Peering into the nature of plant species

Sarah J. Jacobs[1,2], Claudia L. Henriquez[1], Felipe Zapata[1]*

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095

[2]Department of Botany, California Academy of Sciences, San Francisco CA 94118

* Corresponding author: Name: Felipe Zapata. Address: 612 Charles E. Young Dr. South, Los Angeles, CA 90095. Phone: (310) 206 4583. Email: fzapata@ucla.edu

Total number of words in Abstract: 200

Total number of words in Main Text (excluding Methods): 3109

Total number of Figures: 2

Total number of Tables: 4

Keywords

Cryptic species, Escallonia, speciation, species limits, syngameon

Abstract

What we mean by species and whether they have any biological reality has been debated since the early days of evolutionary biology. Consequently, many biologists suggest that species are created by taxonomists as a subjective, artificial division of nature. However, the nature of species has been rarely tested critically with data while ignoring taxonomy. We integrate phenomic and genomic data across hundreds of individuals at a continental scale to investigate this question in a group of angiosperms which includes multiple taxonomic species (the species proposed by taxonomists). Using statistical methods for species delimitation for phenotypic and genomic data, we show that plant species do exist as an objective, discrete property of nature independent of taxonomy. Nonetheless, we show that such species correspond poorly to taxonomic species ($< 20\%$) and that phenomic and genomic data seldom delimit congruent entities ($< 20\%$). We propose that phenomic and genomic data analyzed on an equal footing help build a broader perspective on the nature of species by delineating different 'types of species', which are consistent with speciation theory and emerging patterns across the tree of life. Our results caution studies which take taxonomic species for granted and challenge the notion of plant species without empirical evidence.

Introduction

A perennial question in biology concerns the possibility that plant species are not real, but presumably constructs of the psyche of taxonomists.[1–3] Previous researchers investigating this question through phenotypic data have focused on validating taxonomic species (i.e., the species proposed by taxonomists).[3,4] Taxonomic species are usually considered standard references to gauge the strength of the evidence in support of the reality of species when researchers analyze phenotypic data with numerical taxonomy methods to identify species.[5] The most comprehensive meta-analysis of studies using numerical taxonomy procedures to identify species for plants and animals has revealed that validation of taxonomic species is low ($< 60\%$ of statistically identified discrete clusters are congruent with taxonomic species) even though discrete phenotypic groups apparently exist in most taxonomic groups.[3] However, by using a species validation approach, as opposed to a species discovery approach,[6,7] this meta-analysis assumed that taxonomic species are present. Therefore, this study largely corroborated taxonomic preconceptions about species–entities that have been characterized as arbitrary constructs of the human mind–[2,8]rather than examining their reality. As a consequence, the fundamental question about the reality of plant species independent of the influence of taxonomists remains unanswered.[7] To date, no studies integrating phenotypic and genome-wide DNA data have assessed the reality of plant species for a group including multiple hypothesized taxonomic species at a broad geographic scale. Here we investigate this question through high-density phenotypic (ca. $8,300$ quantitative measurements) and genome-wide (ca. $1,000,000$ DNA sequences) analyses of a large data set of 848 individuals in *Escallonia* (Escalloniaceae), a group of shrubs and trees spanning the montane region of South America (Fig. 1, Supplementary Table S1).

In addition to the limitation described above, the meta-analysis of taxonomic studies[3] presents other shortcomings relevant to understanding the nature of plant species. First, it relies on taxonomic studies which use statistical methods disconnected from biological theory[9] and

hence are compromised in detecting biologically meaningful species. In particular, such studies use methods that rely on graphical analyses that convey little information on phenotype frequencies, exclude phenotypic traits potentially important for species detection, and use measures of central tendency which are inconsequential to assess species distinctiveness.[10] Second, it analyzes studies biased toward 'problematic taxa' (i.e., species complexes, hybrid swarms) in which statistical methods have been historically applied to taxonomy, and thus it may provide a distorted general perspective on the nature of plant species. Third, it does not investigate the question about the nature of plant species directly using genetic data which bear an explicit relationship to evolutionary divergence and gene flow, two relevant criteria in delineating species.[11] Lastly, it does not consider the evidence of species in a geographic context despite the central role of geography in the study of species and speciation.[12–14] We tackle these limitations in examining the nature of plant species by integrating multiple types of data and proper statistical approaches well grounded on evolutionary theory in a typical genus of flowering plants, seemingly composed of 'good' taxonomic species.[15]

Elucidating the nature of plant species has broader implications beyond taxonomy. Species are fundamental units of analysis in ecology and evolution. Therefore, determining whether species are objective biological entities is critical to understanding the origin, evolution, and structure of biodiversity. In particular, discovering discrepancies between phenotypes and genotypes can shed light into how new species are created by understanding how geographic variation within species transitions to variation between distinct species.[16] In addition, examining whether the taxonomic species commonly used by ecologists and evolutionary biologists correspond to the biological units product of natural processes can influence our understanding of the hypotheses explaining the patterns and processes in the natural world.

Results and Discussion

We present and discuss the major findings below in the context of the whole *Escallonia* radiation. Detailed results are presented in the Supplementary Material.

The current state of taxonomic species

We first characterized the evolutionary history of *Escallonia* using different phylogenetic approaches with a subset of specimens spanning the geographic range of these plants across South America (Fig. 1, Supplementary Figures S1, S2). In all of these analyses, we consistently recover six groups of taxonomic species (hereafter, clades I-VI), in line with a previous study based on fewer loci.[17] All clades are markedly restricted to geographic regions, except clade VI; this clade is mainly restricted to southeastern Brazil, Uruguay, and northeastern Argentina, but includes some species in the Andes (Fig. 1). A closer examination of the relationship between clade composition and the geographical as well as elevational distributions of clades reveals that when specimens from different clades co-occur in close spatial proximity (e.g., Clades I, II, III, IV in the Tropical Andes), clades are genetically distinct with no intermixing (Fig. 1, Supplementary Figures S1, S2). Further, all clades have consistent composition and receive strong statistical support when we use different approaches to phylogenetic analysis (See Methods). However, when we include multiple specimens of the same taxonomic species, several of these specimens are not always each other's closest relatives within clades (i.e., taxonomic species are polyphyletic; Supplementary Figure S2). This result, along with the marked phylogenetic geographic concordance and consistent composition of clades, suggests that although clades are evolutionarily distinct, the state of taxonomic species within clades may be in question.[17] Therefore, we focus our subsequent analyses of phenotypic and genome-wide variation to investigate the nature of species in *Escallonia* on a clade by clade basis.

To investigate the current state of taxonomic species through phenotypic data, we used the

morphological characteristics—leaf and floral traits—provided in the taxonomic description of each species.[15] We focused on these traits because taxonomic descriptions include the characters useful in distinguishing all species and in comparing them with other species.[18] We first tabulated the maximum and minimum values of ten quantitative continuous traits provided in each species description (these values are derived from specimens not included in the current dataset). We then used these values as vertices of a 10-cube to represent each species geometrically in phenotypic space and estimated the pairwise overlap among all 10-cubes within clades. This analysis shows that taxonomic species within clades occupy distinct regions of 10-dimensional phenospace with little to no overlap (Table 1, Supplementary Figures S5, S16, S27, S38, S49, S60). We followed these geometric-based analyses with a matching-prediction analysis whereby we assessed whether each specimen identified to a taxonomic species was inside or outside the 10-cube of its corresponding species based on quantitative measurements of the morphological traits defining the 10-cube (See Methods). Contrary to expectations, these analyses show that the majority (99.2%) of specimens fall outside their respective 10-cube. Furthermore, 98.4% specimens fall outside any 10-cube (Table 1, Supplementary Figures S5, S16, S27, S38, S49, S60). Although these results may simply reflect issues emerging from both the statistical and mathematical properties of high-dimensional data spaces,[19,20] it is plausible that taxonomic descriptions do not capture the biological complexity of species (e.g., trait covariances), and hence taxonomic species have low explanatory power because they do not correspond to real entities in nature. Indeed, given that plant species are rarely delimited and described with morphology based on explicit analyses grounded on biological theory,[21,22] our results suggest that investigating the nature of plant species by relying on validating taxonomic species can be problematic.

Evolutionary model-based evidence to identify species as objective entities

We used Gaussian finite mixture modeling (GFMM)[23] within clades to determine both the number of species and the assignment of specimens to species using phenotypic data without

prior information about taxonomy. This modeling framework is well-suited for this problem because it implements the evolutionary model underlying the use of quantitative, continuous phenotypic variation in species discovery and delimitation.[9,10] To perform this analysis, we used the same specimens and the same ten diagnostic morphological traits as in our previous analysis (see above). Importantly, previous studies have used these phenotypic traits to characterize taxonomic species and define species boundaries in *Escallonia*.[15] We rotated the original data matrix into orthogonal axes using robust covariance estimators and reduced the dimensionality of the orthogonal axes to only those that optimized the shape, orientation, and the number of phenotypic-based species (hereafter, phenogroups). We identified the best Gaussian Mixture Model - GMM (Naive model) in each clade in a Bayesian information criterion (BIC) and integrated complete-data likelihood (ICL) framework. In addition, we assessed support for alternative models in which we assigned specimens to groups defined *a priori*, including taxonomic species (Taxonomy model) as well as phenogroups we defined independent of taxonomy (Taxonomy Unaware model). The results from these analyses are shown in Figure 1 and Table 2. The Naive model was the best-supported model for five of the six clades (ΔBIC$>$ 8), while one clade had support (ΔBIC$<$ 1) even though the model was not the best supported for this clade (Supplementary Figure S39). These results were insensitive to model-selection approach (BIC or ICL) (See Supplementary Material). The strong performance of the Naive model is not unexpected owing to the severe limitations of the competing, non-statistical approaches to delimit species without considering the shape, orientation, and arbitrary overlap of phenogroups in multidimensional phenotypic space[10] (Supplementary Figures S6, S17, S28, S39, S50, S61). This is also consistent with the prediction that nature is, in fact, discontinuous[24,25] despite suggestions that species are not discrete objective entities.[2] Furthermore, because the majority of the identified phenogroups within clades co-occur locally in sympatry (Fig. 1, Supplementary Figures S6, S17, S28, S39, S50, S61), species status for these groups is granted under a wide range of species definitions.[10,11,14,26] Yet, phenogroups may conceal distinct species when similar phenotypes

have evolved independently.[27] Thus, incorporating phylogenetic information is beneficial in understanding the nature of species and deciding whether all phenogroups are distinct species.

In order to identify species and assign specimens to species within clades using genetic data, we evaluated the fit of three common species delimitation models. These models implement three different species definitions, namely species defined as genotypic clusters[28,29] (GC model), species defined as the transition point from cladogenesis to anagenesis[30,31] (CA model), and species defined as reproductively isolated lineages[12,32] (RI model). For this analysis, we collected genome-wide data for a subset of the specimens used in our phenotypic analyses and compared competing species delimitation models in a Bayesian framework using Bayes factors[33] to identify genomic-based species (hereafter, genogroups). Because neither taxonomic species nor any other *a priori* groups have been proposed based on genetic data, we did not assess support for any other alternative species delimitation models. Figure 1 and Table 3 show the results of these analyses. In general, the CA model outperformed the alternative models; in five of six clades, the CA model was the best-supported model, while the GC model fit better for only one clade. Across clades, the best fitting model identified the largest number of genogroups. The reason why the models with more genogroups fit better in all clades is likely the result of the higher genetic variation between genogroups than within genogroups, apparent as long branches in the species trees (Fig. 1). This suggests that genogroups are divergent lineages on separate evolutionary trajectories, and is consistent with the hypothesis that such lineages are distinct species.[7,11] Moreover, several of these genogroups within clades co-occur locally in sympatry, and thus species status for such groups is granted under multiple species definitions.[12,14,26] However, in some clades genogroups form isolated, allopatric groups of specimens, which could presumably result from sparse geographic sampling within a single species.[34] Therefore, the weight of the evidence in support of the species status for these genogroups is weak and requires considering other lines of evidence on an equal footing.

Integrating phenotypic and genome-wide variation, spatial information, and evolutionary history

With the phenogroups and genogroups derived from the evolutionary model-based analyses, we were able to examine the nature of species by integrating phenotypic and genome-wide data in an explicit spatial and evolutionary context (Fig. 1, Supplementary Figure S13, S24, S35, S46, S57, S68). For this analysis, we first assigned each specimen to its corresponding phenogroup and genogroup, akin to a two-way contingency table (Fig. 2). This assignment allowed the identification of congruence–or lack thereof– between phenotypic and genomic groups. Some specimens were incomplete (e.g., sterile) and could not be scored for all phenotypic traits, while other specimens failed during processing for genomic work (hereafter, unknown specimens); nevertheless, the geographic distribution of these unknown specimens in relation to the specimens with both kinds of data may inform the most parsimonious pheno- or genogroup assignment (for example, in Clade IV all the unknown specimens from northern South America likely belong to phenogroup 2 and genogroup 1; Fig. 1). Overall, we found that only a small percentage of phenogroups correspond directly to unique genogroups (15%), even assuming concordant group assignment for all unknown specimens (18%). By contrast, we found that in most clades a given phenogroup occurs across multiple genogroups (for example, see phenogroup 2 in clade IV, Fig. 2), and less frequently that a given genogroup occurs across different phenogroups (for example, see genogroup 9 in clade V, Fig. 2). Taken together, our results suggest that the proportion of 'good species' (i.e., phenotypic and genomic distinct and congruent groups) in *Escallonia* is remarkably low, particularly given the widespread notion in biology that 'good species' are the norm, and suggest that other types of species, including 'phenotypic cryptic species'[27] (i.e., one phenogroup across multiple genogroups) and 'genetic cryptic species'[35] (i.e., one genogroup across multiple phenogroups), are more common. The existence of these different types of species is consistent with the idea that the properties of species, such as morphological distinguishability or genealogical exclusivity of

alleles, may evolve at different times and sequential order owing to the heterogeneous nature of the speciation process.[36,37]

Interpreting the species that we identified in an explicit spatial and phylogenetic context can further elucidate the nature of plant species. Most 'good species' co-occur in local sympatry or segregate according to elevation with other species (Fig. 1, Fig. 2, Supplementary Figures S13, S24, S35, S46, S57, S68). This suggests that environmentally-mediated selection in sympatry or along elevational gradients in parapatry may be an important evolutionary force driving speciation[38] or at least maintaining species differences in *Escallonia*. Alternatively, it is possible that these species have evolved later than other species during the speciation continuum and have accumulated enough differences.[39,40] Further sampling in combination with phylogenetic dating approaches and experimental data are desirable to evaluate these hypotheses with increasing rigor. When the genogroups of 'phenotypic cryptic species' are distantly related, a reasonable hypothesis to explain this pattern is the idea of convergent evolution in phenotypes in response to similar selective regimes, either in sympatry or allopatry[41] (for example, see phenogroup 1, genogroups $2, 4, 10, 11$, clade VI; Fig. 1). By contrast, when such genogroups are each other's closest relatives and do not co-occur locally in sympatry (for example, see phenogroup 2, genogroups $1, 2$, clade III; Fig 1), under some species definitions genogroups may correspond to allopatric populations within a single species[12] rather than to distinct species resulting from recent speciation with little time for phenotypic differentiation, or speciation with niche conservatism.[41,42] Exhaustive geographic sampling is necessary before these hypotheses can be confronted confidently and the nature of these species is better understood. In all the 'genetic cryptic species' that we identified, phenogroups do not show a strong geographic structure (for example, see genogroup 10, phenogroups $2, 3, 5, 7$, clade V; Fig. 1). This is consistent with the intriguing possibility that these otherwise phenotypically distinct species may be interconnected via gene interchange, likely facilitated by their broad overlap in geographical space.[43,44] Indeed, genomic evidence for this type of species is rapidly accumulating for other plants[45–47] as well as various taxa

across the tree of life.[35,48] Yet, how these groups of species are initiated and persist, and what portion of their genomes is exchanged freely across species boundaries without species collapse needs to be studied in closer detail.[49] Alternatively, these species may be the result of rapid divergence events driven by strong factors influencing traits relevant for ecological isolation with little time for alleles to sort completely between sister species.[50] Further taxon and genome sampling in combination with explicit population genomic models are required to isolate the signal of incomplete lineage sorting from hybridization between sister species.[51]

Conclusion

In sum, our analyses of a large scale phenotypic and genome-wide dataset using state of the art model-based approaches for species discovery and delimitation reveal that plant species do exist as a property of nature independent of taxonomy.[7,25] However, the observed pattern of excessive discordance between species identified with phenotypic and genomic data suggests that in the absence of evidence the prevalent assumption that phenotypically (or genetically) distinct entities are necessarily 'good species' is not warranted. Furthermore, parallel signatures of such discordance across divergent clades suggest that this may be a widespread phenomenon, which is consistent with the emerging patterns about the nature of species across the tree of life.[27,35,46–48,52] Previous studies have proposed that approximately 70% of plant taxonomic species represent good, biologically real species,[3] but this is not supported in our study. Instead, our results suggest that the percentage of taxonomic species which correspond to 'good species' may be as low as 17% (Table 4, Supplementary Table S4, S7, S10, S13, S16, S19). To the extent that our findings capture any generalizable perspective about the nature of plant species, reinforced by the overall poor theoretical basis underlying plant species delimitation,[21,22] our results suggest that studies in other areas of biology which assume taxonomic species represent good, biologically real entities may need critical evaluation. Our results underscore the need of further comparative studies combining high-throughput phenotypic and genotypic data across taxa and across broad and

narrow spatial scales to comprehensively understand the nature of plant species.[7] Given the unprecedented advances in phenomics, genomics, and computation, there has not been a more thriving time to be a taxonomist than now.

Methods

Taxon sampling and data collection This study complies with local and national regulations. Collecting permits were obtained for field collections. A total of 848 specimens were included in this study (a mix of field collections and herbarium specimens). These specimens covered the entire geographic range of *Escallonia*. To assign specimens to taxonomic species, one of us (Felipe Zapata) identified all plant material using the dichotomous key provided by Sleumer[15] as well as information on habit, habitat, geographic locality, and the available comparative material from ca. $3,500$ herbarium collections. Complete voucher information for all specimens is available in Table S1. On these specimens, we measured 10 quantitative, continuous phenotypic traits (leaf length, leaf width, pedicel length, ovary length, length of calyx tube, length of calyx lobes, petal length, petal width, filament length, style length) to characterize the geographic pattern of phenotypic variation across *Escallonia*. All measurements were log-transformed prior to downstream analysis.

To examine the geographic pattern of genomic variation across *Escallonia*, we used double-digest Restriction-Site Associated DNA Sequencing (ddRAD)[53] for 315 specimens (out of the 848 specimens). We first extracted DNA from silica-dried adult leaves or herbarium specimens and then prepared quadruple-indexed, triple-enzyme RADseq libraries using the *EcoRI*, *XbaI*, and *NheI* restriction enzymes.[54] All libraries were sequenced across multiple lanes of 100PE sequencing on the Illumina HiSeq 4000 Sequencing Platform. We assembled RAD loci and called variants using `iPyrad` v0.7.28,[55] and filtered files for downstream analyses using `VCFtools` v0.1.14[56] and custom-made scripts. To assess the sensitivity of our results to the amount of missing data, we ran analyses using three matrices with different levels of

missing data (25%, 50%, and 75% missing data). Detailed descriptions on sampling and data collection are provided in the Supplementary Material.

The current state of Escallonia taxonomic species We used a subset of specimens to reconstruct the phylogeny of *Escallonia*. We chose these specimens to represent the overall spectrum of morphological variation and the geographic range of each taxonomic species. We used *Valdivia gayana* as outgroup.[17] We built phylogenies with two and four specimens per taxonomic species using the three data matrices with different amounts of missing data. For each dataset, we inferred lineage trees using a matrix of concatenated full loci in `IQ-TREE` v2.0.3 and the edge-proportional partition model with 1000 ultrafast bootstrap replicates.[61] To infer species trees, we used `SVDQuartets`[62] in `PAUP*` v4.0a168[63] by evaluating all possible quartets. Confidence on species trees was assessed with a multilocus bootstrap analysis using 100 replicates. Both the lineage and species tree reconstructions across all subsets consistently recovered six well-supported clades (See Results; clades I-VI). We conducted all downstream analyses within clades considering only ingroup samples.

To examine the state of taxonomic species through phenotypic data, we used the most recent taxonomic monograph of *Escallonia* to tabulate the minimum and maximum values reported for ten quantitative traits used to describe and delimit each taxonomic species.[15] The combination of these values predicts a hypervolume in phenotypic space occupied by each taxonomic species. Therefore, we used these values as vertices to construct a hypervolume (i.e., a 10-cube) to represent geometrically each species in 10 phenotypic dimensions. To determine the distinctiveness of each taxonomic species, we estimated the pairwise asymmetric proportion of overlap of all 10-cubes within clades. To assess whether the specimens that we measured in this study matched the prediction specified by the taxonomic description of each species (i.e., whether specimens were inside the space defined by the hypervolume in phenotypic space), we used the morphological measurements to ask whether specimens assigned to a taxonomic species were inside or outside the 10-cube of their corresponding

taxonomic species. We used this approach because taxonomic descriptions include all the characters useful in distinguishing species and in comparing them with other species in multidimensional phenospace.[18] Therefore, our approach provides a reasonable assessment of the range of variation present in nature predicted to be partitioned by each taxonomic species. We refer to this analysis as 'matching-prediction analysis'. We did not include meristic or qualitative traits in this analysis because we focused on the same traits that we analyzed using explicit methods for species discovery and delimitation with phenotypic data, which are grounded on evolutionary theory (see below). We used the `R` packages `grDevices`[64] and `geometry` v0.4.5[65] to carry out these analyses. Further details are provided in the Supplementary Material.

Model-based evidence for species using phenotypic data To determine the number of phenotypic-based species (hereafter, phenogroups) and the assignment of specimens to phenogroups within clades, we applied the quantitative genetics model for the distribution of continuous quantitative traits within a species.[9] This model states that under the assumption of polygenic architecture for phenotypic traits and random mating, gene frequencies would be close to Hardy–Weinberg equilibrium and phenotypic variation among individuals of a single species would tend to be normally distributed.[66] We applied this Fisherian model employing Gaussian Finite Mixture Modeling (GFMM) to search for the mixture of normal distributions (i.e., phenogroups) that best explains the variation in the data.[23] GFMM is a powerful framework to model the phenotypic variation of species seen in nature because it can combine normal distributions of different shapes and orientations.[10] To define the phenotypic space for GFMM, we first used robust principal components analysis (rPCA)—an approach useful for high dimensional data when outliers could skew the orientation of the rotated axes markedly—[67]on our ten, log-transformed, quantitative traits. We then used automatic variable selection[68,69] to select the most useful set of robust PC axes for GFMM using forward variable selection and no variable transformation. Lastly, we fitted different Gaussian Mixture Models (GMM) specifying 1 to $n + n/2$ number of phenogroups, where

n is equal to the number of taxonomic species currently hypothesized to exist within each clade. This approach aimed to limit the number of phenogroups present in the mixture to a reasonable number informed by current taxonomy and minimize over-differentiation of phenogroups. We evaluated three competing models for phenogroup delimitation:

Naive model The optimal GMM was determined without *a priori* assignment of specimens to phenogroups.

Taxonomy model The GMM used specimens assigned *a priori* to taxonomic species (See above)

Taxonomy unaware model The GMM used specimens assigned *a priori* to groups based on a comparative, non-explicit analysis of phenotypic variation (i.e., phenogroups were determined by eye).

Model Selection To determine the best fit model —including the number, orientation, and shape of phenogroups in the mixture as well as the assignment of specimens to phenogroups— , we used the Bayesian information criterion (BIC)[70] and the integrated complete-data likelihood (ICL) criterion.[71] We used the `R` packages `pcaPP` v1.9-73[72] and `mclust` v5.4.6[73] to perform these analyses. Further details are provided in the Supplementary Material.

Model-based evidence for species using genomic data Because our sensitivity analyses were robust to the amount of missing data (See Supplementary Material), we performed the following analyses using the matrix with the lowest amount of missing data (25% missing data) for computational efficiency. To determine the number of genomic-based species (hereafter, genogroups) and the assignment of specimens to genogroups within clades, we evaluated three competing models for genogroup delimitation. In all analyses, we did not assign specimens to genogroups *a priori*.

GC model (genotypic clusters model) This model is in essence the operational equivalent with genetic data of the Fisherian model described above. It states that the presence of

two or more genotypic clusters in a sample of individuals provides evidence for more than one species because distinct genetic clusters are recognized by a deficit of intermediates, both at single and multiple loci.[28] To delimit these genogroups, we employed GFMM in genotypic space.[29] Using the matrix with a single nucleotide polymorphism (SNP) per locus, we first estimated the shared allele distance,[74] defined as one minus the proportion of alleles shared by 2 individuals averaged over loci. Loci with missing data were not considered in the pairwise distance calculation. To define the genotypic space for GFMM, we followed Huasdorf and Hennig[29] and used non-metric multidimensional scaling (NMDS) to reduce the dimensionality. In all clades, we retained only two dimensions (stress $< 15\%$). In this space, we fitted different GMM specifying 1 to $n + n/2$ number of phenogroups, where n is equal to the number of taxonomic species currently hypothesized to exist within each clade. To determine the best GMM, we used the Bayesian Information Criterion (BIC). We used the `R` package `prabclus` v2.3-2[75] to carry out these analyses.

CA model (cladogenesis to anagenesis model) This model states that species reside at the transition point between evolutionary relationships that are best represented cladogenetically (i.e., between-species branching events) and relationships that are best reflected anagenetically (i.e., within-species branching events).[30] To delimit these genogroups, we applied an explicit phylogenetic model to identify significant changes in the pace of branching events on a phylogeny.[31] Under the assumption that the number of substitutions between species is significantly higher than the number of substitutions within species, these differences are reflected by branch lengths that represent the mean expected number of substitutions per site between two branching events (cladogenesis and anagenesis). We applied this model within clades employing multi-rate Poisson tree process modeling in the `mPTP` software.[76] We used the concatenated matrix with complete sequences for all loci and generated a phylogenetic tree per clade using `IQ-TREE` v2.0.3 with ultrafast bootstrap approximation to assess branch support.[58,59] Because `mPTP` requires a rooted phylogeny, we mid-point rooted each phylogeny using the R package `phytools` v0.6-99.[77] We ran `mPTP` under both a maximum likelihood and

a Bayesian framework with a minimum branch length threshold of 0.0001 for all analyses. For Bayesian runs, we used default priors and generated 500 million samples (i.e., independent delimitations), sampling every 1 million steps and ignoring the first 1 million as burn-in. We summarized all runs to indicate the percentage of delimitations in which a node was identified as a cladogenesis event (nodes with values closer to 1) or a transition to anagenesis (nodes with values closer to 0).

RI model (reproductive isolation model) This model states that species are evolutionarily independent groups of individuals which do not exchange genes.[12] To delimit these genogroups, we used an explicit population genetic framework[78] which, under the assumption of extremely limited to absent gene flow after speciation, models the evolution of gene trees within species and identifies groups of individuals in gene trees that are shared across loci.[79] We applied this model within clades employing a Bayesian modeling framework using the software `BPP` v4.0[80] in the analysis mode `A11`.[81] Because `BPP` requires that specimens are assigned *a priori* to 'genetic populations' (i.e., demes), we used the matrix with one SNP per locus and employed model-based clustering for this initial step. This clustering approach uses multilocus genotypes to find demes that (as far as possible) are in Hardy-Weinberg or linkage equilibrium. We applied this model-based clustering approach in a Bayesian framework using the programs `STRUCTURE` v2.3.4[82] and `rMaverick` v1.0.5,[83] which uses thermodynamic integration instead of the heuristic estimators used in `STRUCTURE`. For both analyses, we fitted different models specifying 1 to $n + n/2$ number of demes, where n is equal to the number of taxonomic species currently hypothesized to exist within each clade. To determine proper exploration across different species delimitation models, we used both algorithms (0 and 1) implemented in `BPP`[79] and compared the results across replicated runs. For each run, we used a random starting tree and a chain with at least $500,000$ steps, sampling every 10 step and discarding the first $1,000$ samples as burn-in. Further details are provided in the Supplementary Material.

Model Selection To determine the best fit model for genogroup delimitation —including the number of genogroups and the assignment of specimens to genogroups—, we used Bayes factor delimitation (*with genomic data; BFD*).[84] For this analysis, we used an explicit population genetic model to compute the likelihood of a species tree directly from the SNP datasets, which bypasses the sampling of the gene trees at each locus.[85] To properly compare candidate species delimitation models, we applied the scaling of the marginal likelihood proposed by Leaché et al..[84] We applied this framework employing the Bayesian Markov chain Monte Carlo (MCMC) sampler `SNAPP` v1.4.1,[85] which we ran through the software `BEAST` v2.5.2.[86] BFD* uses path sampling to estimate the marginal likelihood of the species delimitation models.[84] We conducted path sampling with 24 steps, using an MCMC with $250,000$ steps, sampling every 10 steps, and ignoring the first $12,500$ steps as burn-in. If each of the 24 steps achieved an effective sample sizes (ESS) $\geqslant 100$, we inferred convergence; otherwise, we ran a second path sampling with 24 more steps using an MCMC with $500,000$ steps and $25,000$ steps as burn-in. We compared competing models and chose the best model fit for genogroup delimitation using Bayes factors according to the framework provided by Kass and Raftery.[87] A Bayes factor (BF) statistic ($2 \times log_e$) > 10 provides decisive evidence favoring the highest ranked model. To carry out these analyses, we ran `BEAST` v2.5.2 on the CIPRES Science Gateway v3.3.[88] Further details are provided in the Supplementary Material.

Integrating phenotypic and genome-wide variation, spatial information, and evolutionary history Based on the best fit models for phenogroup and genogroup delimitation, we assigned all specimens to their corresponding phenogroup and genogroup. Because each specimen was necessarily assigned to a single phenogroup and a single genogroup, we determined three types of species according to the possible combinations of phenogroup and genogroup assignment. First, specimens assigned to a single phenogroup and a single genogroup delineated species that we determined as 'good species'. Second, specimens assigned to a single phenogroup across multiple genogroups delineated species that we determined as 'phenotypic cryptic species'. Third, specimens assigned to a single genogroup across

multiple phenogroups delineated species that we determined as 'genetic cryptic species'. Several specimens did not have overlapping phenotypic and genomic data (e.g., old herbarium specimens for which only phenotypic data were available, sterile specimens for which only genomic data were available). Therefore, we assigned these specimens only to their corresponding phenogroup or genogroup, accordingly. We referred to these specimens as 'unknown specimens'. To interpret the different types of species and the 'unknown specimens' in an evolutionary context, we mapped the phenogroup and genogroup assignments onto the tips of the phylogenies inferred in the CA model analysis (See above). Similarly, we interpreted the different types of species and the 'unknown specimens' in a spatial context using the geolocation data available for each specimen. Both the evolutionary and spatial contexts provided insight into the nature of plant species by illustrating patterns of common ancestry and patterns of sympatry/allopatry across geography and elevation.

Correspondence between taxonomic species and model-based species To compare the taxonomic species with the species we delimited based on phenotypic and genomic data, we assigned all specimens to their corresponding taxonomic species, and to their best fit phenogroup and genogroup. Because each specimen was necessarily assigned to a single taxonomic species, phenogroup, and genogroup, we counted the number of 'perfect matches'. A perfect match is defined as a symmetrical match between a unique taxonomic species and a unique phenogroup, genogroup, or combination of phenogroup and genogroup. For instance, specimens assigned to species x and uniquely to phenogroup a as well as assigned uniqely to phenogroup a and species x represent a perfect match. This assessment enabled us to determine the number of taxonomic species that represent 'good species'.

References

1. Lewis, H. The nature of plant species. *Journal of the Arizona Academy of Science* **1**, 3–7 (1959).

2. Levin, D. A. The nature of plant species. *Science* **204**, 381–384 (1979).

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

4. Mayr, E. A local flora and the biological species concept. *American Journal of Botany* **79**, 222–238 (1992).

5. Sneath, P. H. & Sokal, R. R. *Numerical taxonomy. The principles and practice of numerical classification.* (1973).

6. Carstens, B. C., Pelletier, T. A., Reid, N. M. & Satler, J. D. How to fail at species delimitation. *Molecular ecology* **22**, 4369–4383 (2013).

7. Barraclough, T. G. *The evolutionary biology of species.* (Oxford University Press, 2019).

8. Ehrlich, P. R. & Raven, P. H. Differentiation of populations. *Science* 1228–1232 (1969).

9. Fisher, R. A. XV.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52**, 399–433 (1919).

10. Cadena, C. D., Zapata, F. & Jiménez, I. Issues and perspectives in species delimitation using phenotypic data: Atlantean evolution in darwin's finches. *Systematic Biology* **67**, 181–194 (2018).

11. Queiroz, K. de. The General Lineage Concept of Species, Species Criteria, and the Process. in *Endless forms: Species and speciation* (eds. Harrison, R. G. & Berlocher, S. H.) 57–75 (1998).

12. Mayr, E. *Populations, species, and evolution: An abridgment of animal species and evolution*. vol. 19 (Harvard University Press, 1970).

13. Levin, D. A. *The origin, expansion, and demise of plant species*. (Oxford University Press on Demand, 2000).

14. Coyne, J. A. & Orr, H. A. *Speciation*. vol. 37 (Sinauer Associates Sunderland, MA, 2004).

15. Sleumer, H. O. Die Gattung Escallonia. *Verhandelingen der Koninklijke Nederlandse Akademie van Wetenschappen, Afd. Natuurkunde* 1–149 (1968).

16. Nosil, P., Feder, J. L. & Gompert, Z. How many genetic changes create new species? *Science* **371**, 777–779 (2021).

17. Zapata, F. A multilocus phylogenetic analysis of escallonia (escalloniaceae): Diversification in montane south america. *American Journal of Botany* **100**, 526–545 (2013).

18. Winston, J. E. *Describing species: Practical taxonomic procedure for biologists*. (Columbia University Press, 1999).

19. Bellman, R. Dynamic programming and stochastic control processes. *Information and control* **1**, 228–239 (1958).

20. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction*. (Springer Science & Business Media, 2009).

21. McDade, L. A. Species concepts and problems in practice: Insight from botanical monographs. *Systematic Botany* 606–622 (1995).

22. Stevens, P. F. Botanical systematics 1950-2000: Change, progress, or both? *Taxon* **49**, 635–659 (2000).

23. McLachlan, G. J. & Peel, D. *Finite mixture models.* (John Wiley & Sons, 2004).

24. Dobzhansky, T. *Genetics and the origin of species.* (Columbia university press, 1937).

25. Barraclough, T. G. & Humphreys, A. M. The evolutionary reality of species and higher taxa in plants: A survey of post-modern opinion and evidence. *New Phytologist* **207**, 291–296 (2015).

26. Mallet, J. Hybridization, ecological races and the nature of species: Empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 2971–2986 (2008).

27. Fišer, C., Robinson, C. T. & Malard, F. Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology* **27**, 613–635 (2018).

28. Mallet, J. A species definition for the modern synthesis. *Trends in Ecology & Evolution* **10**, 294–299 (1995).

29. Hausdorf, B. & Hennig, C. Species delimitation using dominant and codominant multilocus markers. *Systematic Biology* **59**, 491–503 (2010).

30. Baum, D. A. & Shaw, K. L. Genealogical perspectives on the species problem. in *Experimental and molecular approaches to plant biosystematics* (ed. Hoch, P. C.) 289–303 (Missouri Botanical Garden Press, 1995).

31. Zhang, J., Kapli, P., Pavlidis, P. & Stamatakis, A. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* **29**, 2869–2876 (2013).

532 32. Yang, Z. & Rannala, B. Bayesian species delimitation using multilocus sequence data.
533 *Proceedings of the National Academy of Sciences* **107**, 9264–9269 (2010).

534 33. Leaché, A. D., Fujita, M. K., Minin, V. N. & Bouckaert, R. R. Species delimitation
535 using genome-wide SNP data. *Systematic biology* **63**, 534–542 (2014).

536 34. Mason, N. A., Fletcher, N. K., Gill, B. A., Funk, W. C. & Zamudio, K. R. Coalescent-
537 based species delimitation is sensitive to geographic sampling and isolation by distance.
Systematics and Biodiversity **18**, 269–280 (2020).

538 35. Cadena, C. D. & Zapata, F. The genomic revolution and species delimitation in birds
539 (and other organisms): Why phenotypes should not be overlooked. *Ortnithology* **138**,
1–18 (2021).

540 36. Baum, D. A. Individuality and the existence of species through time. *Systematic*
541 *Biology* **47**, 641–653 (1998).

542 37. De Queiroz, K. Species concepts and species delimitation. *Systematic biology* **56**,
543 879–886 (2007).

544 38. Filatov, D. A., Osborne, O. G. & Papadopulos, A. S. Demographic history of speciation
in a senecio altitudinal hybrid zone on mt. etna. *Molecular Ecology* **25**, 2467–2481
545 (2016).

546 39. Weir, J. T. & Price, T. D. Limits to speciation inferred from times to secondary
sympatry and ages of hybridizing species along a latitudinal gradient. *The American*
547 *Naturalist* **177**, 462–469 (2011).

548 40. Singhal, S. & Moritz, C. Reproductive isolation between phylogeographic lineages
scales with divergence. *Proceedings of the Royal Society B: Biological Sciences* **280**,
549 20132246 (2013).

41. Struck, T. H. *et al.* Finding evolutionary processes hidden in cryptic species. *Trends in Ecology & Evolution* **33**, 153–163 (2018).

42. Wiens, J. J. Speciation and ecology revisited: Phylogenetic niche conservatism and the origin of species. *Evolution* **58**, 193–197 (2004).

43. Lotsy, J. Species or linneon. *Genetica* **7**, 487–506 (1925).

44. Cronk, Q. C. & Suarez-Gonzalez, A. The role of interspecific hybridization in adaptive potential at range margins. *Molecular Ecology* **27**, 4653–4656 (2018).

45. Novikova, P. Y. *et al.* Sequencing of the genus arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics* **48**, 1077–1082 (2016).

46. Cannon, C. H. & Petit, R. J. The oak syngameon: More than the sum of its parts. *New Phytologist* **226**, 978–983 (2020).

47. Wang, X., He, Z., Shi, S. & Wu, C.-I. Genes and speciation: Is it time to abandon the biological species concept? *National Science Review* **7**, 1387–1397 (2020).

48. Mallet, J., Besansky, N. & Hahn, M. W. How reticulated are species? *BioEssays* **38**, 140–149 (2016).

49. Harrison, R. G. & Larson, E. L. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity* **105**, 795–809 (2014).

50. Rundell, R. J. & Price, T. D. Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends in Ecology & Evolution* **24**, 394–399 (2009).

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

572 52. Barth, J. M. *et al.* Stable species boundaries despite ten million years of hybridization
573 in tropical eels. *Nature Communications* **11**, 1–13 (2020).

574 53. 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
575 model and non-model species. *PloS ONE* **7**, e37135 (2012).

576 54. Bayona-Vásquez, N. J. *et al.* Adapterama III: Quadruple-indexed, double/triple-
577 enzyme RADseq libraries (2RAD/3RAD). *PeerJ* **7**, e7724 (2019).

578 55. Eaton, D. A. & Overcast, I. Ipyrad: Interactive assembly and analysis of RADseq
579 datasets. *Bioinformatics* **36**, 2592–2594 (2020).

580 56. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
581 (2011).

582 57. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Haeseler, A. von & Jermiin, L.
S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature*
583 *Methods* **14**, 587–589 (2017).

584 58. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2:
Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*
585 **35**, 518–522 (2018).

586 59. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic
587 inference in the genomic era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

588 61. Minh, B. Q., Hahn, M. W. & Lanfear, R. New methods to calculate concordance factors for phylogenomic datasets. *Molecular biology and evolution* **37**, 2727–2733 589 (2020).

590 61. Minh, B. Q., Hahn, M. W. & Lanfear, R. New methods to calculate concordance factors for phylogenomic datasets. *Molecular biology and evolution* **37**, 2727–2733 591 (2020).

592 62. Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317–3324 (2014). 593

594 63. Swofford, D. L. PAUP*: Phylogenetic analysis using parsimony (and other methods) version 4.0 beta. (2003). 595

596 64. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020). 597

598 65. Habel, K., Grasman, R., Gramacy, R. B., Mozharovskyi, P. & Sterratt, D. C. *Geometry: Mesh generation and surface tessellation*. (2019). 599

600 66. Templeton, A. R. *Population genetics and microevolutionary theory*. (John Wiley & Sons, 2006). 601

602 67. Croux, C., Filzmoser, P. & Oliveira, M. R. Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **87**, 603 218–225 (2007).

604 68. Raftery, A. E. & Dean, N. Variable selection for model-based clustering. *Journal of the American Statistical Association* **101**, 168–178 (2006). 605

69. Maugis, C., Celeux, G. & Martin-Magniette, M.-L. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis* **53**, 3872–3882 (2009).

70. Fraley, C. & Raftery, A. E. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal* **41**, 578–588 (1998).

71. Biernacki, C., Celeux, G. & Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* **22**, 719–725 (2000).

72. Filzmoser, P., Fritz, H. & Kalcher, K. *pcaPP: Robust PCA by projection pursuit.* (2018).

73. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal* 289–317 (2016).

74. Bowcock, A. M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).

75. Hennig, C. & Hausdorf, B. *Prabclus: Functions for clustering of presence-absence, abundance and multilocus genetic data.* (2019).

76. Kapli, P. *et al.* Multi-rate poisson tree processes for single-locus species delimitation under maximum likelihood and markov chain monte carlo. *Bioinformatics* **33**, 1630–1638 (2017).

77. Revell, L. J. Phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).

78. Rannala, B. & Yang, Z. Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics* **164**, 1645–1656 (2003).

79. Yang, Z., Rannala, B. & Edwards, S. V. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences* **107**, 9264–9269 (2010).

80. Flouri, T., Jiao, X., Rannala, B. & Yang, Z. Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution* **35**, 2585–2593 (2018).

81. Yang, Z. & Rannala, B. Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution* **31**, 3125–3135 (2014).

82. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

83. Verity, R. & Nichols, R. A. Estimating the number of subpopulations (k) in structured populations. *Genetics* **203**, 1827–1839 (2016).

84. Leache, A. D., Fujita, M. K., Minin, V. N. & Bouckaert, R. R. Species Delimitation using Genome-Wide SNP Data. *Systematic Biology* **63**, 534–542 (2014).

85. Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution* **29**, 1917–1932 (2012).

86. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology* **10**, e1003537–6 (2014).

87. Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the american statistical association* **90**, 773–795 (1995).

88. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES science gateway for inference of large phylogenetic trees. in *2010 gateway computing environments workshop (GCE)* 1–8 (Ieee, 2010).

Acknowledgements

We are grateful to Michael Grundler for feedback on an earlier version of this manuscript. We thank Thomas Huggins (LA), James Solomon (MO), and Andrea Voyer (MO) for help with herbarium loans from the following herbaria: CORD, CTES, E, F, GH, GOET, K, L, LIL, MO, NY, RB, REU, RSA, SP, UC, and US; thanks to the collections' managers of those herbaria for granting access to their collections. For help in the lab, we thank Mary Sarkinan and Dana McCarney. For support in the field or providing samples, we thank Barry Hammel, Rosa Oritz, Alfredo Navas, Carmen Ulloa, Pamela Puppo, Efraín Suclli, Luis Valenzuela, Isidoro Sánchez, Angelina Laura, Víctor Quipuscoa, Stephan Beck, Arely Palabral, Félix Huanca, Teresa Ortuño, Silvana Sede, Lone Aagesen, Fernando Zuloaga, Cintia Cornelius, Fernanda Salinas, Pablo Necochea, Alicia Marticorena, Lúcia Lohmann, Susana Alcântara, Luis Henrique Fonseca, and Wesley Pires. We thank the UCLA Institute for Digital Research and Education for use of the research computing infrastructure, specifically the Hoffman2 HPC cluster. This work was supported in part by the National Science Foundation (OISE-0738118), the Whitney R. Harris World Ecology Center, the Federated Garden Club of Missouri, the American Society of Plant Taxonomists, the Garden Club of America, Idea Wild, the University of Missouri–St. Louis, the Missouri Botanical Garden, and the Hellman Fellows Fund (award to F.Z.).

Author contributions

F.Z and S.J.J. conceived this study. F.Z. supervised the project. S.J.J., C.L.H., and F.Z. generated the data and conducted analyses. S.J.J. and F.Z. wrote the paper. All authors discussed the results and implications and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information for this paper is available at https://github.com/zapata-lab/ms_nature_of_species

Data availability

Raw FASTQ reads for this study have been deposited in the SRA under Bioproject accession number TBD. All other data, including raw morphological measurements and intermediate files are available in a pubic repository at: https://github.com/zapata-lab/ms_nature_of_species

Code availability

Code repository is available at: https://github.com/zapata-lab/ms_nature_of_species

Figure Legends

Figure 1 Phylogenetic history, taxon sampling, and evolutionary model-based species delimitation. Maximum Likelihood (ML) tree of *Escallonia* based on genome wide data (left) with tips indicating the six focal clades (Clade I-VI) of our study. For each clade, the first column on the left shows the taxon sampling, with filled symbols indicating specimens used in phenotypic analyses and empty symbols specimens used in genomic analyses; the insets show the distribution of specimens along elevation. The second column to the right shows results of the best fit model for species delimitation with phenotypic data (i.e., phenogroups); phenogroups are shown with different shapes in geographic space. The third column shows results of the best fit model for species delimitation with genomic data (i.e., genogroups); genogroups are indicated with different colors as tips of unrooted ML trees based on matrices of concatenated loci and mapped in geographic space. The fourth column shows the integration of phenogroups and genogroups with evolutionary history and geographic distribution to elucidate the nature of plant species; specimens without overlapping phenotypic and genomic data are designated as unknown specimens.

Figure 2. Integration of phenotypic and genome-wide variation to delimit species. For each clade (See Fig. 1), we assigned specimens to their corresponding phenogroup and genogroup based on the best fit models for each type of data. Shaded cells show specimens assigned to a particular combination of best fit phenogroup and genogroup (i.e., each shaded cell is a species). Three type of species are recognized. First, specimens assigned uniquely to a single phenogroup and a single genogroup are recognized as 'good species' (e.g., phenogroup 4, genogrpup 3 in Clade III). Second, specimens assigned to a single phenogroup across multiple genogroups are recognized as 'phenotypic cryptic species' (e.g., phenogroup 2, genogroups $1, 2$ in Clade III). Third, specimens assigned to a single genogroup across multiple phenogroups are recognized as 'genetic cryptic species' (e.g., phenogroups $1, 3$, genogroup 5, in Clade III). Empty rows or columns correspond to specimens which did

not have overlapping phenotypic and genomic data and thus were assigned only to their corresponding phenogroup or genogroup, accordingly (e.g., genogroup 2 in Clade I).

Tables

Table 1: **Current state of taxonomic species.**

Clade	Taxonomic species	Specimens	Minimum proportion overlap among 10-cubes	Maximum proportion overlap among 10-cubes	Percent specimens matching any 10-cube	Percent specimens matching correct 10-cube
I	2	33	0	0.00	0.0	0.0
II	2	33	0	0.00	0.0	0.0
III	6	130	0	0.02	1.6	0.8
IV	2	74	0	0.00	0.0	0.0
V	7	214	0	0.13	0.0	0.0
VI	10	195	0	0.00	0.0	0.0

Table 2: **Gaussian finite mixture modeling (GFMM) for phenogroup delimitation and model selection using the Bayesian information criterion (BIC)**

Clade	Model	Phenogroups	BIC	Rank	ΔBIC
I	Naive	2	54.03099	1	0.00000
	Taxonomy	2	45.80586	2	8.22513
	Taxonomy unaware	1	33.45654	3	20.57445
II	Naive	3	71.72976	1	0.00000
	Taxonomy unaware	1	47.52785	2	24.20191
	Taxonomy	2	17.77346	3	53.95630
III	Naive	5	387.15280	1	0.00000
	Taxonomy unaware	4	170.83930	2	216.31350
	Taxonomy	6	53.38527	3	333.76753
IV	Taxonomy	2	-115.00390	1	0.00000
	Taxonomy unaware	2	-115.00390	1	0.00000
	Naive	3	-115.89910	2	0.89520
V	Naive	8	-516.72340	1	0.00000
	Taxonomy unaware	4	-648.03900	2	131.31560
	Taxonomy	7	-791.45350	3	274.73010
VI	Naive	8	231.24780	1	0.00000
	Taxonomy unaware	10	200.30380	2	30.94400
	Taxonomy	10	-517.76350	3	749.01130

Table 3: **Genomic modeling for genogroup delimitation and model selection using Bayes factors (BF)**

Clade	Model	Genogroups	Marginal Likelihood (log_e)	Rank	BF (2 x log_e)
I	GC	3	-6580.495	1	
	AC	2	-6754.495	2	348.000
	RI	2	-6754.495	2	348.000
II	AC	4	-13460.917	1	
	GC	3	-15036.438	2	3151.042
	RI[a]	3	-15036.438	2	3151.042
	RI[b]	2	-18963.342	3	11004.850
III	AC	7	-8985.782	1	
	RI[a]	5	-10014.260	2	2056.955
	RI[b]	3	-12233.131	3	6494.698
	GC	3	-12233.131	3	6494.698
IV	AC	6	-9601.514	1	
	GC	3	-11546.649	2	3890.271
	RI[a]	2	-12017.878	3	4832.728
	RI[b]	2	-12017.878	3	4832.728
V	AC	10	-4588.693	1	
	GC	6	-5381.361	2	1585.336
	RI[a]	3	-5601.058	3	2024.730
	RI[b]	2	-6085.998	4	2994.610
VI	AC	11	-2921.024	1	
	GC	7	-3627.806	2	1413.564
	RI[a]	4	-4661.351	3	3480.654
	RI[b]	4	-4661.351	3	3480.654

[a] specimens assigned to demes using MAVERICK

[b] specimens assigned to demes using STRUCTURE

Table 4: **Correspondence between taxonomic species and best-fit phenogroups and genogroups.**

Clade	Taxonomic species	Phenogroups	Perfect match taxonomic species to phenogroups	Genogroups	Perfect match taxonomic species to genogroups	Perfect match taxonomic species to phenogroup and genogroup
I	2	2	2	3	1	1
II	2	3	0	4	1	0
III	6	5	1	7	3	1
IV	2	2	2	6	1	1
V	7	8	0	10	0	0
VI	10	8	2	11	5	2