenthes* for 283 non-*Nepenthes*), overrepresenting this taxa in comparison to the others. Thus, 17 pairs of *Nepenthes* were randomly subsampled and kept for the analysis to ensure a more genuine representativity of each genera. The species total decreased to 300 once the *Nepenthes* pairs were subsampled (Fig. 15).

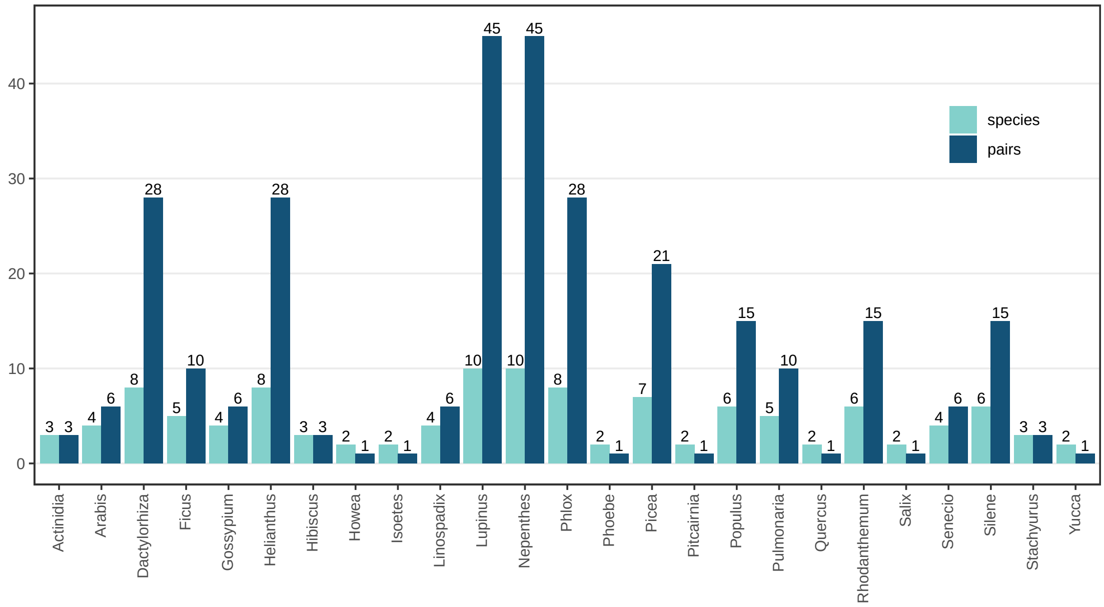


Figure 15: **Species and pairs of species distributions.**

Distribution of the number of species per genera. The x-axis presents the 25 genera and their respective number of species (blue bar) is represented on the y-axis. The number of pairs of species that can be formed in each genus is represented by an adjacent darker blue bar.

Raw data processing

The data from BioProjects were processed through a common workflow, with specific steps for RAD-seq data (Fig. 16, detailed commands in supplementary).

Unviable SRA accessions (individual sample) were filtered using SRA Run Selector (NCBI) to conserve only samples meeting the criteria listed in the previous paragraph (e.g. only wild sample, at least two individuals per population...). BioProject's data were retrieved from NCBI database (using *prefetch* from Sra-toolkit 2.11) and were stored on an IFB project (Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013)) for the whole processing. SRA files were uncompressed into fastq files (using *fasterq-dump* from Sra-toolkit 2.11). WGS and RNA fastq files could be used as is, but the preparation of RAD fastq files required a

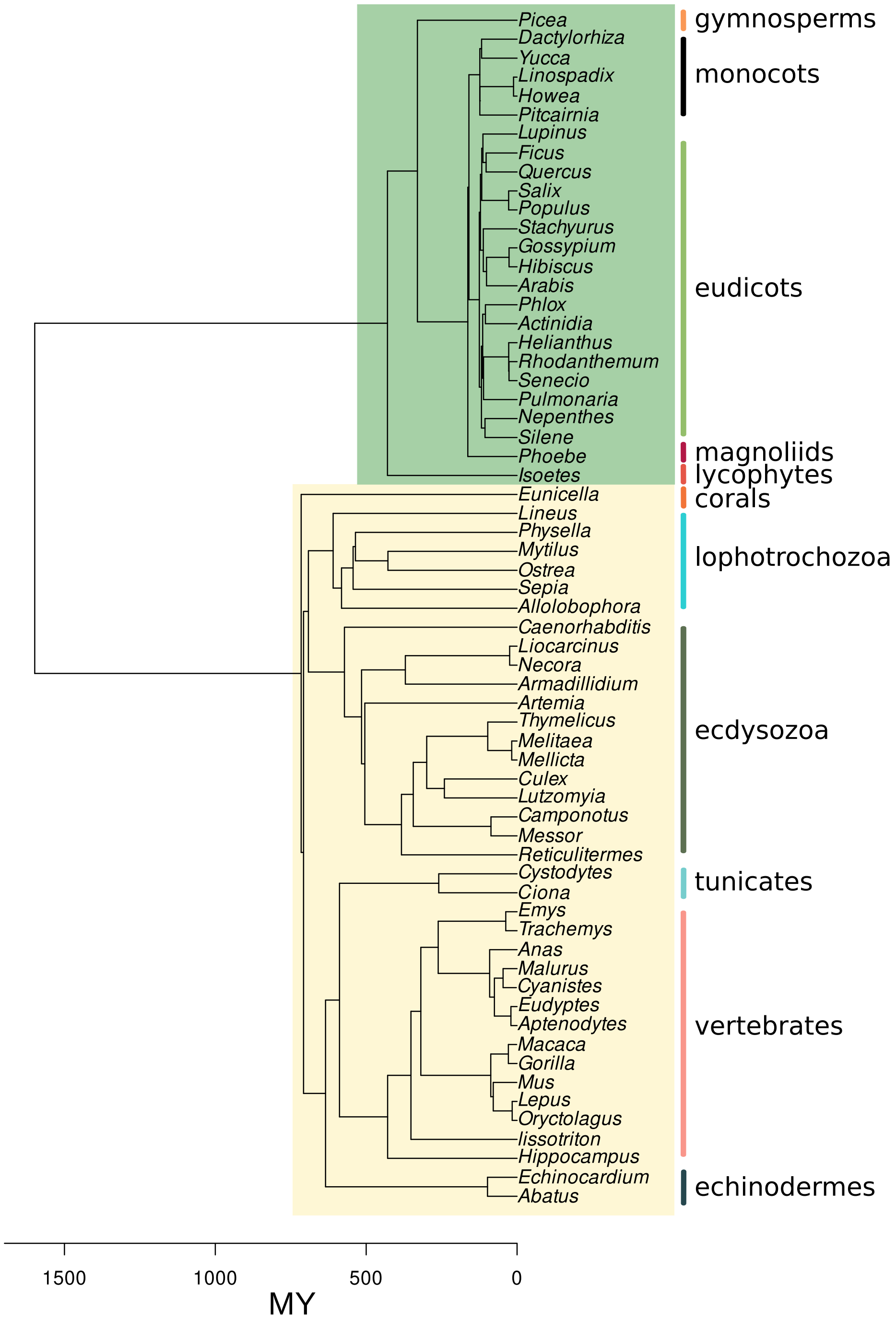


Figure 14: **Plants and animals phylogenetic relationships.**

The scale represents the time from present expressed in million years (MY) according to TimeTree (Kumar et al., 2017). Animals (yellow square) are from (Roux et al., 2016). Plants (green square) are included in the current study.

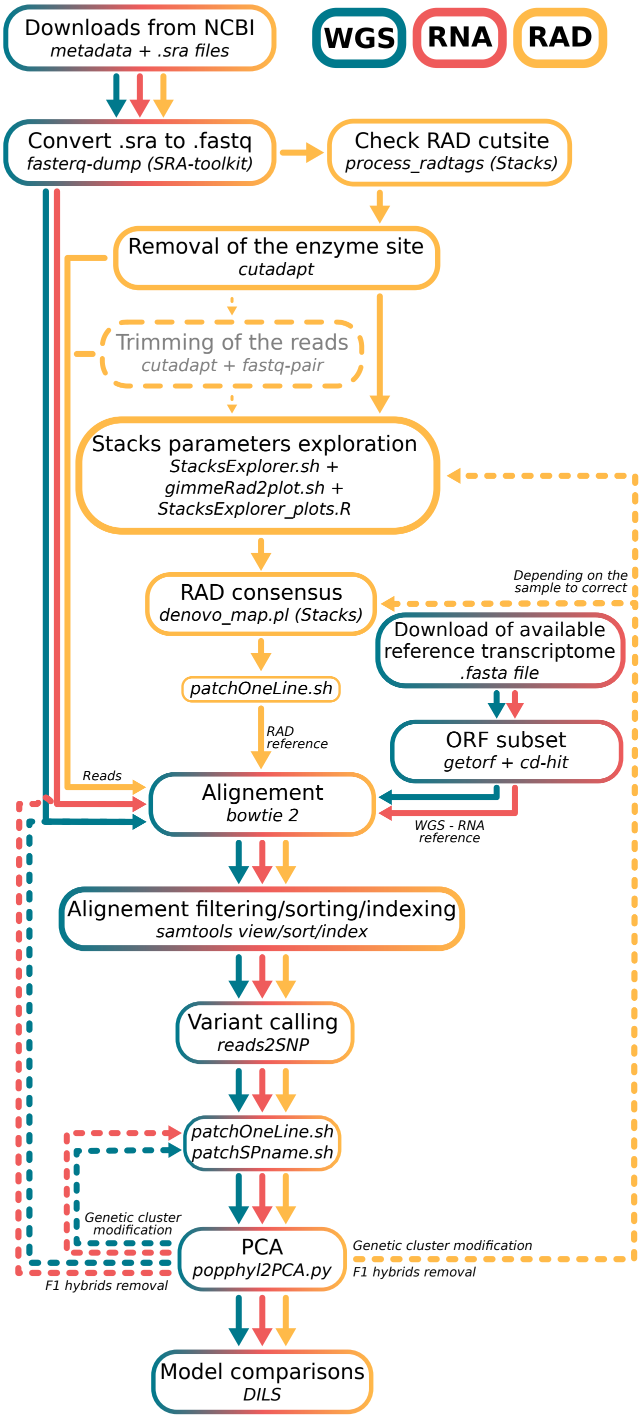


Figure 16: **Workflow.**

Diagram of the workflow presented in this method section. Each box represents a step, with in italic the type of files involved or the script used for the step. The workflow for the data issued from RAD sequencing (in yellow) is different from the two other types of sequencing (blue and red). Dotted arrows indicate additional steps that were necessary for specific datasets. These additional steps include the trimming of RAD reads when this was not already done before the upload on NCBI, the modification of the species of samples according to PCA analysis (split of population into genetic clusters or incorrect classification of a sample) or removal of hybrids of first generation.

few additional steps (see Fig. 16). The integrity of the RAD cut sites were checked (using *process_radtags* from Stacks 2.6, Catchen et al., 2011, 2013). This step was done just by safety, as it was supposed to be already done pre-NCBI upload. The enzyme sites were removed (using *cutadapt 4.0*, Martin, 2011) to avoid any bias in the divergence estimation (since the enzyme sites cannot have any polymorphism). The software used later to build mappable markers from RAD-seq data (Stacks 2.6, Catchen et al., 2011, 2013) requires reads of identical length. When this was not the case, reads were trimmed to a single length by removing reads too short and trimming the others (using *cutadapt 4.0*, Martin, 2011). The optimal length value allowing the conservation of the maximum amount of information was simply estimated with the formula

$$n = x. \sum_{k=x}^{lmax} n_{k'}$$

which gives the overall number of nucleotides kept after discarding reads that are shorter than the length *x* and trimming reads that are longer than *x*. For each length *k*, the number of reads of this length was computed using this command on two randomly selected samples[3]. In the case of paired datasets, single reads having lost their match were removed (using *Fastq_pair 1.0*, Edwards and Edwards, 2019).

At this stage, reads of three sequencing types were ready for the alignment and variant calling. When available, references from source articles were used for RNA-seq and WGS-seq datasets. Otherwise, transcriptome data from the 1KP project (Carpenter et al., 2019; Leebens-Mack et al., 2019) were used to produce references. ORF were extracted using *getORF* (Rice et al., 2000) and similar sequences were merged with *CD-Hit 4.8.1* ((Fu et al., 2012; Li and Godzik, 2006). References (mappable markers) for RAD datasets were produced using Stacks 2.6, Catchen et al., 2011, 2013. As recommended in Paris et al. (2017), combinations of parameters were explored to obtain a good enough combination of arguments for *denovo_map.pl* (from Stacks). Using a subset of individuals, mappable markers were build with differents combinations of the three parameters: the minimum number of reads required to form a stack ($-m$ ranging from 3 to 5), the number of mismatches allowed between stacks ($-M$ ranging from 1 to

[3] For paired datasets, the estimation was made on each end of a sample, therefore two thresholds were used.

6) and the number of differences allowed among stacks during the construction of the catalogue ($-n$ equal to M or $M+1$), for a total of 36 combinations (with arguments *–min-samples-per-pop 0.80* and *–rm-pcr-duplicates*). A combination of parameters was then selected based on the trade-off between maximising the number of polymorphic loci shared by at least 80% of the individuals and minimising the value of parameters for m, M and n. This combination was used to produce an assembly with all individuals per dataset (using *denovo_map.pl*).

References were indexed and aligned using *bowtie2 2.5.1* (Langmead and Salzberg, 2012). Alignment files were filtered for a minimum quality of 20, sorted and indexed with *samtools 1.15.1* (Danecek et al., 2021). Variant calling was made with *reads2snp* (Gayral et al., 2013; Tsagkogeorga et al., 2012) with a minimum of eight reads to call a genotype. Fasta files produced were then slightly modified to fit the format required by the software used in the analysis (sequences on one line with name of the species in their headers).

Before conducting the analysis, the correct match between the genetic clustering and the taxonomy provided was verified. PCA were produced for each dataset (using *popPhyl_PCA*) in order to exclude obvious F1 hybrids, and to split species into populations when clear genetic clusters could be visually identified.

Demographic analysis

The analysis consisted in comparisons of demographic models. The comparisons were performed with DILS (Fraïsse et al., 2021), an ABC framework (Approximate Bayesian Computation) to conduct demographic inferences, using the two or four individuals of each genetic cluster with the most information. Compared models included ongoing migration models (Secondary Contact **SC** + Isolation with Migration **IM**) and ongoing isolation models (Strict Isolation **SI** + Ancient Migration **AM**) (Fig. 8), each divided into sub-models (heterogeneous or homogeneous effective size along the genome, and heterogeneous or homogeneous migration in the case of ongoing migration models). For each pair of species of a genus was inferred the probability of ongoing migration, i.e. the add up probabilities of models with ongoing migration (**SC** submodels + **IM** submodels) versus ongoing isolation models (**SI** submodels + **AM** submodels). The parameters of *DILS*, supplied through a .yaml file, were set as follow :

region: *noncoding* (for RAD-seq) or *coding* (for WGS-seq and RNA-seq)
useSFS: 1
population_growth: *variable*
modeBarrier: *bimodal*
max_N_tolerated: 0.25
Lmin: 10
nMin: 4
mu: 7.31×10^{-9}
rho_over_theta: 0.2
N_min: 0
N_max: $5 \times max(\frac{\pi_A}{4_\mu}; \frac{\pi_B}{4_\mu})$
Tsplit_min: 0
Tsplit_max: $5 \times \frac{d_a}{2\mu}$
M_min: 0
M_max: 40

with π_A and π_B being Tajima's θ (Tajima, 1989) of the two species of the pair, and d_a being the net divergence (Nei and Li, 1979). The parameters respectively correspond to:

region: define if *DILS* should use all the positions or only the third 4-fold degenerate positions.
useSFS: define if *DILS* should include the site frequency spectrum to summarise the data.
population_growth: define if the population size can vary or not with time.
modeBarrier: define the type of model from which samples the rates of heterogeneous migration.
max_N_tolerated: define the maximum proportion of N/gaps allowed in the sequence of a locus.
Lmin: define the minimum length of a sequence to be considered.
nMin: define the minimum of sequences per species for a locus to be considered. If a locus has more than nMin sequences, then *DILS* samples nMin of those sequences for the analysis.
mu: define the mutation rate per site per generation.
rho_over_theta: define the rate of recombination over mutation.
N_min and **N_max**: define the minimum and maximum prior of population size.
Tsplit_min and **Tsplit_max**: define the minimum and maximum prior number of generations for the time of demographic changes.
M_min and **M_max**: define the minimum and maximum prior number of migrants per generation.

Once the analysis were conducted, and similarly as in the source study of the animals data (Roux et al., 2016), the pairs with an inferred probability of ongoing migration between 0.1304 and 0.6419 were considered not sufficiently reliable and were therefore withdrawn from the rest of the analysis.

In order to test for the difference between the plant and animal dynamics of speciation, the probability of ongoing migration was transformed into a binary variable. Pairs with ongoing migration were set as 1 and pairs with ongoing isolation set as 0. Animal and plant ongoing migration status were then modelled as a function of the net divergence, using a Generalised Linear Model (GLM) with a logit function :

$$g\left(\mathbb{E}(Y_i|\mathbf{X}_i)\right) = g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1,i}$$

where $\beta 0$ represents the intercept and $\beta 1$ the coefficient reflecting the effect of genomic divergence on the isolation/migration status coded as 0 and 1, respectively. The probability p_i to observe an ongoing migration status for a pairs of species at a level of divergence of X_i, can be calculated with :

$$p_i = \frac{\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)} = \frac{1}{1+\exp\left(-\mathbf{X}_i\boldsymbol{\beta}\right)}$$

With these two expressions, the level of net divergence X_i of a given probability p_i can be obtained as follow :

$$X = -\frac{1}{2\beta_1}\left(\beta_0 + \sqrt{\beta_0^2 + 4\beta_1 \log\left(\frac{p_i}{1-p_i}\right)}\right)$$

This expression can be used to retrieve the inflection point of a sigmoid. This point corresponds to the level of net divergence X for which the model predicts a probability p equal to 0.5, in other words it corresponds to the level of net divergence from which we are more likely to observe pairs of species isolated than pairs with ongoing migration. For $p = 0.5$, the expression can be simplified as $X = -\frac{\beta_0}{\beta_1}$. The log-likelihood function ℓ of the migration/isolation status Y given the average net molecular divergence X is then obtained to evaluate the fit of a model to the observed data:

$$\begin{aligned}
\ell(\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}) &= \log\left(\mathcal{L}(\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x})\right) \\
&= \sum_{i=1}^{N}\left[y_i\log(p_i)+(1-y_i)\log(1-p_i)\right] \\
&= \sum_{i=1}^{N}\left[y_i\log\left(\frac{p_i}{1-p_i}\right)+\log(1-p_i)\right] \\
&= \sum_{i=1}^{N}\left[y_i\cdot\mathbf{x}_i\boldsymbol{\beta}-\log(1+\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right))\right] \\
&= \sum_{i=1}^{N}\left[y_i\cdot(\beta_0+\beta_1X_{1,i})-\log(1+\exp\left(\beta_0+\beta_1X_{1,i}\right))\right]
\end{aligned}$$

To test for a significant difference between the plants and animals sigmoids (i.e. speciation dynamic), the log-likelihood ℓ of models with different subset of data (minus inferences with insufficient probability of ongoing migration (again, as in Roux et al., 2016) were compared:

1. M_0 : both plants and animals share the same logistic relationship between X_i and Y_i .
2. M_{plants} : model fitted to the plants data only.
3. $M_{animals}$: model fitted to the animals data only.

The log-likelihood $\ell(M_0)$ was then estimated for the whole dataset comprising both plants and animals by using the above formula where:

- β_0 and β_1 represent for M_0 the coefficient of the model fitted to the whole plants and animals dataset by using the glm function (family = 'binomial') implemented in R.
- $X_{1,i}$ represents the series of observed divergence values for a single kingdom (plants or animals).
- y_i represents the series of inferred isolation/migration status for a single kingdom (plants or animals).

For M_{plants} and $M_{animals}$, GLM models were fitted with only the data from the corresponding kingdom. We then estimated the log-likelihoods $ell(M_{plants})$ and $\ell(M_{animals})$ as for M_0. Finally, comparisons were conducted between the log-likelihood $\ell(M_0)$ and the combined log-likelihood $\ell(M_{plants})+\ell(M_{animals})$, which is derived from the summation of log-likelihoods obtained by fitting independent models to each respective kingdom. The significance of the difference between $\ell(M_0) and \ell(M_{plants}) + \ell(M_{animals})$ was evaluated using a log-likelihood ratio test. Specifically, twice the absolute difference of the log-likelihood between $\ell(M_0)$ and $\ell(M_{plants}) + \ell(M_{animals})$ is approximately χ-squared distributed. The *P*-value returned by the R function *pchisq* corresponds to the probability of observing $2.|\ell(M_0) - \ell(M_{plants}) - \ell(M_{animals})|$ in a χ-squared distribution with two degrees of freedom.

Geographical effect

Geographical coordinates were collected to investigate the role of geographical distances between sampled populations in the observed dynamic of speciation. The information came from NCBI metadata, source article or from exchange with the articles's corresponding author. Not all sample coordinates could be obtained. For every pair of species, the distance between each pair of inter-specific samples was calculated using the Vincenty Ellipsoid method in the R package *geosphere* (Hijmans et al., 2022) and compared to retaining only the lower distance (only the individuals retained for the analyses were used, i.e. two or four samples per species). The choice of using the minimum distance (rather than another unit of distance, the mean distance for example) was motivated by the idea that this was the safest way to assess the 'strength of allopatry' between populations. The difference of median minimum distance between the plants and animals datasets was tested with a Wilcoxon test (R function), and both net divergence and geographic distance were tested as explaining variables of the migration status fitting a GLM (R function, binomial family).

Sequencing technology effect

The animals and plants datasets differ in their distribution of sequencing technologies. Looking at the pairs of species which were inferred with a sufficiently strong probability of ongoing migration, animal pairs are mostly issued from RNA-sequencing with 46 pairs plus 8 pairs from Sanger sequencing. In contrast, pairs of plant species are more diversified in sequencing technologies with 90 pairs from RAD-sequencing, 86 pairs from RNA-sequencing and 34 pairs from whole genome sequencing. The influence of this asymmetry was investigated with a log-likelihood ratio test between the plants and animals RNA-seq pairs of species.

Table 2: **Log-likelihood ratio test for logit models fitted to plant and animal datasets.**

model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	**df**	***P*-value**
M_0	-108.4313	1.977	-508.150	0.0039		
M_{plants}	-74.61307	2.532	-802.545	0.0031		
$M_{animals}$	-6.991024	3.967	-237.160	0.0167		
					2	2.23×10^{-12}

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{plants} and $M_{animals}$.
$\boldsymbol{\beta_0}$: estimated intercept.
$\boldsymbol{\beta_1}$: estimated coefficient.
$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{plants}) - \ell(M_{animals})|$ in a χ-squared distribution with two degrees of freedom.

Results

Of the 300 pairs of species investigated, only 210 obtained a sufficiently robust estimate of the probability of ongoing migration (with a value outside of [0.1304;0.6419], based on the robustness test of Roux et al. (2016)), to be kept for the rest of the analysis. Six species were withdrawn because of their absence of pairs with robust estimates, including the two species of *Isoetes*: *I.lacustris* and *I.echinospora*, resulting in 118 species and 24 genera after removal (Fig. Inferred Probability of ongoing migration). The net divergence of pairs of plant species ranged from null[4] to $\approx 5.52\%$, where animal's pairs ranged from $\approx 0\%$ to $\approx 31\%$ of net divergence. In contrast with the grey-zone observed for animals data between $\approx 0.5\%$ and 2% of net divergence), the plant's grey-zone ranged from ≈ 0.16 to $\approx 0.71\%$ (Fig. 17). The inflection points of the plant and animal sigmoids were measured respectively at $\approx 0.32\%$ and $\approx 1.67\%$ of net divergence (Fig. 18). The significance of this difference was confirmed with a *P*-value of 2.23×10^{-12} (Tab. 2). The log-likelihood ratio test with only the RNA-seq pairs of species was significant with a *P*-value of 5.38×10^{-8} (Tab. 3).

[4] Negative net divergences were considered as null. This negative values are observed for the pairs of species with a mean polymorphism higher than their divergence, since $d_a = d_{XY} - (\pi_X + \pi_Y)/2$

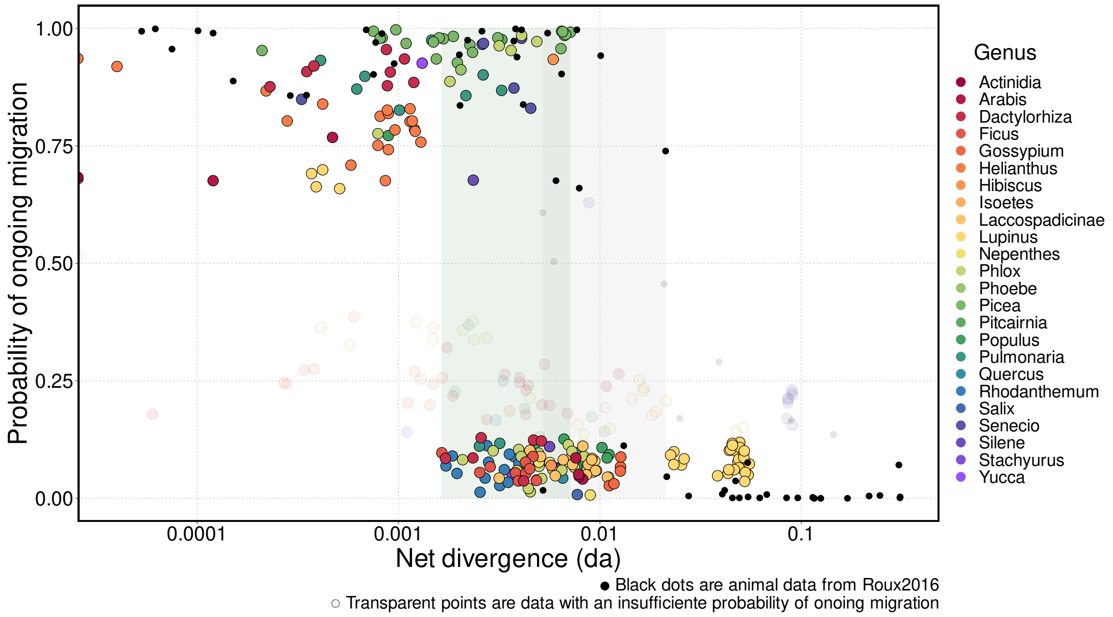


Figure 17: **Inferred probability of ongoing migration.**

Plot of the probability of ongoing migration inferred (y-axis) on the net divergence (x-axis). Each dot represents a pair of species (i.e two genetic clusters) and the colour indicates the genus of the two species (black is for the animal pairs from Roux et al. (2016)). The probability (y-axis) represents the confidence of the ABC inference in the choice between the 'super model' of ongoing migration scenarios (SC + IM) and the 'super model' of ongoing isolation (AM + SI). Pairs of species with insufficient probability of ongoing migration (therefore not used in the rest of the analyses) are transparent. The grey-zone (i.e. range of net divergence with both genetically connected and isolated pairs of species) of the animal data is represented as a grey area (range from $\approx 0.5\%$ to $\approx 2\%$ of net divergence), and the equivalent for plant data is represented as a green area (range from $\approx 0.16\%$ to $\approx 0.71\%$).

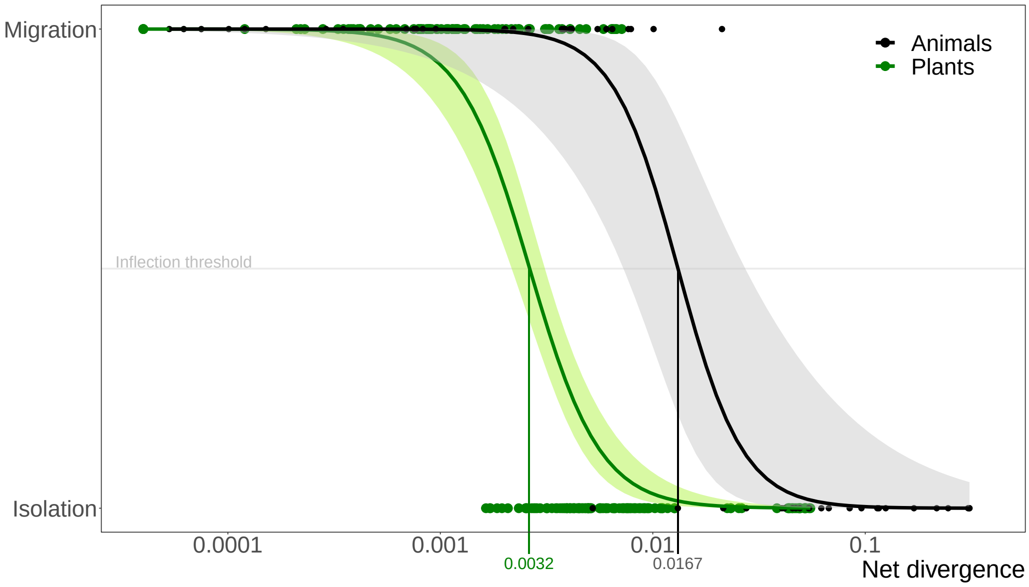


Figure 18: **Plants and animals sigmoids.**

A plot depicting the ongoing migration status as a function of net divergence (*da*, Nei and Li, 1979). Green dots are pairs of congenera plant species, black dots are pairs of animals (data from Roux et al., 2016). The ongoing migration status, inferred using *DILS* (Fraïsse et al., 2021), is shown on the y-axis, while net divergence is represented on the x-axis. The sigmoid curves represent the probability of current or recent gene exchange for pairs at various levels of net divergence (*da*). This probability is determined using linear regressions (GLM) and is accompanied by a 95% confidence interval. The inflection points, which signify the threshold between a probability of ongoing migration $>$ 0.5 and $<$ 0.5 (represented as a horizontal grey line), are indicated by vertical bars. These inflection points are approximately 0.3% (95% CI: [0.27%-0.47%]) of net divergence for plants and about 1.7% (95% CI: [1.52%-2%]) for animals.

Table 3: **Log-likelihood ratio test for logit models fitted to plant and animal RNA-seq datasets.**

model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	**df**	***P*-value**
M_0	-41.92664	2.413	-320.743	0.007		
M_{plants}	-20.8281	4.031	-766.155	0.005		
$M_{animals}$	-4.361694	5.347	-271.134	0.0197		
					2	5.38×10^{-8}

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{plants} and $M_{animals}$.
$\boldsymbol{\beta_0}$: estimated intercept.
$\boldsymbol{\beta_1}$: estimated coefficient.
$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{plants}) - \ell(M_{animals})|$ in a χ-squared distribution with two degrees of freedom.

Species pairs were divided into three categories based on the best scenario of speciation inferred. Then, average levels of divergence were compared between plants and animals within each category. Homogeneous migrations (i.e ongoing migration with an unimodal rate of migration among the loci) were inferred for plants and animals at respectively $\approx 0.45\%$ and $\approx 0.41\%$ of net divergence (Fig. 19). Heterogeneous migrations (i.e. detection of putative reproductive barriers) were first detected at 0.037% of net divergence for plants and at 0.0751% for animals, and inferred at a maximum of $\approx 0.714\%$ and $\approx 2.11\%$ respectively. Interestingly, the range of evolution of the reproductive barriers, that spawn from the level of net divergence of the first pair with heterogeneous migration to the last pair with homogeneous migration, is wider for plants $(+\sim 0.08\%)$, a result consistent with the less 'clock-like' (i.e. more diverse) accumulation of RI reported in the scientific literature (Baack et al., 2015). The first pair of isolated species was inferred at $\approx 0.16\%$of net divergence for plants and $\approx 0.5\%$ for animals.

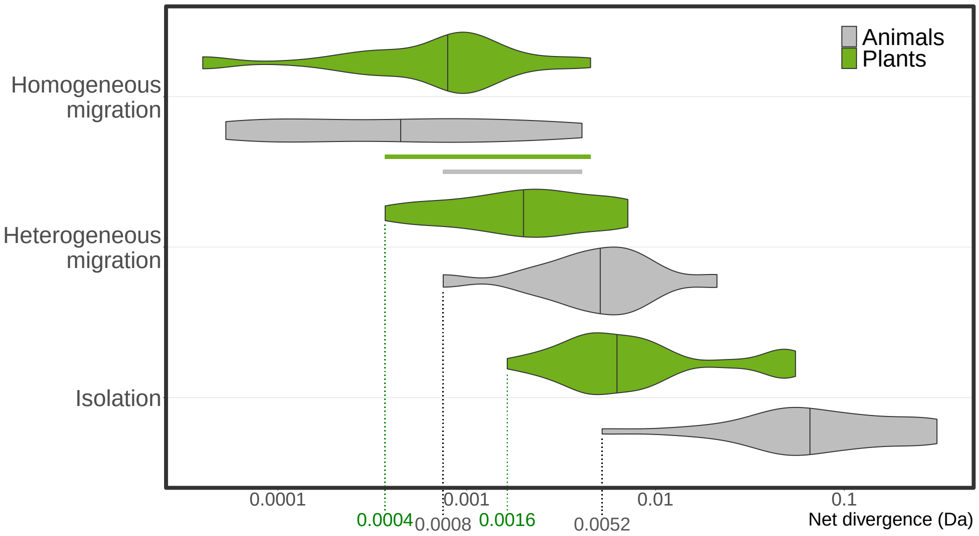


Figure 19: **Distribution of the type of migration inferred in function of the net divergence.**

Distribution of the pair's inferences. Ongoing migration scenarios (isolation with migration and secondary contact) are distributed in the two first categories, depending on the homogeneity of the migration rate along the genomes. Ongoing isolation scenarios (ancient migration and strict isolation) are all distributed in the last category. The range of detectable first reproductive barriers, as indicated by the net divergence between the first pair with heterogeneous migration and the last pair with homogeneous migration, is represented by vertical bars, each corresponding to a specific kingdom and colour.

More than 75% of the pairs of animal species inferred with ongoing migration have experienced a continuous migration since their split. This contrasts with the proportion observed for plants where the majority of the pairs ($\approx 55\%$) have initiated their accumulation of divergence in allopatry (Fig. 20).

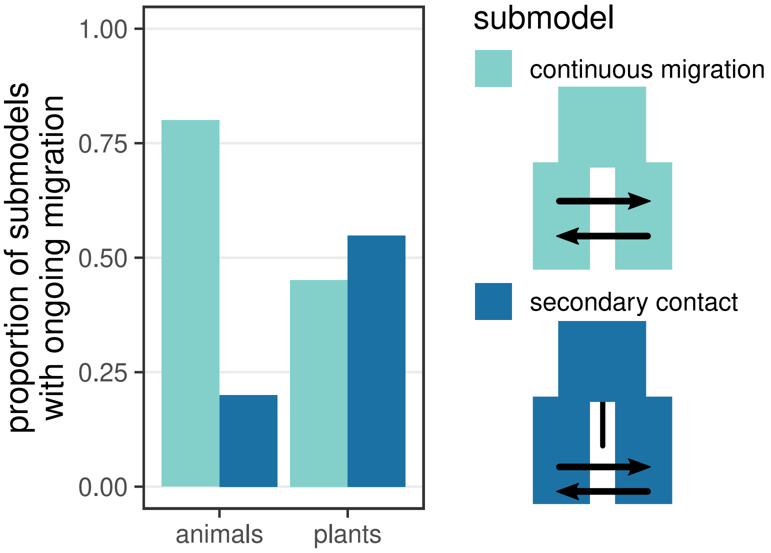


Figure 20: **Distribution of ongoing migration models.**

Barplot presents the proportion of ongoing migration submodels for the animals and for the plants.

Plant's minimal geographical distance was available for 121 of the 210 pairs of plant species (but geographical distance was available for each pair of animal species). For these 121 plant pairs, minimal geographical distance goes up to 2,000 km (10,000 km for animals), with a median at $\approx$ 230 km (significatively different of the 770 km for animals, Wilcoxon *P*-value $= 3.547 \times 10^{-5}$) (Fig. 21). The pattern observed for the green or grey-zone (sigmoid curve) is not observed with geographical distance (km) in place of genetic distance (d_a) (Fig. 22). The glm model (using plant data) found significant effects for genetic distance but not for geographical distance (*P*-value $= 1.19 \times 10^{-7}$ vs 0.866).

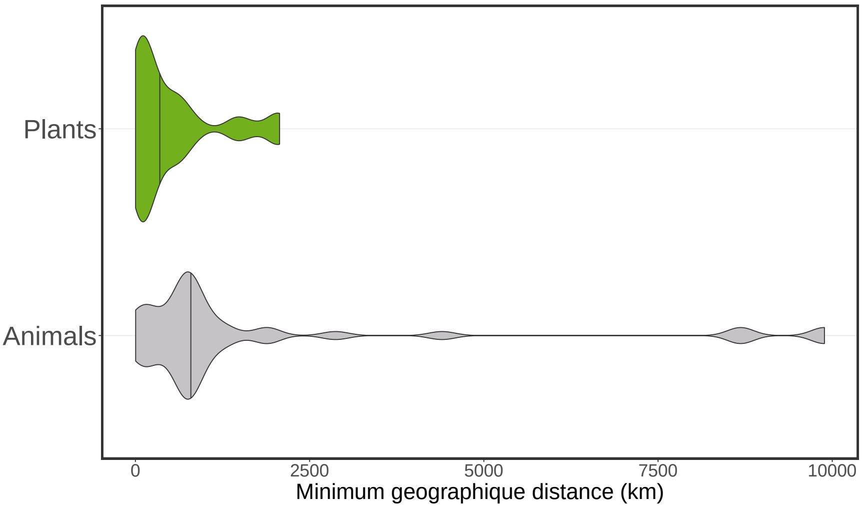


Figure 21: **Minimal geographical distances.**

Violin plot showing the minimum geographical distance between samples from two populations within a pair. A distance of *x* kilometres for a pair of species indicates that the two closest samples from each population are separated by *x* kilometres.

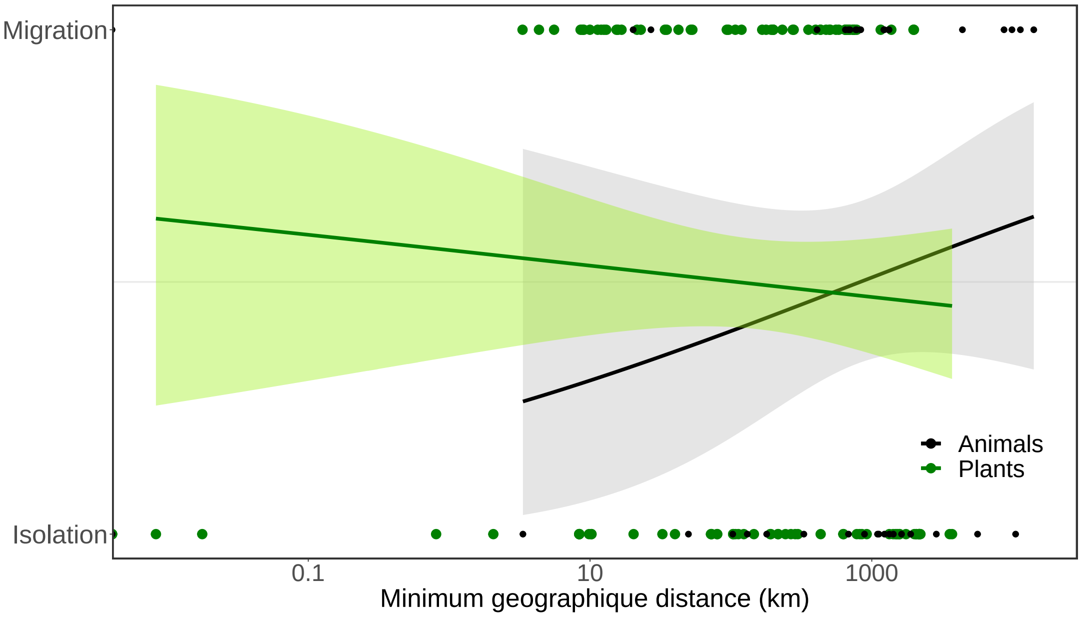


Figure 22: **Sigmoids of minimal geographical distances.**

Plot illustrating the ongoing migration status as a function of the minimum distance between samples from the two species within each pair. Each dot represents a species pair for which GPS coordinates of samples were available. The sigmoids were obtained by fitting linear models (GLM with a binomial link function) to the data for both plants and animals.

Discussion

Representativity of the dataset

The first challenge of this study was to build from open science a dataset that adequately represents the dynamics of speciation of the plant taxa. The good representativity can be reduced to two criteria: enough pairs of species to cover the entire divergence continuum of the speciation process, and a sampling of species that cover all the major taxa of the plant kingdom. The former criterion seemed to be fulfilled as the net divergence of the 300 pairs of species were distributed along a divergence continuum covering the whole speciation process, from almost zero divergence to levels of net divergence where ongoing migration was strictly absent. The completion of the second criterion is more critical, as most of the species of the dataset belong to the eudicots. Missing some of the major plant taxa exposed the analysis to the risk of inferring a dynamics of speciation nonrepresentative of the overall plant kingdom. This dataset provides a large scope of combination of traits linked with speciation dynamics, such as the life forms (tree, herb..), mating system (some species are known selfers, such as *Arabis nemorensis* and *A.sagittata*, Dittberner et al., 2022), were other carries self-incompatibility system, such as some *Helianthus* species (Ferrer and Good-Avila, 2007) or lifespan (*Silene dioica* is annual or biannual were other *Silene* of this dataset are perennial (based on world flora online). However, some interesting taxa with specific traits (or combination of traits) are still poorly or not represented at all, such as algae or gymnosperms, as data for such taxa was more difficult to obtain from open science. The comparison of dynamics of speciation in the present study will therefore be limited to seed plants.

Comparison of the plants and animals speciation dynamics

Roux et al. (2016) uncovered a grey-zone of speciation in animals between $\approx 0.5\%$ and $\approx 2\%$ of net divergence, that respectively correspond to the pair of isolated species with the lower net divergence and the pair of introgressing species with the highest net divergence of their animal dataset. This contrasts with the results obtained in this study as, using a similar approach (same ABC framework, same robustness thresholds), we observed a plant's grey-zone shifted at a significantly lower level of net divergence. Not only the plant's grey-zone begins at a level of net divergence around 3 times lower than the one of animals, but the plant's grey-zone is also around 3 times narrower. Solely based on those grey-zone characteristics (animals and plants), this could indicate a faster speciation for plants. However, the difference observed could be mostly explained by outliers as the grey-zones are defined uniquely on two extreme pairs of species and do not account for the complete distributions of pairs of species. For this reason, we have opted for an approach that could account for the whole distribution of pairs of species, that is the comparison of linear models fitted to the plants and animals data. Comparison of the two sigmoids inflection points indicated that plants speciate at a ≈ 5 times lower level of net divergence than for animals, corroborating an earlier cessation of the gene flow in the plant kingdom. In addition, heterogeneous migration (i.e. semi-isolated species) was inferred at a lower level of net divergence for plants (Fig. 19), suggesting an earlier development of reproductive barriers. Together, these elements indicate that complete RI is reached at a lower level of divergence and might suggest a faster speciation process for plants in contrast with animals. However, given that the speciation is not thought to be an unidirectional process (Stankowski and Ravinet, 2021), the level of net divergence should not be considered as the true age of the speciation events. Thus, it is possible that the establishment of a complete RI takes more absolute time (i.e. in years) for plants even if reproductive barriers and complete RI are observed at lower levels of net divergence for plants. Noteworthy, the complete RI approximated from the absence of detectable recent gene flow might be recklessly interpreted as a completed speciation event. The absence of recent introgression can be explained by an achieved speciation process (i.e. reproductive barriers completely and durably prevent exchange of genetic material), but can also be explained by a sufficiently long allopatric state for those pairs of species. If a secondary contact was to happen, the speciation thought completed could turn out to be still in progress. Similarly, the collapsing of reproductive barriers could also reveal the incompleteness (i.e. non-perennial) of a speciation process (e.g. Taylor et al., 2005; Xiong and Mallet, 2022). This mistakenly assumption of a mature speciation through the inference of recent genetic isolation is predominantly expected for low diverged pairs of species, but could also be particularly encountered for plants, as it seems that plant speciation events involve more often allopatry than those of animals (Fig. 20).

Genetic and morphological clusters

The accumulation of divergence is initiated more often in allopatry (SC) than in sympatry (IM) for the pairs of plant species (Fig. 20), a result concordant with the scientific literature (Abbott, 2017). The opposite is observed for the animals where most of the pairs initially diverged in sympatry. This dissymmetry could explain in part the contrast between the earlier emergence of complete RI for plants observed in this analysis and the (historical) assumption that plants tend to hybridise more often than animals (see introduction of this chapter). This assumption is mostly based on observations in the field where natural hybridization is reported for more species of plants than for animals (T. Dobzhansky, 1951; Gottlieb, 1984; P. R. Grant and B. R. Grant, 1992; Mallet, 2005; Mayr, 1963; Stebbins, 1959). These identifications of hybrids rely on morphological recognition, therefore it is possible that part of the hybrids went disregarded as their morphological features were not sufficiently contrasted with parental lineages. These omissions are particularly expected in taxa with relatively low morphological divergence, or in backcross hybrids that are often difficult to distinguish from parental species (Abbott, 2017; Mallet, 2005). Similarly, we might except hybrids from recent secondary contact (SC) to display more contrasted morphology than hybrids from perennial hybrid zones (IM, Fig. 23). The latter being composed of a higher proportion of hybrids backcross, it might form a more continuous gradient of morphological disparities. Accordingly, we might hypothesise a bias in the recognition of natural hybrids as a majority of sympatric scenarios were inferred for animals, suggesting a lessened disparity of morphological traits among hybrid backcrosses. In contrast, plants mostly experienced allopatric divergence accumulations, which would be compatible with an easier detection of hybrids in zones with few backcrosses. The inconsistency between the results of the present study and the higher proportion of hybridising plant species encountered in nature could therefore be explained (at least in part) by the fact that this analysis is based on genetic clusters rather than morphological groups.

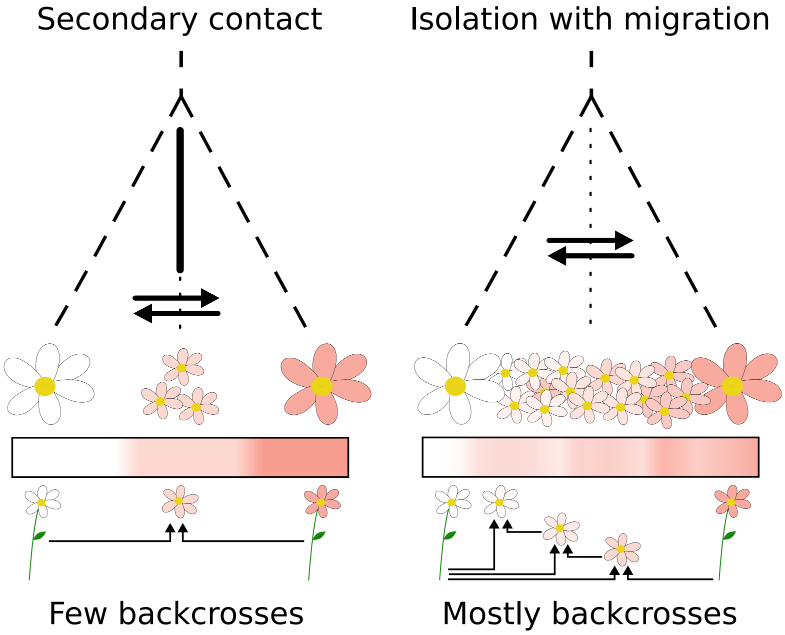


Figure 23: **Hybrid morphological disparity.**

Figure illustrating the hybrid morphological disparities that could be expected in hybrids from idealised recent SC (left) or IM (right) scenario. Hybrids from perennial hybrid zones are more likely to be backcrosses of different degrees, thus lessening morphological disparities among close individuals.

Influence of the geographical factor on the dynamics

The dissimilarity of the plants and animals dynamics of speciation is probably mostly explained by differences in speciation factors (e.g. selfing rate, presence of chloroplast...), but could also be induced by differences in geographical distance among sampling. The implication of geographical distance in the speciation process have been widely studied (Mayr, 1963, to name but one), and geographical distance might be seen as a part component of RI (e.g. Mallet and Mullen, 2022). However, this thesis focuses on the genetic components in the evolution of RI and geography is only seen as a putative bias as physical distance between two populations can affect our ability to accurately estimate how the speciation dynamic is shaped by reproductive genetic barriers, this for two reasons: first, geographically distant populations might have lower gene flow, and thus higher divergence, because of the increasing difficulty of physical gamete exchange, without a necessary increase in the number/strength of reproductive barriers. Thus, the geographic distance could mimic the effect of a complete RI (i.e. an absence of recent gene flow). A significant correlation between ongoing migration status and geographic distance would therefore prevent conclusion on the genetic basis of observed RI (but neither would it prove the nonexistence of such a link). Secondly, the geographical distance between samples of a hybriding population and a hybrid zone (zone of physical exchange of gametes between this population and another, in a simple two demes model) may influence the foreign diversity that is detected in the genomic data of those samples. In fact, the further away from the hybrid zone, the more generations are required for a foreign neutral allele to diffuse from the hybrid zone to local individuals (Barton and Hewitt, 1989). In case of a secondary contact, the number of generations required for an allele to diffuse should be negatively correlated with the dispersal capacity, e.g. a neutral allele should diffuse slowly in a population of selfer in comparison with a population of SI because of the reduced gene flow (Wright et al., 2013). For these two reasons, the effect of geographic distance was investigated. Distance's medians between plants and animals was found significantly different, with a lower minimum distance for plants. This result suggests that pairs of plant species have an earlier genetic isolation despite a geographical isolation less important than for animals. Furthermore, the effect of the geographical variable was found to be not significant in the GLM (RI $\sim$ geography + divergence). These results do not indicate the absence of the geographic effects previously discussed, but support for negligible effects of geography on our demographic inferences. Furthermore, the geographic localisation was available for only $\approx 58\%$ of the pairs of plant species (considering only pairs with sufficient probability of ongoing migration). A more exhaustive analysis of the geographical distance effect is still needed to confirm these results.

Speciation-related factors

If the difference in speciation dynamics between plants and animals is not explained by the geographical factor, consideration should be given to which speciation-related factors might explain it. Plants and animals exhibit various distinctions, including factors exclusive to their respective kingdoms (e.g., the presence of chloroplasts in plants) and others more prevalent in one kingdom (e.g. hermaphroditism, and consequently, autogamy, is more common in plants). As a consequence, the influence of these exclusive kingdom-specific factors, such as the presence of chloroplasts, remains constant within the analysis. Regardless of the sampling process, our dataset will inherently consist of plant data with chloroplasts and animal data without chloroplasts, mirroring the natural distribution. However, for other factors, the representativeness of our dataset in the analysis is subject to variations based on the sampling process (e.g. does the selfing rate of species in our dataset accurately represent the global selfing patterns in both plants and animals? Are all levels of speciation-related factors adequately represented?). It is furthermore complicated to genuinely characterise the effects of the speciation related factors as they can be dependent on the environment and/or other speciation related factors, i.e. interaction effects (Anderson et al., 2023). Even for well-represented factors in our dataset, it remains insufficient to accurately disentangle the influence of each factor on speciation dynamics in plants and animals. The complete understanding of how each factor shapes speciation dynamics in animals and plants is expected to be a lengthy and complex journey. This endeavour will necessitate a comprehensive dataset or, more likely, the amalgamation of multiple datasets and studies encompassing all possible factor combinations. Eventually, comparison studies will have to be conducted on (ideally) one factor at a time, starting with evident candidate factors. Therefore, we put forward a list of these factors for further investigation.

Plants/animals contrasting speciation's factors

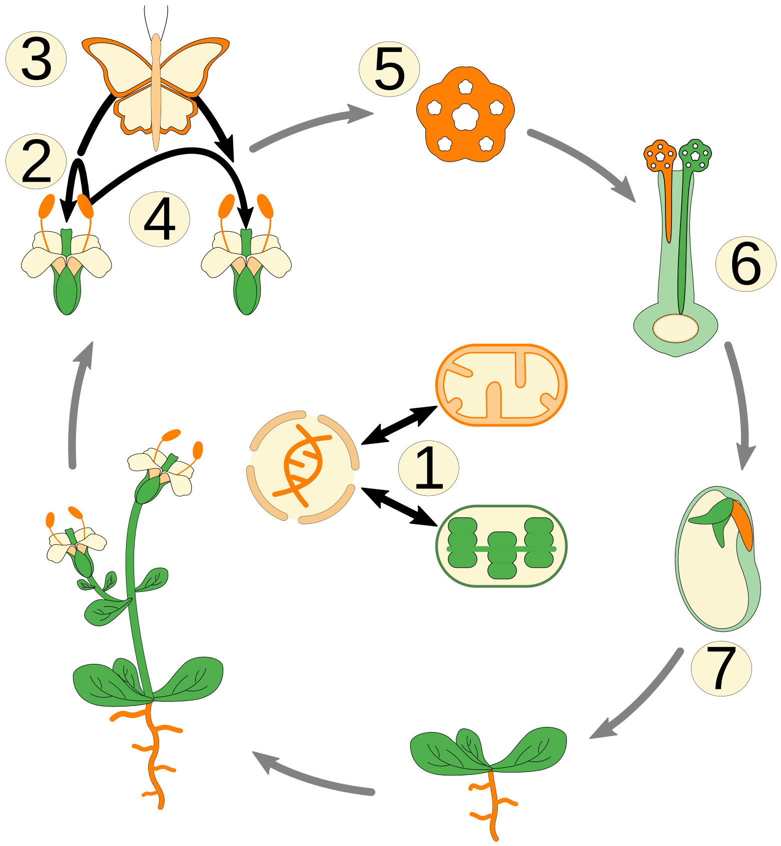


Figure 24: **Contrasting factors between plants and animals.**

This figure presents the different speciation related traits that characterise the plants in contrast to the animals. ① presence of chloroplast; ② predominant self-fertilisation in plants; ③ dependence on external pollinators; ④ different dispersion modalities; ⑤ differences in the strength of haploid selection; ⑥ predominant exposure to environmental pollen; ⑦ parental conflicts due to ubiquitous polyandry in plants.
Credit: Camille Roux.

1. Cyto-nuclear incompatibilities

Mitochondria and plastids are organelles respectively in charge of cellular respiration and photosynthesis (among other things), both essential functions for the eukaryotic cell. They originated after endosymbiosis events billions of years ago (Greiner and Bock, 2013). Along their coevolution with the nucleus genome, an important share of the organellar genes translocated to the nucleus genome, some of them ending up lost due to reduction of redundant function (Rand et al., 2004; Sloan et al., 2018). Even if most of the proteins required for the organellar genome functioning are encoded by the nuclear genome (e.g. nucleus code for at least 90% of plastid proteins, Ferreira de Carvalho et al., 2019), some crucial genes remain coded by organellar genomes (Rand et al., 2004; Sloan et al., 2014, 2018). Because of this entanglement of cytoplasmic/nuclear genomes functioning mechanisms, selection will promote coevolution between organellar and nuclear genomes to achieve proper functioning of the plant cells (Greiner and Bock, 2013; Rand et al., 2004). As cytonuclear coadaptation patterns are specific for each lineage, hybridization will break coadapted genes pairing which may reveal cytonuclear incompatibilities (Sloan et al., 2018). Plastid-nuclear incompatibilities seem common (Barnard-Kubow et al., 2016; Greiner et al., 2011), and may explain in part the difference of speciation dynamic pattern observed between animals and plants since plastids are only found in plants. Mitochondria may also induce cytonuclear incompatibility (Hill, 2016), mitochondrial-associated sterility seems much more common in plants than animals (Rieseberg and Blackman, 2010; Weeks, 2012). Male cytoplasmic sterility (CMS) is 'the maternally-inherited inability to produce functional pollen (male gametes) in individuals from an otherwise hermaphroditic species' (Budar et al., 2003). This cytoplasm-induced sterility is found in gynodioecious species, characterised by the co-occurence of female and hermaphrodite individuals in their populations (Gouyon and Couvet, 1987). CMS arise from a conflict between mitochondrial and nuclear genes, which differ in their mode of transmission (maternal inheritance for the mitochondria) (Murlas Cosmides and Tooby, 1981). Cytoplasmic mutation re-allocating resources from pollen to ovule production will be selected as they increase organellar transmission (Barr and Fishman, 2011; Murlas Cosmides and Tooby, 1981). Once a CMS appears in a population, it may rise in frequency, fixe and lead to the extinction of the population (no male gamete available). Nuclear mutations restoring fecundity (R_f) allows persistence of gynodioecious populations (Kheyr-Pour, 1980, 1981; van Damme, 1983), they should be selected with a strength dependent on the rarity of the male function and on a possible cost of the restoration (Frank, 1989). Different CMS cytotypes can co-occurs in a species, which may differ by the quantity and quality of their CMS mutations as co-adapted Rf mutations (Frank, 1989). Like in plastid-nuclear incompatibilities, hybridization may break pairing of co-adapted mito-nuclear genomes. It has been hypothesised that this hybrid CMS could participate in the creation of post-zygotic RI (Case et al., 2016; Fishman and Willis, 2006; Tiffin et al., 2001). By lowering nuclear gene transmission (male sterile offspring), or if expression or cost of CMS/Rf are linked to environmental factors, or if nucleo-mitochondrial co-evolution set up DMIs, hybrids should have lower fitness than pure individuals of the parental lineages (Greiner and Bock, 2013). This hypothesis is however questioned as it is unlikely that this RI persists. This RI is incomplete (some of the F1, F2 and backcross remain fertile), allowing transmission of CMS cytoplasm, or even nuclear restorer (which would promote introgression) (Fishman and Willis, 2006). The Rf fitness cost should determine in part the strength of RI (or the ease of introgression) (Greiner and Bock, 2013).

2. Selfing rate

An important characteristic of plants lies in their remarkable diversity in mating systems, particularly evident among hermaphrodites, where a broad spectrum of self-fertilisation rates exists, ranging from high-selfers to high-outcrossers (Goodwillie et al., 2005). This may influence speciation as selfing is known to have different roles in the accumulation of RI (Castillo et al., 2016; Cutter, 2019; Grundt et al., 2006; Hu, 2015; Ishizaki et al., 2013; Levin, 1971; Rausher, 2017; Wendt et al., 2002; Wright et al., 2013). For instance, selfing can reduce depth/breath of fitness valley in case of underdominant (B. Charlesworth, 1992) or compensatory mutation (Marie-Orleach et al., 2022), leading to the emergence of two populations separated by hybrid incompatibilities (an alternative scenario to the Bateson-Dobzhansky-Muller model (BDM)). Selfing also eases the emergence of underdominant, compensatory and BDM incompatibilities by reducing the effective population size and increasing genetic linkage (Marie-Orleach et al., 2022). Chromosomal rearrangements are more likely to appear in selfing populations (B. Charlesworth, 1992) and may promote RI (Lynch and Force, 2000; Rieseberg, 2001, see chromosome rearrangement section). Selfing and associated pollen discounting may also promote the emergence of RI by facilitating genomic isolation between selfer populations (Wright et al., 2013). Beside the effect of stable rate of selfing, the process of speciation is also expected to be affected by the transition to autogamous systems. This is especially true as shifts in mating systems tend predominantly towards selfing (Barrett et al., 1996; Goldberg et al., 2010; Stebbins, 1974). The increase of selfing rate is associated with multiple trait modifications associated with an increase in RI, like the loss of nectar production or pollen discounting (Shimizu and Tsuchimatsu, 2015), the so-called 'selfing syndrome' (Sicard and Lenhard, 2011). Genetic element modifications are also associated with selfing transition: reduced individual heterozygosity and polymorphism, increased deleterious mutation accumulation and population differentiation, to cite the most predominantly encountered in plant studies (Cutter, 2019). Transition from SI to SC is also associated with a reduction of sexual conflict. This reduction may lead to the loss of male-specific and sexually selected genes (as other genomic modifications), thereby building divergence between close lineages (Cutter, 2019). By promoting divergence (different selection regime or drift), transition to selfing may increase the chance of DMI emergence (Orr and Turelli, 2001). Selfing may also promote reinforcement, in particular when pollen discounting is associated with selfing (Rausher, 2017). Reinforcement (here in the form of selfing) may be selected in populations with different ecological adaptations as selfer's offsprings should be less exposed to outbreeding depression (Epinat and Lenormand, 2009).

3. Animal pollinator dependence

Despite the historical emphasis on prezygotic isolation through behavioural mate choice as a determinant of reduced interspecies gene flow in animals, it is worth noting the occurrence of specific prezygotic isolation mechanisms in plants too. First, phenological shifts between related plant taxa can cause allochronic isolation (Devaux and Lande, 2009). Second, plant species with animal pollination can exhibit pollinator shifts between closely-related species, playing in such situations a significant role in reproductive isolation (Kay and Sargent, 2009). Most of the angiosperms rely on pollinators for their fertilisation (Friedman and Barrett, 2009; Ollerton et al., 2011). Shift in pollinators-plant associations is thought to be an important driver of plant speciation (Crepet, 1984; Darwin, 1862; Grant, 1949; S. D. Johnson, 2010; Nosil, 2012; Stebbins, 1970; Vamosi and Vamosi, 2010; Van der Niet et al., 2014) and have been investigated in numerous study (V. Grant and K. A. Grant, 1965; Schemske and Bradshaw, 1999; Van der Niet et al., 2014; van der Niet and Johnson, 2012). Differences in pollinator ecology of a species range can promote adaptation of linked floral traits, which in turn may lead to the apparition of pollinator-mediated RI between differentially adapted lineages (Van der Niet et al., 2014), an effect on reproductive isolation reflected at the macroevolutionary scale with higher diversification rates in plants which are associated with fewer pollinator species (Schiestl and Schlüter, 2009). This type of ecological speciation is expected with an initial allopatric phase since sympatry would require a sudden switch in pollinator adaptation to allow RI to persist (Coyne and Orr, 2004; Kay and Sargent, 2009). Consistently, most of the pollinator-mediated RI reported in the literature are case of allopatry (Armbruster and Muchhala, 2009; Kay and Sargent, 2009), with a few case of putative emergence in sympatric (for example with orchids Xu et al. (2012)). Although pollinator-mediated barriers are expected to be insufficient for a complete RI (Kay and Sargent, 2009), and are usually reported with other reproductive barriers (Armbruster and Muchhala, 2009), pollinator-mediated RI remain considered as an important contributor to angiosperm speciation and most likely the more common mechanism of ecological speciation (Van der Niet et al., 2014).

4. Dispersal capacity

Plants and animals obviously differ in their modes of dispersal, with two types of propagules, pollen and seeds, passively dispersed through abiotic and biotic vectors in plants, as compared to individual (or mother-mediated dispersal for some mammals, Tiedemann et al., 2004) mobility in animals. The overall effect of such differences is that the extent of gene flow among populations within plant species is lower (and genetic differentiation is higher) on average than in animals (Morjan and Rieseberg, 2004). This may result from overall differences in dispersal kernels between plants and animals, but also from the stronger stochasticity of dispersal in plants (Nathan, 2006) as compared to animals (excluding passive dispersers, e.g., marine animals with a planktonic larval dispersal), which results in higher effective migration in the latter. Stronger genetic differentiation among conspecific populations may trigger the evolution of partial reproductive isolation, a phenomenon known as outbreeding depression, which is thought to be more common among plant than animal populations (Edmands, 2007).

5. Haploid pollen gene expression

During their reproductive cycle, plants and animals alternate between diploid and haploid phases. Haploid genomes undergo selection differently than diploids in that selection purifies more efficiently recessive deleterious mutations revealed by homozygosity (Mable and Otto, 1998). It has been argued that haploid selection might be stronger for plants as their gametic phase differs by many factors (Immler, 2019; Joseph and Kirkpatrick, 2004; Otto et al., 2015; Rieseberg, 2001; Turelli and Moyle, 2007). On one hand, animal sperm cells are genetically haploid but remain phenotypically diploid for the major part of their development as they share mRNA and protein through cytoplasmic bridges (Caldwell and Handel, 1991; Dym and Fawcett, 1971; Jeon, 2004; Joseph and Kirkpatrick, 2004). DNA is also expected to be highly packed into animal sperm cells, with a greatly reduced gene expression (Steger, 1999). Haploid selection is therefore expected to be low in animal male gametophyte, even if recent studies temper this statement (Immler, 2019; Joseph and Kirkpatrick, 2004; Otto et al., 2015). On the other hand, pollen is thought to express a higher percentage of its genome comparatively with sperm cells (Arunkumar et al., 2013; Immler, 2019; Joseph and Kirkpatrick, 2004; Otto et al., 2015; Rieseberg, 2001), mostly for germination, pollen tube growth and interaction with the pistil (Rutley and Twell, 2015, also see pollen-pistil interaction section), providing more material for haploid selection to work on.

6. Pollen-pistil reproductive barriers

In order to fertilise an oocyte, a sporophyte has to produce a viable male gametophyte, this pollen grain has to reach the stigma of a pistil and successfully deliver the two sperm cells to the oocyte. This last stage is not devoid of potential reproductive incompatibilities as the pollen has to successfully interact with three parts of the pistil (stigma, style, ovary). Pollen-pistil interactions can be ineffective in crosses between distant species, either by mechanical mismatch (e.g. between the style and the pollen tube growth rate, Kuboyama et al., 1994) or genetic mismatch (e.g. maize recognition genes, Kermicle and Evans, 2005). These passive mismatches are referred to as incongruity (Hogenboom et al., 1997). In contrast, pollen-pistil mechanisms of inbreeding avoidance actively reject genetically close pollen and are referred to as self-incompatible (SI, Takayama and Isogai, 2005). Pollen-pistil incongruity and incompatibilities are not exclusives and can both cause RI specific to plants (Broz and Bedinger, 2021). Once the pollen lands on the pistil, it has to stick to the stigma, hydrate, germinate and grow its pollen tube to reach the transmitting tissue of the style. All of these steps can potentially passively or actively (SI) contribute to reproductive barriers (Broz and Bedinger, 2021). Incompatibilities are also encountered in pollen-style interaction, with rejection in SI systems (e.g. Baek et al., 2015) and SC systems (e.g. Broz et al., 2017). Finally, pollen-ovary interactions can also promote reproductive barriers by species-specific pollen tube attraction by ovules (Higashiyama et al., 2006; Takeuchi and Higashiyama, 2012; Uebler et al., 2013) or pollen tube perception (Escobar-Restrepo et al., 2007; Williams et al., 1986). Most of the studies that reveal pollen-pistil incompatibilities are based on Brassicaceae, Poaceae or Solanaceae and some mechanisms are not fully understood (Broz and Bedinger, 2021), but this reproductive barriers could be widespread in plants and contribute in the difference of reproductive isolation pattern observed between plants and animals. The role of pollen-pistil interactions in the difference of speciation pattern observed between plants and animals might be tempered by the existence of similar features in animals (e.g. pollen tube growth rate $\approx$ sperm speed, Cutter, 2019).

7. Endosperm incompatibility

One characteristic of angiosperms is the usual presence of endosperm, a crucial seed tissue which provides nutrient reserves to support embryo development (Yan et al., 2014). Deficiency in this tissue can lead to the inviability of seeds and have been observed in different crossing studies (Coughlan and Matute, 2020; İltaş et al., 2021; Lafon-Placette et al., 2017; Oneal et al., 2016; Rebernig et al., 2015; Roth et al., 2018; Sandstedt and Sweigart, 2022). Parental conflict has been hypothesised as the cause of these deficiencies (Haig and Westoby, 1991). In outcrossing species, optimal seed's resource allocation strategy differs between maternal and paternal gamete donors. From the father's perspective, gene expression promoting a preferential maternal resource allocation in their seeds should be advantageous, where gene expression restoring equality in the resource allocation between seeds should increase maternal gene transmission (Brandvain and Haig, 2005). Selection should therefore favour the accumulation of parent-of-origin biassed gene expression by imprinting mechanism (Batista and Köhler, 2020; Haig and Westoby, 1991; Kinoshita, 2007; Reik and Walter, 2001). When brought together in crosses, genomic backgrounds of different parental conflict histories can induce seed development failure, a postzygotic barrier common in angiosperms (Lafon-Placette et al., 2017).

Polyploidization and speciation

Polyploidization, or Whole Genome Duplication (WGD), are phenomenons known from more than a century (Lutz, 1907) and which have been extensively studied in plant speciation (Bock et al., 2023). Hybrids triploids formed by the cross of tetraploids (resulting from a WGD event) and diploids will often suffer reduction of fitness caused by the failure of endosperm development, and meiosis's failure (Baack et al., 2015), thus promoting speciation. Conversely, diploidization is the process through which a polyploid lineage returns to a diploid-like state. This process relies on different mechanisms such as genome rearrangements or genome downsizing (i.e. reduction of paralogs) (Doyle et al., 2008; Wendel, 2015). These mechanisms can also promote speciation (e.g. differential loss of gene copies can result in the emergence of incompatibilities, Scannell et al., 2006). The completion of a diploidization can extend from thousands to millions of generations (Bock et al., 2023).

Here, we only consider diploid species to match the framework of Roux et al. (2016), thus, we excluded those factors from the list of the best candidate factors to explain the difference of dynamics of speciation between plants and animals.Although polyploidization is thought to be involved in up to $\approx 15\%$ of angiosperm speciation events (T. E. Wood et al., 2009), and diploidization is expected to be faster in plants than animals (Z. Li et al., 2021) and could be achieved in a few generations (e.g. Shi et al., 2023), we consider it unlikely that an undetected WGD swiftly followed by a return to diploidization concerns numerous pairs of species of this dataset. For this reason, we do not extend further on this factor as a putative explanation for the observed difference of dynamics of speciation between plants and animals.

Conclusion

Based on a novel approach, the present comparative study revealed that (i) complete reproductive isolation emerges at a level of net divergence ~ 5 times lower for plants in comparison with animals, suggesting a faster speciation in the plant kingdom, and that (ii) contrastively with animals, plant species seem to initially accumulate divergence mostly in allopatry. These results challenge the historical assumption on the better capacity of plants (in contrast to animals) to introgresse despite the accumulation of genetic divergence. They also support further investigation in a list of speciation related factors that could explain the difference observed between the dynamics of speciation of plants and animals. Finally, this study emphasised the usefulness of a new approach, encouraging its application in future comparison studies.

Chapter II

Introduction

The previous chapter presented the dynamic of speciation observed in plants. This dynamic was compared to the one observed for animals, and different traits were proposed as non-exclusive explanation for the difference observed between the dynamics of speciation of the plants and of the animals. However, studying the effect of these factors on the dynamic of speciation is not trivial as it requires the development of appropriate approaches to analyse the distribution of ongoing migration status of pairs of species. In this chapter, the effects of selfing rate and of the life form (two traits related to speciation) are tested as variables to explain the ongoing migration status of the plant pairs of species from the previous chapter. An approach is proposed to realise this type of analysis.

Plants reproduce through a variety of different mating systems. Some species are dioecious, where each individual carries only male or female reproductive organs, but most of the species ($\approx 94\%$, Renner and Ricklefs, 1995) are monoecious (individuals are functionally bisexual) and are called hermaphroditic when their flowers bear both sexes. Mating systems 'in between' monoecy and dioecy can also be encountered, such as gynodioecy, where populations are made of hermaphroditic and female individuals, but they remain rare in contrast with hermaphroditism (e.g. only 2 % of angiosperm genera are gynodioecious, Dufaÿ et al., 2014). The rate of outcrossing (the proportion of fertilisation with non-self individuals) varies widely among hermaphroditic species but can also vary substantially among populations within species (Whitehead et al., 2018). The distribution of outcrossing rate among species was found to be U shaped in seed plants, with only 36 to 42% of the species with an outcrossing rate between 0.2 and 0.8 (Goodwillie et al., 2005; Igic and Kohn, 2006; Whitehead et al., 2018). On one side, outcrossing is frequently enforced by Self-Incompatibility (SI) systems, where self-fertilisation is prevented by a rejection of the pollen by the pistil (De Nettancourt, 2001) and which ensures inbreeding avoidance. On the other side, many self-compatible (SC) species exhibit high rates of self-fertilisation, which provide two benefits: the transmission advantage, that is the double transmission of alleles from a parent to its selfed offspring (Fisher, 1941), and the reproductive assurance that the parent will find a mating partner (itself) even in cases of low population density or absence of pollinators (Jain, 1976). Although SI provides a long-term advantage by reducing inbreeding, the SI to SC transition is the most commonly observed (Igic et al., 2008; Stebbins, 1974). The loss of self-incompatibility and transition to selfing is associated with a documented increase in extinction rate (Goldberg et al., 2010) and is sometimes referred to as an 'evolutionary dead end' (Stebbins, 1957). The higher extinction rate has been suggested to result from a reduced effective recombination and effective population size that drive the accumulation of weakly deleterious mutations and reduce environmental adaptation capacity (Burgarella and Glémin, 2017; Wright et al., 2013). The increase of selfing was also found to be associated with an increase in speciation rate (Goldberg and Igić, 2012; Goldberg et al., 2010). Indeed, the most direct effect of selfing on the speciation process is the reduction of gene flow between populations, as individuals favour autogamy and lessen pollen dispersal (Barrett et al., 1996; Levin, 1971). Populations of selfing lineages tend to be more structured (Hamrick and Godt, 1996), with reduced effective size and recombination rate. This favours the evolution of RI by drift, but can also reduce the accumulation of RI due to local adaptation as selection will be less efficient (Gavrilets, 2004). Overall, underdominant mutations, compensatory mutations and BDMi should accumulate more easily in selfing allopatric populations, even in the face of local adaptation (Marie-Orleach et al., 2022). Aside of the effect of reinforcing allopatry between extant populations, selfing is also expected to improved colonising ability after long distance dispersal by the purge of inbreeding depression (Sachdeva, 2019) and the reproductive assurance property (Baker, 1955; Pannell and Barrett, 1998), promoting furthermore circumstances of allopatry in selfing lineages. Besides differences in patterns of speciation between selfing and outcrossing lineages, speciation is also expected to be influenced differently depending on the mating system combinations (Pickup et al., 2019). The most well-known combination being the SI x SC rule, an unilateral incompatibility where the SI species are more likely to reject pollen from selfer species than the opposite (Lewis and Crowe, 1958). SC x SC hybridisations might also result in unilateral gene flow but the influence of ecology or demography on this asymmetry is still poorly understood, and SI x SI hybridations where found to favour introgression through the advantage of rare S alleles (Castric et al., 2008; Pickup et al., 2019). Even if complete RI is not expected from the simple effect of any of these combinations, speciation should generally be promoted in any combination involving SC.

Based on all these arguments, we expect a faster dynamic of speciation in selfing lineages as compared to outcrossers, an expected pattern that we could potentially test with our plant population genomic dataset provided that we could get reliable information on the mating system and that substantial variation in mating system does occur in our dataset.

Life form is one of the most investigated traits to explain variation in plant speciation rate (Helmstetter et al., 2023), most likely because of its simplicity of identification and because it is correlated with multiple traits affecting speciation. Plants are usually categorised into three life forms: tree, shrub and herb (Petit and Hampe, 2006). The speciation rate is expected to differ particularly in trees comparatively to shrubs and herbs as several tree traits are more prone to promote speciation. First, most of their reproductive modes are allogamous (Hamrick and Godt, 1996), suggesting that their speciation rate should be lower comparatively to other life forms since selfing mostly promotes speciation (for the reasons discussed previously). Second, trees experience greater gene flow among their populations due to both higher pollen and seed dispersal (Petit and Hampe, 2006), hence limiting allopatry and by extension the accumulation of RI. Genetic diversity, which is not unrelated to previously discussed traits, is also expected to be greater in trees (Carvalho et al., 2019), suggesting lower incidence of genetic drift and higher effective recombination, which would delay the process of accumulation of RI.

From the different factors that are expected to shape the dynamics of speciation in plants, the mating system is particularly interesting as it can be investigated through selfing rates directly estimated from genetic data. This is convenient as the botanical literature lacks information for most of the species in the present data set (e.g. the information of the sexual system was found for less than 50% of the species investigated here, the pollination mode/vector for less than 15%; Sylvain Glémin, personal observation). Information about life form, however, is more easy to obtain from the botanical literature. In this study, we use our plant population genomic dataset described in chapter I (210 pairs of congenera plant species[5]) to investigate the effects of the selfing rate and of the life form on the dynamics of speciation by testing their use as explanatory variables to predict the probability of ongoing migration of pairs of species in relation to molecular divergence. As selfing tends globally to promote speciation, it is expected that, for similar levels of divergence, the probability of ongoing migration within pairs would decrease with the increase in selfing rate (a tendency that might be re-enforced by the higher extinction rate of selfing lineages). The life form is known to be correlated with factors linked to speciation (outcrossing, dispersal capacity...; Anderson et al., 2023; Petit and Hampe, 2006. The effect of the life form on the probability of ongoing migration was thus investigated with the a priori that tree pairs of species should have a slower dynamics of speciation.

Methods

Obtaining estimates of the selfing rate for individual species of the dataset

Two different approaches were used to obtain indirect estimates of the mating system of the studied plant species based on the polymorphism data gathered for demographic inferences. First, the selfing rate was estimated from measures of the inbreeding coefficient (F_{is}), a statistic that aims to capture the share of homozygosity attributable to inbreeding. Once an estimate of the F_{is} of a population has been obtained, the selfing rate s can be inferred from the formula

$$s = \frac{2F_{\mathrm{is}}}{1 + F_{\mathrm{is}}}$$

(Wright, 1984). Second, the selfing rate was inferred from estimates of the identity disequilibrium, a statistic measuring the departure from random association between alleles at different

[5] Note that only the pairs of plant species with sufficiently strong probability of ongoing migration inferred were used for the analysis of this chapter.

loci (David et al., 2007). Identity disequilibrium, captured by the statistic g_2 (the heterozygosity disequilibrium between two loci), can be used to estimate s through the formula

$$\hat{s}_{g_2} = \frac{1 + 5\hat{g}_2 - \sqrt{1 + 10\hat{g}_2 + 9\hat{g}_2^2}}{2\hat{g}_2}$$

(David et al., 2007). As the data from this study consist of a mix of datasets with different characteristics (differences in sequencing method, number of samples per population...) that might differentially influence the selfing estimation depending on the method, three tools were tested to estimate the selfing rate of our populations/species. Individual F_{is} were calculated either with the Rpackage *Hierfstat* (Goudet, 2005) using the fasta files produced with *reads2snp* (Gayral et al., 2013; Tsagkogeorga et al., 2012), or with *VCFtools* (Danecek et al., 2011) using the vcf files (also produced with *reads2snp*). Identity disequilibrium (estimated through the $\hat{g}_2$ statistic, see David et al., 2007) was calculated using the Rpackage *InbreedR* (Stoffel et al., 2016). The multilocus average F_{is} was calculated for each species from individual F_{is} obtained with *Hierfstat* and *VCFtools*, then transformed into selfing rate using the formula given above. As it, most of these F_{is} values could not be used since their transformation into selfing rate would result in negative or infinite s values (see Fig. 25). Negative values of F_{is} were therefore considered as 0.

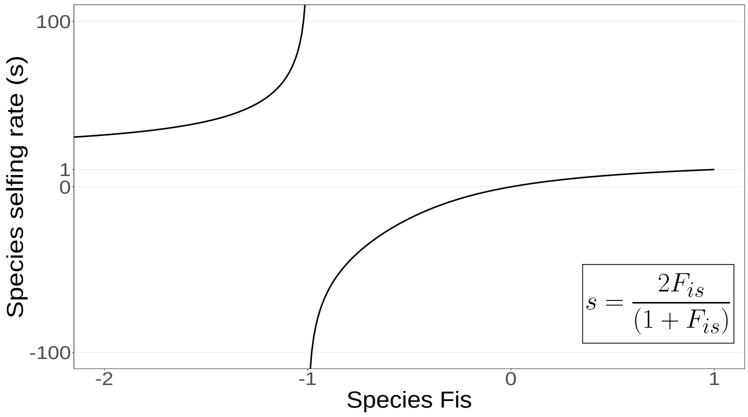


Figure 25: **Selfing rate function.**

This graph illustrates how the selfing rate s varies with F_{is} using the equation applied in this analysis. For negative F_{is} values, as often seen with *Hierfstat* or *VCFtools*, selfing rates can exceed 1, drop below 0, or become undefined at $F_{\text{is}} = 1$. Notably, this equation yields particularly anomalous selfing rates as F_{is} approaches -1.

Obtaining life form status

Information on traits relevant to the dynamics of speciation (e.g. pollination mode, dispersal capacity...) can be difficult to obtain, in particular for non-model species. For the data set of the present study, only the information of form (simplified in three categories: tree, herb or liana/shrub) could be found for each species and was provided by Sylvain Glémin by querying http://www.worldfloraonline.org and scientific sources.

The net divergence is an explanatory variable associated to pairs of species, where selfing rate and life form are associated to species individually. As it was necessary for the statistical analysis to have a single value per pair and per variable, selfing mean was calculated for each pair and the identities of life forms of each pair were simplified as a single value (since the life form was always shared by all the species of a genus).

Statistical analyses

An analysis of the effect of selfing rate and life form on the dynamics of plant speciation was conducted using the approach presented in Chapter I. Because this approach relies on the comparison of sigmoids of taxa, the continuous selfing rate had to be transformed into a categorical variable with three quantiles of selfing rate, thus dividing the dataset into three groups of equal size with increasing selfing rates. Linear models (GLM, binomial family) were fitted to these datasets ($M_{1st\ quantile}$, $M_{2nd\ quantile}$ and $M_{3nd\ quantile}$), as for the complete dataset (M_0), and the difference of sigmoids was tested with a log-likelihood ratio test. To ensure the absence of 'random lineage-specific effects,' i.e. the effect of other variables non-homogeneously distributed in each quantile or life form, additional log-likelihood ratio tests were conducted with mixed lineages ($M_{1\&2\ quantiles}$, $M_{1\&3\ quantiles}$ and $M_{2\&3\ quantiles}$). Similarly, linear models were fitted to the life form datasets (M_{herb} and M_{tree}), and their differences were tested with a log-likelihood ratio test. Note that the samples with liana or shrub life form were not used as they were too few for their divergence range to cover the speciation continuum.

In a second phase, to compare the results obtained with the sigmoid comparison approach with a more traditional framework of statistical analysis and to assess the effects of both variables in models that account for both variables, a linear regression analysis was conducted using PGLMM (Phylogenetic Generalised Linear Mixed Model, from the R package *phyr 1.1* (D. Li et al., 2020). The choice of using PGLMM was based on its capacity to be used with a binomial distribution (in this case, the ongoing migration status of each pair) and to include random effects to account for phylogenetic non-independence. Taking phylogenetic non-independence into account is important because not every trait that could influence the dynamics of speciation was measured. As closer lineages are more likely to share similar trait values, their estimated probabilities of ongoing migration are expected to diverge from the global mean not only because of the effect of the measured traits (fixed effect), but also because of unmeasured traits for which non-independent lineages share similar values. Not accounting for phylogenetic effect can result in underestimated standard errors, which in turn decrease the *P*-value, and increase the risk of type I error (wrongly rejecting the null hypothesis). The results are non-independent for several reasons. First, as mentioned previously, results shared different levels of phylogenetic covariance. Part of this covariance is measured in the fixed effect of net divergence between species of each pair. As for the covariance between genera, it can be accounted for with a Variance CoVariance matrix (VCV matrix). The production of this VCV matrix requires a phylogeny which was obtained with Timetree Kumar et al., 2022. As this phylogeny is used to account for the phylogenetic covariance among genera, it should ideally only contain distance among genera since the genetic distance among species of each pair is a fixed effect of the model. Unfortunately, the information of resolved phylogeny at the genus level is not available for the genera of this study. Therefore, a single species (with taxonomic information available on NCBI) was randomly selected per genus, and the list of species was used on Timetree to obtain a phylogeny in newick format. This approach modifies the genera VCV matrix. To illustrate this with an example, a VCV matrix can be calculated for an example of phylogeny (Fig. 26). A VCV matrix is obtained by summing the branch length between the most recent common ancestor of the two tips and the global common ancestor of the tree (Tab. 4).

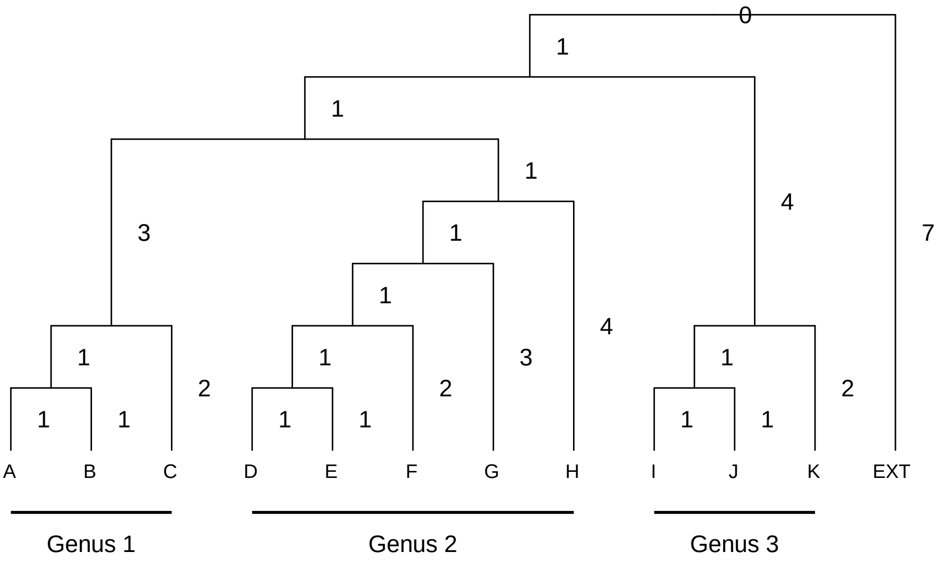


Figure 26: **Example of phylogeny.**

Example of a dendrogram with 3 genera of different sizes. The numbers on the branches represent arbitrary phylogenetic distances with no units.

Table 4: **VCV matrix of the example tree.**

	A	B	C	D	E	F	G	H	I	J	K
A	7	6	5	2	2	2	2	2	1	1	1
B	6	7	5	2	2	2	2	2	1	1	1
C	5	5	7	2	2	2	2	2	1	1	1
D	2	2	2	7	6	5	4	3	1	1	1
E	2	2	2	6	7	5	4	3	1	1	1
F	2	2	2	5	5	7	4	3	1	1	1
G	2	2	2	4	4	4	7	3	1	1	1
H	2	2	2	3	3	3	3	7	1	1	1
I	1	1	1	1	1	1	1	1	7	6	5
J	1	1	1	1	1	1	1	1	6	7	5
K	1	1	1	1	1	1	1	1	5	5	7

Note that this is the VCV matrix for the species phylogeny. Two approaches can be used to obtain a genus VCV matrix. Either by approximating it using the tip of a species in place of its genus node (the current approach) or by using the genera nodes (not possible in the present study) (Fig. 27, based on species tips (left) or genus nodes (right)).

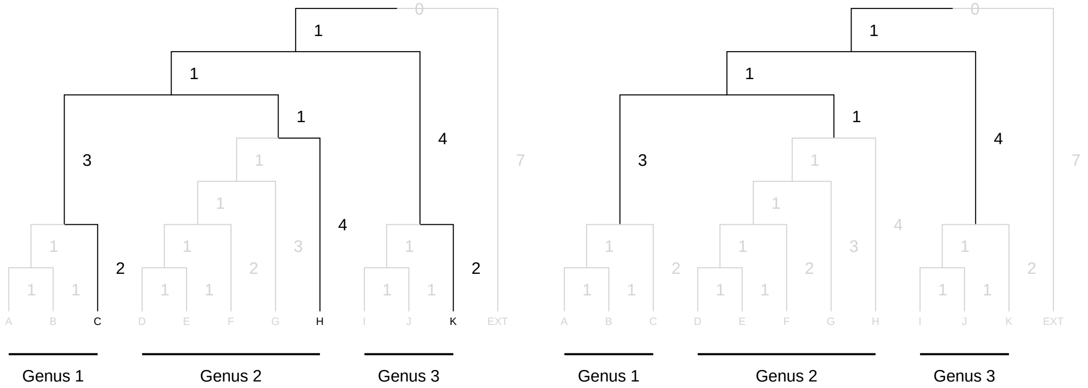


Figure 27: **Genera phylogenies, based on species tips (left) or genus nodes (right).**

Both of those phylogenies are based on the example of phylogeny previously presented (Fig. 26). The left tree illustrates a 'genus tree' based on the tips, thus accounting for the length of the external branches. This tree keeps the ultrametric configuration. The right tree illustrates a genus tree based on the genus nodes, where the external branches are not used.

Table 5: **Genera VCV matrices (left: by tips, right: by nodes).**

	Genus 1	Genus 2	Genus 3		Genus 1	Genus 2	Genus 3
Genus 1	7	2	1	Genus 1	5	2	1
Genus 2	2	7	1	Genus 2	2	3	1
Genus 3	1	1	7	Genus 3	1	1	5

Only the diagonal of the VCV matrices directly depends on the approach. However, the VCV matrix has to be standardised to have its determinant equal to 1 before its use in a PGLMM (Ives, 2019). The standardisation applied by the function *PGLMM()* is

$$matrix_{standard} = \frac{matrix}{det(matrix)^{\frac{1}{n}}}$$

with n the number of species/genus of the matrix.

Table 6: **Standardized genera VCV matrices (left: by tips, right: by nodes).**

	Genus 1	Genus 2	Genus 3
Genus 1	1.04	0.30	0.15
Genus 2	0.30	1.04	0.15
Genus 3	0.15	0.15	1.04

	Genus 1	Genus 2	Genus 3
Genus 1	1.35	0.54	0.27
Genus 2	0.54	0.81	0.27
Genus 3	0.27	0.27	1.35

Once standardised, the matrices differ completely in their values, and by extension in their way of influencing the estimation of the PGLMM. However, the difference of diagonals between these two approaches is considered negligible for the current data as on average diagonal values of the 'node phylogeny' are only 3% lower than those of the 'tips phylogeny'.

Another non-independence issue stems from the very nature of the results. Each result is an ongoing migration status estimated from a pair of species, and each species (except when a genus has only two species) is present in multiple pairs. Similarly with the phylogenetic non-independence, the ongoing migration status of pairs is dependent not only from the measured effects, but also from unmeasured effects. For a genus comprising three species (A, B and C), three pairs of species can be built (A-B, A-C, and B-C). The species A might carry particular values of unmeasured effect which greatly influence the migration status of the pairs including A. Not accounting for this could result in incorrect coefficient estimations and/or type I error. The non-independence due to the 'paired nature' of the results could in principle be accounted for in the PGLMM the same way the phylogeny (at the genus level) is. PGLMM not only accepts phylogenetic VCV matrices but any VCV matrix (for example correlation between samples induced by the distance between sites of sampling). The model would therefore include either two matrices, one for the phylogeny and one for the species correlation, or one matrix made of the sum of both matrices with coefficients to weight the influence of each. Unfortunately, the attempt to apply this solution failed because of the VCV matrix requirement for PGLMM. The VCV matrix needs to be transformed with the Cholesky decomposition (Benoit, 1924) which requires a positive-definite matrix. Although it may be perhaps possible to do so, attempts to obtain a positive-definite pairs matrix were unsuccessful.

Another possible solution might be incorporating a random effect associated with the species. This can be easily achieved in a PGLMM by adding a random effect based on the information of species. This random effect has to be split into two random effects because each pair of species 'belong to two species'. This causes an issue since each species must be specified in a single random effect. As pictured in an example with a genus made of three species (Tab. 7), there is no configuration for which each species only belongs to a single column. In the first configuration of the example, the non-independence of the pairs A-B and B-C caused by the shared B is not accounted for because B does not belong to a single column (i.e. a single random effect).

Table 7: **Impossibility of mono-column for each species.**

Issue.with.B	Issue.with.C	Issue.with.A
A - B	A - B	A - B
A - C	A - C	C - A
B - C	C - B	C - B

A way of getting around this issue is to subsample results in order to ensure that each species is only represented in one random effect. This implies a loss of nearly half of the results, with large genera more affected than small ones (Tab. 8, Fig. 28). The subsampling of the data might change the estimations obtained with PGLMM. To ensure that the conclusion deduced from the PGLMM results remains the same independently of the sampling, multiple subsampling were drawn and their PGLMM results were summarised.

Table 8: **Pair sub-sampling to ensure a single column per species.**

Species	AB|C	AB|CD	ABC|DE	ABC|DEF	ABCD|EFG
unique pairs	A - B A - C B - C	A - B A - C A - D B - C B - D C - D	A - B A - C A - D A - E B - C B - D B - E C - D C - E D - E	A - B A - C A - D A - E A - F B - C B - D B - E B - F C - D C - E C - F D - E D - F E - F	A - B A - C A - D A - E A - F A - G B - C B - D B - E B - F B - G C - D C - E C - F C - G D - E D - F D - G E - F E - G F - G
Proportion of remaining pairs	2/3 0.667	4/6 0.667	6/10 0.6	9/15 0.6	12/21 0.571

Note: pairs in red are removed

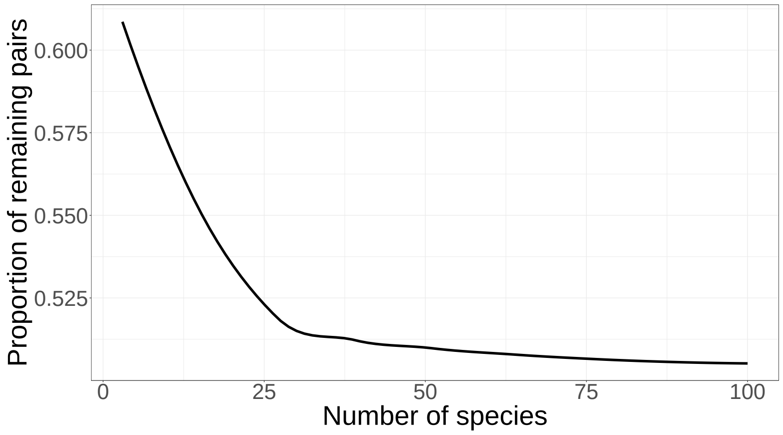


Figure 28: **Proportion of remaining pairs in function of the number of species in a genus.**

As the number of species in a genus (x-axis) increases, the proportion of remaining pairs (y-axis) for that genus, once subsampled as in the table 8, decreases. This reduction in the proportion of remaining pairs is less significant for a higher number of species.

PGLM models were runs with different sub sampling using this command:

```
pglmm(ongoing_migration_status ~ net_divergence + selfing_rate + form + (1|spA) +
(1|spB) + (1|genus__), data = subsample, cov_ranef = list(genus = phylogeny), family
= 'binomial')
```

The statistical significance of the fixed effects (selfing rate and life form) were directly extracted from the output of PGLMM, the *P*-values of the random effect were estimated with the function *pglmm_profile_LRT* from the R package *phyr* (D. Li et al., 2020). The R^2 were estimated with the function *R2* from the R package *rr2* (Ives, 2019; Ives and Li, 2018). The residual model assumptions were checked with the function *simulateResiduals* from the R package *DHARMa* (Hartig, 2022).

Results

The F_{i} values calculated ranged from -6.4 to 1 with *Hierfstat*, and from -0.83 to 0.83 with *VCFtools* (Fig. 29). The mean of Hierfstat results was at -0.01 (median at 0.04) and the mean of VCFtools results were at 0.148 (median at 0.14). Once transformed into selfing rate (s), *Hierfstat*'s s values ranged from -50.98 to 112.02 and *VCFtools*' s ranged from -1.32 to 0.70 (Fig. 30). The mean of *Hierfstat* s values was -0.16 (median ≈ -0.28) and the mean of *VCFtools* s values was 0.01 (median ≈ 0.03). Distributions of the three sets of s estimates (with *Hierfstat* and *VCFtools*, negative F_{i} values were set to 0) were similar (Fig. 31), but as the negative values of F_{i} could be in part due to sequencing or mapping/variant calling error, the *InbreedR* method was preferred as it is supposed to be less sensitive to these artefacts (David et al., 2007). Noticeably, the U shape of outcrossing rate usually reported in the scientific literature (Goodwillie et al., 2005) was not found in these distributions.

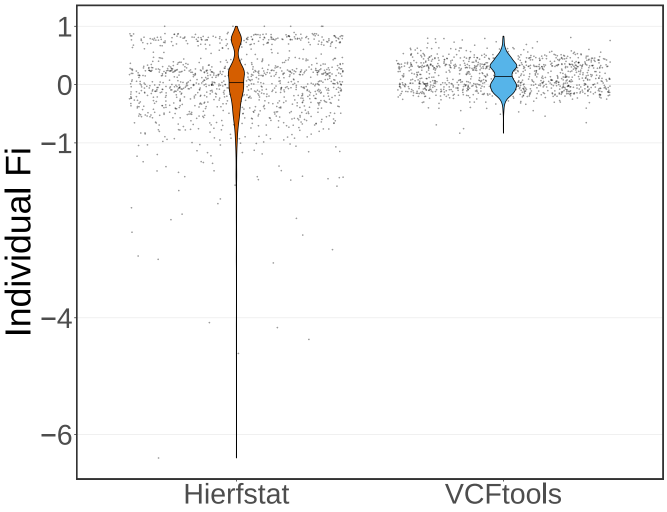


Figure 29: **individual F_i per method.**

Distribution of the Fi of each sample (y-axis, peudolog10) per method (x-axis). Each dot is a sample, and the median of each distribution is represented with a horizontal bar, at ≈ 0.04 for *Hierfstat* and ≈ 0.14 for *VCFtools*.

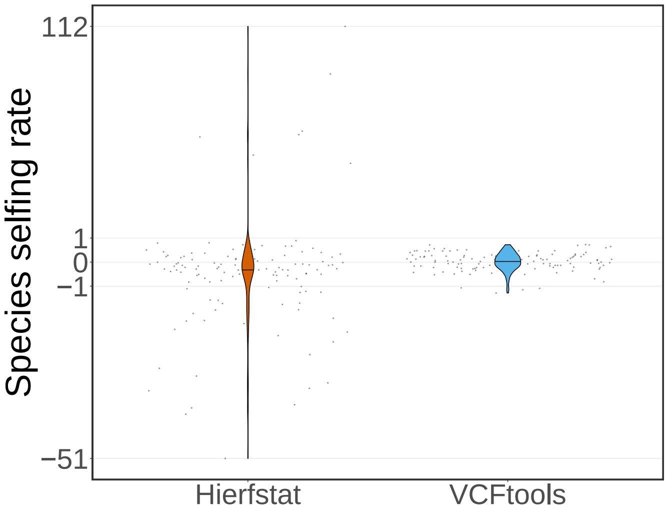


Figure 30: **species' *s* per method.**

Distribution of the mean selfing rate s for each species (y-axis, pseudolog10) per method (x-axis). Each dot is a species, and the median of each distribution is represented with a horizontal bar, at ≈ -0.28 for *Hierfstat* and ≈ 0.03 for *VCFtools*.

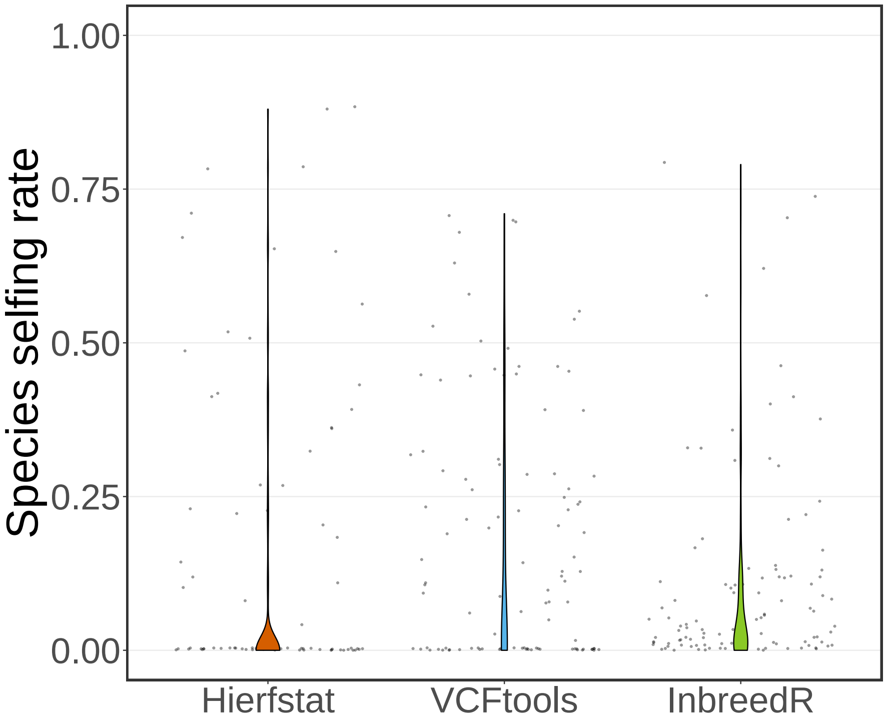


Figure 31: **selfing rate per method, transformed negative values.**

Distribution of the mean selfing rate s for each species (y-axis) per method (x-axis), each dot is a species. The negatives values of *Hierfstat* and *VCFtools* have been transformed into 0.

The estimates of mean selfing rates estimated with *InbreedR* for each pair of species were distributed in 3 quantiles. The quantile ranges were [0,0.04], [0.04,0.1] and [0.11,0.75], with their respective median at 0.016, 0.066 and 0.221 (Fig. 32). The net divergence of the pairs of species in each quantile ranged from, [0.0000, 0.0111] for the 1st quantile, [0.0000, 0.0117] for the second and [0.0000, 0.0551] for the third (Fig. 33).

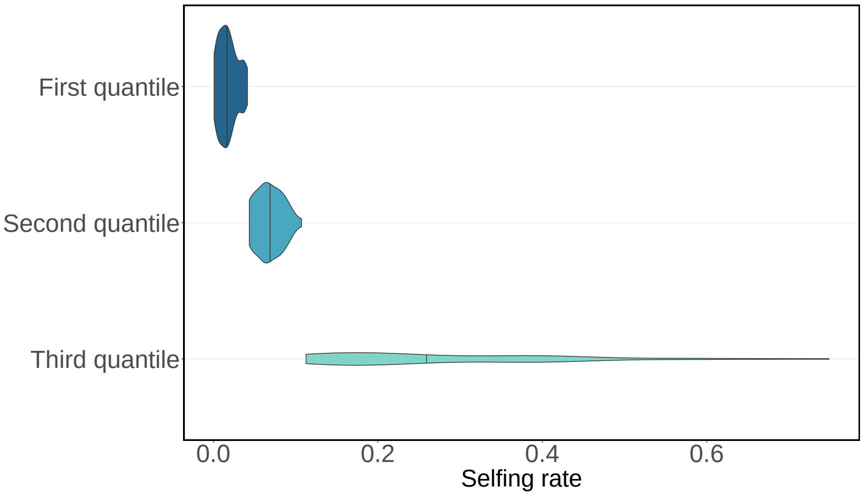


Figure 32: **Mean selfing quantiles distribution.**

Distribution of the mean selfing rate of the pairs of species in the three quantiles (from low to high selfing rates). The medians of the quantiles are represented with a vertical line, with $\tilde{s}_{1st\ quantile} \approx 0.016$, $\tilde{s}_{2nd\ quantile} \approx 0.066$ and $\tilde{s}_{3rd\ quantile} \approx 0.221$.

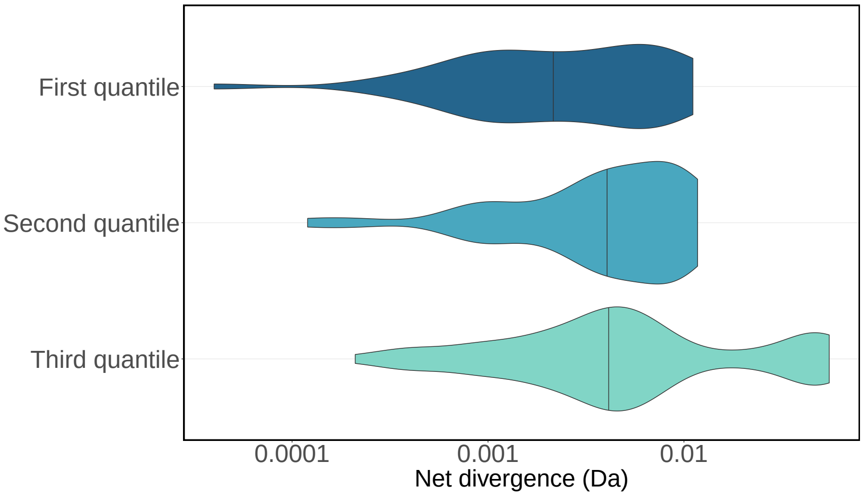


Figure 33: **Divergence quantiles distribution.**

Distribution of the selfing rate quantiles in function of the net divergence. The medians of the quantiles are represented with a vertical line, with $\tilde{da}_{1st\ quantile} \approx 0.00254$, $\tilde{da}_{2nd\ quantile} \approx 0.00453$ and $\tilde{da}_{3rd\ quantile} \approx 0.0043$.

Based on the linear models (GLM, binomial family) fitted to the datasets of each quantile, the inflection point of the sigmoids were calculated at ≈ 0.0027 of net divergence for the model of the first quantile ($M_{1st\ quantile}$), $\approx 0.0028\ d_a$ for the second quantile ($M_{2nd\ quantile}$) and ≈ 0.0021 d_a for the third quantile ($M_{3nd\ quantile}$) (Fig. 34).

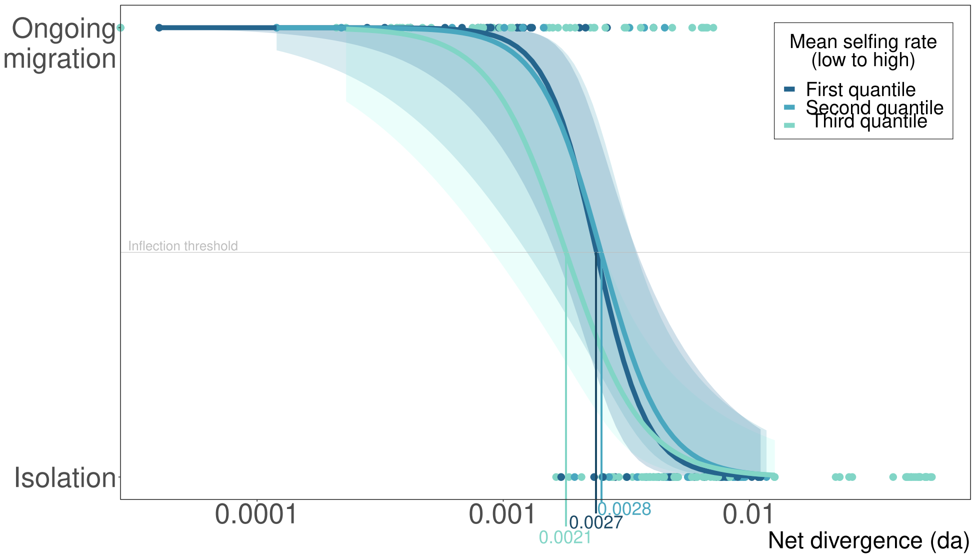


Figure 34: **Ongoing migration in function of the net divergence for the selfing rate quantiles.**

Ongoing migration status as a function of net divergence (or da, Nei and Li, 1979). Each dot is a pair of congenera plant species whose colour represents one of three quantiles of average selfing rate. The ongoing migration status, inferred using *DILS* (Fraisse et al., 2021), is shown on the y-axis, while net divergence is represented on the x-axis. The sigmoid curves represent the probability of current or recent gene exchange for pairs at various levels of net divergence (d_a). This probability is determined using linear regressions (GLM) and is accompanied by a 95% confidence interval. The inflection points, which signify the threshold between a probability of ongoing migration $>$ 0.5 and $<$ 0.5 (represented as a horizontal grey line), are indicated by vertical bars. These inflection points are approximately 0.27% (95% CI: [0.017%-0.515%]) of net divergence for the first quantile, approximately 0.28% (95% CI: [0.005%-0.561%]) of net divergence for the second quantile and about 0.21% (95% CI: [0.011%-0.412%]) for the third quantile.

The log-likelihood ratio test was not significant with a *P*-value of ≈ 0.255 (Tab. 9). No significant *P*-value were found either for the log-likelihood ratio tests with the models fitted on mixed datasets (Tab. S3).

No pairs of species exhibited differences in life forms. Therefore, the life form variable was simplified as a combination of life forms for pairs of species (herb, liana & shrub, tree), rather

Table 9: **Log-likelihood ratio test for logit models fitted to the selfing rate quantiles datasets.**

model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	**df**	***P*-value**
M_0	-37.34021	3.288	-1295.263	0.0025		
$M_{1st\ quantile}$	-9.273642	4.387	-1646.129	0.0027		
$M_{2nd\ quantile}$	-10.90222	3.495	-1234.369	0.0028		
$M_{3nd\ quantile}$	-15.7987	2.432	-1148.878	0.0021		
					2	0.255

$\boldsymbol{\ell}$: log-likelihoods of models M_0, $M_{1st\ quantile}$, $M_{2nd\ quantile}$ and $M_{3nd\ quantile}$.
$\boldsymbol{\beta_0}$: estimated intercept.
$\boldsymbol{\beta_1}$: estimated coefficient.
$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{1st\ quantile}) - \ell(M_{2nd\ quantile}) - \ell(M_{3nd\ quantile})|$ in a χ-squared distribution with two degrees of freedom.

than a single life form for each pair of species (e.g., herb-tree). Most of the 210 pairs of congenera plant species were described as herbs (64%) or trees (30%), with only 11 pairs identified as liana or shrub (Fig. 35). The range of net divergence among the herb species pairs extended to approximately 0.05516 of net divergence, while the range for the tree species pairs extended to around 0.01248 of net divergence. With the exception of a single pair of species falling below 0.26% of net divergence (*Actinidia arguta* and *A. argutaGiraldii*, with a net divergence of 0), the range of divergence for the lianas or shrub pairs of species extended over only about 0.00683 d_a, which is approximately 8 times lower than that of the herb species pairs and about 2 times lower than that of the tree species pairs (Fig. 36).

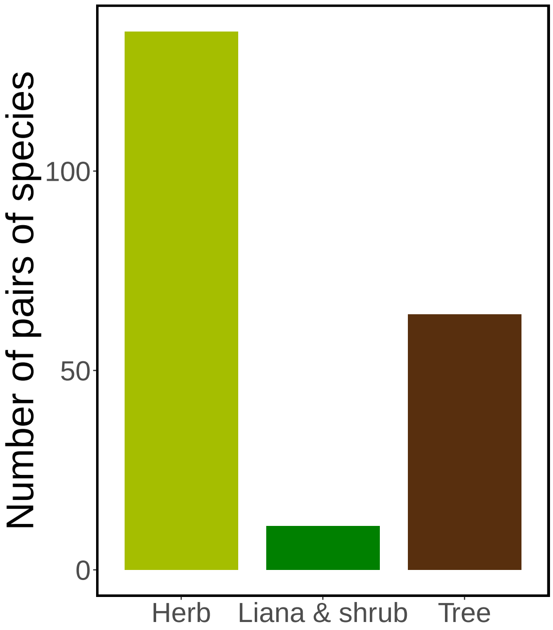


Figure 35: **Life form's distribution.**

Distribution of the pairs of species (y-axis) in function of the life form (x-axis). The pairs of congera species are distributed in 3 life forms, herb with 135 pairs, liana and shrub with 11 pairs and tree with 64 pairs.

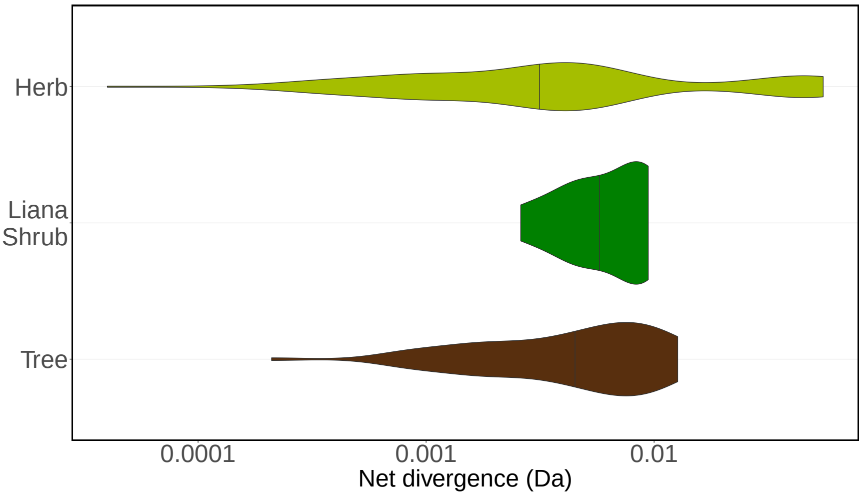


Figure 36: **Life forms' divergence distribution.**

Distribution of the life form in function of the net divergence (d_a). The medians of the quantiles are represented with a vertical line, with $\tilde{d_a}_{\ herb} \approx 0.00374$, $\tilde{d_a}_{\ liana\&shrub} \approx 0.00564$ and $\tilde{d_a}_{\ tree} \approx 0.00495$. Note that due to the logarithmic scale of the x-axis, one pair of the 'liana & shrub' life form is not represented on the chart because its net divergence level is 0.

Considering the underrepresentation of 'liana & shrub' pairs of species (approximately 5%, Fig. 35), and the limited net divergence range of this life form (Fig. 36), which makes it unlikely to have a complete speciation continuum represented, we have decided to exclude these pairs from the remaining analysis. Based on the linear models (GLM, binomial family) fitted to the datasets of each life form (herb or tree), the inflection point of the sigmoids were calculated at ≈ 0.0028 of net divergence for the model fitted to the plant dataset (M_{herb}) and $\approx 0.004\ d_a$ for the model fitted to the tree dataset (M_{tree}) (Fig. 37). These results suggest that, on average, the pairs of herb species reach a complete RI at a lower level of net divergence than the pairs of tree species, with a difference of 0.0012 of net divergence. This difference was tested significant with the log-likelihood ratio test (*P*-value = 0.009, Tab. 10).

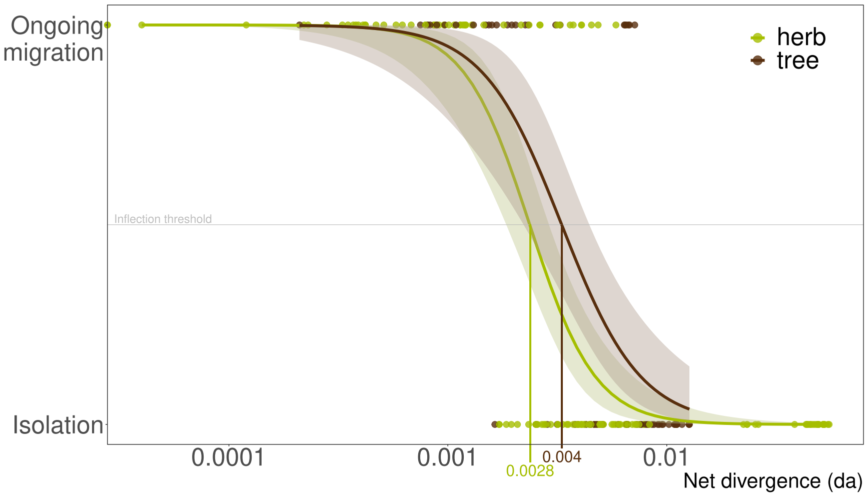


Figure 37: **Ongoing migration in function of the net divergence for the life forms.**

Ongoing migration status as a function of net divergence (or da, Nei and Li, 1979). Each point on the graph represents a pair of congeneric plant species, with herb pairs in green and tree pairs in brown. The y-axis displays the inferred ongoing migration status, determined using *DILS* (Fraïsse et al., 2021), while the x-axis represents net divergence. The sigmoid curves on the graph illustrate the probability of current or recent gene exchange for species pairs at different levels of net divergence (d_a). These probabilities are established using linear regressions (GLM) and are accompanied by a 95% confidence interval. Vertical bars indicate the inflection points, signifying the threshold where the probability of ongoing migration shifts from $>$ 0.5 to $<$ 0.5 (represented as a horizontal grey line). For the herb model, these inflection points are at approximately 0.28% of net divergence (95% CI: [0.137%-0.42%]), and for the tree model, they are approximately 0.4% of net divergence (95% CI: [0.098%-0.7%]).

Table 10: **Log-likelihood ratio test for logit models fitted to the life form datasets.**

model	ℓ	β_0	β_1	$X_{p=0.5}$	df	*P*-value
M_0	-72.81199	2.469	-773.902	0.0032		
M_{herb}	-40.68626	3.096	-1111.101	0.0028		
M_{tree}	-27.38298	2.248	-562.997	0.0040		
					2	0.009

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{herb} and M_{tree}.
$\boldsymbol{\beta_0}$: estimated intercept.
$\boldsymbol{\beta_1}$: estimated coefficient.
$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{herb}) - \ell(M_{tree})|$ in a χ-squared distribution with two degrees of freedom.

Considering the PGLMM analysis, the effect of net divergence on the status of current migration was found to be significant in all of the 100 PGLMM runs (each time with a random subsampling of pairs of taxa), while the rest of the fixed effects were found to be invariably non-significant, although with variances that were strongly affected by the subsampling (Fig. 38). The categorical factor of life form as a whole was tested with a χ^2 and found to be non significant in each of the runs. The average R^2 was around 0.923 with a standard deviation of 0.025, indicating a strong explanatory power for the net divergence factor.

Species random effects (1|spA and 1|spB) were found to be non-significant in all runs (Fig. 39). The phylogenetic random effect is split into two random effects, 1|genus and 1|genus__ (PGLMM syntax). The first one is a simple random effect without phylogenetic covariance, the second is the actual covariance of the phylogeny. This separation is required so that variation among genera is not completely captured by the variance diagonal of the VCV matrix (D. Li et al., 2020). Independent genus covariance (1|genus) was found to be significant in almost all of the runs, but no random effect of phylogenetic covariance (1|genus__) was found to be significant.

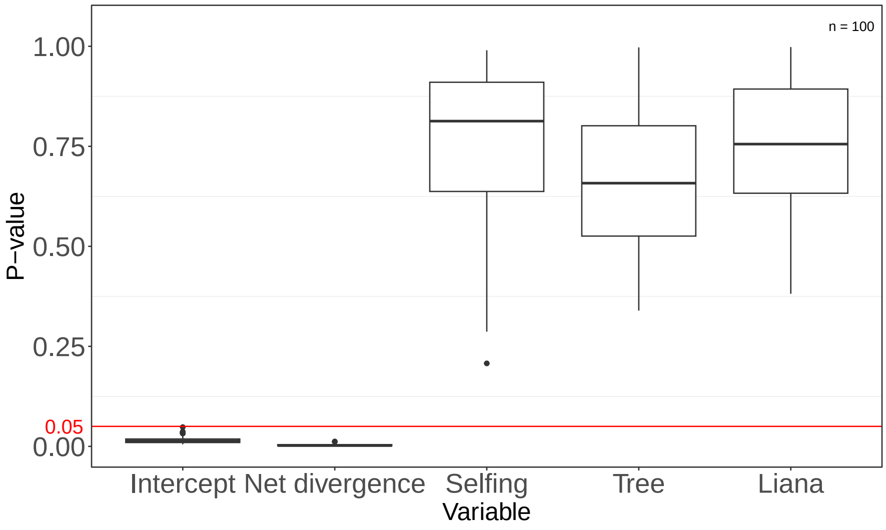


Figure 38: **Fixed effect *P*-values.**

Boxplot of the *P*-values (y-axis) obtained for each of the fixed effects (x-axis). PGLMM analyses were runned for 100 different subsampling of pairs of species.Significant effects were only obtained with the net divergence variable.

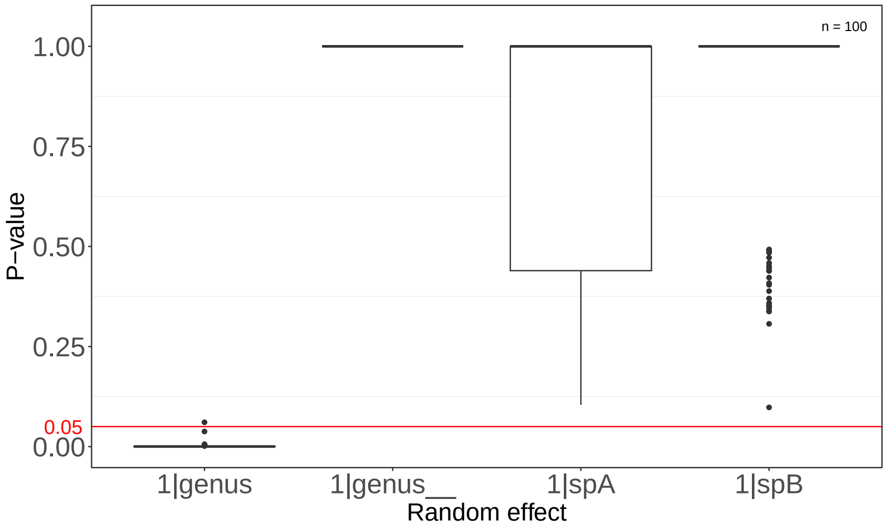


Figure 39: **Random effect *P*-values.**

Boxplot of the *P*-values (y-axis) obtained for each of the random effects (x-axis). PGLMM analyses were runned for 100 different subsampling of pairs of species. Significant effects were only obtained with one of the two phylogenetic random effect (1|genus).

Discussion

The mating system and life form are two traits linked to speciation dynamics. A high rate of selfing is expected to promote the emergence of reproductive barriers (Goldberg et al., 2010; Marie-Orleach et al., 2022; Wright et al., 2013), and the dynamics of speciation in tree and herb species are expected to differ due to their correlations with speciation-related factors such as tree allogamy (Hamrick and Godt, 1996), dispersal capacity (Petit and Hampe, 2006) and effective size (Carvalho et al., 2019). The effect of these traits on the dynamics of speciation in our plant dataset was initially examined using the sigmoid comparison approach proposed in Chapter I of this thesis. Non-significant *P*-values from the log-likelihood ratio tests on the selfing rate quantiles indicated that the three datasets (1st, 2nd, and 3rd quantiles) shared a common speciation dynamics. This result contrasts with the existing scientific literature on the subject (see 2. Selfing rate). Conversely, the log-likelihood ratio test for the herb (M_{herb}) and tree (M_{tree}) models revealed a significant difference in sigmoids, suggesting distinct dynamics of speciation between the two groups. Although the difference in the inflection points of the sigmoids was only 0.0012 d_a (approximately 11 times lower than the one measured between plants and animals in Chapter I (see Fig. 18), it is in line with the expected dynamics of speciation between trees and herbs. A valid criticism of this analysis is the complete absence of consideration for other speciation-related factors in the comparisons. It is unlikely that the three datasets of selfing rate quantiles, as well as the two datasets of herbs and trees, are similar in their speciation-related factors aside from selfing rates or life forms. This limitation could be addressed with an appropriate dataset that allows for control of other factors influencing dynamics. The PGLMM approach enabled the testing of selfing rate and life form effects on speciation dynamics, taking into account the potential influence of other traits through the inclusion of a phylogenetic random effect. In all the PGLMM runs, the effect of the net divergence variable was found significant effect, supporting the classical view about the speciation process, i.e. that RI mostly increases with divergence accumulation (Coyne and Orr, 2004; Roux et al., 2016; Wu, 2001). Conservatively, the effect of the selfing rate and of the life form were neither found significant in any of the PGLMM runs. This aligns with the absence of a significant *P*-value in the log-likelihood ratio tests for the selfing rate but contradicts the significant *P*-value from the life form log-likelihood ratio test. One possible explanation for this inconsistency is that the significant effect of life form detected in the sigmoid comparison may partly be due to correlations between life form and other speciation-related factors. The addition of a random effect accounting for phylogenetic effects, the inclusion of the selfing rate variable, or a combination of both, might have corrected for the significance of the life form factor. While these results might suggest a minimal impact of selfing rates and life forms on the dynamics of speciation, it's important to consider that the inconsistencies with the current scientific literature may also stem from limitations in this study.

Limitations

Non-independency of the pairs

Linear model approaches are commonly employed due to their advantages, such as simplicity and versatility, and they offer valuable modelling as long as their assumptions remain unviolated. One of these assumptions is the independence of the results, which means that each measure of the response variable must be independent of any other measure. This assumption is not validated because the probabilities of ongoing migration are correlated at different levels. Pairs of species are linked by their genus affiliation, and, more broadly, genera are correlated to varying degrees based on their phylogenetic distances. Moreover, pairs of specific genera might share species (e.g., A - B and A - C pairs sharing the species A), further complicating the independence among results. To address these non-independencies, random effects were considered, but only the genus random effect was found to be significant (the intercept of pairs grouped by genus significantly differs from the global intercept). These results do not support an effect of phylogeny or a correlation between pairs that share species. The estimation of correlation between pairs sharing a species might be influenced by the genera random effect, as pairs sharing a species also tend to share a genus. Since these different non-independencies are themselves correlated, it may be challenging, if not impossible, to account for them without affecting the estimation of each other's effects. This is due to the inherent nature of the result variable, which consists of pairs of species. Nevertheless, only the individual estimation of these random effects might be affected. Fixed effects should be correctly estimated, at least concerning these non-independencies.

Quality and representativeness of selfing rate estimates

Where life form information is undoubtedly correctly identified, selfing rates acquisition can be more difficult as their estimates rely on different assumptions (see Bürkli et al., 2017). Given the nature of the data sampling in this study (i.e. open science, data collected from various studies), it was not possible to verify the assumptions used for estimating selfing rates based on identity disequilibria. These assumptions include negligible biparental inbreeding and outbreeding depression, as well as populations being at inbreeding equilibrium (Wang et al., 2012). Moreover, the sample sizes per population were constrained by the available data, which further complicated the estimation of selfing rates. Notably, inconsistencies were observed between the selfing estimates in this study and the known mating systems of certain species, based on scientific literature or expert knowledge. For example, both *Arabis nemorensis* and *A. sagittata* are predominantly selfers (Dittberner et al., 2022), yet none of the *Arabis* populations had selfing rates estimated above 0.21. Furthermore, the distribution of selfing rate estimates did not align with the U-shaped pattern reported in scientific literature (Goodwillie et al., 2005; Igic and Kohn, 2006; Whitehead et al., 2018). This suggests that the dataset may not be representative of the actual selfing rate distribution found in nature. Instead, it appears to be skewed toward predominantly outcrossing species, potentially explaining the absence of significant differences among selfing rate quantile sigmoids. Given these considerations, it is possible that the selfing rate estimates in this study deviate from the true selfing rates of these species or from the typical selfing rates observed in plants. Incorrect selfing rate estimates can lead to erroneous statistical analyses.

Traits non-independence and opposing effect on speciation

In the most common scenario, speciation dynamics are primarily influenced by a 'molecular clock dynamic'. Allopatric lineages accumulate genetic differences at a constant rate, which, in turn, promotes the emergence and accumulation of reproductive barriers (Baack et al., 2015). Additionally, the rate of genetic divergence and the emergence of reproductive barriers are influenced by traits and environmental factors. The understanding of the effects of these factors on speciation remains a subject of ongoing research (Helmstetter et al., 2023). These investigations face several challenges, as noted in a recent article (Anderson et al., 2023). First, the effects of a trait on speciation are not always unidirectional. For example, selfing promotes reproductive isolation through reduced pollen dispersal capacity (Barrett and Harder, 1996), but it may reduce the effect of reproductive isolation linked to local adaptation by decreasing effective population size and recombination rate (Gavrilets, 2004). Second, traits influencing speciation are often interrelated. For instance, there is an association between selfing and polyploidy (Stebbins, 1950). In this study, we examined only selfing rate and life form as explanatory variables for the ongoing migration status of species pairs. These two traits are known to be correlated (Mitchell et al., 2019), which challenges the assumption of their independence. Moreover, they could independently complicate the statistical analysis of their effects, as discussed earlier. Anderson et al. (2023) has identified seven different pathways through which selfing can influence reproductive isolation and genetic divergence, either positively or negatively. Furthermore, it emphasises the correlation between life form and traits directly related to speciation. For example, the association between vertebrate dispersal and speciation is found in woody plants but absent in herbaceous plants (de Queiroz, 2002; Tiffney and Mazer, 1995). Given the conflicting effects of individual traits on speciation, the intercorrelation among traits, and the lack of information on other traits linked to speciation, the results of this study should be interpreted with caution.

Conclusion

Contrary to expectations based on current scientific knowledge, this study found no significant effects of selfing rate or life form. These results should be interpreted with caution due to potential factors that may affect the estimation of the true impact of these traits on speciation. While the use of demographic inferences with linear modeling offers advantages and has the potential to elucidate the effects of selfing, life form, or other traits on speciation, future analyses would benefit from more phylogenetically diverse datasets, improved selfing estimates (increased sample sizes per lineage), and more comprehensive trait information.

Discussion

Speciation is a knotty process, shaped by contingency and a rich complex of factors. Our comprehension of this phenomena has progressed a lot since the origin of Species (Darwin, 1859), but much remains to be understood (Butlin et al., 2011). For instance, numerous factors have been linked with speciation, but their relative importance in the process remains to be assessed (Helmstetter et al., 2023). This can be investigated with various methods, notably by comparing the RI between pairs of taxa (Stankowski and Ravinet, 2021). By observing the difference of RI between taxa along a continuum of divergence, comparative analysis can provide elements of solution to questions such as by what type of reproductive barriers does the process of evolution usually begin ? What are the factors that mainly influence the emergence and strength of reproductive barriers ? What is the importance of intrinsic postzygotic barriers in the coexistence of sympatric species ? (Butlin et al., 2011; Coughlan and Matute, 2020; Stankowski and Ravinet, 2021). In this perspective, this thesis analysed and compared the dynamics of speciation of animals and plants, two clades for which the dynamics of speciation were expected to differ. Historically, the dynamics of speciation of animals was assumed to be faster (i.e. RI building faster with time), an assumption based on arguments such as the existence of mating behaviour or on the greater complexity of animal organisms and thus more easily affected by hybridization (but see introduction of the chapter I). This assumption was supported by recent studies based on morphological (Mallet, 2005) or molecular criteria (*D* statistics, Dagilis et al., 2022), although the opposing tendency is also supported by molecular analysis as higher F_{ST} among conspecific populations were observed for plants relative to animals (Frankham et al., 2014; Morjan and Rieseberg, 2004). Here, the results obtained in this thesis challenged the historical assumption as a complete RI between genetic clusters is observed on average at a lower level of net divergence for plants than for animals. In addition, plant speciation events appear to involve allopatry more frequently than for animals. Indeed, among species pairs with ongoing migration, demographic patterns of secondary contact were inferred for a majority of pairs of plant species but only for a quarter of the animal pairs. Together, these results suggest that one or many factors that differ between plants and animals are involved in the dynamics of speciation of those taxa, and that the overall effect of those factors promote a faster complete RI for the plants. This tendency is in accordance with the higher F_{ST} among conspecific populations observed in the plant kingdom (Morjan and Rieseberg, 2004). Besides their intrinsic significance for science knowledge, those results highlight the usefulness of the novel comparative approach proposed to investigate dynamics of speciation. Thanks to the flexibility of the approach on the type of input data, numerous multilocus NGS data freely accessible online could be used to build a dataset sufficiently consequent to compare two clades without the need of additional sequencing effort. Once the data set is formed, the approach consists in the production of demographic inferences with an ABC method designed to be swift and user-friendly (Fraïsse et al., 2021). Comparison of the dynamics of speciation inferred with the ABC method on each clade can be performed with a simple log-likelihood ratio test as illustrated in the Chapter I. With those features, a large variety of comparisons can be considered beyond the one of this thesis. For example, the plants or animals dynamic of speciation could be compared to fungi's, another kingdom where the dynamics of speciation have been investigated (Giraud and Gourbière, 2012). Such comparison could provide new combinations of 'speciation related factors' to further depict the dynamics of speciation. For example, fungi's richness of reproductive strategies is closer to plants (Nieuwenhuis and James, 2016), while their lack of chloroplast make them more similar to animals, both these traits being putatively linked to speciation dynamics (e.g. B. Charlesworth, 1992; Greiner et al., 2011; Marie-Orleach et al., 2022; Pickup et al., 2019; Postel and Touzet, 2020. Furthermore, comparisons of monophyletic taxa could highlight the indirect effect of key life-history traits, for example the ability to fly developed in the Aves lineage (birds), an innovation that improves their dispersal capacity which in turn affects their dynamics of speciation (Claramunt et al., 2011)(. Another idea could be to conduct comparisons between groups of species that differ in extrinsic characteristics (i.e. not of genetic origin). For example, the dynamics of speciation of continental/oceanic populations could be compared with islander/lake populations to search for effects of reduced effective population size, differences in genetic variation or co-occurrence as suggested by Marques et al. (2019). Another example, of particular interest in regards to the current biodiversity crisis (Newbold et al., 2015; Sponsel, 2013), is the comparison of dynamics of speciation between populations in habitats with various levels of human disturbance. It is probable that current speciation events are affected by human activity through change in distribution, dispersal capacity and interbreeding (i.e. change between allopatric and sympatric condition) (Crispo et al., 2011; McFarlane and Pemberton, 2019). However, the proposed approach relies on inference of 'recent' gene flow between genetic clusters and it is therefore not suitable to study effects at the scale of a few generations. The comparison should therefore involve habitats highly preserved from human influence (e.g. Kerguelen Islands) to ensure a difference of 'treatment' on sufficiently large time scales. Another strategy could be to split the taxa according to the IUCN Red List endangered status (2022) to compare speciation events involving species highly impacted by human activity.

In the second part of the thesis, the dynamics of speciation of the plants was further investigated by testing the explanatory power of two traits, the selfing rate and the life form, selected on the basis of their availability and their link with speciation. By impeding gene flow through different pathways, high levels of selfing and tree-like forms are expected to promote speciation (Helmstetter et al., 2023; Pickup et al., 2019, but see chapter II). The analysis was conducted with a GLM, with the approximated RI of the pairs of species as response variable and the divergence, the selfing rate and the life form as explanatory variables. Of the two approaches that were explored to account for the intrinsic non-independence of the pairs of species, only the approach based on multiple re-sampling was achieved as the VCV matrix approach was more challenging to fully implement during the time of this thesis. Although the re-sampling approach limits the effect of non-independency, it also limits the power of the analysis as only a subsample of the data are used. Therefore, future linear regression using the RI of pairs of species as response variable would benefit from further investigation on the mathematical plausibility of the VCV matrix approach. No effect was found significant for any tested variables, except for the net divergence, a surprising result that contrasts with the scientific literature on the subject. It may be tempting to attribute the absence of significant effect to limitations peculiar to this analysis, such as the unreliable estimates of selfing rate, and the verification of this uncertainty calls for further analysis with better estimates and bigger dataset to enhance the power of the models.

With the growing availability of sequencing data, the approach presented in this thesis promises numerous interesting taxa comparisons. However, relying on open science is not devoid of concerns. First, researchers do not have control on the data acquisition. Protocol standards vary from one study to another and detailed descriptions are not necessarily available, compelling the researcher to trust the available data. Secondly, the different sequencing methods available are not equivalently suitable for genomic studies, and researchers might be forced to deal with sequencing methods introducing bias to complete their dataset. For example, in the analysis of this thesis, sequences issued from RAD sequencing were considered neutral as the majority of the plant genomes are non-coding (Heslop-Harrison and Schmidt, 2012), thus the proportion of sequences that belong to coding loci were considered as negligible. This might introduce a bias in the estimation of net divergence as *DILS* will consider as neutral some positions that are under purifying selection (i.e. that mainly drive the evolution of coding sequences). The net divergence is therefore most likely underestimated for the pairs of species with RAD sequencing. A last issue of open science is the phylogenetic non-independence of a part of the pairs of species. Most of the sequencing data originates from population genetic studies that usually sequence populations or species closely related. In consequence, species of the comparison dataset will be clustered in genera or families. Fortunately, it is possible that the ongoing development of open science will lessen those issues by encouraging the normalisation of exhaustive metadata and by increasing the availability of whole genome sequencing methods[6]. Another limitation of the proposed approach is the choice of samples to form the genetic clusters. To illustrate this, we can imagine a simple two demes model where the diverging populations hybridise at a contact zone (Fig. 40). Although samples could be collected at different locations of the range of the populations, they would not be equally adequate to study the RI of the pair. Samples collected near the contact zone (red cross, Fig. 40) would have a greater chance to be early generation hybrids, hybrids that are usually avoided for RI study as

[6] Additionally, the release of annotated reference genomes could also lessen the RAD bias by it would be possible to remove from the analysis the sequences from coding loci.

they may not represent the actual RI (e.g. sterile hybrids). On the other hand, samples that are collected a long way from the contact zone (red dot, Fig. 40) may carry a weak signal of the actual introgression. The diffusion of alleles in populations can be a slow process (Barton, 1979) and the signal of introgression after a secondary contact may not have reached yet the location of sampling. Accordingly, the sampling should be done at a reasonable distance from contact zones, sufficiently away to capture foreign allele frequencies after the effect of reproductive barriers, but close enough to detect the maximum of the introgression signal. This may be difficult to settle, and a simple common protocol for sampling could be useful to increase the comparability of the different studies.

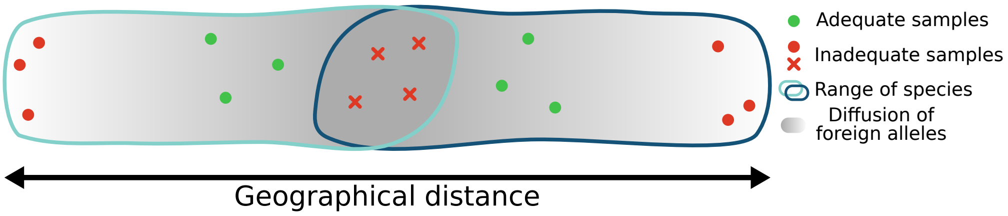


Figure 40: **geographical distance.**

Cartoon of the adequacy of sampling as a function of their geographical distance with a contact zone in an ideal two-demes model. In principle, in the context of the analysis of this thesis, the samples should not be collected too close to the contact zone (red crosses) to avoid early-generation hybrids, and not too far (red points) to efficiently capture the signal of introgression.

A last example of limitation that could be improved can be found in the binary approximation of the RI. In our analyses, we considered that either the species of a pair remained genetically isolated since a certain time, or they experienced introgression in the recent past. In the latter case, any level of RI different from a full reproductive isolation is considered as null. This obviously deviates from reality as RI progressively builds up during speciation events (Wu, 2001). However, estimating a continuous level of RI with *DILS* can be tricky. It could be approximated by measuring the proportion of the genome linked to reproductive barriers, but this requires whole genome sequencing if we cannot consider that RAD and RNA sequencing datasets represent a uniform subsampling of the whole genome. This lessens the advantage of data compatibility of *DILS* (e.g. only 4 of the 29 accessions are WGS in the dataset of this thesis). Furthermore, a more precise estimation of RI would require an estimation of the migration rate of loci linked with reproductive barriers, a task where *DILS* performed poorly (figure 5e, Fraïsse et al., 2021).

Even with good estimates of migration rates along the genome, the migration rate free of the effect of any reproductive barrier (m) would not be measurable with this method because strong reproductive barrier effects impact the gene flow at the genome scale (i.e. even with maximal recombination rate between loci), preventing from true absolute measure of the RI (Westram et al., 2022). Because of these limits, this approach is therefore restricted to correlative investigations of the relationship between dynamics of speciation and speciation factors. Hence, further improvement would require different approaches (Ravinet et al., 2017; Stankowski and Ravinet, 2021).

But finally ... what's behind our sigmoids?

Our approach to the study of speciation dynamics involves investigating the reduction in gene flow along a continuum of divergence, ranging from minimal barriers to gene flow on the left to progressively more impediments on the right. This approach yields a sigmoidal relationship that serves as a highly valuable tool for conducting comparative analyses among various taxa. Its primary purpose is to assess variations in the rates at which barriers to gene flow accumulate. During the course of my doctoral thesis, I conducted comparative analysis comparing plants and animals, plants of different levels of selfing rate, as well as herbs and trees. In the future, additional analyses will encompass plant *versus* animal *versus* fungi comparisons or delve into the study of the impact of life-history traits such as haplodiploid cycles *versus* diplobiontic cycles, free-living *versus* parasitic lifestyles, and more.

Intuitively, we can readily grasp the significance of this relationship: the greater the number of divergent mutations between two lineages, the more pronounced the expression of Dobzhansky Muller incompatibilities within hybrid genomes. Furthermore, as the number of divergent mutations increases, the pairing of homologous chromosomes during meiosis decreases. Additionally, with divergence, structural differences accumulate as well as disparities in gene expression patterns and more. Therefore, intuitively, we understand that divergence exerts a detrimental effect on gene flow between species. However, the exact nature of this sigmoidal curve remains misunderstood, because it has not yet been studied theoretically. What factors shape its profile?

In other words, what can we infer about the biology of speciation from the inflection point and slope of the sigmoid?

In this section, I present some non-comprehensive avenues for interpreting the relationships between divergence and reproductive isolation. My analysis is purely conceptual, relying exclusively on functionalist intuitions concerning the resilience of biological pathways in the face of incompatibilities and the significance of individual pathways for the fitness of organisms.

Moving from left to right along the curve, the initial portion of the sigmoid appears as a plateau in which divergent mutations do not seem to create effective barriers against gene flow. From a functionalist perspective, the duration of this latent phase before the effects of barriers become apparent suggests that organism genomes possess qualities that mitigate the deleterious impacts of these barriers. Among these qualities, I emphasize the interconnection of biological and metabolic pathways, which, in turn, give rise to functional redundancies (see Figure 41.A). Variations in the levels of redundancy in nature could lead to shifts in the grey zone of speciation. For instance, a highly reticulated metabolic network characterized by substantial redundancy might require a higher number of barriers to completely disrupt its functionality (as depicted by the orange curve in Figure 41.A), thus adversely affecting the fitness of hybrids. Assuming that the number of barriers is linearly related to species divergence, greater functional redundancy, on average, in an organism could extend this initial plateau to higher levels of divergence.

Expanding this line of reasoning to the individual contributions of networks to the fitness of an (hybrid) individual, it may be possible, after numerous hypotheses, to find a biological explanation for the slope of the sigmoid (as shown in Figure 41.B). I must admit that I am presenting an empirical intuition rather than precise quantitative expectations. My intuition suggests that, given similar levels of complexity in metabolic networks across compared groups, an organism in which each network is indispensable for survival (depicted in orange in Figure 41.B) would likely experience hybrid depression as soon as any of its networks become saturated with incompatibilities. Such a network architecture would result in a sudden reduction in introgression once a metabolic pathway is disrupted. Conversely (in red), other organisms might tolerate the loss of individual networks more easily, prolonging the speciation process to higher levels of divergence, up to the point where the entire metabolic pathway is severely compromised.

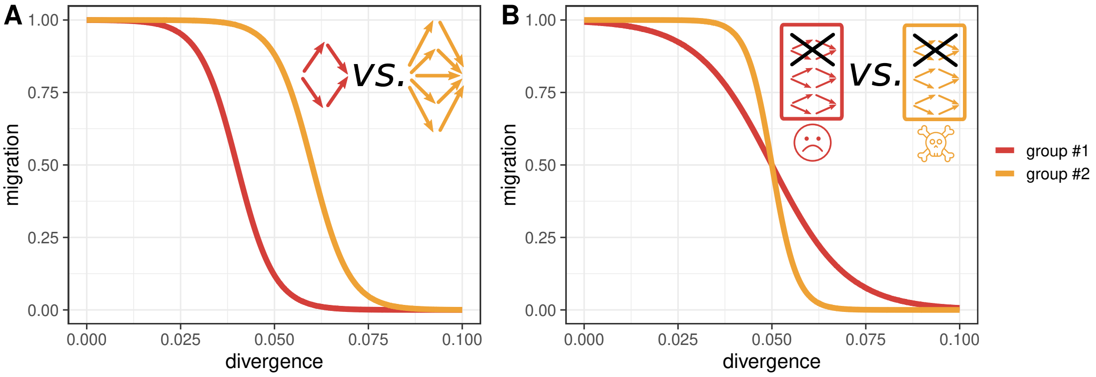


Figure 41: **What biological factors shape the ’sigmoid’?**

Verbal proposal of an effect of “metabolic networks’ on **A)** the inflection point of the sigmoid, and **B)** the slope of the sigmoid. The red and orange colours indicate sigmoids fitted to data from two groups distinguished by a factor. In panel **A**, red symbols represent organisms with low levels of functional redundancy (represented by a network of red arrows, requiring at least two barriers to interrupt the pathway in hybrids), compared with orange organisms with a higher level of functional redundancy (requiring at least five barriers to interrupt the pathway). In panel **B**, red symbols represent organisms with networks whose individual deletion has little impact on hybrid fitness. In Orange, organisms in which the loss of a single network due to a barrier would have a lethal effect in hybrids.

Therefore, it seems plausible to consider exploring the role of metabolic network architecture in speciation dynamics as an initial step. Specifically, employing theoretical approaches to comprehend the impact of network properties (the number of networks, network reticulation, redundancy, and the relative importance of a given network in completing the life cycle) on the shape of the sigmoid. This exploration will provide insights into why the comparison between plants and animals appears to be more of a shift in the grey zone (Figure 41.A) rather than a difference in slope (Figure 41.B). However, it is essential to acknowledge that alternative comparisons, rather than plant versus animal, may in the future exhibit variations in the slope of the sigmoid. Before embarking on multiple comparative analyses yielding multiple sigmoidal curves, it is imperative to swiftly undertake theoretical investigations into the diverse factors influencing the shape of these curves to gain a deeper comprehension of the central element employed in our comparisons.

Take home message

By comparing the dynamics of speciation of two kingdoms, this thesis reveals a faster completion of the reproductive isolation of plants in contrast with animals, challenging an opposite historical assumption. This result illustrates the interest of the novel comparative framework developed to conduct easy, rapid and flexible future comparisons of dynamics of speciation, and test of putative explanatory factors.

References

Abbott, R. J. (2017). “Plant Speciation across Environmental Gradients and the Occurrence and Nature of Hybrid Zones”. In: *Journal of Systematics and Evolution* 55.4, pp. 238–258. ISSN: 1759-6831. DOI: 10.1111/jse.12267.

Anderson, B. et al. (2023). “Opposing Effects of Plant Traits on Diversification”. In: *iScience* 26.4, p. 106362. ISSN: 25890042. DOI: 10.1016/j.isci.2023.106362.

Anderson, E. and G. L. Stebbins (1954). “Hybridization as an Evolutionary Stimulus”. In: *Evolution* 8.4, p. 378. ISSN: 00143820. DOI: 10.2307/2405784. JSTOR: 2405784.

Anderson, E. (1948). “Hybridization of the Habitat”. In: *Evolution* 2.1, pp. 1–9. ISSN: 0014-3820. DOI: 10.2307/2405610. JSTOR: 2405610.

Anderson, E. and L. Hubricht (1938). “Hybridization in Tradescantia. III. The Evidence for Introgressive Hybridization”. In: *American Journal of Botany* 25.6, pp. 396–402. ISSN: 0002-9122. DOI: 10.2307/2436413. JSTOR: 2436413.

Armbruster, W. S. and N. Muchhala (2009). “Associations between Floral Specialization and Species Diversity: Cause, Effect, or Correlation?” In: *Evolutionary Ecology* 23.1, pp. 159–179. ISSN: 1573-8477. DOI: 10.1007/s10682-008-9259-z.

Arunkumar, R. et al. (2013). “Pollen-Specific, but Not Sperm-Specific, Genes Show Stronger Purifying Selection and Higher Rates of Positive Selection Than Sporophytic Genes in Capsella Grandiflora”. In: *Molecular Biology and Evolution* 30.11, pp. 2475–2486. ISSN: 0737-4038. DOI: 10.1093/molbev/mst149.

Baack, E. et al. (2015). “The Origins of Reproductive Isolation in Plants”. In: *New Phytologist* 207.4, pp. 968–984. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.13424.

Baek, Y. S. et al. (2015). “Testing the SI × SC Rule: Pollen–Pistil Interactions in Interspecific Crosses between Members of the Tomato Clade (Solanum Section Lycopersicon, Solanaceae)”. In: *American Journal of Botany* 102.2, pp. 302–311. ISSN: 1537-2197. DOI: 10.3732/ajb.1400484.

Baker, H. G. (1955). “Self-Compatibility and Establishment after’long-Distance’dispersal”. In: *Evolution* 9.3, pp. 347–349. ISSN: 0014-3820.

Barnard-Kubow, K. B., N. So, and L. F. Galloway (2016). “Cytonuclear Incompatibility Contributes to the Early Stages of Speciation”. In: *Evolution* 70.12, pp. 2752–2766. ISSN: 0014-3820. DOI: 10.1111/evo.13075.

Barr, C. M. and L. Fishman (2011). “Cytoplasmic Male Sterility in Mimulus Hybrids Has Pleiotropic Effects on Corolla and Pistil Traits”. In: *Heredity* 106.5 (5), pp. 886–893. ISSN: 1365-2540. DOI: 10.1038/hdy.2010.133.

Barrett, S. C. and L. D. Harder (1996). “Ecology and Evolution of Plant Mating”. In: *Trends in Ecology & Evolution* 11.2, pp. 73–79. ISSN: 01695347. DOI: 10.1016/0169-5347(96)81046-9.

Barrett, S. C. H., L. D. Harder, and A. C. Worley (1996). “The Comparative Biology of Pollination and Mating in Flowering Plants”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 351.1345, pp. 1271–1280. ISSN: 0962-8436.

Barton, N. H. (1979). “Gene Flow Past a Cline”. In: *Heredity* 43.3, pp. 333–339. ISSN: 0018-067X, 1365-2540. DOI: 10.1038/hdy.1979.86.

— (1980). “The Hybrid Sink Effect”. In: *Heredity* 44.2 (2), pp. 277–278. ISSN: 1365-2540. DOI: 10.1038/hdy.1980.23.

Barton, N. H. and G. M. Hewitt (1985). “Analysis of Hybrid Zones”. In: *Annual Review of Ecology and Systematics* 16, pp. 113–148. ISSN: 0066-4162. JSTOR: 2097045.

Barton, N. H., K. S. Gale, and R. G. Harrison (1993). “Genetic Analysis of Hybrid Zones”. In: *Hybrid zones and the evolutionary process*, pp. 13–45.

Barton, N. H. and G. M. Hewitt (1989). “Adaptation, Speciation and Hybrid Zones”. In: *Nature* 341.6242, pp. 497–503. ISSN: 0028-0836.

Barton, N. and B. O. Bengtsson (1986). “The Barrier to Genetic Exchange between Hybridising Populations”. In: *Heredity* 57.3 (3), pp. 357–376. ISSN: 1365-2540. DOI: 10.1038/hdy.1986.135.

Batista, R. A. and C. Köhler (2020). “Genomic Imprinting in Plants—Revisiting Existing Models”. In: *Genes & Development* 34.1-2, pp. 24–36. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.332924.119.

Bawa, K. S. (1980). “Evolution of Dioecy in Flowering Plants”. In: *Annual Review of Ecology and Systematics* 11.1, pp. 15–39. ISSN: 0066-4162. DOI: 10.1146/annurev.es.11.110180.000311.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). “Approximate Bayesian Computation in Population Genetics”. In: *Genetics* 162.4, pp. 2025–2035. ISSN: 1943-2631. DOI: 10.1093/genetics/162.4.2025.

Benoit, C. (1924). “Note sur une méthode de résolution des équations normales provenant de l’application de la méthode des moindres carrés à un système d’équations linéaires en nombre inférieur à celui des inconnues (Procédé du Commandant Cholesky)”. In: *Bulletin géodésique* 2.1, pp. 67–77.

Bierne, N. et al. (2011). “The Coupling Hypothesis: Why Genome Scans May Fail to Map Local Adaptation Genes”. In: *Molecular Ecology* 20.10, pp. 2044–2072. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2011.05080.x.

Bock, D. G. et al. (2023). “Genomics of Plant Speciation”. In: *Plant Communications*, p. 100599. ISSN: 2590-3462. DOI: 10.1016/j.xplc.2023.100599.

Bolnick, D. I. et al. (2023). “A Multivariate View of the Speciation Continuum”. In: *Evolution* 77.1, pp. 318–328. ISSN: 0014-3820. DOI: 10.1093/evolut/qpac004.

Brandrud, M. K. et al. (2020). “Phylogenomic Relationships of Diploids and the Origins of Allotetraploids in Dactylorhiza (Orchidaceae)”. In: *Systematic Biology* 69.1, pp. 91–109. ISSN: 1063-5157. DOI: 10.1093/sysbio/syz035.

Brandvain, Y. and D. Haig (2005). “Divergent Mating Systems and Parental Conflict as a Barrier to Hybridization in Flowering Plants”. In: *The American Naturalist* 166.3, pp. 330–338. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/432036.

Broz, A. K. and P. A. Bedinger (2021). “Pollen-Pistil Interactions as Reproductive Barriers”. In: *Annual Review of Plant Biology* 72.1, pp. 615–639. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-080620-102159.

Broz, A. K. et al. (2017). “Mating System Transitions in Solanum Habrochaites Impact Interactions between Populations and Species”. In: *New Phytologist* 213.1, pp. 440–454. ISSN: 1469-8137. DOI: 10.1111/nph.14130.

Budar, F., P. Touzet, and R. D. Paepe (2003). “The Nucleo-Mitochondrial Conflict in Cytoplasmic Male Sterilities Revisited”. In.

Burgarella, C. and S. Glémin (2017). “Population Genetics and Genome Evolution of Selfing Species”. In: *eLS*. 1st ed. Wiley, pp. 1–8. ISBN: 978-0-470-01617-6 978-0-470-01590-2. DOI: 10.1002/9780470015902.a0026804.

Bürkli, A. et al. (2017). “Comparing Direct and Indirect Selfing Rate Estimates: When Are Population-Structure Estimates Reliable?” In: *Heredity* 118.6 (6), pp. 525–533. ISSN: 1365-2540. DOI: 10.1038/hdy.2017.1.

Butlin, R. et al. (2011). “What Do We Need to Know about Speciation?” In: *Trends in ecology & evolution* 27.1, pp. 27–39. ISSN: 0169-5347.

Cabot, E. L. et al. (1994). “Genetics of Reproductive Isolation in the Drosophila Simulans Clade: Complex Epistasis Underlying Hybrid Male Sterility.” In: *Genetics* 137.1, pp. 175–189. ISSN: 1943-2631. DOI: 10.1093/genetics/137.1.175.

Caldwell, K. A. and M. A. Handel (1991). “Protamine Transcript Sharing among Postmeiotic Spermatids.” In: *Proceedings of the National Academy of Sciences* 88.6, pp. 2407–2411. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.88.6.2407.

Carpenter, E. J. et al. (2019). “Access to RNA-sequencing Data from 1,173 Plant Species: The 1000 Plant Transcriptomes Initiative (1KP)”. In: *GigaScience* 8.10, giz126. ISSN: 2047-217X. DOI: 10.1093/gigascience/giz126.

Carvalho, Y. G. S. et al. (2019). “Recent Trends in Research on the Genetic Diversity of Plants: Implications for Conservation”. In: *Diversity* 11.4 (4), p. 62. ISSN: 1424-2818. DOI: 10.3390/d11040062.

Case, A. L. et al. (2016). “Selfish Evolution of Cytonuclear Hybrid Incompatibility in Mimulus”. In: *Proceedings of the Royal Society B: Biological Sciences* 283.1838, p. 20161493. DOI: 10.1098/rspb.2016.1493.

Castillo, D. M., A. K. Gibson, and L. C. Moyle (2016). “Assortative Mating and Self-Fertilization Differ in Their Contributions to Reinforcement, Cascade Speciation, and Diversification”. In: *Current Zoology* 62.2, pp. 169–181. ISSN: 1674-5507. DOI: 10.1093/cz/zow004.

Castric, V. et al. (2008). “Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection”. In: *PLOS Genetics* 4.8, e1000168. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000168.

Catchen, J. et al. (2011). “Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences”. In: *G3 Genes—Genomes—Genetics* 1.3, pp. 171–182. ISSN: 2160-1836. DOI: 10.1534/g3.111.000240.

Catchen, J. et al. (2013). “Stacks: An Analysis Tool Set for Population Genomics”. In: *Molecular Ecology* 22.11, pp. 3124–3140. ISSN: 1365-294X. DOI: 10.1111/mec.12354.

Charlesworth, B. (1992). “Evolutionary Rates in Partially Self-Fertilizing Species”. In: *The American Naturalist* 140.1, pp. 126–148. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/285406.

Charlesworth, D. (1993). “Why Are Unisexual Flowers Associated with Wind Pollination and Unspecialized Pollinators?” In: *The American Naturalist* 141.3, pp. 481–490. ISSN: 0003-0147. DOI: 10.1086/285485.

Chazdon, R. L. et al. (2003). “Community and Phylogenetic Structure of Reproductive Traits of Woody Species in Wet Tropical Forests”. In: *Ecological Monographs* 73.3, pp. 331–348. ISSN: 1557-7015. DOI: 10.1890/02-4037.

Claramunt, S. et al. (2011). “High Dispersal Ability Inhibits Speciation in a Continental Radiation of Passerine Birds”. In: *Proceedings of the Royal Society B: Biological Sciences* 279.1733, pp. 1567–1574. DOI: 10.1098/rspb.2011.1922.

Coughlan, J. M. and D. R. Matute (2020). “The Importance of Intrinsic Postzygotic Barriers throughout the Speciation Process”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1806, p. 20190533. DOI: 10.1098/rstb.2019.0533.

Coyne, J. A. and H. A. Orr (1989). “Patterns of speciation in drosophila”. In: *Evolution* 43.2, pp. 362–381. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/j.1558-5646.1989.tb04233.x.
— (1997). “”Patterns of Speciation in Drosophila” Revisited”. In: *Evolution* 51.1, pp. 295–303. ISSN: 0014-3820. DOI: 10.2307/2410984. JSTOR: 2410984.
— (1998). “The Evolutionary Genetics of Speciation”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 353.1366, pp. 287–305. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.1998.0210.
— (2004). *Speciation*. Sunderland, Mass: Sinauer Associates. 545 pp. ISBN: 978-0-87893-091-3 978-0-87893-089-0.

Crepet, W. L. (1984). “Advanced (Constant) Insect Pollination Mechanisms: Pattern of Evolution and Implications Vis-à-Vis Angiosperm Diversity”. In: *Annals of the Missouri Botanical Garden*, pp. 607–630. ISSN: 0026-6493.

Crispo, E. et al. (2011). “Broken Barriers: Human-induced Changes to Gene Flow and Introgression in Animals: An Examination of the Ways in Which Humans Increase Genetic Exchange among Populations and Species and the Consequences for Biodiversity”. In: *BioEssays* 33.7, pp. 508–518. ISSN: 02659247. DOI: 10.1002/bies.201000154.

Cruickshank, T. E. and M. W. 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. ISSN: 1365-294X. DOI: 10.1111/mec.12796.

Csilléry, K. et al. (2010). “Approximate Bayesian Computation (ABC) in Practice”. In: *Trends in Ecology & Evolution* 25.7, pp. 410–418. ISSN: 01695347. DOI: 10.1016/j.tree.2010.04.001.

Cutter, A. D. (2019). “Reproductive Transitions in Plants and Animals: Selfing Syndrome, Sexual Selection and Speciation”. In: *New Phytologist* 224.3, pp. 1080–1094. ISSN: 1469-8137. DOI: 10.1111/nph.16075.

Dagilis, A. J. et al. (2022). “A Need for Standardized Reporting of Introgression: Insights from Studies across Eukaryotes”. In: *Evolution Letters* 6.5, pp. 344–357. ISSN: 2056-3744. DOI: 10.1002/evl3.294.

Danecek, P. et al. (2011). “The Variant Call Format and VCFtools”. In: *Bioinformatics* 27.15, pp. 2156–2158. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330.

Danecek, P. et al. (2021). “Twelve Years of SAMtools and BCFtools”. In: *GigaScience* 10.2, giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. pmid: 33590861.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life.* J.Murray. London.

— (1862). “On the Various Contrivances by Which British and Foreign Orchids Are Fertilized”. In: *Murray, London* 365.

David, P. et al. (2007). “Reliable Selfing Rate Estimates from Imperfect Population Genetic Data”. In: *Molecular Ecology* 16.12, pp. 2474–2487. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2007.03330.x.

Davis, A. W., E. G. Noonburg, and C. I. Wu (1994). “Evidence for Complex Genic Interactions between Conspecific Chromosomes Underlying Hybrid Female Sterility in the Drosophila Simulans Clade.” In: *Genetics* 137.1, pp. 191–199. ISSN: 1943-2631. DOI: 10.1093/genetics/137.1.191.

De Nettancourt, D. (2001). *Incompatibility and Incongruity in Wild and Cultivated Plants*. Vol. 3. Springer Science & Business Media. ISBN: 3-540-65217-5.

De Queiroz, A. (2002). “Contingent Predictability in Evolution: Key Traits and Diversification”. In: *Systematic Biology* 51.6, pp. 917–929. ISSN: 1063-5157. DOI: 10.1080/10635150290102627.

Devaux, C. and R. Lande (2009). “Displacement of Flowering Phenologies among Plant Species by Competition for Generalist Pollinators”. In: *Journal of Evolutionary Biology* 22.7, pp. 1460–1470. ISSN: 1420-9101. DOI: 10.1111/j.1420-9101.2009.01762.x.

Ding, X. et al. (2019). “Congruent Species Delimitation of Two Controversial Gold-Thread Nanmu Tree Species Based on Morphological and Restriction Site-Associated DNA Sequencing Data”. In: *Journal of Systematics and Evolution* 57.3, pp. 234–246. ISSN: 1759-6831. DOI: 10.1111/jse.12433.

Dittberner, H., A. Tellier, and J. de Meaux (2022). “Approximate Bayesian Computation Untangles Signatures of Contemporary and Historical Hybridization between Two Endangered Species”. In: *Molecular Biology and Evolution* 39.2, msac015. ISSN: 1537-1719. DOI: 10.1093/molbev/msac015.

Dobzhansky, T. H. (1936). “Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids”. In: *Genetics* 21.2, p. 113.

Dobzhansky, T. (1951). *Genetics and the Origin of Species*. 3rd. Columbia Univ. Press, New York.

— (1958). "Species after Darwin". In: *A century of Darwin. Heinemann, London*, pp. 19–55.

Dowling, T. E. and C. L. Secor (1997). "The Role of Hybridization and Introgression in the Diversification of Animals". In: *Annual Review of Ecology and Systematics* 28.1, pp. 593–619. ISSN: 0066-4162. DOI: 10.1146/annurev.ecolsys.28.1.593.

Doyle, J. J. et al. (2008). "Evolutionary Genetics of Genome Merger and Doubling in Plants". In: *Annual Review of Genetics* 42.1, pp. 443–461. ISSN: 0066-4197, 1545-2948. DOI: 10.1146/annurev.genet.42.110807.091524.

Dufaÿ, M. et al. (2014). "An angiosperm-wide analysis of the gynodioecy–dioecy pathway". In: *Annals of botany* 114.3, pp. 539–548.

Dunning, L. T. et al. (2016). "Ecological Speciation in Sympatric Palms: 1. Gene Expression, Selection and Pleiotropy". In: *Journal of Evolutionary Biology* 29.8, pp. 1472–1487. ISSN: 1420-9101. DOI: 10.1111/jeb.12895.

Durand, E. Y. et al. (2011). "Testing for Ancient Admixture between Closely Related Populations". In: *Molecular Biology and Evolution* 28.8, pp. 2239–2252. ISSN: 0737-4038. DOI: 10.1093/molbev/msr048.

Dym, M. and D. W. Fawcett (1971). "Further Observations on the Numbers of Spermatogonia, Spermatocytes, and Spermatids Connected by Intercellular Bridges in the Mammalian Testis1". In: *Biology of Reproduction* 4.2, pp. 195–215. ISSN: 0006-3363. DOI: 10.1093/biolreprod/4.2.195.

Edmands, S. (2007). "Between a Rock and a Hard Place: Evaluating the Relative Risks of Inbreeding and Outbreeding for Conservation and Management". In: *Molecular Ecology* 16.3, pp. 463–475. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2006.03148.x.

Edwards, J. A. and R. A. Edwards (2019). "Fastq-pair: efficient synchronization of paired-end fastq files". In: *BioRxiv*, p. 552885.

Elena, S. F. and R. E. Lenski (2001). "Epistasis between new mutations and genetic background and a test of genetic canalization". In: *Evolution* 55.9, pp. 1746–1752. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/j.0014-3820.2001.tb00824.x.

Ellegren, H. et al. (2012). "The Genomic Landscape of Species Divergence in Ficedula Flycatchers". In: *Nature* 491.7426 (7426), pp. 756–760. ISSN: 1476-4687. DOI: 10.1038/nature11584.

Endler, J. A. (1977). *Geographic Variation, Speciation, and Clines*. Princeton University Press. ISBN: 0-691-08192-1.

Epinat, G. and T. Lenormand (2009). "The evolution of assortative mating and selfing with in- and outbreeding depression". In: *Evolution* 63.8, pp. 2047–2060. ISSN: 00143820, 15585646. DOI: 10.1111/j.1558-5646.2009.00700.x.

Escobar-Restrepo, J.-M. et al. (2007). "The FERONIA Receptor-like Kinase Mediates Male-Female Interactions During Pollen Tube Reception". In: *Science* 317.5838, pp. 656–660. DOI: 10.1126/science.1143562.

Etienne, R. S., H. Morlon, and A. Lambert (2014). “Estimating the duration of speciation from phylogenies”. In: *Evolution* 68.8, pp. 2430–2440. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/evo.12433.

Excoffier, L. et al. (2021). “Fastsimcoal2: Demographic Inference under Complex Evolutionary Scenarios”. In: *Bioinformatics* 37.24, pp. 4882–4885. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab468.

Fagundes, N. J. R. et al. (2007). “Statistical Evaluation of Alternative Models of Human Evolution”. In: *Proceedings of the National Academy of Sciences* 104.45, pp. 17614–17619. DOI: 10.1073/pnas.0708280104.

Feder, J. L., S. P. Egan, and P. Nosil (2012). “The Genomics of Speciation-with-Gene-Flow”. In: *Trends in Genetics* 28.7, pp. 342–350. ISSN: 01689525. DOI: 10.1016/j.tig.2012.03.009.

Feng, Y., H. P. Comes, and Y.-X. Qiu (2020). “Phylogenomic Insights into the Temporal-Spatial Divergence History, Evolution of Leaf Habit and Hybridization in Stachyurus (Stachyuraceae)”. In: *Molecular Phylogenetics and Evolution* 150, p. 106878. ISSN: 10557903. DOI: 10.1016/j.ympev.2020.106878.

Ferreira de Carvalho, J. et al. (2019). “Cytonuclear Interactions Remain Stable during Allopolyploid Evolution despite Repeated Whole-Genome Duplications in Brassica”. In: *The Plant Journal* 98.3, pp. 434–447. ISSN: 1365-313X. DOI: 10.1111/tpj.14228.

Ferrer, M. M. and S. V. Good-Avila (2007). “Macrophylogenetic Analyses of the Gain and Loss of Self-Incompatibility in the Asteraceae”. In: *New Phytologist* 173.2, pp. 401–414. ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.2006.01905.x.

Fisher, R. A. (1941). “Average Excess and Average Effect of a Gene Substitution”. In: *Annals of Eugenics* 11.1, pp. 53–63. ISSN: 2050-1420.

Fishman, L. and J. H. Willis (2006). “A cytonuclear incompatibility causes anther sterility in *Mimulus* hybrids”. In: *Evolution* 60.7, pp. 1372–1381. ISSN: 00143820, 15585646. DOI: 10.1111/j.0014-3820.2006.tb01216.x.

Fraïsse, C., J. a. D. Elderfield, and J. J. Welch (2014). “The Genetics of Speciation: Are Complex Incompatibilities Easier to Evolve?” In: *Journal of Evolutionary Biology* 27.4, pp. 688–699. ISSN: 1010-061X. DOI: 10.1111/jeb.12339.

Fraïsse, C. et al. (2021). “DILS: Demographic Inferences with Linked Selection by Using ABC”. In: *Molecular Ecology Resources* 21.8, pp. 2629–2644. ISSN: 1755-098X, 1755-0998. DOI: 10.1111/1755-0998.13323.

Frank, S. A. (1989). “The Evolutionary Dynamics of Cytoplasmic Male Sterility”. In: *The American Naturalist* 133.3, pp. 345–376. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/284923.

Frankham, R., C. J. A. Bradshaw, and B. W. Brook (2014). “Genetics in Conservation Management: Revised Recommendations for the 50/500 Rules, Red List Criteria and Population Viability Analyses”. In: *Biological Conservation* 170, pp. 56–63. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2013.12.036.

Friedman, J. and S. C. H. Barrett (2009). “Wind of Change: New Insights on the Ecology and Evolution of Pollination and Mating in Wind-Pollinated Plants”. In: *Annals of Botany* 103.9, pp. 1515–1527. ISSN: 0305-7364. DOI: 10.1093/aob/mcp035.

Fu, L. et al. (2012). “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data”. In: *Bioinformatics (Oxford, England)* 28.23, pp. 3150–3152. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts565. pmid: 23060610.

Fukuda, A. et al. (2021). “DDBJ Update: Streamlining Submission and Access of Human Data”. In: *Nucleic Acids Research* 49.D1, pp. D71–D75. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa982. pmid: 33156332.

Galtier, N. (2019). “Delineating Species in the Speciation Continuum: A Proposal”. In: *Evolutionary Applications* 12.4, pp. 657–663. ISSN: 1752-4571, 1752-4571. DOI: 10.1111/eva.12748.

Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species (MPB-41)*. Princeton University Press. ISBN: 978-0-691-18705-1. DOI: 10.1515/9780691187051.

Gayral, P. et al. (2013). “Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap”. In: *PLOS Genetics* 9.4, e1003457. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003457.

Giraud, T. and S. Gourbière (2012). “The Tempo and Modes of Evolution of Reproductive Isolation in Fungi”. In: *Heredity* 109.4 (4), pp. 204–214. ISSN: 1365-2540. DOI: 10.1038/hdy.2012.30.

Goldberg, E. E. and B. Igić (2012). “Tempo and mode in plant breeding system evolution”. In: *Evolution* 66.12, pp. 3701–3709. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.2012.01730.x.

Goldberg, E. E. et al. (2010). “Species Selection Maintains Self-Incompatibility”. In: *Science* 330.6003, pp. 493–495. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1194513. pmid: 20966249.

Goodwillie, C., S. Kalisz, and C. G. Eckert (2005). “The Evolutionary Enigma of Mixed Mating Systems in Plants: Occurrence, Theoretical Explanations, and Empirical Evidence”. In: *Annual Review of Ecology, Evolution, and Systematics* 36.1, pp. 47–79. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev.ecolsys.36.091704.175539.

Gottlieb, L. D. (1984). “Genetics and Morphological Evolution in Plants”. In: *The American Naturalist* 123.5, pp. 681–709. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/284231.

Goudet, J. (2005). “Hierfstat, a Package for r to Compute and Test Hierarchical F-statistics”. In: *Molecular Ecology Notes* 5.1, pp. 184–186. ISSN: 1471-8286. DOI: 10.1111/j.1471-8286.2004.00828.x.

Goulet-Scott, B. E., A. G. Garner, and R. Hopkins (2021). “Genomic Analyses Overturn Two Long-standing Homoploid Hybrid Speciation Hypotheses”. In: *Evolution* 75.7, pp. 1699–1710. ISSN: 0014-3820. DOI: 10.1111/evo.14279.

Gouyon, P.-H. and D. Couvet (1987). “A Conflict between Two Sexes, Females and Hermaphrodites”. In: *The Evolution of Sex and Its Consequences*. Experientia Supplementum. Basel: Birkhäuser, pp. 245–261. ISBN: 978-3-0348-6273-8. DOI: 10.1007/978-3-0348-6273-8_11.

Gramlich, S., N. D. Wagner, and E. Hörandl (2018). “RAD-seq Reveals Genetic Structure of the F2-generation of Natural Willow Hybrids (Salix L.) and a Great Potential for Interspecific Introgression”. In: *BMC Plant Biology* 18.1, p. 317. ISSN: 1471-2229. DOI: 10.1186/s12870-018-1552-6.

Grant, P. R. and B. R. Grant (1992). “Hybridization of Bird Species”. In: *Science* 256.5054, pp. 193–197. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.256.5054.193.

Grant, V. (1949). “Pollination Systems as Isolating Mechanisms in Angiosperms”. In: *Evolution*, pp. 82–97. ISSN: 0014-3820.

— (1971). *Plant Speciation*. New York: Columbia Univ. Press. 435 pp.

Grant, V. and K. A. Grant (1965). “Flower Pollination in the Phlox Family”. In.

Green, R. E. et al. (2010). “A Draft Sequence of the Neandertal Genome”. In: *Science* 328.5979, pp. 710–722. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1188021.

Greiner, S. and R. Bock (2013). “Tuning a Ménage à Trois: Co-evolution and Co-Adaptation of Nuclear and Organellar Genomes in Plants”. In: *BioEssays* 35.4, pp. 354–365. ISSN: 02659247. DOI: 10.1002/bies.201200137.

Greiner, S. et al. (2011). “The Role of Plastids in Plant Speciation”. In: *Molecular Ecology* 20.4, pp. 671–691. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2010.04984.x.

Grover, C. E. et al. (2022). “Dual Domestication, Diversity, and Differential Introgression in Old World Cotton Diploids”. In: *Genome Biology and Evolution* 14.12, evac170. ISSN: 1759-6653. DOI: 10.1093/gbe/evac170. pmid: 36510772.

Grundt, H. H. et al. (2006). “High Biological Species Diversity in the Arctic Flora”. In: *Proceedings of the National Academy of Sciences* 103.4, pp. 972–975. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0510270103.

Grünig, S., M. Fischer, and C. Parisod (2021). “Recent Hybrid Speciation at the Origin of the Narrow Endemic Pulmonaria Helvetica”. In: *Annals of Botany* 127.1, pp. 21–31. ISSN: 0305-7364. DOI: 10.1093/aob/mcaa145.

Gutenkunst, R. et al. (2010). “Diffusion Approximations for Demographic Inference: DaDi”. In: *Nature Precedings*, pp. 1–1. ISSN: 1756-0357. DOI: 10.1038/npre.2010.4594.1.

Hahn, M. W. and L. Nakhleh (2016). “Irrational Exuberance for Resolved Species Trees”. In: *Evolution* 70.1, pp. 7–17. ISSN: 0014-3820. DOI: 10.1111/evo.12832.

Haig, D. and M. Westoby (1991). “Genomic Imprinting in Endosperm: Its Effect on Seed Development in Crosses between Species, and Its Implications for the Evolution of Apomixis”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 333.1266, pp. 1–13. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.1991.0057.

Hamrick, J. and M. Godt (1996). “Effects of Life History Traits on Genetic Diversity in Plant Species”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 351.1345, pp. 1291–1298. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.1996.0112.

Harr, B. (2006). “Genomic Islands of Differentiation between House Mouse Subspecies”. In: *Genome Research* 16.6, pp. 730–737. ISSN: 1088-9051. DOI: 10.1101/gr.5045006.

Hartig, F. (2022). *DHARMa: Residual Diagnostis for Hierarchical (Multi-Level / Mixed) Regression Models*. Version R package version 0.4.6.

Hatfield, T., N. Barton, and J. B. Searle (1992). “A model of a hybrid zone between two chromosomal races of the common shrew (*Sorex araneus*)”. In: *Evolution* 46.4, pp. 1129–1145. ISSN: 00143820. DOI: 10.1111/j.1558-5646.1992.tb00624.x.

Helmstetter, A. J. et al. (2023). “Trait-dependent Diversification in Angiosperms: Patterns, Models and Data”. In: *Ecology Letters* 26.4, pp. 640–657. ISSN: 1461-023X.

Heslop-Harrison, J. (and T. Schmidt (2012). “Plant Nuclear Genome Composition”. In: *eLS*. 1st ed. Wiley. ISBN: 978-0-470-01617-6 978-0-470-01590-2. DOI: 10.1002/9780470015902.a0002014.pub2.

Hewitt, G. M. (1988). “Hybrid Zones-Natural Laboratories for Evolutionary Studies”. In: *Trends in Ecology & Evolution* 3.7, pp. 158–167. ISSN: 01695347. DOI: 10.1016/0169-5347(88)90033-X.

Hey, J. and R. Nielsen (2007). “Integration within the Felsenstein Equation for Improved Markov Chain Monte Carlo Methods in Population Genetics”. In: *Proceedings of the National Academy of Sciences* 104.8, pp. 2785–2790. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0611164104.

Hibbins, M. S. and M. W. Hahn (2022). “Phylogenomic Approaches to Detecting and Characterizing Introgression”. In: *Genetics* 220.2, iyab173. ISSN: 1943-2631. DOI: 10.1093/genetics/iyab173.

Higashiyama, T. et al. (2006). “Species Preferentiality of the Pollen Tube Attractant Derived from the Synergid Cell of Torenia Fournieri”. In: *Plant Physiology* 142.2, pp. 481–491. ISSN: 0032-0889. DOI: 10.1104/pp.106.083832.

Hijmans, R. J. et al. (2022). *Geosphere: Spherical Trigonometry*. Version 1.5-18.

Hill, G. E. (2016). “Mitonuclear Coevolution as the Genesis of Speciation and the Mitochondrial DNA Barcode Gap”. In: *Ecology and Evolution* 6.16, pp. 5831–5842. ISSN: 2045-7758. DOI: 10.1002/ece3.2338.

Hogenboom, N. G. et al. (1997). “Incompatibility and Incongruity: Two Different Mechanisms for the Non-Functioning of Intimate Partner Relationships”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 188.1092, pp. 361–375. DOI: 10.1098/rspb.1975.0025.

Hohenlohe, P. A. et al. (2012). “Extensive Linkage Disequilibrium and Parallel Adaptive Divergence across Threespine Stickleback Genomes”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587, pp. 395–408. DOI: 10.1098/rstb.2011.0245.

Hu, X.-S. (2015). “Mating System as a Barrier to Gene Flow: MATING SYSTEM AND ISOLATING BARRIER”. In: *Evolution* 69.5, pp. 1158–1177. ISSN: 00143820. DOI: 10.1111/evo.12660.

Hu, X.-S. and D. A. Filatov (2016). “The Large-X Effect in Plants: Increased Species Divergence and Reduced Gene Flow on the *Silene* X-chromosome”. In: *Molecular Ecology* 25.11, pp. 2609–2619. ISSN: 09621083. DOI: 10.1111/mec.13427.

Hudson, R. R. (1983). “Properties of a neutral allele model with intragenic recombination”. In: *Theoretical population biology* 23.2, pp. 183–201.

Igic, B. and J. R. Kohn (2006). “The Distribution of Plant Mating Systems: Study Bias Against Obligay Outcrossing Species”. In: *Evolution* 60.5, pp. 1098–1103. ISSN: 1558-5646. DOI: 10.1111/j.0014-3820.2006.tb01186.x.

Igic, B., R. Lande, and J. R. Kohn (2008). “Loss of Self-Incompatibility and Its Evolutionary Consequences”. In: *International Journal of Plant Sciences* 169.1, pp. 93–104. ISSN: 1058-5893. DOI: 10.1086/523362.

İltaş, Ö. et al. (2021). “Early Evolution of Reproductive Isolation: A Case of Weak Inbreeder/Strong Outbreeder Leads to an Intraspecific Hybridization Barrier in *Arabidopsis Lyrata*”. In: *Evolution* 75.6, pp. 1466–1476. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/evo.14240.

Immler, S. (2019). “Haploid Selection in “Diploid” Organisms”. In: *Annual Review of Ecology, Evolution, and Systematics* 50.1, pp. 219–236. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev-ecolsys-110218-024709.

Ishizaki, S., T. Abe, and M. Ohara (2013). “Mechanisms of Reproductive Isolation of Interspecific Hybridization between *Trillium Camschatcense* and *T. Tschonoskii* (Melanthiaceae): REPRODUCTIVE ISOLATIONS OF *TRILLIUM*”. In: *Plant Species Biology* 28.3, pp. 204–214. ISSN: 0913557X. DOI: 10.1111/j.1442-1984.2012.00378.x.

Ives, A. R. (2019). “R^2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs”. In: *Systematic Biology* 68.2, pp. 234–251. ISSN: 1063-5157. DOI: 10.1093/sysbio/syy060.

Ives, A. R. and D. Li (2018). “‘rr2‘: An R Package to Calculate \$R2̂\$s for Regression Models”. In: *Journal of Open Source Software* 3.30, p. 1028. ISSN: 2475-9066. DOI: 10.21105/joss.01028.

Jain, S. K. (1976). “The Evolution of Inbreeding in Plants”. In: *Annual Review of Ecology and Systematics* 7.1, pp. 469–495. DOI: 10.1146/annurev.es.07.110176.002345.

Jeon, K. W. (2004). *International Review of Cytology: A Survey of Cell Biology*. Elsevier. 298 pp. ISBN: 978-0-08-049564-4. Google Books: 7K8Hh5YdXVkC.

Johnson, N. A. (2000). “Speciation: Dobzhansky–Muller Incompatibilities, Dominance and Gene Interactions”. In: *Trends in Ecology & Evolution* 15.12, pp. 480–482. ISSN: 01695347. DOI: 10.1016/S0169-5347(00)01961-3.

Johnson, S. D. (2010). “The Pollination Niche and Its Role in the Diversification and Maintenance of the Southern African Flora”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1539, pp. 499–516. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2009.0243.

Joseph, S. and M. Kirkpatrick (2004). “Haploid Selection in Animals”. In: *Trends in Ecology & Evolution* 19.11, pp. 592–597. ISSN: 01695347. DOI: 10.1016/j.tree.2004.08.004.

Jouganous, J. et al. (2017). “Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation”. In: *Genetics* 206.3, pp. 1549–1567. ISSN: 1943-2631. DOI: 10.1534/genetics.117.200493.

Katz, K. et al. (2022). “The sequence read archive: a decade more of explosive growth”. In: *Nucleic acids research* 50.D1, pp. D387–D390.

Kay, K. M. and R. D. Sargent (2009). “The Role of Animal Pollination in Plant Speciation: Integrating Ecology, Geography, and Genetics”. In: *Annual Review of Ecology, Evolution, and Systematics* 40.1, pp. 637–656. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev.ecolsys.110308.120310.

Keinan, A. and D. Reich (2010). "Human Population Differentiation Is Strongly Correlated with Local Recombination Rate". In: *PLOS Genetics* 6.3, e1000886. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000886.

Kermicle, J. L. and M. M. Evans (2005). "Pollen–Pistil Barriers to Crossing in Maize and Teosinte Result from Incongruity Rather than Active Rejection". In: *Sexual Plant Reproduction* 18.4, pp. 187–194. ISSN: 1432-2145. DOI: 10.1007/s00497-005-0012-2.

Kern, A. D. and F. A. Kondrashov (2004). "Mechanisms and Convergence of Compensatory Evolution in Mammalian Mitochondrial tRNAs". In: *Nature Genetics* 36.11 (11), pp. 1207–1212. ISSN: 1546-1718. DOI: 10.1038/ng1451.

Kheyr-Pour, A. (1980). "Nucleo-Cytoplasmic Polymorphism for Male Sterility in Origanum Vulgare L." In: *Journal of Heredity* 71.4, pp. 253–260. ISSN: 1465-7333, 0022-1503. DOI: 10.1093/oxfordjournals.jhered.a109359.

— (1981). "Wide Nucleo-Cytoplasmic Polymorphism for Male Sterility in Origanum Vulgare L." In: *Journal of Heredity* 72.1, pp. 45–51. ISSN: 1465-7333, 0022-1503. DOI: 10.1093/oxfordjournals.jhered.a109424.

Kinoshita, T. (2007). "Reproductive Barrier and Genomic Imprinting in the Endosperm of Flowering Plants". In: *Genes & Genetic Systems* 82.3, pp. 177–186. ISSN: 1341-7568, 1880-5779. DOI: 10.1266/ggs.82.177.

Kondrashov, A. S., S. Sunyaev, and F. A. Kondrashov (2002). "Dobzhansky–Muller Incompatibilities in Protein Evolution". In: *Proceedings of the National Academy of Sciences* 99.23, pp. 14878–14883. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.232565499.

Kuboyama, T., C. Chung, and G. Takeda (1994). "The Diversity of Interspecific Pollen-Pistil Incongruity in Nicotiana". In: *Sexual Plant Reproduction* 7.4. ISSN: 0934-0882, 1432-2145. DOI: 10.1007/BF00232744.

Kumar, S. et al. (2017). "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times". In: *Molecular biology and evolution* 34.7, pp. 1812–1819. ISSN: 0737-4038.

Kumar, S. et al. (2022). "TimeTree 5: an expanded resource for species divergence times". In: *Molecular Biology and Evolution* 39.8, msac174.

Lafon-Placette, C. et al. (2017). "Endosperm-Based Hybridization Barriers Explain the Pattern of Gene Flow between Arabidopsis Lyrata and Arabidopsis Arenosa in Central Europe". In: *Proceedings of the National Academy of Sciences* 114.6, E1027–E1035. DOI: 10.1073/pnas.1615123114.

Lamichhaney, S. et al. (2018). "Rapid Hybrid Speciation in Darwin's Finches". In: *Science* 359.6372, pp. 224–228. DOI: 10.1126/science.aao4593.

Langmead, B. and S. L. Salzberg (2012). "Fast Gapped-Read Alignment with Bowtie 2". In: *Nature Methods* 9.4 (4), pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923.

Leducq, J.-B. et al. (2013). "Intriguing Small-Scale Spatial Distribution of Chloropastic and Nuclear Diversity in the Endangered Plant Biscutella Neustriaca (Brassicaceae)". In: *Conservation Genetics* 14.1, pp. 65–77. ISSN: 1572-9737. DOI: 10.1007/s10592-012-0426-y.

Leebens-Mack, J. H. et al. (2019). "One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants". In: *Nature* 574.7780 (7780), pp. 679–685. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1693-2.

Levin, D. A. (2012). “The Long Wait for Hybrid Sterility in Flowering Plants”. In: *New Phytologist* 196.3, pp. 666–670. ISSN: 0028-646X.

— (1971). “The origin of reproductive isolating mechanisms in flowering plants”. In: *TAXON* 20.1, pp. 91–113. ISSN: 0040-0262, 1996-8175. DOI: 10.2307/1218538.

Lewis, D. and L. K. Crowe (1958). “Unilateral Interspecific Incompatibility in Flowering Plants”. In: *Heredity* 12.2 (2), pp. 233–256. ISSN: 1365-2540. DOI: 10.1038/hdy.1958.26.

Lewontin, R. C. and J. Krakauer (1973). “Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms”. In: *Genetics* 74.1, pp. 175–195. ISSN: 1943-2631. DOI: 10.1093/genetics/74.1.175.

Li, D. et al. (2020). “Phyr: An r Package for Phylogenetic Species-Distribution Modelling in Ecological Communities”. In: *Methods in Ecology and Evolution* 11.11, pp. 1455–1463. ISSN: 2041-210X. DOI: 10.1111/2041-210X.13471.

Li, W. and A. Godzik (2006). “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences”. In: *Bioinformatics (Oxford, England)* 22.13, pp. 1658–1659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl158. pmid: 16731699.

Li, Z. et al. (2021). “Patterns and Processes of Diploidization in Land Plants”. In: *Annual Review of Plant Biology* 72.1, pp. 387–410. DOI: 10.1146/annurev-arplant-050718-100344. pmid: 33684297.

Liu, Y. et al. (2017). “Rapid Radiations of Both Kiwifruit Hybrid Lineages and Their Parents Shed Light on a Two-Layer Mode of Species Diversification”. In: *New Phytologist* 215.2, pp. 877–890. ISSN: 1469-8137. DOI: 10.1111/nph.14607.

Lohse, K. (2017). “Come on Feel the Noise - from Metaphors to Null Models”. In: *Journal of Evolutionary Biology* 30.8, pp. 1506–1508. ISSN: 1010061X. DOI: 10.1111/jeb.13109.

Loveless, M. D. and J. L. Hamrick (1984). “Ecological Determinants of Genetic Structure in Plant Populations”. In: *Annual Review of Ecology and Systematics* 15.1, pp. 65–95. ISSN: 0066-4162. DOI: 10.1146/annurev.es.15.110184.000433.

Lutz, A. M. (1907). “A Preliminary Note on the Chromosomes of Oenothera Lamarckiana and One of Its Mutants, O. Gigas”. In: *Science* 26.657, pp. 151–152. ISSN: 0036-8075.

Lynch, M. and L. A. Real (1994). “Neutral Models of Phenotypic Evolution”. In: *Ecological genetics*, pp. 86–108.

Lynch, M. and A. G. Force (2000). “The Origin of Interspecific Genomic Incompatibility via Gene Duplication”. In: *The American Naturalist* 156.6, pp. 590–605. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/316992.

Mable, B. K. and S. P. Otto (1998). “The Evolution of Life Cycles with Haploid and Diploid Phases”. In: *BioEssays* 20.6, pp. 453–462. ISSN: 02659247. DOI: 10.1002/(SICI)1521-1878(199806)20:6<453::AID-BIES3>3.0.CO;2-N.

Mallet, J. (2005). “Hybridization as an Invasion of the Genome”. In: *Trends in Ecology & Evolution* 20.5, pp. 229–237. ISSN: 01695347. DOI: 10.1016/j.tree.2005.02.010.

Mallet, J. and S. P. Mullen (2022). “Reproductive Isolation Is a Heuristic, Not a Measure: A Commentary on Westram et al., 2022”. In: *Journal of Evolutionary Biology* 35.9, pp. 1175–1182. ISSN: 1420-9101. DOI: 10.1111/jeb.14052.

Marie-Orleach, L., C. Brochmann, and S. Glémin (2022). “Mating System and Speciation I: Accumulation of Genetic Incompatibilities in Allopatry”. In: *PLOS Genetics* 18.12, e1010353. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1010353.

Marques, D. A., J. I. Meier, and O. Seehausen (2019). “A Combinatorial View on Speciation and Adaptive Radiation”. In: *Trends in Ecology & Evolution* 34.6, pp. 531–544. ISSN: 01695347. DOI: 10.1016/j.tree.2019.02.008.

Martin, M. (2011). “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads”. In: *EMBnet.journal* 17.1 (1), pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200.

Martin, S. H., J. W. Davey, and C. D. Jiggins (2015). “Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci”. In: *Molecular Biology and Evolution* 32.1, pp. 244–257. ISSN: 0737-4038. DOI: 10.1093/molbev/msu269.

Matthey-Doret, R. and M. C. Whitlock (2019). “Background Selection and F_{ST} : Consequences for Detecting Local Adaptation”. In: *Molecular Ecology* 28.17, pp. 3902–3914. ISSN: 0962-1083, 1365-294X. DOI: 10.1111/mec.15197.

Matute, D. R. et al. (2010). “A Test of the Snowball Theory for the Rate of Evolution of Hybrid Incompatibilities”. In: *Science* 329.5998, pp. 1518–1521. ISSN: 0036-8075.

Mayr, E. (1963). “Animal Species and Evolution”. In: p. 824.

McFarlane, S. E. and J. M. Pemberton (2019). “Detecting the True Extent of Introgression during Anthropogenic Hybridization”. In: *Trends in Ecology & Evolution* 34.4, pp. 315–326. ISSN: 01695347. DOI: 10.1016/j.tree.2018.12.013.

Mitchell, N. et al. (2019). “Correlates of Hybridization in Plants”. In: *Evolution Letters* 3.6, pp. 570–585. ISSN: 2056-3744. DOI: 10.1002/evl3.146.

Morjan, C. L. and L. H. Rieseberg (2004). “How Species Evolve Collectively: Implications of Gene Flow and Selection for the Spread of Advantageous Alleles”. In: *Molecular Ecology* 13.6, pp. 1341–1356. ISSN: 09621083, 1365294X. DOI: 10.1111/j.1365-294X.2004.02164.x.

Moyle, L. C. and T. Nakazato (2010). “Hybrid Incompatibility “Snowballs” Between Solanum Species”. In: *Science*.

Mugal, C. F., B. Nabholz, and H. Ellegren (2013). “Genome-Wide Analysis in Chicken Reveals That Local Levels of Genetic Diversity Are Mainly Governed by the Rate of Recombination”. In: *BMC Genomics* 14.1, p. 86. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-86.

Murlas Cosmides, L. and J. Tooby (1981). “Cytoplasmic Inheritance and Intragenomic Conflict”. In: *Journal of Theoretical Biology* 89.1, pp. 83–129. ISSN: 00225193. DOI: 10.1016/0022-5193(81)90181-8.

Muyle, A. et al. (2021). “Dioecy Is Associated with High Genetic Diversity and Adaptation Rates in the Plant Genus *Silene*”. In: *Molecular Biology and Evolution* 38.3, pp. 805–818. ISSN: 1537-1719. DOI: 10.1093/molbev/msaa229.

Nadeau, N. J. et al. (2012). “Genomic Islands of Divergence in Hybridizing Heliconius Butterflies Identified by Large-Scale Targeted Sequencing”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences*. DOI: 10.1098/rstb.2011.0198.

Nagylaki, T. (1976). “Clines with variable migration”. In: *Genetics* 83.4, pp. 867–886. ISSN: 1943-2631. DOI: 10.1093/genetics/83.4.867.

Nathan, R. (2006). “Long-Distance Dispersal of Plants”. In: *Science* 313.5788, pp. 786–788. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1124975.

Navarro, A. and N. H. Barton (2003). “Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation”. In: *Evolution* 57.3, pp. 447–459. ISSN: 0014-3820. DOI: 10.1111/j.0014-3820.2003.tb01537.x.

Nei, M. and W. H. Li (1979). “Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases”. In: *Proceedings of the National Academy of Sciences* 76.10, pp. 5269–5273. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.76.10.5269. pmid: 291943.

Nevado, B. et al. (2016). “Widespread Adaptive Evolution during Repeated Evolutionary Radiations in New World Lupins”. In: *Nature Communications* 7.1 (1), p. 12384. ISSN: 2041-1723. DOI: 10.1038/ncomms12384.

Nevado, B. et al. (2020). “Rapid Homoploid Hybrid Speciation in British Gardens: The Origin of Oxford Ragwort (Senecio Squalidus)”. In: *Molecular Ecology* 29.21, pp. 4221–4233. ISSN: 1365-294X. DOI: 10.1111/mec.15630.

Newbold, T. et al. (2015). “Global Effects of Land Use on Local Terrestrial Biodiversity”. In: *Nature* 520.7545 (7545), pp. 45–50. ISSN: 1476-4687. DOI: 10.1038/nature14324.

Nielsen, R. and J. Wakeley (2001). “Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach”. In: *Genetics* 158.2, pp. 885–896. ISSN: 1943-2631. DOI: 10.1093/genetics/158.2.885.

Nieuwenhuis, B. P. S. and T. Y. James (2016). “The Frequency of Sex in Fungi”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1706, p. 20150540. DOI: 10.1098/rstb.2015.0540.

Noor, M. a. F. and S. M. Bennett (2009). “Islands of Speciation or Mirages in the Desert? Examining the Role of Restricted Recombination in Maintaining Species”. In: *Heredity* 103.6 (6), pp. 439–444. ISSN: 1365-2540. DOI: 10.1038/hdy.2009.151.

Norrell, J. (2017). “Differentiating the neches river rose mallow (*Hibiscus dasycalyx*) from its congeners by means of phylogenetics and population genetics”.

Nosil, P. (2012). *Ecological Speciation*. Oxford University Press. ISBN: 0-19-958711-6.

Ollerton, J., R. Winfree, and S. Tarrant (2011). “How Many Flowering Plants Are Pollinated by Animals?” In: *Oikos* 120.3, pp. 321–326. ISSN: 00301299. DOI: 10.1111/j.1600-0706.2010.18644.x.

Oneal, E., J. H. Willis, and R. G. Franks (2016). “Disruption of Endosperm Development Is a Major Cause of Hybrid Seed Inviability between Mimulus Guttatus and Mimulus Nudatus”. In: *New Phytologist* 210.3, pp. 1107–1120. ISSN: 1469-8137. DOI: 10.1111/nph.13842.

Orr, H. A. (1995). “The Population Genetics of Speciation: The Evolution of Hybrid Incompatibilities.” In: *Genetics* 139.4, pp. 1805–1813. ISSN: 1943-2631. DOI: 10.1093/genetics/139.4.1805.

Orr, H. A. and S. Irving (2001). “Complex Epistasis and the Genetic Basis of Hybrid Sterility in the Drosophila Pseudoobscura Bogota-USA Hybridization”. In: *Genetics* 158.3, pp. 1089–1100. ISSN: 1943-2631. DOI: 10.1093/genetics/158.3.1089.

Orr, H. A. and M. Turelli (2001). “The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities”. In: *Evolution* 55.6, pp. 1085–1094. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/j.0014-3820.2001.tb00628.x.

Ortego, J. and L. L. Knowles (2020). “Incorporating Interspecific Interactions into Phylogeographic Models: A Case Study with Californian Oaks”. In: *Molecular Ecology* 29.23, pp. 4510–4524. ISSN: 1365-294X. DOI: 10.1111/mec.15548.

Osborne, O. G. et al. (2019). “Speciation in Howea Palms Occurred in Sympatry, Was Preceded by Ancestral Admixture, and Was Associated with Edaphic and Phenological Adaptation”. In: *Molecular Biology and Evolution* 36.12, pp. 2682–2697. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msz166.

Otto, S. P., M. F. Scott, and S. Immler (2015). “Evolution of Haploid Selection in Predominantly Diploid Organisms”. In: *Proceedings of the National Academy of Sciences* 112.52, pp. 15952–15957. DOI: 10.1073/pnas.1512004112.

Owens, G. L. et al. (2021). “Standing Variation Rather than Recent Adaptive Introgression Probably Underlies Differentiation of the Texanus Subspecies of Helianthus Annuus”. In: *Molecular Ecology* 30.23, pp. 6229–6245. ISSN: 0962-1083.

Palopoli, M. F. and C.-I. Wu (1994). “Genetics of Hybrid Male Sterility between Drosophila Sibling Species: A Complex Web of Epistasis Is Revealed in Interspecific Studies.” In: *Genetics* 138.2, pp. 329–341. ISSN: 1943-2631.

Pamilo, P. and M. Nei (1988). “Relationships between gene trees and species trees.” In: *Molecular biology and evolution* 5.5, pp. 568–583.

Pannell, J. R. and S. C. Barrett (1998). “Baker’s Law Revisited: Reproductive Assurance in a Metapopulation”. In: *Evolution* 52.3, pp. 657–668. ISSN: 0014-3820.

Paris, J. R., J. R. Stevens, and J. M. Catchen (2017). “Lost in Parameter Space: A Road Map for Stacks”. In: *Methods in Ecology and Evolution* 8.10, pp. 1360–1373. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12775.

Parker, G. A. and L. Partridge (1998). “Sexual Conflict and Speciation”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 353.1366, pp. 261–274. ISSN: 0962-8436.

Payseur, B. A. and L. H. Rieseberg (2016). “A Genomic Perspective on Hybridization and Speciation”. In: *Molecular Ecology* 25.11, pp. 2337–2360. ISSN: 0962-1083, 1365-294X. DOI: 10.1111/mec.13557.

Perez, D. E. and C.-I. Wu (1995). “Further Characterization of the Odysseus Locus of Hybrid Sterility in Drosophila: One Gene Is Not Enough.” In: *Genetics* 140.1, pp. 201–206. ISSN: 1943-2631.

Petit, R. J. and A. Hampe (2006). “Some Evolutionary Consequences of Being a Tree”. In: *Annual Review of Ecology, Evolution, and Systematics* 37.1, pp. 187–214. ISSN: 1543-592X. DOI: 10.1146/annurev.ecolsys.37.091305.110215.

Pickup, M. et al. (2019). “Mating System Variation in Hybrid Zones: Facilitation, Barriers and Asymmetries to Gene Flow”. In: *New Phytologist* 224.3, pp. 1035–1047. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.16180.

Postel, Z. and P. Touzet (2020). “Cytonuclear Genetic Incompatibilities in Plant Speciation”. In: *Plants* 9.4 (4), p. 487. ISSN: 2223-7747. DOI: 10.3390/plants9040487.

Presgraves, D. C. (2002). “Patterns of postzygotic isolation in lepidoptera”. In: *Evolution* 56.6, pp. 1168–1183. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/j.0014-3820.2002.tb01430.x.

Price, T. D. and M. M. Bouvier (2002). “The evolution of F $_1$ postzygotic incompatibilities in birds”. In: *Evolution* 56.10, pp. 2083–2089. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/j.0014-3820.2002.tb00133.x.

Pritchard, J. K. et al. (1999). “Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites.” In: *Molecular Biology and Evolution* 16.12, pp. 1791–1798. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026091.

Rand, D. M., R. A. Haney, and A. J. Fry (2004). “Cytonuclear Coevolution: The Genomics of Cooperation”. In: *Trends in Ecology & Evolution* 19.12, pp. 645–653. ISSN: 01695347. DOI: 10.1016/j.tree.2004.10.003.

Rausher, M. D. (2017). “Selfing, Local Mate Competition, and Reinforcement”. In: *The American Naturalist* 189.2, pp. 87–104. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/690009.

Ravinet, M. et al. (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. ISSN: 1420-9101. DOI: 10.1111/jeb.13047.

Rebernig, C. A. et al. (2015). “Non-Reciprocal Interspecies Hybridization Barriers in the Capsella Genus Are Established in the Endosperm”. In: *PLOS Genetics* 11.6, e1005295. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005295.

Reik, W. and J. Walter (2001). “Genomic Imprinting: Parental Influence on the Genome”. In: *Nature Reviews Genetics* 2.1, pp. 21–32. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/35047554.

Renner, S. S. and R. E. Ricklefs (1995). “Dioecy and Its Correlates in the Flowering Plants”. In: *American Journal of Botany* 82.5, pp. 596–606. ISSN: 1537-2197. DOI: 10.1002/j.1537-2197.1995.tb11504.x.

Rice, P., I. Longden, and A. Bleasby (2000). “EMBOSS: The European Molecular Biology Open Software Suite”. In: *Trends in Genetics* 16.6, pp. 276–277. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(00)02024-2. pmid: 10827456.

Rieseberg, L. H. (2001). “Chromosomal Rearrangements and Speciation”. In.

Rieseberg, L. H. and B. K. Blackman (2010). “Speciation Genes in Plants”. In: *Annals of Botany* 106.3, pp. 439–455. ISSN: 0305-7364. DOI: 10.1093/aob/mcq126.

Roth, M. et al. (2018). “Incidence and Developmental Timing of Endosperm Failure in Post-Zygotic Isolation between Wild Tomato Lineages”. In: *Annals of Botany* 121.1, pp. 107–118. ISSN: 0305-7364. DOI: 10.1093/aob/mcx133.

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. ISSN: 1420-9101. DOI: 10.1111/jeb.12425.

Roux, C. et al. (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. ISSN: 0737-4038. DOI: 10.1093/molbev/mst066.

Roux, C. et al. (2016). “Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence”. In: *PLOS Biology* 14.12, e2000234. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2000234.

Royer, A. M., M. A. Streisfeld, and C. I. Smith (2016). “Population Genomics of Divergence within an Obligate Pollination Mutualism: Selection Maintains Differences between Joshua Tree Species”. In: *American Journal of Botany* 103.10, pp. 1730–1741. ISSN: 1537-2197. DOI: 10.3732/ajb.1600069.

Ru, D. et al. (2018). “Population Genomic Analysis Reveals That Homoploid Hybrid Speciation Can Be a Lengthy Process”. In: *Molecular Ecology* 27.23, pp. 4875–4887. ISSN: 09621083. DOI: 10.1111/mec.14909.

Rutley, N. and D. Twell (2015). “A Decade of Pollen Transcriptomics”. In: *Plant Reproduction* 28.2, pp. 73–89. ISSN: 2194-7961. DOI: 10.1007/s00497-015-0261-7.

Sachdeva, H. (2019). “Effect of Partial Selfing and Polygenic Selection on Establishment in a New Habitat”. In: *Evolution* 73.9, pp. 1729–1745. ISSN: 0014-3820. DOI: 10.1111/evo.13812.

Sandstedt, G. D. and A. L. Sweigart (2022). “Developmental Evidence for Parental Conflict in Driving Mimulus Species Barriers”. In: *New Phytologist* 236.4, pp. 1545–1557. ISSN: 1469-8137. DOI: 10.1111/nph.18438.

Sasa, M. M., P. T. Chippindale, and N. A. Johnson (1998). “PATTERNS OF POSTZYGOTIC ISOLATION IN FROGS”. In: *Evolution* 52.6, pp. 1811–1820. ISSN: 00143820. DOI: 10.1111/j.1558-5646.1998.tb02258.x.

Sayers, E. W. et al. (2022). “Database Resources of the National Center for Biotechnology Information”. In: *Nucleic Acids Research* 50.D1, pp. D20–D26. ISSN: 1362-4962. DOI: 10.1093/nar/gkab1112. pmid: 34850941.

Scannell, D. R. et al. (2006). “Multiple Rounds of Speciation Associated with Reciprocal Gene Loss in Polyploid Yeasts”. In: *Nature* 440.7082 (7082), pp. 341–345. ISSN: 1476-4687. DOI: 10.1038/nature04562.

Scharmann, M., A. Wistuba, and A. Widmer (2021). “Introgression Is Widespread in the Radiation of Carnivorous Nepenthes Pitcher Plants”. In: *Molecular Phylogenetics and Evolution* 163, p. 107214. ISSN: 1055-7903. DOI: 10.1016/j.ympev.2021.107214.

Schemske, D. W. and H. D. Bradshaw (1999). “Pollinator Preference and the Evolution of Floral Traits in Monkeyflowers (Mimulus)”. In: *Proceedings of the National Academy of Sciences* 96.21, pp. 11910–11915. DOI: 10.1073/pnas.96.21.11910.

Schiestl, F. P. and P. M. Schlüter (2009). “Floral Isolation, Specialized Pollination, and Pollinator Behavior in Orchids”. In: *Annual Review of Entomology* 54.1, pp. 425–446. ISSN: 0066-4170, 1545-4487. DOI: 10.1146/annurev.ento.54.110807.090603.

Seehausen, O., J. J. van Alphen, and F. Witte (1997). “Cichlid Fish Diversity Threatened by Eutrophication That Curbs Sexual Selection”. In: *Science* 277.5333, pp. 1808–1811. ISSN: 1095-9203.

Shang, H. et al. (2020). “Evolution of Strong Reproductive Isolation in Plants: Broad-Scale Patterns and Lessons from a Perennial Model Group”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1806, p. 20190544. DOI: 10.1098/rstb.2019.0544.

Shi, Q. et al. (2023). “Autoploid Origin and Rapid Diploidization of the Tetraploid *Thinopyrum Elongatum* Revealed by Genome Differentiation and Chromosome Pairing in Meiosis”. In: *The Plant Journal* 113.3, pp. 536–545. ISSN: 0960-7412, 1365-313X. DOI: 10.1111/tpj.16066.

Shimizu, K. K. and T. Tsuchimatsu (2015). “Evolution of Selfing: Recurrent Patterns in Molecular Adaptation”. In: *Annual Review of Ecology, Evolution, and Systematics* 46.1, pp. 593–622. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev-ecolsys-112414-054249.

Sicard, A. and M. Lenhard (2011). “The Selfing Syndrome: A Model for Studying the Genetic and Evolutionary Basis of Morphological Adaptation in Plants”. In: *Annals of Botany* 107.9, pp. 1433–1443. ISSN: 0305-7364. DOI: 10.1093/aob/mcr023.

Sloan, D. B. et al. (2014). “Cytonuclear Interactions and Relaxed Selection Accelerate Sequence Evolution in Organelle Ribosomes”. In: *Molecular Biology and Evolution* 31.3, pp. 673–682. ISSN: 0737-4038. DOI: 10.1093/molbev/mst259.

Sloan, D. B. et al. (2018). “Cytonuclear Integration and Co-Evolution”. In: *Nature Reviews Genetics* 19.10, pp. 635–648. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-018-0035-9.

Smadja, C. M. and R. K. Butlin (2011). “A Framework for Comparing Processes of Speciation in the Presence of Gene Flow”. In: *Molecular Ecology* 20.24, pp. 5123–5140. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2011.05350.x.

Sobel, J. M. and G. F. Chen (2014). “Unification of methods for estimating the strength of reproductive isolation”. In: *Evolution* 68.5, pp. 1511–1522. ISSN: 0014-3820. DOI: 10.1111/evo.12362.

Souto-Vilarós, D. et al. (2018). “Pollination along an Elevational Gradient Mediated Both by Floral Scent and Pollinator Compatibility in the Fig and Fig-Wasp Mutualism”. In: *Journal of Ecology* 106.6, pp. 2256–2273. ISSN: 1365-2745. DOI: 10.1111/1365-2745.12995.

Sponsel, L. E. (2013). “Human Impact on Biodiversity, Overview”. In: *Encyclopedia of Biodiversity*. Elsevier, pp. 137–152. ISBN: 978-0-12-384720-1. DOI: 10.1016/B978-0-12-384719-5.00250-1.

Städler, T., A. M. Florez-Rueda, and M. Paris (2012). “Testing for “Snowballing” Hybrid Incompatibilities in Solanum: Impact of Ancestral Polymorphism and Divergence Estimates”. In: *Molecular Biology and Evolution* 29.1, pp. 31–34. ISSN: 0737-4038. DOI: 10.1093/molbev/msr218.

Stankowski, S. and M. Ravinet (2021). “Defining the Speciation Continuum”. In: *Evolution* 75.6, pp. 1256–1273. ISSN: 1558-5646. DOI: 10.1111/evo.14215.

Stebbins, G. L. (1950). *Variation and Evolution in Plants*. Columbia University Press. ISBN: 0-231-89916-5.

— (1957). “Self Fertilization and Population Variability in the Higher Plants”. In: *The American Naturalist* 91.861, pp. 337–354. ISSN: 0003-0147.

— (1970). “Adaptive Radiation of Reproductive Characteristics in Angiosperms, I: Pollination Mechanisms”. In: *Annual Review of Ecology and Systematics* 1.1, pp. 307–326. ISSN: 0066-4162.

— (1959). “The Role of Hybridization in Evolution”. In: *Proceedings of the American Philosophical Society* 103.2, pp. 231–251. ISSN: 0003-049X. JSTOR: 985151.

— (1974). *Flowering Plants: Evolution Above the Species Level*. Harvard University Press. ISBN: 978-0-674-86484-9. DOI: 10.4159/harvard.9780674864856.

Steger, K. (1999). “Transcriptional and Translational Regulation of Gene Expression in Haploid Spermatids”. In: *Anatomy and Embryology* 199.6, pp. 471–487. ISSN: 0340-2061, 1432-0568. DOI: 10.1007/s004290050245.

Stoffel, M. A. et al. (2016). “inbreedR: An R Package for the Analysis of Inbreeding Based on Genetic Markers”. In: *Methods in Ecology and Evolution* 7.11, pp. 1331–1339. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12588.

Sun, Y. et al. (2018). “Reticulate Evolution within a Spruce (*Picea*) Species Complex Revealed by Population Genomic Analysis”. In: *Evolution* 72.12, pp. 2669–2681. ISSN: 0014-3820, 1558-5646. DOI: 10.1111/evo.13624.

Szymura, J. M. and N. H. Barton (1986). “Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland”. In: *Evolution* 40.6, pp. 1141–1159. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.1986.tb05740.x.

— (1991). “The genetic structure of the hybrid zone between the fire-bellied toads Bombina bombina and B. variegata: comparisons between transects and between loci”. In: *Evolution* 45.2, pp. 237–261. ISSN: 0014-3820. DOI: 10.1111/j.1558-5646.1991.tb04400.x.

Tajima, F. (1989). “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” In: *Genetics* 123.3, pp. 585–595. ISSN: 1943-2631. DOI: 10.1093/genetics/123.3.585.

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

Takayama, S. and A. Isogai (2005). “Self-incompatibility in plants”. In: *Annual Review of Plant Biology* 56.1, pp. 467–489. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev.arplant.56.032604.144249.

Takeuchi, H. and T. Higashiyama (2012). “A Species-Specific Cluster of Defensin-Like Genes Encodes Diffusible Pollen Tube Attractants in Arabidopsis”. In: *PLOS Biology* 10.12, e1001449. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001449.

Tavaré, S. et al. (1997). “Inferring Coalescence Times From DNA Sequence Data”. In: *Genetics* 145.2, pp. 505–518. ISSN: 1943-2631. DOI: 10.1093/genetics/145.2.505.

Tavares, M. M. et al. (2022). “Speciation with Gene Flow between Two Neotropical Sympatric Species (Pitcairnia Spp.: Bromeliaceae)”. In: *Ecology and Evolution* 12.5, e8834. ISSN: 2045-7758. DOI: 10.1002/ece3.8834.

Taylor, E. B. et al. (2005). “Speciation in Reverse: Morphological and Genetic Evidence of the Collapse of a Three-Spined Stickleback (Gasterosteus Aculeatus) Species Pair: collapse of a stickleback species pair”. In: *Molecular Ecology* 15.2, pp. 343–355. ISSN: 09621083. DOI: 10.1111/j.1365-294X.2005.02794.x.

Taylor, S. A. and E. L. Larson (2019). “Insights from Genomes into the Evolutionary Importance and Prevalence of Hybridization in Nature”. In: *Nature Ecology & Evolution* 3.2, pp. 170–177. ISSN: 2397-334X. DOI: 10.1038/s41559-018-0777-y.

Tiedemann, R. et al. (2004). “Mitochondrial DNA and Microsatellite Variation in the Eider Duck (Somateria Mollissima) Indicate Stepwise Postglacial Colonization of Europe and Limited Current Long-Distance Dispersal”. In: *Molecular Ecology* 13.6, pp. 1481–1494. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2004.02168.x.

Tiffin, P., S. Olson, and L. C. Moyle (2001). “Asymmetrical Crossing Barriers in Angiosperms”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1469, pp. 861–867. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2000.1578.

Tiffney, B. H. and S. J. Mazer (1995). “Angiosperm Growth Habit, Dispersal and Diversification Reconsidered”. In: *Evolutionary Ecology* 9.1, pp. 93–117. ISSN: 1573-8477. DOI: 10.1007/BF01237700.

Tsagkogeorga, G., V. Cahais, and N. Galtier (2012). “The Population Genomics of a Fast Evolver: High Levels of Diversity, Functional Constraint, and Molecular Adaptation in the Tunicate Ciona Intestinalis”. In: *Genome Biology and Evolution* 4.8, pp. 852–861. ISSN: 1759-6653. DOI: 10.1093/gbe/evs054.

Turelli, M. and L. C. Moyle (2007). “Asymmetric Postmating Isolation: Darwin’s Corollary to Haldane’s Rule”. In: *Genetics* 176.2, pp. 1059–1088. ISSN: 1943-2631. DOI: 10.1534/genetics.106.065979.

Turner, T. L. and M. W. Hahn (2010). “Genomic Islands *of* Speciation or Genomic Islands *and* Speciation?” In: *Molecular Ecology* 19.5, pp. 848–850. ISSN: 09621083, 1365294X. DOI: 10.1111/j.1365-294X.2010.04532.x.

Uebler, S., T. Dresselhaus, and M. Márton (2013). “Species-Specific Interaction of EA1 with the Maize Pollen Tube Apex”. In: *Plant Signaling & Behavior* 8.10, e25682. ISSN: null. DOI: 10.4161/psb.25682. pmid: 23887497.

Ungerer, M. C., C. R. Linder, and L. H. Rieseberg (2003). “Effects of Genetic Background on Response to Selection in Experimental Populations of Arabidopsis Thaliana”. In: *Genetics* 163.1, pp. 277–286. ISSN: 1943-2631. DOI: 10.1093/genetics/163.1.277.

Vamosi, J. C. and S. M. Vamosi (2010). “Key Innovations within a Geographical Context in Flowering Plants: Towards Resolving Darwin’s Abominable Mystery”. In: *Ecology Letters* 13.10, pp. 1270–1279. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2010.01521.x.

Van der Niet, T., R. Peakall, and S. D. Johnson (2014). “Pollinator-Driven Ecological Speciation in Plants: New Evidence and Future Perspectives”. In: *Annals of Botany* 113.2, pp. 199–212. ISSN: 0305-7364. DOI: 10.1093/aob/mct290.

Van der Niet, T. and S. D. Johnson (2012). “Phylogenetic Evidence for Pollinator-Driven Diversification of Angiosperms”. In: *Trends in Ecology & Evolution* 27.6, pp. 353–361. ISSN: 01695347. DOI: 10.1016/j.tree.2012.02.002.

Van Damme, J. M. M. (1983). “Gynodioecy in Plantago Lanceolata L. II Inheritance of Three Male Sterility Types”. In: *Heredity* 50.3, pp. 253–273. ISSN: 0018-067X, 1365-2540. DOI: 10.1038/hdy.1983.28.

Via, S. (2012). “Divergence Hitchhiking and the Spread of Genomic Isolation during Ecological Speciation-with-Gene-Flow”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences*. DOI: 10.1098/rstb.2011.0260.

Virdee, S. R. and G. M. Hewitt (1994). “Clines for hybrid dysfunction in a grasshopper hybrid zone”. In: *Evolution* 48.2, pp. 392–407. ISSN: 00143820. DOI: 10.1111/j.1558-5646.1994.tb01319.x.

Wagner, F. et al. (2020). “Taming the Red Bastards: Hybridisation and Species Delimitation in the Rhodanthemum Arundanum-Group (Compositae, Anthemideae)”. In: *Molecular Phylogenetics and Evolution* 144, p. 106702. ISSN: 1055-7903. DOI: 10.1016/j.ympev.2019.106702.

Wang, J., Y. A. EL-KASSABY, and K. Ritland (2012). “Estimating Selfing Rates from Reconstructed Pedigrees Using Multilocus Genotype Data”. In: *Molecular ecology* 21.1, pp. 100–116. ISSN: 0962-1083.

Weeks, S. C. (2012). “The Role of Androdioecy and Gynodioecy in Mediating Evolutionary Transitions between Dioecy and Hermaphroditism in the Animalia: Reproductive Transitions in the Animalia”. In: *Evolution* 66.12, pp. 3670–3686. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2012.01714.x.

Welch, J. J. (2004). “Accumulating Dobzhansky-Muller Incompatibilities: Reconciling Theory and Data”. In: *Evolution* 58.6, pp. 1145–1156. ISSN: 1558-5646.

Wendel, J. F. (2015). “The Wondrous Cycles of Polyploidy in Plants”. In: *American Journal of Botany* 102.11, pp. 1753–1756. ISSN: 1537-2197. DOI: 10.3732/ajb.1500320.

Wendt, T. et al. (2002). “Selfing Facilitates Reproductive Isolation among Three Sympatric Species of Pitcairnia (Bromeliaceae)”. In: *Plant Systematics and Evolution* 232.3-4, pp. 201–212. ISSN: 0378-2697, 1615-6110. DOI: 10.1007/s006060200043.

Westram, A. M. et al. (2022). “What Is Reproductive Isolation?” In: *Journal of Evolutionary Biology* 35.9, pp. 1143–1164. ISSN: 1420-9101. DOI: 10.1111/jeb.14005.

White, B. J. et al. (2010). “Genetic Association of Physically Unlinked Islands of Genomic Divergence in Incipient Species of *Anopheles Gambiae*”. In: *Molecular Ecology* 19.5, pp. 925–939. ISSN: 09621083, 1365294X. DOI: 10.1111/j.1365-294X.2010.04531.x.

Whitehead, M. R. et al. (2018). “Plant Mating Systems Often Vary Widely Among Populations”. In: *Frontiers in Ecology and Evolution* 6, p. 38. ISSN: 2296-701X. DOI: 10.3389/fevo.2018.00038.

Whitney, K. D. et al. (2010). “Patterns of Hybridization in Plants”. In: *Perspectives in Plant Ecology, Evolution and Systematics* 12.3, pp. 175–182. ISSN: 1433-8319. DOI: 10.1016/j.ppees.2010.02.002.

Williams, E. G. et al. (1986). “Overgrowth of Pollen Tubes in Embryo Sacs of Rhododendron Following Interspecific Pollinations”. In: *Australian Journal of Botany* 34.4, pp. 413–423. ISSN: 1444-9862. DOI: 10.1071/bt9860413.

Wood, D. P. et al. (2018). "Contrasting Phylogeographic Structures between Freshwater Lycopods and Angiosperms in the British Isles". In: *Botany Letters* 165.3-4, pp. 476–486. ISSN: 2381-8107. DOI: 10.1080/23818107.2018.1505545.

Wood, H. M. et al. (2008). "Sequence Differentiation in Regions Identified by a Genome Scan for Local Adaptation". In: *Molecular Ecology* 17.13, pp. 3123–3135. ISSN: 09621083, 1365294X. DOI: 10.1111/j.1365-294X.2008.03755.x.

Wood, T. E. et al. (2009). "The Frequency of Polyploid Speciation in Vascular Plants". In: *Proceedings of the National Academy of Sciences* 106.33, pp. 13875–13879. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0811575106.

Wright, S. (1984). *Evolution and the Genetics of Populations, Volume 2: Theory of Gene Frequencies*. Vol. 2. University of Chicago press. ISBN: 0-226-91039-3.

Wright, S. I., S. Kalisz, and T. Slotte (2013). "Evolutionary Consequences of Self-Fertilization in Plants". In: *Proceedings of the Royal Society B: Biological Sciences* 280.1760. ISSN: 0962-8452. DOI: 10.1098/rspb.2013.0133. pmid: 23595268.

Wu, C.-I. (2001). "The Genic View of the Process of Speciation". In: *Journal of Evolutionary Biology* 14.6, pp. 851–865. ISSN: 1420-9101. DOI: 10.1046/j.1420-9101.2001.00335.x.

Xiong, T. and J. Mallet (2022). "On the Impermanence of Species: The Collapse of Genetic Incompatibilities in Hybridizing Populations". In: *Evolution* 76.11, pp. 2498–2512. ISSN: 1558-5646. DOI: 10.1111/evo.14626.

Xu, S., P. M. Schlüter, and F. P. Schiestl (2012). "Pollinator-Driven Speciation in Sexually Deceptive Orchids". In: *International Journal of Ecology* 2012, e285081. ISSN: 1687-9708. DOI: 10.1155/2012/285081.

Yan, D. et al. (2014). "The functions of the endosperm during seed germination". In: *Plant and Cell Physiology* 55.9, pp. 1521–1533.

Zheng, Y. and A. Janke (2018). "Gene Flow Analysis Method, the D-statistic, Is Robust in a Wide Parameter Space". In: *BMC Bioinformatics* 19.1, p. 10. ISSN: 1471-2105. DOI: 10.1186/s12859-017-2002-4.

Acknowledgments

Je tiens à remercier de mon mieux toutes les personnes et institutions qui ont contribué, de près ou de loin, à la production de ma thèse. *The parts in English of those acknowledgments are intended for the Ghent side of my joint thesis. I apologise for not venturing into Dutch.*

À ceux qui m'ont financé / *To those who funded me*

Tout d'abord, je remercie grandement la fondation I-SITE ULNE d'avoir financé ma thèse au travers de l'appel à projets "*AAP co-tutelles de thèse avec l'Université de Gand 2020*". Je remercie aussi l'Université de Lille ainsi que l'école doctorale SMRE pour leur soutien. *Thanks to the Universiteit Gent for this joint thesis and for their support.* Je remercie aussi l'équipe Évolution et Écologie d'avoir financé mes missions et les derniers mois de ma thèse. Sans le financement et le soutien de tous ces acteurs, je n'aurais pas pu faire cette thèse dans de bonnes conditions et je leur en suis très reconnaissant / *Without the funding and support of all these actors, I wouldn't have been able to do this thesis, and I'm very grateful to them.*

À mes encadrants de thèse / *To my thesis supervisors*

I would like to thank you, Yves Van de Peer, for co-supervise my thesis and for welcoming me to your laboratory. This thesis would not have been possible without your participation. Despite the challenges posed by COVID-19, which impacted my stay in Ghent, I am truly grateful for your valuable feedback and suggestions on my work.

Xavier, merci d'avoir été présent tout au long de ma thèse malgré un emploi du temps de ~~ministre~~ directeur de laboratoire. J'ai pu compter sur ton encadrement bienveillant, et j'ai apprécié ton recul sur ce projet et tes remarques constructives tout autant que tes blagues et touches de bonne humeur. Ma thèse aurait été moins complète et moins agréable sans ton encadrement, et je te remercie d'avoir accepté d'en faire partie.

Camille, je te remercie tout particulièrement de m'avoir proposé de tenter cette aventure (alors que nos échanges se limitait à l'époque à quelques emails et un ou deux zooms), de m'avoir confié ce projet qui découle de ton travail et de m'avoir accompagné durant toute ma thèse. Merci pour tout ce que tu m'as transmis, pour toutes nos discussions, pour tes patientes explications, et de m'avoir ~~forcé~~ encouragé à apprendre de nouveaux outils (tel que LaTeX[7]...). Enfin, et surtout, merci d'avoir ponctué ces trois années de thèse avec moultes blagues et débats houleux[8]. Merci beaucoup, Camille.

À mon jury de thèse / *To my thesis jury*

Je remercie l'ensemble des membres de mon jury d'avoir accepté de participer à ma soutenance de thèse. En particulier, Tatiana Giraud et Maud Tenaillon, merci d'avoir accepté d'être rapportrices de mon manuscrit de thèse ! Merci Myriam Valero d'avoir accepté de présider ce jury, et merci Nicolas Bierne d'avoir accepté d'être membre de ce jury. *Thank you Zhen Li, for agreeing to be part of my thesis jury despite being organised so late.*

À ceux qui m'ont fourni ressources et aide

Mon travail de thèse a été grandement facilité par tous ceux qui m'ont partagé leurs ressources. Merci,

Zoé Postel et Pascal Touzet d'avoir bien voulu me partager vos données Silene.

[7] *insert Elden Ring boss theme.

[8] Hallyday *versus* Brassens, escalade *versus* badminton, pluviométrie/latitude et ce fameux point de rosée...

Sylvain Glémin d'avoir pris le temps de me partager ses connaissances sur les espèces de mon jeu de données.

à l'IFB de m'avoir permis d'utiliser leurs clusters, ainsi qu'au plateau bioinformatique de l'équipe Évolution et Écologie (Mathieu Genete et Clément Mazoyer) pour leur aide.

à tous les chercheurs qui ont pris la peine de rendre leurs données librement accessibles. Je n'aurais jamais pu constituer mon jeu de données sans ces efforts. Vive la science ouverte.

Aux services administratifs

Merci aux personnes qui m'ont aidé à naviguer dans l'administratif, un univers pour moi bien plus mystérieux que la biologie évolutive. . .
Merci / *thanks*,

à Christophe Van Brussel, pour sa bienveillance et sa patience qui facilite grandement les démarches administratives des doctorants.

to Sophie Maebe, for all her work in getting me through the Ghent's administrative procedure.

à Sandrine et Séverine, qui m'ont accompagné dans ma lutte face à l'administratif francais à chacune de mes missions.

Aux grands

Je suis reconnaissant d'avoir pu faire ma thèse au sein de l'équipe EE. Pouvoir affronter trois ans de stress dans un environnement aussi accueillant est une chance !
Merci,

à Sylvain pour les enrichissantes discussions scientifiques, statistiques et philosophiques.

à Anne et Nina, d'avoir rendu mes services d'enseignements particulièrement agréables.

à Pierre, pour tous les services qu'il a assuré au sein du laboratoire. Je demeure convaincu que le bâtiment de l'équipe EE n'aurait pas survécu à plus d'une semaine de son absence...

à Christelle, Cécile, Laurence et Anne-Cat', pour leur bienveillance et bonne humeur précieuse.

à toutes ces personnes ainsi qu'à celles que je n'ai pas citées. À mes yeux, vous avez fait honneur à l'emblème tout en courbe de notre Université, qui rappel l'hospitalité chaleureuse du Nord de la France.

Aux jeunes

Ma thèse n'aurait définitivement pas été la même sans vous. Je suis heureux d'avoir pu traverser ces trois années entouré d'ami.e.s plutôt que de simples collègues. Je n'imagine pas ce qu'aurait été ma thèse sans tous ces restaurants, ces séances de bloc ou ces sorties culturelles ou sportives. Merci au bureau 206 de m'avoir fait me sentir chez moi au travail. Merci au bureau 222 d'avoir organisé la moitié de nos activités, vous méritez **amplement** votre titre de '2ème bureau le plus cool de l'étage'. Merci à celle qui m'a accompagné durant cet été solitaire. Merci à ceux qui m'ont accompagné à chaque séance de bloc. Merci pour toutes ces pintes partagées, à la MDE du mardi soir ou à Wazemme. Merci de m'avoir fait découvrire le carnaval le plus accueillant du monde. Et merci pour tous les autres moments de bonheur que j'ai passés avec vous et dont je ne parle pas pour éviter d'ajouter quelques dizaines de pages supplémentaires à ce manuscrit. Les copain.ine.s, je ne vous cite pas ici, mais sachez que vous faites partie intégrante de ma thèse ;) ;) ;).

Un merci tout particulier à Flavia, d'avoir été ma complice de bureau depuis le premier jour, d'avoir été cette étincelle de malice qui m'a communiqué tant de joie, et pour toutes ces déclarations, gênantes mais si drôle, que la décence m'interdit de rapporter dans ces pages. Et un merci tout aussi particulier à Émilie, d'avoir été mon *safe space* durant ces trois ans, pour toutes nos discussions qui m'ont enrichi bien au-delà de cette thèse, pour tous ces restaurants et concerts, pour ces gifs réconfortants d'animaux mignons (chats non inclus), et bien entendu de m'avoir aidé à faire mes mathématiques !

Aux étudiants et enseignants qui m'ont accompagné jusque-là.

Merci à Sébastien Ibanez qui m'a convaincu, par son enseignement et par son conseil, de poursuivre mes études vers la biologie évolutive. Et merci aux Disciples, ami.e.s de licence, qui m'ont accompagné dans ces études de plus ou moins près ;)

Je remercie aussi mes DARWIIIIIINs, d'avoir été de super camarades tout au long de mon master, et d'avoir pris soin de ne pas laisser mourir cette amitié à la fin de notre cursus. Je ne suis pas sûr que j'aurais pu finir ce master sans cette chouette ambiance. Je remercie aussi les chercheurs qui m'ont encadré pendant mes deux stages, ces expériences m'ont poussées à continuer en thèse. Merci à Mathilde Dufaÿ ainsi qu'à Patrice David pour cette expérience au CEFE, et merci à Jonathan Romiguier ainsi qu'à Arthur Weyna pour cette expérience à l'ISEM. Merci aussi à Emmanuel Douzery, et à l'équipe pédagogique du master, pour ce super enseignement :)

Merci à la colloc' montpelliéraine, qui m'a accueillie de nombreux mois et qui m'a permis de compenser l'isolation (ou l'isolement, on peut dire les deux) de ma forteresse de solitude nordique par une vie en communauté pendant une bonne partie de ma thèse. Merci pour la compagnie, pour les rires, pour les jeux, pour les pizzas, pour vos expressions revisitées dont je suis si friand, pour les sorties naturalistes... Ces remerciements ne seraient pas complets si vous ne faisiez pas partie de ce document ;) ;) ;).

Merci Mathilde d'avoir été là pour moi depuis le master, d'avoir été l'Atlas de mon monde de manque de confiance en moi, pour toutes nos discussions sur la spéciation et la délimitation d'espèce (et bout à bout, ça doit représenter un petit paquet d'heures...), pour toutes ces touches de réconfort et de rire, et d'être encore là.

À ma famille

Merci à la DBC, d'être mon roc depuis maintenant la plus grande partie de ma vie. Yoann, Romain, Suzie, Manon, Matthieu, Benoit, Sébastien, vous êtes les 7 doigts de ma main, et je vous remercie d'incarner si bien l'expression "les ami.e.s, c'est la famille qu'on choisit".

Finalement, ma plus grande reconnaissance va à mes parents, pour être les premiers financeurs de mes études (et de loin...) et pour tout l'amour avec lequel vous me portez depuis toutes ces années et sans lequel je ne serais pas aller bien loin dans mes études. Maman, Papa, merci.

Supplementary figures

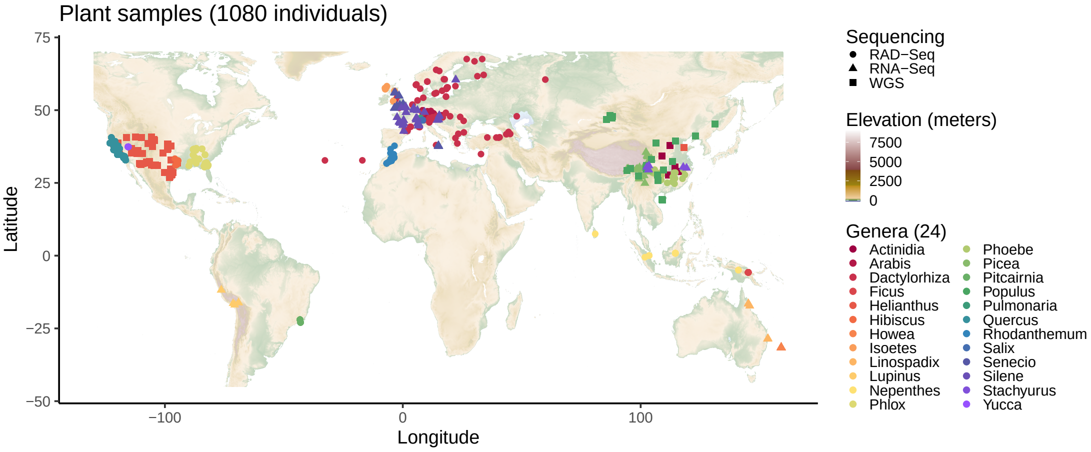


Figure S1: **Geographical location of plant samples and sequencing methods.**

The depicted symbol represents an entity for which geographic data was retrievable in the source publications, with the exception of the genus *Gossypium*. The varying shapes of the symbols serve to differentiate between distinct sequencing technologies.

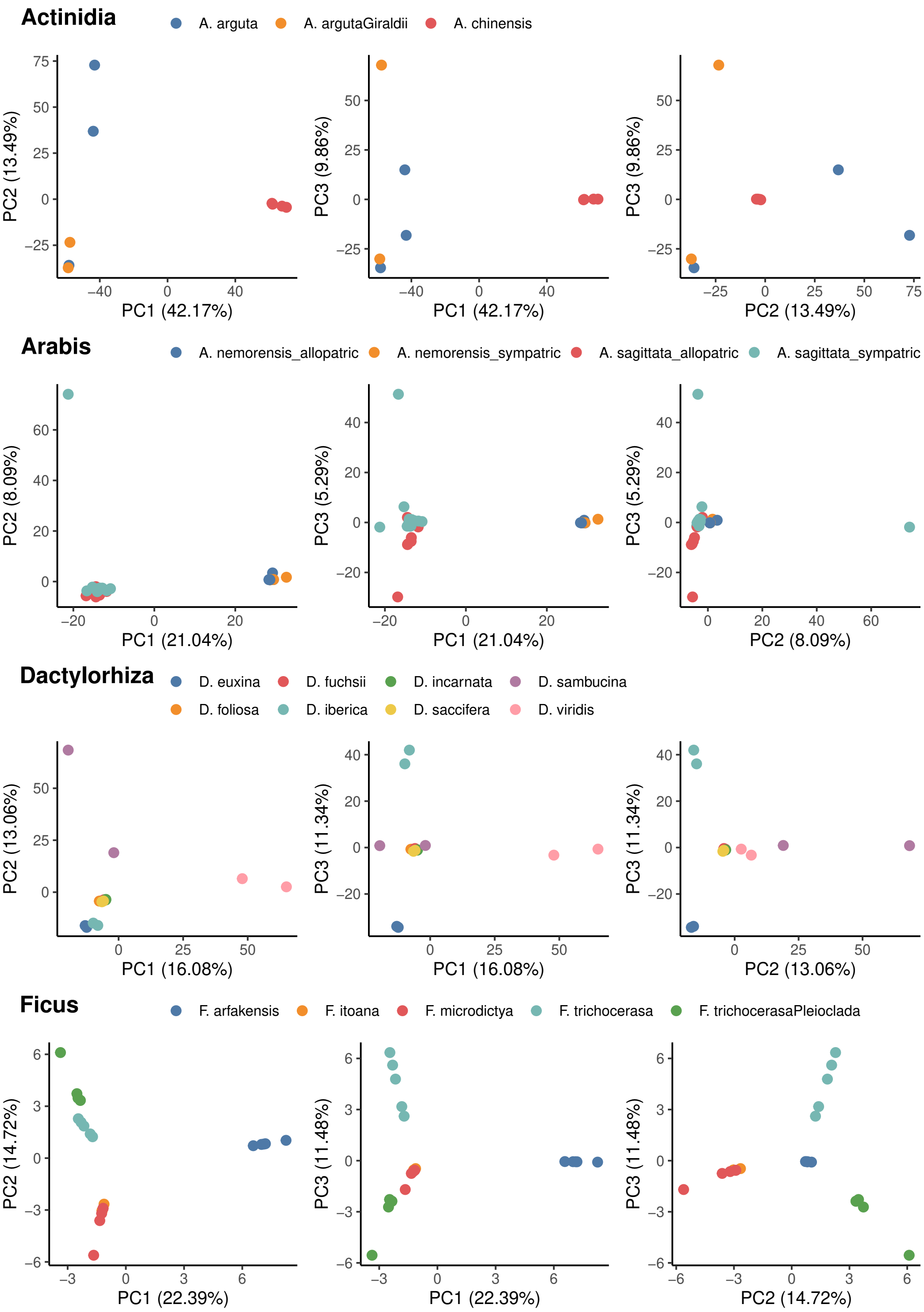


Figure S2: **Principal component analyses on genotypes for all SNPs.**

Each point represents an individual. The colours represent the different populations/species named by the authors of the studies from which the data originated.

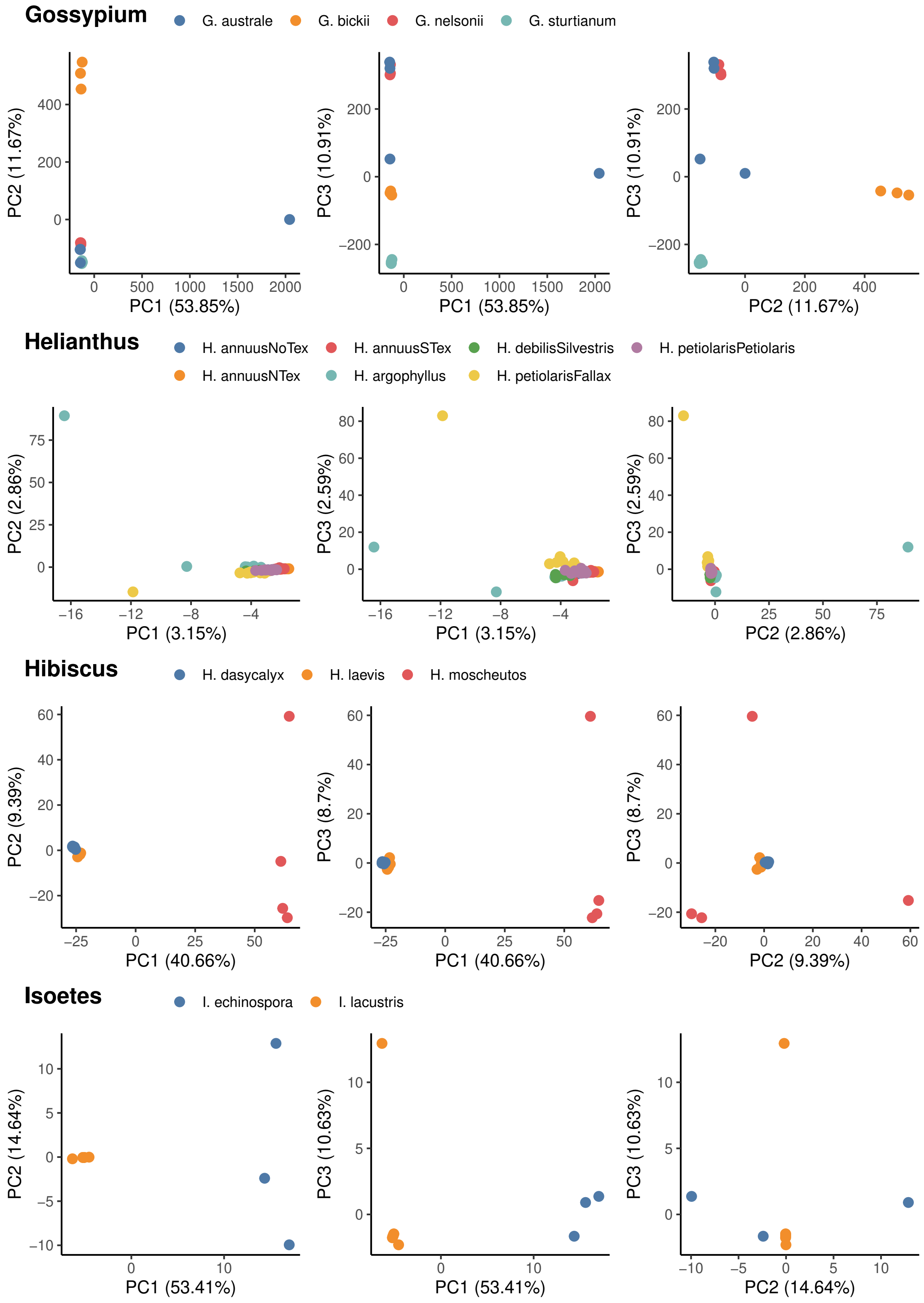


Figure S2: (continued).

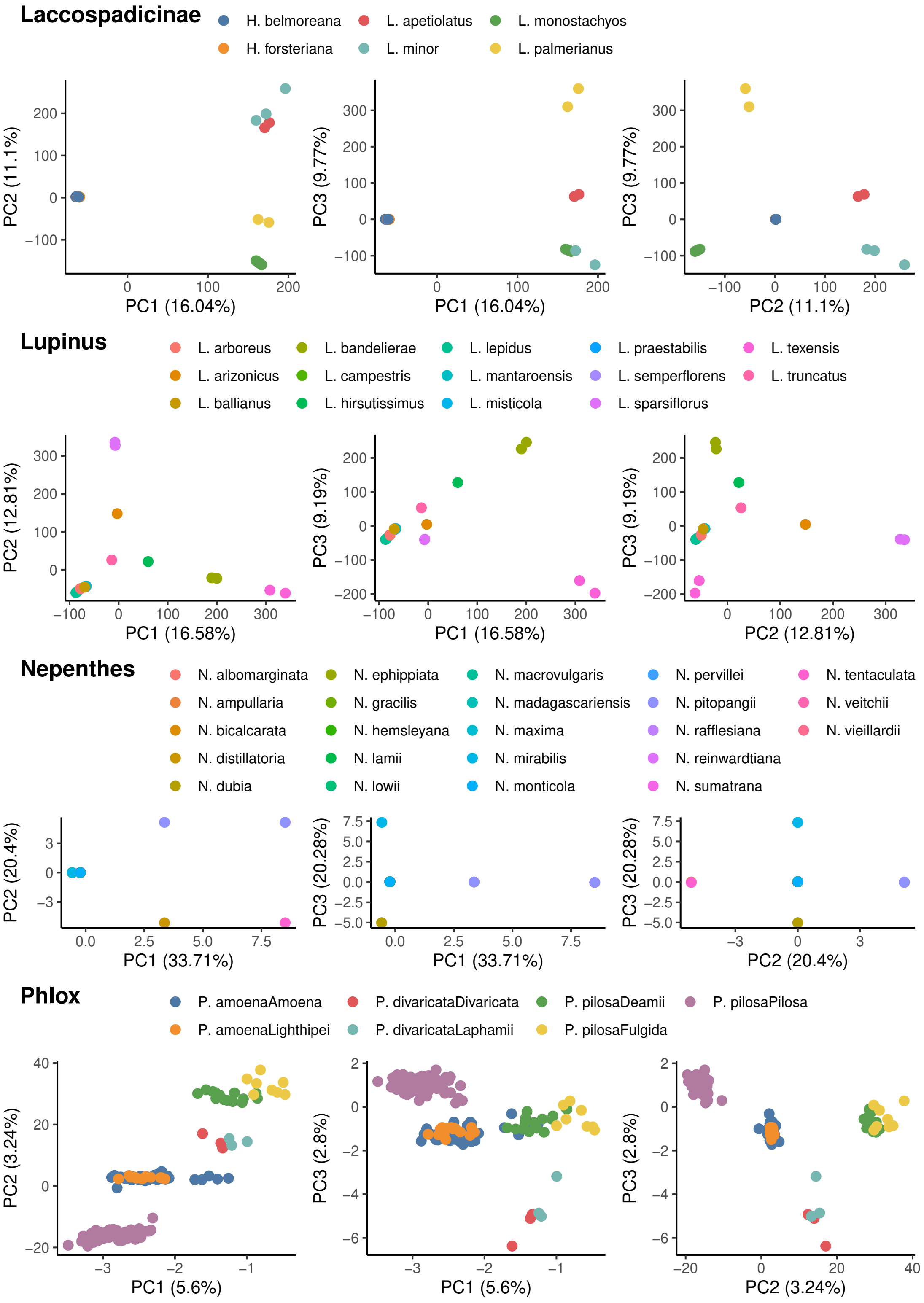
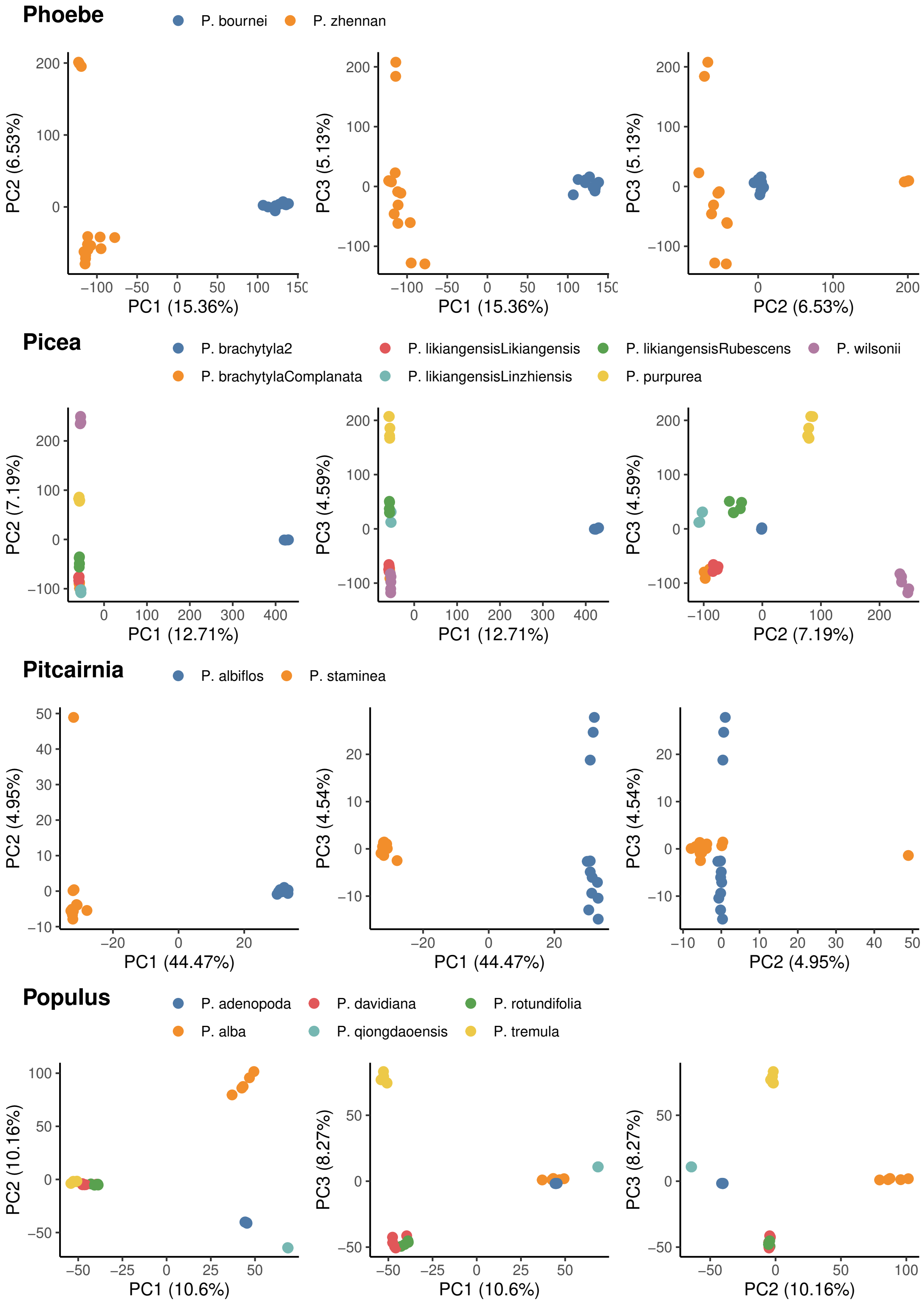
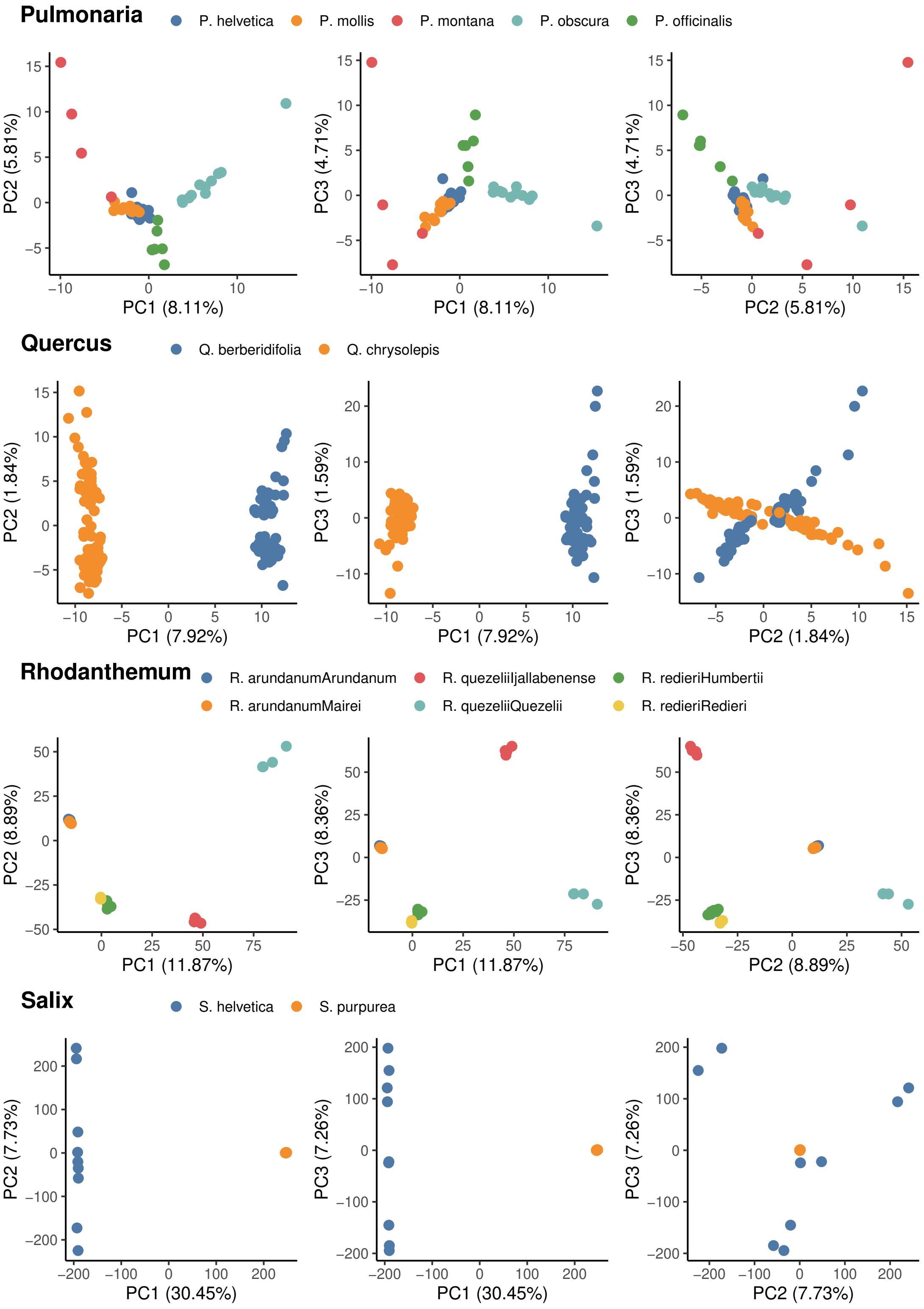


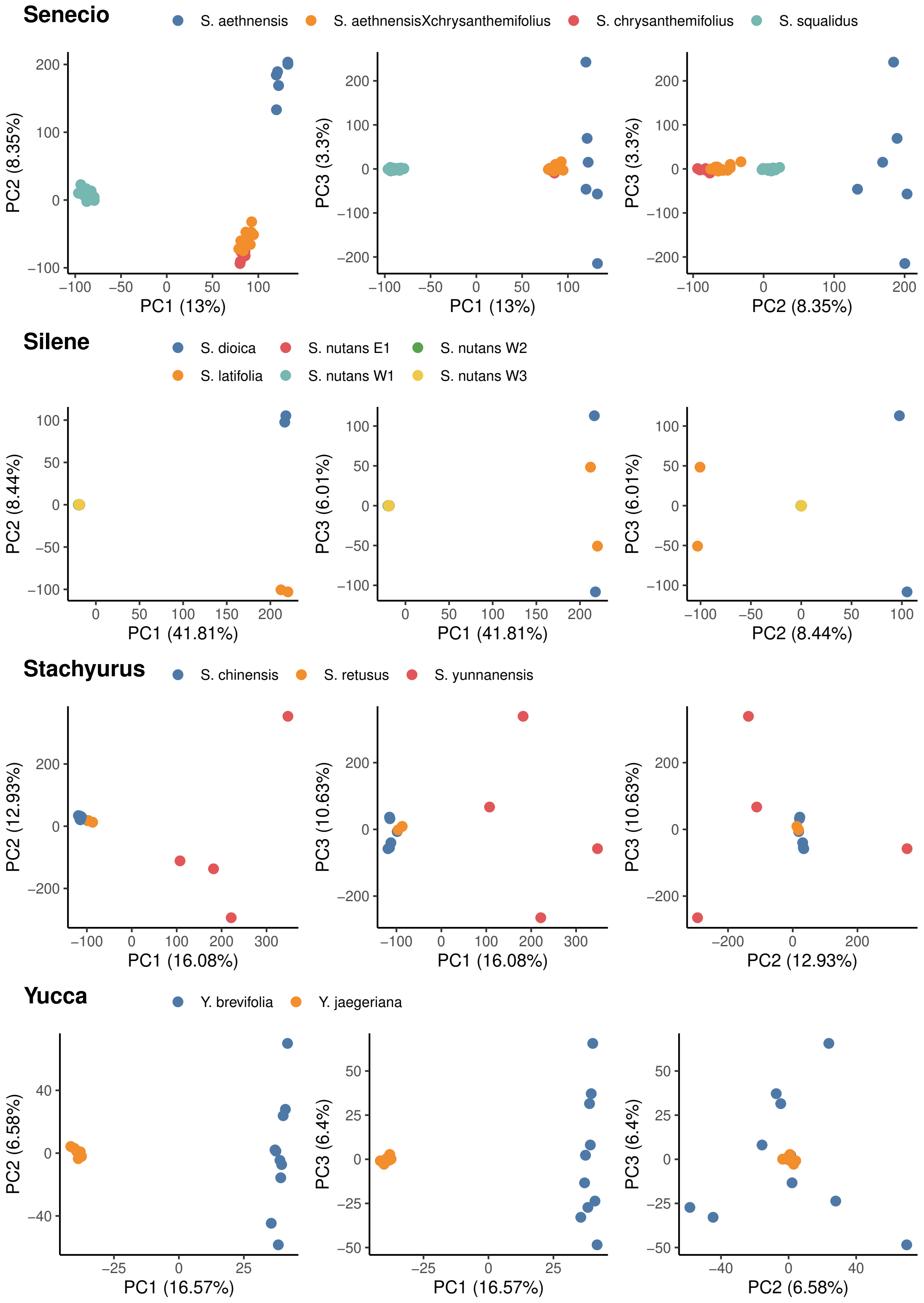
Figure S2: (continued).



(Figure S2: continued).



(Figure S2: continued).



(Figure S2: continued).

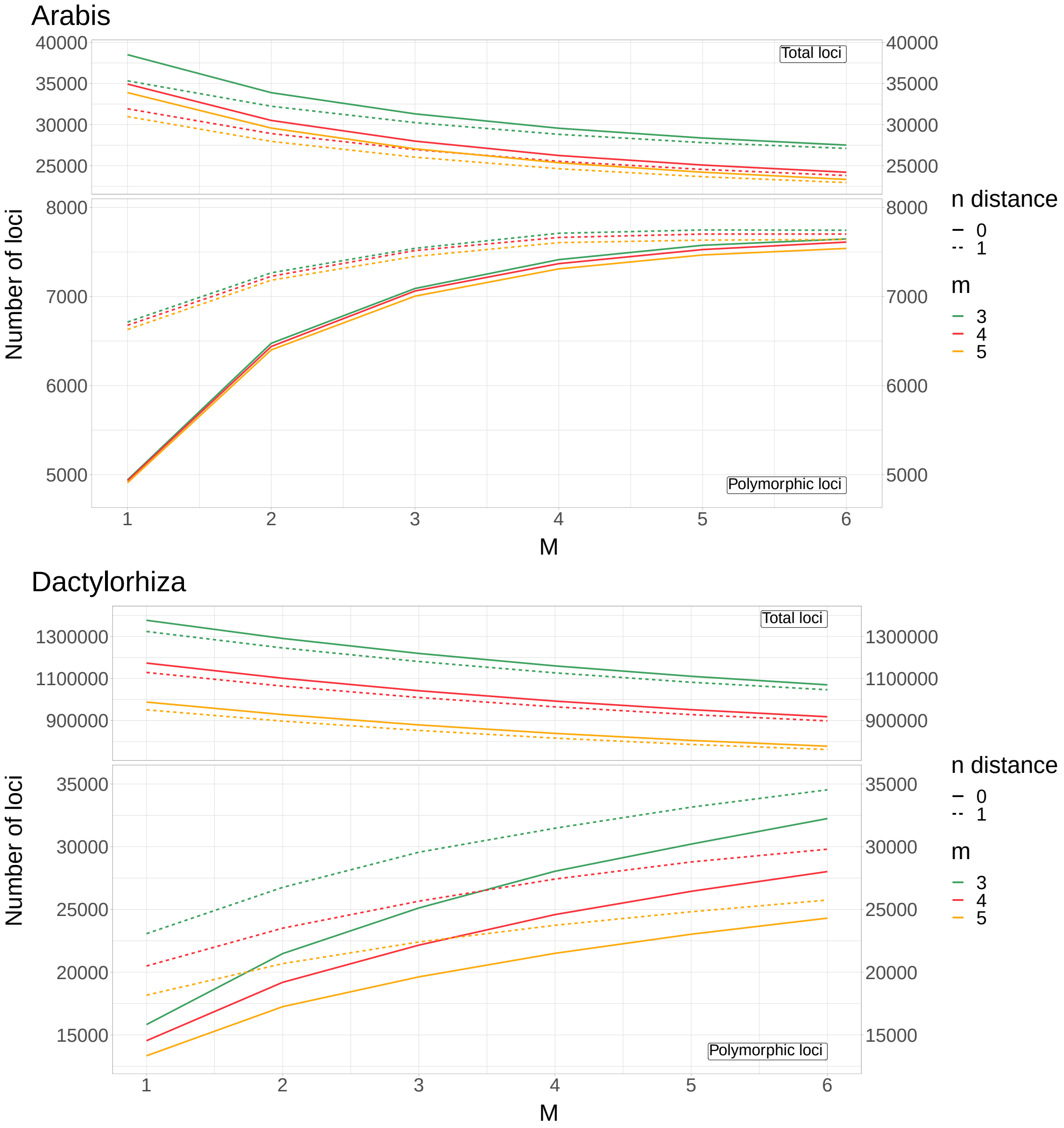


Figure S3: **Polymorphic loci in function of the *Stacks* parameters.**

Number of loci in function of the parameters provided to *denovo_map.pl* (*Stacks*). A figure is presented for each genus and each figure is divided into two parts. The upper part show the total number of loci assembled, and the lower part the number of *r80 poly loci*, both on the y-axis. The three parameters tested (*M*, *m* and *n)*, are respectively represented on the x-axis, by the color and by the type of line.
% r80 poly loci = proportion of polymorphic loci present in at least 80% of the population.
m = minimum number of raw reads required to form a putative allele.
M = number of mismatches allowed between putative alleles to merge them into a putative locus
n = number of mismatches allowed between putative loci to form the catalog.

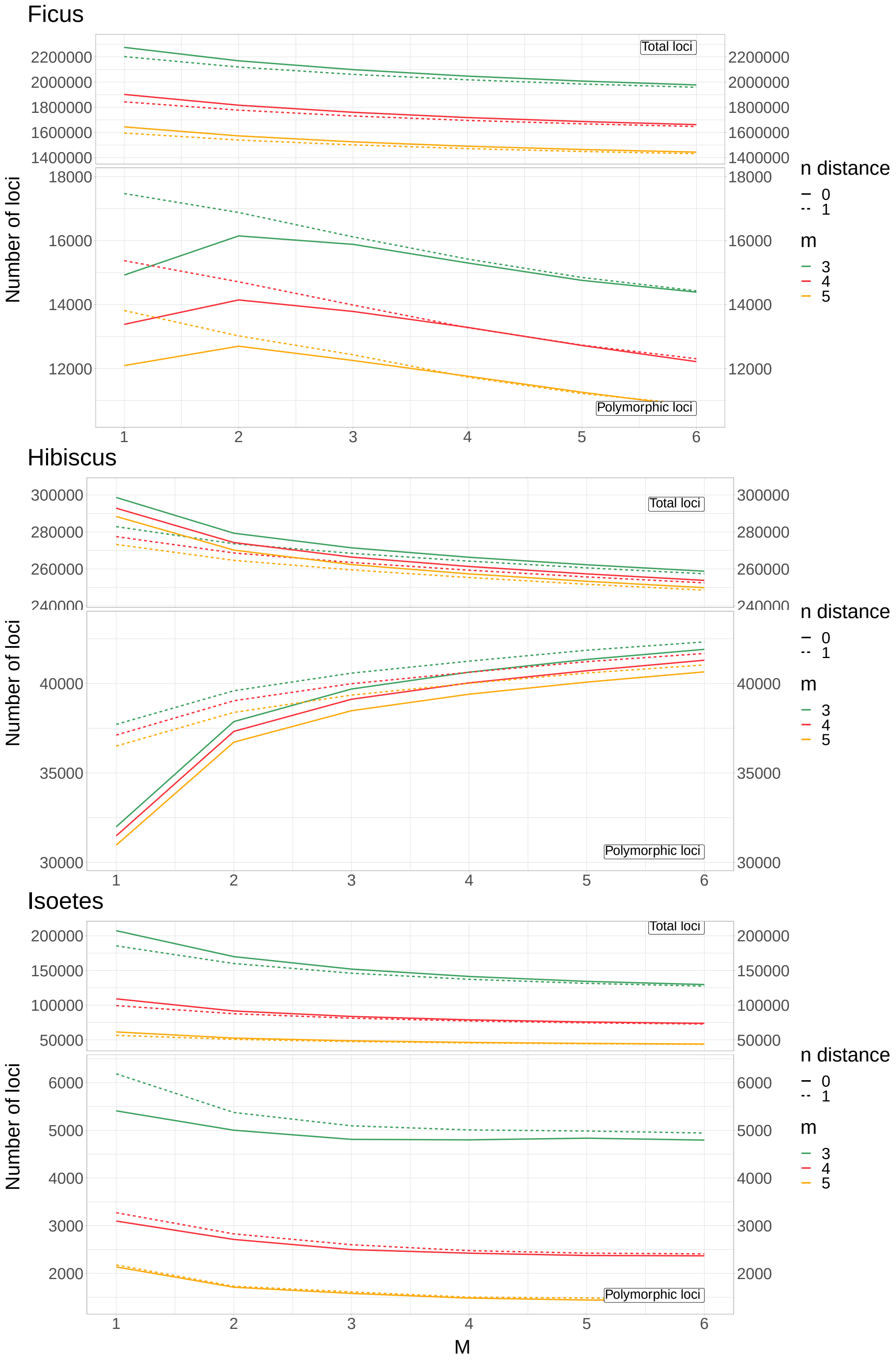


Figure S3: (continued).

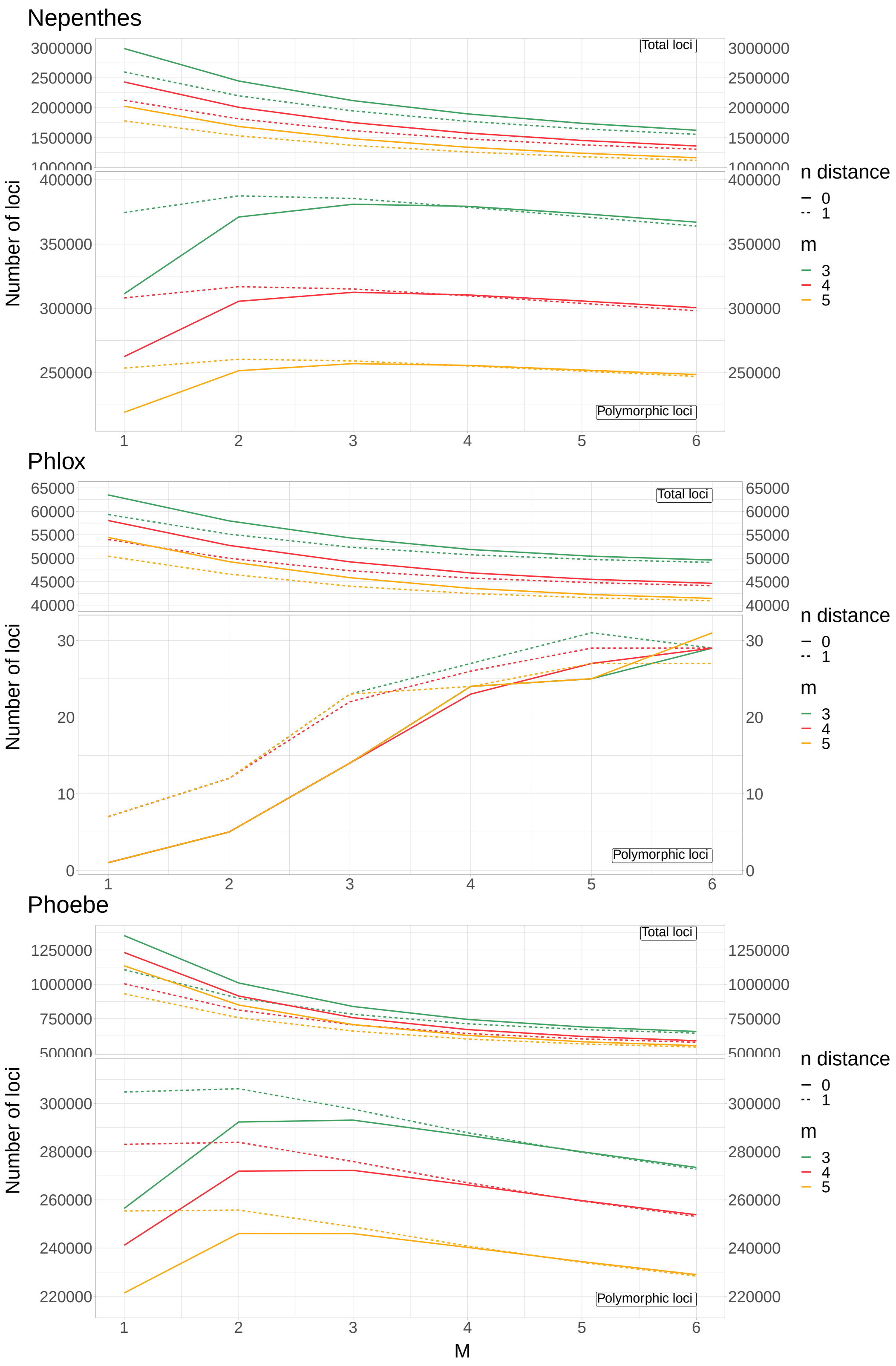


Figure S3: (continued).

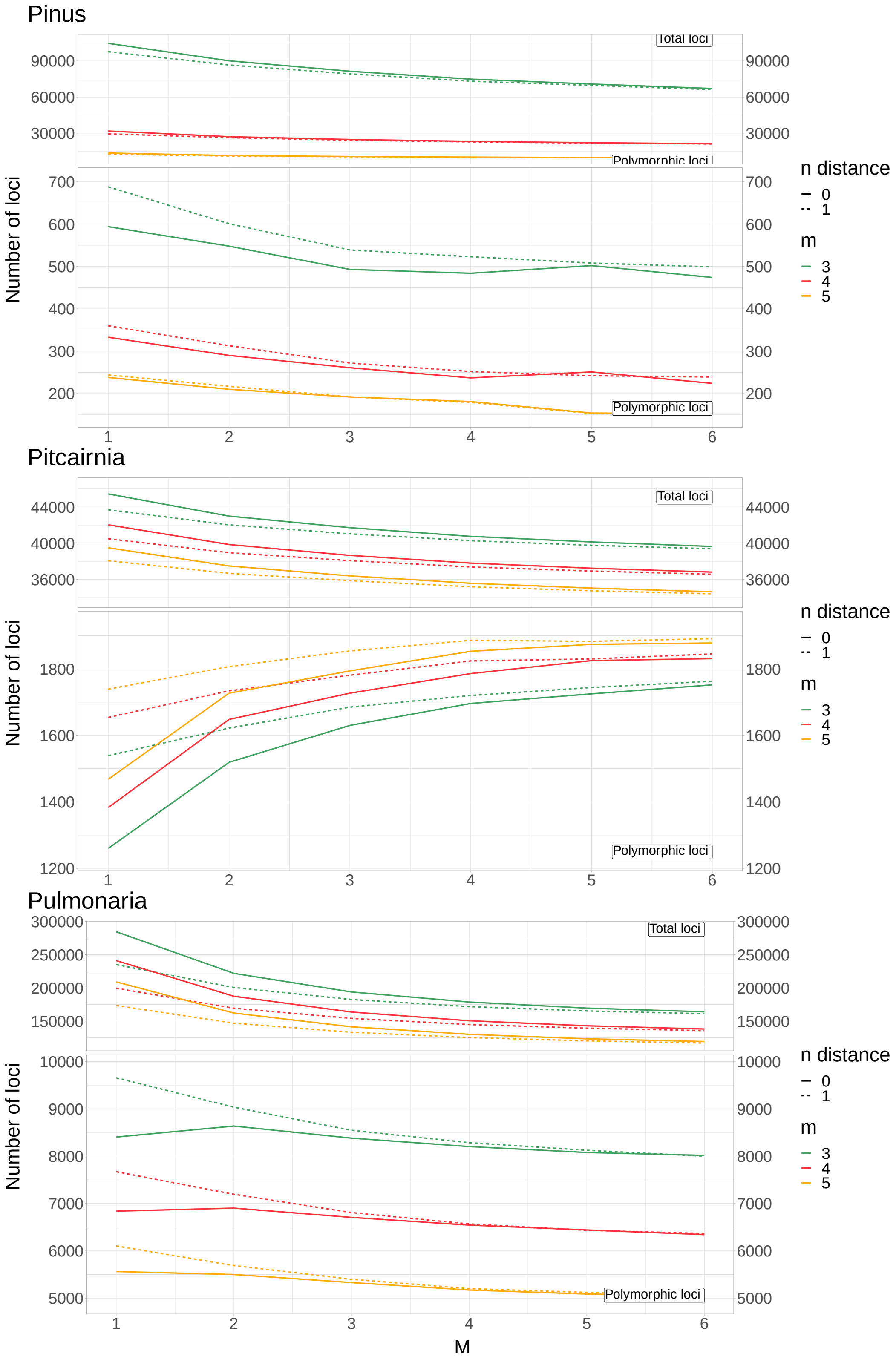
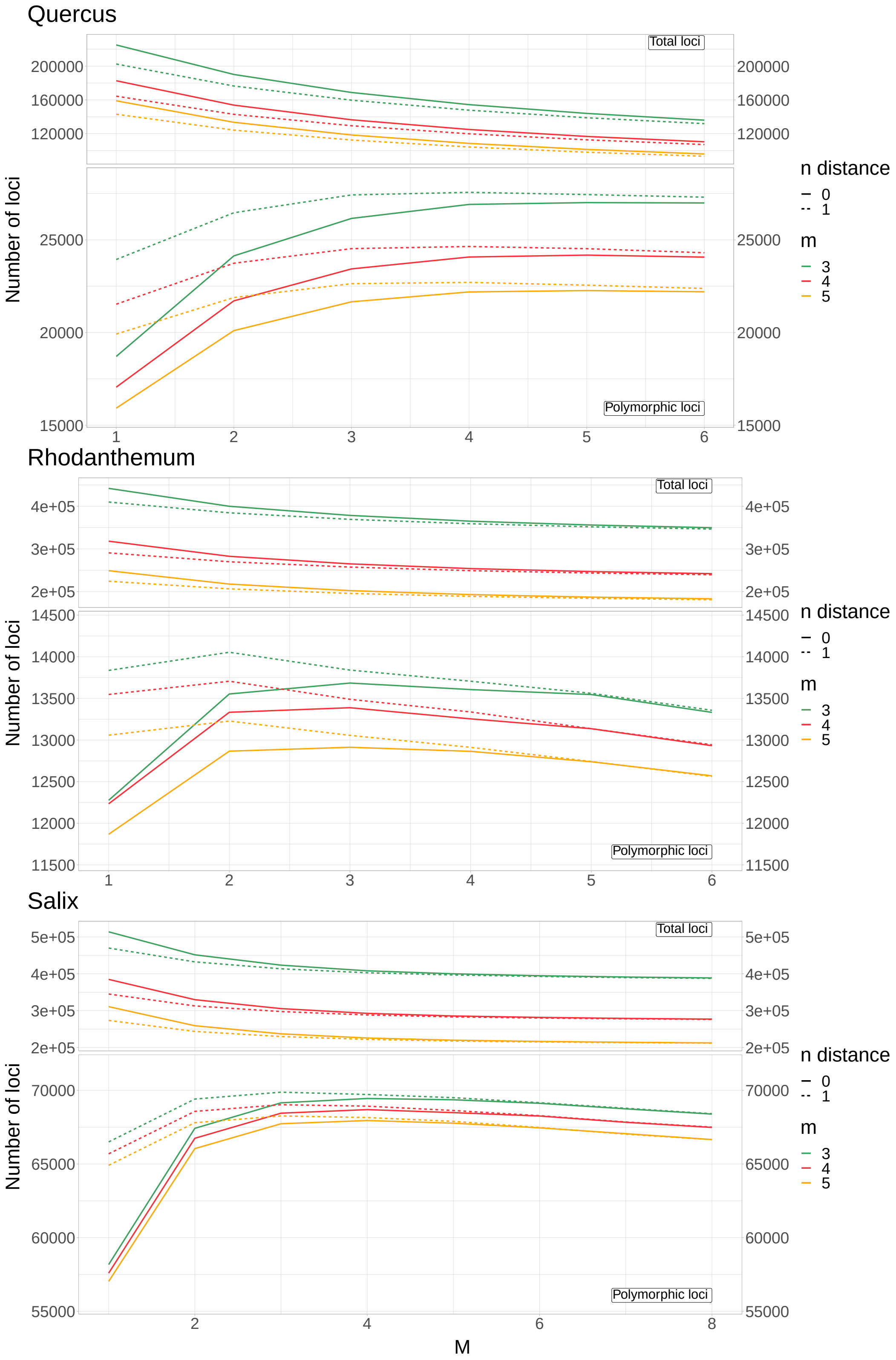


Figure S3: (continued).



Quercus

Rhodanthemum

Salix

Figure S3: (continued).

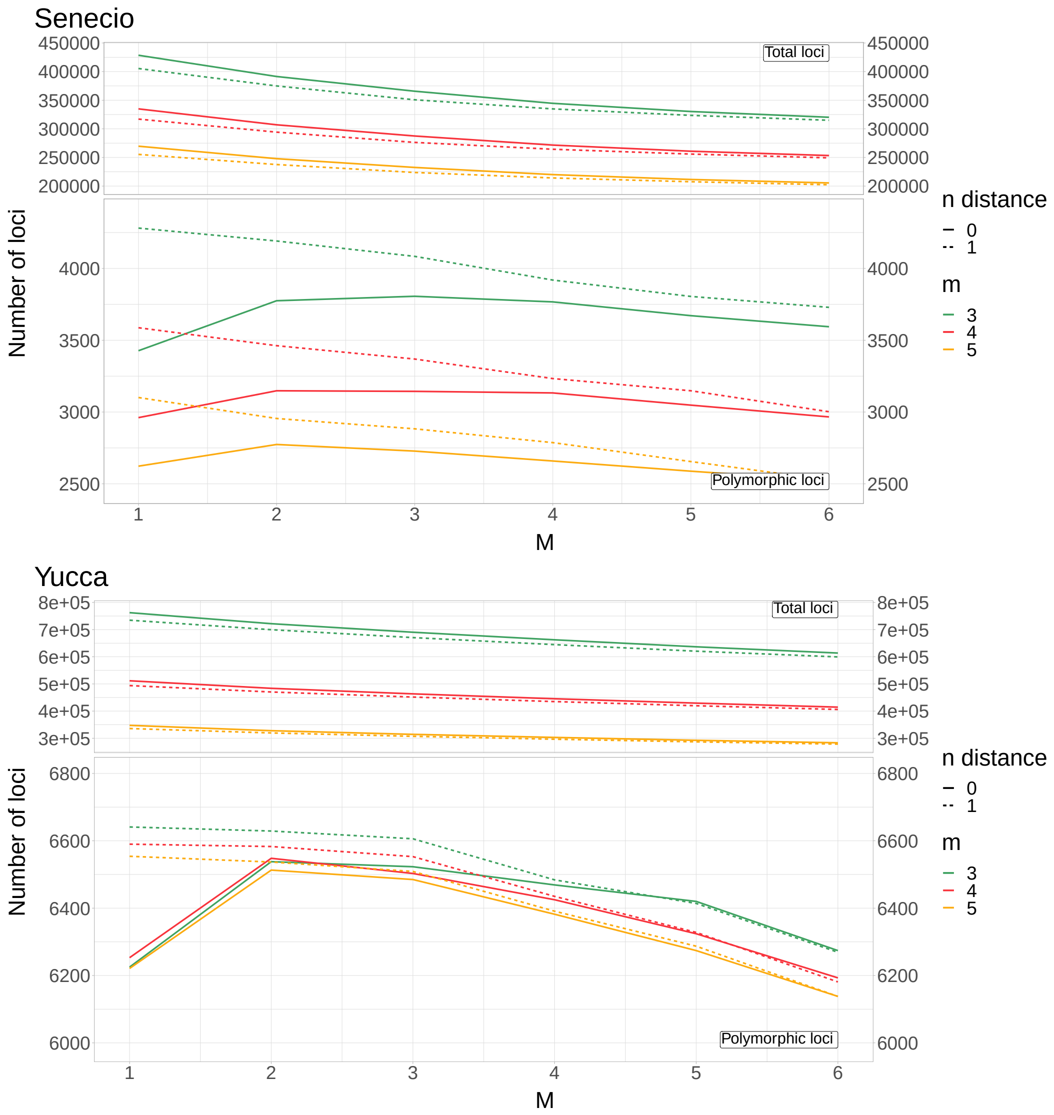


Figure S3: (continued).

Supplementary tables

Table S1: **List of retained NCBI datasets.**

bioproject	**genus**	**species**	**n**	**type of data**	**source**
PRJNA318567	*Actinidia*	*arguta*	3	WGS	Liu et al., 2017
		arguta giraldii	2		
		chinensis	4		
PRJEB33482,	*Arabis*	*nemorensis allop.*	6	RNA	Dittberner et al., 2022
PRJEB39992		*nemorensis symp.*	6		
		sagittata allop.	10		
		sagittata symp.	15		
PRJNA489792	*Dactylorhiza*	*euxina*	5	RAD	Brandrud et al., 2020
		foliosa	2		
		fuchsii	30		
		iberica	2		
		incarnata	31		
		saccifera	4		
		sambucina	3		
		viridis	3		
PRJNA445222	*Ficus*	*arfakensis*	14	RAD	Souto-Vilarós et al., 2018
		itoana	13		
		microdictya	15		
		trichocerasa	15		
		t. pleioclada	26		
PRJNA539957	*Gossypium*	*australe*	4	WGS	Grover et al., 2022
		bickii	3		
		nelsonii	3		
		robinsonii	2		
		sturtianum	6		
PRJNA532579	*Helianthus*	*annuus NoTex*	15	WGS	Owens et al., 2021
		annuus NTex	15		
		annuus STex	15		
		argophyllus	10		

		debilis silvestris	5		
		niveus canescens	8		
		petiolaris fallax	10		
		p. petiolaris	10		
PRJNA382435	*Hibiscus*	*dasycalyx*	6	RAD	Norrell, 2017
		laevis	4		
		moscheutos	5		
PRJNA483403	*Isoetes*	*lacustris*	9	RAD	D. P. Wood et al., 2018
		echiospora	3		
PRJNA244607	*Howea*	*belmoreana*	40	RNA	Dunning et al., 2016
	(*Laccospadix)*	*forsteriana*	39		
PRJNA528594	*Linospadix*	*monostachyos*	18		Osborne et al., 2019
	(*Laccospadix)*	*minor*	9		
		apetiolatus	6		
		palmerianus	6		
PRJNA318864	*Lupinus*	*ballianus*	2	RNA	Nevado et al., 2016
		bandelierae	2		
		misticola	2		
PRJEB37794	*Nepenthes*	*albomarginata*	3	RAD	Scharmann et al., 2021
		ampullaria	8		
		bicalcarata	6		
		distillatoria	2		
		dubia	2		
		ephippiata	2		
		gracilis	8		
		hemsleyana	4		
		lamii	2		
		lowii	2		
		macrovulgaris	2		
		madagascariensis	2		
		maxima	10		
		mirabilis	10		
		monticola	2		
		pervillei	16		
		pitopangii	2		
		rafflesiana	9		
		reinwardtiana	2		
		sumatrana	2		
		tentaculata	2		
		veitchii	3		
		vieillardii	2		
PRJNA701424	*Phlox*	*amoena amoena*	48	RAD	Goulet-Scott et al., 2021
		a. lighthipei	14		
		divaricata divaricata	3		

		d. laphamii	3		
		pilosa deamii	15		
		p. fulgida	8		
		p. pilosa	59		
		subulata	2		
PRJNA464259	*Phoebe*	*zhennan*	9	RAD	Ding et al., 2019
		bournei	12		
PRJNA807675	*Pitcairnia*	*albiflos*	9	RAD	Tavares et al., 2022
		staminea	12		
PRJNA392950,	*Picea*	*brachytyla*	4	RNA	Ru et al., 2018
PRJNA401149,		*b. complanata*	5		Sun et al., 2018
PRJNA378930,		*likiangensis likiangensis*	5		
PRJNA301093		*l. linzhiensis*	5		
		l. rubescens	5		
		purpurea	5		
		wilsoni	5		
PRJNA612655	*Populus*	*adenopoda*	5	WGS	Shang et al., 2020
		alba	5		
		davidiana	5		
		qiongdaoensis	3		
		rotundifolia	4		
		tremula	5		
PRJNA544114	*Pulmonaria*	*helvetica*	24	RAD	Grünig et al., 2021
		mollis	10		
		montana	4		
		obscur	11		
		officinalis	6		
PRJNA639507	*Quercus*	*berberidifolia*	63	RAD	Ortego and Knowles, 2020
		chrysolepis	80		
PRJNA554975	*Rhodanthemum*	*redieri redieri*	4	RAD	Wagner et al., 2020
		r. humbertii	7		
		quezelii quezelii	4		
		q. jallabenense	4		
		arundanum mairei	8		
		a. arundanum	27		
PRJNA429746	*Salix*	*helvetica*	10	RAD	Gramlich et al., 2018
		purpurea	10		
PRJNA549571	*Senecio*	*aethnensis*	6	RNA	Nevado et al., 2020
		aethn. X chrys.	14		
		chrysanthemifolius	6		
		squalidus	28		
PRJNA295359	*Silene*	*dioica*	2	RNA	Hu and Filatov, 2016
		latifolia	2		Muyle et al., 2021
		nutans E1	4		

		n. W1	4		
		n. W2	4		
		n. W3	4		
PRJNA553020	*Stachyurus*	*chinensis*	6	RNA	Feng et al., 2020
		retusus	2		
		yunnanensis	4		
PRJNA329381	*Yucca*	*brevifolia*	24	RAD	Royer et al., 2016
		jaegeriana	39		

Table S2: **List of sample accessions.**

genus	accessions
Actinidia	SRR3524809,SRR3471379,SRR3471380,SRR3531964,SRR3537231,SRR3537232,SRR3406787,SRR3407084,SRR3407085
Arabis	ERR4557693,ERR4557601,ERR4557355,ERR4557859,ERR4557861,ERR4557870,ERR4558014,ERR4558023,ERR4558453,ERR4557573,ERR4557889,ERR4558457,ERR4559969,ERR4557579,ERR4557587,ERR4557148,ERR4557616,ERR4557617,ERR4557659,ERR4557675,ERR4557694,ERR4557728,ERR4557360,ERR4557772,ERR4557786,ERR4557866,ERR4557389,ERR4557868,ERR4557871,ERR4558017,ERR4558019,ERR4558021,ERR4557394,ERR4558455,ERR4557864
Dactylorhiza	SRR7802056,SRR7802057,SRR7802058,SRR7802060,SRR7802061,SRR7802062,SRR7802063,SRR7802064,SRR7802065,SRR7802066,SRR7802067,SRR7802068,SRR7802069,SRR7802070,SRR7802071,SRR7802074,SRR7802075,SRR7802076,SRR7802077,SRR7802079,SRR7802080,SRR7802081,SRR7802082,SRR7802083,SRR7802092,SRR7802093,SRR7802094,SRR7802096,SRR7802097,SRR7802108,SRR7802111,SRR7802173,SRR7802175,SRR7802176,SRR7802177,SRR7802178,SRR7802179,SRR7802180,SRR7802183,SRR7802184,SRR7802193,SRR7802194,SRR7802195,SRR7802196,SRR7802197,SRR7802198,SRR7802199,SRR7802200,SRR7802201,SRR7802202,SRR7802203,SRR7802205,SRR7802206,SRR7802207,SRR7802208,SRR7802209,SRR7802210,SRR7802212,SRR7802213,SRR7802214,SRR7802217,SRR7802218,SRR7802219,SRR7802220,SRR7802229,SRR7802230,SRR7802243,SRR7802245,SRR7802256,SRR7802257,SRR7802259,SRR7802260,SRR7802261,SRR7802262,SRR7802263,SRR7802264,SRR7802265,SRR7802266,SRR7802267,SRR7802268
Ficus	SRR6910686,SRR6910691,SRR6910694,SRR6910697,SRR6910684,SRR6910687,SRR6910689,SRR6910690,SRR6910692,SRR6910693,SRR6910696,SRR6910705,SRR6910708,SRR6910710,SRR6910750,SRR6910752,SRR6910753,SRR6910704,SRR6910706,SRR6910707,SRR6910709,SRR6910711,SRR6910712,SRR6910713,SRR6910665,SRR6910666,SRR6910668,SRR6910670,SRR6910671,SRR6910673,SRR6910676,SRR6910679,SRR6910682,SRR6910695,SRR6910699,SRR6910700,SRR6910703,SRR6910738,SRR6910739,SRR6910741,SRR6910664,SRR6910667,SRR6910669,SRR6910672,SRR6910677,SRR6910678,SRR6910680,SRR6910681,SRR6910683,SRR6910698,SRR6910701,SRR6910702,SRR6910714,SRR6910723,SRR6910724,SRR6910726,SRR6910728,SRR6910729,SRR6910731,SRR6910733,SRR6910735,SRR6910743,SRR6910757,SRR6910759,SRR6910722,SRR6910725,SRR6910727,SRR6910730,SRR6910732,SRR6910734,SRR6910736,SRR6910737,SRR6910740,SRR6910742,SRR6910754,SRR6910755,SRR6910756,SRR6910758
Gossypium	SRR8979898,SRR8979899,SRR8979928,SRR8979929,SRR8979990,SRR8979991,SRR8979902,SRR8979903,SRR8979905,SRR8979904,SRR8979906,SRR8979907,SRR8979992,SRR8979993,SRR8979996,SRR8979997
Helianthus	SRR8892273,SRR8892299,SRR8892310,SRR8892312,SRR8892313,SRR8892315,SRR8892327,SRR8892338,SRR8892342,SRR8892355,SRR8892360,SRR8892372,SRR8892408,SRR8892447,SRR8892456,SRR8892458,SRR8892468,SRR8892469,SRR8892471,SRR8892473,SRR8892475,SRR8892487,SRR8892514,SRR8892518,SRR8892558,SRR8892591,SRR8892636,SRR8892639,SRR8892641,SRR8892653,SRR8892730,SRR8892732,SRR8892736,SRR8892769,SRR8892778,SRR8892785,SRR8892815,SRR8895820,SRR8895872,SRR8895895,SRR8895919,SRR8896003,SRR8896089,SRR8896151,SRR8896155,SRR8896245,SRR8896292,SRR8896317,SRR8896364,SRR8896381,SRR8896382,SRR8896386,SRR8896405,SRR8896414,SRR8896421,SRR8896469,SRR8896479,SRR8896481,SRR8896499,SRR8896504,SRR8896517,SRR8896534,SRR8896547,SRR8896590,SRR8896605,SRR8896626,SRR8896639,SRR8896683,SRR8896711,SRR8896737,SRR8896765,SRR8896794,SRR8896839,SRR8896844,SRR8888512,SRR8888560,SRR8888584,SRR8888625,SRR8888659,SRR8888703,SRR8888715,SRR8888734,SRR8888772,SRR8888812,SRR8888848,SRR8888926,SRR8888998,SRR8889030
Hibiscus	SRR6790655,SRR6790656,SRR6790658,SRR6790659,SRR6790660,SRR6790661,SRR6790662,SRR6790663,

	SRR6790664,SRR6790665,SRR6790666,SRR6790667,SRR6790668,SRR6790669,SRR6790670,SRR6790673
Isoetes	SRR7618766,SRR7618767,SRR7618764,SRR7618757,SRR7618763,SRR7618765,SRR7618769,SRR7618756,SRR7618762,SRR7618768,SRR7618770,SRR7618771
Howea & Linospadix (*Laccospadix*)	SRR1266982,SRR1266983,SRR1266984,SRR1266985,SRR1266986,SRR1266987,SRR1266988,SRR1266989,SRR1266990,SRR1266991,SRR1266992,SRR1266993,SRR1266994,SRR1266995,SRR1266996,SRR1266997,SRR1266998,SRR1266999,SRR1267000,SRR1267001,SRR1267002,SRR1267003,SRR1267004,SRR1267005,SRR1267006,SRR1267007,SRR1267008,SRR1267009,SRR1267010,SRR1267011,SRR1267012,SRR1267013,SRR1267014,SRR1267015,SRR1267016,SRR1267017,SRR1267018,SRR1267019,SRR1267020,SRR1267021,SRR1267022,SRR1267023,SRR1267024,SRR1267025,SRR1267026,SRR1267027,SRR1267028,SRR1267029,SRR1267030,SRR1267031,SRR1267032,SRR1267033,SRR1267034,SRR1267035,SRR1267036,SRR1267037,SRR1267038,SRR1267039,SRR1267040,SRR1267041,SRR1267042,SRR1267043,SRR1267044,SRR1267045,SRR1267046,SRR1267047,SRR1267048,SRR1267049,SRR1267050,SRR1267051,SRR1267052,SRR1267053,SRR1267054,SRR1267055,SRR1267056,SRR1267057,SRR1267058,SRR1267059,SRR1267060,SRR8772276,SRR8772277,SRR8772278,SRR8772279,SRR8772280,SRR8772281,SRR8772282,SRR8772283,SRR8772284,SRR8772285,SRR8772286,SRR8772287,SRR8772288,SRR8772289,SRR8772290,SRR8772291,SRR8772292,SRR8772293,SRR8772294,SRR8772297,SRR8772298,SRR8772299,SRR8772310,SRR8772311,SRR8772312,SRR8772313,SRR8772314,SRR8772315,SRR8772316,SRR8772317,SRR8772318,SRR8772319,SRR8772320,SRR8772322,SRR8772323,SRR8772324,SRR8772325,SRR8772326,SRR8772327
Lupinus	SRR3422973,SRR3423047,SRR3422976,SRR3423048,SRR3422972,SRR3423008
Nepenthes	ERR4027975,ERR4027976,ERR4027982,ERR4027983,ERR4027985,ERR4027986,ERR4027987,ERR4027988,ERR4027990,ERR4027993,ERR4027995,ERR4027996,ERR4027998,ERR4028000,ERR4028001,ERR4028004,ERR4028005,ERR4028008,ERR4028011,ERR4028012,ERR4028014,ERR4028020,ERR4028024,ERR4028026,ERR4028027,ERR4028029,ERR4028031,ERR4028032,ERR4028037,ERR4028038,ERR4028055,ERR4028056,ERR4028090,ERR4028091,ERR4028104,ERR4028105,ERR4028106,ERR4028107,ERR4028108,ERR4028116,ERR4028117,ERR4028118,ERR4028130,ERR4028136,ERR4028139,ERR4028143,ERR4028144,ERR4028145,ERR4028146,ERR4028147,ERR4028148,ERR4028149,ERR4028151,ERR4028152,ERR4028155,ERR4028158,ERR4028159,ERR4028160,ERR4028161,ERR4028162,ERR4028163,ERR4028165,ERR4028166,ERR4028178,ERR4028223,ERR4028224,ERR4028225,ERR4028226,ERR4028227,ERR4028228,ERR4028229,ERR4028230,ERR4028233,ERR4028241,ERR4028244,ERR4028248,ERR4028249,ERR4028366,ERR4028369,ERR4028370,ERR4028371,ERR4028404,ERR4028407,ERR4028408,ERR4076701,ERR4076702,ERR4076707,ERR4028368,ERR4028234,ERR4028245,ERR4028022,ERR4076708,ERR4028134,ERR4027991,ERR4028002,ERR4028156,ERR4028142,ERR4028367,ERR4028405,ERR4028019,ERR4028128,ERR4028138,ERR4028240
Phoebe	SRR7141988,SRR7141989,SRR7141990,SRR7141991,SRR7141993,SRR7141994,SRR7141995,SRR7141996,SRR7141997,SRR7141998,SRR7142001,SRR7142005,SRR7141981,SRR7141982,SRR7141985,SRR7141986,SRR7141999,SRR7142002,SRR7142003,SRR7142006,SRR7142007
Picea	SRR5807743,SRR5807745,SRR5807757,SRR6023856,SRR6023857,SRR6023860,SRR6023864,SRR6023874,SRR6023879,SRR6023881,SRR6023882,SRR6023897,SRR6023919,SRR5351825,SRR5351827,SRR5351828,SRR5351835,SRR5351836,SRR5351843,SRR5351845,SRR5351848,SRR5351853,SRR5351854,SRR5351859,SRR2903110,SRR2903123,SRR2903324,SRR2903330,SRR2903347,SRR2903373,SRR2903386,SRR2905452,SRR2905718,SRR2905757
Pitcairnia	SRR18052673,SRR18052674,SRR18052676,SRR18052678,SRR18052684,SRR18052687,SRR18052688,SRR18052689,SRR18052690,SRR18052691,SRR18052692,SRR18052693,SRR18052694,SRR18052695,SRR18052680,SRR18052681,SRR18052682,SRR18052683,SRR18052685,SRR18052686,SRR18052675
Phlox	SRR13694934,SRR13694935,SRR13694929,SRR13694930,SRR13694931,SRR13694999,SRR13695000,

	SRR13695001,SRR13694933,SRR13694936,SRR13694937,SRR13694938,SRR13694939,SRR13695002,SRR13695003,SRR13695004,SRR13694862,SRR13694873,SRR13694932,SRR13694940,SRR13694942,SRR13694943,SRR13694944,SRR13694945,SRR13694972,SRR13694983,SRR13694994,SRR13695005,SRR13695006,SRR13695007,SRR13694856,SRR13694857,SRR13694858,SRR13694859,SRR13694860,SRR13694988,SRR13694989,SRR13694990,SRR13694991,SRR13694992,SRR13694993,SRR13694995,SRR13694996,SRR13694997,SRR13694998,SRR13694848,SRR13694849,SRR13694850,SRR13694851,SRR13694853,SRR13694854,SRR13694872,SRR13694874,SRR13694875,SRR13694876,SRR13694877,SRR13694878,SRR13694879,SRR13694880,SRR13694881,SRR13694882,SRR13694883,SRR13694884,SRR13694885,SRR13694887,SRR13694888,SRR13694889,SRR13694890,SRR13694908,SRR13694909,SRR13694910,SRR13694911,SRR13694912,SRR13694913,SRR13694914,SRR13694918,SRR13694919,SRR13694920,SRR13694921,SRR13694922,SRR13694926,SRR13694941,SRR13694961,SRR13694971,SRR13694973,SRR13694974,SRR13694975,SRR13694976,SRR13694977,SRR13694978,SRR13694979,SRR13694980,SRR13694981,SRR13694852,SRR13694855,SRR13694861,SRR13694863,SRR13694864,SRR13694865,SRR13694866,SRR13694867,SRR13694868,SRR13694869,SRR13694870,SRR13694871,SRR13694886,SRR13694891,SRR13694892,SRR13694893,SRR13694894,SRR13694895,SRR13694896,SRR13694897,SRR13694898,SRR13694899,SRR13694900,SRR13694901,SRR13694902,SRR13694903,SRR13694905,SRR13694906,SRR13694907,SRR13694924,SRR13694946,SRR13694947,SRR13694948,SRR13694949,SRR13694950,SRR13694951,SRR13694952,SRR13694953,SRR13694954,SRR13694955,SRR13694956,SRR13694957,SRR13694958,SRR13694959,SRR13694960,SRR13694962,SRR13694963,SRR13694964,SRR13694965,SRR13694966,SRR13694967,SRR13694968,SRR13694969,SRR13694970,SRR13694982,SRR13694984,SRR13694985,SRR13694986,SRR13694987
Populus	SRR11308190,SRR11308191,SRR11308192,SRR11308193,SRR11308194,SRR11308195,SRR11308196,SRR11308197,SRR11308198,SRR11308199,SRR11308200,SRR11308201,SRR11308202,SRR11308203,SRR11308204,SRR11308205,SRR11308206,SRR11308207,SRR11308208,SRR11308209,SRR11308210,SRR11308211,SRR11308212,SRR11308213,SRR11308214,SRR11308215,SRR11308216
Pulmonaria	SRR9112547,SRR9112549,SRR9112555,SRR9112557,SRR9112558,SRR9112560,SRR9112562,SRR9112566,SRR9112567,SRR9112569,SRR9112571,SRR9112573,SRR9112575,SRR9112577,SRR9112581,SRR9112585,SRR9112586,SRR9112590,SRR9112592,SRR9112593,SRR9112595,SRR9112596,SRR9112597,SRR9112599,SRR9112601,SRR9112603,SRR9112604,SRR9112606,SRR9112607,SRR9112608,SRR9112610,SRR9112546,SRR9112554,SRR9112556,SRR9112559,SRR9112561,SRR9112563,SRR9112564,SRR9112565,SRR9112568,SRR9112570,SRR9112572,SRR9112574,SRR9112576,SRR9112578,SRR9112580,SRR9112584,SRR9112587,SRR9112589,SRR9112591,SRR9112594,SRR9112598,SRR9112602,SRR9112605,SRR9112609
Quercus	SRR12015666,SRR12015667,SRR12015669,SRR12015672,SRR12015674,SRR12015676,SRR12015678,SRR12015680,SRR12015683,SRR12015685,SRR12015686,SRR12015687,SRR12015688,SRR12015689,SRR12015690,SRR12015691,SRR12015692,SRR12015693,SRR12015694,SRR12015695,SRR12015696,SRR12015697,SRR12015698,SRR12015699,SRR12015700,SRR12015701,SRR12015702,SRR12015703,SRR12015704,SRR12015705,SRR12015706,SRR12015707,SRR12015708,SRR12015709,SRR12015710,SRR12015711,SRR12015712,SRR12015713,SRR12015714,SRR12015715,SRR12015716,SRR12015717,SRR12015718,SRR12015719,SRR12015721,SRR12015723,SRR12015727,SRR12015728,SRR12015729,SRR12015730,SRR12015731,SRR12015733,SRR12015735,SRR12015737,SRR12015738,SRR12015741,SRR12015743,SRR12015745,SRR12015747,SRR12015749,SRR12015751,SRR12015752,SRR12015753,SRR12015755,SRR12015756,SRR12015758,SRR12015760,SRR12015763,SRR12015764,SRR12015765,SRR12015767,SRR12015768,SRR12015769,SRR12015770,SRR12015772,SRR12015774,SRR12015776,SRR12015778,SRR12015780,SRR12015782,SRR12015784,SRR12015786,SRR12015787,SRR12015788,

	SRR12015790,SRR12015792,SRR12015794,SRR12015796,SRR12015797,SRR12015798,SRR12015799,SRR12015801,SRR12015802,SRR12015660,SRR12015661,SRR12015662,SRR12015663,SRR12015664,SRR12015665,SRR12015668,SRR12015670,SRR12015671,SRR12015673,SRR12015675,SRR12015677,SRR12015679,SRR12015681,SRR12015682,SRR12015684,SRR12015720,SRR12015722,SRR12015724,SRR12015725,SRR12015726,SRR12015732,SRR12015734,SRR12015736,SRR12015739,SRR12015740,SRR12015742,SRR12015744,SRR12015746,SRR12015748,SRR12015750,SRR12015754,SRR12015757,SRR12015759,SRR12015761,SRR12015762,SRR12015766,SRR12015771,SRR12015773,SRR12015775,SRR12015777,SRR12015779,SRR12015781,SRR12015783,SRR12015785,SRR12015789,SRR12015791,SRR12015793,SRR12015795,SRR12015800
Rhodanthemum	SRR9707525,SRR9707546,SRR9707547,SRR9707548,SRR9707549,SRR9707552,SRR9707554,SRR9707555,SRR9707504,SRR9707599,SRR9707600,SRR9707602,SRR9707515,SRR9707516,SRR9707517,SRR9707518,SRR9707520,SRR9707544,SRR9707545,SRR9707508,SRR9707509,SRR9707512,SRR9707519,SRR9707521,SRR9707522,SRR9707527,SRR9707528,SRR9707530,SRR9707505,SRR9707506,SRR9707507,SRR9707510,SRR9707514,SRR9707541,SRR9707556,SRR9707557,SRR9707559,SRR9707560,SRR9707561,SRR9707562,SRR9707563,SRR9707564,SRR9707565,SRR9707566,SRR9707568,SRR9707569,SRR9707577,SRR9707581,SRR9707586,SRR9707587,SRR9707596,SRR9707597
Salix	SRR6790655,SRR6790656,SRR6790658,SRR6790659,SRR6790660,SRR6790661,SRR6790662,SRR6790663,SRR6790664,SRR6790665,SRR6790666,SRR6790667,SRR6790668,SRR6790669,SRR6790670,SRR6790673
Senecio	SRR9326601,SRR9326602,SRR9326603,SRR9326604,SRR9326605,SRR9326606,SRR9326607,SRR9326608,SRR9326609,SRR9326610,SRR9326611,SRR9326612,SRR9326613,SRR9326614,SRR9326615,SRR9326616,SRR9326617,SRR9326618,SRR9326619,SRR9326620,SRR9326621,SRR9326622,SRR9326623,SRR9326624,SRR9326625,SRR9326626,SRR9326627,SRR9326628,SRR9326629,SRR9326630,SRR9326631,SRR9326632,SRR9326633,SRR9326634,SRR9326635,SRR9326636,SRR9326637,SRR9326638,SRR9326639,SRR9326640,SRR9326641,SRR9326642,SRR9326643,SRR9326644,SRR9326645,SRR9326646,SRR9326647,SRR9326648,SRR9326649,SRR9326650,SRR9326651,SRR9326652,SRR9326653,SRR9326654
Silene	SRR2351390,SRR2351382,SRR2351395,SRR2351456
Stachyurus	SRR9671160,SRR9671166,SRR9671155,SRR9671156,SRR9671157,SRR9671158,SRR9671154,SRR9671159,SRR9671161,SRR9671162,SRR9671165,SRR9671169
Yucca	SRR3930934,SRR3930739,SRR3930928,SRR3930706,SRR3930737,SRR3930732,SRR3930707,SRR3930933,SRR3930743,SRR3930932,SRR3930850,SRR3930672,SRR3930839,SRR3930847,SRR3930849,SRR3930825,SRR3930833,SRR3930829,SRR3930848,SRR3930673

Table S3: **Log-likelihood ratio test for logit models fitted to the selfing rate mixed quantiles datasets.**

model	ℓ	β_0	β_1	$X_{p=0.5}$	df	*P*-value
M_0	-37.34021	3.288	-1295.263	0.0025		
$M_{1\&2\ quantiles}$	-20.36679	3.888	-1390.980	0.0028		
$M_{3nd\ quantile}$	-15.7987	2.432	-1148.878	0.0021		
					2	0.309

ℓ: log-likelihoods of models M_0, $M_{1\&2\ quantiles}$ and $M_{3nd\ quantile}$.
β_0: estimated intercept.
β_1: estimated coefficient.
$X_{p=0.5}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{1\&2\ quantiles}) - \ell(M_{3nd\ quantile})|$ in a χ-squared distribution with two degrees of freedom.

model	ℓ	β_0	β_1	$X_{p=0.5}$	df	*P*-value
M_0	-37.34021	3.288	-1295.263	0.0025		
$M_{1\&3\ quantiles}$	-26.06544	3.32	-1373.56	0.0024		
$M_{2nd\ quantile}$	-10.90222	3.495	-1234.369	0.0028		
					2	0.689

ℓ: log-likelihoods of models M_0, $M_{1\&3\ quantiles}$ and $M_{2nd\ quantile}$.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{1\&3\ quantiles}) - \ell(M_{2nd\ quantile})|$ in a χ-squared distribution with two degrees of freedom.

model	ℓ	β_0	β_1	$X_{p=0.5}$	df	*P*-value
M_0	-37.34021	3.288	-1295.263	0.0025		
$M_{1st\ quantile}$	-9.273642	4.387	-1646.129	0.0027		
$M_{2\&3\ quantiles}$	-27.41773	2.772	-1150.472	0.0024		
					2	0.523

ℓ: log-likelihoods of models M_0, $M_{1st\ quantile}$ and $M_{2\&3\ quantiles}$.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{1st\ quantile}) - \ell(M_{2\&3\ quantiles})|$ in a χ-squared distribution with two degrees of freedom.

Supplementary code snippets

Listing 1: StacksExplorer.sh

```
#/bin/bash
# StacksExplorer.sh *directory with samples* *popmap file*

### This script launch a set of denovo_map.pl (from Stacks) runs with different parameters to
↪  found out the best configuration ###

# Check if the required arguments are provided
if [ $# -ne 2 ]; then
        echo -e "\nYou need to provide the directory containing the sample files (.fq.gz) and the
        ↪  popmap file as arguments.\n" && exit
fi

### setup of the ressources for the script
module load stacks/ # if you use Environment Modules (often the case on cluster)
function exists_in_list() {
    LIST=$1
    DELIMITER=$2
    VALUE=$3
    echo $LIST | tr "$DELIMITER" '\n' | grep -F -q -x "$VALUE"
} # function to check the answer of the user, used somewhere in the script.

### set the different arguments
## The three next lines explain the three parameters
# m = Minimum number of homomorphe reads for a stack to be valid (homomorphe under this treshold
↪  are considered secondary reads)
# M = Maximum distance between two stacks to be considered from the same polymorphe locus
↪  (distance = number of position with a different nucleotide)
# n = Maximum distance between two locus from differents populations to be considered to be the
↪  same locus (distance = number of position with a different nucleotide)

min_m=3 ### Hard coded value, can be modified ###
max_m=5 ### Hard coded value, can be modified ###
min_M=1 ### Hard coded value, can be modified ###
max_M=6 ### Hard coded value, can be modified ###
```

```
n_diff=1 ### Hard coded value, can be modified ### Distance from M, for exampple : if n_diff=1 and
↪  M=4 then n will take the value 3,4 and 5 (or 4 and 5 if you only test for value of n equal or
↪  superior to M)

samples=${1} # Directory with the individuals to analyzes
popmap=${2} # File of population mapping -> per line: name of the sampling\tpopulation belonging
cpus=10 ### Hard coded value, can be modified ### number of cpus per run
mem=40 ### Hard coded value, can be modified ### number of GB of ram per run
part="fast" ### Hard coded value, can be modified ### if run less than 24h = fast; if run longer
↪  than 24h = long

### check if the script is correctly set :
nb_m=$(expr $max_m - $(expr $min_m - 1)) # number of different values that m will take
nb_M=$(expr $max_M - $(expr $min_M - 1)) # number of different values that M will take
nb_n=$(expr 1 + $n_diff) # number of different values that n will take
nb_run=$(expr $nb_m \* $nb_M \* $nb_n) # total number of run with the current configuration

echo -e "\nPlease take a moment to check that this script is correctly set up :
Parameters for slurm (per run): $cpus cpus, $mem GB mem, partition set as $part
Range of values explored : m {${min_m}:${max_m}}, M {${min_M}:${max_M}}, n will be distant from
↪  $n_diff of M (only superior to M in this version of of the script).
With these values, we're in for $nb_run runs ! Are you sure you want this ? y or n ?\n"
read -p '' tmp # Ask the user if the setup is okay

okay_answer="y yes Y YES okay yup oui o OUI O"

# if the parameters and number of runs are okay, proceed, else exit
if exists_in_list "$okay_answer" " " "$tmp";then
        echo -e "\nOkay, launching the runs.\n"
else
        echo -e "\nPlease correct the value of parameters directly in the script (look for -Hard
        ↪  coded value-, then try again.\n"
        exit
fi

# create a stacksExplorer repertory if it doesn't already exist
if [[ ! -d "stacksExplorer" ]];then
        mkdir stacksExplorer
fi

### launch the different analysis
for m_int in $(seq ${min_m} ${max_m});do # for each value of m
        for M_int in $(seq ${min_M} ${max_M});do # for each value of M
                # to set the min max value of n, if n_diff = 0 then n = M
                #min_n=$(expr $M_int - $n_diff) # You can comment/uncomment this line, depending
                ↪  if you want to test for value of n inferior to M (or not)
                max_n=$(expr $M_int + $n_diff)
                for n_int in $(seq ${M_int} ${max_n});do # for each value of n
                        new_dir="stacksExplorer/stacksExplorer_m${m_int}_M${M_int}_n${n_int}" #
                        ↪  delete the directory for this configuration, if it already exist, else
                        ↪  create it
                        if [[ -d "$new_dir" ]];then
```

```
                                rm -r $new_dir
                                mkdir stacksExplorer/stacksExplorer_m${m_int}_M${M_int}_n${n_int}
                        else
                                mkdir stacksExplorer/stacksExplorer_m${m_int}_M${M_int}_n${n_int}
                        fi
                        echo "Running the m${m_int}_M${M_int}_n${n_int} configuration."
                        sbatch --mem ${mem}GB --cpus-per-task $cpus -p $part --wrap="time
                        ↪   denovo_map.pl -T 10 -M $M_int -m $m_int -n $n_int -o
                        ↪   ./stacksExplorer/stacksExplorer_m${m_int}_M${M_int}_n${n_int}/
                        ↪   --samples $samples --popmap $popmap --min-samples-per-pop 0.80 && rm
                        ↪   stacksExplorer/stacksExplorer_m${m_int}_M${M_int}_n${n_int}/*bam && rm
                        ↪   stacksExplorer/stacksExplorer_m${m_int}_M${M_int}_n${n_int}/*tags*"
                        done
        done
done

echo -e "\nThe runs have been launch.\nWhen done, the next soft to run will be gimmeRad2plot.sh\n"
```

Listing 2: gimmeRad2plot.sh

```
#!/bin/bash
# gimmeRad2plot.sh *stacksExplorer.sh directory*

### Used after the script StacksExplorer.sh to produce a file ready to be plot with
↪  StacksExplorer_plots.R
# it doesn't take any argument, and output a dataframe (stacksExplorer_rdy2plot.csv) with the m,
↪  M, n, r80 (nb of loci present in at least 80% of individuals of a pop), the number of
↪  polymorphic loci and the r80 polymorphic loci (% of loci polymorphe per population).


stacksExplorer=${1:-stacksExplorer}

# check if in the right place
[[ ! -d $stacksExplorer ]] && echo -e "\nNo $stacksExplorer directory found.\nDid you run the
↪  StacksExplorer.sh ?\nAre you in the parent directory of the $stacksExplorer ?\n" && exit

# check the number of fasta files produced (error if there is missing file)
nb_comb=$(find $stacksExplorer/ -maxdepth 1 -wholename '*stacksExplorer_m*' | wc -l)
nb_done=$(find $stacksExplorer/ -wholename '*/catalog.fa.gz' | wc -l)

# informe the user if there is missing files
if [ $nb_comb != $nb_done ];then
        echo -e "\nThe number of catalog.fa.gz doesn't seem to match the number of combinaisons
        ↪  tested.
Are you sure that everything went well with the denono_map.pl ?\n"
#        exit
fi

# get each populations and produce the corresponding headers
pop_file=$(find $stacksExplorer -wholename '*populations.log' | head -1) # to have a
↪  populations.log file, any does the job.
sp_list=$(grep "defaultgrp" $pop_file)
sp_list=${sp_list##*: }
sp_list=${sp_list// /_} # to change space between genus and species if the genus is provided
sp_list=$(echo $sp_list | sed 's/,_/\t/g') # to modified what is between species to tab
nb_sp=$(echo $sp_list | wc -w)
# create the dataframe to plot
echo -e "m\tM\tn\tr80\tpoly_loci\t$sp_list" > $stacksExplorer/stacksExplorer_rdy2plot.tsv

nb_dir=$(find ./$stacksExplorer -wholename '*/catalog.fa.gz' | wc -l)
echo -e "\nThere is $nb_dir analyses to collect.\n"

cnt=1
for i in $(find ./$stacksExplorer -wholename '*/catalog.fa.gz');do # catalog.fa.gz is only produce
↪  if denovo_map.pl worked fine
        comb=${i%/*} # get rid of the "catalog.fa.gz" in the i variable
        # nb of loci present in at least 80% of the individual of the population
        summary_path="${comb}/populations.sumstats_summary.tsv"
        start_line=$(expr $(expr $(cat $summary_path | wc -l) / 2) + 2) #line from which we start
        ↪  the loop, because there is two dataframe in one in this file
        comb2=${comb##*/}
        m_value=$(echo $comb2 | cut -d'_' -f2) && m_value=${m_value/m/}
        M_value=$(echo $comb2 | cut -d'_' -f3) && M_value=${M_value/M/}
        n_value=$(echo $comb2 | cut -d'_' -f4) && n_value=${n_value/n/}
        r80=$(zcat $i  | grep ">" | wc -l)
        # nb of polymorphic loci in the final conscencus file
        popu_log_distr="${comb}/populations.log.distribs" # file containing the distribution of nb
        ↪  of SNP per loci
        poly_loci_tot=$(grep -A5000 "BEGIN snps_per_loc_postfilters" $popu_log_distr |  sed '/^END
        ↪  snps_per_loc_postfilters$/,$d' | tail -n +5 | cut -d$'\t' -f 2 | awk '{Total=Total+$1}
        ↪  END{print Total}') # to sum the number of loci with at least one SNP (aka polymorphic
        ↪  loci)
        if [ -z "$poly_loci_tot" ];then # if the variable is empty, set it to "0", otherwise the
        ↪  script StacksExplorer_plotteur.R bugs.
                poly_loci_tot=0
        fi

        line_toadd="${m_value}\t${M_value}\t${n_value}\t${r80}\t${poly_loci_tot}"
        for u in $(seq $nb_sp);do # to collect value for each populations, it collect the value of
        ↪  the column % of polymo loci for each line (population) starting at the line of the
        ↪  first population (there is 2 dataframes in one file so the loop need to start at the
        ↪  right line)
                start_line=$(($start_line + 1))
                r80_loci_tmp=$(sed -n ${start_line}p $summary_path | cut -d$'\t' -f 6)
                line_toadd="${line_toadd}\t$r80_loci_tmp"
        done
        echo -e "$line_toadd" >> $stacksExplorer/stacksExplorer_rdy2plot.tsv
        echo -e "\e[1A\e[KDone $cnt on ${nb_dir}." # weird part to overwrite previous echo
        cnt=$(($cnt + 1))
done
echo -e "\nThe file stacksExplorer_rdy2plot.tsv has been produce in the $stacksExplorer
↪  directory.\nThe next script to use will be StacksExplorer_plots.R\n"
```

Listing 3: StacksExplorer_plots.R

```
#!/usr/bin/env Rscript
# StacksExplorer_plots.R *genus_name*

### To plot the dataframe from StacksExplorer.sh and
### gimmeRad2plot.sh (plot of r80 loci from rad data).
### Take as single argument the name of the genus.

library(data.table)
library(ggthemes)
library(ggbreak)
library(ggplot2)

args = commandArgs(trailingOnly=TRUE)
# A function factory for getting integer axis values.
integer_breaks <- function(n = 5, ...) {
  fxn <- function(x) {
    breaks <- floor(pretty(x, n, ...))
    names(breaks) <- attr(breaks, "labels")
    breaks
  }
  return(fxn)
}

genus = args[1]
plotme = fread("stacksExplorer_rdy2plot.tsv",header = T,sep = '\t')

# to collect the name of each species
species_names = names(plotme)[-(1:5)]

# to add a column to plot not the absolute value of n but the relative value from M
n_diff=c()
for (row in 1:nrow(plotme)) {
  M_tmp = as.integer(plotme[row,"M"])
  n_tmp = as.integer(plotme[row,"n"])
  diff_tmp = n_tmp - M_tmp
  n_diff = append(n_diff,diff_tmp)
  }
plotme = cbind(plotme,n_diff)

set_colors = c("#3fa261","#f8333c","#fcab10","#2b9eb3","#7712ba")

awesomePlot = ggplot(plotme,aes(M, r80)) +
  geom_line(aes(colour=factor(m), linetype=factor(n_diff)),size=1)+
  ggtitle(paste("r80 in function of m,M and n (",genus," data)",sep="")) +
  ylim(min(plotme$poly_loci)  * 0.975, max(plotme$r80) * 1.025) +
  labs(colour="m",linetype="n distance") + theme_light() +
  scale_x_continuous(breaks = integer_breaks())+
  scale_color_manual(values = set_colors) +
  labs(caption = "r80 = loci found in at least 80% of the population
  m = Minimum stack depth / minimum depth of coverage
  M = Distance allowed between stacks
  n = Distance allowed between catalog loci (express as a distance from M)") +
  scale_y_break(c(max(plotme$poly_loci) * 1.025, min(plotme$r80) * 0.975), scale = 0.5) +
  geom_line(aes(y=poly_loci,colour=factor(m),linetype=factor(n_diff)),size=1) +
  ylab("Number of loci") +
  geom_label( aes(x=max(plotme$M) ,y=min(plotme$poly_loci) * 0.975,hjust=1,vjust=0,
                label="Polymorphic loci"),size = 5,show.legend =F,stat = "unique") +
  geom_label( aes(x=max(plotme$M) ,y=max(plotme$r80) * 0.975,hjust=1,vjust=0,
                label="Total loci"),size = 5,show.legend =F,stat = "unique")

awesomePlot
name_pdf = paste('r80_',genus,'_plot.pdf',sep='')
pdf(name_pdf,width = 19,height = 10,onefile=F)
print(awesomePlot)
dev.off()

# to create specific ggplot (one per species) with % of polymorphe sites
for (species in species_names){
  print(species)
  tmp_plot = ggplot(plotme,aes_string("M",species)) +
    geom_line(aes(colour=factor(m),linetype=factor(n_diff)),size=1)+
    ggtitle(paste("% of r80 polymo loci in function of m,M and n (",species," data)",sep="")) +
    labs(colour="m",shape="n distance",y = "% r80 poly loci") +
    theme_light() +
    scale_x_continuous(breaks = integer_breaks()) +
    scale_color_manual(values = set_colors) +
    labs(caption = "% r80 poly loci = Share of porlymorphic loci in the loci present in at least
        80% of the population.
  m = Minimum stack depth / minimum depth of coverage
  M = Distance allowed between stacks
  n = Distance allowed between catalog loci (express as a distance from M)")
  name_pdf = paste(species,"Plot.pdf",sep="")
  pdf(name_pdf,width = 12,height = 6)
  print(tmp_plot)
  dev.off()
}
```

Listing 4: patchOneLine.sh

```
#!/bin/bash
# patchOneLine.sh *fasta to patch*

### to change sequence on multiple lines to sequence on one line ###

awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}  END {printf("\n");}' < $1 > doomed #
↪  made by Mathilde Barthe
rm $1
mv doomed $1

# to remove a possible empty first line

line=$(sed -n 1p $1)
if [ "$line" == "" ];then

        echo "The first line of the patched file was empty, it has been removed."
        sed -i '1d' $1

fi

echo ""
echo "Success, the file $1 has been one lined."
echo "Don't forget to check if the first line of your fasta is not an empty line !"
echo "(in which case it is necessary to delete it for reads2snp)"
echo ""
```

Listing 5: patchSPname.sh

```bash
#!/bin/bash
# patchSPname_v2.sh *fastaFile* *SraRunTable.txt*
## To change the field |sp| to the actual species name of the individuals

# This script take as input the fasta file from fastq2fasta and the metadata file
↪  (SraRuntable.txt)

# Check if there is all the arguments needed
if [ -z $1 ]; then
        echo ""
        echo "The first argument is not provided. It should indicate the fasta file to correct."
        echo ""
        exit 2
fi

if [ -z $2 ]; then
        echo ""
        echo "The second argument is not provided. It should indicate the SraRunTable.txt."
        echo ""
        exit 2
fi

# Collect the ID of the different individuals present in the fasta file, put it in a list.txt

echo "Initialization."

# Find the number of column with the Organism and Sample Name
rmSpace=$(head $2 -n 1)
rmSpace=${rmSpace// /_}
rmSpace=${rmSpace//,/ } #to loop on a list of strings
# For Organism
count=0
for header in $rmSpace ; do
        count=$((count+1))
        if [ $header == "Organism" ]; then

                spColumnNb=$count

        fi
done

#For Sample_Name
count=0
for header in $rmSpace ; do

        count=$((count+1))
        if [ $header == "Sample_Name" ]; then
```

```
                SnColumnNb=$count

        fi
done
field=$SnColumnNb","$spColumnNb
# Produce a list of individuals with their species from the metadatafile
cut $2 -d ',' -f ${field} | tail -n +2 | sort | uniq > infoSP.txt # the tail allow to skip the
↪  header of the metadata file.

# Proceed to correct/update the fasta file by running a sed for each individual in order to
↪  correct the
# species name
sed -i 's/ /_/' infoSP.txt # to set the name of species as Genus_species
sed -i "s/,/\\t/" infoSP.txt # for the awk cmd

# if the line is a header change the sp into the correct species name
awk -F'[\t,|]' 'BEGIN { OFS = "|" }; NR==FNR { spname[$2]=$1; next} {if ($1 ~ ">") $2=spname[$3];
↪  print $0}' infoSP.txt $1 > file.out

# to inform that the file has been corrected

name=${1/fas/fasta}
mv file.out $name
rm $1

echo ""
echo "$1 have been corrected."
echo "The name of the file has been changed for $name"
echo ""
```

Supplementary texts

Detailed commands of chapter I

Raw reads

Metadata were obtained directly from the SRA Run Selector of NCBI (Sayers et al., 2022). The SraRunTable.txt document was modified by replacing the intra-cells comma by underscore to respect .csv syntax. A few specific modifications were also made when required (sample name in the wrong field, or with an unusual form...). The accession list was also obtained from SRA Run Selector and used with *sra-toolkit 2.11* (https://github.com/ncbi/sra-tools) to download the .sra files :

```
prefetch optionfile SRR_Acc_List.txt
```

The accession list used is available for each dataset (genus) in the table S2. *Sra-toolkit 2.11* (https://github.com/ncbi/sra-tools) was also used to decompressed .sra files into .fastq with *fasterq-dump 2.10.3*:

```
fasterqdump splitfiles[9] file.sra
```

WGS and RNA fastq files could be used as is. RAD fastq files were first checked with process_radtags from *Stacks 2.6* (Catchen et al., 2011, 2013) to verify RAD cut sites (this step was done just by safety, as it was supposed to be already done pre-NCBI upload):

```
process_radtags -p input_directory -p output_directory -e enzyme1 renz2 enzyme2
-c -q
```

[9] Only for paired data.

Enzyme sites were then removed with *cutadapt 4.0* (Martin, 2011) to avoid any bias in the divergence estimation (since the enzyme sites cannot have any polymorphism):

```
cutadapt -g enzyme_sites -o output_directory input_fastq
```

Stacks 2.6 (Catchen et al., 2011, 2013), used later in this workflow, require reads of identical length. When this was not the case, reads were trimmed to a single length by removing reads too short and trimming the others using *cutadapt 4.0* (Martin, 2011):

```
cutadapt -l min_length -m min_length -o output_directory input_directory
```

The optimal length value allowing the conservation of the maximum amount of information was simply estimated with this formula

$$n = x. \sum_{k=x}^{lmax} n_{k'}$$

which gives the overall number of nucleotides kept after discarding reads that are shorter than the length x and trimming reads that are longer than x. For each length k, the number of reads of this length was computed using this command on two randomly selected samples[10]:

```
zgrep length sample.fastq.gz | cut -d'=' -f2 | sort | uniq -c | sort -n -r
```

Fastq_pair 1.0 (Edwards and Edwards, 2019) was used to remove single reads for paired datasets (since reads were independently removed in both ends of paired files):

```
fastq_pair pair_1.fastq pair_2.fastq
```

References

References for RNA and WGS datasets were collected from the dataset source article or from the 1KP project (Carpenter et al., 2019; Leebens-Mack et al., 2019). References from source article were used as is. References from 1KP were produced by extracting ORF from transcriptome files using *getORF* (Rice et al., 2000):

```
getorf -sequence 1kp.fasta -outseq getorf_output.fasta -min 300 -find 3 -reverse N
```

[10] For paired datasets, the estimation was made on each end of a sample, therefore two thresholds were used.

Then by merging similar sequences with *CD-Hit 4.8.1* (Fu et al., 2012; Li and Godzik, 2006):

```
cd-hit -i getorf_output.fasta -o cdhit_output.fasta -c 0.9 -T 7
```

References for RAD datasets were produced using *Stacks 2.6* (Catchen et al., 2011, 2013). As recommended in Paris et al. (2017), combinations of parameters were explored to obtain a good enough combination of arguments for denovo_map.pl (from *Stacks*). Assemblies on a subset of individuals were conducted with differents combinations of the three factors: -m (3 to 5), -M (1 to 6) and -n (equal to M or M + 1), for a total of 36 combinations (with arguments –min-samples-per-pop 0.80 and –rm-pcr-duplicates)), using a homemade script (script 1):

```
StacksExplorer.sh directory_with_samples popmap_file_subset
```

To explore the output of the different combinations, results were collected then plotted with a R script (script 3).

```
gimmeRad2plot.sh results_directory StacksExplorer_plots.R
```

A combination was then selected based on the trade-off between high number of polymorphic loci shared by at least 80% of the individuals and low value of parameters for m, M and n. This combination was later used to produce an assembly with all individuals per dataset (using *denovo_map.pl* from *Stacks*):

```
denovo_map.pl -T thread_number -M M_value -m m_value -n n_value -o output_directory –samples prepared_samples_directory –popmap popmap_file
```

Alignment and variant calling

References were indexed and aligned using *bowtie 2 2.5.1* (Langmead and Salzberg, 2012):

```
bowtie2-build reference_file project_name bowtie2 -x project_name -p thread_number
-U sample.fq.gz -S output.sam
```

Alignment files were cleaned (minimum quality of 20), sorted and indexed with *samtools 1.15.1* (Danecek et al., 2011):

```
samtools view -q 20 -@ thread_number -bS output.sam ¿ output_cleaned.bam samtools sort -@ thread_number output_cleaned.bam -o output_sorted.bam samtools index -@ thread_number output_sorted.bam
```

Variant calling was then made with *reads2snp* (Gayral et al., 2013; Tsagkogeorga et al., 2012) with a minimum of eight reads to call a genotype :

```
reads2snp -nbth thread_number -min 8 -out project_name -bamlist bam_list_file -bamref reference_file
```

The fasta outputs were then slightly modified with homemade scripts (scripts 4 and 5) to convert into the suitable fasta format and to add the name of the species for each read:

```
patchOneLine.sh fasta_file patchSPname.sh fasta_file SraRunTable.txt
```

Genetic clustering

A PCA was produced using *popPhyl_PCA* for each dataset for : 1) exclude F1 hybrids, since possible sterile F1 hybrids does not allow introgression, and 2) to split datasets into genetic clusters, regardless of the species taxonomy provided by the metadata. The genetic clusters were used as populations rather than the species taxonomy when considered appropriate.

```
python3 popphyl2PCA.py output_name fasta_file
```

Supplementary article

The following article, entitled *Rapid establishment of species barriers in plants compared to animals*, was submitted as is on 18 October 2023 to the journal SCIENCE. To date, we have not received any feedback from this submission. In the meantime, it is available on the BIORXIV platform at the following link `https://www.biorxiv.org/content/10.1101/2023.10.16.562535v1`:

Rapid establishment of species barriers in plants compared to animals

François Monnet[1,2,3], Zoé Postel[1], Pascal Touzet[1], Christelle Fraïsse[1], Yves Van de Peer[2,3,4,5], Xavier Vekemans[1], Camille Roux[1*]

[1]Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000, Lille, France
[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium
[3]VIB-UGent Center for Plant Systems Biology, Ghent, Belgium
[4]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa
[5]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

*Corresponding author: camille.roux@univ-lille.fr

Speciation, the process by which new reproductively isolated species arise from ancestral populations, occurs because of genetic changes that accumulate over time. To date, the notion that interspecific genetic exchange occurs more frequently between plant species than animals species has gained a strong footing in the scientific discourse, albeit primarily relying on verbal arguments centered on mating behavior. By examining the dynamics of gene flow across a continuum of divergence in both kingdoms, we observe the opposite relationship: plants experience less introgression than animals at the same level of genetic divergence, suggesting that species barriers are established more

rapidly in plants. This pattern questions the differences in microevolutionary processes between plants and animals that impact genetic exchange at the macroevolutionary scale.

One Sentence Summary

Genetic exchange is more frequent between animal species than plants, challenging historical views.

Introduction

Genetic exchange between populations or between speciating lineages has long been considered an important evolutionary process (*1*). The number of genetic novelties brought by introgression in a population can exceed the contribution of mutation alone, thus increasing both neutral and selected diversity, which can be the source of major evolutionary advances (*2*). One of the consequences of such introgression events is to facilitate the diffusion on a large scale (geographical and/or phylogenetic) of mutations that were originally locally beneficial (*3*). Evidently, genetic exchanges do not occur freely throughout the Tree of Life but are interrupted by species barriers that are progressively established in their genome as the divergence between evolutionary lineages increases. These genetic barriers to gene flow directly act by reducing the production of hybrids, or by affecting their fitness. The consequences of reproductive isolation can therefore be captured through the long-term effect of barriers on reducing introgression locally in the genomes, which provides a useful quantitative metric applicable to any organism (*4*). Thus, the genomes of speciating lineages go through a transitional stage, the so-called 'semi-isolated species', where they form mosaics of genomic regions more or less linked to barriers to gene flow (*5*). The consideration of this 'semi-isolated' status is key to better understanding the dynamics of the speciation process: *i)* When does the transition from populations to semi-isolated species occur? *ii)* At what level of molecular divergence do species become fully isolated?

One approach to studying these speciation dynamics *in natura* is to empirically explore a large continuum of molecular divergence composed by multiple pairs of sister lineages and to determine with model-based demographic inference which ones are currently genetically connected by gene flow (*6*). Introgression leaves detectable signatures in genomes, quantified by statistics commonly used in population genetics, including F_{ST} (*7*) and derivatives of the ABBA-BABA test (*8, 9*). Although they serve as a foundation for testing the hypothesis of strict allopatry between two lineages, they are not sufficient on their own to provide a quantification of the timing of gene flow and thus to estimate the current status of reproductive isolation of a pair of taxa. Recent computational methods have allowed the explicit evaluation of alternative evolutionary scenarios, notably to test the occurrence of ongoing gene flow, as well as to test the semi-permeability of species barriers in the genomes (*10, 11*). Applied to 61 pairs of animal taxa, these methods revealed that introgression is frequent until 2% of net divergences (*6*), and can even take place between lineages 14 times more divergent than the human-chimpanzee divergence (*12*).

The role of hybridisation and introgression in evolution benefited enormously from the efforts of botanists during the mid-20th century, but the patterns of speciation dynamics described above in animals are still unknown in plants. A historical overview of the literature suggests that plants would be more susceptible to hybridisation, and even introgression, than animals (*2, 13–17*). Despite a lack of comparative studies, this notion has been extensively adopted by the scientific community and is supported by some shortcuts. Primarily, the few empirical investigations comparing the dynamics of speciation in plants *versus* animals solely rely on morphological traits to arbitrarily define species (*16*). The emergence of molecular data has now rendered this issue surmountable, as it enables substituting the human-made species concept with genetic clusters that quantitatively vary in their level of genetic distance (*18*) and level of reproductive isolation (*4*). Secondly, the assertion of a higher magnitude of gene flow in plants relative to animals was established without any indication that comparable levels of divergence have been studied. Here, we undertake a comparative genomic approach using molecular datasets from the literature to challenge, with a unified statistical framework, the view that gene flow would be more prevalent in plants than animals for a given level of divergence.

Results

The present investigation examines the decrease in ongoing gene flow between lineages as a function of their genetic divergence, and compares its dynamics between two main kingdoms of the Tree of Life: plants and animals. For this purpose, we empirically explore a continuum of genetic divergence represented by 61 animal pairs and 280 plant pairs. Genomic data from each pair allows the quantification of molecular patterns of polymorphism and divergence by measuring 39 summary statistics commonly used in population genetics and the joint Site Frequency Spectrum (jSFS). For each observed dataset, we then tested whether the observed set of summary statistics was better reproduced by scenarios of speciation with or without migration by using an approximate Bayesian computation framework (ABC; (*10*)). The same ABC methodology for demographic inferences was employed for both the animal dataset (analyzed in (*6*)) and the plant dataset. The new plant dataset was produced from sequencing reads publicly available for 25 genera distributed in the plant phylogeny (212 pairs of eudicots, 45 of monocots, 21 of gymnosperms, 1 lycophyte and 1 magnoliid; Table S1) and were not chosen on the basis of a preconceived idea of their speciation mode (see supplementary materials A.2). The posterior probability of models with ongoing migration computed by the ABC framework is used to assign a status of isolation or migration to each pair along a continuum of divergence (Fig.1-A), allowing the comparison of speciation dynamics between plants and animals. In contrast to the expected outcomes reported in previous studies (*2, 13–17*), our findings suggest that in comparison to animals, plants exhibit a more rapid cessation of genetic exchange at lower levels of genetic divergence. This is characterized by a swifter transition from population pairs that are best-supported by migration models to those that are best described by isolation models ($P = 4.88 \times 10^{-15}$; Fig.1-A and table S2). Therefore, by fitting a generalized linear model for the migration/isolation status to the plant and animal datasets, as a function of the net molecular divergence, we determined that at a net divergence of $\approx 0.3\%$ (95% CI: [0.27%-0.47%]), the probability that two plant lineages are connected by gene flow falls below 50%, while in animals this inflection point occurs at higher levels of divergence close to 1.8% (95% CI: [1.52%-2%]; Fig.1-A and table S2). The plant dataset comprises genomic data derived from diverse sequencing methodologies, including RAD-sequencing ($n = 117$ pairs), RNA-sequencing ($n = 111$) and whole genome sequencing ($n = 52$), while the animal dataset predominantly consists of RNA-sequencing data ($n = 52$). To control for potential bias in sequencing technologies, we restricted our analysis solely to plant and animal datasets acquired through RNA-sequencing. The key result of a faster cessation of gene flow in plants than in animals is still supported ($P = 5.38 \times 10^{-8}$ and table S3), allowing us to reject the idea that our conclusions are derived from such a methodological bias. The number of pairs within a genus showing robust statistical support for either ongoing migration or current isolation in plants ranged from one to 31 pairs. Therefore, we also investigated a possible effect of sampling bias within the plant dataset. Through random sub-sampling involving a single pair of lineages per plant and animal genus, we demonstrate that the contrast in speciation dynamics between plants and animals consistently persists, also rejecting the idea that our result stems from the over-representation of a genus of plants with highly reproductively isolated lineages (Fig. S7).

To investigate the build-up of species barriers within the genomes of both plant and animal species, we now focus towards pairs supported by ongoing gene flow. Within the range of speciation scenarios considered, the rate of gene flow can be uniform across genomes (i.e., homogeneous) or it can vary locally from one genomic region to another (i.e., heterogeneous; see Fig. S5), contingent, respectively, upon the absence or presence of barrier genes that are expressed (*5*). The ABC framework described earlier allows us to classify animal and plant pairs as experiencing either genomically homogeneous or heterogeneous introgression (*19*). We find that plants experience a faster shift from the absence of barriers to semi-permeable barriers, the latter occurring at a net divergence of $\approx 0.2\%$ (compared to $\approx 0.6\%$ in animals; Fig.1-B). These findings demonstrate that, in plants, the initial species barriers that generate genomic heterogeneity of introgression rates, as well as the establishment of complete isolation between species, manifest at relatively lower levels of divergence than in animals. This suggests that the speciation process may require fewer mutations in plants than in animals for reproductive isolation to be both initiated and completed.

Finally, we conducted a comparative analysis of the temporal patterns of gene flow during divergence in plants *vs* animals. We specifically examine whether ongoing gene flow predominantly arises from a continuous migration model, initiated since the subdivision of the ancestral population (as illustrated in Fig. 2), or if it is a consequence of secondary contact following an initial period of geographic isolation and divergence. This model comparison using ABC is restricted to pairs for which we previously found a strong statistical support for ongoing gene flow. Our analysis shows that plants and animals differ in their primary mode of historical divergence, specifically in the extent of gene flow during the initial generations after the lineage split. Indeed, among animals, roughly 80% of the pairs that exhibit robust statistical evidence of ongoing migration diverged in the face of continuous gene flow since the initial split from their ancestor (Fig. 2). A minority of animal pairs (20%) underwent primary divergence in allopatry before coming into secondary contact. Conversely, in the case of plants, pairs that display ongoing gene flow have more frequently experienced secondary contacts ($\approx 55\%$; Fig. 2), in line with what is commonly assumed in plants (*20*). To control for the effect of geography, we computed the minimum geographic distance between taxa within each pair using the GPS data of the studied individuals (Fig. S1). Strikingly, our analyses reveal that ongoing migration is less frequent in pairs of plant lineages despite their closer average minimum geographic distance (≈ 488 km) than in animals ($\approx 2,230$ km), confirming that current geography is a poor predictor of genetic introgression in the history of sister species both in plants ($P = 0.155$) and animals ($P = 0.371$).

Discussion

The historical literature on hybridisation defined hybrids as the offspring of crosses between individuals from genetic lineages "*which are distinguishable on the basis of one or more heritable characters*" (*21*). Within this conceptual framework, examinations of numerous wild species have demonstrated a greater incidence of interspecific hybridization in plants than in animals (*16*), thus supporting the original assumption that plants are indeed more likely to hybridize than animals (*2, 13*). However, the advent of molecular markers to measure genetic differentiation in the early 2000s provided results in contrast to morphological studies, particularly by illuminating the higher F_{ST} values within plant species relative to animals (*22, 23*), indicating higher gene flow at the intraspecific level in the latter. Moreover, Morjan and Rieseberg (*22*) showed that this difference between kingdoms persists regardless of the mating system (from outcrossing to selfing) or the geographical distribution (local, regional, or biregional ranges). In our methodological approach, we depart from the human-made conception of 'species' and instead focus on genetic clusters that exhibit varying degrees of divergence and varying degrees of connectivity due to gene flow (Fig. S4). We could only attain this level of resolution because our methodology explicitly models the divergence history between lineages and captures the effect of species barriers on genomic patterns of gene flow. In doing so, we unravel the apparent paradox between studies of reproductive isolation between morphologically differentiated entities that suggest more frequent hybridization events in plants than in animals, and the greater genetic differentiation observed within plant species with molecular markers. Indeed, our explicit comparisons of ongoing migration models support the idea that scenarios of secondary contact are particularly frequent among the surveyed lineages in plants (*20*), whereas pairs of closely related animal species tend to experience gene flow more continuously over time. Secondary contact scenarios involve a preliminary phase of allopatry which affects the divergence of sister lineages on every marker: molecular and morphological. Such a historical context may thus engender the misconception that plants undergo hybridization events more frequently than animals simply because these events are more conspicuous in plants, as introgression happens more often between morphologically distinct lineages experiencing a secondary contact. This result implies, conversely, that genetic introgression appears to manifest with greater crypticity in animals.

The speciation process clearly does not follow a universal molecular clock, although certain molecular constraints inevitably make the process irreversible once a certain level of divergence is reached (*16*). While light has recently been shed on the rarity of hybrid zones found in plants (*20*), another mystery has now been added: why is the probability of being reproductively isolated greater in plants than in animals given identical genomic divergence? The multi-factorial nature of the speciation process (*24*) exacerbates the methodological limitations of our current approach in producing a simple explanation for the differences observed between plants and animals. Following the first reports based on morphological detection of hybrids and suggesting a greater occurrence of hybridization in plants than in animals, a range of hypotheses have been proposed to explain these observations. One commonly raised argument relates to pre-zygotic isolation, which is believed to exert greater influence on animals, primarily driven by behavioral preferences for reproductive partners (*16*). Another argument is based on the scarcity of heteromorphic sex chromosomes in plants, while they are common in animals (*17*), renowned for their influential role as a preferential sink for genetic barriers to introgression (*25*). While acknowledging the undeniable involvement of these processes, it is crucial to emphasize that they do not serve as definitive or all-encompassing mechanisms governing speciation. Distinctive attributes inherent to plants also provide a favorable context for the accumulation of species barriers within their genomes: *i)* the additional presence of chloroplasts in plant cells (*26, 27*), *ii)* the prevalence of selfing (*28–31*), *iii)* a certain dependency for reproduction on external pollinators (*32–34*), *iv)* less efficient dispersal modalities as illustrated by the higher intra-specific plant differentiation (*22, 23, 35, 36*) and *v)* stronger haploid selection (*37*). The proposed factors presented here are evidently not mutually exclusive, and it would be misleading to assert that the differences in speciation dynamics between plants and animals can be attributed to a single, easily testable factor. Understanding which properties of plants and animals, acting at the micro-evolutionary scale, lead to such a great disparity in speciation patterns at the macro-evolutionary scale, would benefit from a long-term community-based initiative for integrative speciation research across fields and taxa. Finally, we propose that the methodology employed herein to scrutinize variations in speciation dynamics between plants and animals could be extended to examine other contrasts encompassing diverse life-history traits, such as distinctions between external and internal fertilization, reproductive modes involving self-fertilization *versus* allo-fertilization, or variations in life cycles (haplobiontic *versus* diplobiontic), among others. Such prospective studies would be extremely valuable to better understand the respective roles of these various biological factors in influencing the establishment and maintenance of reproductive isolation.

References

1. R. Abbott, *et al.*, *Journal of evolutionary biology* **26**, 229 (2013).

2. G. L. Stebbins, *Proceedings of the American Philosophical Society* **103**, 231 (1959).

3. M. Slatkin, *Population genetics and ecology* (Elsevier, 1976), pp. 767–780.

4. A. M. Westram, S. Stankowski, P. Surendranadh, N. Barton, *Journal of evolutionary biology* **35**, 1143 (2022).

5. C.-I. Wu, *Journal of evolutionary biology* **14**, 851 (2001).

6. C. Roux, *et al.*, *PLoS biology* **14**, e2000234 (2016).

7. S. Wright, *Annals of eugenics* **15**, 323 (1949).

8. S. H. Martin, J. W. Davey, C. D. Jiggins, *Molecular biology and evolution* **32**, 244 (2015).

9. A. J. Dagilis, *et al.*, *Evolution Letters* **6**, 344 (2022).

10. C. Fraïsse, *et al.*, *Molecular Ecology Resources* **21**, 2629 (2021).

11. D. R. Laetsch, *et al.*, *bioRxiv* pp. 2022–10 (2022).

12. C. Fraïsse, *et al.*, *Peer Community Journal* **2** (2022).

13. E. Mayr, *Animal species and evolution* (Harvard University Press, 1963).

14. L. Gottlieb, *The American Naturalist* **123**, 681 (1984).

15. P. R. Grant, B. R. Grant, *Science* **256**, 193 (1992).

16. J. Mallet, *Trends in ecology & evolution* **20**, 229 (2005).

17. B. A. Payseur, L. H. Rieseberg, *Molecular ecology* **25**, 2337 (2016).

18. N. Galtier, *Evolutionary applications* **12**, 657 (2019).

19. C. Roux, G. Tsagkogeorga, N. Bierne, N. Galtier, *Molecular biology and evolution* **30**, 1574 (2013).

20. R. J. Abbott, *Journal of Systematics and Evolution* **55**, 238 (2017).

21. R. G. Harrison, *et al.*, *Oxford surveys in evolutionary biology* **7**, 69 (1990).

22. C. L. Morjan, L. H. Rieseberg, *Molecular ecology* **13**, 1341 (2004).

23. R. Frankham, C. J. Bradshaw, B. W. Brook, *Biological Conservation* **170**, 56 (2014).

24. D. I. Bolnick, *et al.*, *Evolution* **77**, 318 (2023).

25. C. Fraïsse, H. Sachdeva, *Genetics* **217**, iyaa025 (2021).

26. S. Greiner, U. Rauwolf, J. Meurer, R. G. Herrmann, *Molecular ecology* **20**, 671 (2011).

27. Z. Postel, *et al.*, *Molecular Phylogenetics and Evolution* **169**, 107436 (2022).

28. C. Goodwillie, S. Kalisz, C. G. Eckert, *Annu. Rev. Ecol. Evol. Syst.* **36**, 47 (2005).

29. E. E. Goldberg, *et al.*, *Science* **330**, 493 (2010).

30. M. Pickup, *et al.*, *New Phytologist* **224**, 1035 (2019).

31. L. Marie-Orleach, C. Brochmann, S. Glémin, *PLoS Genetics* **18**, e1010353 (2022).

32. C. Devaux, R. Lande, *Journal of evolutionary biology* **22**, 1460 (2009).

33. K. M. Kay, R. D. Sargent, *Annual Review of Ecology, Evolution, and Systematics* **40**, 637 (2009).

34. F. P. Schiestl, P. M. Schlüter, *Annual review of entomology* **54**, 425 (2009).

35. R. Tiedemann, *et al.*, *Molecular Ecology* **13**, 1481 (2004).

36. S. Edmands, *Molecular ecology* **16**, 463 (2007).

37. S. Immler, S. P. Otto, *The American Naturalist* **192**, 241 (2018).

38. F. Monnet, *et al.*, Rapid establishment of species barriers in plants compared to animals (2023).

39. Y. Liu, *et al.*, *New Phytologist* **215**, 877 (2017).

40. H. Dittberner, A. Tellier, J. de Meaux, *Molecular Biology and Evolution* **39**, msac015 (2022).

41. M. K. Brandrud, *et al.*, *Systematic Biology* **69**, 91 (2020).

42. D. Souto-Vilarós, *et al.*, *Journal of Ecology* **106**, 2256 (2018).

43. C. E. Grover, *et al.*, *Genome Biology and Evolution* **14**, evac170 (2022).

44. G. L. Owens, *et al.*, *Molecular Ecology* **30**, 6229 (2021).

45. J. Norrell (2017).

46. D. P. Wood, J. K. Olofsson, S. W. McKenzie, L. T. Dunning, *Botany Letters* **165**, 476 (2018).

47. L. Dunning, *et al.*, *Journal of evolutionary biology* **29**, 1472 (2016).

48. O. G. Osborne, *et al.*, *Molecular Biology and Evolution* **36**, 2682 (2019).

49. M. Scharmann, A. Wistuba, A. Widmer, *Molecular Phylogenetics and Evolution* **163**, 107214 (2021).

50. B. E. Goulet-Scott, A. G. Garner, R. Hopkins, *Evolution* **75**, 1699 (2021).

51. X. Ding, J. H. Xiao, L. Li, J. G. Conran, J. Li, *Journal of Systematics and Evolution* **57**, 234 (2019).

52. M. M. Tavares, M. Ferro, B. S. S. Leal, C. Palma-Silva, *Ecology and Evolution* **12**, e8834 (2022).

53. Y. Sun, *et al.*, *Evolution* **72**, 2669 (2018).

54. D. Ru, *et al.*, *Molecular Ecology* **27**, 4875 (2018).

55. H. Shang, *et al.*, *Philosophical Transactions of the Royal Society B* **375**, 20190544 (2020).

56. S. Grünig, M. Fischer, C. Parisod, *Annals of botany* **127**, 21 (2021).

57. J. Ortego, L. L. Knowles, *Molecular Ecology* **29**, 4510 (2020).

58. F. Wagner, *et al.*, *Molecular phylogenetics and evolution* **144**, 106702 (2020).

59. S. Gramlich, N. D. Wagner, E. Hörandl, *BMC Plant Biology* **18**, 1 (2018).

60. B. Nevado, S. A. Harris, M. A. Beaumont, S. J. Hiscock, *Molecular Ecology* **29**, 4221 (2020).

61. X.-S. Hu, D. A. Filatov, *Molecular ecology* **25**, 2609 (2016).

62. A. Muyle, *et al.*, *Molecular Biology and Evolution* **38**, 805 (2021).

63. Y. Feng, H. P. Comes, Y.-X. Qiu, *Molecular phylogenetics and evolution* **150**, 106878 (2020).

64. A. M. Royer, M. A. Streisfeld, C. I. Smith, *American journal of botany* **103**, 1730 (2016).

65. B. Langmead, S. L. Salzberg, *Nature methods* **9**, 357 (2012).

66. N. Matasci, *et al.*, *Gigascience* **3**, 2047 (2014).

67. J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, *G3: Genes— genomes— genetics* **1**, 171 (2011).

68. J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, W. A. Cresko, *Molecular ecology* **22**, 3124 (2013).

69. J. R. Paris, J. R. Stevens, J. M. Catchen, *Methods in Ecology and Evolution* **8**, 1360 (2017).

70. B. Charlesworth, M. Morgan, D. Charlesworth, *Genetics* **134**, 1289 (1993).

71. N. Barton, B. O. Bengtsson, *Heredity* **57**, 357 (1986).

72. M. A. Beaumont, *Annual review of ecology, evolution, and systematics* **41**, 379 (2010).

73. F. Tajima, *Genetics* **105**, 437 (1983).

74. G. Watterson, *Theoretical population biology* **7**, 256 (1975).

75. F. Tajima, *Genetics* **123**, 585 (1989).

76. M. Nei, W.-H. Li, *Proceedings of the National Academy of Sciences* **76**, 5269 (1979).

77. S. Wright, *Genetics* **28**, 114 (1943).

78. J. Wakeley, J. Hey, *Genetics* **145**, 847 (1997).

79. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, *Molecular biology and evolution* **34**, 1812 (2017).

80. B. Nevado, G. W. Atchison, C. E. Hughes, D. A. Filatov, *Nature communications* **7**, 12384 (2016).

Acknowledgments

We are grateful to the Institut Français de Bioinformatique (IFB; ANR-11-INBS-0013) thanks to whom bioinformatics analyses and demographic inferences have been carried out.
The authors would like to thank Hélène Morlon, Bert Van Boxlaer, John Pannell and Martin Lascoux for their constructive feedback.

Funding: This work was supported by grant from I-SITE ULNE – Ghent University.

Y.VdP. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). **Authors contributions:** F.M (data curation, formal analysis, investigation, visualization, writing), Z.P (data curation), P.T (resources, data curation, supervision), C.F (writing), Y.V.d.P (funding acquisition, supervision, writing), X.V (funding acquisition, supervision, writing), C.R (funding acquisition, conceptualization, methodology, supervision, investigation, visualization, writing). **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. The assembled datasets, the list of references used for mapping and the results of demographic inference are deposited in `Zenodo` with the DOI `10.5281/zenodo.8028615` (*38*). Scripts for bioinformatic treatment of raw sequencing data are available from GitHub `https://github.com/Ladarwall/Greenworld`.

Supplementary Materials

Tables S1-S3

Figures S1-S7

Figures

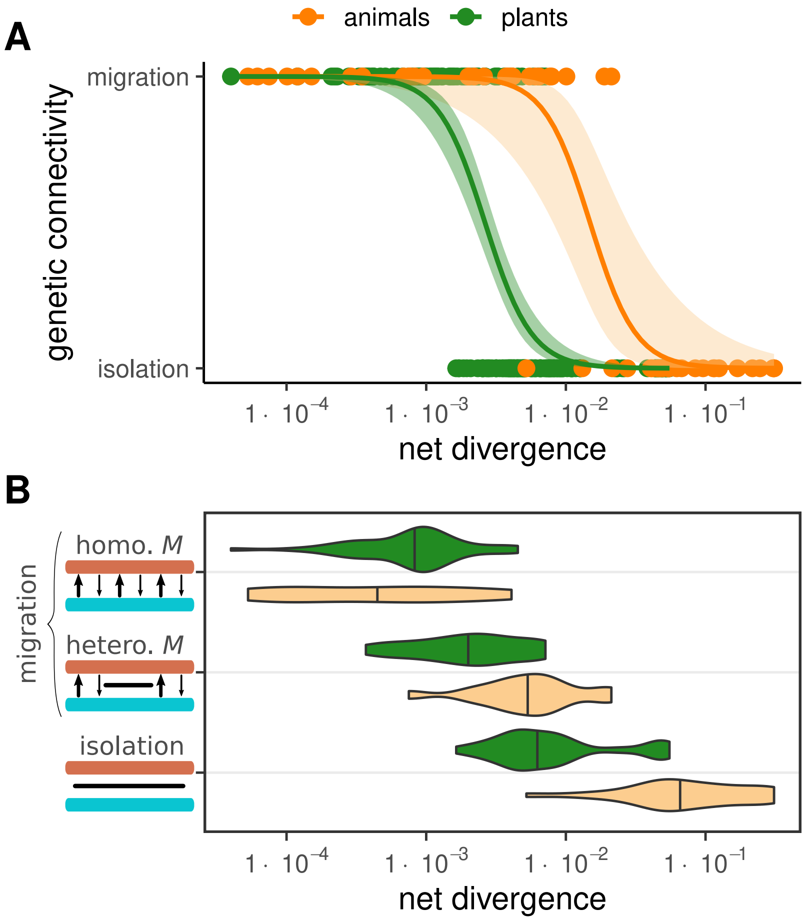


Figure 1: **Genomic patterns of introgression along a divergence continuum in plants *versus* animals.**

Estimation of the average genomic divergence and the migration/isolation status was performed for 280 pairs of plant species/populations (green) and compared to 61 animal pairs (orange) analysed using the same ABC procedure (*6*).

A. x-axis: average net divergence within a pair. y-axis: best supported model in a comparison between ongoing migration and current isolation. Each point represents a pair of populations/species. Curves represent the logit models fitted to the plant and animal data.

B. Distribution of the average net divergence of plant (green) and animal (orange) pairs whose genomic data are best explained by homogeneous (homo. M) or heterogeneous (hetero. M) distributions of migration rates across the genome, or by complete genetic isolation (isolation). y-axis: blue and brown bars symbolize homologous chromosomes within a studied pair. Black arrows symbolize genome regions connected by gene flow. Black bars symbolize local effects of barriers against gene flow.

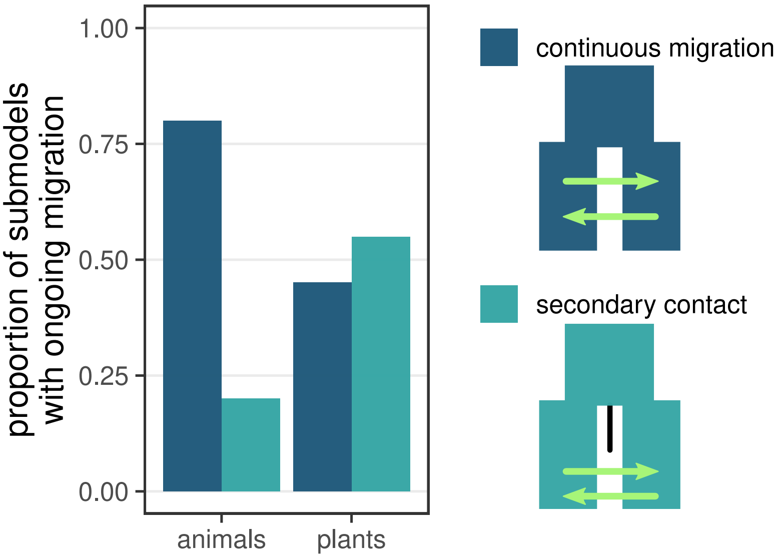


Figure 2: **Temporal models of ongoing migration in plants and animals.**
Strong statistical support of ongoing migration was observed in 82 pairs of plants and 30 pairs of animals. For every pair supported by ongoing gene flow, sub-models were compared to distinguish between continuous migration (dark blue) *versus* secondary contact (light blue).

Supplementary Materials for: Rapid establishment of species barriers in plants compared to animals

François Monnet[1,2,3], Zoé Postel[1], Pascal Touzet[1], Christelle Fraïsse[1],
Yves Van de Peer[2,3,4,5], Xavier Vekemans[1], Camille Roux[1*]

[1]Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000, Lille, France
[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium
[3]VIB-UGent Center for Plant Systems Biology, Ghent, Belgium
[4]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa
[5]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

[*]Corresponding author: camille.roux@univ-lille.fr

A Materials and Methods

A.1 Animal dataset

The animal data come from the Roux et *al.* (2016) study (*6*). They consist essentially of non-model animal populations/species, initially selected without any particular knowledge about the demographic history, and were sampled from natural populations. These data were produced by RNA sequencing, and only synonymous positions were retained for statistical inferences.

A.2 Plant sampling

Raw data used in this work comes from previously published studies (*39–64*). The following criteria were applied to identify datasets in plants:

1. Currently diploid genomes.
2. High-throughput sequencing, i.e, RNA-seq, RAD-seq or whole genome sequencing (WGS).
3. Freely available from `NCBI`.
4. Individuals sampled from natural populations (geographic distribution represented in Fig. S1).
5. A minimum of two sampled populations/species per genus.
6. A minimum of two sequenced individuals per sampled population/species.

Datasets fitting these criteria were examined through exploration of literature found *via* Google Scholar (`https://scholar.google.com`), `NCBI` (`https://www.ncbi.nlm.nih.gov/Traces/study/`) and `DDBJ` (`https://ddbj.nig.ac.jp/search`).

Finally, 118 different plant species/populations from 25 different genera were retained for the demographic analysis according to our criteria (Table S1), allowing 280 pairwise demographic analyses to be carried out. These comparisons cover all possible pairs within each genus. No comparisons are made between different genera, with the exception of comparisons within the *Laccospadicinae* (*Howea* and *Linospadix*) due to their relatively small genetic distance.

A.3 Assembly, read mapping and genotype calling

For the plant datasets: reads and metadata were downloaded using SRA-Toolkit, version 2.11.0 (`https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit`). Here we separate plant projects for which we worked with synonymous positions (from RNA-seq: *n*=7 genera and WGS: *n*=4) from those for which we could not (from RAD sequencing: *n*=13):

A.3.1 Reads from RNA-seq and WGS.

In line with the animal dataset (*6*), the bioinformatic strategy applied to the plant data is to retain synonymous positions. Reads for a given population/species pair were therefore mapped to a reference transcriptome with the bowtie2 program version 2.4.2 (*65*): either taken from the `1KP project` (*66*) if a species of the same genus is represented there (`https://db.cngb.org/onekp/search/`), or taken from the data associated with the original articles when available (Table SS1). Every position (variants and invariants) were called with a minimum of 8 reads using Reads2SNP 2.0, the uncalled low-quality positions were then coded as “N”. The resulting fasta file was used for each population/species as the input file for the demographic inferences.

A.3.2 Reads from RAD-seq.

Loci were assembled for each RAD-seq dataset using `Stacks` 2.6 (*67, 68*). Combinations of parameters were explored following Paris et *al.* 2017 (*69*) to maximise the amount of biological information retained. Using the two or four samples with the highest amount of available data per lineage, assemblies were built using *denovo_map.pl* (`Stacks`) with different combinations of parameters: the minimum depth for a stack to be valid (-*m*, ranging from 3 to 5), the number of mismatches allowed between stacks within individuals (-*M*, ranging from 1 to 6) and the number of mismatches allowed between stacks between individuals (-*n*, set to *M* or *M* + 1), for a total of 36 combinations. In addition, loci that were missing in at least 20% of the samples per population were withdrawn with the argument *–min-samples-per-pop 0.80* (i.e. only loci with the information for all samples were kept, as populations were composed of two or four samples). The number of polymorphic loci was plotted as a function of the different combinations of parameters using a homemade `R` script. For each dataset, a combination was selected in function of the trade-off between maximising the number of polymorphic loci and minimising the parameter values to produce a reference set of loci for each species/population pair. Reads were mapped on this reference with bowtie2 version 2.5.1, and variants were called with Reads2SNP 2.0 in the same way as "RNA-seq and WGS" datasets.

A.4 Demographic inferences

Model comparisons were carried out using the approximate Bayesian computation (ABC) framework applied in the animal study (*6*) and distributed under the name DILS (for Demographic Inferences with Linked Selection (*10*)). Here we describe how DILS works.

A.4.1 Compared models

The primary objective of our demographic analysis is to determine which historical scenario explain the best a given dataset. The term dataset here refers to a pair of populations/species (comprising either two animal or two plant lineages, for which genomic data are described by an array of summary statistics (see section A.4.2). In our ABC methodology, we discern two categories of models.

Four demographic Models: Each of these models describes the subdivision of an ancestral population into two daughter populations (Fig. S5-A). The three populations have independently assigned effective population sizes. The differences between these four models concern the historical patterns of gene flow between two divergent populations, as depicted in figure S5. These models encompass continuous migration (CM), and secondary contact (SC), strict isolation (SI) and ancient migration (AM) :

- models with ongoing migration
  - continuous migration (CM)
  - secondary contact (SC)
- models with current isolation
  - strict isolation (SI)
  - ancient migration (AM)

Notably, the former two models entail ongoing gene flow between the two populations, while the latter two do not. Models with past (AM) or recent (CM and SC) migration assume gene flow between sister populations/species in both directions, at two independently assigned rates.

Models of Linked Selection: Effects of linked selection have been taken into account using a genomic model that encompasses: (a) heterogeneous effective population size across the genome (*hetero. N*), which closely approximates the influence of background selection by down-scaling *Ne* (*70*); and/or (b) heterogeneous migration rate across the genome (*hetero. M*) to account for the effects of selection against hybrids (*71*). The modeling framework employed in this study does not consider the effects of positive selection on linked loci (i.e., genetic hitchhiking).

Within the *hetero. N* genomic model, the variable effective size among loci is assumed to conform to a re-scaled Beta distribution. In essence, all populations share a common Beta distribution with two shape parameters drawn from uniform distributions. However, each population is independently re-scaled by distinct *Ne* values, which are drawn from uniform distributions. Conversely, the *homo. N* genomic model assumes that all loci from the same genome share the same effective population size, and this parameter is independently estimated in all populations. This homogeneous model implies that the genomic landscape remains unaffected (or is uniformly affected) by background selection.

The *hetero. M* genomic model implements local reduction of gene flow in the genome. Variation in migration rates among loci is thus modeled by employing a bimodal distribution where a proportion of loci, drawn from a uniform distribution in]0-1[, is linked to barriers (i.e., $N.m = 0$), while the loci unaffected by species barriers are associated to an effective migration rate *N.m* drawn from a uniform distribution. In the *homo. M* model, a single migration rate *N.m* per direction is universally shared by all loci in the genome.

Subdivisions of the four demographic models (CM, SC, SI and AM) into various genomic submodels were made to accommodate for the effect of linked selection. Heterogeneity in effective population size was a universal consideration across all four models, while heterogeneity in migration rate was specifically accounted for in models exhibiting gene flow (i.e., CM, AM, and SC). Therefore, the SI model was divided into two submodels (*homo. N* or *hetero. N*), while the AM, CM, and SC models were divided into four submodels:

1. *homo. N* and *homo. M*

2. *homo. N* and *hetero. M*

3. *hetero. N* and *homo. M*

4. *hetero. N* and *hetero. M*

For a comprehensive description of all prior distributions employed in this study, please refer to Section A.4.3.

A.4.2 Summary statistics

ABC is a statistical inferential approach based on the comparison of summary statistics derived from simulated and observed datasets (*72*). We present a comprehensive description of the statistics computed within our framework. The following summary statistics are calculated for each locus:

- The number of bi-allelic polymorphisms in the alignment including all sequenced copies in the 2 species/populations

- Pairwise nucleotide diversity π (*73*)

- Watterson's θ (*74*)

- Tajima's D (*75*)

- The proportion of sites displaying fixed differences between the populations/species (S_f)

- The proportion of sites featuring polymorphisms exclusive to a specific population/species (S_{xA} and S_{xB})

- The fraction of sites with polymorphisms shared between the two populations/species (S_s)

- The number of successive shared polymorphic sites

- Raw divergence D_{xy} between the two populations/species (*76*)

- Net divergence D_a between the two populations/species (*76*)

- Relative genetic differentiation between the two populations/species quantified by F_{ST} (*77*)

For the ABC analysis, we used the means and variances of these statistics calculated over all the available loci. Additionally, we utilize the joint Site Frequency Spectrum (jSFS (*78*)) to summarize the data, specifically capturing the count of single-nucleotide polymorphisms (SNPs) where the minor allele occurs in each bin covering the jSFS. Because of the absence of outgroup lineages, jSFS were folded. Singletons are deliberately excluded from the jSFS to mitigate potential inference biases arising from sequencing errors. Each of the non-excluded bin of the jSFS is used as a descriptive statistics in the ABC analysis.

We supplement this set of summary statistics with measures taken on all the loci:

- Pearson's correlation coefficient for π between species

- Pearson's correlation coefficient for θ between species

- Pearson's correlation coefficient between D_{xy} and D_a

- Pearson's correlation coefficient between D_{xy} and F_{ST}

- Pearson's correlation coefficient between D_a and F_{ST}

- Proportion of loci with both S_s and S_f sites

- Proportion of loci with S_s sites but no S_f

- Proportion of loci without S_s sites but with S_f

- Proportion of loci with neither S_s nor S_f sites

The summary statistics obtained from both the empirical data sets (i.e., plants and animals) and the data sets simulated under the demographic models (Fig. S5) were calculated with the same scripts implemented in DILS.

A.4.3 Configuration file

DILS was run using the following parameter values:

- mu = 7.31×10^{-9}

- useSFS = 1

- barrier = bimodal

- $max\_N\_tolerated = 0.25$

- $L_{min} = 10$

- $n_{min} = 4$

- $rho\_over\_theta = 0.2$

- uniform prior for N between 0 and N_{max} individuals

- uniform prior for T_{split} between 0 and T_{max} generations

- uniform prior for migration rate $N.m$ between 0 and 10 migrants per generations

Where:

$$N_{max} = 5 \times max\left(\frac{\pi_A}{4\mu}; \frac{\pi_B}{4\mu}\right)$$

π_A and π_B being the Tajima's θ (*73*) for species A and B respectively (for a given pair).

$$T_{max} = 5 \times \frac{D_a}{2\mu}$$

D_a being the net divergence (*76*).

A.4.4 Returned quantities

At the end of the analysis, DILS returns the posterior probability of ongoing migration *versus* of current isolation. The probability of ongoing migration corresponds to the relative probability of all models including ongoing migration (Secondary Contact, Continuous Migration) and their sub-models (heterogeneity and genomic homogeneity for migration and effective size); while the probability of current isolation corresponds to all models and sub-models with current isolation (Strict Isolation, Ancient Migration). These quantities are used to produce the relationships between the net divergence and the posterior probability of migration (Fig. S6). For each pair of populations/species, three statuses are then assigned:

1. Strong support for genetic isolation: we identify strong statistical support for genetic isolation when our ABC framework yields a posterior probability $P_{\text{migration}} < 0.1304$. This threshold was empirically determined by the robustness test conducted in (*6*).

2. Strong support for ongoing migration: strong statistical support for ongoing migration is indicated when the posterior probability $P_{\text{migration}} > 0.6419$, also empirically determined in (*6*).

3. Ambiguity: statistical ambiguity, denoting situations where our ABC framework does not strongly support either migration or isolation, i.e, when the risk of assigning an analysed pair to a wrong status is greater than 5%.

Pairs for which support was inconclusive were excluded from further analysis. The remaining pairs were categorized either as exhibiting 'migration' or 'isolation,' as illustrated in Figure 1-A, allowing the 'migration' status to be treated in a logistic regression (see section A.5).

A.5 Logistic regression

To study speciation dynamics, we examine reduction in the proportion of plant or animal pairs receiving strong support for models with migration as a function of time (measured here by the net molecular divergence). For this purpose, we modeled $\mathbf{Y}_i$ (the binary status 'isolation' or 'migration' best fitting the data) as a function of $\mathbf{X}_i$ (the average net genomic divergence) by using a generalized linear model (GLM) *via* a linked binomial function:

$$g\left(E(Y_i|\mathbf{X}_i)\right) = g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1,i}$$

where β_0 represents the intercept and β_1 the coefficient reflecting the effect of genomic divergence on the isolation/migration status coded as 0 and 1, respectively. The fitted model is used to predict p_i, the proportion of pairs of populations/species that are currently connected by gene flow (migration status) for a given level of divergence $\mathbf{X}_i$.

$$p_i = \frac{\exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)} = \frac{1}{1 + \exp\left(-\mathbf{X}_i\boldsymbol{\beta}\right)}$$

Reversely, we can determine the divergence level $\mathbf{X}$ for which a given proportion p_i of pairs are connected by gene flow:

$$X = -\frac{1}{2\beta_1}\left(\beta_0 + \sqrt{\beta_0^2 + 4\beta_1 \log\left(\frac{p_i}{1-p_i}\right)}\right)$$

We are interested in comparing the inflection point, i.e, the level of divergence above which more than 50% of species pairs are genetically isolated, between plants and animals. Thus, for a given fitted model, this point corresponds to a divergence level $X = -\frac{\beta_0}{\beta_1}$.

The log-likelihood function ℓ of the migration/isolation status $\mathbf{Y}$ given the average net molecular divergence $\mathbf{X}$ is then obtained to evaluate the fit of a model to the observed data:

$$\begin{aligned}
\ell(\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}) &= \log\left(\mathcal{L}(\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x})\right) \\
&= \sum_{i=1}^{N}\left[y_i\log(p_i)+(1-y_i)\log(1-p_i)\right] \\
&= \sum_{i=1}^{N}\left[y_i\log\left(\frac{p_i}{1-p_i}\right)+\log(1-p_i)\right] \\
&= \sum_{i=1}^{N}\left[y_i\cdot\mathbf{x}_i\boldsymbol{\beta}-\log(1+\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right))\right] \\
&= \sum_{i=1}^{N}\left[y_i\cdot(\beta_0+\beta_1X_{1,i})-\log(1+\exp\left(\beta_0+\beta_1X_{1,i}\right))\right] \qquad (1)
\end{aligned}$$

We can now test whether the sigmoid of plants is significantly different from that of animals, thereby testing if plants and animals share the same speciation dynamic. For this purpose, three models are fitted and associated to log-likelihood ℓ:

1. M_0: both plants and animals share the same logistic relationship between $\mathbf{X}_i$ and $\mathbf{Y}_i$.

2. M_{plants}: model fitted to the plants data only.

3. $M_{animals}$: model fitted to the animals data only.

Thus, for M_0 we fitted a GLM to the entire dataset comprising both plants and animals, after having retained only demographic inferences for which the ABC analysis produced strong statistical support for ongoing migration or current isolation, following the test of robustness applied in Roux et al. (*6*). In that sense, pairs of plants and animals with ambiguous support for isolation or migration were excluded from all GLM regressions. The log-likelihood $\ell(M_0)$ was then estimated for the whole dataset comprising both plants and animals by using formula **1** where:

- $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ represent for M_0 the coefficient of the model fitted to the whole plants and animals dataset by using the **glm** function (family = 'binomial') implemented in R.
- $\mathbf{X}_{1,i}$ represents the series of observed net divergence values.
- $\mathbf{y}_i$ represents the series of inferred isolation/migration status.

For M_{plants} and $M_{animals}$, we fitted a GLM model only to data from the corresponding kingdom. We then estimated the log-likelihoods $\ell(M_{plants})$ and $\ell(M_{animals})$ as for M_0.

Finally, we conducted a comparison between the log-likelihood $\ell(M_0)$ and the combined log-likelihood $\ell(M_{plants})+\ell(M_{animals})$, which is derived from the summation of log-likelihoods obtained by fitting independent models to each respective kingdom. The significance of the difference between $\ell(M_0)$ and $\ell(M_{plants}) + \ell(M_{animals})$ was evaluated using a log-likelihood ratio test. Specifically, twice the absolute difference of the log-likelihood between $\ell(M_0)$ and $\ell(M_{plants})+\ell(M_{animals})$ is approximately χ-squared distributed. The *P*-value returned by the R function **pchisq** corresponds to the probability of observing $2.|\ell(M_0)-\ell(M_{plants})-\ell(M_{animals})|$ in a χ-squared distribution with two degrees of freedom (Table S2).

A.6 Testing for a phylogenetic effect

To control for the variation in the number of pairs between genera, we carried out 5,000 animal-plant comparisons as for Fig. 1 but by randomly selecting a single pair per animal and plant genus. Over these 5,000 sub-samples, the relative positions of the sigmoids were compared *via* the inflection points of the models fitted to the plant *versus* animal sub-samples. The inflection point was estimated as being $-\frac{\beta_0}{\beta_1}$. We find that the inflection point is found systematically at lower levels of divergence in plants than in animals (Fig. S7).

A.7 Testing for a sequencing technology effect

Out of the total dataset comprising 280 pairs of plants and 61 pairs of animals, 210 plant pairs and 54 animal pairs exhibited strong statistical support for migration or isolation based on ABC model comparison. These retained datasets encompass a diversity of sequencing methodologies. Specifically, within plants, among the 210 retained pairs: 90 pairs were acquired through RAD-sequencing, 86 pairs through RNA-sequencing, and 34 pairs through whole genome sequencing. In the case of animals: 46 pairs were derived from RNA-sequencing, while 8 pairs were the result of Sanger sequencing. To assess the potential influence of sequencing techniques, we determined whether the observed differences in dynamics between plants and animals, as previously reported for the entire dataset, remained consistent when considering only the data generated exclusively through RNA sequencing. This choice was motivated by the fact that RNA-sequencing is the sole sequencing technique shared by both biological kingdoms under study. By retaining only the data from RNAseq, we maintain a statistically significant support for a more rapid cessation of gene flow in plants than in animals, despite a *P*-value that increases from 4.88×10^{-15} (Table S2) to 5.38×10^{-8} (Table S3).

A.8 Geography

Geographical (geodesic) distance in meters was measured using GPS coordinates provided in the metadata when available, using the **distGeo** function in the `R` package *geosphere*. For a given pair of populations/species A and B, this distance corresponds to the distance between the two geographically closest individuals. In the case of sampled sympatric pairs, and if a single coordinate was provided by the authors for all individuals A and B, we consider a distance of

10m in line with current sampling practices to reduce relatedness. Among the 25 plant genera under examination, our review of the literature has not yielded information pertaining to the geographical origins of specimens from *Gossypium*.

A.9 Data availability

All the assembled datasets, the list of references used for mapping and the results of demographic inference are deposited in `Zenodo` with the DOI `10.5281/zenodo.8028615` (*38*).

B Supplementary Figures

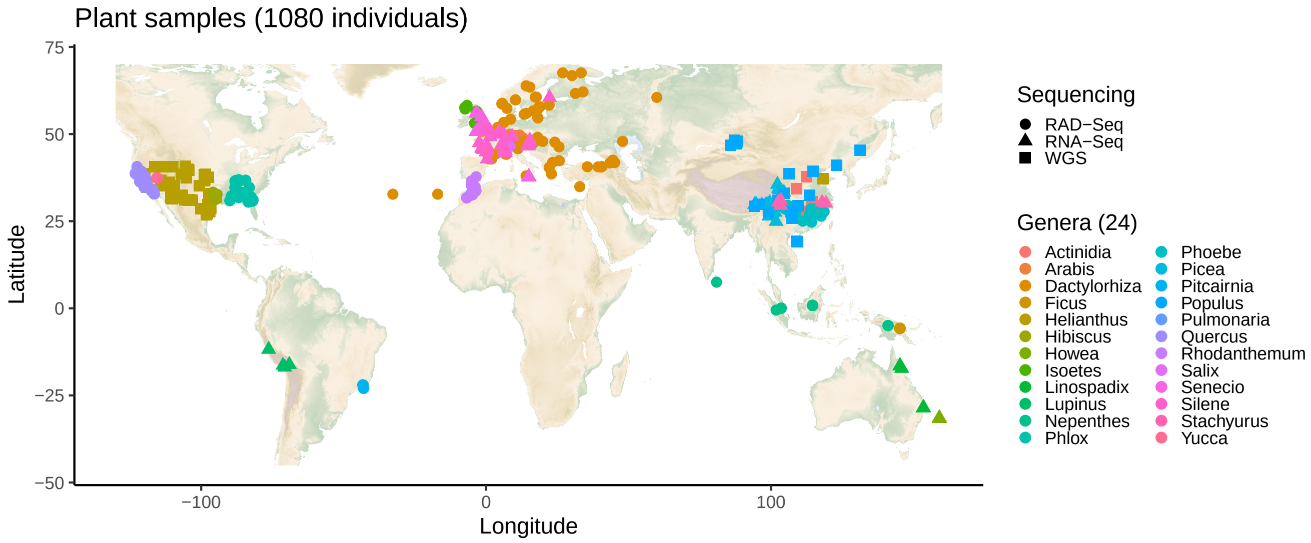


Figure S1: **Geographical location of plant samples and sequencing methods**
The depicted symbol represents an entity for which geographic data was retrievable in the source publications, with the exception of the genus *Gossypium*. The varying shapes of the symbols serve to differentiate between distinct sequencing technologies.

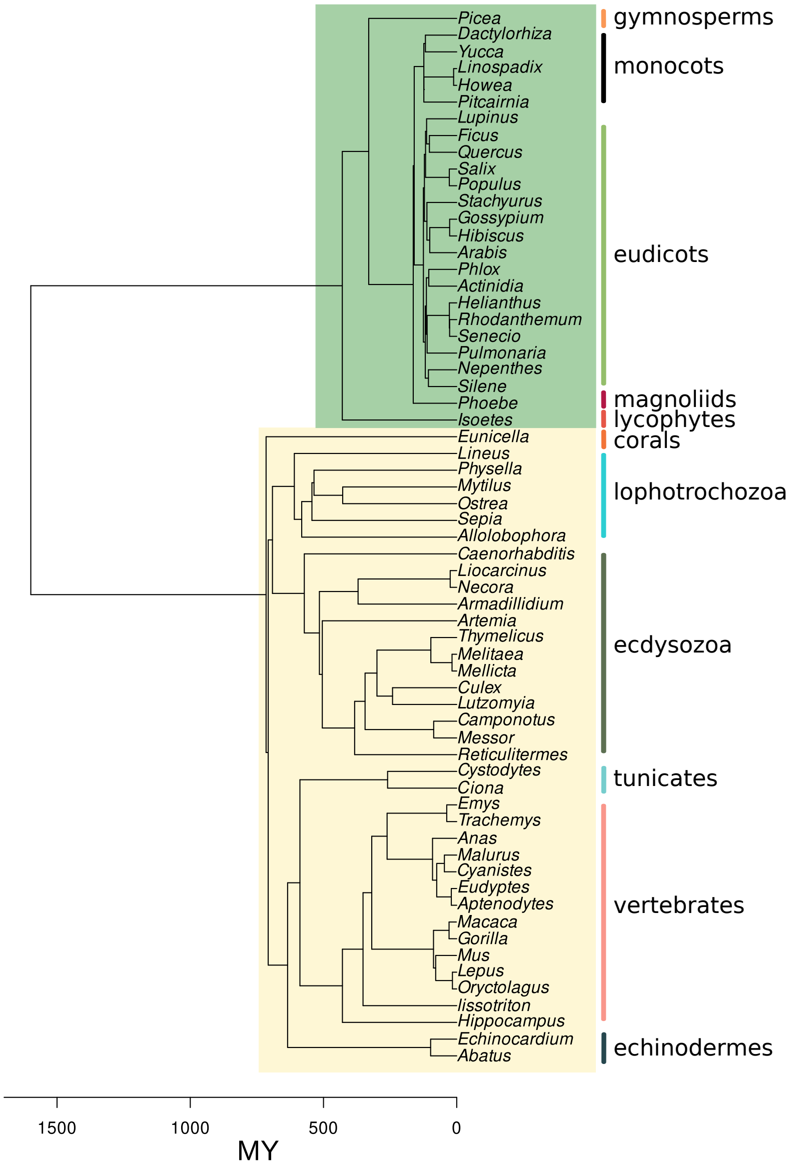


Figure S2: **Phylogenetic relationships between species included in the current study.** Plants and animals are indicated by green and yellow rectangles respectively. The scale represent the time from present expressed in million years (MY) according to `TimeTree` (*79*). Animals (yellow square) are from (*6*). Plants (green square) are included in the current study.

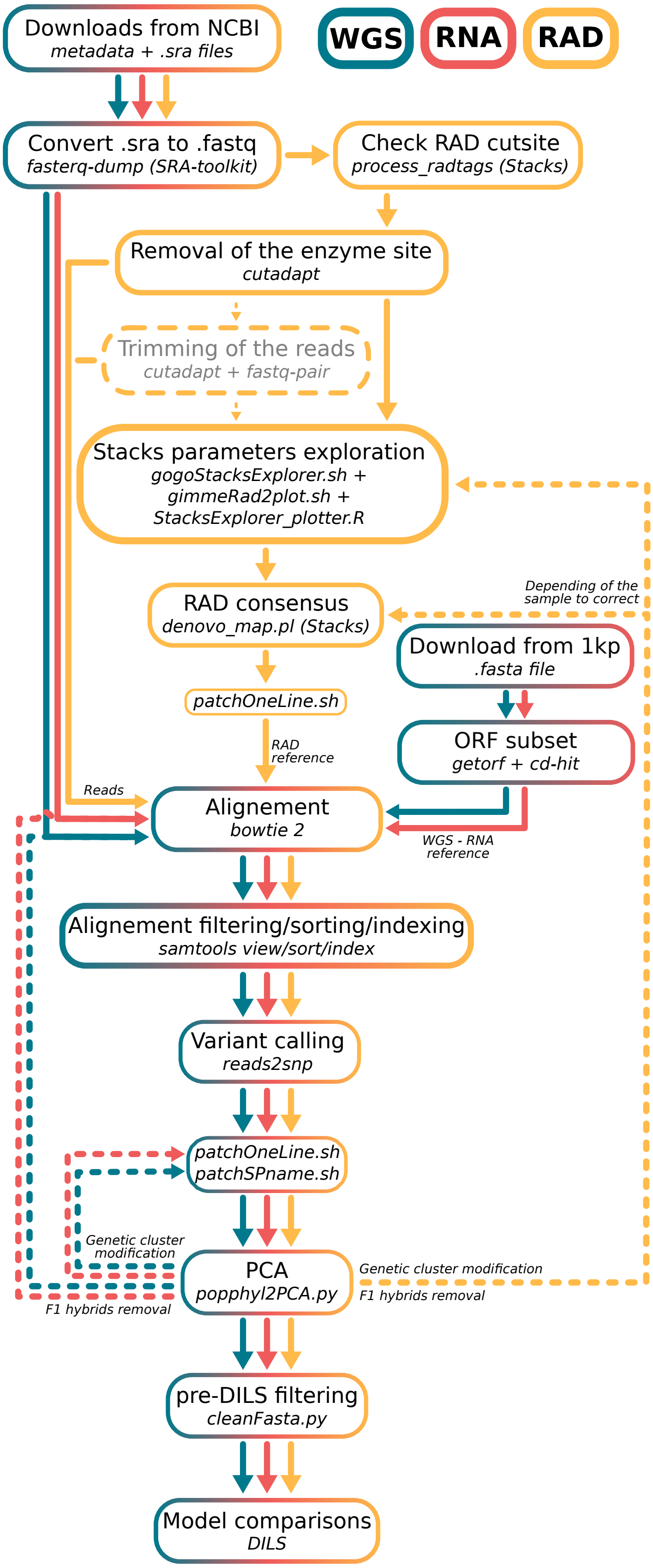


Figure S3: **Bioinformatics steps from the raw reads to demographic analysis**
Within each box, the upper line delineates an information technology procedure employed for data processing, while the lower line specifies the program or script utilized for its execution. The coloration denotes the specific sequencing technology concerned by each step.

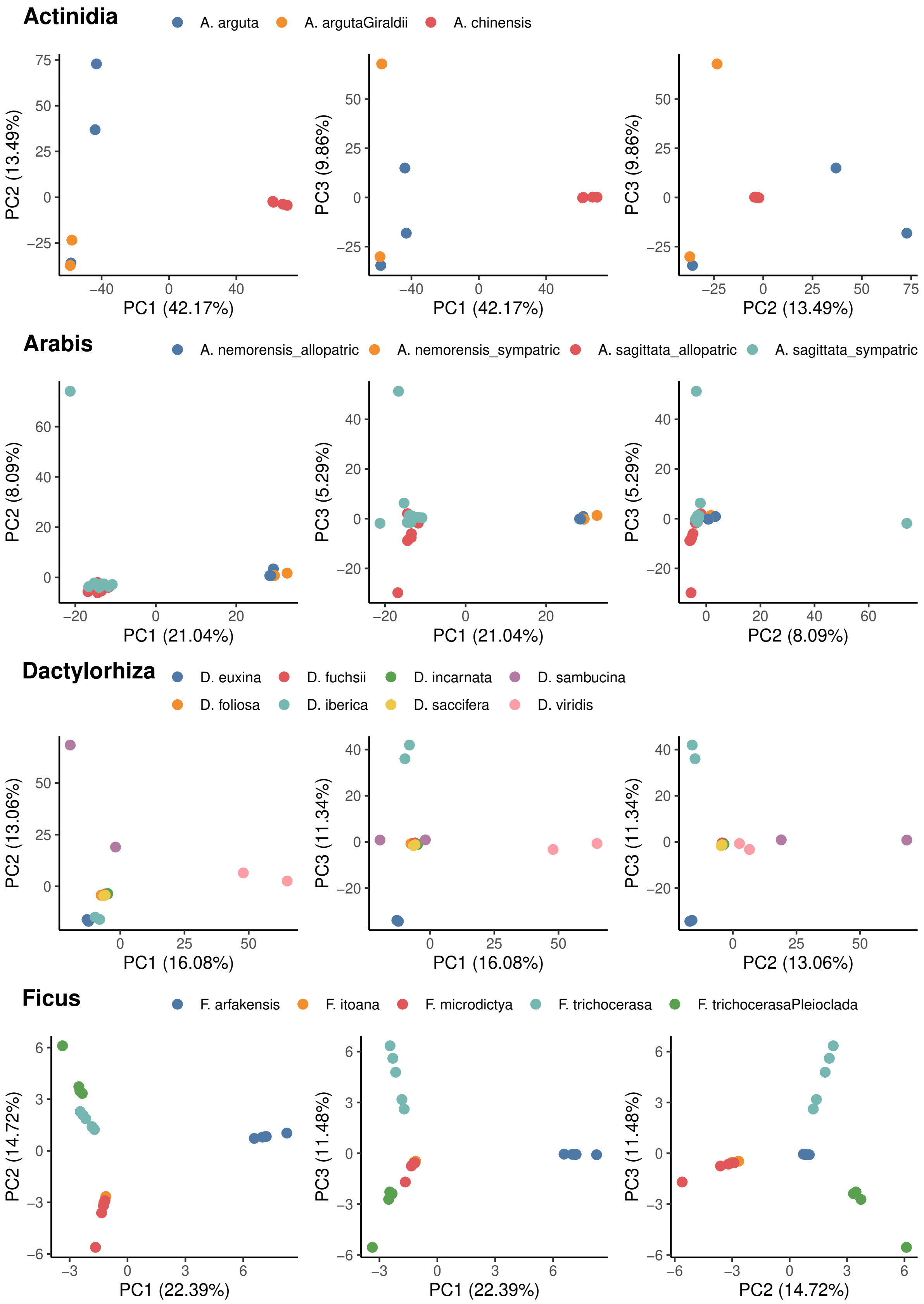


Figure S4: **Principal component analyses on genotypes for all SNPs.**
Each point represents an individual. The colours represent the different populations/species named by the authors of the studies from which the data originated.

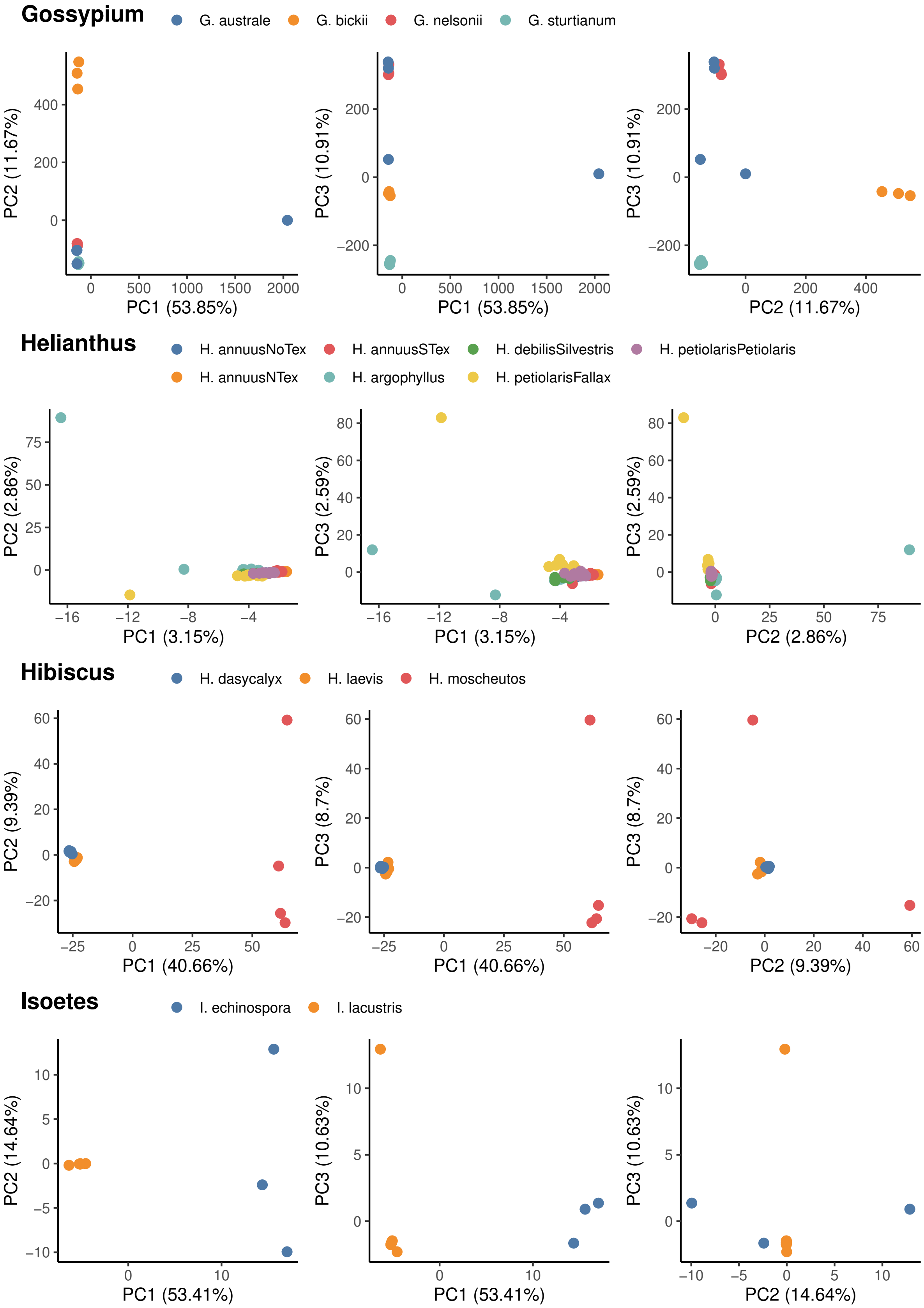


Figure S4: (continued).

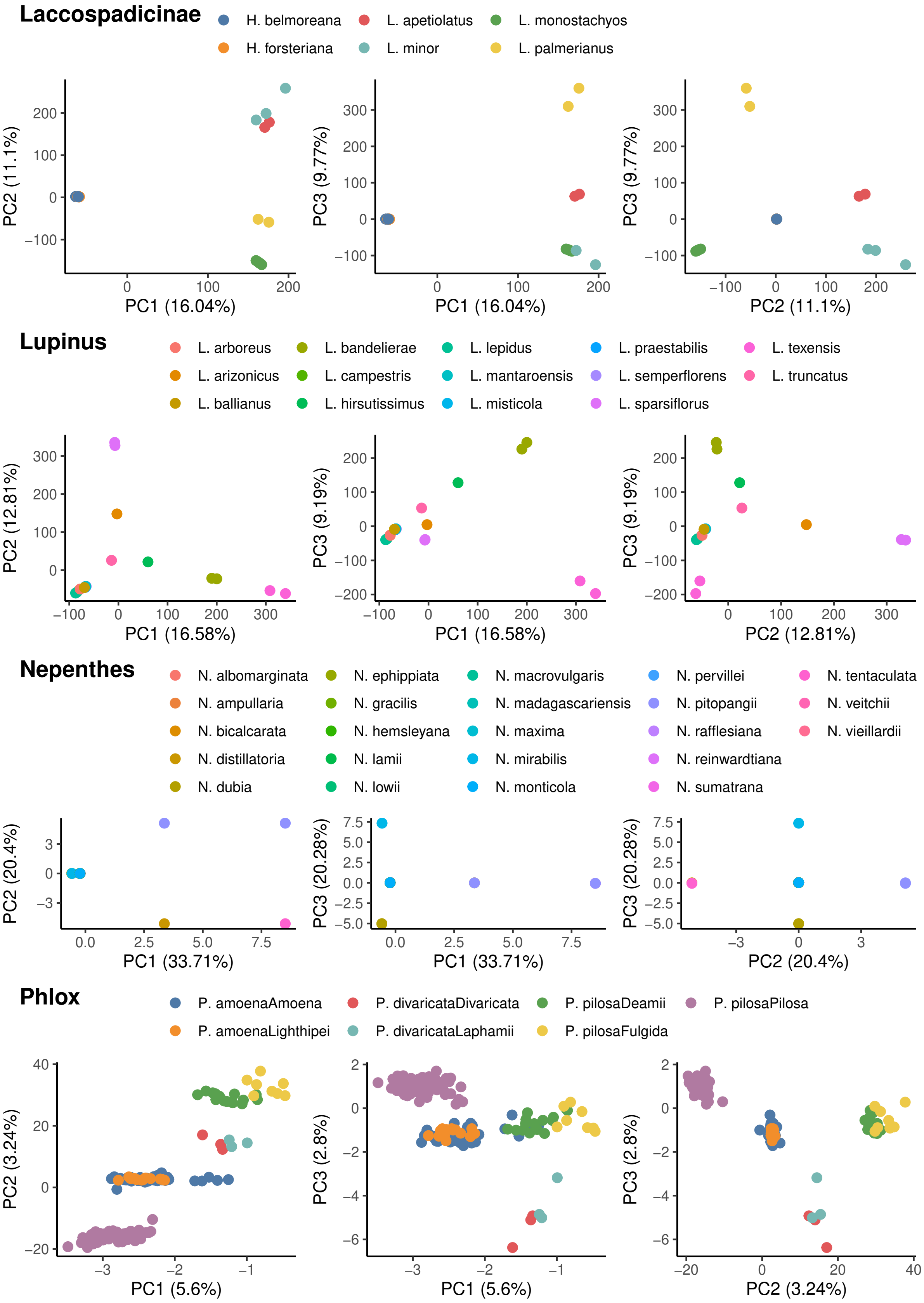


Figure S4: (continued).

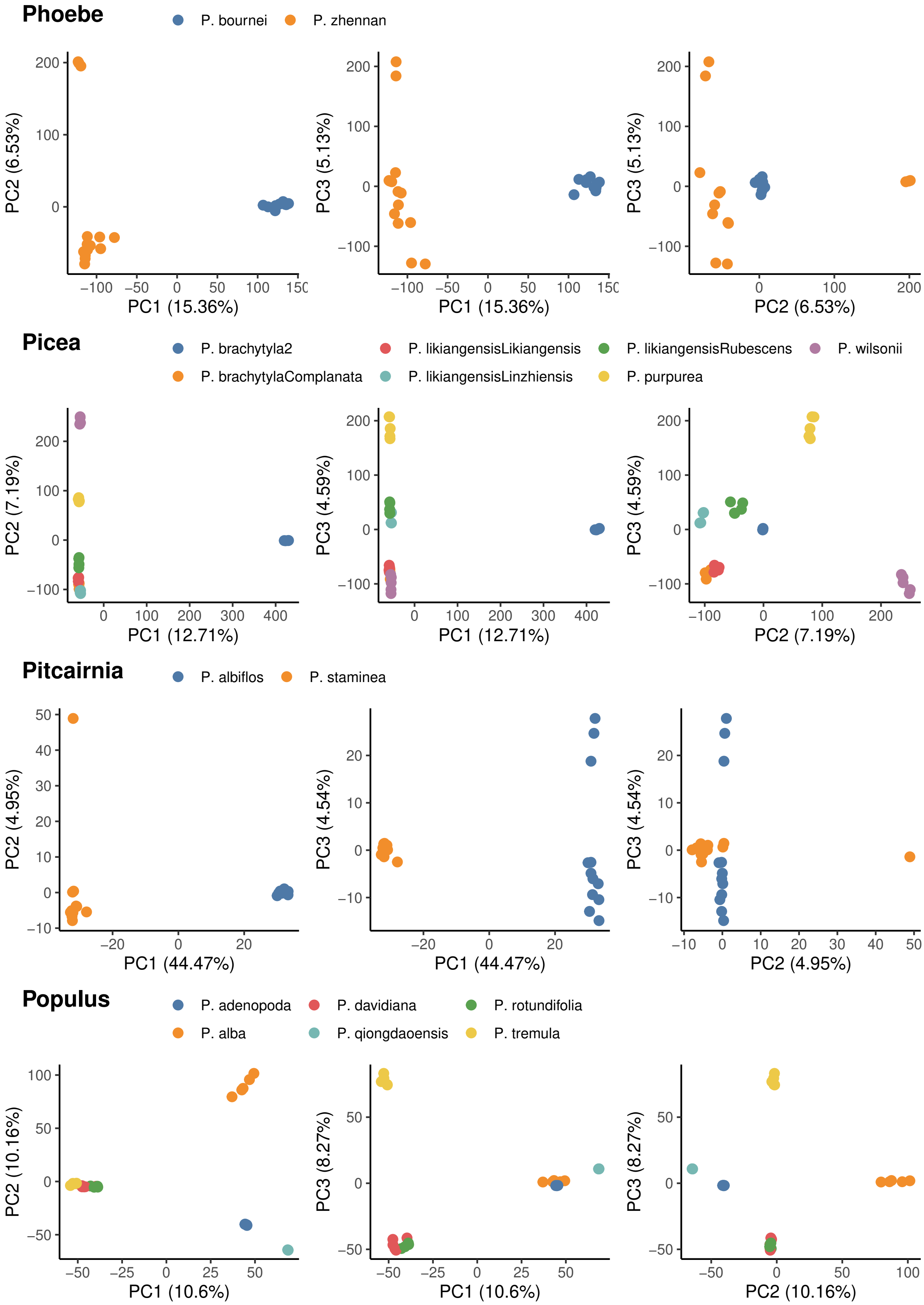


Figure S4: (continued).

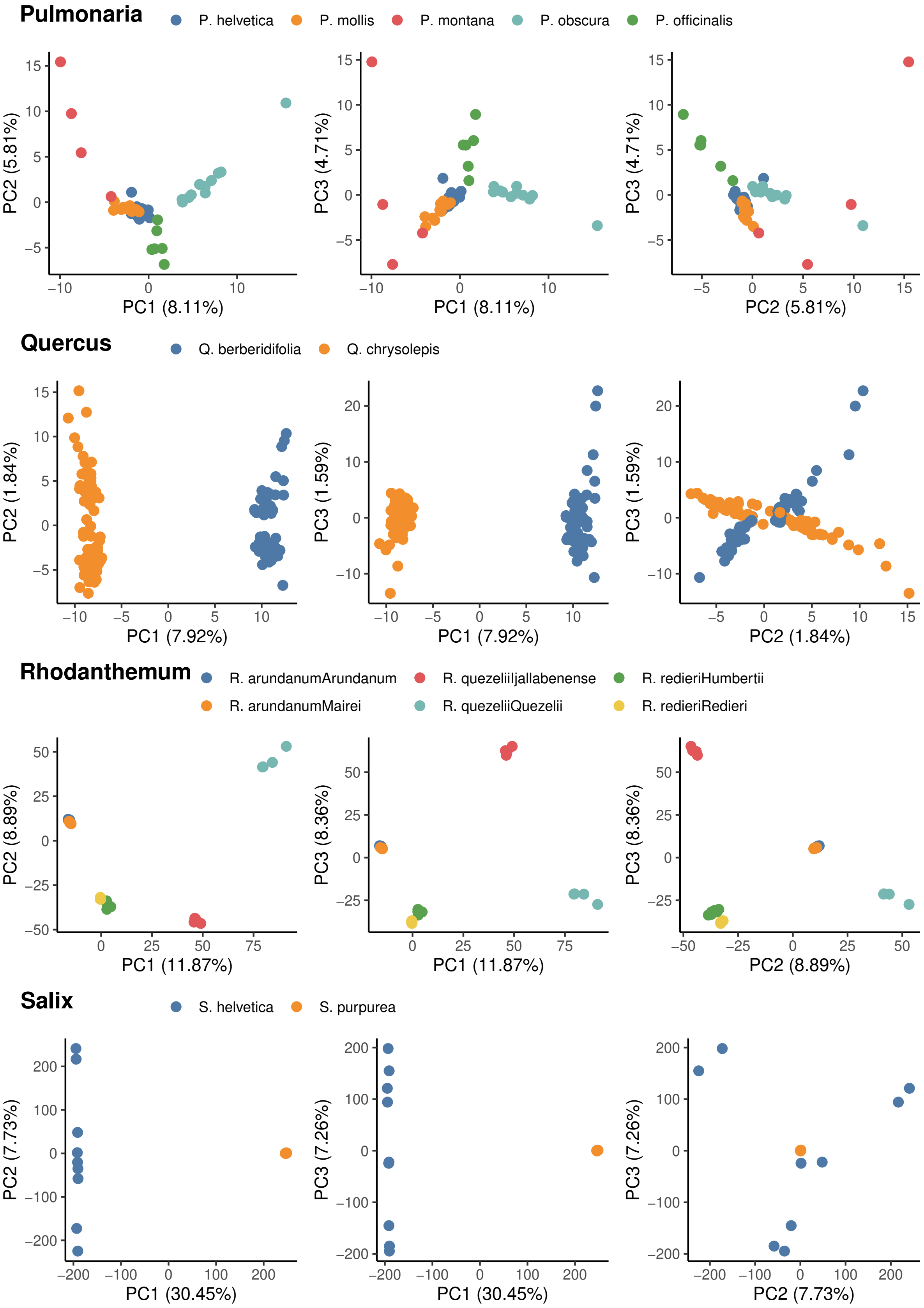


Figure S4: (continued).

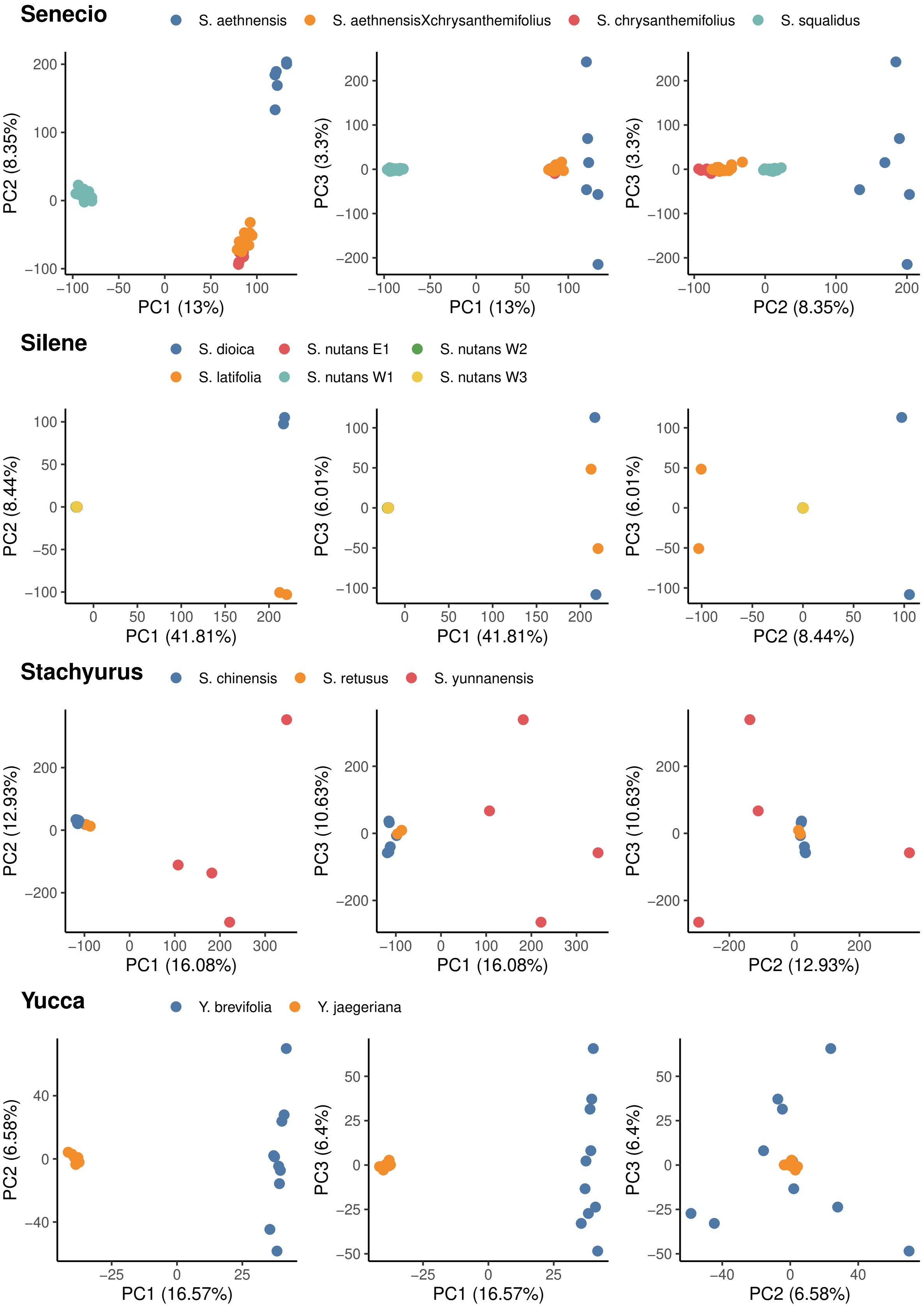


Figure S4: (continued).

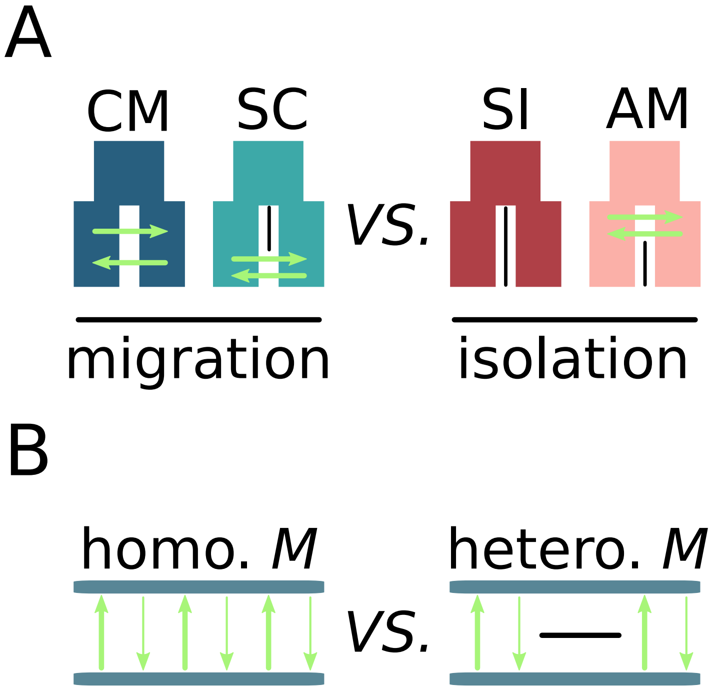


Figure S5: **Compared models using approximate Bayesian computation (ABC).**

A. Models with ongoing migration correspond to all CM (Continuous Migration) and SC (Secondary Contact) models. Models with current isolation correspond to all SI (Strict Isolation) and AM (Ancestral Migration) models. The first step in our ABC classification is to compare the set of CM+SC *versus* SI+AM models in order to assign a migration or isolation status to each of the 341 pairs of lineages (61 animals, 280 plants) according to the computed posterior probability.

B. Pairs of plants or animals, for which our ABC framework has provided strong statistical evidence of ongoing migration, are subsequently subjected to analysis aimed at discerning the uniformity of gene flow across the genome, whether it exhibits homogeneity (characterized by the absence of local genomic barriers) or heterogeneity (signifying genetic linkage to species barriers). The comparison between homo. *M* *versus* hetero. *M* was carried out using the same ABC framework as in the previous step.

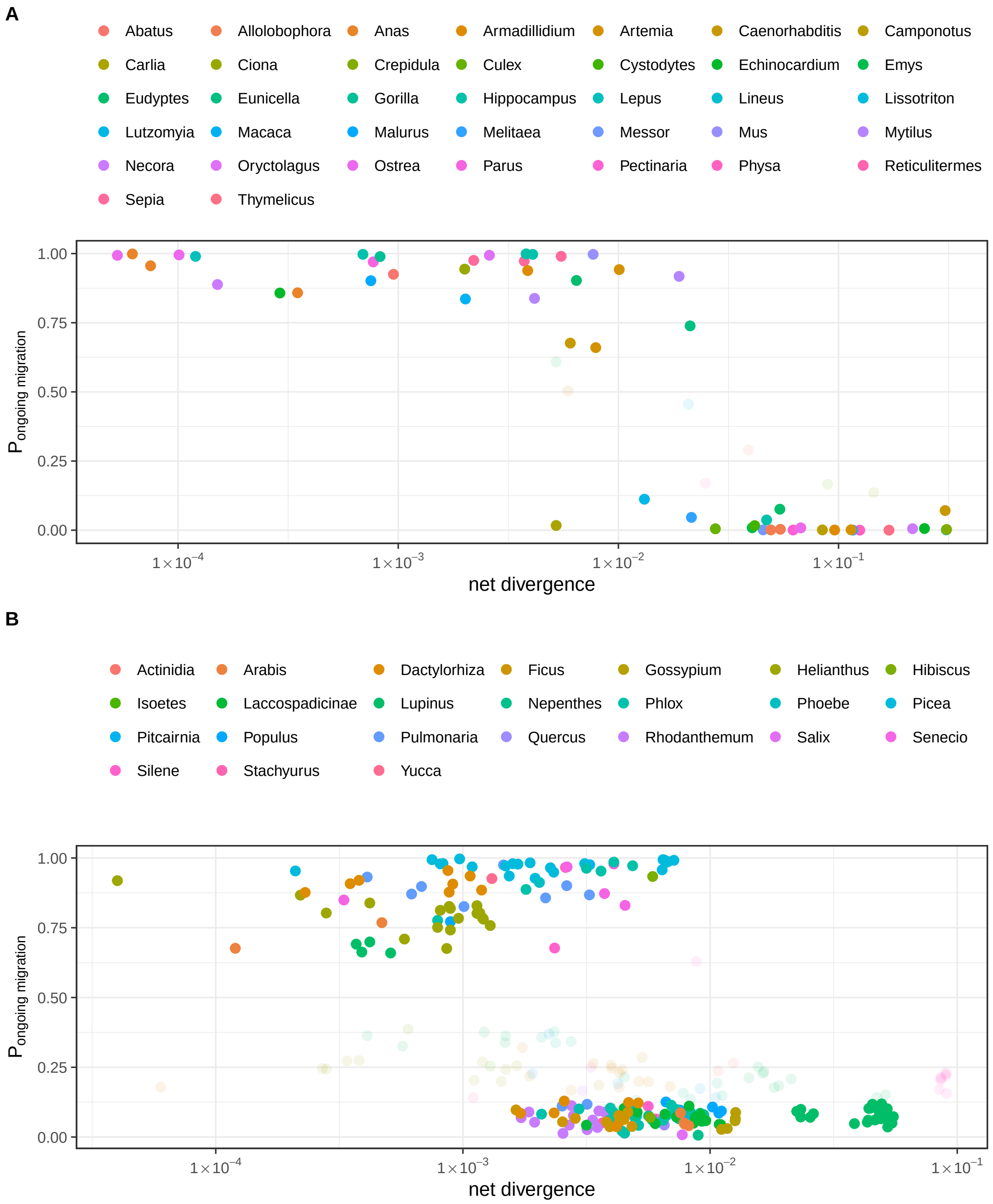


Figure S6: **Relationship between mean net divergence and posterior probability for ongoing migration.**

Each point corresponds to a pair of animals (A) or plants (B). x-axis: average net divergence. y-axis: posterior probability for ongoing migration attributed by our ABC framework. Colours correspond to surveyed genera. Solid points represent pairs for which there is strong statistical evidence either supporting or rejecting the ongoing migration model, as determined by the robustness test outlined in (*6*). In contrast, transparent points indicate pairs for which the comparison between the migration and isolation models yields an inconclusive result. Pairs for which support was inconclusive were excluded from further analysis. The remaining pairs were categorized either as exhibiting ‘migration’ or ‘isolation’, as illustrated in Figure 1-A (see section A.4.4).

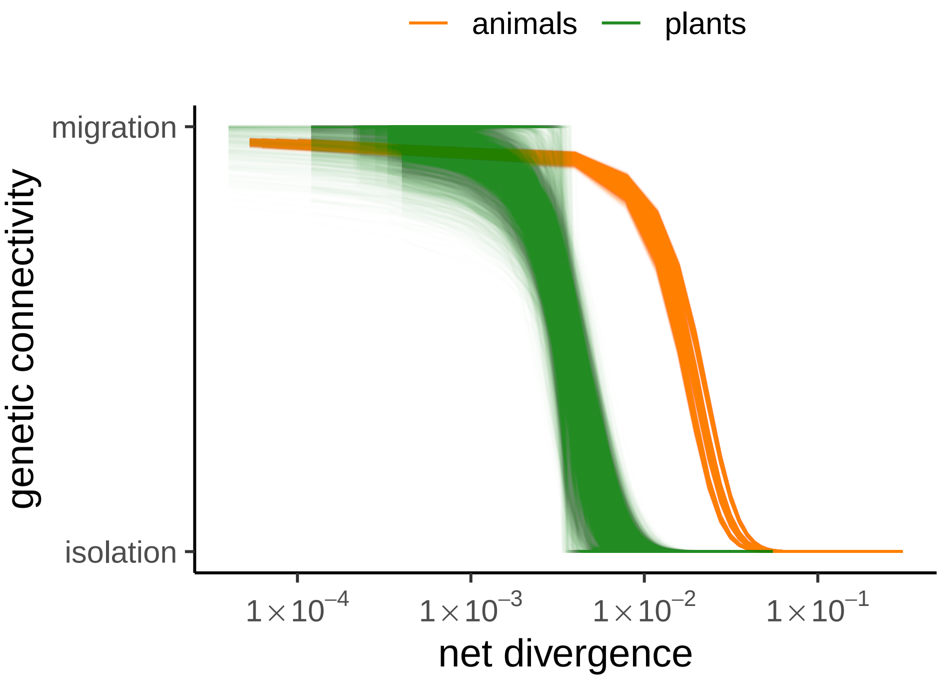


Figure S7: **Relationship between mean net divergence and migration/isolation status controlled by a genus effect.**
The relationship was established as for the entire dataset shown in Figure 1, but by randomly sub-sampling a single pair of populations/species within each plant and animal genus. Each line represents one of the 5,000 random iterations.

C Supplementary Tables

Table S1: List of retained NCBI datasets.

bioproject	**genus**	**species**	**n**	**type of data**	**source**
PRJNA318567	*Actinidia*	*arguta*	3	WGS	(*39*)
		arguta giraldii	2		
		chinensis	4		
PRJEB33482, PRJEB39992	*Arabis*	*nemorensis allop.*	6	RNA	(*40*)
		nemorensis symp.	6		
		sagittata allop.	10		
		sagittata symp.	15		
PRJNA489792	*Dactylorhiza*	*euxina*	5	RAD	(*41*)
		foliosa	2		
		fuchsii	30		
		iberica	2		
		incarnata	31		
		saccifera	4		
		sambucina	3		
		viridis	3		
PRJNA445222	*Ficus*	*arfakensis*	14	RAD	(*42*)
		itoana	13		
		microdictya	15		
		trichocerasa	15		
		t. pleioclada	26		
PRJNA539957	*Gossypium*	*australe*	4	WGS	(*43*)
		bickii	3		
		nelsonii	3		
		robinsonii	2		
		sturtianum	6		
PRJNA532579	*Helianthus*	*annuus NoTex*	15	WGS	(*44*)
		annuus NTex	15		
		annuus STex	15		
		argophyllus	10		
		debilis silvestris	5		
		niveus canescens	8		
		petiolaris fallax	10		
		p. petiolaris	10		
PRJNA382435	*Hibiscus*	*dasycalyx*	6	RAD	(*45*)
		laevis	4		
		moscheutos	5		
PRJNA483403	*Isoetes*	*lacustris*	9	RAD	(*46*)
		echiospora	3		

PRJNA244607	*Howea*	*belmoreana*	40	RNA	(*47*)
		forsteriana	39		
PRJNA528594	*Linospadix*	*monostachyos*	18		(*48*)
		minor	9		
		apetiolatus	6		
		palmerianus	6		
PRJNA318864	*Lupinus*	*ballianus*	2	RNA	(*80*)
		bandelierae	2		
		misticola	2		
PRJEB37794	*Nepenthes*	*albomarginata*	3	RAD	(*49*)
		ampullaria	8		
		bicalcarata	6		
		distillatoria	2		
		dubia	2		
		ephippiata	2		
		gracilis	8		
		hemsleyana	4		
		lamii	2		
		lowii	2		
		macrovulgaris	2		
		madagascariensis	2		
		maxima	10		
		mirabilis	10		
		monticola	2		
		pervillei	16		
		pitopangii	2		
		rafflesiana	9		
		reinwardtiana	2		
		sumatrana	2		
		tentaculata	2		
		veitchii	3		
		vieillardii	2		
PRJNA701424	*Phlox*	*amoena amoena*	48	RAD	(*50*)
		a. lighthipei	14		
		divaricata divaricata	3		
		d. laphamii	3		
		pilosa deamii	15		
		p. fulgida	8		
		p. pilosa	59		
		subulata	2		

PRJNA464259	*Phoebe*	*zhennan*	9	RAD	(*51*)
		bournei	12		
PRJNA807675	*Pitcairnia*	*albiflos*	9	RAD	(*52*)
		staminea	12		
PRJNA392950,	*Picea*	*brachytyla*	4	RNA	(*53, 54*)
PRJNA401149,		*b. complanata*	5		
PRJNA378930,		*likiangensis likiangensis*	5		
PRJNA301093		*l. linzhiensis*	5		
		l. rubescens	5		
		purpurea	5		
		wilsoni	5		
PRJNA612655	*Populus*	*adenopoda*	5	WGS	(*55*)
		alba	5		
		davidiana	5		
		qiongdaoensis	3		
		rotundifolia	4		
		tremula	5		
PRJNA544114	*Pulmonaria*	*helvetica*	24	RAD	(*56*)
		mollis	10		
		montana	4		
		obscur	11		
		officinalis	6		
PRJNA639507	*Quercus*	*berberidifolia*	63	RAD	(*57*)
		chrysolepis	80		
PRJNA554975	*Rhodanthemum*	*redieri redieri*	4	RAD	(*58*)
		r. humbertii	7		
		quezelii quezelii	4		
		q. jallabenense	4		
		arundanum mairei	8		
		a. arundanum	27		
PRJNA429746	*Salix*	*helvetica*	10	RAD	(*59*)
		purpurea	10		
PRJNA549571	*Senecio*	*aethnensis*	6	RNA	(*60*)
		aethn. X chrys.	14		
		chrysanthemifolius	6		
		squalidus	28		
PRJNA295359	*Silene*	*dioica*	2	RNA	(*61, 62*)
		latifolia	2		
		nutans E1	4		
		n. W1	4		

		n. W2	4		
		n. W3	4		
PRJNA553020	*Stachyurus*	*chinensis*	6	RNA	(*63*)
		retusus	2		
		yunnanensis	4		
PRJNA329381	*Yucca*	*brevifolia*	24	RAD	(*64*)
		jaegeriana	39		

Table S2: Log-likelihood Ratio Test for logit models fitted to plant and animal datasets (Fig.1)

model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	**df**	$\boldsymbol{P}$**-value**
M_0	-115.2766	1.736	-433.504	0.004		
M_{plants}	-74.5078	2.517	-799.021	0.003		
$M_{animals}$	-7.808659	3.935	-209.252	0.0188		
					2	4.88×10^{-15}

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{plants} and $M_{animals}$.
$\boldsymbol{\beta_0}$: estimated intercept.
$\boldsymbol{\beta_1}$: estimated coefficient.
$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{plants}) - \ell(M_{animals})|$ in a χ-squared distribution with two degrees of freedom.

Table S3: Log-likelihood Ratio Test for logit models fitted to plant and animal datasets obtained by RNA-sequencing only

model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	**df**	$\boldsymbol{P}$**-value**
M_0	-41.92664	2.413	-320.743	0.007		
M_{plants}	-20.82818	4.031	-766.155	0.005		
$M_{animals}$	-4.361694	5.347	-271.134	0.0197		
					2	5.38×10^{-8}

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{plants} and $M_{animals}$.
$\boldsymbol{\beta_0}$: estimated intercept.
$\boldsymbol{\beta_1}$: estimated coefficient.
$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$).
df: number of degrees of freedom.
P**-value**: probability to observe $2.|\ell(M_0) - \ell(M_{plants}) - \ell(M_{animals})|$ in a χ-squared distribution with two degrees of freedom.

Dynamiques de spéciation: contrastes entre plantes et animaux.

Résumé:
La spéciation, le processus conduisant à l'émergence d'espèces reproductivement isolées par l'accumulation de barrières reproductives génétiques, est étudiée depuis *l'origine des espèces* et reste un sujet de recherche actif. L'un des principaux objectifs de ces études est d'élucider les processus microévolutifs qui façonnent la dynamique de la spéciation. Dans cette thèse, nous avons introduit une nouvelle approche comparative visant à démêler l'effet des facteurs liés à la spéciation. Cette approche est illustrée par l'investigation d'une hypothèse historique : la spéciation supposée plus rapide des animaux par rapport aux plantes. En comparant la dynamique de la spéciation entre les plantes et les animaux, nous avons observé que l'isolement reproductif complet apparaissait, en moyenne, à un niveau de divergence plus faible pour les plantes. Nous avons également analysé la dynamique de la spéciation chez les plantes à l'aide de modèles linéaires, mais nous n'avons pas trouvé d'effets significatifs pour les deux facteurs testés: le taux d'autofécondation et la forme de vie. Dans l'ensemble, ces résultats soulignent le potentiel de notre nouvelle approche comparative pour effectuer des comparaisons faciles, rapides et flexibles de dynamiques de spéciation pour de futures recherches.

Mots clés: Spéciation, Plantes, Génétique des populations, Inférence démographique, Isolement reproducteur, Analyse comparative

Speciation dynamics: contrasts between plants and animals.

Abstract:
Speciation, the process leading to the emergence of reproductively isolated species through the accumulation of genetic reproductive barriers, has been a subject of study since *the origin of species* and remains an active topic of research. One primary goal of these studies is to elucidate which microevolutionary processes shape the dynamics of speciation. In this thesis, we introduced a novel comparative approach aimed at disentangling the effect of several speciation-related factors. This approach is illustrated by an investigation tackling an historical assumption: the supposed faster speciation of animals in contrast to plants. When comparing the dynamics of speciation between plants and animals, we observed that complete reproductive isolation occurred, on average, at a lower level of divergence for plants. We further analysed the dynamics of speciation in plants using linear modelling but did not find any significant effects for the two factors tested: selfing rate and life form. Overall, these results highlight the potential of our novel comparative approach to conduct easy, rapid and flexible comparisons of speciation dynamics in future research.

Keywords: Speciation, Plants, Population genetics, Demographic inference, Reproductive isolation, Comparative analysis

Université de Lille

GHENT UNIVERSITY