

DNA-based identification of plants and the genomic nature of plant species differences

Received: 4 March 2025

Accepted: 3 March 2026

Cite this article as: Huang, W., Li, D.-Z., Antonelli, A. *et al.* DNA-based identification of plants and the genomic nature of plant species differences. *Commun Biol* (2026). <https://doi.org/10.1038/s42003-026-09858-7>

Wu Huang, De-Zhu Li, Alexandre Antonelli, Christine D. Bacon, Lian-Ming Gao, Catherine Kidner, R. Toby Pennington, Douglas E. Soltis, Pamela Soltis, Jeannine Cavender-Bares, Camille Christe, Kyle G. Dexter, Yanqian Ding, Mario Durán-Castillo, Mario Fernández-Mazuecos, Elliot M. Gardner, Morgan R. Gostel, Margarita Hernandez, Andrew L. Hipp, Johanna R. Jantzen, Jacob B. Landis, Xiu-Qun Liu, Andre A. Naranjo, James Nicholls, Han-Tao Qin, Jeffrey P. Rose, Nicolas Salamin, Rowan Schley, Philipp M. Schlüter, Jessica D. Stephens, Matthew A. Streisfeld, Natascha D. Wagner, Xiao-Quan Wang, Qiu-Yun Jenny Xiang, Alex D. Twyford & Peter M. Hollingsworth

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

DNA-based Identification of Plants and the Genomic Nature of Plant Species Differences

Wu Huang^{1,2,6,*}, De-Zhu Li^{3,4,5}, Alexandre Antonelli^{6,7,8,9}, Christine D. Bacon⁷, Lian-Ming Gao^{3,4,5}, Catherine Kidner^{1,10}, R. Toby Pennington^{1,11}, Douglas E. Soltis^{12,13}, Pamela Soltis^{12,13}, Jeannine Cavender-Bares¹⁴, Camille Christe¹⁵, Kyle G. Dexter^{1,16,17}, Yanqian Ding^{1,2}, Mario Durán-Castillo², Mario Fernández-Mazuecos¹⁸, Elliot M. Gardner¹⁹, Morgan R. Gostel²⁰, Margarita Hernandez²¹, Andrew L. Hipp²², Johanna R. Jantzen^{12,13}, Jacob B. Landis²³, Xiu-Qun Liu²⁴, Andre A. Naranjo²⁵, James Nicholls^{1,26}, Han-Tao Qin^{3,4}, Jeffrey P. Rose²⁷, Nicolas Salamin²⁸, Rowan Schley¹¹, Philipp M. Schlüter²⁹, Jessica D. Stephens³⁰, Matthew A. Streisfeld³¹, Natascha D. Wagner³², Xiao-Quan Wang³³, Qiu-Yun Jenny Xiang³⁴, Alex D. Twyford^{1,2,*}, Peter M. Hollingsworth^{1,*}

¹Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, EH3 5LR, UK

²Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK

³Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China

⁴Lijiang Forest Diversity National Observation and Research Station, Kunming Institute of Botany, Chinese Academy of Sciences, Lijiang 674100, China

⁵Center for Interdisciplinary Biodiversity Research & College of Forestry, Shandong Agricultural University, Tai'an, Shandong 271018, China

⁶Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK

⁷Göteborg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, 41390 Göteborg, Sweden

⁸Department of Biology, University of Oxford, Oxford, OX1 3RB, UK

⁹Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China

¹⁰Institute of Molecular Plant Sciences, University of Edinburgh, Max Born Crescent, Edinburgh EH9 3BF, UK

¹¹Department of Geography, University of Exeter, North Park Road, Exeter EX4 4QE, UK

¹²Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

¹³Biodiversity Institute, University of Florida, Gainesville, FL 32611, USA

¹⁴Harvard University Herbaria, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

¹⁵Conservatoire et Jardin botaniques de Genève, 1 ch. de L'Impératrice, 1292 Pregny-Chambésy, Switzerland

¹⁶School of Geosciences, University of Edinburgh, Grant Institute, King's Buildings West Mains Road, Edinburgh, EH9 3JW, UK

¹⁷Department of Life Sciences and Systems Biology, University of Turin, Via Giuseppe Verdi, 8, 10124 Torino TO, Italy

¹⁸Department of Biology / Centro de Investigación en Biodiversidad y Cambio Global (CIBC-UAM), Universidad Autónoma de Madrid, 28049 Madrid, Spain

¹⁹Department of Biology, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, USA

²⁰Botanical Research Institute of Texas, Fort Worth, TX 76132 -3400, USA

²¹Department of Anthropology, Penn State University, State College, PA 16801, USA

²²Herbarium and Center for Tree Science, The Morton Arboretum, Lisle, IL, USA

²³School of Integrative Plant Science, Section of Plant Biology and the L.H. Bailey Hortorium, Cornell University, Ithaca, NY, USA

²⁴Key Laboratory of Horticultural Plant Biology (Ministry of Education), College of Horticulture and Forestry Science, Huazhong Agricultural University, Wuhan 430070, China

²⁵International Center for Tropical Botany, Institute of Environment, Florida International University, Miami, FL, USA

²⁶Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australian National Insect Collection, Clunies Ross Street, Acton, ACT, 2601, Australia

²⁷Department of Botany, University of Wisconsin-Madison, Madison, WI 53706, USA

²⁸Department of Computational Biology, University of Lausanne, Quartier Centre, 1015 Lausanne, Switzerland

²⁹Institute of Biology, University of Hohenheim, Garbenstr. 30, 70599 Stuttgart, Germany

³⁰Biology Department, Westfield State University, 577 Western Ave, Westfield, MA 01086 USA

³¹Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403, USA

³²Department of Systematics, Biodiversity and Evolution of Plants (with Herbarium), University of Goettingen, Untere Karspüle 2, 37073 Göttingen, Germany

³³State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

³⁴Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27695-7612, USA

* Wu Huang, Alex D. Twyford, Peter M. Hollingsworth

Email: whuang@rbge.org.uk, Alex.Twyford@ed.ac.uk, phollingsworth@rbge.org.uk

Author Contributions: W.H., P.M.H., and A.D.T. designed research; W.H. designed and developed bioinformatic pipeline and performed data collection and analysis; D.Z.L., A.A., C.D.B., L.M.G., C.K., R.T.P., D.E.S., P.S., J.C.B., C.C., K.G.D., Y.Q.D., M.D.C., M.F.M., E.M.G., M.R.G., M.H., A.L.H., J.R.J., J.B.L., X.Q.L., A.A.N., J.N., H.T.Q., J.P.R., N.S., R.S., P.M.S., J.D.S., M.A.S., N.D.W., X.Q.W., Q.Y.J.W. contributed sequence data; and W.H., P.M.H., and A.D.T. wrote the paper with input from D.Z.L., L.M.G., P.S., D.E.S., R.T.P., and A.A. All authors reviewed the final version of the manuscript.

Inclusion & Ethics statement: This research was conducted in accordance with all relevant institutional, national, and international guidelines and regulations. No human participants, identifiable human data, or animal experiments were involved. All biological materials were obtained under appropriate collection permits and material transfer agreements, and their use complies with the Convention on Biological Diversity and the Nagoya Protocol on Access and Benefit-Sharing.

We are committed to promoting transparency, open scientific exchange, and equitable collaboration. Local partners and institutions involved in sample access and knowledge exchange are acknowledged as research contributors and co-authors where appropriate. The project did not exclude any individuals or groups from participation on the basis of personal characteristics, and efforts were made to ensure accessible communication of the research findings.

Competing Interest Statement: The authors declare no competing interests.

Classification: Evolution (major), Genetics (minor)

Keywords: DNA barcoding, plant species discrimination, comparative genomics, multi-locus nuclear DNA sequence

ARTICLE IN PRESS

Abstract

Telling species apart using DNA sequence data plays a key role in understanding, monitoring, and managing biodiversity. However, plant species discrimination is often difficult due to the complex nature of plant species boundaries. To inform future strategies for DNA-based identification of plants using the nuclear genome and to gain fundamental insights into the genomic nature of differences between plant species, we conducted a large-scale analysis mining data from 151 studies. Of the 1,713 multiple-sampled species evaluated, 1,202 resolved as monophyletic (70.2%). We then assessed the density of species-specific SNPs (SSSNP) in the DNA sequence data - of the 462 species from 27 genera assessed in detail, there was a median density of 193 SSSNPs per Mb and 412 species (89.2%) had at least one SSSNP. Randomly sub-sampling the SNP data showed an asymptote in species discrimination with around 3,000 randomly selected SNPs. Finally, we undertook a resampling of 6 target-capture datasets and showed that 1-9 pre-selected loci provided equivalent levels of species discrimination compared to hundreds of nuclear loci. These findings provide an important quantitative assessment of the genomic nature of differences between plant species and provide foundations for the development of enhanced approaches for high-resolution DNA-based plant species discrimination.

Introduction

Accelerating species delimitation and identification is a pressing challenge for biodiversity science and conservation^{1,2}, given the importance of species as a fundamental unit of biodiversity and the rapidly accelerating rate of biodiversity decline³. DNA barcoding, using the mitochondrial *cytochrome oxidase 1 (COI)* gene has proved remarkably successful in discriminating among animal species^{4,5}. In plants, the standard DNA barcoding approach based on two or three plastid genes⁶, i.e. *rbcL*⁷, *matK*⁸, *trnH-psbA*⁹ and the internal transcribed spacers (ITS)¹⁰⁻¹² of nuclear ribosomal DNA has proved useful for many different applications¹³. This barcoding approach, however, often provides resolution to species-groups, rather than uniquely discriminating closely related species^{14,15}. A number of attributes of plant species and the current plant barcodes all contribute to this imperfect resolution. Plant species hybridise frequently¹⁶ and introgression and incomplete lineage sorting (ILS) of the barcode-region-containing plastid genomes and rDNA loci are well documented¹⁷⁻²⁰ and this undoubtedly contributes to DNA barcode sharing among species. Furthermore, rapid speciation mechanisms including polyploidy and/or breeding system transitions can lead to morphologically recognised taxa that do not show corresponding divergence in standard barcoding loci¹³. Additional intrinsic limitations of the standard DNA barcode regions include (a) the relatively slow mutation rate of plastid genomes which impacts the rate of generation of taxonomically informative species-specific substitutions²¹ and (b) the significant asymmetry in seed:pollen dispersal in most plant species, where limited seed dispersal constrains the likelihood of substitutions in predominantly maternally inherited plastid DNA from spreading throughout a species range^{13,22}.

Efforts to extend the standard plant barcode to include a wider coverage of the plastid genome provides a modest increase in discriminatory power, but is fundamentally limited by the fact that plastid genomes are typically maternally inherited and often do not track species boundaries^{19,23}. Thus, although there is a multitude of benefits and insights garnered from the use of standard plant barcodes, there is also an outstanding research challenge to develop improved scalable DNA-based methods that can more reliably distinguish plant species^{18,24}.

Addressing this challenge involves tackling two closely related issues. Firstly, at a practical level, there is a need to identify alternative DNA sequencing strategies that target genomic regions with increased discriminatory power, such as multiple loci from the nuclear genome¹⁸. Secondly, and at a more conceptual level, there is a need to better understand the genomic nature of the differences between plant species, such that the design of any improved DNA assay is informed by the nature of the problem. Outstanding issues here include understanding the relative proportions of taxonomically recognised plant species that resolve as monophyletic versus those whose evolutionary history and mode of speciation result in non-monophyletic relationships among related species^{25,26}, even with numerous nuclear markers²⁷. Likewise, the frequency and distribution of species-specific nucleotide substitutions in the nuclear genome are unknown: are species-specific substitutions common throughout the genome, or is the predominant signal between species one of frequency differences across multiple loci, with fixed differences being restricted to a small number of loci under strong selection²⁸? Refining DNA-based identification approaches using the nuclear genome will be inherently easier if species monophyly is the norm and taxon-specific substitutions are common.

In this study, we compile and analyse datasets that have generated large amounts of nuclear sequence data from multiple individuals of multiple congeneric species to provide a first synthetic assessment of the generalities of the genomic nature of the differences among plant species. These groups span a range of attributes, such as including woody and non-woody species, where differing rates of molecular evolution might affect the frequencies of species-specific substitutions, and genera of various sizes, including recent species radiations where previous molecular marker studies found limited genetic differences at the species-level.

First, we investigate the general nature of nuclear genomic differences between plant species, asking: (i) Whether most named plant species resolve as monophyletic as assumed in many standard tests of species discrimination, or whether non-monophyly is common due to a lack of resolution, polyphyly or paraphyly (as expected with recent species divergence and reticulation, or parapatric speciation especially where the progenitor species have large effective population sizes²⁷). (ii) Whether most plant species are characterised by the presence of frequent species-specific Single Nucleotide Polymorphisms (SSSNPs) in the nuclear genome, or whether SSSNPs are infrequent in most cases. We then examine the implications of these findings for the development of practical approaches for plant species discrimination using DNA data by (iii) assessing the minimum number of nuclear loci that are required, on average, to provide maximal levels of species discrimination. Collectively these investigations provide insights into the nature and patterns of genomic differences among plant species and will facilitate the future development of genomic barcoding approaches in plants.

Results

How often are named plant species monophyletic?

We compiled 151 studies corresponding to 134 plant genera in which multiple nuclear loci of 1713 species were sequenced (Supplementary Data 1), including 123 angiosperm, 6 gymnosperm, 3 fern, and 2 moss genera. We calculated the Monophyletic Ratio, defined as the proportion of monophyletic species out of the total number of species with multiple sampled individuals. Overall, 70.2% of species resolved as monophyletic. Among the 151 studies, 37 (24.5%) had 100% of the recognised species resolve as monophyletic, while 77 (51.0%) had less than 75% (Fig. 1A), with 70% of species resolving as monophyletic when averaged across studies. The Pearson Correlation Coefficient (r) between study size (number of species sampled) and monophyletic ratio was very weak ($r = 0.057$). No significant difference in monophyletic ratio was observed between plant groups with different lifeforms ($p = 0.81$, woody vs. herbaceous, Supplementary Data 2, Fig. 1B) or between studies using different sequencing techniques such as target capture, genome skimming, transcriptome sequencing, and RAD/GBS (two-sided Wilcoxon rank-sum test, p -values all > 0.05 , Supplementary Data 2, Fig. 1C)²⁹.

Are species-specific SNPs the norm or the exception?

To understand the genetic basis of species differences, we characterised the abundance of species-specific SNPs (SSSNPs) in 27 datasets, for which the sequence alignment data are accessible, representing 20 different seed plant families (Supplementary Data 3).

Overall, 16 datasets (59.3%) had at least one SSSNP in all species studied (Supplementary Data 4, Fig. S4). The density of SSSNPs ranges from 0 to 27,263 per Mb, with a median density of 193 SSSNPs per Mb over all assessed species (Fig. 2). In species resolved as monophyletic, the density of SSSNPs typically ranges from 22 - 5,623 per Mb (excluding outlier taxa in the top or bottom 5%). In contrast, most non-monophyletic species have SSSNPs at a density between 0 - 648 SSSNPs per Mb (excluding outlier taxa in the top or bottom 5%) and the density of SSSNPs is as expected significantly different between monophyletic species and non-monophyletic species (Fig. S1, Wilcoxon signed-rank p -value < 0.001). There is also notable variation in the results for species from different genera, with SSSNP densities ranging from 650 - 7,693 SSSNPs per Mb in genera such as the angiosperm *Linaria* (Plantaginaceae) to 0 - 4 SSSNPs per Mb in genera that have been considered recently diverged such as the orchid *Ophrys* (Orchidaceae)^{30,31} (Supplementary Data 4, Fig. 2).

To check whether the observed density of SSSNPs was due to genuine biological signal, as opposed to random shared mutations among samples arising due to the large number of sampled nucleotides, we randomised the taxon assignment of species within genera and conducted the same analysis. The randomised-labelled data resulted in a significantly smaller number of SSSNPs for all 27 datasets compared with the original assignment (Wilcoxon signed-rank test average p -value < 0.01 , Supplementary Data 5) and in most cases these randomised samples had zero SSSNPs.

How much data are needed to achieve maximal species discrimination success?

To test the minimal number of loci needed to tell named species apart, we randomly and repeatedly sub-sampled SNP loci in 23 of the datasets (selected on the basis that the full dataset resolved at least two species as monophyletic) (Supplementary Data 6). From randomly subsampling 10 SNP loci and progressively increasing the number of loci, we evaluated the point at which maximal species discrimination success (monophyly) was achieved (i.e. reaching the point where an equivalent number of monophyletic species was resolved as in the full dataset).

Sub-sampling of these datasets (Fig. 3) showed that the number of species discriminated based upon monophyly increases sharply from 100 to 500 randomly drawn SNPs. The proportion of species resolved as monophyletic then begins to plateau at around 500 to 1500 SNPs. Over all genera, at ~1,500 SNPs c. 90% of the species that resolve as monophyletic in the full datasets are resolved as monophyletic (Fig. 3A). At the within-genus level, at ~3,000 SNPs almost all genera exhibit an asymptote in their levels of species discrimination and 21/23 genera (91%) have >85% of their distinguishable species being discriminated with 3,000 randomly selected SNPs (Fig. 3B, Supplementary Data 7). Genera with a small number of multiple-sampled species (such as the gymnosperms *Taxus* and *Tsuga*, as well as the angiosperm *Mimulus*) hit the maximum species discrimination at an earlier stage, i.e. increasing the data volume beyond c. 3,000 randomly selected SNPs does not increase the number of species resolved as monophyletic. Genera with more than 20 sampled species with multiple intra-specific samples, such as *Salix*, *Artocarpus*, *Geonoma*, and *Inga*, show a continued but slower increase from 500 to 1,000 SNPs in the proportion of species resolving as monophyletic. The proportion of species resolving as monophyletic in some genera such as *Geonoma*, *Linaria*, and *Quercus* continued to slowly increase even after 7,000 SNPs were sampled (Fig. S5).

Do some genes show exceptional performance in species discrimination?

For six target capture datasets (Supplementary Data 8) where individual locus information was available, we performed sub-sampling at the level of individual genes rather than just the SNPs (this was not done on datasets with only small 'loci' such as RADseq). In five of the target capture datasets that had >400 genes sampled in their full dataset, four (*Inga*, *Tsuga*, *Polemonium*, and *Geonoma*) showed an asymptote in species discrimination at 100 genes or less, whereas in the fifth genus (*Artocarpus*) there was a slightly more protracted curve with 100 genes on average recovering 26 species as monophyletic and 200 genes recovering 28 species (Fig. S5).

To assess the performance of single loci in telling species apart, the frequency distribution of the number of species resolved by individual genes was plotted for the six genera with available datasets (Fig. S2). In five of these six genera, it approximates a normal distribution in the spread of performance of individual loci. In the remaining genus, *Polemonium*, most loci showed low levels of species discrimination with only a few loci being individually able to distinguish more than one species. In all datasets, there are clearly some loci that are much better than others in distinguishing species. For example, in *Geonoma*, the locus *LOC105045005* alone resolves 30 species as monophyletic (Supplementary Data 9), which is more than the number of species distinguished by using all 795 loci (which resolved 28 species as monophyletic). In four of the genera, the best single locus gave equivalent resolving power as the full dataset (Fig. S2; Table 1).

Only in *Inga* and *Capurodendron* is there a bigger discrepancy between the efficacy of the best-performing locus and the total dataset. In *Inga*, the best-performing locus distinguished 31 species compared to 45 species from the full dataset; in *Capurodendron*, the best-performing locus distinguished 13 species, compared to 20 species with the full dataset. But even in the case of these two genera, the nine and seven most informative genes resolved as many species as monophyletic as 810 genes for *Inga* and 615 genes for *Capurodendron* (Table 1).

We then examined whether there was a correlation between loci that were maximally informative in terms of discriminatory power (i.e. efficacy in recovering monophyly, density of SSSNPs) versus levels of nucleotide diversity (Supplementary Data 10, Fig. S3). Of the five genera tested (*Capurodendron* was excluded due to variation in the number of individuals per locus), there was a variable and only partially significant relationship between nucleotide diversity of genes within a genus and the density of SSSNPs in those genes (average correlation coefficient $r = 0.269$; this was statistically significant in only 3/5 tested datasets: *Geonoma* $r = 0.154$, *Inga* $r = 0.493$, *Tsuga* $r = 0.529$; two-sided Wilcoxon rank-sum test, $p < 0.001$). In terms of the relationship between nucleotide diversity and the number of species resolving as monophyletic, this was again a weak and only partially significant relationship; the average correlation coefficient ($r = 0.193$) and was statistically significant in three of the five datasets (namely *Geonoma* $r = 0.159$, *Inga* $r = 0.423$, *Tsuga* $r = 0.185$; two-sided Wilcoxon rank-sum test, $p < 0.001$). As expected, there was a stronger relationship between the density of SSSNPs and species monophyly for individual genes (average correlation coefficient $r = 0.413$ and statistically significant in all five genera ($p < 0.001$, two-sided Wilcoxon rank-sum test)).

Discussion

This study provides a synthetic quantification of the species discrimination signal in plant multi-locus nuclear DNA sequence datasets. It also provides foundational information for the future development of nuclear DNA barcoding approaches in plants.

The first question to address is what proportion of plant species resolve as monophyletic based on multi-locus nuclear sequence data? The concept of species-level monophyly is central to many methods of species discrimination (e.g. molecular operational taxonomic units – MOTUs³² or barcode index numbers - BINS³³) and is indeed central to certain species concepts (e.g. the phylogenetic species concept per Mishler and Donoghue³⁴; referred to as the monophyletic species concept by Judd *et al.*³⁵). In plants, various authors have identified high levels of non-monophyly from standard plastid and ribosomal barcoding regions^{36,37} but it is difficult to know how much of this relates to the attributes of barcoding loci, versus a more general point about the species-level monophyly of plant species. In our synthesis of 1,713 named species from 151 studies, we found 1,202 species resolving as monophyletic (70.2%) based on nuclear genomic data. At the level of individual genera, only 37 of the 151 studies (24.5%) had all multiple-sampled species resolved as monophyletic. The theoretical expectations that a substantial portion of plant species will not resolve as monophyletic are well established^{25,38}. Key factors likely to explain such non-monophyly are: (a) recent speciation and or long generation times leading to predicted long time to monophyly and the retention of ancestral polymorphisms (e.g. incomplete lineage sorting)³⁹, (b) founder speciation / peripheral isolate speciation⁴⁰, (c) hybridisation and homoploid hybrid speciation⁴¹, (d) auto- and allo-polyloid speciation⁴².

A wider point, and one not assessed in the current paper, is the degree to which non-monophyly is attributable to imperfect taxonomy. A study of 41,583 museum specimens of European Lepidoptera estimated ~23% of species to be non-monophyletic for the *COI* barcode, with 57% of non-monophyly at the species level attributed to taxonomic issues including under-splitting, over-splitting, or other identification difficulties⁴³. In plants, a recent study of *Capurodendron* shows that some polyphyletic species were in fact different species supported by good morphological characters with an added identification key⁴⁴. Taxonomic issues will impact this study, with different taxonomic concepts applied to different groups, particularly in cases where species boundaries are controversial (e.g. in *Euphrasia*⁴⁵ and *Antirrhinum*⁴⁶).

Notwithstanding the unquantified component of taxonomic error in this study, the estimate of ~30% non-monophyly quantified here is consistent with earlier hypotheses of the expected occurrence of monophyly based on morphological and/or geographical data. Crisp & Chandler⁴⁷ undertook a partial survey of two angiosperm families (Fabaceae, Proteaceae) and reported that c 20% of plant species resolved as paraphyletic. Likewise, in a more general perspective, Rieseberg & Brouillet²⁵ hypothesised that about 50% of plant species are “products of geographically local speciation, and that close to one half of these are likely not to be monophyletic”.

The estimate of 30% non-monophyly of named plant species recovered in the current study may change with additional sampling, and increased sampling of individuals within species, and species within genera. This has the potential for further disrupting patterns of monophyly and decreasing the proportion of species resolving as monophyletic. On the other hand, further sampling of monotypic or species-poor genera and/or genera with highly divergent species, would lead to higher levels of species monophyly, especially if such genera are currently under-represented in the literature due to being of lower intrinsic interest to systematists. Another factor which may ultimately reduce the overall proportion of plant species resolving as non-monophyletic is the possibility that extensive nuclear sequence datasets will identify cases where a substantial proportion of plant species non-monophyly is due to imperfect taxonomy, and that subsequent taxonomic clarification and taxonomic revisions will act in turn to increase the proportion of plant species names that are associated with monophyletic genetic lineages. Finally, there is likely to be a technical/analytical component to the proportion of species resolving as non-monophyletic. For multiple-sampled species, adding more sequence data or conducting detailed analyses for each study could improve resolution, potentially showing some species currently classified as non-monophyletic to be monophyletic.

Quantifying the frequency distribution of species-specific SNPs provides baseline information on the genomic nature of interspecific differences in plants and practical information regarding the ease of designing species-specific diagnostic assays.

In this study, the majority (89%) of tested species had at least one SSSNP (412/462 species from 27 datasets) even in complex groups such as *Salix* (willows, Fig. 2). This was true even for many species that resolved as non-monophyletic and 116 non-monophyletic species had at least some SSSNPs. In these cases, these SSSNPs may simply reflect plesiomorphies in the ancestral species, and/or may be linked to regions of the genome under selection and thus linked to the cohesiveness of a species⁴⁵.

The cases where we did not detect SSSNPs may be caused by imperfect taxonomy as outlined above or reflect the genuinely complex nature of species boundaries themselves^{48,49}. In particular,

recently formed species may not show fixed differences throughout much of the genome due to homogenising gene flow from relatives or maintenance of ancestral polymorphism, with species-specific differences limited to a few regions of the genome underlying divergence. Such clustered SNPs underlying species differences are unlikely to be represented in most RAD or target capture datasets and are only likely to be identified with whole-genome sequencing⁵⁰.

In terms of the density of SSSNPs, there is clearly a wide distribution in the frequency of SSSNPs (Fig. 2), although the median value of 193 SSSNP per Mb (IQR 20 – 834) is a useful summary catching the fact that SSSNPs are not so common they are routinely found every few 100 bp, nor are they so rare that they require megabases of sequence data to detect. This distribution of SSSNPs is of interest, as a first approximation of what proportion of the genome shows fixed differences between species and how this relates to wider patterns of variation. It is of course subject to sampling density, sequencing approach and the species divergence patterns among the 27 datasets. We stress that the SSSNP density distribution presented here should be interpreted as a very broad-brush first approximation rather than a robust point estimate, and it will undoubtedly be refined as more whole genome datasets become available across a wide range of densely sampled plant groups.

There is a consistent and growing demand for effective species identification, from elucidating the species composition of mixtures such as in food authentication, pollen metabarcoding and diet analyses, to species-level diagnostics and identification in forensic cases and species monitoring programmes⁵¹. In all these examples, there is a clear benefit to enhancing levels of species discrimination above and beyond that achievable with the current standard plant barcodes.

The rapid pace of development in sequencing technologies and the associated flood of data from the nuclear genome provide great potential for improvements in DNA-based species identification and gives optimism for the future design and definition of new nuclear barcoding approaches for plants. A comparison of available datasets which compare multi-locus nuclear sequencing approaches with standard DNA barcodes (Supplementary Data 12) shows clear improvement in species discrimination in 10 out of 12 cases (with the remaining two cases showing no difference). In terms of designing optimal future standard methodologies for exploiting the nuclear genome, when there is wide availability of reference genomes and associated whole-genome resequencing data for all species, it may ultimately be possible to identify species simply by genome resequencing⁵². However, until that future is realised there is a premium on more selective approaches. The current study indicates that without a priori selection, ~200 loci / 3000 SNPs represents the general point from which increases in species discrimination tails off. It is noteworthy that this asymptote is obtained relatively steeply as opposed to a slow progressive gain of more species being discriminated when more data are added. However, this amount of sequence data is clearly substantially higher than any of the current barcoding approaches and thus currently challenging to deploy at scale to very large sample sets. One promising non-targeted approach is kmer-based methods of species discrimination based on genome skimming, and recent studies show clear promise for exploiting shallow-pass genome skimming data from the nuclear genome of plants⁵²⁻⁵⁴.

An alternative perspective is to focus on more targeted approaches. The current study shows that, for a specific group, 1 to 10 best-performing loci can be equally effective as the full datasets from which they were derived. This result is encouraging and there is a high premium on follow up studies to assess whether there are some regions of the genome that are consistently informative

across taxonomic groups and will allow the maximal discrimination of species with a much smaller pool of loci. An obvious starting point for this approach is to explore whether any of the loci in the Angiosperms353 target capture set⁵⁵, which were developed for (and initially applied to) deeper-level phylogenetic studies, have sufficient resolving power at the species level. This targeted approach using a small number of loci is a promising route for development, though any assay with a modest number of loci will be unlikely to provide complete species resolution in complex groups characterised by contrasting modes and tempos of divergence. And regardless of the number of loci used, some situations will remain challenging, such as distinguishing recently formed autopolyploid species from their diploid progenitors⁵⁶.

Concluding remarks

The standard plant DNA barcodes remain a powerful and widely used method for understanding and characterising plant species diversity. To increase levels of resolution by exploiting signal in the nuclear genome of plants, there is a pressing need for community collaboration to address outstanding data and infrastructure needs. Immediate priorities include focused densely sampled genome resequencing studies aimed at understanding the genomic nature of plant species differences, and optimisation of pipelines and analytical methods for routinely and robustly quantifying the degree to which multi-locus nuclear sequence data can tell plant species apart in the most cost-efficient fashion. Moving beyond the currently available barcodes will bring considerable benefits in revealing the proportion of named plant species that correspond to coherent genetic groupings and also offer enhanced resolution in environmental and ecological biomonitoring applications. This latter point is of pressing global importance given the ambitious targets articulated in the Kunming Montreal Global Biodiversity Framework.

Materials and Methods

Our workflow involves the following key steps

- 1) Compile datasets which sample multiple individuals from multiple congeneric species for multiple nuclear markers
- 2) Assess % species monophyly based on the original published analyses
- 3) For a subset of these datasets evaluate the density and distribution of species-specific SNPs (SSSNPs)
- 4) For a subset of these datasets randomly and repeatedly subsample the data and execute a rapid tree building process to see how many randomly selected SNPs/loci are required to recover the same levels of monophyly as the full dataset
- 5) For target capture datasets assess whether some genes show higher species discriminatory power than others and assess whether discriminatory power is correlated with nucleotide diversity.

Compiling datasets

We compiled studies to assess the extent of plant species monophyly, with the criteria that they were published after 2013, sequenced three or more unlinked nuclear loci, included at least three individuals from multiple congeneric species and had a phylogenetic tree where species monophyly could be inferred. The average number of multiple-sampled species per study was 12 (range from 2 - 53). A total of 151 plant groups from the published literature or from collaborators were included (Supplementary Data 1). Studies were categorised by sequencing techniques, namely 1) Restriction site-Associated DNA sequencing (RAD-seq)⁵⁹ and its derivatives, (e.g., GBS⁶⁰, ddRAD-seq⁶¹, 2b-RAD⁶²); 2) Target Capture⁶³; 3) Genome skimming⁶⁴; 4) Transcriptome or exon sequencing^{65,66}. For full criteria for inclusion see this protocol <http://dx.doi.org/dx.doi.org/10.17504/protocols.io.kxygx3z9og8j/v1>⁵⁷.

Among these studies, 27 datasets were selected to further assess genomic differences between plant species (Supplementary Data 3). They were chosen because they have relevant metadata that link sequences of individuals to their species identities, a sequence alignment file in .fasta, .phylip, or .nex format or SNP matrix in .vcf or .fasta format, and a phylogenetic tree.

Assessing general patterns of species monophyly

In this study we focused on assessing the proportion of species where there is a clear phylogenetic signal of species-level monophyly, e.g. where all individuals within a species group together as a monophyletic unit in a phylogenetic tree. We evaluated 151 published datasets and manually recorded the number of species represented by more than one sampled individual that resolved as monophyletic, as a proportion of the total number of species in the dataset with more than one sampled individual. This assessment of monophyly was based on the phylogenetic trees presented in the original publications which represented the original authors' best estimates of phylogenetic relationships.

Assessing the frequency distribution of species-specific Single Nucleotide Polymorphisms (SSSNPs)

To estimate the density and abundance of species-specific SNPs (SSSNPs) in the 27 datasets with relevant sequence files, we calculated the number (and density per Mb) of SSSNPs from the total dataset with a SSSNP defined as a SNP that was fixed in one species and different from all other congeners. More details on these analyses are provided in this protocol <http://dx.doi.org/dx.doi.org/10.17504/protocols.io.5qpvo33rzv4o/v1>⁵⁸.

To group related genera within Figure 2, a phylogenetic tree was estimated based on version 3.0 of the Tree Of Life explorer <https://treeoflife.kew.org/tree-of-life>.

Assessing the minimum number of nuclear loci that are required to provide maximal levels of species discrimination

To evaluate the minimum number of random SNPs or loci required for species discrimination, subsampling of SNPs or loci was performed on the 23 genera where at least three species resolved as monophyletic (Supplementary Data 6). For each dataset, a UPGMA tree was generated for 50 random subsets of 200 SNPs and the proportion of species resolving as monophyletic recorded with Monophy⁶⁷. This process was repeated by incrementally adding more data, with 200 additional SNPs until 2000 SNPs were obtained, followed by adding 1000 SNPs at each step, until an asymptote in species discrimination was reached. Datasets where multiple SNPs could be recovered as loci (e.g. some target capture datasets) used an initial dataset of 10 loci, with a step size of 10 loci until 100 loci were reached, then 40 loci until 300 loci were reached, and subsequently 100 additional loci at each further step. We selected UPGMA as the tree building method based on its speed, and our basic requirement for estimating species monophyly from large numbers of trees from the simulations across multiple datasets with no requirement for estimating branch lengths or resolving the deeper nodes in the trees.

Additional analyses were performed to assess whether a small number of *selected* loci could result in species discrimination (measured by species monophyly) that was equivalent to a larger *random* selection of loci. These analyses were performed on the six datasets with defined loci such as target capture and which had little missing data, meaning that RAD data were excluded. These datasets include target capture data for *Artocarpus*, *Capurodendron*, *Geonoma*, *Inga* and *Polemonium*, and transcriptome sequencing for *Tsuga* (Supplementary Data 8). Specifically, we assessed which individual genes recovered the maximum number of species as monophyletic, and which minimal combinations of individual genes could recover the same number of monophyletic species as the complete dataset. Further details are provided in sections 6 and 7 of our analysis protocol, available at: <http://dx.doi.org/10.17504/protocols.io.5qpvo33rzv4o/v1>.

Statistics and Reproducibility

All statistical analyses were conducted to quantify patterns of species discrimination and genomic differentiation using multi-locus nuclear DNA datasets compiled from published studies and collaborator-provided data. Statistical tests were applied where appropriate to assess differences between groups or to evaluate whether observed patterns differed from random expectations.

Comparisons of monophyletic ratios among studies, life forms (woody vs. herbaceous), and sequencing approaches were conducted using two-sided Wilcoxon rank-sum tests. Correlations between study attributes (e.g. number of species sampled) and monophyletic ratios were assessed using Pearson correlation coefficients. Differences in the density of species-specific single nucleotide polymorphisms (SSSNPs) between monophyletic and non-monophyletic species were evaluated using two-sided Wilcoxon signed-rank tests. To assess whether observed SSSNP densities exceeded expectations under random species assignment, species labels were permuted within genera and SSSNP densities recalculated; observed and randomized values were compared using paired Wilcoxon signed-rank tests.

For analyses involving data resampling, reproducibility was assessed through repeated random subsampling. SNP- and locus-level subsampling analyses were performed with 50 independent replicates at each subsampling interval. For each replicate, species discrimination success was quantified as the number of species resolving as monophyletic, using UPGMA trees and the *MonoPhy* R package. Reported values represent the distribution across replicates, with medians and interquartile ranges shown where appropriate.

Replicates were defined as independent biological samples (individuals) within species, as provided in the original published datasets. Only species represented by two or more individuals were included in analyses of species monophyly and SSSNPs. Sample sizes therefore varied among studies and genera, reflecting the original experimental designs, with an average of 12 multi-sampled species per study (range 2–53). No technical replicates were generated as part of this study; all analyses were conducted on biological replicates or on resampled subsets of existing genomic data.

To assess the relationship between genes that showed high levels of discriminatory power, the density of SSSNPs and levels of nucleotide divergence, we calculated the correlation coefficient using *CORREL* function in excel and tested the significance via a Wilcoxon rank-sum test. The nucleotide divergence of each locus was calculate using *nuc.div* function in R package *pegas*⁶⁸.

Sequence data generation and alignment for *Linanthus* and *Leptosiphon*

Leaf material and flower colour information for 172 taxa of *Leptosiphon* and *Linanthus*, as well as taxa from *Gilia* and *Phlox* which serve as outgroup accessions, were collected from both field collection and herbarium specimens at the following herbaria: California Academy of Sciences (CAS), Jepson Herbarium – UC Berkeley (JEPS), Rancho Santa Ana Botanic Garden (RSA), University Herbarium – UC Berkeley (UC), and University of California Riverside (UCR). Samples were targeted for those in which flower colour was documented at the time of collections, and leaf material includes multiple colour morphs for most taxa. DNA was extracted from all samples following a modified CTAB extraction protocol⁷⁰.

Resuspended DNA from fresh material was then sonicated to a targeted length of 300 bp using a Covaris S220 sonicator (Covaris Inc., Woburn, MA, USA) following the manufacturer's suggested protocol. Resuspended DNA from herbarium accessions showed sufficient fragmentation so that sonication was not necessary. A total of 3–5 μ g of DNA from each sample

was sent to RapidGenomics (Gainesville, FL, USA) for Illumina library preparation with dual indexed barcodes. Targeted exon capture was conducted using MYbaits probes (MYcroarray, Ann Arbor, MI, USA) described by Landis et al. (2016)⁷¹. Capture products were pooled and distributed across three individual Illumina runs: 10 samples in a HiSeq 2000 (2 × 100 bp), 22 samples in a MiSeq 2 × 150 bp, and 148 samples in a NextSeq 2 × 150 mid-throughput run.

Raw reads were processed using the custom scripts described by Landis et al. (2016). Briefly, these scripts trim and filter the reads using cutadapt⁷². To pass the filtering parameters, sequences must have a minimum score of 20 and minimum length of 20 bp. Cleaned reads were then used in a BLAT⁷³ analysis to isolate plastome reads and on-target nuclear reads. For the 100 nuclear genes that were targeted, we obtained an average of 14,048 on-target sequences for 172 taxa, ranging from 6 to 106,937 paired-end reads per taxon. In addition to the targeted nuclear genes, we obtained on average 31,154 reads matching the plastome, with a range of 48 to 249,734 paired-end reads per taxon.

Nuclear and plastid coding genes were then assembled to individual genes using default parameters of HybPiper⁷⁴. For both nuclear and plastome data, genes were separated by using Burrows-Wheeler Aligner (BWA)⁷⁵ analysis against a reference file. Each gene was then aligned using MAFFT⁷⁶ (version 7.245) installed on the University of Florida Research Computing cluster using pairwise comparisons with 1,000 iterations. Aligned sequences were then concatenated using SequenceMatrix⁷⁷. Each concatenated sequence was then analyzed using PartitionFinder⁷⁸ (version 2.0) using a greedy algorithm and RAxML⁷⁹ (version 8) to find the best partition scheme for RAxML analyses. Raw reads for each accession were deposited in GenBank's Short Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) under the NCBI Bioproject number PRJNA322057.

Further analyses consisted of separate matrices for both *Leptosiphon* and *Linanthus* of the following data sets: (1) concatenated matrix of 61 nuclear loci for both genera (referred to hereafter as the nuclear total evidence approach), (2) concatenated matrix of 22 nuclear loci for *Leptosiphon* and a concatenated matrix of 14 nuclear loci for *Linanthus* that each produced a well-resolved phylogeny in preliminary analyses (hereafter referred to as the reduced nuclear data set), and (3) concatenated matrix of 80 plastid protein-coding regions for each genus. All analyses were conducted using RAxML on the University of Florida Research Computing cluster using the partitions identified by PartitionFinder and a GTR+G substitution model. Support values were determined by running 1,000 bootstrap replicates. The two species of *Gilia*, *G. brecciarum* subsp. *brecciarum* and *G. nevinii*, were used as outgroups.

To incorporate phylogenetic uncertainty in the character evolution analyses, separate Bayesian runs were conducted using MrBayes⁸⁰ (version 3.2.6) using the nuclear reduced data sets for each genus with the identified partitions from above. The *Linanthus* analysis was conducted for 5 million generations, while the *Leptosiphon* analysis was conducted for 7 million generations, each sampling every 1,000 generations. The final 1,000 trees of the posterior distribution were used as a sample of the Bayesian posterior distribution.

Data Availability

The list of datasets used in this study can be found in supplementary Supplementary Data 13 with reference to publications listed in supplementary Supplementary Data 11. Sequence data, alignments, phylogenetic trees, and metadata are archived and publicly available in Zenodo <https://zenodo.org/records/17603347> (DOI 10.5281/zenodo.17603347). The repositories for raw sequences of the published studies are specified in the cited publications, with the exception of *Linanthus* and *Leptosiphon*, *Attalea* and *Syagrus*, of which raw reads for each accession were deposited in GenBank's Short Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) under the NCBI Bioproject number PRJNA322057 and PRJNA1074667 respectively.

Code availability

All scripts for computational analyses used in this study is released on GitHub and archived on Zenodo. The archived release is available at <https://doi.org/10.5281/zenodo.18034794>.

Acknowledgments

This work was supported by a Darwin Trust of Edinburgh PhD studentship to Wu Huang; and also part funded by Horizon Europe under the Biodiversity, Circular Economy and Environment (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00173, and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme, supported by Wellcome through a Darwin Tree of Life Discretionary Award (218328). The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division. The large-scale analysis was conducted on the UK crop diversity bioinformatics HPC facility (<https://www.cropdiversity.ac.uk>). For a full list of data resources and providers please see Supplementary Data 11.

ARTICLE IN PRESS

References

- 1 Wiens, J. J. Species delimitation: new approaches for discovering diversity. *Systematic Biology* **56**, 875-878 (2007). <https://doi.org/10.1080/10635150701748506>
- 2 Steele, P. R. & Pires, J. C. Biodiversity assessment: state-of-the-art techniques in phylogenomics and species identification. *Am J Bot* **98**, 415-425 (2011). <https://doi.org/10.3732/ajb.1000296>
- 3 Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51-57 (2011). <https://doi.org/10.1038/nature09678>
- 4 Folmer, O., Black, M., Wr, H., Lutz, R. & Vrijenhoek, R. DNA primers for amplification of mitochondrial Cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**, 294-299 (1994).
- 5 Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* **270**, 313-321 (2003). <https://doi.org/10.1098/rspb.2002.2218>
- 6 CBOL, Plant Working Group, A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106**, 12794-12797 (2009). <https://doi.org/10.1073/pnas.0905845106>
- 7 Chase, M. W. *et al.* Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Ann Mo Bot Gard* **80**, 528-580 (1993). <https://doi.org/10.2307/2399846>
- 8 Hilu, K. W. & Liang, g. The *matK* gene: sequence variation and application in plant systematics. *Am J Bot* **84**, 830-839 (1997). <https://doi.org/10.2307/2445819>
- 9 Kress, W. J. & Erickson, D. L. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One* **2**, e508 (2007). <https://doi.org/10.1371/journal.pone.0000508>
- 10 China Plant BOL Group *et al.* Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences* **108**, 19641-19646 (2011). <https://doi.org/10.1073/pnas.1104551108>
- 11 Yao, H. *et al.* Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* **5** (2010). <https://doi.org/10.1371/journal.pone.0013102>
- 12 Hollingsworth, P. M. Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **108**, 19451-19452 (2011). <https://doi.org/10.1073/pnas.1116812108>
- 13 Hollingsworth, P. M., Graham, S. W. & Little, D. P. Choosing and using a plant DNA barcode. *PLoS One* **6**, e19254 (2011). <https://doi.org/10.1371/journal.pone.0019254>
- 14 Blaxter, M. Imagining Sisyphus happy: DNA barcoding and the unnamed majority. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371** (2016). <https://doi.org/10.1098/rstb.2015.0329>
- 15 Page, R. D. DNA barcoding and taxonomy: dark taxa and dark texts. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371** (2016). <https://doi.org/10.1098/rstb.2015.0334>
- 16 Rieseberg, L. H. & Carney, S. E. Plant hybridization. *The New Phytologist* **140**, 599-624 (1998). <https://doi.org/DOI 10.1046/j.1469-8137.1998.00315.x>

- 17 Edelman, N. B. & Mallet, J. Prevalence and adaptive impact of introgression. *Annual Review Genetics* **55**, 265-283 (2021). <https://doi.org/10.1146/annurev-genet-021821-020805>
- 18 Hollingsworth, P. M., Li, D. Z., van der Bank, M. & Twyford, A. D. Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371** (2016). <https://doi.org/10.1098/rstb.2015.0338>
- 19 Percy, D. M. *et al.* Understanding the spectacular failure of DNA barcoding in willows (*Salix*): does this result from a trans-specific selective sweep? *Molecular Ecology* **23**, 4737-4756 (2014). <https://doi.org/10.1111/mec.12837>
- 20 Schley, R. J. *et al.* Introgression across evolutionary scales suggests reticulation contributes to Amazonian tree diversity. *Molecular Ecology* (2020). <https://doi.org/10.1111/mec.15616>
- 21 Smith, D. R. Mutation rates in plastid genomes: they are lower than you might think. *Genome Biology and Evolution* **7**, 1227-1234 (2015). <https://doi.org/10.1093/gbe/evv069>
- 22 Birky, C. W., Jr. Uniparental inheritance of organelle genes. *Current Biology* **18**, R692-695 (2008). <https://doi.org/10.1016/j.cub.2008.06.049>
- 23 Dexter, K. G. *et al.* Dispersal assembly of rain forest tree communities across the Amazon basin. *Proceedings of the National Academy of Sciences* **114**, 2645-2650 (2017). <https://doi.org/10.1073/pnas.1613655114>
- 24 Zhang, L. *et al.* DNA barcoding of Cymbidium by genome skimming: call for next-generation nuclear barcodes. *Mol Ecol Resour* (2022). <https://doi.org/10.1111/1755-0998.13719>
- 25 Rieseberg, L. H. & Brouillet, L. Are many plant species paraphyletic? *Taxon* **43**, 21-32 (1994). <https://doi.org/10.2307/1223457>
- 26 Rieseberg, L. H., Wood, T. E. & Baack, E. J. The nature of plant species. *Nature* **440**, 524-527 (2006). <https://doi.org/10.1038/nature04402>
- 27 Pennington, R. T. & Lavin, M. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *The New Phytologist* **210**, 25-37 (2016). <https://doi.org/10.1111/nph.13724>
- 28 Roux, C. *et al.* Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology* **14**, e2000234 (2016). <https://doi.org/10.1371/journal.pbio.2000234>
- 29 Knapp, H. *Introductory statistics using SPSS ; 9* (SAGE Publications Ltd, United Kingdom, 2017).
- 30 Sedeek, K. E. *et al.* Genic rather than genome-wide differences between sexually deceptive *Ophrys* orchids with different pollinators. *Molecular Ecology* **23**, 6192-6205 (2014). <https://doi.org/10.1111/mec.12992>
- 31 Bateman, R. M., Sramko, G. & Paun, O. Integrating restriction site-associated DNA sequencing (RAD-seq) with morphological cladistic analysis clarifies evolutionary relationships among major species groups of bee orchids. *Annals of Botany* **121**, 85-105 (2018). <https://doi.org/10.1093/aob/mcx129>
- 32 Blaxter, M. L. The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**, 669-679 (2004). <https://doi.org/10.1098/rstb.2003.1447>

- 33 Ratnasingham, S. & Hebert, P. D. N. A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS One* **8**, e66213-e66213 (2013). <https://doi.org/10.1371/journal.pone.0066213>
- 34 Mishler, B. D. & Donoghue, M. J. Species concepts: a case for pluralism. *Systematic Zoology* **31** (1982). <https://doi.org/10.2307/2413371>
- 35 Judd, W. S., Campbell, C. S., Kellogg, E. A., Stevens, P. F. & Donoghue, M. J. *Plant systematics: a phylogenetic approach*. Vol. 25(2) 215 (Sinauer Associates, 1999).
- 36 Gernandt, D. S. *et al.* Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection *Australes*. *Am J Bot* **105**, 711-725 (2018). <https://doi.org/10.1002/ajb2.1052>
- 37 Yan, L. J. *et al.* DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya-Hengduan Mountains. *Mol Ecol Resour* **15**, 932-944 (2015). <https://doi.org/10.1111/1755-0998.12353>
- 38 Lavin, M. 197-212 (Cambridge University Press, 2022).
- 39 Kong, H. *et al.* Phylogenomic and macroevolutionary evidence for an explosive radiation of a plant genus in the Miocene. *Systematic Biology* **71**, 589-609 (2022). <https://doi.org/10.1093/sysbio/syab068>
- 40 Crawford, D. J. Progenitor-derivative species pairs and plant speciation. *Taxon* **59**, 1413-1423 (2010). <https://doi.org/https://doi.org/10.1002/tax.595008>
- 41 Soltis, P. S. & Soltis, D. E. The role of hybridization in plant speciation. *Annual Review Plant Biology* **60**, 561-588 (2009). <https://doi.org/10.1146/annurev.arplant.043008.092039>
- 42 Alix, K., Gérard, P. R., Schwarzacher, T. & Heslop-Harrison, J. S. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany* **120**, 183-194 (2017). <https://doi.org/10.1093/aob/mcx079>
- 43 Mutanen, M. *et al.* Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology* **65**, 1024-1040 (2016). <https://doi.org/10.1093/sysbio/syw044>
- 44 Munoz-Rodriguez, P. *et al.* A taxonomic monograph of *Ipomoea* integrated across phylogenetic scales. *Nature Plants* **5**, 1136-1144 (2019). <https://doi.org/10.1038/s41477-019-0535-4>
- 45 Becher, H. *et al.* Maintenance of species differences in closely related tetraploid parasitic *Euphrasia* (Orobanchaceae) on an isolated island. *Plant Communications* **1**, 100105 (2020). <https://doi.org/https://doi.org/10.1016/j.xplc.2020.100105>
- 46 Durán-Castillo, M., Hudson, A., Wilson, Y., Field, D. L. & Twyford, A. D. A phylogeny of *Antirrhinum* reveals parallel evolution of alpine morphology. *The New Phytologist* **233**, 1426-1439 (2022). <https://doi.org/10.1111/nph.17581>
- 47 Crisp, M. & Chandler, G. Paraphyletic species. *Telopea* **6**, 813-844 (1996). <https://doi.org/10.7751/telopea19963037>
- 48 Grant, V. *Plant Speciation*. (Columbia University Press, 1981), 10.7312/gran92318.
- 49 Grant, V. Plant speciation, the book: perspectives and paradigms. *The New Phytologist* **161**, 8-11 (2003). <https://doi.org/10.1111/j.1469-8137.2004.00964.x>
- 50 Lowry, D. B. *et al.* Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol Ecol Resour* **17**, 142-152 (2017). <https://doi.org/https://doi.org/10.1111/1755-0998.12635>

- 51 Hebert, P. D., Hollingsworth, P. M. & Hajibabaei, M. From writing to reading the
encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological
Sciences* **371** (2016). <https://doi.org/10.1098/rstb.2015.0321>
- 52 Sarmashghi, S., Bohmann, K., MT, P. G., Bafna, V. & Mirarab, S. Skmer: assembly-free
and alignment-free sample identification using genome skims. *Genome Biology* **20**, 34
(2019). <https://doi.org/10.1186/s13059-019-1632-4>
- 53 Duan, H. N. *et al.* Skmer approach improves species discrimination in taxonomically
problematic genus *Schima* (Theaceae). *Plant Divers* **46**, 713-722 (2024).
<https://doi.org/10.1016/j.pld.2024.06.003>
- 54 de Medeiros, B. A. S. *et al.* A composite universal DNA signature for the tree of life. *Nat
Ecol Evol* (2025). <https://doi.org/10.1038/s41559-025-02752-1>
- 55 Johnson, M. G. *et al.* A universal probe set for targeted sequencing of 353 nuclear genes
from any flowering plant designed using k-medoids clustering. *Systematic Biology*
(2018). <https://doi.org/10.1093/sysbio/syy086>
- 56 Pillon, Y. *et al.* Evolution and temporal diversification of western European polyploid
species complexes in *Dactylorhiza* (Orchidaceae). *Taxon* **56**, 1185-1208 (2007).
<https://doi.org/10.2307/25065911>
- 57 Huang, W., Twyford, A. D. & Hollingsworth, P. M. Acquiring and integrating data from
multiple resources for meta-analysis. (2023). <[https://www.protocols.io/view/acquiring-
and-integrating-data-from-multiple-resou-kxygx3z9og8j/v1](https://www.protocols.io/view/acquiring-and-integrating-data-from-multiple-resou-kxygx3z9og8j/v1)>.
- 58 Huang, W., Twyford, A. D. & Hollingsworth, P. M. NucBarcoder - a bioinformatic
pipeline to characterise the genetic basis of plant species differences. (2023).
<[https://www.protocols.io/view/nucbarcoder-a-bioinformatic-pipeline-to-characteri-
5qpvo33rzv4o/v1](https://www.protocols.io/view/nucbarcoder-a-bioinformatic-pipeline-to-characteri-5qpvo33rzv4o/v1)>.
- 59 Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD
markers. *PLoS One* **3**, e3376 (2008). <https://doi.org/10.1371/journal.pone.0003376>
- 60 Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high
diversity species. *PLoS One* **6**, e19379 (2011).
<https://doi.org/10.1371/journal.pone.0019379>
- 61 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double digest
RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model
and non-model species. *PLoS One* **7**, e37135 (2012).
<https://doi.org/10.1371/journal.pone.0037135>
- 62 Wang, S., Meyer, E., McKay, J. K. & Matz, M. V. 2b-RAD: a simple and flexible
method for genome-wide genotyping. *Nature Methods* **9**, 808-810 (2012).
<https://doi.org/10.1038/nmeth.2023>
- 63 Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nature
Methods* **7**, 111-118 (2010). <https://doi.org/10.1038/nmeth.1419>
- 64 Dodsworth, S. Genome skimming for next-generation biodiversity analysis. *Trends in
Plant Science* **20**, 525-527 (2015). <https://doi.org/10.1016/j.tplants.2015.06.012>
- 65 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics.
Nature Review Genetics **10**, 57-63 (2009). <https://doi.org/10.1038/nrg2484>
- 66 One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and
the phylogenomics of green plants. *Nature* **574**, 679-685 (2019).
<https://doi.org/10.1038/s41586-019-1693-2>

- 67 Schwery, O. & O'Meara, B. C. MonoPhy: a simple R package to find and visualize
monophyly issues. *PeerJ Computer Science* **2** (2016). <https://doi.org/10.7717/peerj-cs.56>
- 68 Paradis, E. pegas: an R package for population genetics with an integrated–modular
approach. *Bioinformatics* **26**, 419–420 (2010).
- 69 Jiménez, M. F. T., *et al.* Phylogenomics of the palm tribe Lepidocaryeae (Calamoideae:
Arecaceae) and description of a new species of *Mauritiella*. *Systematic Botany* **46**, 863–
874 (2021). <https://doi.org/10.1600/036364421X16312067913543>
- 70 Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh
leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1978).
- 71 Landis, J. B., O'Toole, R. D., Ventura, K. L., Gitzendanner, M. A., Oppenheimer, D. G.,
Soltis, D. E. & Soltis, P. S. The phenotypic and genetic underpinnings of flower size in
Polemoniaceae. *Frontiers in Plant Science* **6**, 1144 (2016).
<https://doi.org/10.3389/fpls.2015.01144>
- 72 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
EMBnet Journal **17**, 10–12 (2011). <https://doi.org/10.14806/ej.17.1.200>
- 73 Kent, W. J. BLAT – the BLAST-like alignment tool. *Genome Research* **12**, 656–664
(2002).
- 74 Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega,
N. J. C. & Wickett, N. J. HybPiper: extracting coding sequence and introns for
phylogenetics from high-throughput sequencing reads using target enrichment.
Applications in Plant Sciences **4**, 1600016 (2016).
- 75 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 76 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–
780 (2013).
- 77 Vaidya, G., Lohman, D. J. & Meier, R. SequenceMatrix: a concatenation software for the
fast assembly of multi-gene datasets with character set and codon information. *Cladistics*
27, 171–180 (2011).
- 78 Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: combined selection
of partitioning schemes and substitution models for phylogenetic analyses. *Molecular
Biology and Evolution* **29**, 1695–1701 (2012).
- 79 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 80 Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D., Darling, A., Höhna, S., Larget,
B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. MrBayes 3.2: efficient Bayesian
phylogenetic inference and model choice across a large model space. *Systematic Biology*
61, 539–542 (2012).

Figures and tables:

Figure 1. Quantification of the proportion of plant species within a genus that resolve as monophyletic based on nuclear genomic data for 151 studies. (A) Proportion of species resolving as monophyletic in each study (B) Comparison of plant groups with herbaceous and woody life forms (C) Comparison of different sequencing methods. Boxplots show the median, lower, and upper quartiles, with whiskers extending to 1.5 times the interquartile range; the dots in (C) are the monophyletic ratio for each study.

Figure 2. Distribution of the density of species-specific SNPs (SSSNPs) for 27 datasets from 26 genera (left panel, log-transformed), with the phylogenetic relationships among study taxa (middle panel), and the proportion of species that have more than one SSSNP (right panel). The right panel is the proportion of species in each genus that have more than one SSSNP (orange bars) and the proportion of species that resolve as monophyletic in that genus (blue bars).

Figure 3. The proportion of species discriminated using different numbers of sub-sampled SNPs. The x-axis is the number of SNPs randomly sub-sampled with 50 replicates for each sub-sampling interval. The intervals are increments of 200 SNPs between 100 to 1,900 SNPs and 1,000 SNPs between 3,000 to 10,000 SNPs. (A) Percentage of species discriminated across 23 datasets (n=23). The y-axis shows the percentage of species discriminated across all genera as a proportion of the total number of species resolved in the full dataset based upon monophyly. (B) The proportion of species discriminated with increasing numbers of randomly drawn SNPs for each genus separately.

Table 1. The minimal number of best-performing loci required to match the resolution of the full dataset, as measured by the number of species resolving as monophyletic.

| Genus | Number of loci in the dataset | Number of species resolved as monophyletic with all loci/Total number of multiple-sampled species | Minimum numbers of loci to achieve maximum species resolution | Number of species resolved as monophyletic by minimum number of loci |
|----------------------|-------------------------------|---|---|--|
| <i>Artocarpus</i> | 517 | 34 / 42 | 1 | 27 |
| <i>Capurodendron</i> | 615 | 13 / 20 | 7 | 14 |
| <i>Geonoma</i> | 795 | 30 / 44 | 1 | 30 |
| <i>Inga</i> | 810 | 45 / 69 | 9 | 44 |
| <i>Polemonium</i> | 360 | 7 / 12 | 1 | 7 |
| <i>Tsuga</i> | 881 | 7 / 8 | 1 | 7 |

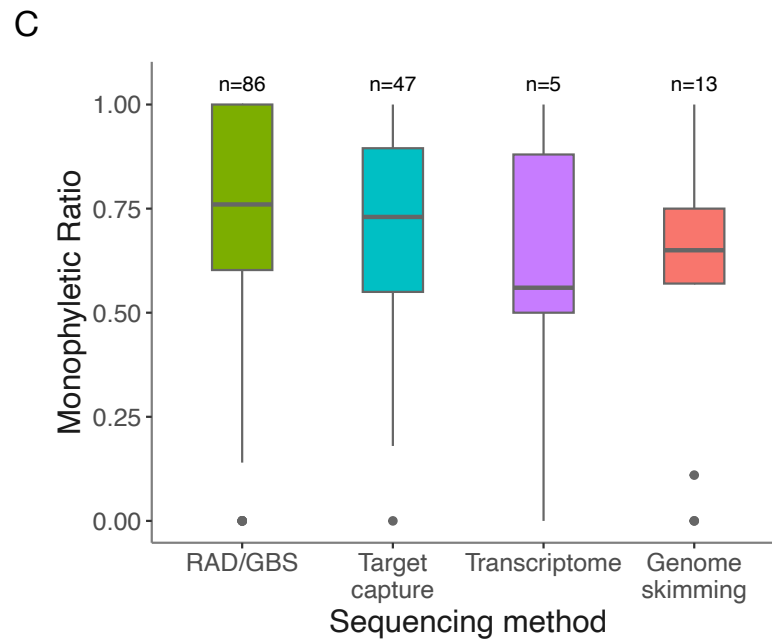
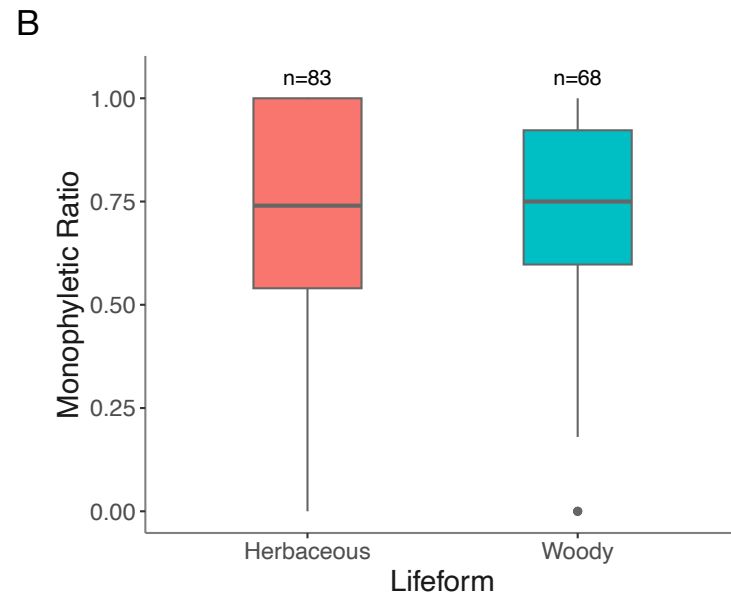
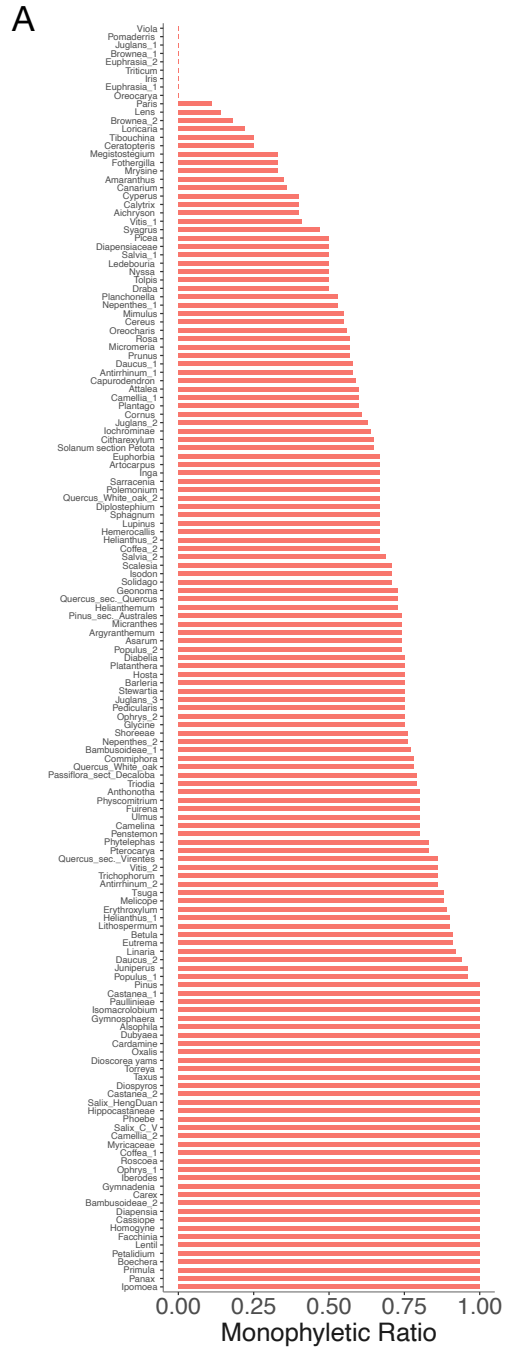
Editor Summary

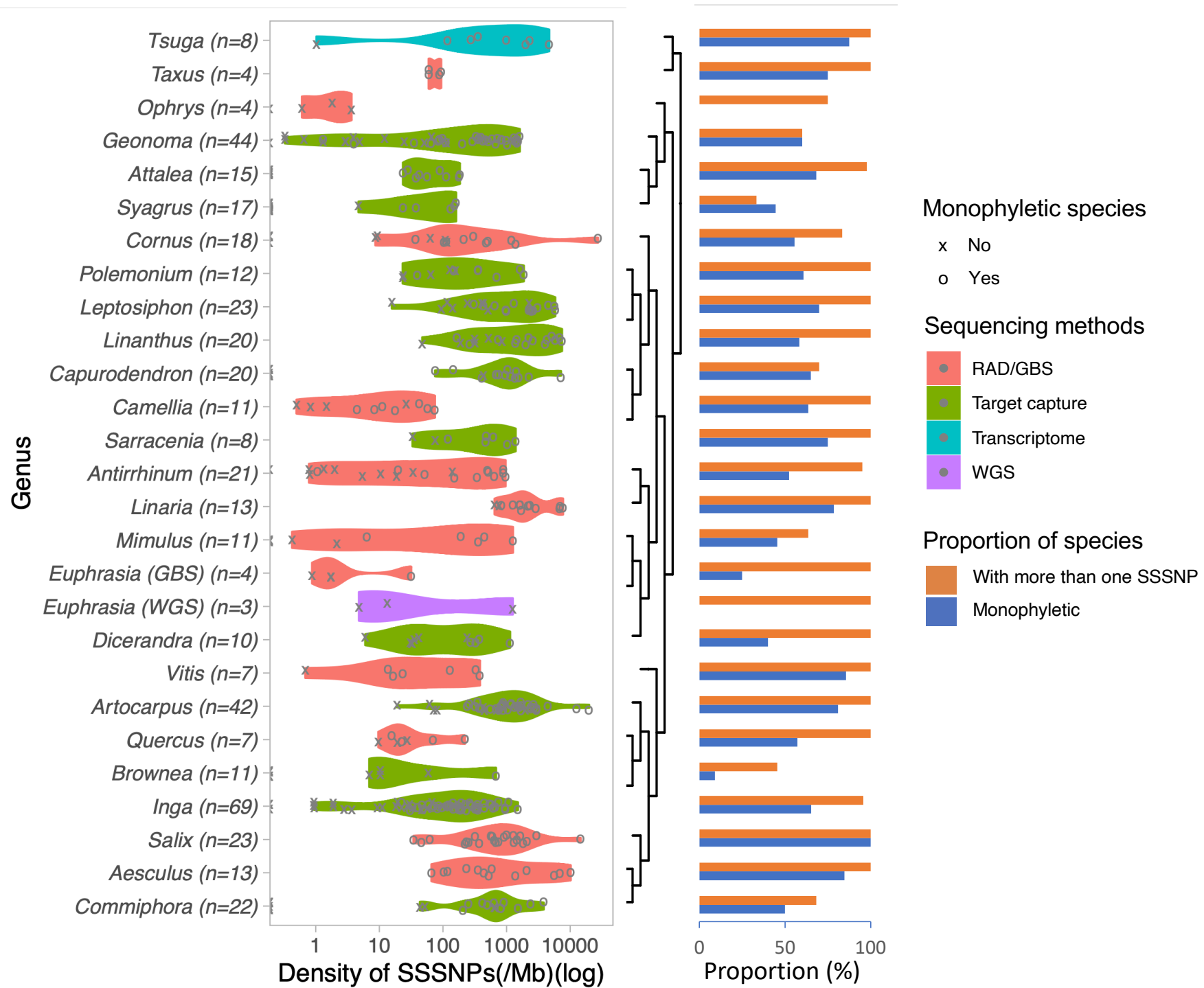
A meta-analysis of plant nuclear genomic data provides fundamental insights into the genomic nature of differences between plant species and informs future strategies for DNA-based identification of plants using nuclear genomes.

Peer Review Information

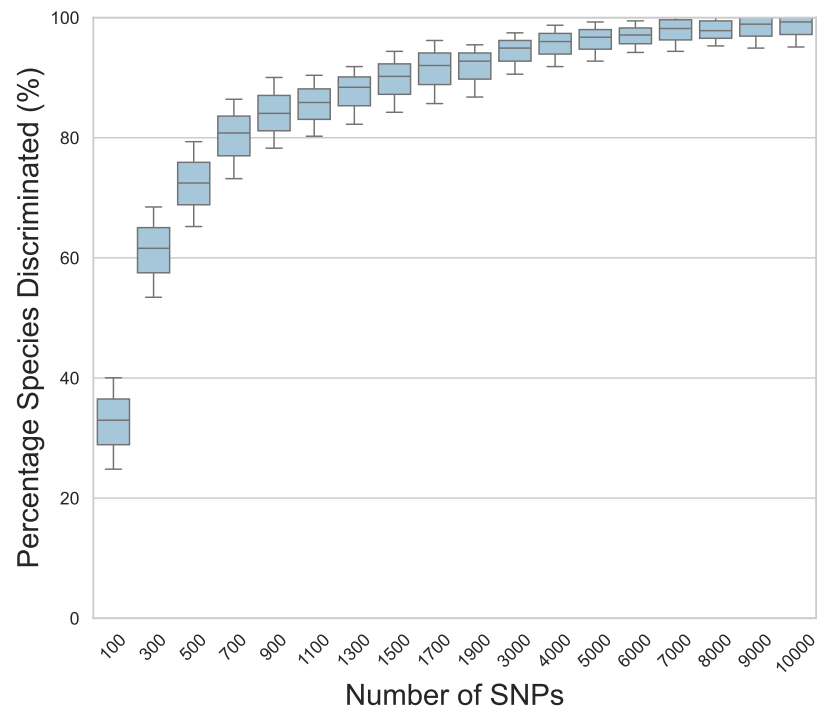
Communications Biology thanks Paula Reyes, Quentin Mauvisseau, Yves Bawin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Jorge Duitama and David Favero. A peer review file is available.

ARTICLE IN PRESS





A



B

