# Rapid establishment of species barriers in plants compared to animals

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

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

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

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

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

## One Sentence Summary

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

# Introduction

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

One approach to studying these speciation dynamics *in natura* is to empirically explore a large continuum of molecular divergence composed by multiple pairs of sister lineages and to determine with model-based demographic inference which ones are currently genetically connected by gene flow (*6*). Introgression leaves detectable signatures in genomes, quantified by statistics commonly used in population genetics, including $F_{\mathrm{ST}}$ (*7*) and derivatives of the ABBA-BABA test (*8, 9*). Although they serve as a foundation for testing the hypothesis of strict

allopatry between two lineages, they are not sufficient on their own to provide a quantification of the timing of gene flow and thus to estimate the current status of reproductive isolation of a pair of taxa. Recent computational methods have allowed the explicit evaluation of alternative evolutionary scenarios, notably to test the occurrence of ongoing gene flow, as well as to test the semi-permeability of species barriers in the genomes (*10, 11*). Applied to 61 pairs of animal taxa, these methods revealed that introgression is frequent until $2\%$ of net divergences (*6*), and can even take place between lineages 14 times more divergent than the human-chimpanzee divergence (*12*).

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

# Results

The present investigation examines the decrease in ongoing gene flow between lineages as a function of their genetic divergence, and compares its dynamics between two main kingdoms of the Tree of Life: plants and animals. For this purpose, we empirically explore a continuum of genetic divergence represented by 61 animal pairs and 280 plant pairs. Genomic data from each pair allows the quantification of molecular patterns of polymorphism and divergence by measuring 39 summary statistics commonly used in population genetics and the joint Site Frequency Spectrum (jSFS). For each observed dataset, we then tested whether the observed set of summary statistics was better reproduced by scenarios of speciation with or without migration by using an approximate Bayesian computation framework (ABC; (*10*)). The same ABC methodology for demographic inferences was employed for both the animal dataset (analyzed in (*6*)) and the plant dataset. The new plant dataset was produced from sequencing reads publicly available for 25 genera distributed in the plant phylogeny (212 pairs of eudicots, 45 of monocots, 21 of gymnosperms, 1 lycophyte and 1 magnoliid; Table S1) and were not chosen on the basis of a preconceived idea of their speciation mode (see supplementary materials A.2). The posterior probability of models with ongoing migration computed by the ABC framework is used to assign a status of isolation or migration to each pair along a continuum of divergence (Fig.1-A), allowing the comparison of speciation dynamics between plants and animals. In contrast to the expected outcomes reported in previous studies (*2, 13–17*), our findings suggest that in comparison to animals, plants exhibit a more rapid cessation of genetic exchange at lower levels of genetic divergence. This is characterized by a swifter transition from population pairs that are best-supported by migration models to those that are best described by isolation models ($P = 4.88 \times 10^{-15}$; Fig.1-A and table S2). Therefore, by fitting a generalized linear model for the migration/isolation status to the plant and animal datasets, as a function of the net molecular

divergence, we determined that at a net divergence of $\approx 0.3\%$ (95% CI: [0.27%-0.47%]), the probability that two plant lineages are connected by gene flow falls below $50\%$, while in animals this inflection point occurs at higher levels of divergence close to $1.8\%$ (95% CI: [1.52%-2%]; Fig.1-A and table S2). The plant dataset comprises genomic data derived from diverse sequencing methodologies, including RAD-sequencing ($n = 117$ pairs), RNA-sequencing ($n = 111$) and whole genome sequencing ($n = 52$), while the animal dataset predominantly consists of RNA-sequencing data ($n = 52$). To control for potential bias in sequencing technologies, we restricted our analysis solely to plant and animal datasets acquired through RNA-sequencing. The key result of a faster cessation of gene flow in plants than in animals is still supported ($P = 5.38 \times 10^{-8}$ and table S3), allowing us to reject the idea that our conclusions are derived from such a methodological bias. The number of pairs within a genus showing robust statistical support for either ongoing migration or current isolation in plants ranged from one to 31 pairs. Therefore, we also investigated a possible effect of sampling bias within the plant dataset. Through random sub-sampling involving a single pair of lineages per plant and animal genus, we demonstrate that the contrast in speciation dynamics between plants and animals consistently persists, also rejecting the idea that our result stems from the over-representation of a genus of plants with highly reproductively isolated lineages (Fig. S7).

To investigate the build-up of species barriers within the genomes of both plant and animal species, we now focus towards pairs supported by ongoing gene flow. Within the range of speciation scenarios considered, the rate of gene flow can be uniform across genomes (i.e., homogeneous) or it can vary locally from one genomic region to another (i.e., heterogeneous; see Fig. S5), contingent, respectively, upon the absence or presence of barrier genes that are expressed (*5*). The ABC framework described earlier allows us to classify animal and plant pairs as experiencing either genomically homogeneous or heterogeneous introgression (*19*). We find that plants experience a faster shift from the absence of barriers to semi-permeable barriers,

the latter occurring at a net divergence of $\approx 0.2\%$ (compared to $\approx 0.6\%$ in animals; Fig.1-B). These findings demonstrate that, in plants, the initial species barriers that generate genomic heterogeneity of introgression rates, as well as the establishment of complete isolation between species, manifest at relatively lower levels of divergence than in animals. This suggests that the speciation process may require fewer mutations in plants than in animals for reproductive isolation to be both initiated and completed.

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

animals ($P = 0.371$).

# Discussion

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

ondary contact scenarios involve a preliminary phase of allopatry which affects the divergence of sister lineages on every marker: molecular and morphological. Such a historical context may thus engender the misconception that plants undergo hybridization events more frequently than animals simply because these events are more conspicuous in plants, as introgression happens more often between morphologically distinct lineages experiencing a secondary contact. This result implies, conversely, that genetic introgression appears to manifest with greater crypticity in animals.

The speciation process clearly does not follow a universal molecular clock, although certain molecular constraints inevitably make the process irreversible once a certain level of divergence is reached (*16*). While light has recently been shed on the rarity of hybrid zones found in plants (*20*), another mystery has now been added: why is the probability of being reproductively isolated greater in plants than in animals given identical genomic divergence? The multi-factorial nature of the speciation process (*24*) exacerbates the methodological limitations of our current approach in producing a simple explanation for the differences observed between plants and animals. Following the first reports based on morphological detection of hybrids and suggesting a greater occurrence of hybridization in plants than in animals, a range of hypotheses have been proposed to explain these observations. One commonly raised argument relates to pre-zygotic isolation, which is believed to exert greater influence on animals, primarily driven by behavioral preferences for reproductive partners (*16*). Another argument is based on the scarcity of heteromorphic sex chromosomes in plants, while they are common in animals (*17*), renowned for their influential role as a preferential sink for genetic barriers to introgression (*25*). While acknowledging the undeniable involvement of these processes, it is crucial to emphasize that they do not serve as definitive or all-encompassing mechanisms governing speciation. Distinctive attributes inherent to plants also provide a favorable context for the accumulation of species barriers within their genomes: *i)* the additional presence of

chloroplasts in plant cells (*26, 27*), *ii)* the prevalence of selfing (*28–31*), *iii)* a certain dependency for reproduction on external pollinators (*32–34*), *iv)* less efficient dispersal modalities as illustrated by the higher intra-specific plant differentiation (*22, 23, 35, 36*) and *v)* stronger haploid selection (*37*). The proposed factors presented here are evidently not mutually exclusive, and it would be misleading to assert that the differences in speciation dynamics between plants and animals can be attributed to a single, easily testable factor. Understanding which properties of plants and animals, acting at the micro-evolutionary scale, lead to such a great disparity in speciation patterns at the macro-evolutionary scale, would benefit from a long-term community-based initiative for integrative speciation research across fields and taxa. Finally, we propose that the methodology employed herein to scrutinize variations in speciation dynamics between plants and animals could be extended to examine other contrasts encompassing diverse life-history traits, such as distinctions between external and internal fertilization, reproductive modes involving self-fertilization *versus* allo-fertilization, or variations in life cycles (haplobiontic *versus* diplobiontic), among others. Such prospective studies would be extremely valuable to better understand the respective roles of these various biological factors in influencing the establishment and maintenance of reproductive isolation.

# References


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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

45. J. Norrell (2017).

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

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

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

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

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

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

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

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

54. D. Ru, *et al.*, *Molecular Ecology* **27**, 4875 (2018).

55. H. Shang, *et al.*, *Philosophical Transactions of the Royal Society B* **375**, 20190544 (2020).

56. S. Grünig, M. Fischer, C. Parisod, *Annals of botany* **127**, 21 (2021).

57. J. Ortego, L. L. Knowles, *Molecular Ecology* **29**, 4510 (2020).

58. F. Wagner, *et al.*, *Molecular phylogenetics and evolution* **144**, 106702 (2020).

59. S. Gramlich, N. D. Wagner, E. Hörandl, *BMC Plant Biology* **18**, 1 (2018).

60. B. Nevado, S. A. Harris, M. A. Beaumont, S. J. Hiscock, *Molecular Ecology* **29**, 4221 (2020).

61. X.-S. Hu, D. A. Filatov, *Molecular ecology* **25**, 2609 (2016).

62. A. Muyle, *et al.*, *Molecular Biology and Evolution* **38**, 805 (2021).

63. Y. Feng, H. P. Comes, Y.-X. Qiu, *Molecular phylogenetics and evolution* **150**, 106878 (2020).

64. A. M. Royer, M. A. Streisfeld, C. I. Smith, *American journal of botany* **103**, 1730 (2016).

65. B. Langmead, S. L. Salzberg, *Nature methods* **9**, 357 (2012).

66. N. Matasci, *et al.*, *Gigascience* **3**, 2047 (2014).

67. J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, *G3: Genes—genomes— genetics* **1**, 171 (2011).

68. J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, W. A. Cresko, *Molecular ecology* **22**, 3124 (2013).

69. J. R. Paris, J. R. Stevens, J. M. Catchen, *Methods in Ecology and Evolution* **8**, 1360 (2017).

70. B. Charlesworth, M. Morgan, D. Charlesworth, *Genetics* **134**, 1289 (1993).

71. N. Barton, B. O. Bengtsson, *Heredity* **57**, 357 (1986).

72. M. A. Beaumont, *Annual review of ecology, evolution, and systematics* **41**, 379 (2010).

73. F. Tajima, *Genetics* **105**, 437 (1983).

74. G. Watterson, *Theoretical population biology* **7**, 256 (1975).

75. F. Tajima, *Genetics* **123**, 585 (1989).

76. M. Nei, W.-H. Li, *Proceedings of the National Academy of Sciences* **76**, 5269 (1979).

77. S. Wright, *Genetics* **28**, 114 (1943).

78. J. Wakeley, J. Hey, *Genetics* **145**, 847 (1997).

79. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, *Molecular biology and evolution* **34**, 1812 (2017).

80. B. Nevado, G. W. Atchison, C. E. Hughes, D. A. Filatov, *Nature communications* **7**, 12384 (2016).

# Acknowledgments

We are grateful to the Institut Français de Bioinformatique (IFB; ANR-11-INBS-0013) thanks to whom bioinformatics analyses and demographic inferences have been carried out.
The authors would like to thank Hélène Morlon, Bert Van Boxlaer, John Pannell and Martin Lascoux for their constructive feedback.

**Funding:** This work was supported by grant from I-SITE ULNE – Ghent University.

Y.VdP. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). **Authors contributions:** F.M (data curation, formal analysis, investigation, visualization, writing), Z.P (data curation), P.T (resources, data curation, supervision), C.F (writing), Y.V.d.P (funding acquisition, supervision, writing), X.V (funding acquisition, supervision, writing), C.R (funding acquisition, conceptualization, methodology, supervision, investigation, visualization, writing). **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. The assembled datasets, the list of references used for mapping and the results of demographic inference are deposited in `Zenodo` with the DOI `10.5281/zenodo.8028615` (*38*). Scripts for bioinformatic treatment of raw sequencing data are available from GitHub `https://github.com/Ladarwall/Greenworld`.

# Supplementary Materials

Tables S1-S3

Figures S1-S7

# Figures

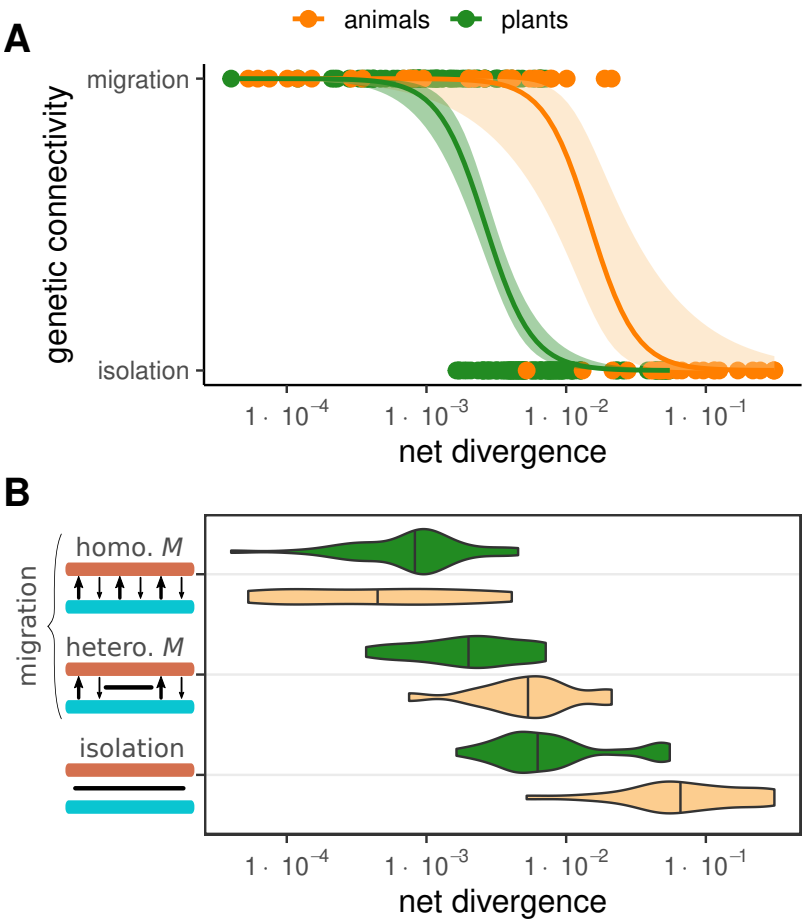


Figure 1: **Genomic patterns of introgression along a divergence continuum in plants *versus* animals.**
Estimation of the average genomic divergence and the migration/isolation status was performed for 280 pairs of plant species/populations (green) and compared to 61 animal pairs (orange) analysed using the same ABC procedure (*6*).
**A.** x-axis: average net divergence within a pair. y-axis: best supported model in a comparison between ongoing migration and current isolation. Each point represents a pair of populations/species. Curves represent the logit models fitted to the plant and animal data.
**B.** Distribution of the average net divergence of plant (green) and animal (orange) pairs whose genomic data are best explained by homogeneous (homo. M) or heterogeneous (hetero. M) distributions of migration rates across the genome, or by complete genetic isolation (isolation). y-axis: blue and brown bars symbolize homologous chromosomes within a studied pair. Black arrows symbolize genome regions connected by gene flow. Black bars symbolize local effects of barriers against gene flow.

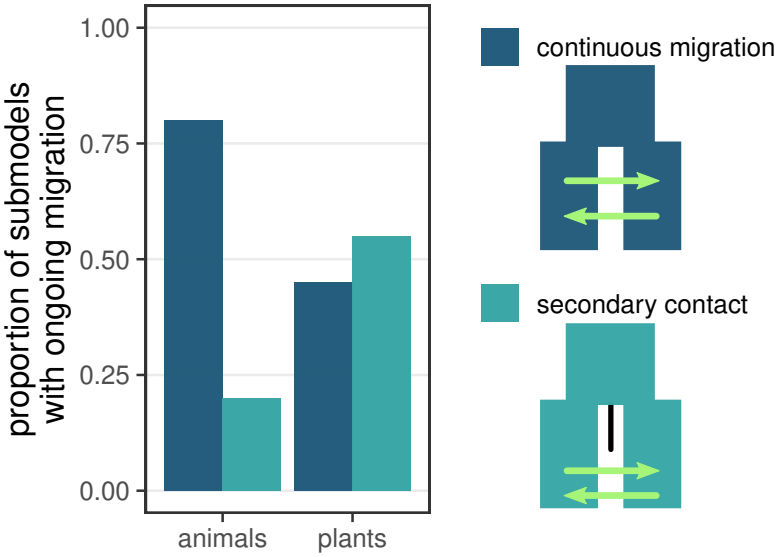


Figure 2: **Temporal models of ongoing migration in plants and animals.**
Strong statistical support of ongoing migration was observed in 82 pairs of plants and 30 pairs of animals. For every pair supported by ongoing gene flow, sub-models were compared to distinguish between continuous migration (dark blue) *versus* secondary contact (light blue).

# **Supplementary Materials for:** Rapid establishment of species barriers in plants compared to animals

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

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

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

# A Materials and Methods

## A.1 Animal dataset

The animal data come from the Roux et *al.* (2016) study (*6*). They consist essentially of non-model animal populations/species, initially selected without any particular knowledge about the demographic history, and were sampled from natural populations. These data were produced by RNA sequencing, and only synonymous positions were retained for statistical inferences.

## A.2 Plant sampling

Raw data used in this work comes from previously published studies (*39–64*). The following criteria were applied to identify datasets in plants:

1. Currently diploid genomes.
2. High-throughput sequencing, i.e, RNA-seq, RAD-seq or whole genome sequencing (WGS).
3. Freely available from `NCBI`.
4. Individuals sampled from natural populations (geographic distribution represented in Fig. S1).
5. A minimum of two sampled populations/species per genus.
6. A minimum of two sequenced individuals per sampled population/species.

Datasets fitting these criteria were examined through exploration of literature found *via* Google Scholar (`https://scholar.google.com`), `NCBI` (`https://www.ncbi.nlm.nih.gov/Traces/study/`) and `DDBJ` (`https://ddbj.nig.ac.jp/search`).

Finally, 118 different plant species/populations from 25 different genera were retained for the demographic analysis according to our criteria (Table S1), allowing 280 pairwise demographic analyses to be carried out. These comparisons cover all possible pairs within each genus. No comparisons are made between different genera, with the exception of comparisons within the *Laccospadicinae* (*Howea* and *Linospadix*) due to their relatively small genetic distance.

## A.3 Assembly, read mapping and genotype calling

For the plant datasets: reads and metadata were downloaded using SRA-Toolkit, version 2.11.0 (`https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit`). Here we separate plant projects for which we worked with synonymous positions (from RNA-seq: $n$=7 genera and WGS: $n$=4) from those for which we could not (from RAD sequencing: $n$=13):

### A.3.1 Reads from RNA-seq and WGS.

In line with the animal dataset (*6*), the bioinformatic strategy applied to the plant data is to retain synonymous positions. Reads for a given population/species pair were therefore mapped to a reference transcriptome with the bowtie2 program version 2.4.2 (*65*): either taken from the `1KP project` (*66*) if a species of the same genus is represented there (`https://db.cngb.org/onekp/search/`), or taken from the data associated with the original articles when available (Table SS1). Every position (variants and invariants) were called with a minimum of 8 reads using Reads2SNP 2.0, the uncalled low-quality positions were then coded as "N". The resulting fasta file was used for each population/species as the input file for the demographic inferences.

### A.3.2 Reads from RAD-seq.

Loci were assembled for each RAD-seq dataset using `Stacks` 2.6 (*67, 68*). Combinations of parameters were explored following Paris et *al.* 2017 (*69*) to maximise the amount of biological information retained. Using the two or four samples with the highest amount of available data per lineage, assemblies were built using *denovo_map.pl* (`Stacks`) with different combinations of parameters: the minimum depth for a stack to be valid (-*m*, ranging from 3 to 5), the number of mismatches allowed between stacks within individuals (-*M*, ranging from 1 to 6) and the number of mismatches allowed between stacks between individuals (-*n*, set to *M* or *M* + 1), for a total of 36 combinations. In addition, loci that were missing in at least 20% of the samples per population were withdrawn with the argument *–min-samples-per-pop 0.80* (i.e. only loci with the information for all samples were kept, as populations were composed of two or four samples). The number of polymorphic loci was plotted as a function of the different combinations of parameters using a homemade `R` script. For each dataset, a combination was selected in function of the trade-off between maximising the number of polymorphic loci and minimising the parameter values to produce a reference set of loci for each species/population pair. Reads were mapped on this reference with bowtie2 version 2.5.1, and variants were called with Reads2SNP 2.0 in the same way as "RNA-seq and WGS" datasets.

## A.4 Demographic inferences

Model comparisons were carried out using the approximate Bayesian computation (ABC) framework applied in the animal study (*6*) and distributed under the name DILS (for Demographic Inferences with Linked Selection (*10*)). Here we describe how DILS works.

### A.4.1 Compared models

The primary objective of our demographic analysis is to determine which historical scenario explain the best a given dataset. The term dataset here refers to a pair of populations/species (comprising either two animal or two plant lineages, for which genomic data are described by an array of summary statistics (see section A.4.2). In our ABC methodology, we discern two categories of models.

**Four demographic Models:** Each of these models describes the subdivision of an ancestral population into two daughter populations (Fig. S5-A). The three populations have independently assigned effective population sizes. The differences between these four models concern the historical patterns of gene flow between two divergent populations, as depicted in figure S5. These models encompass continuous migration (CM), and secondary contact (SC), strict isolation (SI) and ancient migration (AM) :

- models with ongoing migration
  - continuous migration (CM)
  - secondary contact (SC)
- models with current isolation
  - strict isolation (SI)
  - ancient migration (AM)

Notably, the former two models entail ongoing gene flow between the two populations, while the latter two do not. Models with past (AM) or recent (CM and SC) migration assume gene flow between sister populations/species in both directions, at two independently assigned rates.

**Models of Linked Selection:** Effects of linked selection have been taken into account using a genomic model that encompasses: (a) heterogeneous effective population size across the

genome (*hetero. N*), which closely approximates the influence of background selection by down-scaling *Ne* (*70*); and/or (b) heterogeneous migration rate across the genome (*hetero. M*) to account for the effects of selection against hybrids (*71*). The modeling framework employed in this study does not consider the effects of positive selection on linked loci (i.e., genetic hitchhiking).

Within the *hetero. N* genomic model, the variable effective size among loci is assumed to conform to a re-scaled Beta distribution. In essence, all populations share a common Beta distribution with two shape parameters drawn from uniform distributions. However, each population is independently re-scaled by distinct *Ne* values, which are drawn from uniform distributions. Conversely, the *homo. N* genomic model assumes that all loci from the same genome share the same effective population size, and this parameter is independently estimated in all populations. This homogeneous model implies that the genomic landscape remains unaffected (or is uniformly affected) by background selection.

The *hetero. M* genomic model implements local reduction of gene flow in the genome. Variation in migration rates among loci is thus modeled by employing a bimodal distribution where a proportion of loci, drawn from a uniform distribution in ]0-1[, is linked to barriers (i.e., $N.m = 0$), while the loci unaffected by species barriers are associated to an effective migration rate *N.m* drawn from a uniform distribution. In the *homo. M* model, a single migration rate *N.m* per direction is universally shared by all loci in the genome.

Subdivisions of the four demographic models (CM, SC, SI and AM) into various genomic submodels were made to accommodate for the effect of linked selection. Heterogeneity in effective population size was a universal consideration across all four models, while heterogeneity in migration rate was specifically accounted for in models exhibiting gene flow (i.e., CM, AM, and SC). Therefore, the SI model was divided into two submodels (*homo. N* or *hetero. N*), while the AM, CM, and SC models were divided into four submodels:

1. *homo. N* and *homo. M*

2. *homo. N* and *hetero. M*

3. *hetero. N* and *homo. M*

4. *hetero. N* and *hetero. M*

For a comprehensive description of all prior distributions employed in this study, please refer to Section A.4.3.

### A.4.2 Summary statistics

ABC is a statistical inferential approach based on the comparison of summary statistics derived from simulated and observed datasets (*72*). We present a comprehensive description of the statistics computed within our framework. The following summary statistics are calculated for each locus:

- The number of bi-allelic polymorphisms in the alignment including all sequenced copies in the 2 species/populations
- Pairwise nucleotide diversity $\pi$ (*73*)
- Watterson's $\theta$ (*74*)
- Tajima's *D* (*75*)
- The proportion of sites displaying fixed differences between the populations/species ($S_f$)
- The proportion of sites featuring polymorphisms exclusive to a specific population/species ($S_{xA}$ and $S_{xB}$)

- The fraction of sites with polymorphisms shared between the two populations/species ($S_s$)
- The number of successive shared polymorphic sites
- Raw divergence $D_{xy}$ between the two populations/species (*76*)
- Net divergence $D_a$ between the two populations/species (*76*)
- Relative genetic differentiation between the two populations/species quantified by $F_{ST}$ (*77*)

For the ABC analysis, we used the means and variances of these statistics calculated over all the available loci. Additionally, we utilize the joint Site Frequency Spectrum (jSFS (*78*)) to summarize the data, specifically capturing the count of single-nucleotide polymorphisms (SNPs) where the minor allele occurs in each bin covering the jSFS. Because of the absence of outgroup lineages, jSFS were folded. Singletons are deliberately excluded from the jSFS to mitigate potential inference biases arising from sequencing errors. Each of the non-excluded bin of the jSFS is used as a descriptive statistics in the ABC analysis.

We supplement this set of summary statistics with measures taken on all the loci:

- Pearson's correlation coefficient for $\pi$ between species
- Pearson's correlation coefficient for $\theta$ between species
- Pearson's correlation coefficient between $D_{xy}$ and $D_a$
- Pearson's correlation coefficient between $D_{xy}$ and $F_{ST}$
- Pearson's correlation coefficient between $D_a$ and $F_{ST}$
- Proportion of loci with both $S_s$ and $S_f$ sites

- Proportion of loci with $S_s$ sites but no $S_f$
- Proportion of loci without $S_s$ sites but with $S_f$
- Proportion of loci with neither $S_s$ nor $S_f$ sites

The summary statistics obtained from both the empirical data sets (i.e., plants and animals) and the data sets simulated under the demographic models (Fig. S5) were calculated with the same scripts implemented in DILS.

### A.4.3 Configuration file

DILS was run using the following parameter values:

- mu = $7.31 \times 10^{-9}$
- useSFS = 1
- barrier = bimodal
- $max_N_tolerated = 0.25$
- $L_{min} = 10$
- $n_{min} = 4$
- $rho_over_theta = 0.2$
- uniform prior for $N$ between 0 and $N_{max}$ individuals
- uniform prior for $T_{split}$ between 0 and $T_{max}$ generations
- uniform prior for migration rate $N.m$ between 0 and 10 migrants per generations

Where:

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

$\pi_A$ and $\pi_B$ being the Tajima's $\theta$ (*73*) for species A and B respectively (for a given pair).

$$T_{max} = 5 \times \frac{D_a}{2\mu}$$

$D_a$ being the net divergence (*76*).

### A.4.4 Returned quantities

At the end of the analysis, DILS returns the posterior probability of ongoing migration *versus* of current isolation. The probability of ongoing migration corresponds to the relative probability of all models including ongoing migration (Secondary Contact, Continuous Migration) and their sub-models (heterogeneity and genomic homogeneity for migration and effective size); while the probability of current isolation corresponds to all models and sub-models with current isolation (Strict Isolation, Ancient Migration). These quantities are used to produce the relationships between the net divergence and the posterior probability of migration (Fig. S6). For each pair of populations/species, three statuses are then assigned:

1. Strong support for genetic isolation: we identify strong statistical support for genetic isolation when our ABC framework yields a posterior probability $P_{\text{migration}} < 0.1304$. This threshold was empirically determined by the robustness test conducted in (*6*).

2. Strong support for ongoing migration: strong statistical support for ongoing migration is indicated when the posterior probability $P_{\text{migration}} > 0.6419$, also empirically determined in (*6*).

3. Ambiguity: statistical ambiguity, denoting situations where our ABC framework does not strongly support either migration or isolation, i.e, when the risk of assigning an analysed pair to a wrong status is greater than 5%.

Pairs for which support was inconclusive were excluded from further analysis. The remaining pairs were categorized either as exhibiting 'migration' or 'isolation,' as illustrated in Figure 1-A, allowing the 'migration' status to be treated in a logistic regression (see section A.5).

## A.5 Logistic regression

To study speciation dynamics, we examine reduction in the proportion of plant or animal pairs receiving strong support for models with migration as a function of time (measured here by the net molecular divergence). For this purpose, we modeled $\mathbf{Y}_i$ (the binary status 'isolation' or 'migration' best fitting the data) as a function of $\mathbf{X}_i$ (the average net genomic divergence) by using a generalized linear model (GLM) *via* a linked binomial function:

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

where $\beta_0$ represents the intercept and $\beta_1$ the coefficient reflecting the effect of genomic divergence on the isolation/migration status coded as 0 and 1, respectively. The fitted model is used to predict $p_i$, the proportion of pairs of populations/species that are currently connected by gene flow (migration status) for a given level of divergence $\mathbf{X}_i$.

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

Reversely, we can determine the divergence level $\mathbf{X}$ for which a given proportion $p_i$ of pairs are connected by gene flow:

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

We are interested in comparing the inflection point, i.e, the level of divergence above which more than $50\%$ of species pairs are genetically isolated, between plants and animals. Thus, for a given fitted model, this point corresponds to a divergence level $X = -\frac{\beta_0}{\beta_1}$.
The log-likelihood function $\ell$ of the migration/isolation status $\mathbf{Y}$ given the average net molecular divergence $\mathbf{X}$ is then obtained to evaluate the fit of a model to the observed data:

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

We can now test whether the sigmoid of plants is significantly different from that of animals, thereby testing if plants and animals share the same speciation dynamic. For this purpose, three models are fitted and associated to log-likelihood $\ell$:

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

2. $M_{plants}$: model fitted to the plants data only.

3. $M_{animals}$: model fitted to the animals data only.

Thus, for $M_0$ we fitted a GLM to the entire dataset comprising both plants and animals, after having retained only demographic inferences for which the ABC analysis produced strong statistical support for ongoing migration or current isolation, following the test of robustness applied in Roux et al. (*6*). In that sense, pairs of plants and animals with ambiguous support

for isolation or migration were excluded from all GLM regressions. The log-likelihood $\ell(M_0)$ was then estimated for the whole dataset comprising both plants and animals by using formula **1** where:

- $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ represent for $M_0$ the coefficient of the model fitted to the whole plants and animals dataset by using the **glm** function (family = 'binomial') implemented in `R`.
- $\mathbf{X}_{1,i}$ represents the series of observed net divergence values.
- $\mathbf{y}_i$ represents the series of inferred isolation/migration status.

For $M_{plants}$ and $M_{animals}$, we fitted a GLM model only to data from the corresponding kingdom. We then estimated the log-likelihoods $\ell(M_{plants})$ and $\ell(M_{animals})$ as for $M_0$.

Finally, we conducted a comparison between the log-likelihood $\ell(M_0)$ and the combined log-likelihood $\ell(M_{plants})+\ell(M_{animals})$, which is derived from the summation of log-likelihoods obtained by fitting independent models to each respective kingdom. The significance of the difference between $\ell(M_0)$ and $\ell(M_{plants}) + \ell(M_{animals})$ was evaluated using a log-likelihood ratio test. Specifically, twice the absolute difference of the log-likelihood between $\ell(M_0)$ and $\ell(M_{plants})+\ell(M_{animals})$ is approximately $\chi$-squared distributed. The *P*-value returned by the `R` function **pchisq** corresponds to the probability of observing $2.|\ell(M_0)-\ell(M_{plants})-\ell(M_{animals})|$ in a $\chi$-squared distribution with two degrees of freedom (Table S2).

## A.6 Testing for a phylogenetic effect

To control for the variation in the number of pairs between genera, we carried out 5,000 animal-plant comparisons as for Fig. 1 but by randomly selecting a single pair per animal and plant genus. Over these 5,000 sub-samples, the relative positions of the sigmoids were compared *via* the inflection points of the models fitted to the plant *versus* animal sub-samples. The inflection

point was estimated as being $-\frac{\beta_0}{\beta_1}$. We find that the inflection point is found systematically at lower levels of divergence in plants than in animals (Fig. S7).

## A.7 Testing for a sequencing technology effect

Out of the total dataset comprising 280 pairs of plants and 61 pairs of animals, 210 plant pairs and 54 animal pairs exhibited strong statistical support for migration or isolation based on ABC model comparison. These retained datasets encompass a diversity of sequencing methodologies. Specifically, within plants, among the 210 retained pairs: 90 pairs were acquired through RAD-sequencing, 86 pairs through RNA-sequencing, and 34 pairs through whole genome sequencing. In the case of animals: 46 pairs were derived from RNA-sequencing, while 8 pairs were the result of Sanger sequencing. To assess the potential influence of sequencing techniques, we determined whether the observed differences in dynamics between plants and animals, as previously reported for the entire dataset, remained consistent when considering only the data generated exclusively through RNA sequencing. This choice was motivated by the fact that RNA-sequencing is the sole sequencing technique shared by both biological kingdoms under study. By retaining only the data from RNAseq, we maintain a statistically significant support for a more rapid cessation of gene flow in plants than in animals, despite a *P*-value that increases from $4.88 \times 10^{-15}$ (Table S2) to $5.38 \times 10^{-8}$ (Table S3).

## A.8 Geography

Geographical (geodesic) distance in meters was measured using GPS coordinates provided in the metadata when available, using the **distGeo** function in the `R` package *geosphere*. For a given pair of populations/species A and B, this distance corresponds to the distance between the two geographically closest individuals. In the case of sampled sympatric pairs, and if a single coordinate was provided by the authors for all individuals A and B, we consider a distance of

10m in line with current sampling practices to reduce relatedness. Among the 25 plant genera under examination, our review of the literature has not yielded information pertaining to the geographical origins of specimens from *Gossypium*.

## A.9 Data availability

All the assembled datasets, the list of references used for mapping and the results of demographic inference are deposited in `Zenodo` with the DOI `10.5281/zenodo.8028615` (*38*).

# B Supplementary Figures

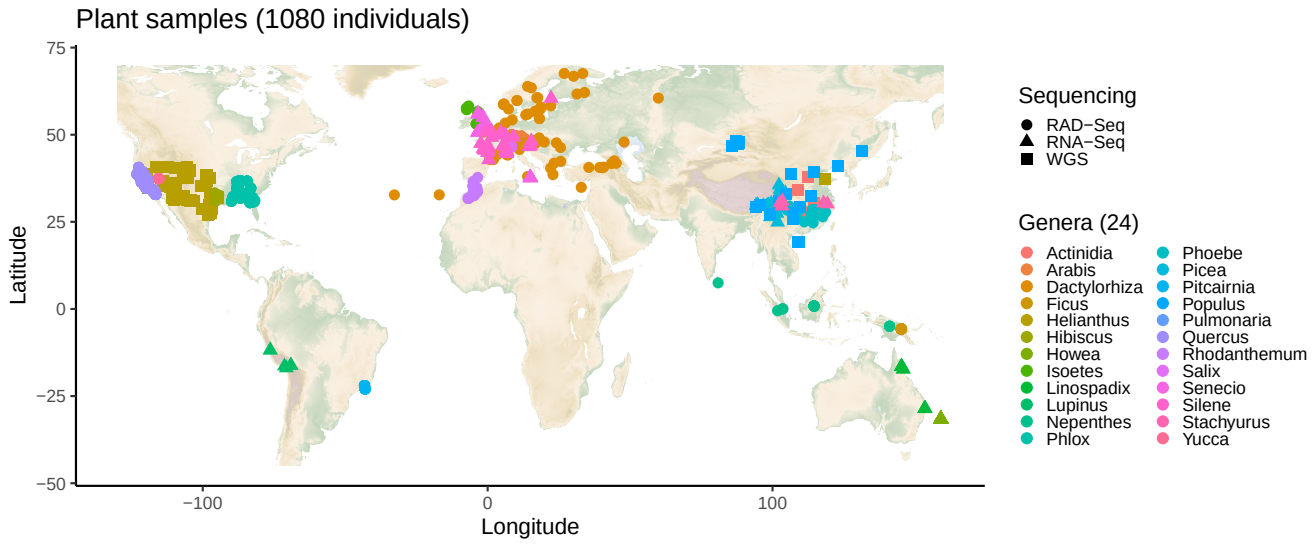


Figure S1: **Geographical location of plant samples and sequencing methods**
The depicted symbol represents an entity for which geographic data was retrievable in the source publications, with the exception of the genus *Gossypium*. The varying shapes of the symbols serve to differentiate between distinct sequencing technologies.

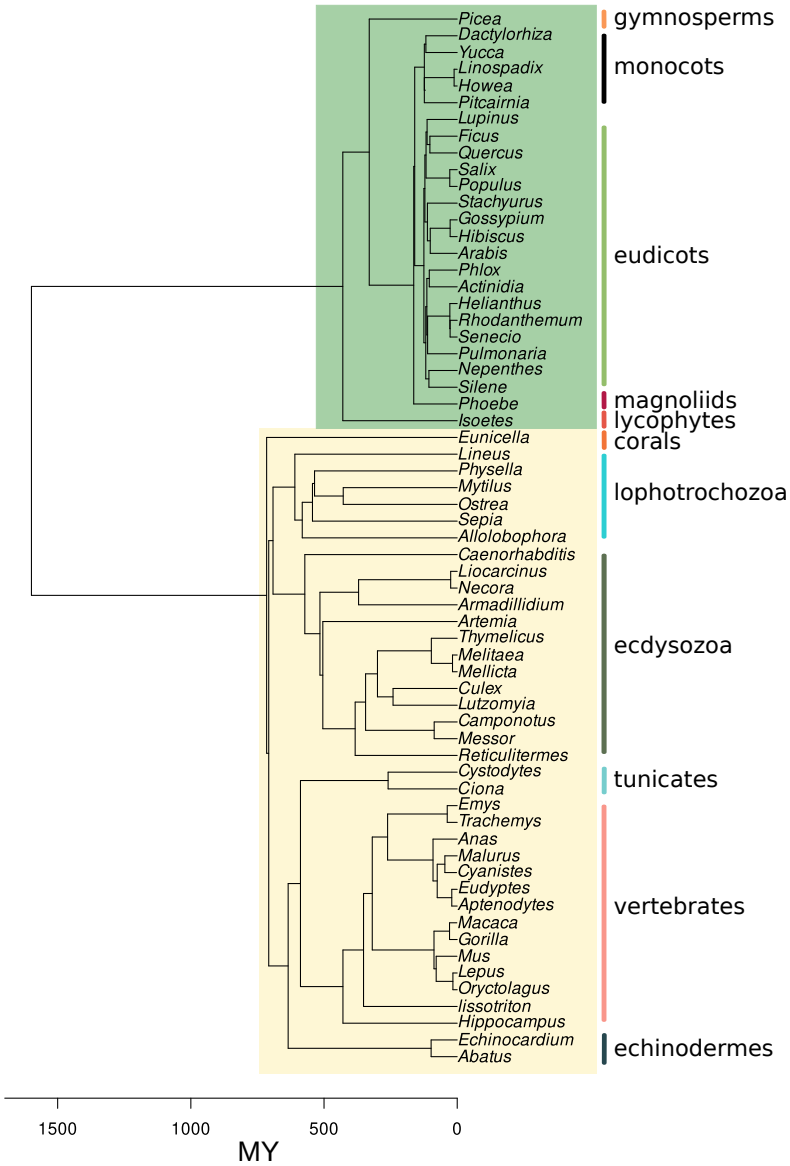


Figure S2: **Phylogenetic relationships between species included in the current study.** Plants and animals are indicated by green and yellow rectangles respectively. The scale represent the time from present expressed in million years (MY) according to `TimeTree` (*79*). Animals (yellow square) are from (*6*). Plants (green square) are included in the current study.

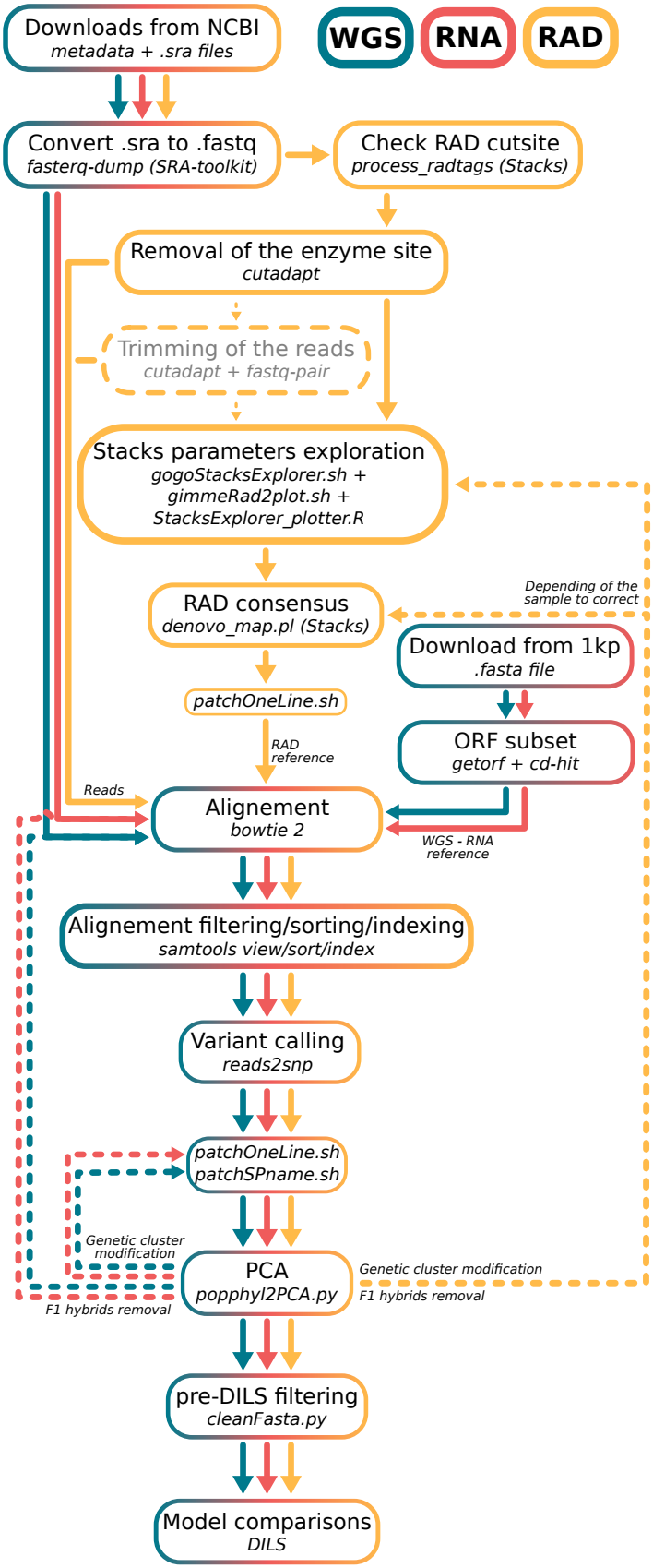


Figure S3: **Bioinformatics steps from the raw reads to demographic analysis**
Within each box, the upper line delineates an information technology procedure employed for data processing, while the lower line specifies the program or script utilized for its execution. The coloration denotes the specific sequencing technology concerned by each step.

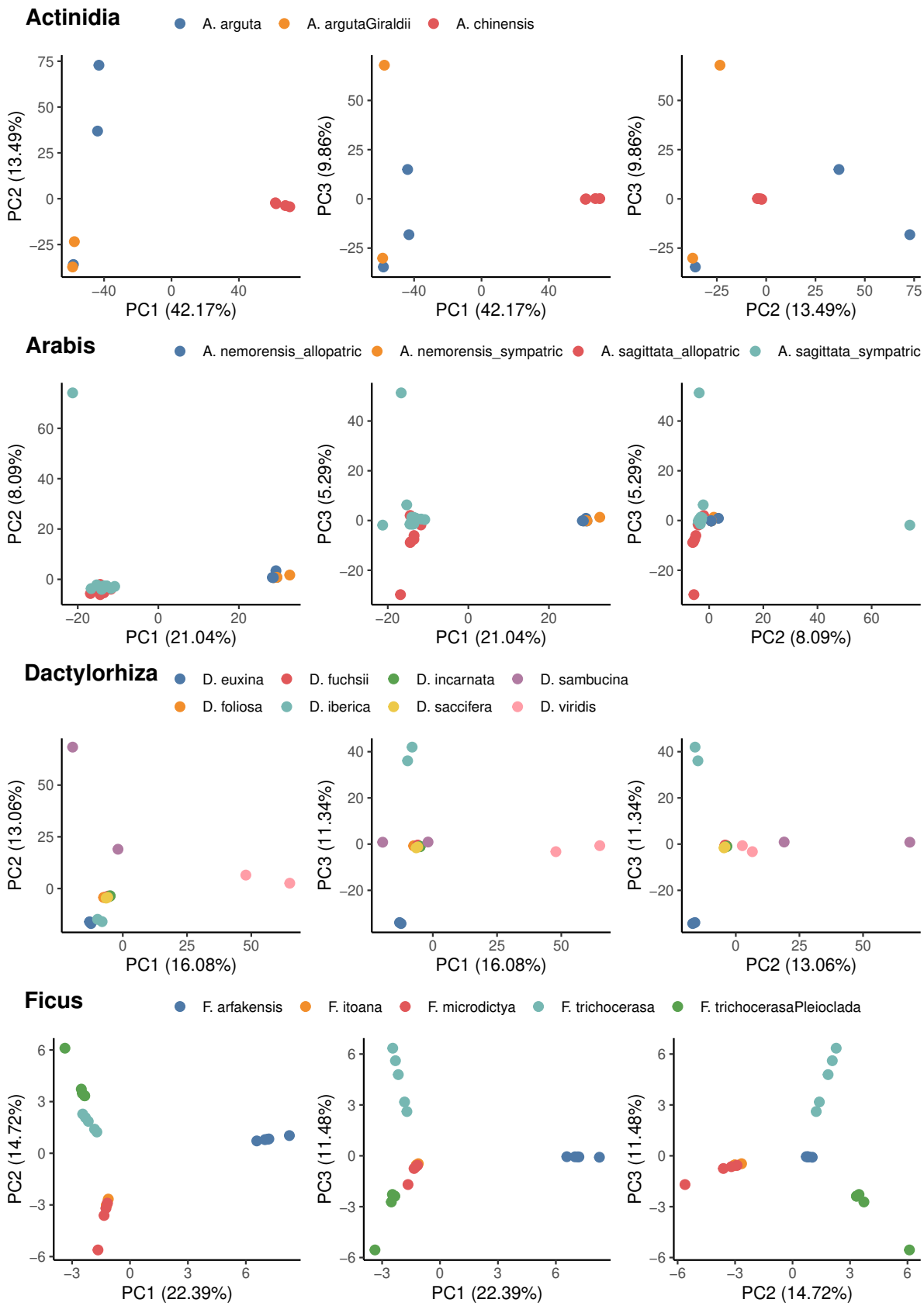


Figure S4: **Principal component analyses on genotypes for all SNPs.**
Each point represents an individual. The colours represent the different populations/species named by the authors of the studies from which the data originated.

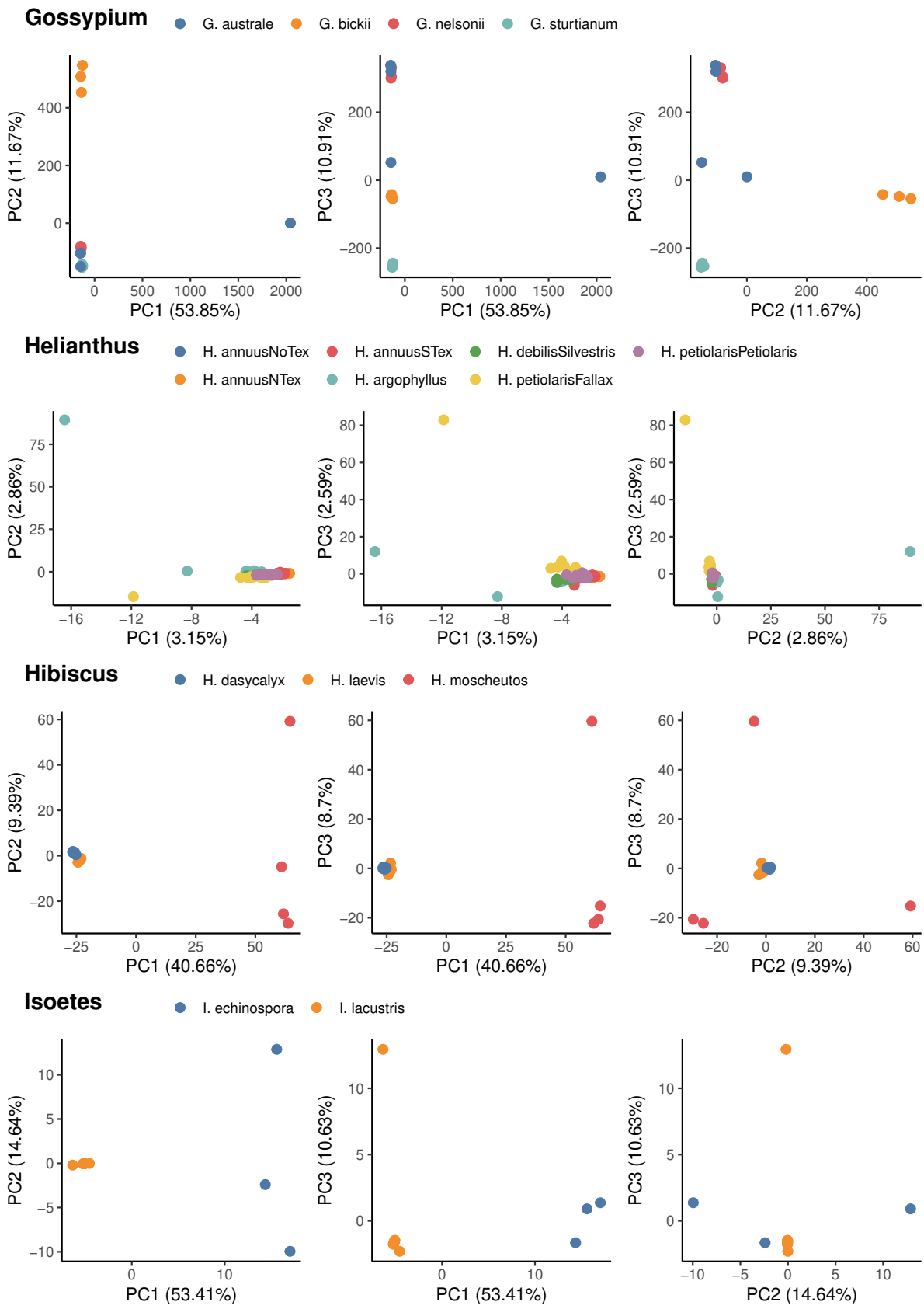


Figure S4: (continued).

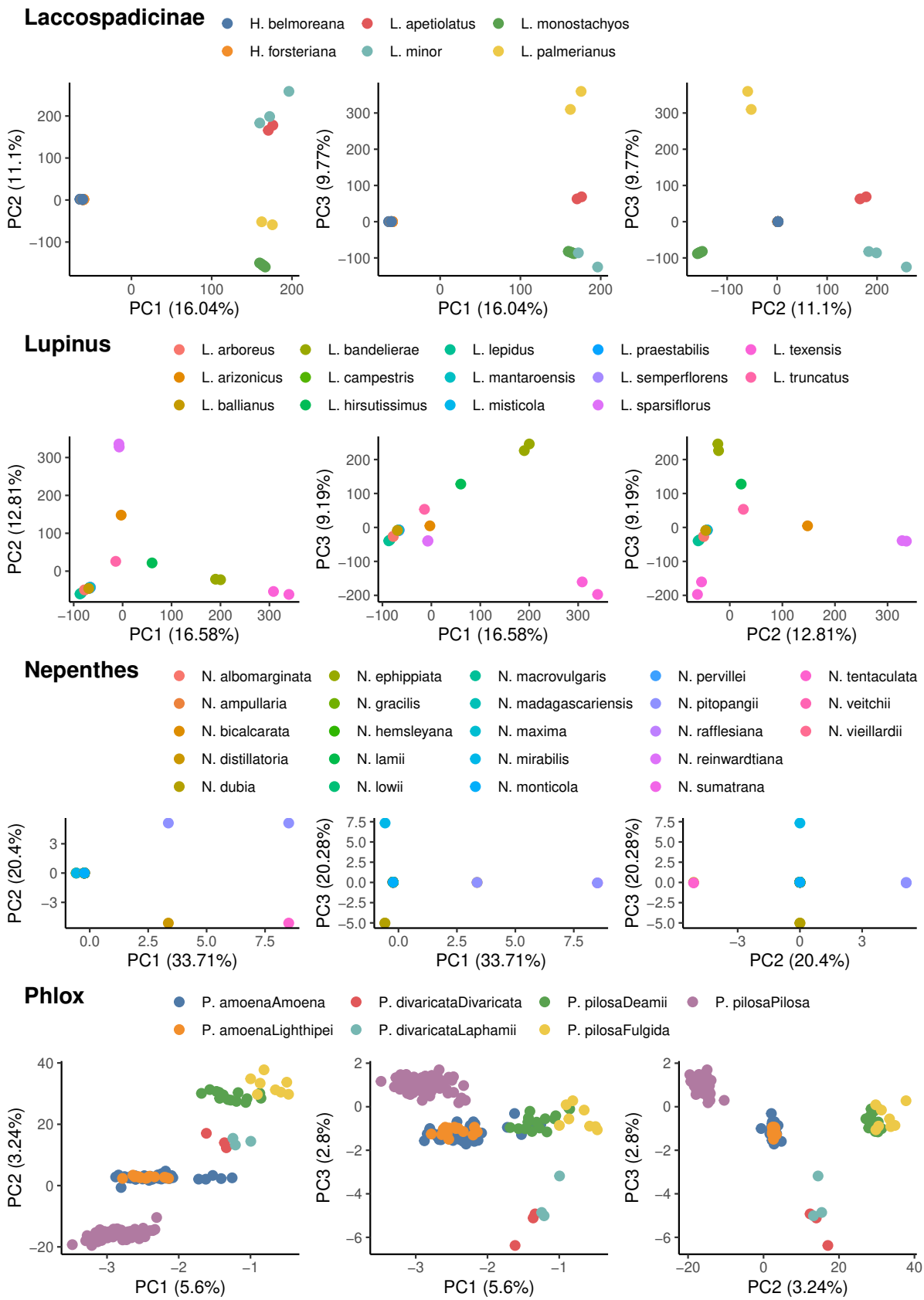


Figure S4: (continued).

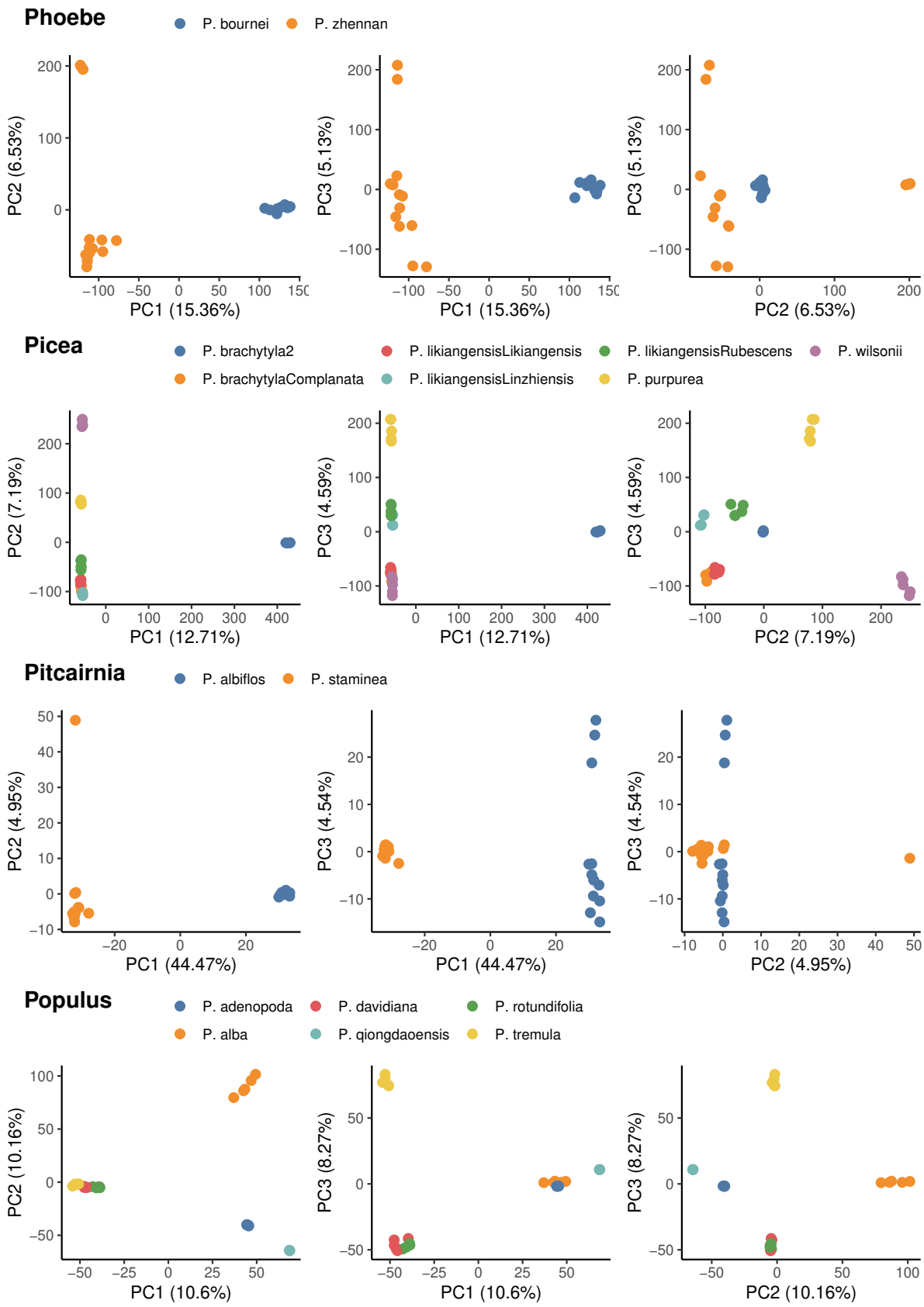


Figure S4: (continued).

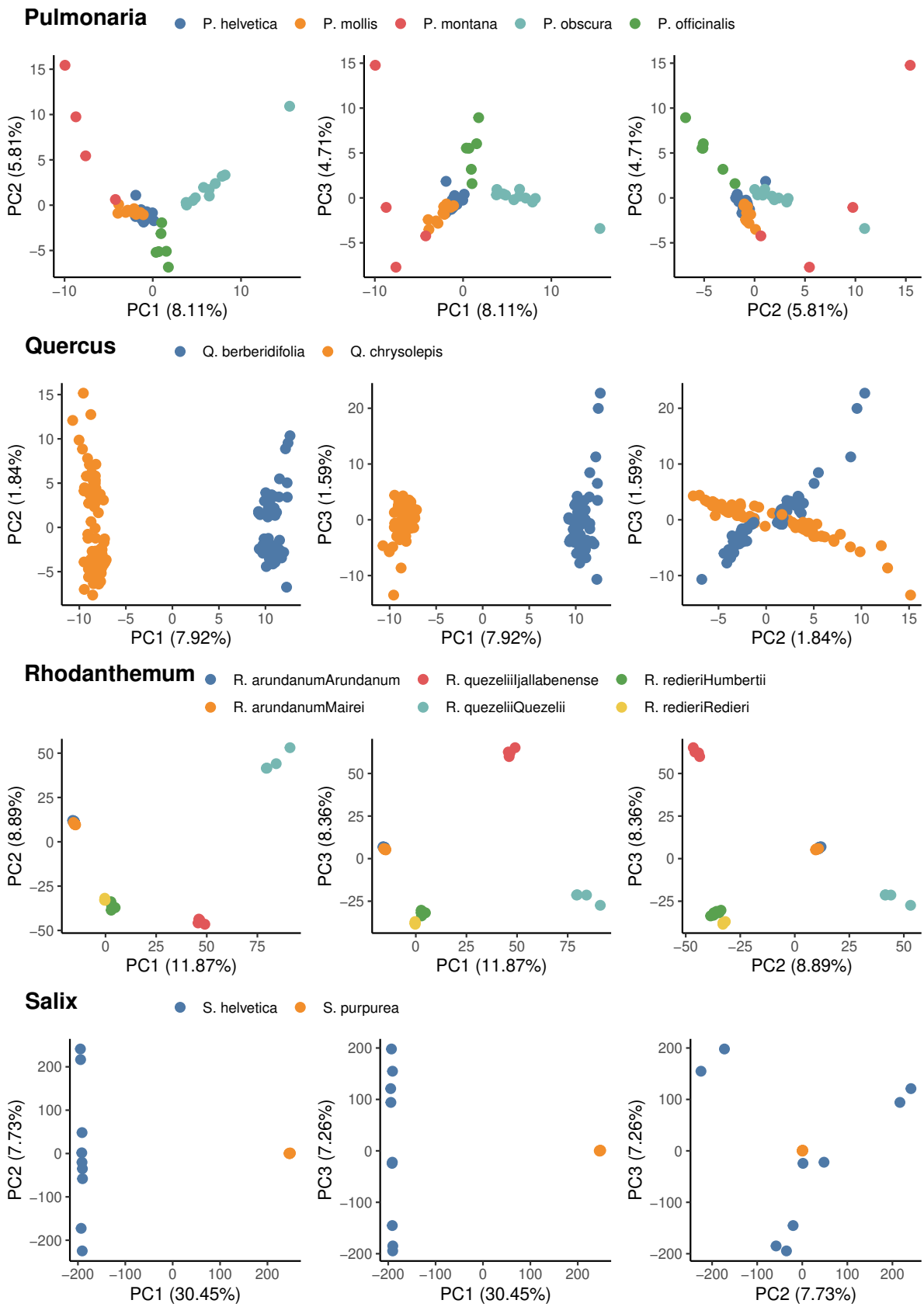


Figure S4: (continued).

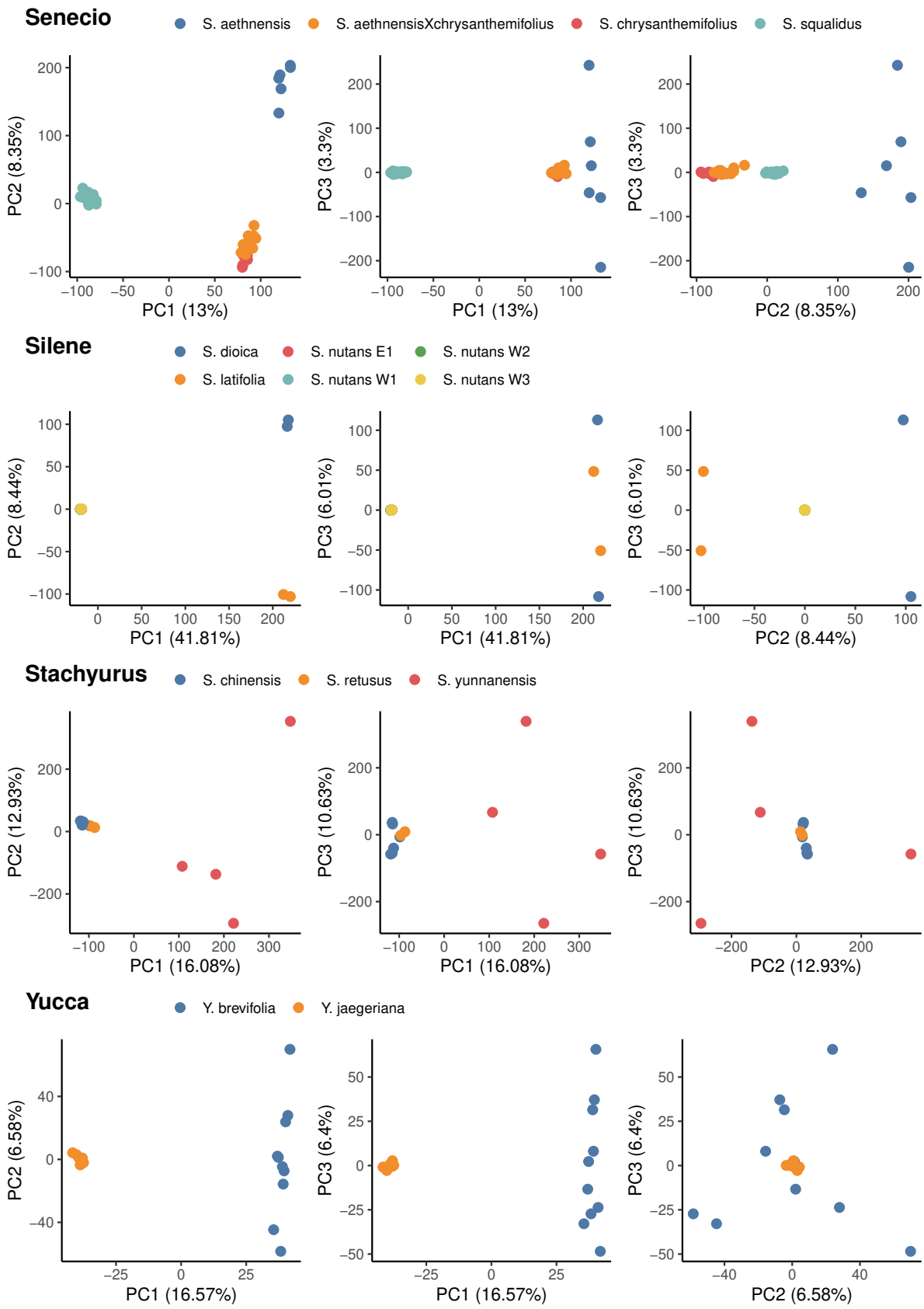


Figure S4: (continued).

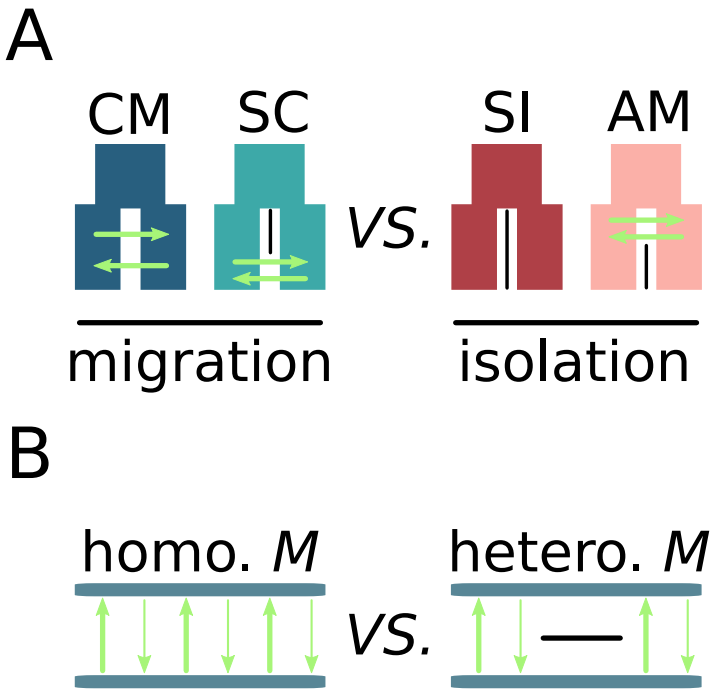


Figure S5: **Compared models using approximate Bayesian computation (ABC).**
**A.** Models with ongoing migration correspond to all CM (Continuous Migration) and SC (Secondary Contact) models. Models with current isolation correspond to all SI (Strict Isolation) and AM (Ancestral Migration) models. The first step in our ABC classification is to compare the set of CM+SC *versus* SI+AM models in order to assign a migration or isolation status to each of the 341 pairs of lineages (61 animals, 280 plants) according to the computed posterior probability.
**B.** Pairs of plants or animals, for which our ABC framework has provided strong statistical evidence of ongoing migration, are subsequently subjected to analysis aimed at discerning the uniformity of gene flow across the genome, whether it exhibits homogeneity (characterized by the absence of local genomic barriers) or heterogeneity (signifying genetic linkage to species barriers). The comparison between homo. *M* *versus* hetero. *M* was carried out using the same ABC framework as in the previous step.

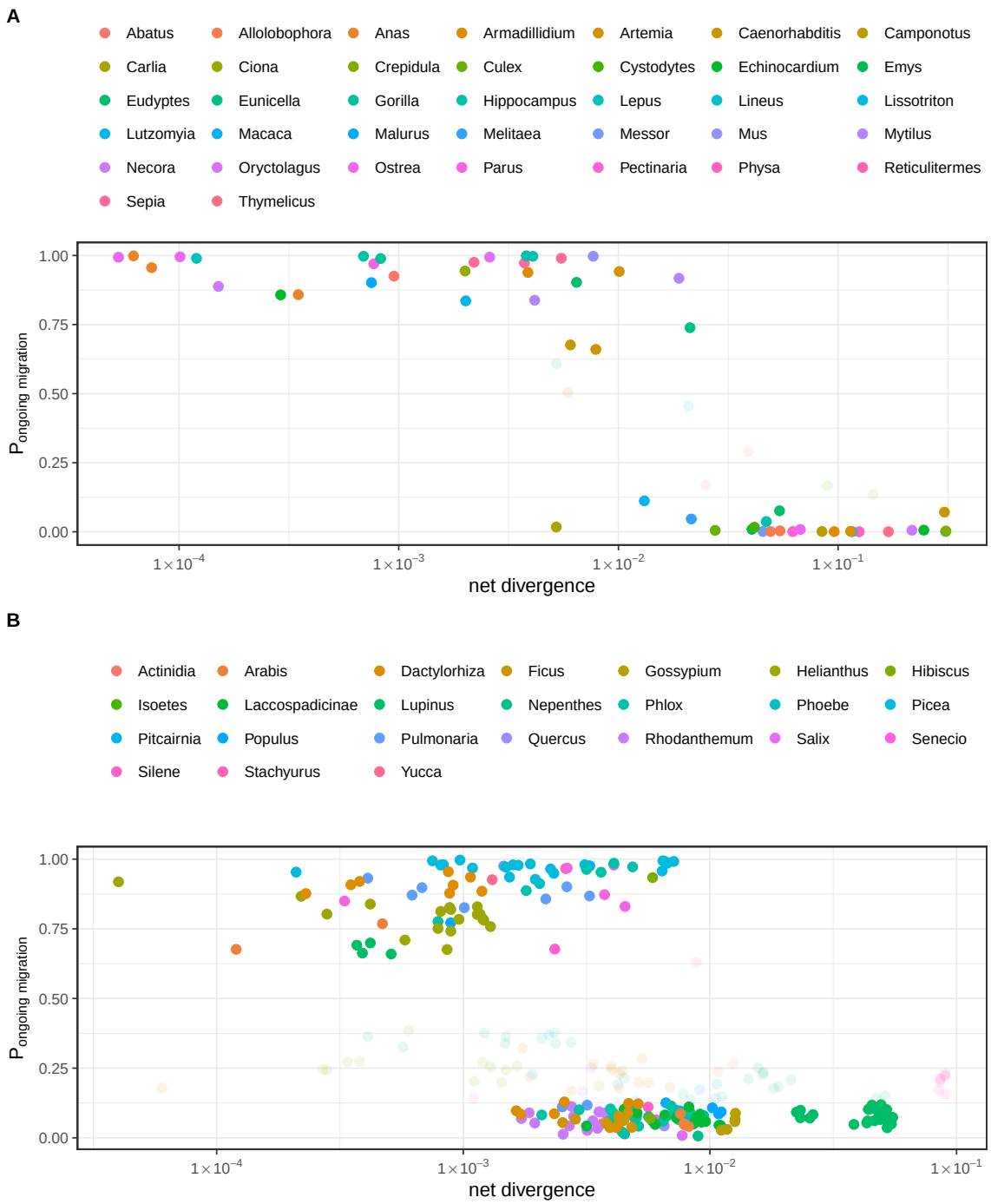


Figure S6: **Relationship between mean net divergence and posterior probability for ongoing migration.**
Each point corresponds to a pair of animals (A) or plants (B). x-axis: average net divergence. y-axis: posterior probability for ongoing migration attributed by our ABC framework. Colours correspond to surveyed genera. Solid points represent pairs for which there is strong statistical evidence either supporting or rejecting the ongoing migration model, as determined by the robustness test outlined in (*6*). In contrast, transparent points indicate pairs for which the comparison between the migration and isolation models yields an inconclusive result. Pairs for which support was inconclusive were excluded from further analysis. The remaining pairs were categorized either as exhibiting 'migration' or 'isolation', as illustrated in Figure 1-A (see section A.4.4).

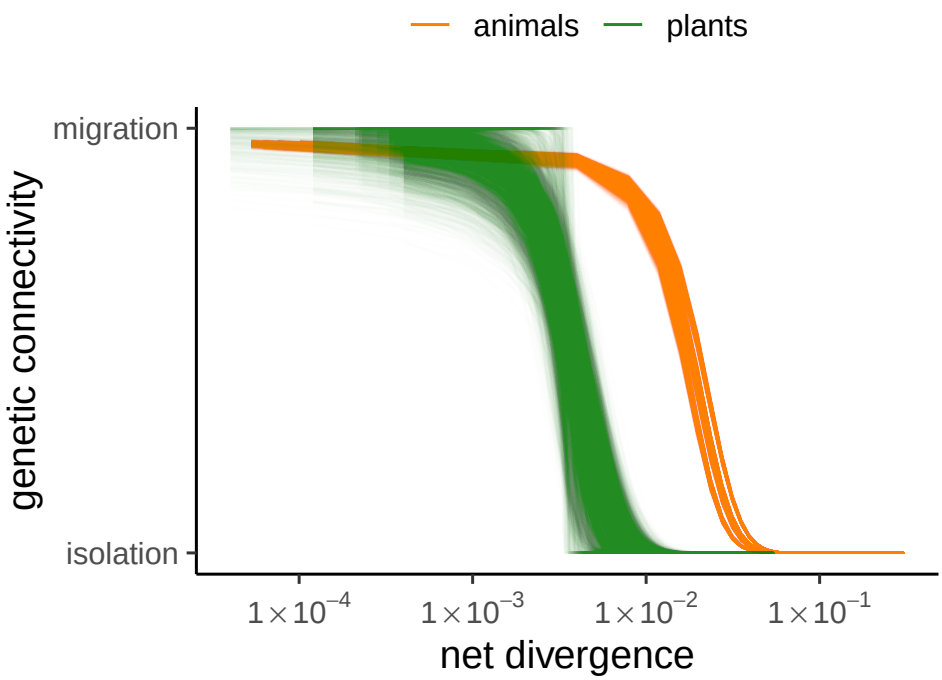


Figure S7: **Relationship between mean net divergence and migration/isolation status controlled by a genus effect.**
The relationship was established as for the entire dataset shown in Figure 1, but by randomly sub-sampling a single pair of populations/species within each plant and animal genus. Each line represents one of the 5,000 random iterations.

# C Supplementary Tables

Table S1: List of retained NCBI datasets.

| **bioproject** | **genus** | **species** | **n** | **type of data** | **source** |
|---|---|---|---|---|---|
| PRJNA318567 | *Actinidia* | *arguta* | 3 | WGS | (*39*) |
| | | *arguta giraldii* | 2 | | |
| | | *chinensis* | 4 | | |
| PRJEB33482, PRJEB39992 | *Arabis* | *nemorensis allop.* | 6 | RNA | (*40*) |
| | | *nemorensis symp.* | 6 | | |
| | | *sagittata allop.* | 10 | | |
| | | *sagittata symp.* | 15 | | |
| PRJNA489792 | *Dactylorhiza* | *euxina* | 5 | RAD | (*41*) |
| | | *foliosa* | 2 | | |
| | | *fuchsii* | 30 | | |
| | | *iberica* | 2 | | |
| | | *incarnata* | 31 | | |
| | | *saccifera* | 4 | | |
| | | *sambucina* | 3 | | |
| | | *viridis* | 3 | | |
| PRJNA445222 | *Ficus* | *arfakensis* | 14 | RAD | (*42*) |
| | | *itoana* | 13 | | |
| | | *microdictya* | 15 | | |
| | | *trichocerasa* | 15 | | |
| | | *t. pleioclada* | 26 | | |
| PRJNA539957 | *Gossypium* | *australe* | 4 | WGS | (*43*) |
| | | *bickii* | 3 | | |
| | | *nelsonii* | 3 | | |
| | | *robinsonii* | 2 | | |
| | | *sturtianum* | 6 | | |
| PRJNA532579 | *Helianthus* | *annuus NoTex* | 15 | WGS | (*44*) |
| | | *annuus NTex* | 15 | | |
| | | *annuus STex* | 15 | | |
| | | *argophyllus* | 10 | | |
| | | *debilis silvestris* | 5 | | |
| | | *niveus canescens* | 8 | | |
| | | *petiolaris fallax* | 10 | | |
| | | *p. petiolaris* | 10 | | |
| PRJNA382435 | *Hibiscus* | *dasycalyx* | 6 | RAD | (*45*) |
| | | *laevis* | 4 | | |
| | | *moscheutos* | 5 | | |
| PRJNA483403 | *Isoetes* | *lacustris* | 9 | RAD | (*46*) |
| | | *echiospora* | 3 | | |

| | | | | | |
|---|---|---|---|---|---|
| PRJNA244607 | *Howea* | *belmoreana* | 40 | RNA | (*47*) |
| | | *forsteriana* | 39 | | |
| PRJNA528594 | *Linospadix* | *monostachyos* | 18 | | (*48*) |
| | | *minor* | 9 | | |
| | | *apetiolatus* | 6 | | |
| | | *palmerianus* | 6 | | |
| PRJNA318864 | *Lupinus* | *ballianus* | 2 | RNA | (*80*) |
| | | *bandelierae* | 2 | | |
| | | *misticola* | 2 | | |
| PRJEB37794 | *Nepenthes* | *albomarginata* | 3 | RAD | (*49*) |
| | | *ampullaria* | 8 | | |
| | | *bicalcarata* | 6 | | |
| | | *distillatoria* | 2 | | |
| | | *dubia* | 2 | | |
| | | *ephippiata* | 2 | | |
| | | *gracilis* | 8 | | |
| | | *hemsleyana* | 4 | | |
| | | *lamii* | 2 | | |
| | | *lowii* | 2 | | |
| | | *macrovulgaris* | 2 | | |
| | | *madagascariensis* | 2 | | |
| | | *maxima* | 10 | | |
| | | *mirabilis* | 10 | | |
| | | *monticola* | 2 | | |
| | | *pervillei* | 16 | | |
| | | *pitopangii* | 2 | | |
| | | *rafflesiana* | 9 | | |
| | | *reinwardtiana* | 2 | | |
| | | *sumatrana* | 2 | | |
| | | *tentaculata* | 2 | | |
| | | *veitchii* | 3 | | |
| | | *vieillardii* | 2 | | |
| PRJNA701424 | *Phlox* | *amoena amoena* | 48 | RAD | (*50*) |
| | | *a. lighthipei* | 14 | | |
| | | *divaricata divaricata* | 3 | | |
| | | *d. laphamii* | 3 | | |
| | | *pilosa deamii* | 15 | | |
| | | *p. fulgida* | 8 | | |
| | | *p. pilosa* | 59 | | |
| | | *subulata* | 2 | | |

| | | | | | |
|---|---|---|---|---|---|
| PRJNA464259 | *Phoebe* | *zhennan* | 9 | RAD | (*51*) |
| | | *bournei* | 12 | | |
| PRJNA807675 | *Pitcairnia* | *albiflos* | 9 | RAD | (*52*) |
| | | *staminea* | 12 | | |
| PRJNA392950, | *Picea* | *brachytyla* | 4 | RNA | (*53, 54*) |
| PRJNA401149, | | *b. complanata* | 5 | | |
| PRJNA378930, | | *likiangensis likiangensis* | 5 | | |
| PRJNA301093 | | *l. linzhiensis* | 5 | | |
| | | *l. rubescens* | 5 | | |
| | | *purpurea* | 5 | | |
| | | *wilsoni* | 5 | | |
| PRJNA612655 | *Populus* | *adenopoda* | 5 | WGS | (*55*) |
| | | *alba* | 5 | | |
| | | *davidiana* | 5 | | |
| | | *qiongdaoensis* | 3 | | |
| | | *rotundifolia* | 4 | | |
| | | *tremula* | 5 | | |
| PRJNA544114 | *Pulmonaria* | *helvetica* | 24 | RAD | (*56*) |
| | | *mollis* | 10 | | |
| | | *montana* | 4 | | |
| | | *obscur* | 11 | | |
| | | *officinalis* | 6 | | |
| PRJNA639507 | *Quercus* | *berberidifolia* | 63 | RAD | (*57*) |
| | | *chrysolepis* | 80 | | |
| PRJNA554975 | *Rhodanthemum* | *redieri redieri* | 4 | RAD | (*58*) |
| | | *r. humbertii* | 7 | | |
| | | *quezelii quezelii* | 4 | | |
| | | *q. jallabenense* | 4 | | |
| | | *arundanum mairei* | 8 | | |
| | | *a. arundanum* | 27 | | |
| PRJNA429746 | *Salix* | *helvetica* | 10 | RAD | (*59*) |
| | | *purpurea* | 10 | | |
| PRJNA549571 | *Senecio* | *aethnensis* | 6 | RNA | (*60*) |
| | | *aethn. X chrys.* | 14 | | |
| | | *chrysanthemifolius* | 6 | | |
| | | *squalidus* | 28 | | |
| PRJNA295359 | *Silene* | *dioica* | 2 | RNA | (*61, 62*) |
| | | *latifolia* | 2 | | |
| | | *nutans E1* | 4 | | |
| | | *n. W1* | 4 | | |

| | | | | | |
|---|---|---|---|---|---|
| | | *n. W2* | 4 | | |
| | | *n. W3* | 4 | | |
| PRJNA553020 | *Stachyurus* | *chinensis* | 6 | RNA | (*63*) |
| | | *retusus* | 2 | | |
| | | *yunnanensis* | 4 | | |
| PRJNA329381 | *Yucca* | *brevifolia* | 24 | RAD | (*64*) |
| | | *jaegeriana* | 39 | | |

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

| **model** | $\boldsymbol{\ell}$ | $\boldsymbol{\beta_0}$ | $\boldsymbol{\beta_1}$ | $\boldsymbol{X_{p=0.5}}$ | **df** | ***P*-value** |
|---|---|---|---|---|---|---|
| $M_0$ | -115.2766 | 1.736 | -433.504 | 0.004 | | |
| $M_{plants}$ | -74.5078 | 2.517 | -799.021 | 0.003 | | |
| $M_{animals}$ | -7.808659 | 3.935 | -209.252 | 0.0188 | | |
| | | | | | 2 | $4.88 \times 10^{-15}$ |

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

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

| **model** | $\boldsymbol{\ell}$ | $\boldsymbol{\beta_0}$ | $\boldsymbol{\beta_1}$ | $\boldsymbol{X_{p=0.5}}$ | **df** | ***P*-value** |
|---|---|---|---|---|---|---|
| $M_0$ | -41.92664 | 2.413 | -320.743 | 0.007 | | |
| $M_{plants}$ | -20.82818 | 4.031 | -766.155 | 0.005 | | |
| $M_{animals}$ | -4.361694 | 5.347 | -271.134 | 0.0197 | | |
| | | | | | 2 | $5.38 \times 10^{-8}$ |

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