Rapid establishment of species barriers in plants compared with that in animals

François Monnet, Zoé Postel, Pascal Touzet, Christelle Fraïsse, Yves van de Peer, Xavier Vekemans, Camille Roux

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. Email: camille.roux@univ-lille.fr

Speciation, the process by which new reproductively isolated species emerge from ancestral populations, results from the gradual accumulation of barriers to gene flow within genomes. To date, the notion that interspecific genetic exchange (introgression) occurs more frequently between plant species than animals has gained a strong footing in scientific discourse. 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 raises questions about which differences in microevolutionary processes between plants and animals influence the dynamics of reproductive isolation establishment at the macroevolutionary scale.

Introduction

Introgression, the genetic exchange between populations or between speciating lineages, has long been recognized as 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 consequence of such introgression events is to facilitate the spread on a large scale (geographical and/or phylogenetic) of mutations that were originally locally beneficial (*3*). However, genetic exchange across the Tree of Life is not unrestricted. It is gradually interrupted by specific mutations, known as species barriers, that reduce hybrid fitness and progressively accumulate in the genome as evolutionary lineages diverge (*4*). These genetic barriers to gene flow contribute to reproductive isolation by either reducing hybrid production or diminishing hybrid fitness. The consequences of reproductive isolation can therefore be captured through the long-term effect of these genetic barriers as a localized reduction of introgression in the genomes, which provides a powerful quantitative metric applicable to any organism (*5*). Thus, the genomes of speciating lineages go through a transitional stage, the so-called 'semi-isolated species', during which they form mosaics of genomic regions: some remain permeable to gene flow while others become tightly linked to barriers (*6*). 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 reproductively isolated?

Historically, reproductive isolation has been studied using complementary approaches. From an 'organismal' perspective, it is evaluated using controlled crosses that measure reductions in hybrid viability or fertility, typically over one or two generations (*5*). In contrast, the 'genetic' approach infers reduction in gene flow between populations over extended evolutionary periods based on the molecular signatures left by reproductive isolation in their genomes (*5*). While organismal studies offer valuable insights into isolating mechanisms and the genomic architecture of reproductive isolation -such as the number of loci involved, their individual phenotypic effects, and the extent of genetic interactions- they have inherent limitations in detecting species barriers as they occur in nature (*7*). To overcome these limitations and investigate the dynamics of reproductive isolation,

we examine patterns of gene flow in natural populations, where introgression leaves detectable genomic signatures. Metrics such as F_{ST} (*8*) and derivatives of the ABBA-BABA test (*9, 10*) are widely employed to assess whether divergence occurred under strict allopatry. However they cannot estimate the timing of gene flow. Recent computational methods have made it possible to explicitly test evolutionary scenarios, including ongoing gene flow, and accounting for the effect of semi-permeable species barriers (*11–13*). Applied to 61 pairs of animal taxa across a continuum of molecular divergence, these methods revealed frequent introgression up to 2% net divergence (*14*) and outstanding cases of gene flow between lineages 14 times more divergent than humans and chimpanzees (*15*).

Whether the dynamics of species barrier formation follow universal patterns across kingdoms remains unknown. In the early 2000s, a landmark study comparing hybridisation rates in plants and animals based on taxonomic records reported higher average frequencies in plants (*16*), in agreement with older assertions on the subject (*2, 17, 18*), but this was contradicted by an analysis using phenotypic rather than taxonomic evidence that found higher crossability in animals (*19*). Since these two studies, few attempts have been made to directly compare plants and animals in this context (*20*). These early approaches were restricted to the occurrence of hybridisation, largely based on morphological species, and lacked information on molecular divergence or introgression. As a result, they could not assess how gene flow decays with molecular divergence, a key component of speciation dynamics (*4*). This limitation can now be overcome using genomic data, which enables the delineation of genetic clusters and the quantification of both divergence and effective gene flow across the genome (*5, 21*). Here, we adopt a comparative genomic approach that integrates datasets across taxa within a unified statistical framework to explicitly test whether the dynamics of gene flow reduction differ between plants and animals at comparable levels of divergence.

Results

The present investigation examines the decrease in ongoing gene flow between closely related lineages as a function of their genetic divergence, and compares this dynamic across two 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 a combination of 39 summary statistics commonly used in population genetics along with the joint Site Frequency Spectrum (jSFS). Analysis of simulated datasets provided evidence that our approach is more powerful than phylogenetic methods for detecting introgression (*22*) (Supplementary Materials A.4). For each dataset, we then applied an approximate Bayesian computation framework (ABC; (*11*)) to assess whether the observed set of summary statistics was better reproduced by scenarios of speciation with or without ongoing migration. The same ABC methodology for demographic inferences was employed for both the animal dataset (analyzed in (*14*)) and the newly compiled plant dataset. The plant dataset was produced from publicly available sequencing reads covering 25 genera spanning the plant phylogeny (212 pairs of eudicots, 45 of monocots, 21 of gymnosperms, 1 lycophyte and 1 magnoliid; Table S1). These pairs were selected without a preconceived idea about their speciation mode (see Supplementary Materials A.2). The posterior probability of models with ongoing migration, as estimated by the ABC framework, was used to classify each pair as either isolated (isolation status) or still exchanging genes (migration status) along a continuum of divergence (Fig. 1-A). This approach allows for a direct comparison of speciation dynamics between plants and animals.

In contrast to qualitative expectations described in earlier studies (*16*), our findings suggest that genetic exchange ceases more rapidly in plants than in animals at lower levels of genetic divergence. This is characterized by a swifter transition in plants from population pairs that are best-supported by migration models to those that are best described by isolation models ($P < 1 \times 10^{-4}$; 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 the probability that two plant lineages are connected by gene flow falls below 50% at a net divergence of $\approx 0.3\%$ (95% CI: [0.26%-0.36%]). In contrast, this inflection point in animals occurs at higher levels of divergence close to 1.5% (95% CI: [1.1%-2.7%]; Fig. 1-A and table S2). This pattern remains robust after accounting for potential methodological and biological biases (Supplementary Materials A.6), including sequencing technology (Table S3), unequal taxonomic sampling (Fig. S1), and geographic distance (Fig. S2). Within plants, speciation dynamics do not appear to be driven by growth form, as no significant difference was detected between trees and herbs using a similar

comparative approach (Fig. S3, Table S4). Likewise, we found no evidence that variation in selfing rates, whether estimated from published methods (*23*), custom approaches (*24*), or approximated from phenotypic data (*25*), influences the dynamics of gene flow cessation within plants across the range of selfing rates represented in our dataset (Supplementary Materials A.7.2, Fig. S4, Table S4).

To further explore the build-up dynamics of species barriers within plant and animal genomes, we focused on species pairs supported by ongoing gene flow and examined the genomic distribution of introgression. Within the range of speciation scenarios considered, the rate of gene flow can be uniform across the genome (i.e., homogeneous) or it can vary locally from one genomic region to another (i.e., heterogeneous; see Fig. S5) when barrier genes establish between the diverging species (*6*). The ABC framework described earlier allows us to classify animal and plant pairs as experiencing either genomically homogeneous or heterogeneous introgression (*26*). We find that plants transition more quickly from no barriers to semi-permeable barriers, with this shift 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, both the onset of the 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, a conclusion consistent with previous experimental crosses unraveling the genomic architecture of reproductive isolation in plants (*27–29*).

Finally, we compared the temporal dynamics of gene flow during divergence in plants *vs* animals. We specifically examine whether ongoing gene flow is more often explained by continuous migration initiated since the subdivision of the ancestral population (as illustrated in Fig. 2), or by secondary contact following an initial period of geographic isolation and divergence. This model comparison using ABC was 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. Among animals, most cases of ongoing gene flow are best explained by models of continuous migration since the subdivision of the ancestral population ($\approx 80\%$; Fig. 2), indicating that gene flow tends to occur more steadily over time. In contrast, plant lineages more

frequently exhibit patterns indicative of secondary contact following an initial phase of allopatry (≈ 52%; Fig. 2). This results aligns with Richard Abbott's synthesis suggesting that secondary contact predominates in plant hybrid zones (*30*).

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*" (*31*). Within this conceptual framework, examinations of numerous wild species based on morphological and taxonomic records have historically supported the assumption that plants are more prone to hybridize than animals (*2, 16, 18*), although this result has been questioned by analyses using phenotypic criteria that found higher crossability in animals (*19*). However, the advent of molecular markers to measure genetic differentiation in the early 2000s provided results challenging morphological studies. Molecular data revealed that plants often show higher F_{ST} values within species than animals (*32, 33*), indicating that gene flow at the intraspecific level is generally more restricted in plants. Moreover, Morjan and Rieseberg (*32*) showed that this difference between kingdoms persists regardless of the mating system (from outcrossing to selfing) as well as across various geographical scales (local, regional, or biregional ranges). Our methodological approach focuses on genetic clusters that vary in both their degrees of pairwise divergence and connectivity through gene flow (Fig. S8). By explicitly modeling the divergence history between lineages and genome-wide heterogeneity of gene flow, we were able to capture the genomic effect of species barriers. This framework help reconcile an apparent contradiction: morphology-based studies of reproductive isolation suggest more frequent hybridization events in plants than in animals (*16*) (although this view was challenged by Rieseberg et al. (*19*)), while molecular data reveal greater genetic differentiation within plant species (*32*). 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 (*30*), relative to closely related animal species that tend to experience gene flow more continuously over time. Secondary contact scenarios imply an initial period of allopatry, which can drive divergence of sister lineages across molecular and morphological markers. 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 due to prior isolation. Conversely, introgression in animals may often be overlooked, particularly if it occurs more frequently between species lacking clear morphological differences compared to plants. However, we strongly emphasize that our findings should not be interpreted as evidence that speciation in animals frequently occurs in the presence of gene flow. Rather, the absence of ongoing gene flow inferred in our study serves as a proxy for reproductive isolation, reflecting the effect of barriers that, on average, accumulate more slowly across animal genomes.

The long-standing hypothesis that plants may experience more frequent hybridization at the population/species level than animals has also been difficult to reconcile with observations from both micro- and macroevolutionary scales:

i) At the experimental level, quantitative genetic studies suggest potential differences in the genomic architecture of reproductive isolation between plants and animals. In plants, reproductive barriers may often involve relatively simpler genetic architectures, typically governed by a small number of loci with large effects (*27–29, 34*). Moreover, it has been proposed that speciation genes may be subject to different evolutionary constraints in plants and animals (*35*). The growing number of speciation genes identified across diverse clades (*36*) will facilitate future empirical tests of whether average effect sizes differ between plants and animals, and whether speciation genes indeed evolve under weaker selective constraints in plants. In parallel with these studies, measurements of postzygotic isolation in interspecific crosses indicate that a greater proportion of phenotypically defined species correspond to reproductively independent lineages in plants (around 70%) than in animals (approximately 40%) (*19*).

ii) At the phylogenetic scale, estimates of diversification rates suggest that plant lineages, on average, give rise to new lineages more rapidly than both vertebrates and invertebrates, as reflected in their higher estimated speciation rates (*37, 38*).

The observed contrast in the accumulation of reproductive isolation between plants and animals may reflect underlying differences in the genetic architecture of reproductive barriers, and/or differences in selective constraints acting on these barriers. Although our analysis focuses on population/species

level processes, it raises the possibility that such microevolutionary differences could partly underlie the higher speciation rates often inferred in plants. Clarifying this connection will require future efforts combining comparative population genomics and macroevolutionary analyses.

Speciation dynamics clearly do not follow a universal molecular clock, although certain molecular constraints inevitably make the process irreversible once a critical level of divergence is reached (*4, 16*). The rarity of hybrid zones in plants has recently received renewed attention (*30*), raising a new question: what genomic properties of plants facilitate the emergence of reproductive isolation between lineages after relatively few genetic changes? The evolution of reproductive barriers is shaped by ecological, genetic, geographical, and cytological factors, among others. Although our comparative population genomics approach reveals a striking contrast between plants and animals, it cannot alone unravel the underlying mechanisms driving this difference. Advancing our understanding of why reproductive isolation appears to accumulate more rapidly in plants will require integrative approaches, combining genomic, ecological, and developmental insights across diverse biological systems and evolutionary timescales (*39, 40*).

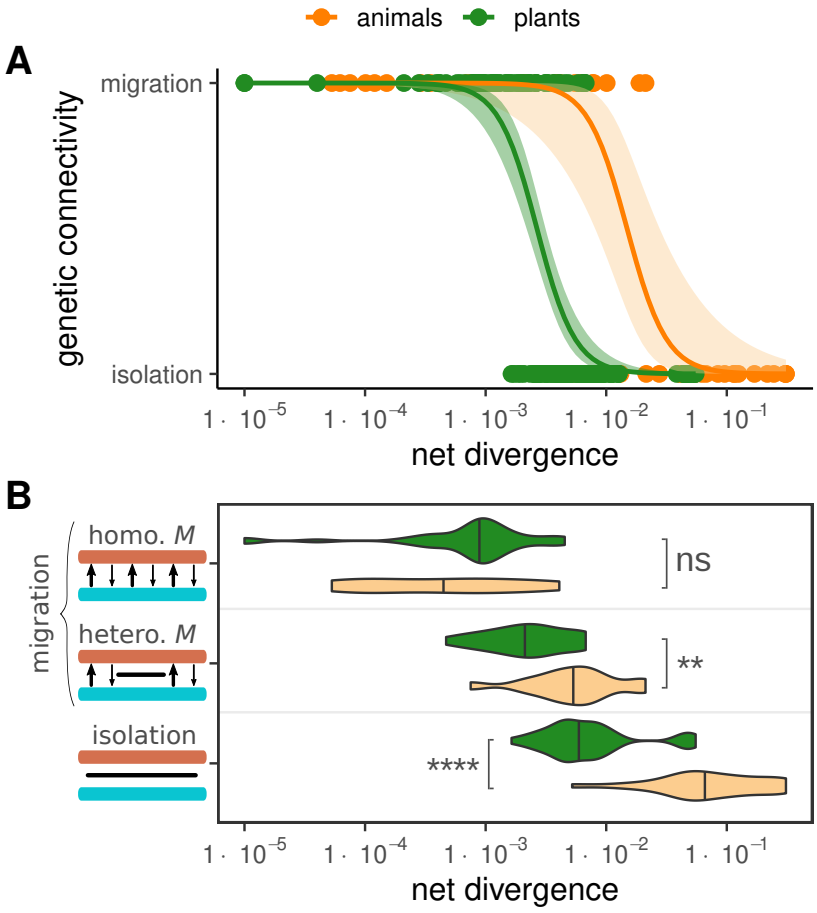

Figure 1: **Genomic patterns of introgression along a divergence continuum in plants *versus* animals.** Average genomic divergence and migration/isolation status were estimated for 280 plant pairs (green) and 61 animal pairs (orange) using ABC (*11, 14*). (**A**) Each point represents a plant (green) or animal (orange) pair, plotted according to its average genomic net divergence on the x-axis and its best-supported model on the y-axis (ongoing migration *versus* isolation). Curves show logit regressions fitted separately for plants and animals. (**B**) Net divergence distributions for plant (green) and animal (orange) pairs under homogeneous (homo. M), heterogeneous (hetero. M), or isolation models. Bars along the y-axis: homologous chromosomes (blue, brown), gene flow (arrows), or barriers (black bars). Statistical significance (Mann–Whitney U test): *ns* ($P > 0.05$), * ($P \leq 0.05$), ** ($P \leq 0.01$), *** ($P \leq 0.001$), **** ($P \leq 0.0001$).

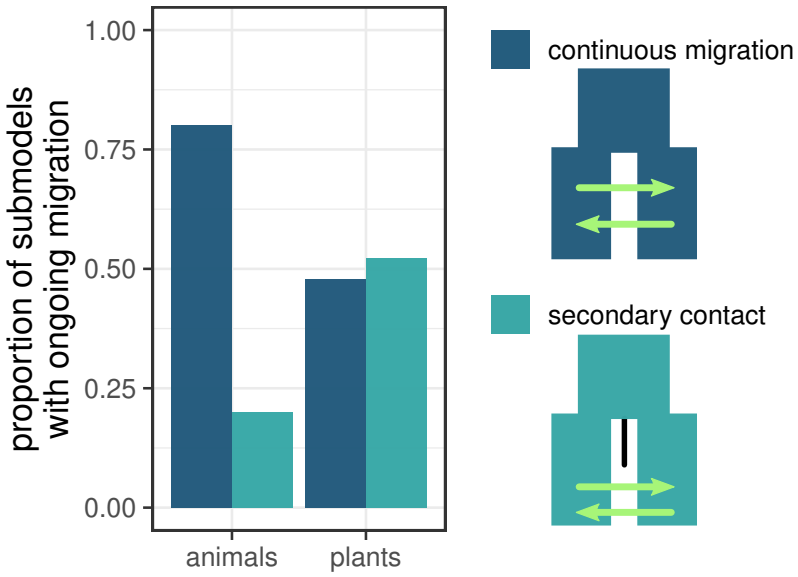

Figure 2: **Temporal models of ongoing migration in plants and animals.** Among the 69 plant pairs and 30 animal pairs that were strongly supported under models with ongoing gene flow by our ABC approach, we further distinguished between alternative temporal models. For each of these supported pairs, we compared sub-models of continuous migration (dark blue) and secondary contact (light blue), and report their relative proportions within each kingdom.

References and Notes

1. R. Abbott, *et al.*, Hybridization and speciation. *Journal of evolutionary biology* **26** (2), 229–246 (2013).

2. G. L. Stebbins, The role of hybridization in evolution. *Proceedings of the American Philosophical Society* **103** (2), 231–251 (1959).

3. M. Slatkin, The rate of spread of an advantageous allele in a subdivided population, in *Population genetics and ecology* (Elsevier), pp. 767–780 (1976).

4. M. R. Brown, *et al.*, Genetic factors predict hybrid formation in the British flora. *Proceedings of the National Academy of Sciences* **120** (16), e2220261120 (2023).

5. A. M. Westram, S. Stankowski, P. Surendranadh, N. Barton, What is reproductive isolation? *Journal of evolutionary biology* **35** (9), 1143–1164 (2022).

6. C.-I. Wu, The genic view of the process of speciation. *Journal of evolutionary biology* **14** (6), 851–865 (2001).

7. M. E. Frayer, B. A. Payseur, Do genetic loci that cause reproductive isolation in the lab inhibit gene flow in nature? *Evolution* **78** (6), 1025–1038 (2024).

8. S. Wright, The genetical structure of populations. *Annals of eugenics* **15** (1), 323–354 (1949).

9. S. H. Martin, J. W. Davey, C. D. Jiggins, Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular biology and evolution* **32** (1), 244–257 (2015).

10. A. J. Dagilis, *et al.*, A need for standardized reporting of introgression: Insights from studies across eukaryotes. *Evolution Letters* **6** (5), 344–357 (2022).

11. C. Fraïsse, *et al.*, DILS: Demographic inferences with linked selection by using ABC. *Molecular Ecology Resources* **21** (8), 2629–2644 (2021).

12. D. R. Laetsch, *et al.*, Demographically explicit scans for barriers to gene flow using gIMble. *PLoS genetics* **19** (10), e1010999 (2023).

13. V. C. Sousa, M. Carneiro, N. Ferrand, J. Hey, Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics* **194** (1), 211–233 (2013).

14. C. Roux, *et al.*, Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS biology* **14** (12), e2000234 (2016).

15. C. Fraïsse, *et al.*, Introgression between highly divergent sea squirt genomes: an adaptive breakthrough? *Peer Community Journal* **2** (2022).

16. J. Mallet, Hybridization as an invasion of the genome. *Trends in ecology & evolution* **20** (5), 229–237 (2005).

17. L. Gottlieb, Genetics and morphological evolution in plants. *The American Naturalist* **123** (5), 681–709 (1984).

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

19. L. H. Rieseberg, T. E. Wood, E. J. Baack, The nature of plant species. *Nature* **440** (7083), 524–527 (2006).

20. S. Wang, J. E. Mank, D. Ortiz-Barrientos, L. H. Rieseberg, Genome architecture and speciation in plants and animals. *Molecular Ecology* p. e70004 (2025).

21. N. Galtier, Delineating species in the speciation continuum: A proposal. *Evolutionary applications* **12** (4), 657–663 (2019).

22. N. B. Edelman, *et al.*, Genomic architecture and introgression shape a butterfly radiation. *Science* **366** (6465), 594–599 (2019).

23. M. A. Stoffel, *et al.*, inbreedR: an R package for the analysis of inbreeding based on genetic markers. *Methods in Ecology and Evolution* **7** (11), 1331–1339 (2016).

24. C. Roux, popgenomics/selfing_ML: selfing_ML, Zenodo (2025), doi:10.5281/zenodo.15296403, `https://doi.org/10.5281/zenodo.15296403`.

25. A. J. Helmstetter, *et al.*, Pollination and mating traits underlie diverse reproductive strategies in flowering plants. *bioRxiv* pp. 2024–02 (2024).

26. C. Roux, G. Tsagkogeorga, N. Bierne, N. Galtier, Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. *Molecular biology and evolution* **30** (7), 1574–1587 (2013).

27. L. C. Moyle, E. B. Graham, Genetics of hybrid incompatibility between Lycopersicon esculentum and L. hirsutum. *Genetics* **169** (1), 355–373 (2005).

28. L. C. Moyle, T. Nakazato, Comparative genetics of hybrid incompatibility: sterility in two Solanum species crosses. *Genetics* **179** (3), 1437–1453 (2008).

29. A. L. Sweigart, L. Fishman, J. H. Willis, A simple genetic incompatibility causes hybrid male sterility in Mimulus. *Genetics* **172** (4), 2465–2479 (2006).

30. R. J. Abbott, Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *Journal of Systematics and Evolution* **55** (4), 238–258 (2017).

31. R. G. Harrison, *et al.*, Hybrid zones: windows on evolutionary process. *Oxford surveys in evolutionary biology* **7**, 69–128 (1990).

32. C. L. Morjan, L. H. Rieseberg, How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular ecology* **13** (6), 1341–1356 (2004).

33. R. Frankham, C. J. Bradshaw, B. W. Brook, Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation* **170**, 56–63 (2014).

34. M. P. Zuellig, A. L. Sweigart, A two-locus hybrid incompatibility is widespread, polymorphic, and active in natural populations of Mimulus. *Evolution* **72** (11), 2394–2405 (2018).

35. L. H. Rieseberg, B. K. Blackman, Speciation genes in plants. *Annals of botany* **106** (3), 439–455 (2010).

36. M. E. Frayer, N. V. Robles, M. J. R. Barrera, J. M. Coughlan, M. Schumer, The molecular evolutionary basis of species formation revisited. *EcoEvoRxiv* (2025).

37. J. P. Scholl, J. J. Wiens, Diversification rates and species richness across the Tree of Life. *Proceedings of the Royal Society B: Biological Sciences* **283** (1838), 20161334 (2016).

38. H. Morlon, *et al.*, Phylogenetic Insights into Diversification. *Annual Review of Ecology, Evolution, and Systematics* **55** (2024).

39. D. I. Bolnick, *et al.*, A multivariate view of the speciation continuum. *Evolution* **77** (1), 318–328 (2023).

40. S. Stankowski, *et al.*, Toward the integration of speciation research. *Evolutionary Journal of the Linnean Society* **3** (1), kzae001 (2024).

41. F. Monnet, *et al.*, Rapid establishment of species barriers in plants compared to animals, Zenodo (2025), doi:10.5281/zenodo.15288584, `https://doi.org/10.5281/zenodo.15288584`.

42. F. Monnet, Ladarwall/Greenworld: Main release (2025), doi:10.5281/zenodo.15881247, `https://doi.org/10.5281/zenodo.15881247`.

43. Y. Liu, *et al.*, Rapid radiations of both kiwifruit hybrid lineages and their parents shed light on a two-layer mode of species diversification. *New Phytologist* **215** (2), 877–890 (2017).

44. H. Dittberner, A. Tellier, J. de Meaux, Approximate Bayesian computation untangles signatures of contemporary and historical hybridization between two endangered species. *Molecular Biology and Evolution* **39** (2), msac015 (2022).

45. M. K. Brandrud, *et al.*, Phylogenomic relationships of diploids and the origins of allotetraploids in Dactylorhiza (Orchidaceae). *Systematic Biology* **69** (1), 91–109 (2020).

46. D. Souto-Vilarós, *et al.*, Pollination along an elevational gradient mediated both by floral scent and pollinator compatibility in the fig and fig-wasp mutualism. *Journal of Ecology* **106** (6), 2256–2273 (2018).

47. C. E. Grover, *et al.*, Dual domestication, diversity, and differential introgression in Old World cotton diploids. *Genome Biology and Evolution* **14** (12), evac170 (2022).

48. G. L. Owens, *et al.*, Standing variation rather than recent adaptive introgression probably underlies differentiation of the texanus subspecies of Helianthus annuus. *Molecular Ecology* **30** (23), 6229–6245 (2021).

49. J. Norrell, Differentiating the Neches River Rose Mallow (hibiscus Dasycalyx) from Its Congeners by Means of Phylogenetics and Population Genetics (2017).

50. D. P. Wood, J. K. Olofsson, S. W. McKenzie, L. T. Dunning, Contrasting phylogeographic structures between freshwater lycopods and angiosperms in the British Isles. *Botany Letters* **165** (3-4), 476–486 (2018).

51. L. Dunning, *et al.*, Ecological speciation in sympatric palms: 1. Gene expression, selection and pleiotropy. *Journal of evolutionary biology* **29** (8), 1472–1487 (2016).

52. O. G. Osborne, *et al.*, Speciation in Howea palms occurred in sympatry, was preceded by ancestral admixture, and was associated with edaphic and phenological adaptation. *Molecular Biology and Evolution* **36** (12), 2682–2697 (2019).

53. M. Scharmann, A. Wistuba, A. Widmer, Introgression is widespread in the radiation of carnivorous Nepenthes pitcher plants. *Molecular Phylogenetics and Evolution* **163**, 107214 (2021).

54. B. E. Goulet-Scott, A. G. Garner, R. Hopkins, Genomic analyses overturn two long-standing homoploid hybrid speciation hypotheses. *Evolution* **75** (7), 1699–1710 (2021).

55. X. Ding, J. H. Xiao, L. Li, J. G. Conran, J. Li, Congruent species delimitation of two controversial gold-thread nanmu tree species based on morphological and restriction site-associated DNA sequencing data. *Journal of Systematics and Evolution* **57** (3), 234–246 (2019).

56. M. M. Tavares, M. Ferro, B. S. S. Leal, C. Palma-Silva, Speciation with gene flow between two Neotropical sympatric species (Pitcairnia spp.: Bromeliaceae). *Ecology and Evolution* **12** (5), e8834 (2022).

57. Y. Sun, *et al.*, Reticulate evolution within a spruce (Picea) species complex revealed by population genomic analysis. *Evolution* **72** (12), 2669–2681 (2018).

58. D. Ru, *et al.*, Population genomic analysis reveals that homoploid hybrid speciation can be a lengthy process. *Molecular Ecology* **27** (23), 4875–4887 (2018).

59. H. Shang, *et al.*, Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group. *Philosophical Transactions of the Royal Society B* **375** (1806), 20190544 (2020).

60. S. Grünig, M. Fischer, C. Parisod, Recent hybrid speciation at the origin of the narrow endemic Pulmonaria helvetica. *Annals of botany* **127** (1), 21–31 (2021).

61. J. Ortego, L. L. Knowles, Incorporating interspecific interactions into phylogeographic models: A case study with Californian oaks. *Molecular Ecology* **29** (23), 4510–4524 (2020).

62. F. Wagner, *et al.*, Taming the Red Bastards: Hybridisation and species delimitation in the Rhodanthemum arundanum-group (Compositae, Anthemideae). *Molecular phylogenetics and evolution* **144**, 106702 (2020).

63. S. Gramlich, N. D. Wagner, E. Hörandl, RAD-seq reveals genetic structure of the F 2-generation of natural willow hybrids (Salix L.) and a great potential for interspecific introgression. *BMC Plant Biology* **18**, 1–12 (2018).

64. B. Nevado, S. A. Harris, M. A. Beaumont, S. J. Hiscock, Rapid homoploid hybrid speciation in British gardens: the origin of Oxford ragwort (Senecio squalidus). *Molecular Ecology* **29** (21), 4221–4233 (2020).

65. X.-S. Hu, D. A. Filatov, The large-X effect in plants: increased species divergence and reduced gene flow on the Silene X-chromosome. *Molecular ecology* **25** (11), 2609–2619 (2016).

66. A. Muyle, *et al.*, Dioecy is associated with high genetic diversity and adaptation rates in the plant genus Silene. *Molecular Biology and Evolution* **38** (3), 805–818 (2021).

67. Y. Feng, H. P. Comes, Y.-X. Qiu, Phylogenomic insights into the temporal-spatial divergence history, evolution of leaf habit and hybridization in Stachyurus (Stachyuraceae). *Molecular phylogenetics and evolution* **150**, 106878 (2020).

68. A. M. Royer, M. A. Streisfeld, C. I. Smith, Population genomics of divergence within an obligate pollination mutualism: Selection maintains differences between Joshua tree species. *American journal of botany* **103** (10), 1730–1741 (2016).

69. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature methods* **9** (4), 357–359 (2012).

70. N. Matasci, *et al.*, Data access for the 1,000 Plants (1KP) project. *Gigascience* **3** (1), 2047–217X (2014).

71. J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes— genomes— genetics* **1** (3), 171–182 (2011).

72. J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, W. A. Cresko, Stacks: an analysis tool set for population genomics. *Molecular ecology* **22** (11), 3124–3140 (2013).

73. J. R. Paris, J. R. Stevens, J. M. Catchen, Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8** (10), 1360–1373 (2017).

74. L. Excoffier, *et al.*, fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* **37** (24), 4882–4885 (2021).

75. R. Gutenkunst, R. Hernandez, S. Williamson, C. Bustamante, Diffusion approximations for demographic inference: DaDi. *Nature precedings* pp. 1–1 (2010).

76. J. Jouganous, W. Long, A. P. Ragsdale, S. Gravel, Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics* **206** (3), 1549–1567 (2017).

77. M. Malinsky, M. Matschiner, H. Svardal, Dsuite-Fast D-statistics and related admixture evidence from VCF files. *Molecular ecology resources* **21** (2), 584–595 (2021).

78. T. E. Cruickshank, M. W. Hahn, Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular ecology* **23** (13), 3133–3157 (2014).

79. C. Roux, *et al.*, Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a Mytilus hybrid zone. *Journal of Evolutionary Biology* **27** (8), 1662–1675 (2014).

80. T. Leroy, *et al.*, Extensive recent secondary contacts between four European white oak species. *New Phytologist* **214** (2), 865–878 (2017).

81. T. Capblancq, *et al.*, Untangling the contribution of adaptive versus non-adaptive processes in the evolution of reproductive isolation between Coenonympha butterflies. *bioRxiv* pp. 2024–11 (2024).

82. T. Koppetsch, M. Malinsky, M. Matschiner, Towards reliable detection of introgression in the presence of among-species rate variation. *Systematic biology* **73** (5), 769–788 (2024).

83. N. Galtier, An approximate likelihood method reveals ancient gene flow between human, chimpanzee and gorilla. *Peer Community Journal* **4** (2024).

84. J. Ross-Ibarra, *et al.*, Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata. *PloS one* **3** (6), e2411 (2008).

85. R. R. Hudson, Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18** (2), 337–338 (2002).

86. B. Charlesworth, M. Morgan, D. Charlesworth, The effect of deleterious mutations on neutral molecular variation. *Genetics* **134** (4), 1289–1303 (1993).

87. N. Barton, B. O. Bengtsson, The barrier to genetic exchange between hybridising populations. *Heredity* **57** (3), 357–376 (1986).

88. M. A. Beaumont, Approximate Bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics* **41**, 379–406 (2010).

89. F. Tajima, Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105** (2), 437–460 (1983).

90. G. Watterson, On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7** (2), 256–276 (1975).

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

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

93. S. Wright, Isolation by distance. *Genetics* **28** (2), 114 (1943).

94. J. Wakeley, J. Hey, Estimating ancestral population parameters. *Genetics* **145** (3), 847–855 (1997).

95. J. L. Hamrick, J. D. Nason, Gene flow in forest trees. *Forest conservation genetics: Principles and practice* pp. 81–90 (2000).

96. F. Austerlitz, S. Mariette, N. Machon, P.-H. Gouyon, B. Godelle, Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics* **154** (3), 1309–1321 (2000).

97. B. Igic, J. R. Kohn, The distribution of plant mating systems: study bias against obligately outcrossing species. *Evolution* **60** (5), 1098–1103 (2006).

98. B. Anderson, *et al.*, Opposing effects of plant traits on diversification. *Iscience* **26** (4) (2023).

99. J. A. Hamlin, M. S. Hibbins, L. C. Moyle, Assessing biological factors affecting postspeciation introgression. *Evolution letters* **4** (2), 137–154 (2020).

100. J. Goudet, Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular ecology notes* **5** (1), 184–186 (2005).

101. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution* **34** (7), 1812–1819 (2017).

102. B. Nevado, G. W. Atchison, C. E. Hughes, D. A. Filatov, Widespread adaptive evolution during repeated evolutionary radiations in New World lupins. *Nature communications* **7** (1), 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, Sylvain Glémin, John Pannell and Martin Lascoux for their constructive feedback.

Funding: This work was supported by University of Lille (grant I-SITE ULNE – Ghent University). The French State under the France-2030 programme and the Initiative of Excellence of the University of Lille are acknowledged for the funding and support granted to the R-CDP-24-002-PIE project. 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).

Author contributions: F.M (data curation, formal analysis, investigation, visualization, writing), Z.P (data curation), P.T (resources, data curation, supervision), C.F (methodology, writing), Y.V.d.P (funding acquisition, supervision, writing), X.V (funding acquisition, supervision, writing), C.R (funding acquisition, conceptualization, methodology, supervision, coding, investigation, visualization, writing).

Competing interests: There are no competing interests to declare.

Data and materials availability: All data described in the manuscript or the supplementary materials are available. The assembled datasets, the list of references used for mapping and the results of demographic inference are deposited in Zenodo (*41*). Scripts for bioinformatic treatment of raw sequencing data and estimate of selfing rates are available from Zenodo (*24, 42*).

Supplementary materials

Materials and Methods

Figs. S1 to S14

Tables S1 to S4

References *(43-102)*

Supplementary Materials for

Rapid establishment of species barriers in plants compared to animals

François Monnet, Zoé Postel, Pascal Touzet, Christelle Fraïsse, Yves Van de Peer,

Xavier Vekemans, Camille Roux

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

This PDF file includes:

Materials and Methods

Figures S1 to S14

Tables S1 to S4

Table of Contents for Supplementary Materials

Materials and Methods

Animal dataset

The animal data come from the Roux et *al.* (2016) study (*14*). 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.

Plant dataset

Raw data used in this work comes from previously published studies (*43–68*). The following criteria were applied to identify datasets in plants:

i) Currently diploid genomes.

ii) High-throughput sequencing, i.e, RNA-seq, RAD-seq or whole genome sequencing (WGS).

iii) Freely available from `NCBI`.

iv) Individuals sampled from natural populations (geographic distribution represented in Fig. S2).

v) A minimum of two sampled populations/species per genus.

vi) 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.

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):

Reads from RNA-seq and WGS.

In line with the animal dataset (*14*), 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 (*69*): either taken from the `1KP project` (*70*) 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.

Reads from RAD-seq.

Loci were assembled for each RAD-seq dataset using Stacks 2.6 (*71, 72*). Combinations of parameters were explored following Paris et *al.* 2017 (*73*) 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.

Demographic inferences

Choice of the inferential method

Several methods exist to test introgression between evolutionary lineages, which can broadly be categorized into population genomic approaches (e.g., FastSimCoal (*74*), $\partial a \partial i$ (*75*), Moments (*76*), DILS (*11*)) and phylogenetic methods (e.g., QuIBL (*22*), Dsuite (*77*)). All these methods are effective within the contexts for which they were designed. The goal of this section is not to demonstrate the superiority of one approach over another but to justify why DILS was chosen as a relevant tool for addressing the specific question posed in this study: whether there is ongoing gene flow between two populations or whether they are currently isolated. The choice of method also depends on the properties of the sampling (i.e., number of sampled individuals, number of sampled populations in the genus, assumptions about demography for certain population pairs, etc.) as well as on the properties of the molecular data (i.e., locus length, presence or absence of intra-locus recombination, etc.). Finally, we also aimed to reduce biases that arise when linked selection is not taken into account, whether from background selection (*78*) or selection against species barriers (*79*). In the next section, we focus on a comparison between QuIBL and DILS because the literature has shown that FastSimCoal and $\partial a \partial i$ produce results largely comparable to DILS when applied to the same datasets (*80, 81*).

Key methodological differences between DILS and QuIBL

In a given phylogenetic tree with more than three lineages (species or populations), QuIBL extracts information from individual gene trees and tests, for a given triplet of lineages, whether the distribution of internal branch lengths can be explained solely by incomplete lineage sorting or whether introgression processes must also be invoked. For a tree of the topology $((1, 2), 3)$, QuIBL is particularly powerful in detecting deviations from strict allopatry between lineages 1 (or 2) and 3. Such deviations are interpreted as evidence of historical introgression events, which may vary in

recency.

In contrast, DILS leverages population genomic summary statistics, particularly those derived from the Site Frequency Spectrum (SFS), to evaluate whether observed patterns are better explained by a model with ongoing migration or one without, using an Approximate Bayesian Computation (ABC) framework. Given that the question addressed here concerns the extent of current reproductive isolation between specific pairs of lineages (in plants and animals), it was important to discriminate between different temporal patterns of introgression. Deviations from strict allopatry caused by ancient migration do not convey the same implications for current reproductive isolation as those caused by secondary contact. This distinction is explicitly addressed by DILS through the analysis of intra-specific polymorphism and interspecific divergence data for a given pair of lineages.

Rationale for Choosing DILS

Our decision to use DILS was motivated by the following considerations:

- DILS accounts for variations in effective population size over time, whereas phylogenetic approaches assume a constant N_e.
- DILS accounts for genome-wide variations in effective population size caused by background selection, as well as variations in effective migration rates driven by linked selection against species barriers.
- DILS accounts for intra-locus recombination, while the distribution of internal branch lengths is derived under the assumption of a simple tree per locus, with no recombination occurring within individual loci.
- DILS does not require a specific sampling scheme to test introgression, whereas phylogenetic methods cannot detect gene flow between lineages 1 and 2 in a tree of the topology $((1, 2), 3)$.
- DILS works directly with measurable quantities derived from sequences (e.g., F_{ST}, θ_W, π, D_a, D_{xy}, etc.) without requiring a preliminary step of inferring phylogenetic trees for each locus.

- DILS can handle multiple individuals within each species, reducing the risk of overlooking shared polymorphisms between species.
- DILS does not require long loci for accurate inference, making it less sensitive to biases introduced by RADseq or RNAseq data. In contrast, phylogenetic methods may perform poorly with short loci and are subject to violations of no-recombination assumptions with long loci (*82, 83*). In addition, very long loci may buffer local genomic variations for N_e and $N_e.m$.
- DILS works directly on pairs of lineages and does not require data from more than three species, allowing its application to genera where only two species have been sequenced.

Comparison between DILS and QuIBL on simulated data

To further illustrate the relative strengths of these approaches, we conducted a comparative analysis of DILS and QuIBL using coalescent simulations under a four populations model (Fig. S9). Data were simulated under various temporal patterns of introgression using `msnsam` (*84*), a modified version of `ms` (*85*). The simulations were designed to satisfy all assumptions made by QuIBL (e.g., long loci, no intra-locus recombination, no migration between lineages 1 and 2 in a tree of the topology $((1, 2), 3)$). For QuIBL, the input consisted of exact, simulated gene trees rather than inferred trees, as would be the case with real biological data.

Simulated Scenarios Four demographic scenarios were simulated (Fig. S9):

- **Strict Isolation ($\mathbf{SI_{4pop}}$)**: No migration.
- **Isolation-Migration ($\mathbf{IM_{4pop}}$)**: Migration between lineages 2 and 3 from the present to T_4.
- **Ancient Migration ($\mathbf{AM_{4pop}}$)**: Migration between lineages 2 and 3 between T_{dem} and T_4.
- **Secondary Contact ($\mathbf{SC_{4pop}}$)**: Migration between lineages 2 and 3 from the present to T_{dem}.

Migration was modeled symmetrically at a rate $M = N_e.m$, where N_e is the effective population size and m is the proportion of migrants in each generation.

The parameters explored included:

- T_4: Speciation time between lineages 1 and 2 ($0.5, 2, 4\, N_e$ generations).
- T_3: Speciation time between $(1, 2)$ and 3 ($T_4 + 0.5, 2, 4\, N_e$ generations).
- T_2: Speciation time between $((1, 2), 3)$ and 4 ($40\, N_e$ generations).
- T_{dem}: Migration onset/end time ($T_4 \cdot [0.1, 0.25, 0.5]$ for $\text{SC}_{4\text{pop}}$; $T_4 \cdot [0.5, 0.75, 0.9]$ for $\text{AM}_{4\text{pop}}$).
- $M (= N_e.m)$: Number of migrants per generation (0.25, 10 migrants).
- L_{mig}: Number of loci affected by migration (100, 500, 1000 loci out of 1,000 total loci).

Migration is assumed to be symmetric, occurring at a rate M corresponding to $N_e.m$, where N_e is the effective population size and m is the proportion of individuals in a population consisting of migrants in each generation. We simulate 1,000 independent loci, with no intra-locus recombination. Among these loci, a subset L_{mig} is influenced by migration, while the remaining loci represent genetic isolation between the species.

It is important to note that the assumptions of a constant N_e over time and across the genome, as well as the absence of intra-locus recombination, are not required by DILS. These constraints are therefore relaxed in the empirical application to plants *versus* animals. However, they are implemented in the simulation study to enable a direct comparison between DILS and QuIBL, as they align with the assumptions required by QuIBL.

Simulations for QuIBL and DILS were performed using the same parameter combinations. Differences lay in sampling schemes: for QuIBL, a single copy per locus was sampled from lineages 1, 2, 3, and 4, with exact gene trees simulated and rooted using lineage 4 (Fig. S9). For DILS, two diploids were sampled from lineages 2 and 3, reflecting the minimal sampling scheme when DILS was applied to the empirical plants *versus* animals dataset.

Simulations were generated using the script `simulations_for_QuIBL_DILS_full.py`. The command lines for simulating data using `msnsam` under different demographic models are as follows:

SI_{4pop}:

- **QuIBL:** `msnsam 4 1000 -T -I 4 1 1 1 1 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`

- **DILS:** `msnsam 8 1000 -t 40 -I 4 0 4 4 0 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`

AM$_{4pop}$:

- **QuIBL:** `msnsam 4 1000 -T -I 4 1 1 1 1 0 -em tbs 2 3 tbs -em tbs 3 2 tbs -eM tbs 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`
- **DILS:** `msnsam 8 1000 -t 40 -I 4 0 4 4 0 0 -em tbs 2 3 tbs -em tbs 3 2 tbs -eM tbs 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`

IM$_{4pop}$:

- **QuIBL:** `msnsam 4 1000 -T -I 4 1 1 1 1 0 -m 2 3 tbs -m 3 2 tbs -eM tbs 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`
- **DILS:** `msnsam 8 1000 -t 40 -I 4 0 4 4 0 0 -m 2 3 tbs -m 3 2 tbs -eM tbs 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`

SC$_{4pop}$:

- **QuIBL:** `msnsam 4 1000 -T -I 4 1 1 1 1 0 -m 2 3 tbs -m 3 2 tbs -eM tbs 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`
- **DILS:** `msnsam 8 1000 -t 40 -I 4 0 4 4 0 0 -m 2 3 tbs -m 3 2 tbs -eM tbs 0 -ej tbs 2 1 -ej tbs 3 1 -ej 10 4 1`

The QuIBL analyses were performed using the following options:

- `-numdistributions: 2`
- `-likelihoodthresh: 0.01`
- `-numsteps: 10`
- `-gradascentscalar: 0.5`
- `-multiproc: False`

- `-maxcores: 1000`
- `-totaloutgroup: 4`

For the DILS inferences on the simulated datasets, we used the following parameters: assuming a constant population size (`population_growth: constant`), a mutation rate μ of 2×10^{-8} mutations per generation per nucleotide (`mu: 0.00000002`), no intra-locus recombination (`rho_over_theta: 0`), and an effective population size N_e uniformly explored between 0 and 200,000 individuals. The split time was uniformly explored between 0 and 1,500,000 generations, and migration ($N_e.m$) was uniformly explored between 0.2 and 10 migrants per generation. No outgroup was used to polarize mutations (`nameOutgroup: NA`).

The final output file, `table_QuIBL_DILS.txt`, summarizes the demographic parameters and the inference results obtained from DILS and QuIBL for each simulated dataset.

Results of inferences performed with DILS and QuIBL on simulated datasets When the simulated model represents current isolation (SI_{4pop} and AM_{4pop}), both methods converge on the correct model in approximately 75% of simulations across all explored parameter combinations (Fig. S10-A, left panel). DILS performs slightly better than QuIBL, with around 14% of datasets where it is the only method to correctly recover the isolation model, compared to approximately 6.5% for QuIBL. However, these performances in supporting isolation are of the same general magnitude for both methods. For the parameter combinations used in these simulations, both methods fail to recover the isolation model in about 4.5% of cases.

The differences between the two methods become more pronounced when the true model involves ongoing migration (IM_{4pop} and SC_{4pop}). In this scenario, both methods converge on the correct model in approximately 36% of simulations under the explored parameter combinations (Fig. S10-A, right panel). QuIBL alone identifies migration in only 0.001% of simulations, while DILS is the sole method to correctly infer migration in about 29% of cases. This analysis suggests that both methods are highly conservative, failing to detect migration in 35% of simulations with ongoing gene flow. However, it is important to note that these results stem from discrete parameter combinations (T_4, T_3, T_{dem}, M, and L_{mig}), where migration can be both low ($N_e.m = 0.1$) and affecting only a small portion of the genome (10% of loci).

When we focus on simulations with higher levels of introgression (Fig. S10-B, right panel), the performance of both methods improves significantly. In these cases, the two methods converge on the correct model with ongoing migration in approximately 72% of simulations. The remaining simulations with ongoing migration are correctly recovered solely by DILS, and no simulations are simultaneously missed by both methods.

Our choice to use DILS in this study was primarily driven by its demonstrated reliability in previous analyses and its ease of use for interpreting temporal patterns of introgression via explicit scenario testing. Unlike QuIBL, which excels in handling large phylogenies, DILS is particularly well-suited for datasets with simpler sampling schemes, such as cases where only two species are available for analysis. Our comparison confirms that our initial choice of DILS did not come at the expense of reduced performance. This validation, combined with DILS's interpretability and suitability for the scope of our study, highlights its relevance for addressing the research questions we aimed to explore.

DILS

Model comparisons were carried out using the approximate Bayesian computation (ABC) framework applied in the animal study (*14*) and distributed under the name DILS (Demographic Inferences with Linked Selection (*11*)). DILS aims to test whether the studied species are connected by gene flow. It incorporates the effects of linked selection by modeling background selection as heterogeneous effective population sizes (hetero-N_e) across the genome and selection associated with barriers to gene flow as genomic heterogeneity in effective introgression rates (hetero-$N_e.m$). Here we describe how DILS works.

Compared models

The primary objective of our demographic analysis is to determine which two populations 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). In our ABC methodology, we discern two categories of models.

Demographic Models: Each of the four models 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_e), which closely approximates the influence of background selection by down-scaling N_e (*86*); and/or (b) heterogeneous migration rate across the genome (*hetero. M*) to account for the effects of selection against hybrids (*87*). 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_e 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 N_e values, which are drawn from uniform distributions. Conversely, the *homo.* N_e 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., $N_e.m = 0$), while the loci unaffected by species barriers are associated to an effective migration rate $N_e.m$ drawn from a uniform distribution. In the *homo. M* model, a single migration rate $N_e.m$ 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:

i) *homo.* N_e and *homo. M*

ii) *homo.* N_e and *hetero. M*

iii) *hetero.* N_e and *homo. M*

iv) *hetero.* N_e and *hetero. M*

For a comprehensive description of all prior distributions employed in this study, please refer to Section .

Summary statistics

ABC is a statistical inferential approach based on the comparison of summary statistics derived from simulated and observed datasets (*88*). We present a comprehensive description of the statistics computed within our framework. Previous studies have demonstrated the effectiveness of these statistics in statistically distinguishing demographic models with and without ongoing migration (*14, 26, 79*). 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 (*89*)
- Watterson's θ (*90*)
- Tajima's D (*91*)
- 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 (*92*)
- Net divergence D_a between the two populations/species (*92*)
- Relative genetic differentiation between the two populations/species quantified by F_{ST} (*93*)

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 (*94*)) 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.

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_e between 0 and $N_{e,max}$ individuals
- uniform prior for T_{split} between 0 and T_{max} generations
- uniform prior for migration rate $N_e.m$ between 0 and 10 migrants per generations

Where:

$$N_{e,max} = 5 \times max\left(\frac{\pi_A}{4\mu}; \frac{\pi_B}{4\mu}\right)$$

π_A and π_B being the Tajima's θ (*89*) 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 (*92*).

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. S11). For each pair of populations/species, three statuses are then assigned:

i) Strong support for genetic isolation: we identify strong statistical support for genetic isolation when our ABC framework yields a posterior probability $P_{\text{mig}} < 0.1304$. This threshold was empirically determined by the robustness test conducted in (*14*), where the robustness R_{mig} was calculated as:

$$R_{\text{mig}} = \frac{P(P_{\text{mig}} = P_i \mid \text{mig})}{P(P_{\text{mig}} = P_i \mid \text{mig}) + P(P_{\text{mig}} = P_i \mid \text{iso})}$$

where:

- P_i: the posterior probability attributed by our ABC framework to migration to a given simulated dataset.
- $P(P_{\text{mig}} = P_i \mid \text{mig})$ is the probability that a dataset simulated under a model with ongoing migration is correctly inferred as a model with migration by our ABC approach, given a posterior probability for migration of P_i.

- $P(P_{\text{mig}} = P_i \mid \text{iso})$ is the probability that a dataset simulated under a model with current isolation is wrongly inferred as a model with migration by our ABC approach, given a posterior probability for migration of P_i.

ii) Strong support for ongoing migration: strong statistical support for ongoing migration is indicated when the posterior probability $P_{\text{mig}} > 0.6419$, also empirically determined in (*14*).

iii) 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 Fig. 1-A, allowing the 'migration' status to be treated in a logistic regression (see section).

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:

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

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 $X_{p=0.5} = -\frac{\beta_0}{\beta_1}$.
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:

$$\begin{aligned}
\ell(\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x}) &= \log\left(\mathcal{L}(\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\mathbf{X}=\mathbf{x})\right) \\
&= \sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \\
&= \sum_{i=1}^{N}\left[y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right] \\
&= \sum_{i=1}^{N}\left[y_i \cdot \mathbf{x}_i\boldsymbol{\beta} - \log(1+\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right))\right] \\
&= \sum_{i=1}^{N}\left[y_i \cdot (\beta_0 + \beta_1 X_{1,i}) - \log(1+\exp\left(\beta_0 + \beta_1 X_{1,i}\right))\right] \qquad (1)
\end{aligned}$$

The sigmoid of plants can then be tested against that of animals to determine whether plants and animals share the same dynamics of reproductive isolation accumulation. For this purpose, three models are fitted and associated to log-likelihood ℓ:

i) M_0: both plants and animals share the same logistic relationship between $\mathbf{X}_i$ and $\mathbf{Y}_i$.

ii) M_{plants}: model fitted to the plants data only.

iii) $M_{animals}$: 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. (*14*). 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 permutation-based approach. Specifically, the absolute difference between $\ell(M_0)$ and $\ell(M_{plants}) + \ell(M_{animals})$ was compared to a null distribution generated from 10,000 random permutations of the data. The *P*-value corresponds to the proportion of permutations in which the absolute difference exceeded the observed value (Table S2).

Controls for methodological and biological biases

To ensure the robustness of our main finding—that plants experience a faster cessation of gene flow compared to animals—we performed additional control analyses addressing three potential sources of bias: sequencing methodology, unequal representation of genera, and geographic distances between lineages. Each of these analyses confirmed that our conclusions are unlikely to result from methodological artifacts or sampling biases, as detailed in the sections below.

Testing for a sequencing technology effect

Out of the total dataset comprising 280 pairs of plants and 61 pairs of animals, 183 plant pairs and 54 animal pairs exhibited high robustness in model comparison and passed the goodness-of-fit test. These retained datasets encompass a diversity of sequencing methodologies. Specifically,

within plants, among the 183 retained pairs: 81 pairs were acquired through RAD-sequencing, 70 pairs through RNA-sequencing, and 32 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 (P-value< 0.0001; Table S3).

Testing for a phylogenetic effect

To control for the variation in the number of pairs between genera, we carried out 1,000 animal-plant comparisons as for Fig. 1 but by randomly selecting a single pair per animal and plant genus (Fig. S1-A). Over these 1,000 sub-samples, the relative positions of the sigmoids were compared *via* the inflection points ($X_{p=0.5} = -\frac{\beta_0}{\beta_1}$) of the models fitted to the plant *versus* animal sub-samples. We consistently find that the inflection point occurs at lower levels of divergence in plants than in animals (Fig. S1-B). However, the difference in likelihood between the global model and the combined likelihoods of the plant and animal models is not significant in 7% of the permutations (Fig. S1-C; blue bars). This lack of significance is attributed to the sub-sampling process, which substantially reduces the number of data points contributing to the likelihood calculations, thereby decreasing the power to discriminate effectively between models. Nevertheless, the distance between the plant and animal inflection points was significantly greater than expected by random chance in 100% of the permutations (Fig. S1-C; orange bars).

Testing for a geographic effect

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*.

To test whether the observed pattern of reduced gene flow in plants compared to animals could be merely explained by greater geographic distances between plant pairs, we performed logistic regression analyses separately for plants and animals. Using the GPS coordinates of sampled individuals, we calculated the minimum geographic distance (`distances_meters`) between taxa within each pair.

We then modeled the probability of ongoing migration (`P_ongoing_migration`) as a function of geographic distance and net divergence (`netdiv_avg`) using binomial logistic regression with a logit link function. The logistic regression results showed that geographic distance (`distances_meters`) was not a significant predictor of ongoing migration for either animals ($P = 0.688$) or plants ($P = 0.937$), confirming that our conclusions are not driven by geographic factors.

Testing factors influencing speciation dynamics within plants

To investigate factors that may influence the dynamics of reproductive isolation in plants, we focused on two life-history traits: (i) growth form, specifically comparing herbs and trees (Section), and (ii) mating systems by comparing different selfing rates (Supplementary Materials A.7.2). The first comparison is motivated by the typically greater pollen dispersal observed in trees compared to annual plants, leading to reduced nuclear genetic differentiation within tree populations (*95, 96*). For the effect of the mating system, we limited our analysis to a comparison between species in our sample with the lowest selfing rates and those with the highest. However, it is important to note that our sample is not representative of the full diversity of selfing rates found in plants (*97*), as it is biased towards high-outcrossing species (Fig. S13). The theoretical effect of selfing rate on the evolution of reproductive isolation remains unclear and appears contradictory (*98*). On one hand, higher selfing rates reduce gene flow, decrease effective recombination, and increase the strength of genetic drift, which can facilitate the fixation of deleterious mutations that may act as barriers when compensated. On the other hand, reduced efficacy of selection in high selfing species can

mitigate intragenomic conflicts, potentially leading to fewer interspecific incompatibilities.

For these analyses, we focused on the plant dataset, subdivided into different groups based on the specific comparisons performed for each tested factor. We first conducted a comparison based on life form (herbs *versus* trees), followed by three comparisons to assess the effect of selfing rate (pairs of selfers *versus* pairs of outcrossers, pairs of selfers *versus* pairs with different systems, and pairs of outcrossers *versus* pairs with systems).

Each comparison was performed similarly to the plant-*versus*-animal comparison: we first fitted a global model on the entire dataset and then tested whether the sum of the likelihoods of the models fitted on each subgroup was significantly greater than that of the global model.

Effects of plant life forms

The plant dataset was divided into four categories based on their life form:

- **Liana** (e.g., *Actinidia*, *Nepenthes*)
- **Herb** (e.g., *Arabis*, *Dactylorhiza*, *Helianthus*, *Hibiscus*, *Isoetes*, *Lupinus*, *Pitcairnia*, *Pulmonaria*, *Rhodanthemum*, *Senecio*, *Silene*, *Phlox*)
- **Tree** (e.g., *Ficus*, *Laccospadicinae*, *Phoebe*, *Picea*, *Populus*, *Quercus*, *Yucca*)
- **Shrub** (e.g., *Gossypium*, *Salix*, *Stachyurus*)

We performed two comparisons: first, between herbs and trees (Fig. S3-A), which showed no significant reduction in gene flow among herbs compared to trees (P = 0.261; Table S4). When herbs, lianas, and shrubs were grouped together (Fig. S3-B), we still did not observe a reduction in gene flow relative to trees (P = 0.1937; Table S4).

Effects of plant mating systems

In line with (*99*), we aimed to test whether the mating system influences the dynamics of introgression. More specifically, we tested whether selfing species exhibit a faster emergence of reproductive isolation compared to outcrossing species, as hypothesized due to reduced dispersal, decreased

effective recombination, and a higher rate of accumulation of genetic incompatibilities. Alternatively, selfing species might experience a slower development of reproductive isolation compared to outcrossers, driven by reduced intragenomic conflicts that could otherwise act as barriers.

To investigate whether the extent of inbreeding influences the rate of reproductive barrier establishment, we categorized plant species according to their selfing rates using three independent sources of information: two quantitative and one qualitative.

i) First, we developed a custom estimator, hereafter referred to as `selfing_ML`, which infers the selfing rate s using a maximum likelihood approach. This method models the probability of observing a genotype given the allele frequency at each polymorphic position and a specified selfing rate, under the assumption of Hardy–Weinberg equilibrium modified for self-fertilization.

ii) Second, we used the R package `inbreedR` (*23*), which estimates inbreeding based on multilocus identity disequilibrium (g_2), a signal of correlated heterozygosity among loci (Fig. S13-A). However, for 11 plant species, `inbreedR` failed to produce estimates.

iii) Finally, we incorporated qualitative data on plant mating systems extracted from an established plant trait database (*25*). Based on this source, we identified in our dataset 18 dioecious species, 19 self-incompatible (SI), 5 distylous, 10 self-compatible (SC), and 4 gynodioecious species. Information was not available for the remaining species.

Genomic data yielded similar results regarding the shape of the selfing rate distribution across species. Both methods produced a unimodal distribution with a peak at $s = 0$ (Fig. S13-A), whereas the literature suggests that, across a broader range of species, the distribution is typically bimodal, with a second peak at $s = 1$ (Fig. S13-B; (*97*)). This indicates that our sampling is likely biased toward a deficit of highly selfing species. A difference between `selfing_ML` and `inbreedR` is that more species were estimated to have $s = 0$ using `selfing_ML`. Qualitative analysis of mating systems reveals that these species are either dioecious or possess a genetic self-incompatibility (SI) system that prevents self-fertilization (Fig. S14), suggesting that this pattern has a biological basis rather than resulting from a methodological artifact. This qualitative analysis also supports the observed deficit of clearly selfing species: among the 10 species (over a total of 56 species

with known mating system) identified as self-compatible (SC), the median selfing rate estimate was below 0.2 (Fig. S14). The gynodioecious taxa in our study correspond to *Silene nutans* lineages in which hermaphrodites are not self-incompatible, resulting in a reproductive system that combines obligatory outcrossing via females and partial selfing from hermaphroditic individuals.

To assess whether selfing influences the dynamics of reproductive isolation, we categorized species separately based on selfing rates estimated by the `selfing_ML` and `inbreedR` methods. For each method, we classified species into two categories based on whether their selfing rate was above (classified as 'highest') or below (classified as 'lowest') the method-specific median across the 105 plant species analyzed (94 species for which `inbreedR` provided estimates). The median selfing rate was 0 for `selfing_ML` and 0.026 for `inbreedR`. Based on these classifications, we grouped pairs of species into three categories: *highest–highest*, *lowest–lowest*, and *mixed* (one 'highest', one 'lowest' selfer). For `selfing_ML`, this resulted in 30 highest–highest pairs, 108 lowest–lowest, and 45 mixed pairs; for `inbreedR`, we obtained 45, 41, and 57 pairs, respectively (Fig. S4-A and B).

We then applied the same analytical framework used to compare reproductive isolation dynamics between plants and animals, testing whether the extent of gene flow differed significantly between the three pair categories. For both selfing rate estimation methods, and across all pairwise comparisons, we found mostly non-significant effect (Table S4). This lack of detectable signal may partly reflect the bias in our dataset toward outcrossing species, as previously noted.

Definitions and Notations concerning `selfing_ML`.

In this paragraph, we detail the rationale and functioning of the `selfing_ML` method.

- s is the equilibrium selfing rate, where $s \in [0, 1]$.
- p is the allele frequency of the alternative allele at a given site.
- $q = 1 - p$ is the frequency of the reference allele.
- F is the inbreeding coefficient, which is related to the selfing rate s at equilibrium by:

$$F = \frac{s}{2 - s}$$

Genotype Probabilities.

For a given site, the probabilities of observing the genotypes are modeled using p, q, and F:

- Probability of observing genotype "11" (homozygous for the alternative allele):

$$P_{11} = p^2 + p \cdot q \cdot F \tag{2}$$

- Probability of observing heterozygous genotype "10" or "01":

$$P_{10} = 2 \cdot p \cdot q \cdot (1 - F) \tag{3}$$

- Probability of observing genotype "00" (homozygous for the reference allele):

$$P_{00} = q^2 + p \cdot q \cdot F \tag{4}$$

Log-Likelihood Calculation.

For an individual i, the log-likelihood of the equilibrium selfing rate s explaining F is computed by summing the log-probabilities of the observed genotypes across all polymorphic sites:

$$\log \mathcal{L}(s \mid \text{data}) = \sum_{l=1}^{L} \sum_{j=1}^{n} \log P_{\text{geno}_{ij}}(f_A, s)$$

Where:

- L is the number of polymorphic loci (i.e, loci containing at least one polymorphic site).
- n is the number of polymorphic sites within each locus.
- $P_{\text{geno}_{ij}}$ is the probability of the observed genotype at site j for individual i, calculated based on equations (2-4).

Selfing Rate Estimation.

We evaluate the log-likelihood for a range of selfing rates s (from 0 to 1 in steps of 0.05). The estimated selfing rate $\hat{s}_i$ for individual i is the value of s that maximizes the log-likelihood:

$$\hat{s}_i = \arg \max_{s \in [0,1]} \log \mathcal{L}(s \mid \text{data})$$

Numerical Stability.

To avoid issues with numerical underflow when computing log-probabilities, a small constant $\epsilon = 10^{-10}$ is added:

$$\log(P_{11}) = \log(p^2 + p \cdot q \cdot F + \epsilon)$$

This ensures that the probabilities are always strictly positive, preventing errors in the log calculations.

Implementation.

The entire pipeline for estimating selfing rates using the maximum likelihood approach described above has been implemented in Python. This method, along with detailed documentation and example datasets, is freely available as an open-source project on `Zenodo` (*24*).

Accuracy of the estimate.

We evaluated the performance of our method across a range of 11 selfing rates (s) from 0 to 1. Genotype data were simulated using the R package `hierfstat` (*100*), which generates diploid genotypes for populations with a specified inbreeding coefficient (F), given a defined number of loci and individuals. The simulated datasets were then analyzed with our custom tool to estimate the selfing rate (s), allowing us to assess both the accuracy and the limitations of the method. We investigated the impact of varying sample sizes (2, 3, 4, 5, and 10 diploid individuals) and explored datasets consisting of 1,000 and 5,000 SNPs. Each combination of selfing rate, SNP count, and sample size was replicated 20 times.

Overall, selfing rate estimates (s) are substantially underestimated when the sample consists of only 2 diploid individuals (Fig. S12). Thus, for simulations with a sample size of 2 diploids and low true selfing rates (below 0.3), the inferred selfing rate is frequently estimated as zero. This bias diminishes rapidly with an increase in sample size. However, even with small sample sizes, our method reliably detects high selfing rates when the true value is elevated, although the estimates tend to be slightly underestimated. This underestimation in small samples can be attributed to the overestimation of true allele frequencies for rare alleles, which are the most common type of alleles in a site frequency spectrum. In a sample of only 2 diploid individuals, a rare allele would have a minimum frequency of 25%, which does not accurately reflect its true population frequency. This

inflated allele frequency in small samples biases the selfing rate estimate downward by making the absence of homozygotes for such alleles unlikely under a non-zero selfing rate. This effect diminishes both with a moderate increase in sample size (beyond 2 individuals) and at higher true selfing rates.

Data availability

All assembled datasets, the reference list used for mapping, the results of demographic inference, and the R scripts for statistical analyses and figure generation are available on Zenodo (*24, 41, 42*).

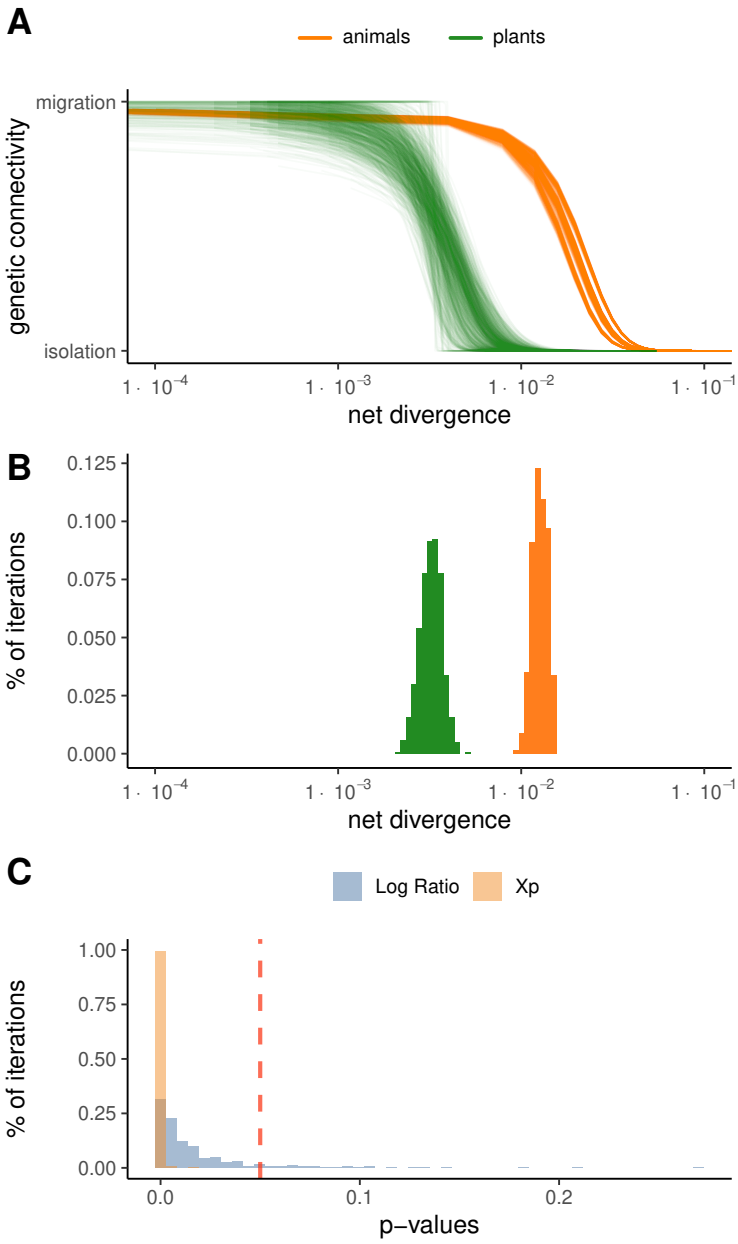

Figure S1: **Test of the robustness of differences between plants and animals to genus effects through random subsampling.** (**A**) Relationship between net divergence and migration/isolation status based on random sub-sampling of one population/species pair per genus (plants: green; animals: orange), repeated 1,000 times. Each line represents one sub-sampling. (**B**) Distribution of inflection points ($X_{p=0.5}$) for plants and animals across 1,000 sub-samplings. (**C**) *P*-value distributions from 1,000 sub-samplings: blue bars represent the log-likelihood ratio tests, and orange bars represent the tests on difference between inflection points ($X_{p=0.5;\text{animals}} - X_{p=0.5;\text{plants}}$). The red dashed line indicates the 0.05 significance threshold.

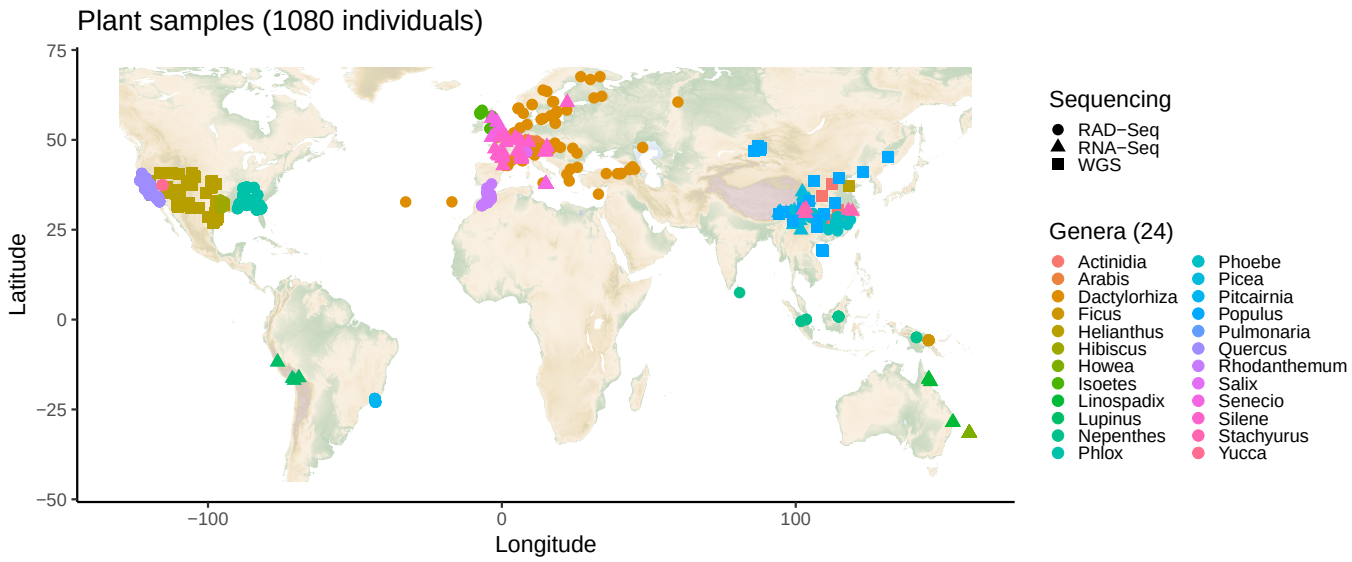

Figure S2: **Geographical location of plant samples and sequencing methods.** Each symbol represents a sequenced plant individual. The shapes correspond to the sequencing technologies used, while the colors indicate the genera.

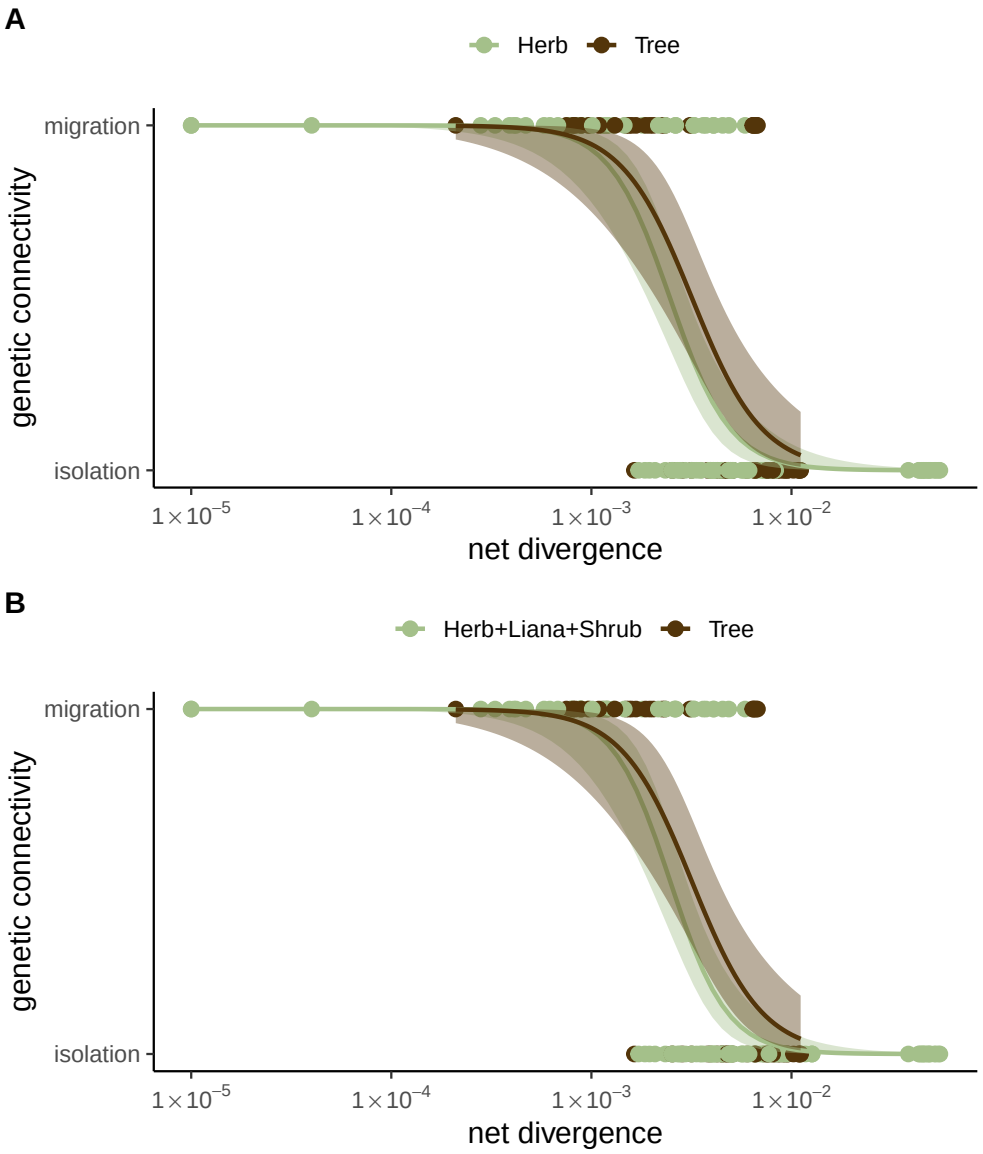

Figure S3: **Relationship between mean net divergence and migration/isolation status across plant life forms.** (**A**) Comparison between herbs and trees. (**B**) Comparison between herbs, lianas, and shrubs combined *versus* trees.

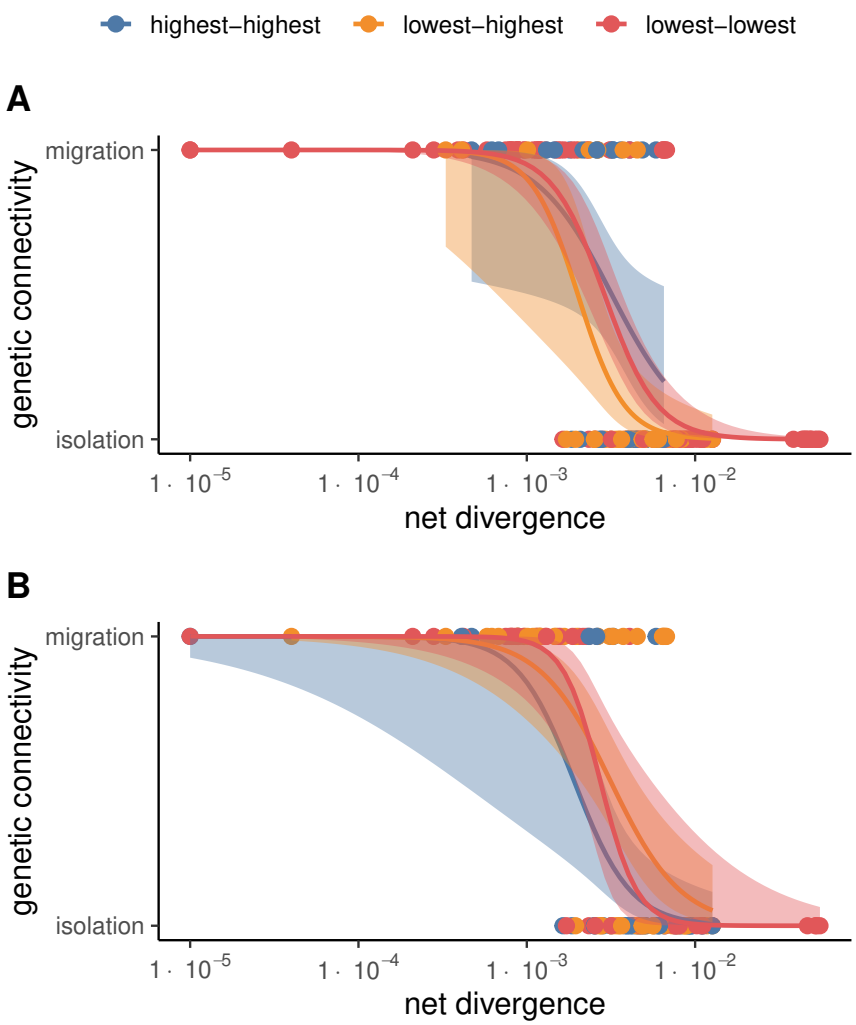

Figure S4: **Relationship between net divergence and genetic connectivity across plant pairs grouped by selfing rates or reproductive systems.** (**A**) Grouping based on species-averaged selfing rates estimated with `selfing_ML` (median across species = 0). Pairs are classified as *highest-highest* (both species with $s > 0$, $n = 30$), *lowest-lowest* (both with $s = 0$, $n = 108$), or mixed (one species with $s > 0$, the other with $s = 0$, $n = 45$). (**B**) Same classification using estimates from `inbreedR` (median = 0.026). Pairs are grouped as *highest-highest* ($s > 0.026$, $n = 45$), *lowest-lowest* ($s \leq 0.026$, $n = 41$), or mixed ($n = 57$). Only pairs with a significant signal of migration or isolation are included as for Figure 1.

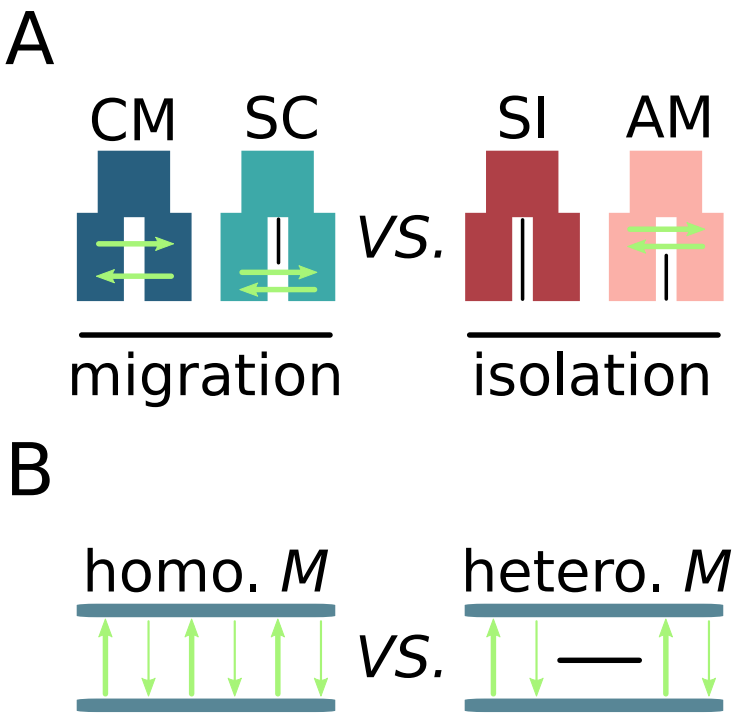

Figure S5: **Compared models using approximate Bayesian computation (ABC).** (**A**) Models with ongoing migration correspond to all CM (Continuous Migration) and SC (Secondary Contact) models. Models with current isolation correspond to all SI (Strict Isolation) and AM (Ancestral Migration) models. 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.

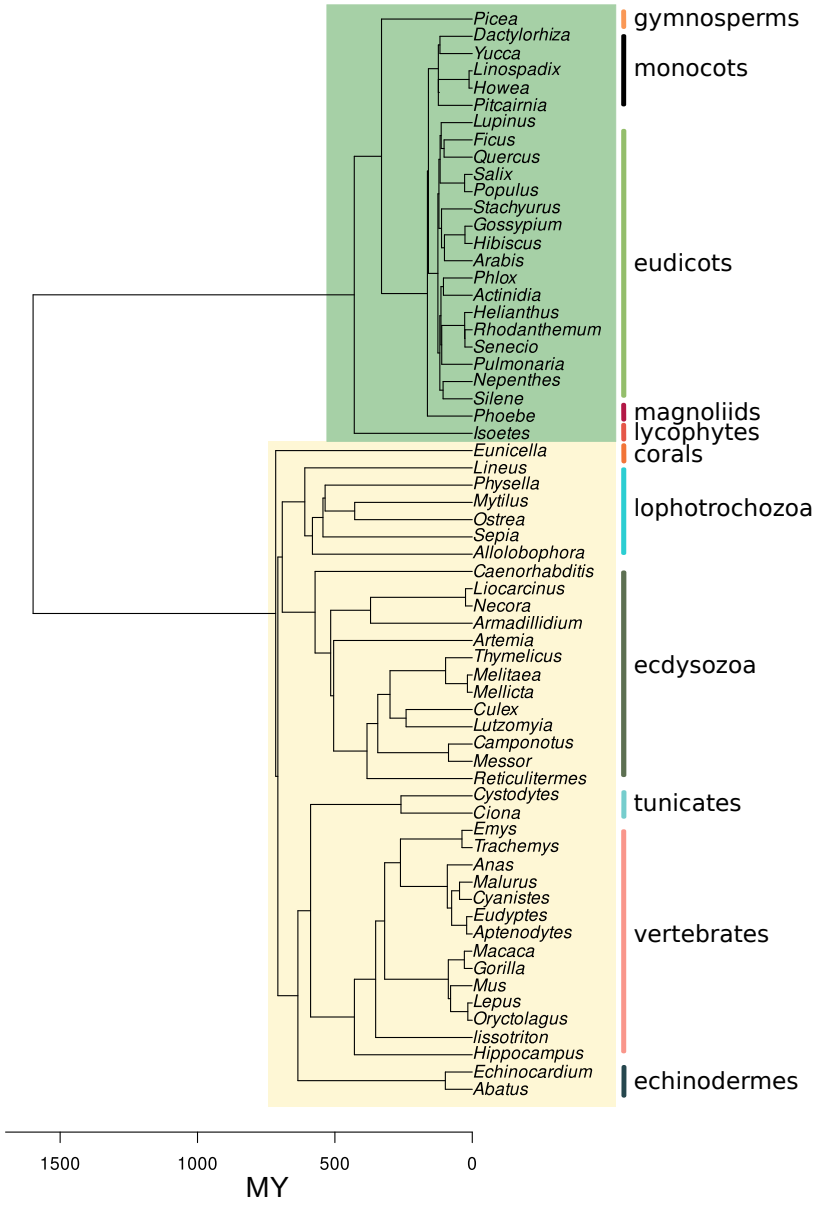

Figure S6: **Phylogenetic relationships between species included in the current study.** Plants and animals are indicated by green and yellow rectangles respectively. The scale represent the time from present expressed in million years (MY) according to `TimeTree` (*101*). Animals (yellow square) are from (*14*). Plants (green square) are included in the current study.

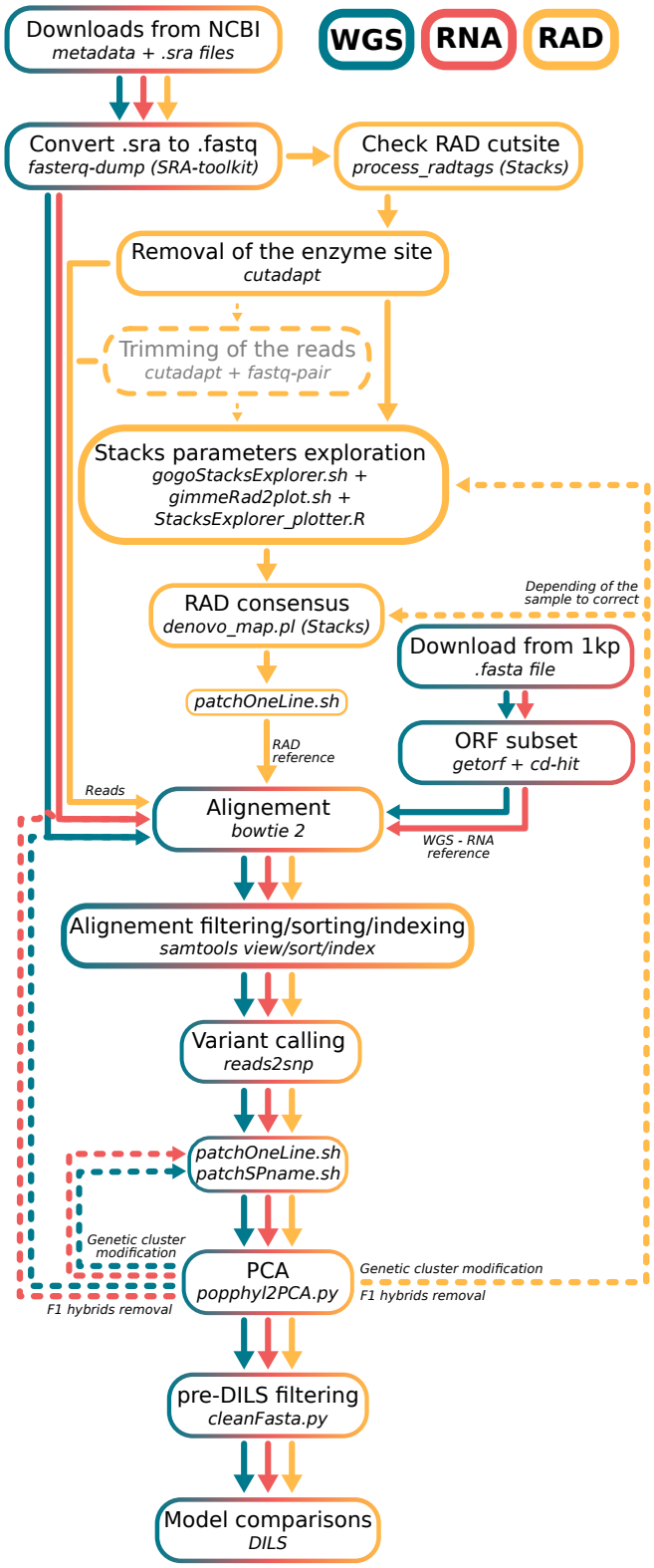

Figure S7: **Bioinformatics steps from the raw reads to demographic analysis.** Within each box, the upper line delineates an information technology procedure employed for data processing, while the lower line specifies the program or script utilized for its execution. The coloration denotes the specific sequencing technology concerned by each step.

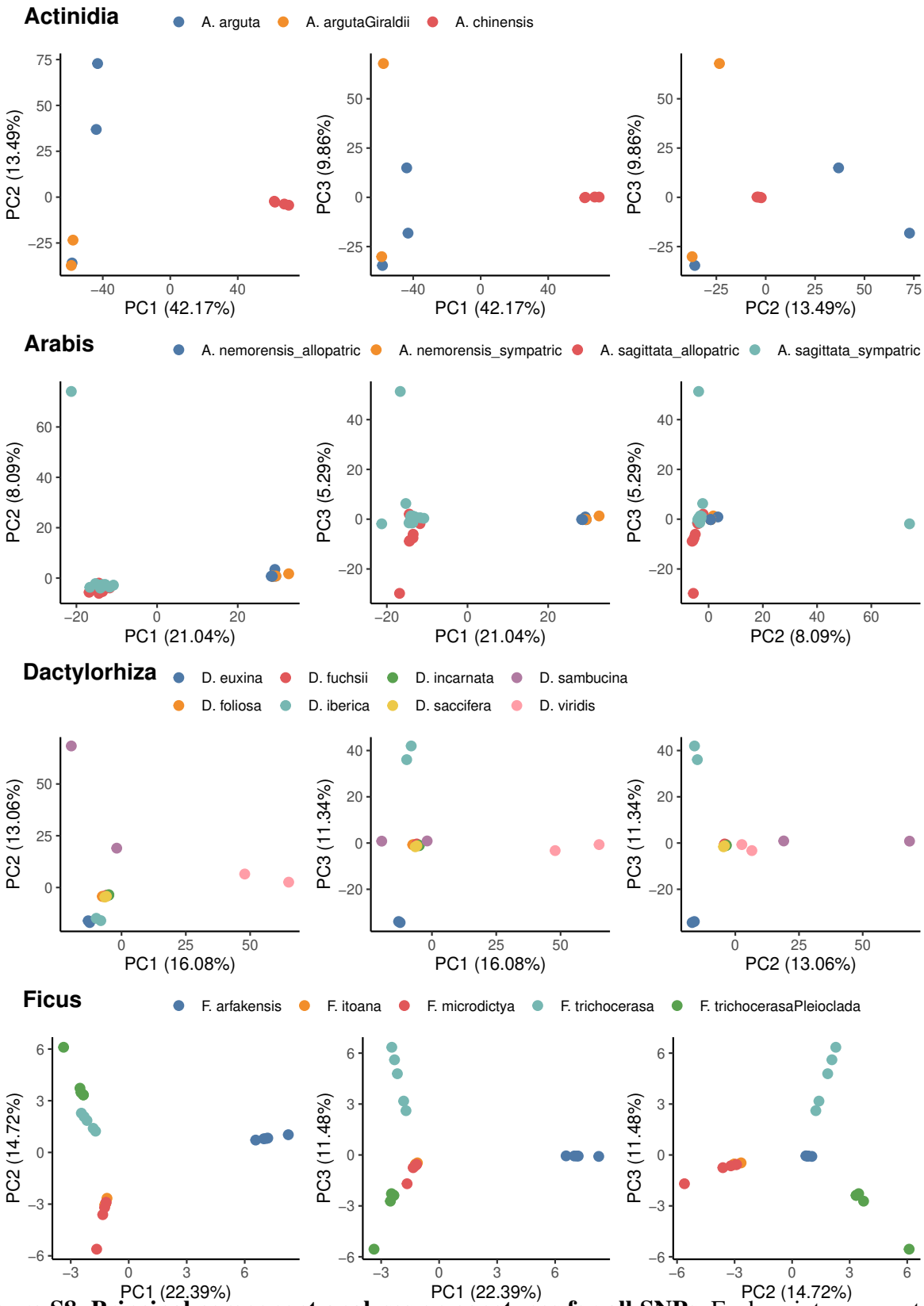

Figure S8: **Principal component analyses on genotypes for all SNPs.** Each point represents an individual. The colours represent the different populations/species named by the authors of the studies from which the data originated. (Page 1 of 6)

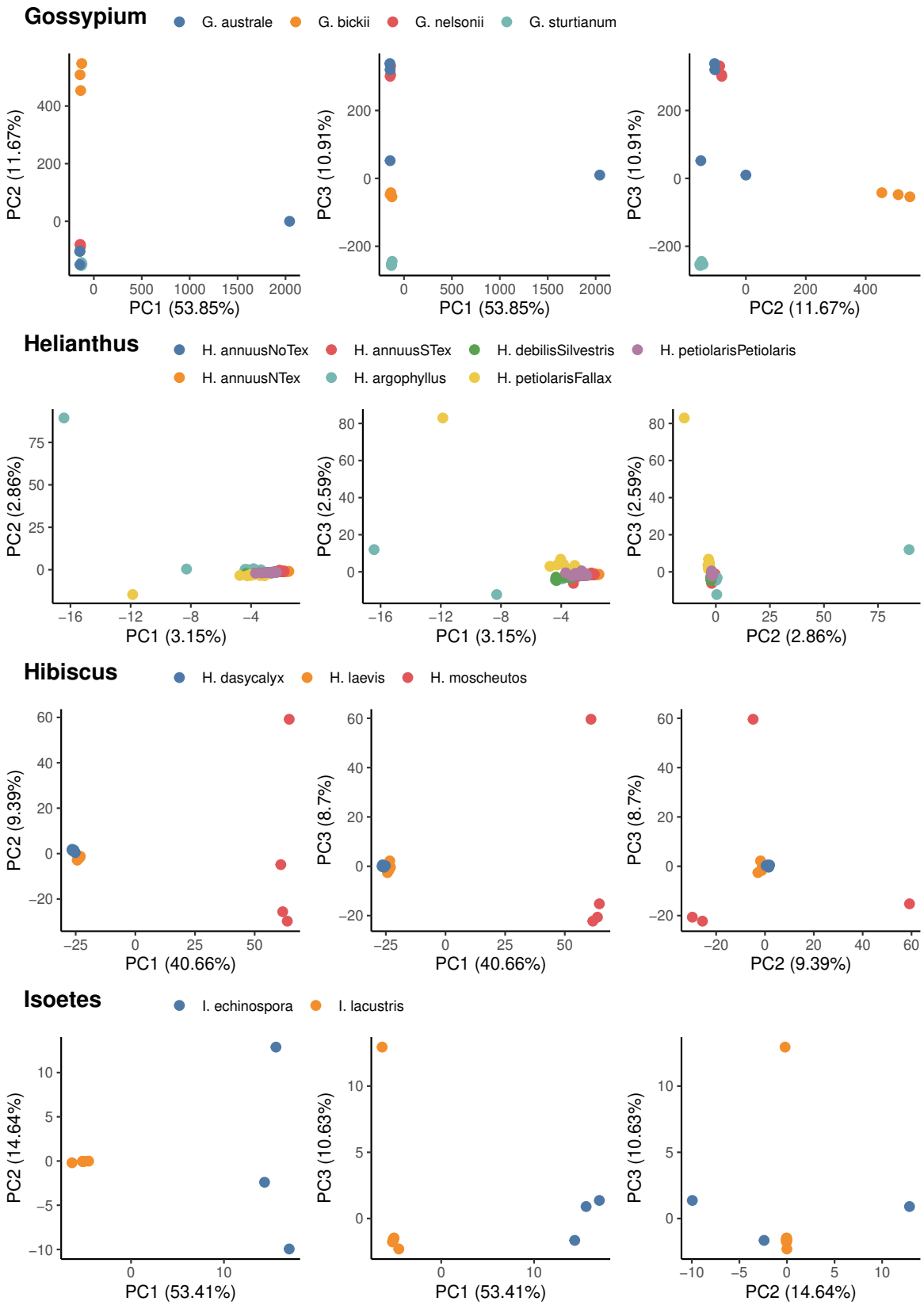

Figure S8 (continued, page 2 of 6)

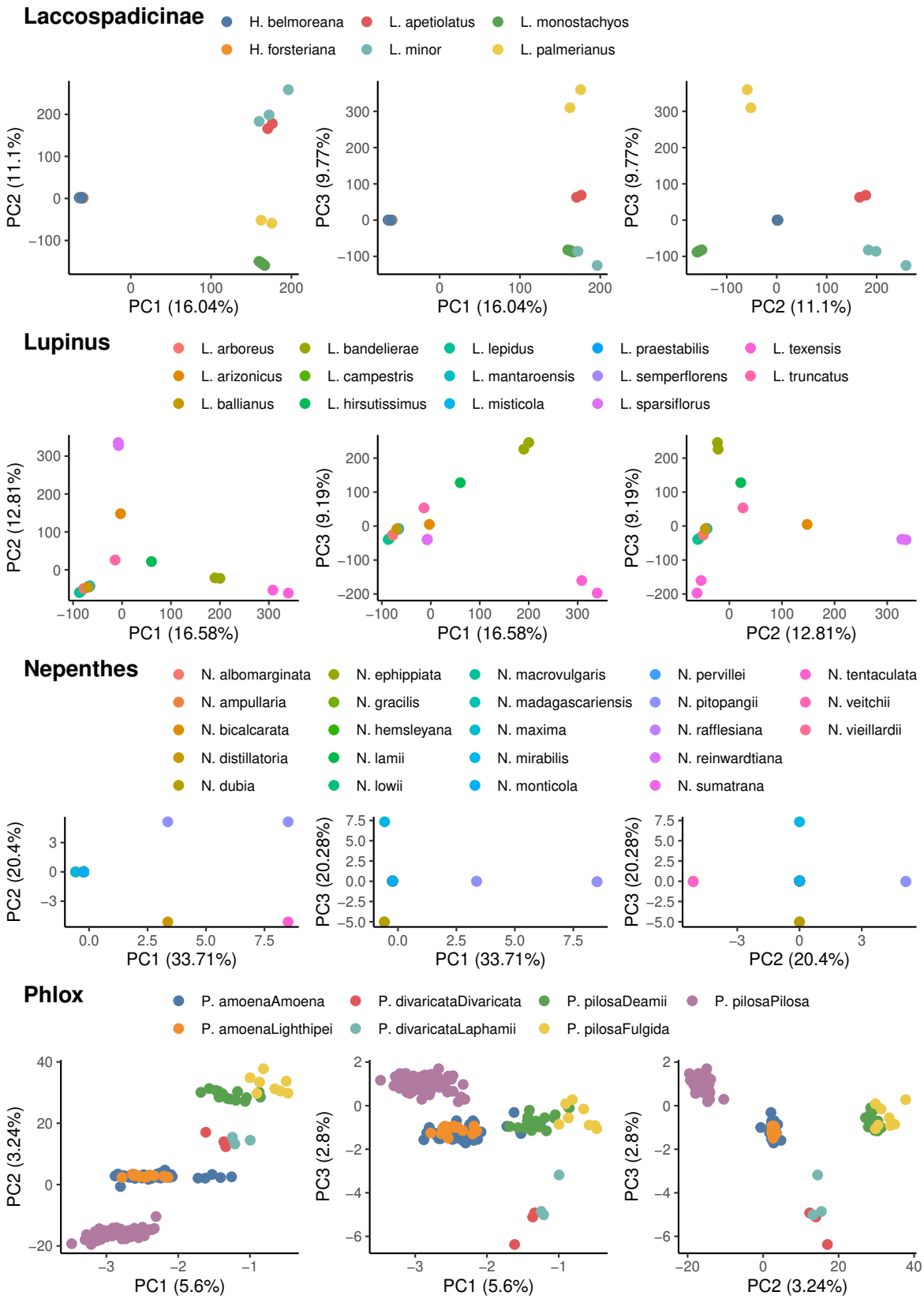

Figure S8 (continued, page 3 of 6)

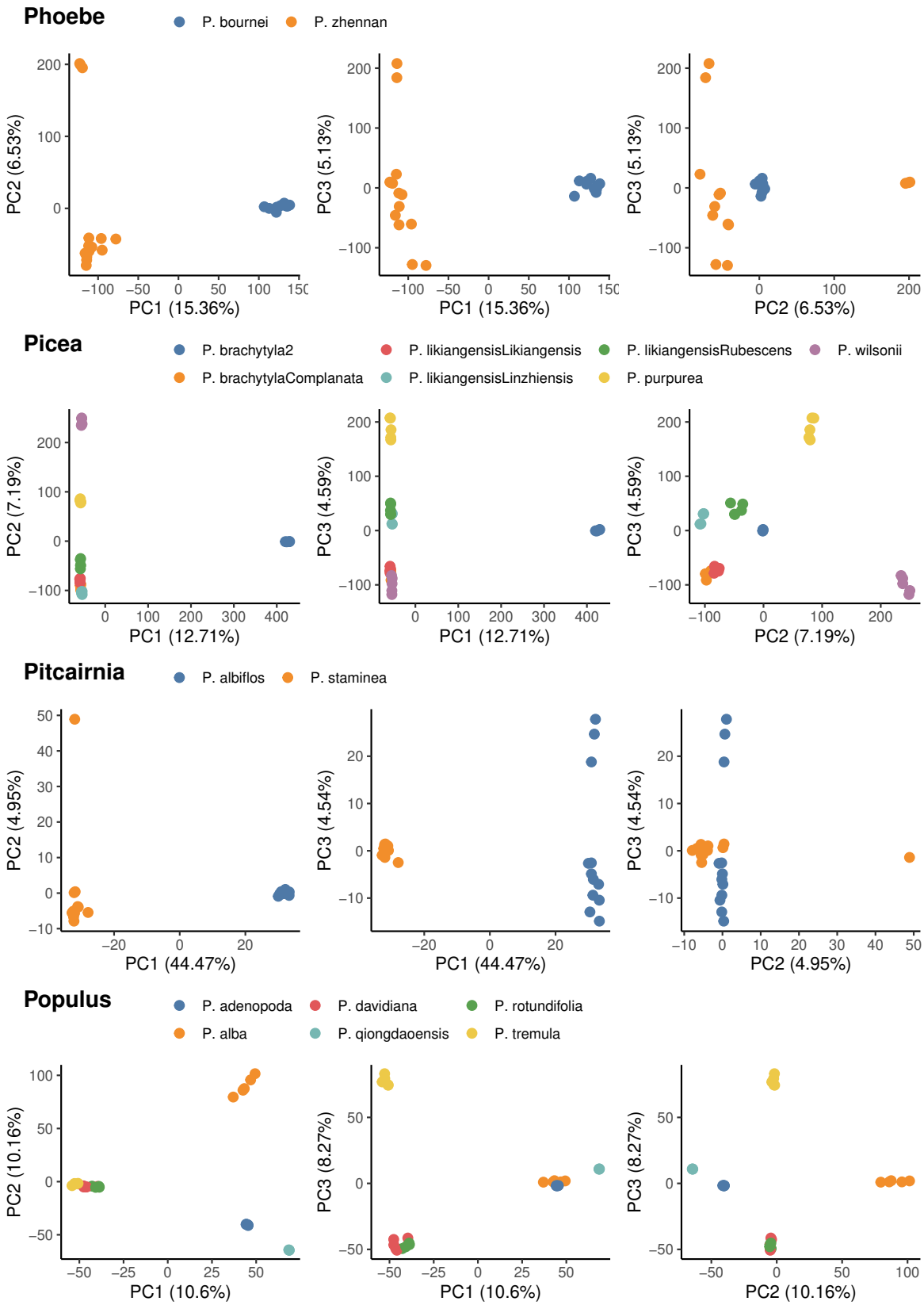

Figure S8 (continued, page 4 of 6)

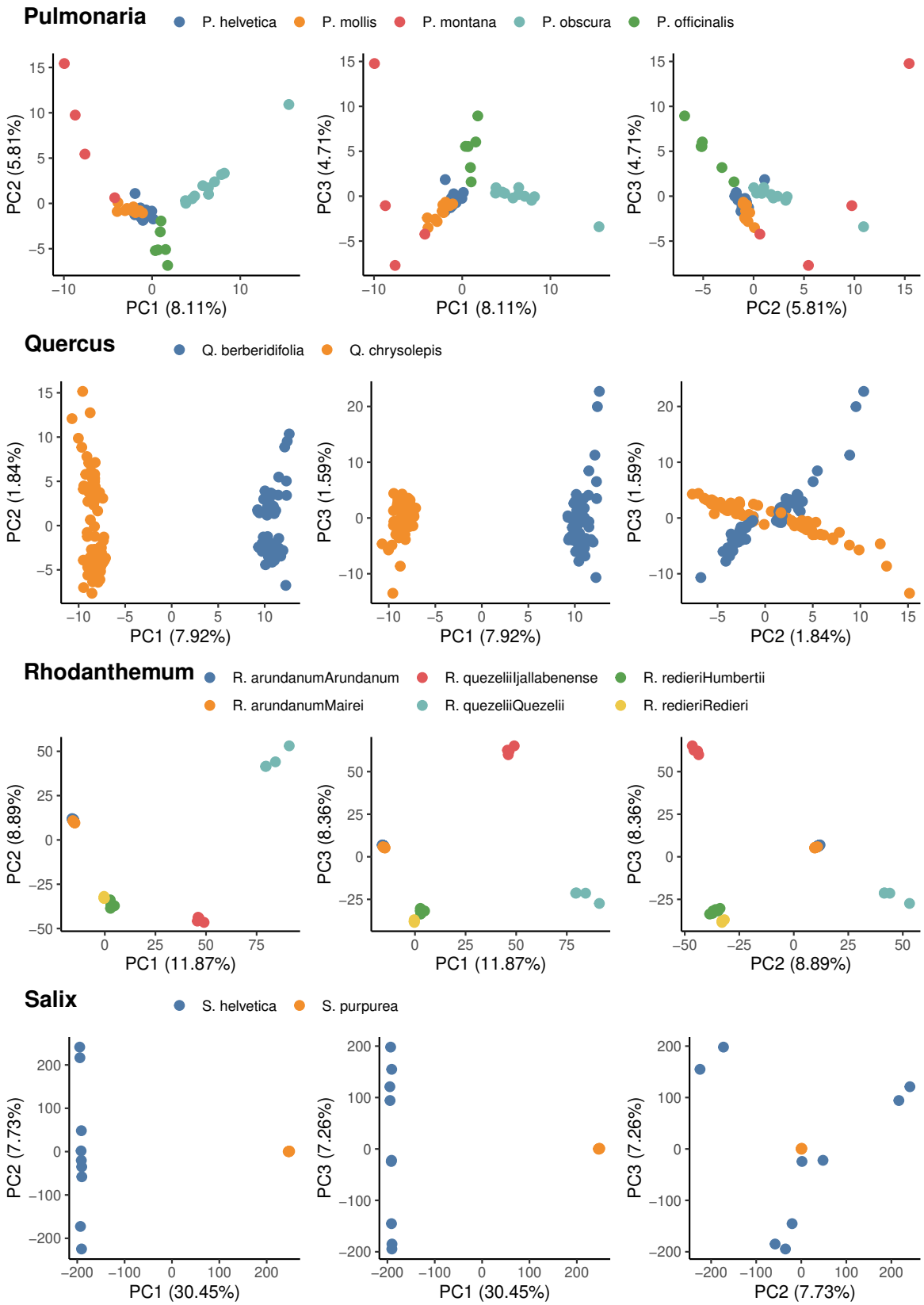

Figure S8 (continued, page 5 of 6)

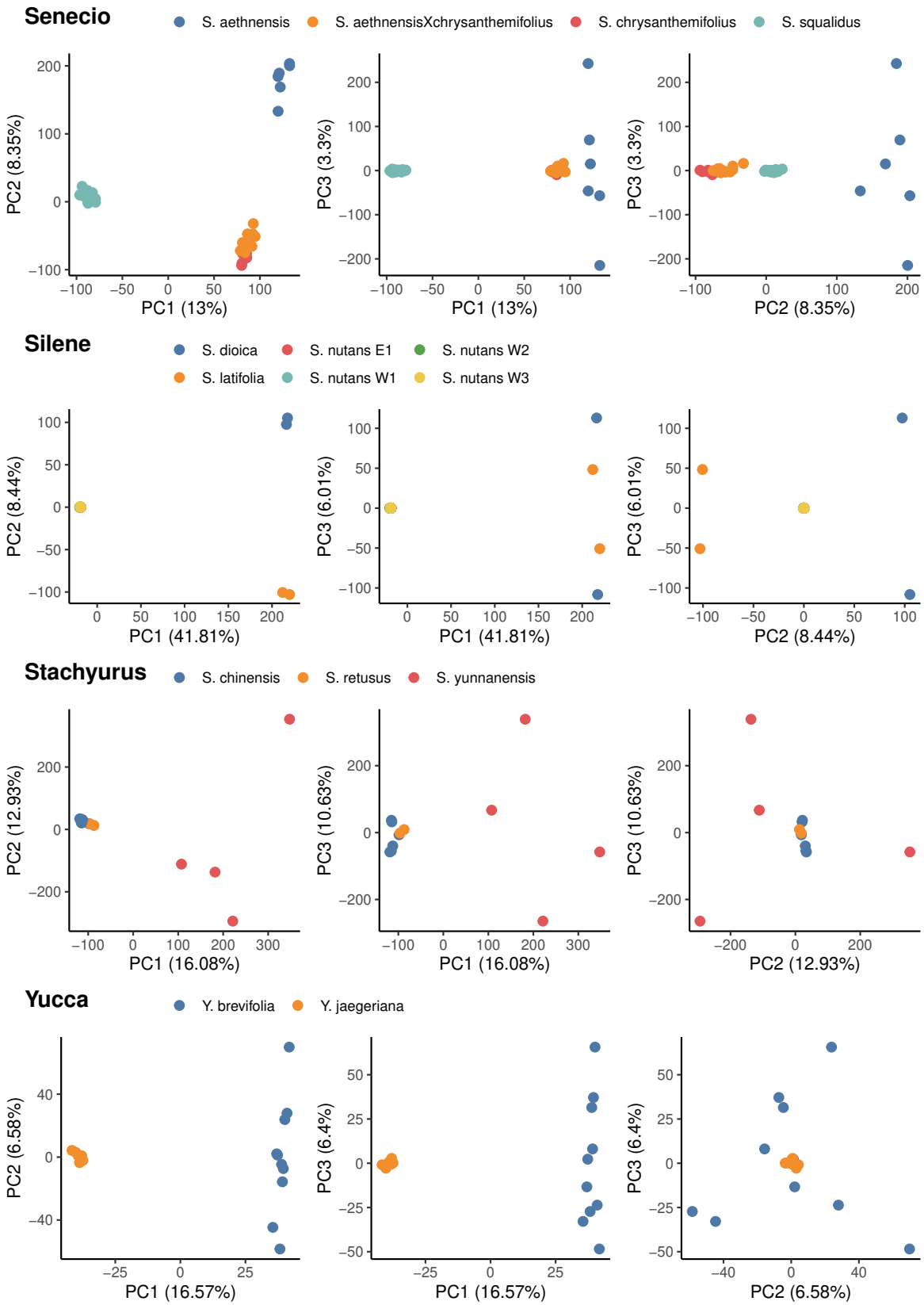

Figure S8 (continued, page 6 of 6)

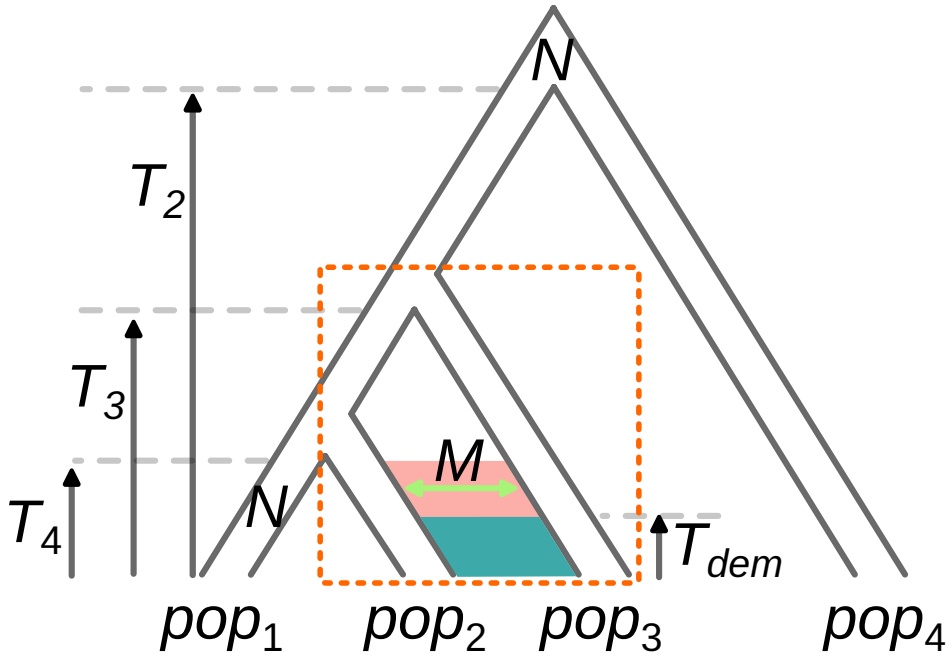

Figure S9: **Simulated demographic model for testing introgression between populations pop_2 and pop_3 (dotted rectangle) using DILS and QuIBL.** Parameters: T_4 (speciation time between 1 and 2), T_3 (speciation time between $(1, 2)$ and 3), T_2 (speciation time between $((1, 2), 3)$ and 4), T_{dem} (migration onset/end), M (number of migrants per generation), N_e (effective population size). Scenarios: $\text{SI}_{4\text{pop}}$ ($M = 0$ during red and blue periods), $\text{AM}_{4\text{pop}}$ ($M > 0$ during red, $M = 0$ during blue), $\text{IM}_{4\text{pop}}$ ($M > 0$ during red and blue), $\text{SC}_{4\text{pop}}$ ($M > 0$ during blue, $M = 0$ during red).

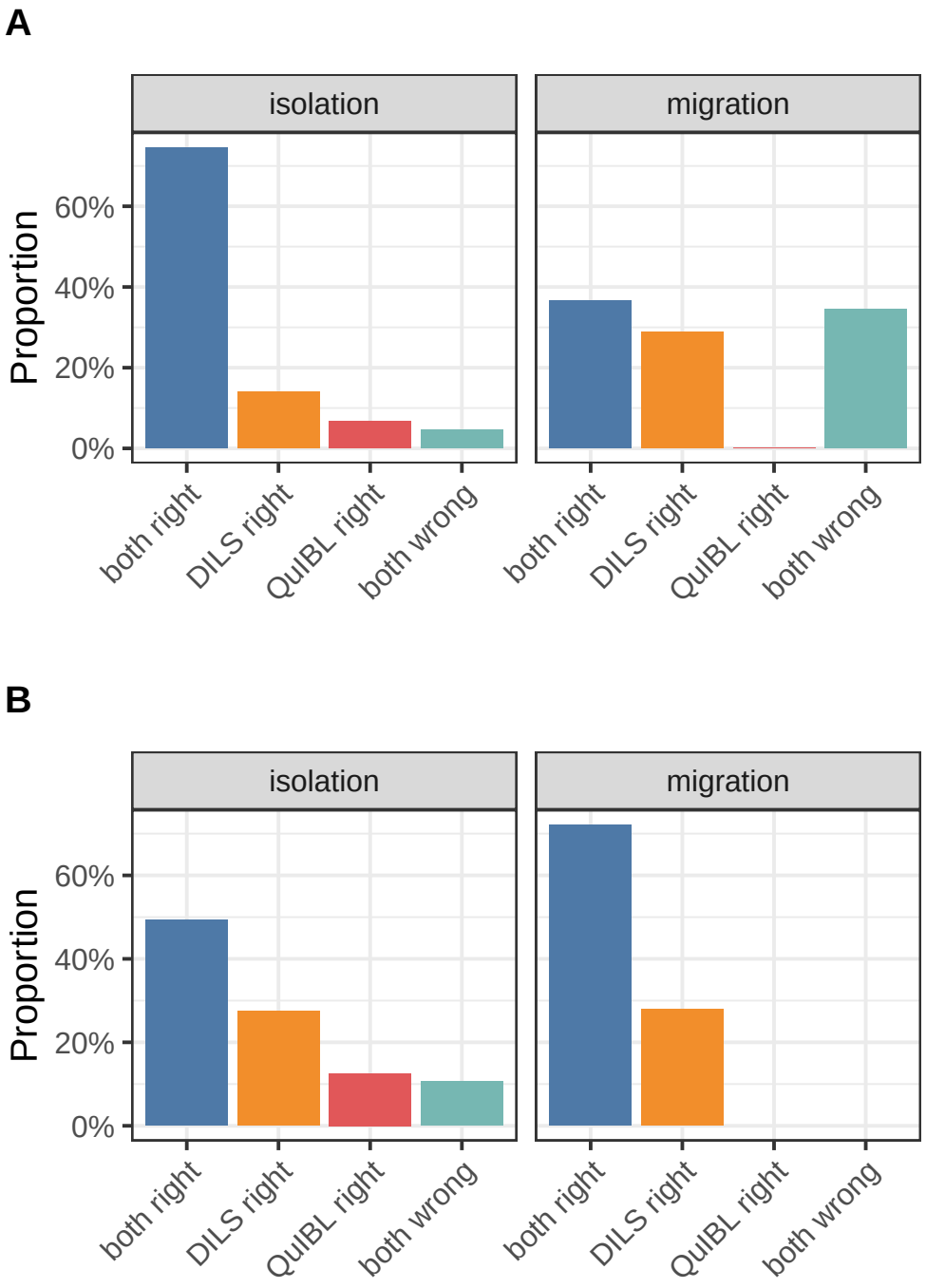

Figure S10: **Comparison of DILS and QuIBL results on simulated datasets with migration (IM_{4pop}, SC_{4pop}; Fig. S9) and without migration (SI_{4pop}, AM_{4pop}; Fig. S9))**. Colors represent the proportions of simulations where: both methods were correct (blue), only DILS was correct (orange), only QuIBL was correct (red), and both methods were incorrect (green). (**A**) Results across all explored parameters. (**B**) Results for parameters where more than 10% of loci are affected by migration and $N_e.m > 0.25$.

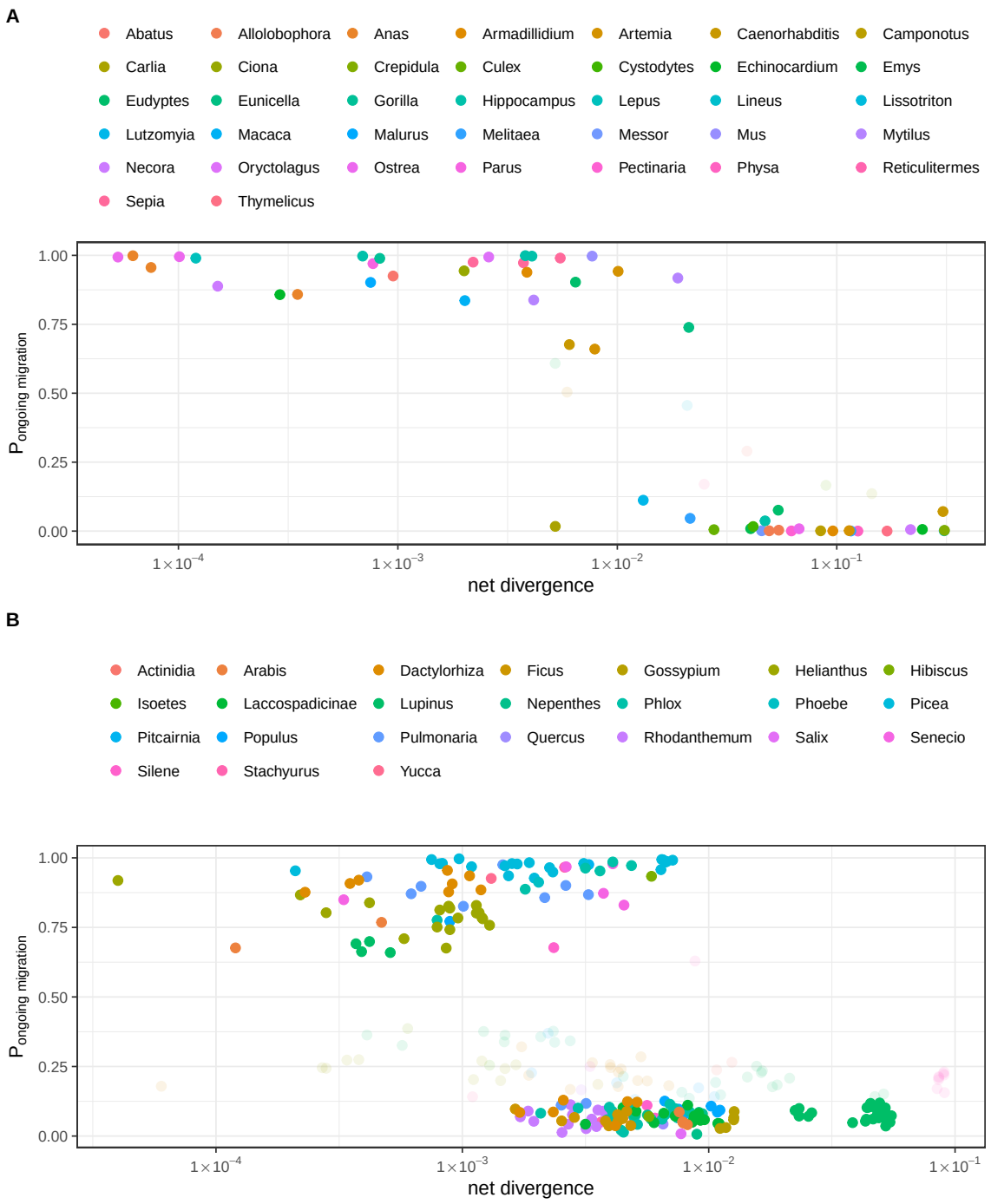

Figure S11: **Relationship between mean net divergence and posterior probability for ongoing migration.** Each point corresponds to a pair of (**A**) animals or (**B**) plants. x-axis: average net divergence. y-axis: posterior probability for ongoing migration attributed by our ABC framework. Colours 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 (*14*). 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).

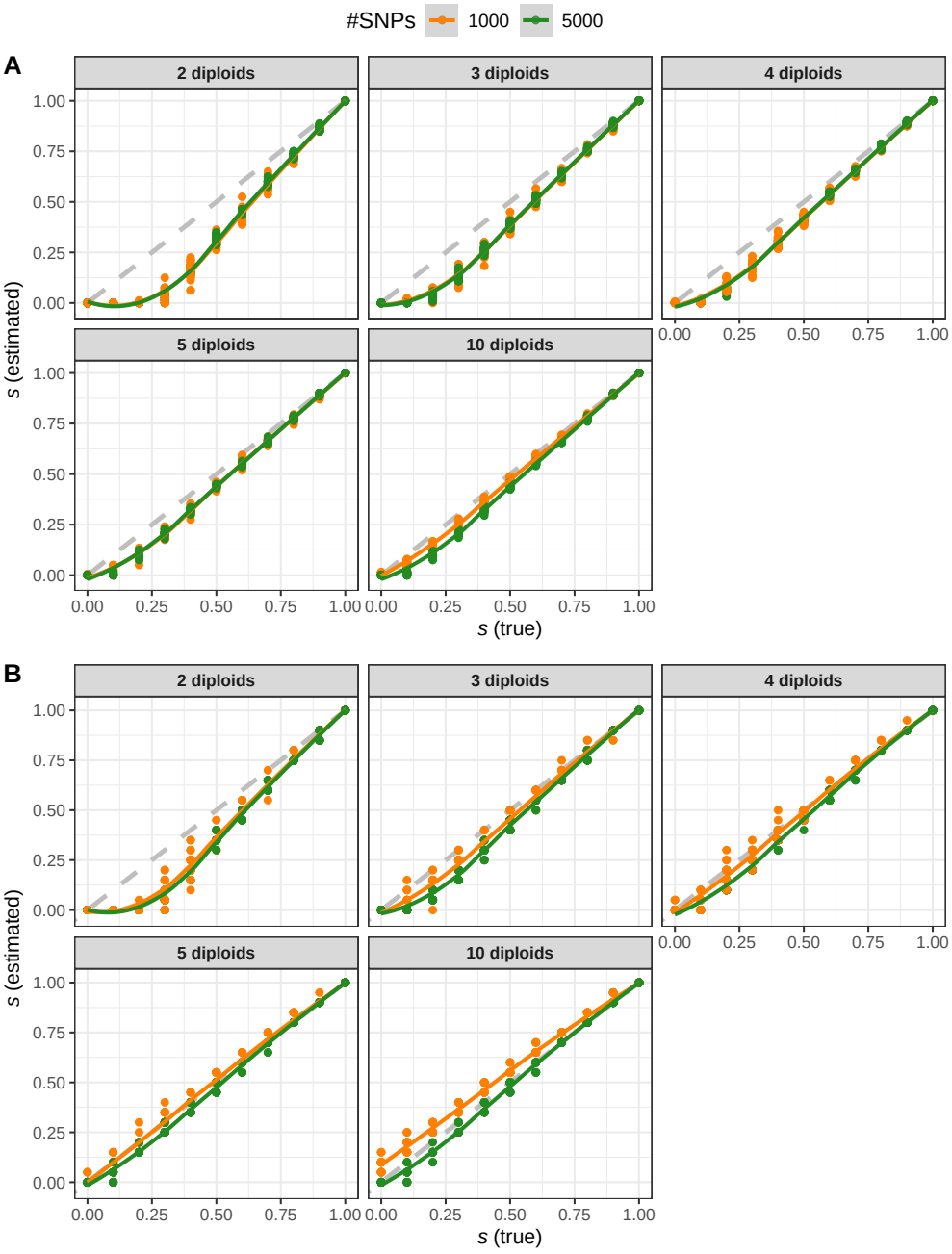

Figure S12: **Evaluation of selfing rate (*s*) estimation from simulated genomic data.**The x-axis represents the true selfing rate (*s*) used for simulations, while the y-axis shows the selfing rate estimated using our custom method (*24*). (**A**) Mean selfing rate estimated across all individuals. (**B**) Maximum selfing rate estimated among individuals. Orange and green points represent simulated datasets with 1,000 and 5,000 SNPs, respectively. The number of sampled diploid individuals is indicated in each sub-panel.

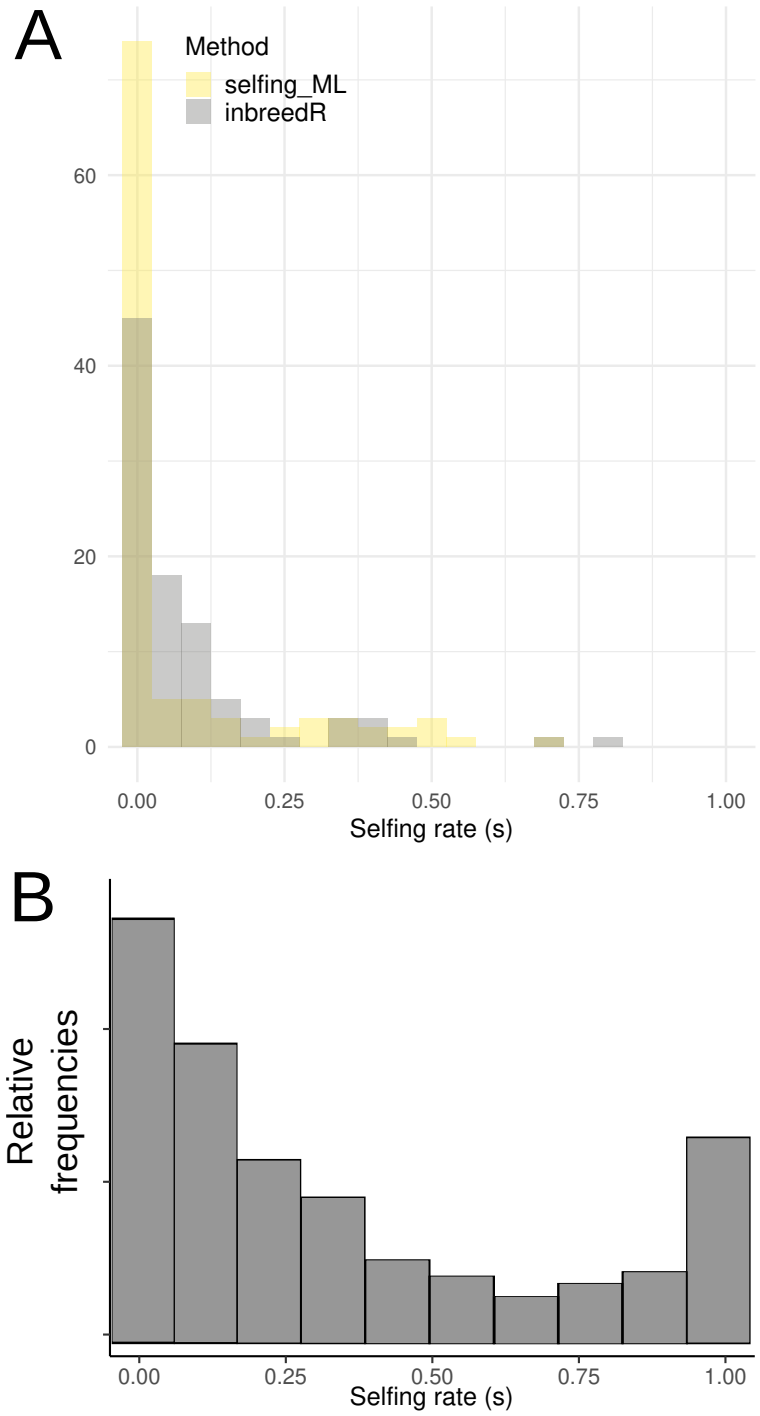

Figure S13: **Distribution of selfing rates (*s*) estimated from molecular data.** (**A**) Distribution of species-averaged selfing rates among plant species used in the plant-*versus*-animal comparison, estimated using `selfing_ML` (yellow) and `inbreedR` (*23*) (grey). (**B**) Meta-analysis of selfing rates across 329 plant species not related to our study, modified from the study by Igic and Kohn in 2006 (*97*).

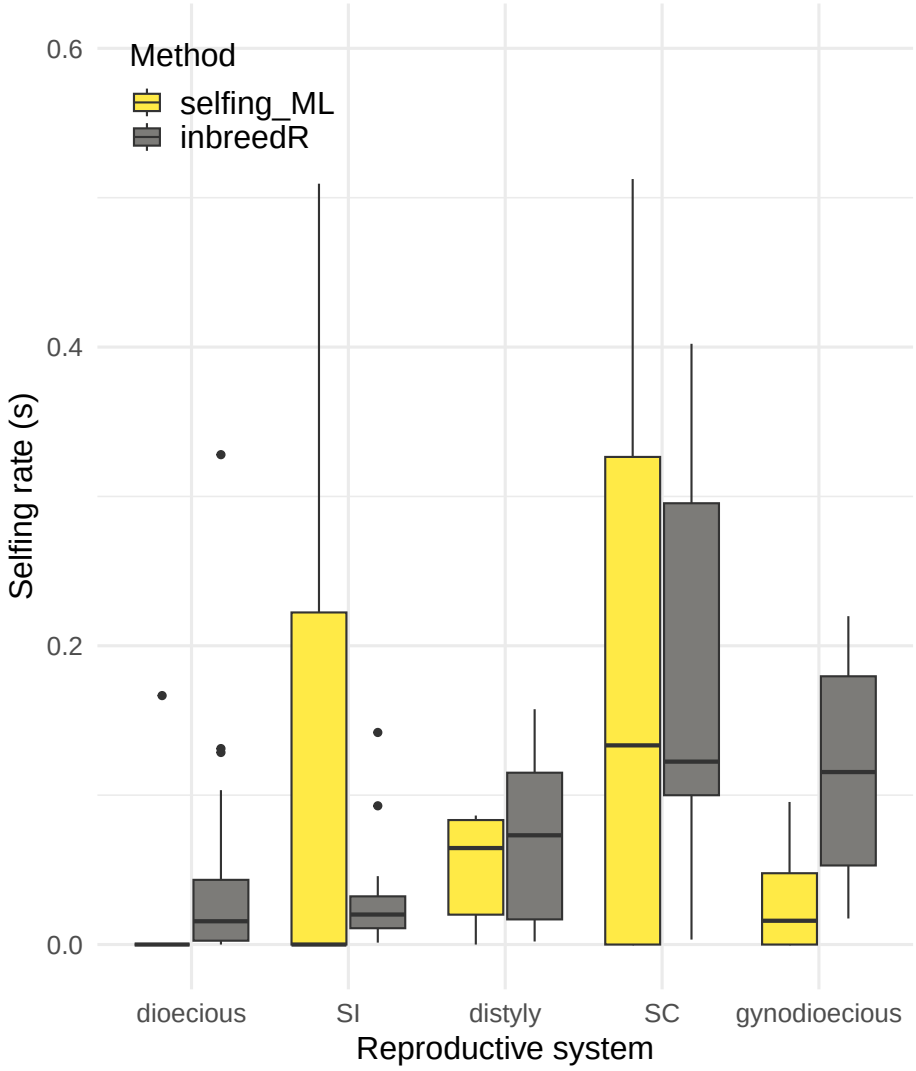

Figure S14: **Distribution of estimated selfing rates (*s*) across reproductive systems in 56 plant species.**

The reproductive systems include: 18 dioecious, 19 self-incompatible (SI), 5 distylous, 10 self-compatible (SC), and 4 gynodioecious species. Selfing rate estimates were obtained using two methods: **yellow**, the custom-developed `selfing_ML` method; and **grey**, the `inbreedR` package (*23*). Each boxplot summarizes the distribution of species-averaged selfing rates per reproductive system (*25*) and method.

Table S1: List of retained NCBI datasets.

bioproject	genus	species	n	type of data	source
PRJNA318567	*Actinidia*	*arguta*	3	WGS	(*43*)
		arguta giraldii	2		
		chinensis	4		
PRJEB33482,	*Arabis*	*nemorensis allop.*	6	RNA	(*44*)
PRJEB39992		*nemorensis symp.*	6		
		sagittata allop.	10		
		sagittata symp.	15		
PRJNA489792	*Dactylorhiza*	*euxina*	5	RAD	(*45*)
		foliosa	2		
		fuchsii	30		
		iberica	2		
		incarnata	31		
		saccifera	4		
		sambucina	3		
		viridis	3		
PRJNA445222	*Ficus*	*arfakensis*	14	RAD	(*46*)
		itoana	13		
		microdictya	15		
		trichocerasa	15		
		t. pleioclada	26		
PRJNA539957	*Gossypium*	*australe*	4	WGS	(*47*)
		bickii	3		
		nelsonii	3		
		robinsonii	2		
		sturtianum	6		
PRJNA532579	*Helianthus*	*annuus NoTex*	15	WGS	(*48*)
		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	(*49*)
		laevis	4		
		moscheutos	5		
PRJNA483403	*Isoetes*	*lacustris*	9	RAD	(*50*)
		echiospora	3		
PRJNA244607	*Howea*	*belmoreana*	40	RNA	(*51*)
		forsteriana	39		
PRJNA528594	*Linospadix*	*monostachyos*	18		(*52*)
		minor	9		
		apetiolatus	6		
		palmerianus	6		
PRJNA318864	*Lupinus*	*ballianus*	2	RNA	(*102*)
		bandelierae	2		
		misticola	2		
PRJEB37794	*Nepenthes*	*albomarginata*	3	RAD	(*53*)
		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	(*54*)
		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	(*55*)
		bournei	12		
PRJNA807675	*Pitcairnia*	*albiflos*	9	RAD	(*56*)
		staminea	12		
PRJNA392950,	*Picea*	*brachytyla*	4	RNA	(*57, 58*)
PRJNA401149,		*b. complanata*	5		
PRJNA378930,		*likiangensis likiangensis*	5		

PRJNA301093		*l. linzhiensis*	5		
		l. rubescens	5		
		purpurea	5		
		wilsoni	5		
PRJNA612655	*Populus*	*adenopoda*	5	WGS	(*59*)
		alba	5		
		davidiana	5		
		qiongdaoensis	3		
		rotundifolia	4		
		tremula	5		
PRJNA544114	*Pulmonaria*	*helvetica*	24	RAD	(*60*)
		mollis	10		
		montana	4		
		obscur	11		
		officinalis	6		
PRJNA639507	*Quercus*	*berberidifolia*	63	RAD	(*61*)
		chrysolepis	80		
PRJNA554975	*Rhodanthemum*	*redieri redieri*	4	RAD	(*62*)
		r. humbertii	7		
		quezelii quezelii	4		
		q. jallabenense	4		
		arundanum mairei	8		
		a. arundanum	27		
PRJNA429746	*Salix*	*helvetica*	10	RAD	(*63*)
		purpurea	10		
PRJNA549571	*Senecio*	*aethnensis*	6	RNA	(*64*)
		aethn. X chrys.	14		
		chrysanthemifolius	6		
		squalidus	28		

PRJNA295359	*Silene*	*dioica*	2	RNA	(*65, 66*)
		latifolia	2		
		nutans E1	4		
		n. W1	4		
		n. W2	4		
		n. W3	4		
PRJNA553020	*Stachyurus*	*chinensis*	6	RNA	(*67*)
		retusus	2		
		yunnanensis	4		
PRJNA329381	*Yucca*	*brevifolia*	24	RAD	(*68*)
		jaegeriana	39		

Table S2: Log-likelihood Ratio Test for logit models fitted to plant and animal datasets (Fig. 1)

Model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	***P*-value**
M_0	−91.55841	−11.28733	−4.50495	0.00312	
M_{plants}	−61.60692	−16.15385	−6.27316	0.00266	
M_{animals}	−8.01012	−11.13402	−6.09245	0.01488	
					$< 1 \times 10^{-4}$

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{plants}, and M_{animals}.

$\boldsymbol{\beta_0}$: estimated intercept ($\log_{10}$ scaled).

$\boldsymbol{\beta_1}$: estimated coefficient ($\log_{10}$ scaled).

$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = 10^{-\beta_0/\beta_1}$).

***P*-value**: probability that by random chance, the absolute difference between $\ell(M_{\text{plants}}) + \ell(M_{\text{animals}})$ and $\ell(M_0)$ exceeds the observed value, estimated from 10,000 random permutations.

Table S3: Log-likelihood Ratio Test for logit models fitted to plant and animal datasets obtained by RNA-sequencing only

Model	$\boldsymbol{\ell}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{X_{p=0.5}}$	***P*-value**
M_0	−34.28871	−11.79074	−5.31189	0.00603	
M_{plants}	−18.28419	−19.81859	−8.47952	0.00460	
M_{animals}	−4.29462	−16.75334	−9.58339	0.01786	
					$< 1 \times 10^{-4}$

$\boldsymbol{\ell}$: log-likelihoods of models M_0, M_{plants}, and M_{animals}.

$\boldsymbol{\beta_0}$: estimated intercept ($\log_{10}$ scaled).

$\boldsymbol{\beta_1}$: estimated coefficient ($\log_{10}$ scaled).

$\boldsymbol{X_{p=0.5}}$: inflection point beyond which, for any level of divergence, less than 50% of pairs are expected to be connected by gene flow ($X_{p=0.5} = 10^{-\beta_0/\beta_1}$).

***P*-value**: probability that by random chance, the absolute difference between $\ell(M_{\text{plants}}) + \ell(M_{\text{animals}})$ and $\ell(M_0)$ exceeds the observed value, estimated from 10,000 random permutations.

Table S4: Log-likelihood Ratio Test for logit models assessing factor effects on reproductive isolation dynamics within plants

Factor	**Comparison**	**Group 1**	**Group 2**	$\ell(M_0)$	$\ell(M_1)$	$\ell(M_2)$	***P*-value**
Life form	Herbs *versus* Trees	Herb	Tree	−59.78	−36.095	−22.433	0.261
	H–L–S *versus* Trees	H–L–S	Tree	−61.607	−37.662	−22.433	0.1937
Selfing rate	s_h–s_h *vs.* s_l–s_l	s_h–s_h	s_l–s_l	−46.2	−17.2	−28.7	0.63
(`selfing_ML`)	s_h–s_h *vs.* s_h–s_l	s_h–s_h	s_h–s_l	−32.53	−17.17	−12.47	0.045
	s_l–s_l *vs.* s_h–s_l	s_l–s_l	s_h–s_l	−43.12	−28.66	−12.47	0.1515
Selfing rate	s_h–s_h *vs.* s_l–s_l	s_h–s_h	s_l–s_l	−22.21	−12.79	−8.38	0.4
(`inbreedR`)	s_h–s_h *vs.* s_h–s_l	s_h–s_h	s_h–s_l	−41.58	−12.79	−26.23	0.07
	s_l–s_l *vs.* s_h–s_l	s_l–s_l	s_h–s_l	−35.69	−8.38	−26.23	0.316

H–L–S: group of species including herbs, lianas, and shrubs.

s_h and s_l: species with selfing rates above and below the median, respectively.

Pairs are grouped as s_h–s_h (high–high), s_l–s_l (low–low), or s_h–s_l (mixed).

Selfing rates were estimated using two approaches: the custom method `selfing_ML` (*24*) and the `inbreedR` package (*23*).

$\ell(M_0)$: log-likelihood of the model fitted to all species pairs; $\ell(M_1)$ and $\ell(M_2)$: log-likelihoods of models fitted separately to each group.

P-value: probability that, by random chance, the absolute difference between $\ell(M_1) + \ell(M_2)$ and $\ell(M_0)$ exceeds the observed value, estimated from 10,000 permutations.