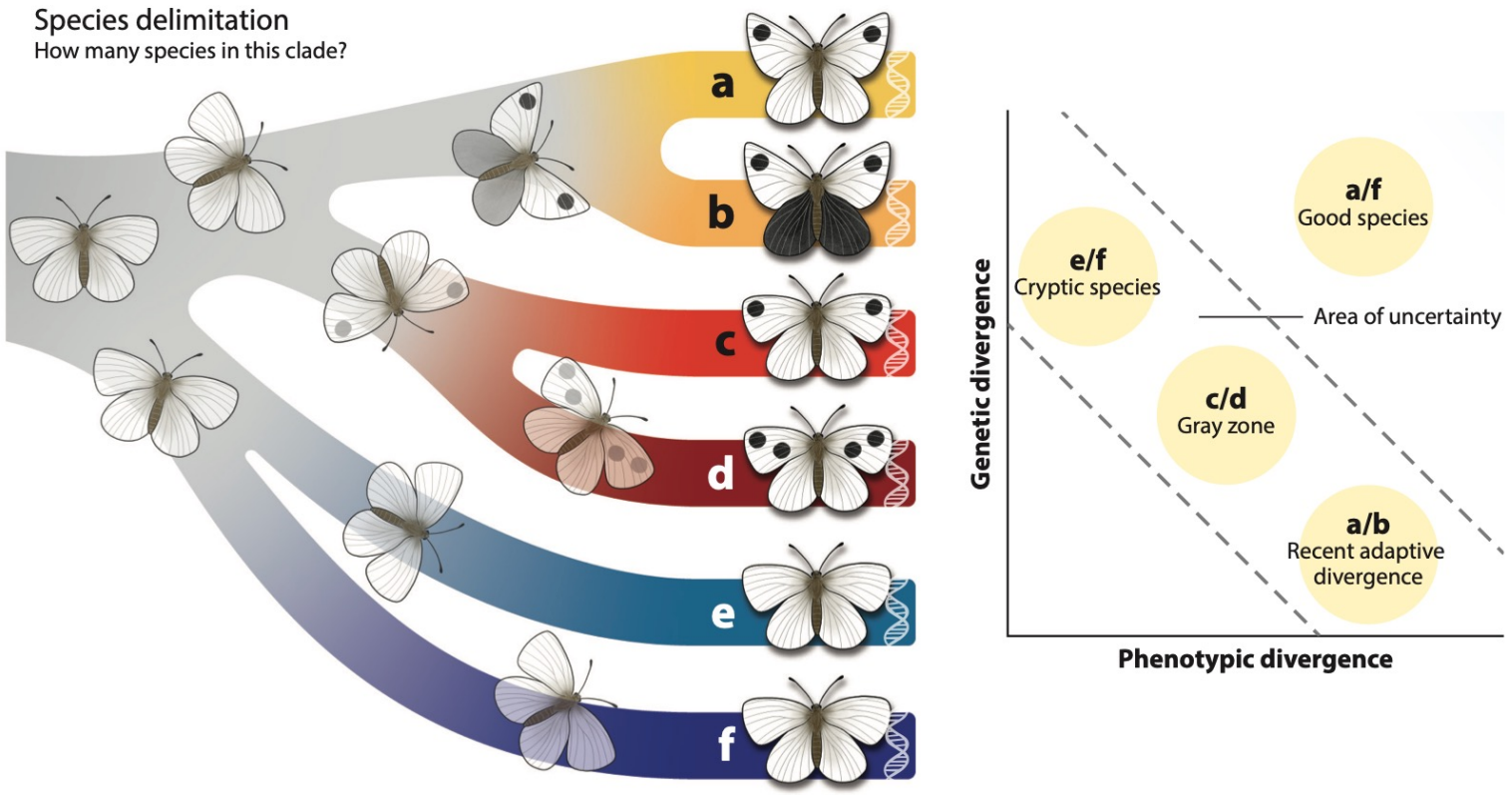


Figure 1

The goal of species delimitation is to define the number of species in a clade and the boundaries between them. Here we focus on how genetic and phenotypic divergence can be used to define species. When species are markedly different genetically and phenotypically (e.g., butterflies *a* vs. *f*), then they are clearly good species. However, when levels of genetic and phenotypic divergence are uncorrelated (e.g., butterflies *e* vs. *f* and *a* vs. *b*) or when divergence is intermediate (e.g., butterflies *c* vs. *d*), species status can be ambiguous.

Sonal Singhal, Adam D. Leaché, Matthew K. Fujita, Carlos Daniel Cadena, and Felipe Zapata. 2025. "A Genomic Perspective on Species Delimitation" Annu. Rev. Ecol. Evol. Syst. 2025. 56:467–89 https://doi.org/10.1146/annurev-ecolsys-102723-055311

Class focus area: Species delimitation

EEB603: Brian O'Meara

All quotes and images from the above paper unless otherwise noted

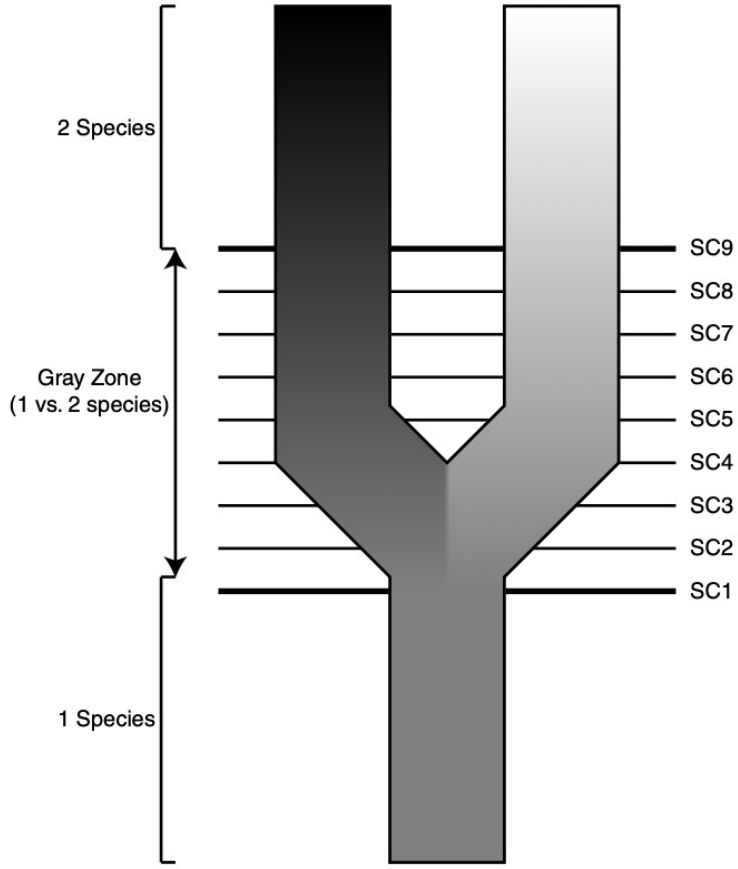


FIGURE 1. Lineage separation and divergence (speciation) and species concepts (after de Queiroz, 1998, 1999, 2005a). This highly simplified diagram represents a single lineage (species) splitting to form two lineages (species). The gradations in shades of gray represent the daughter lineages diverging through time, and the horizontal lines labeled SC (species criterion) 1 to 9 represent the times at which they acquire different properties (i.e., when they become phenetically distinguishable, diagnosable, reciprocally monophyletic, reproductively incompatible, ecologically distinct, etc.). The entire set of properties forms a gray zone within which alternative species concepts come into conflict. On either side of the gray zone, there will be unanimous agreement about the number of species. Before the acquisition of the first property, everyone will agree that there is a single species, and after the acquisition of the last property, everyone will agree that there are two. In between, however, there will be disagreement. The reason is that different contemporary species concepts adopt different properties (represented by the horizontal lines) as their species criteria—that is, as their cutoffs for considering a separately evolving lineage to have become a species.

de Queiroz 2007

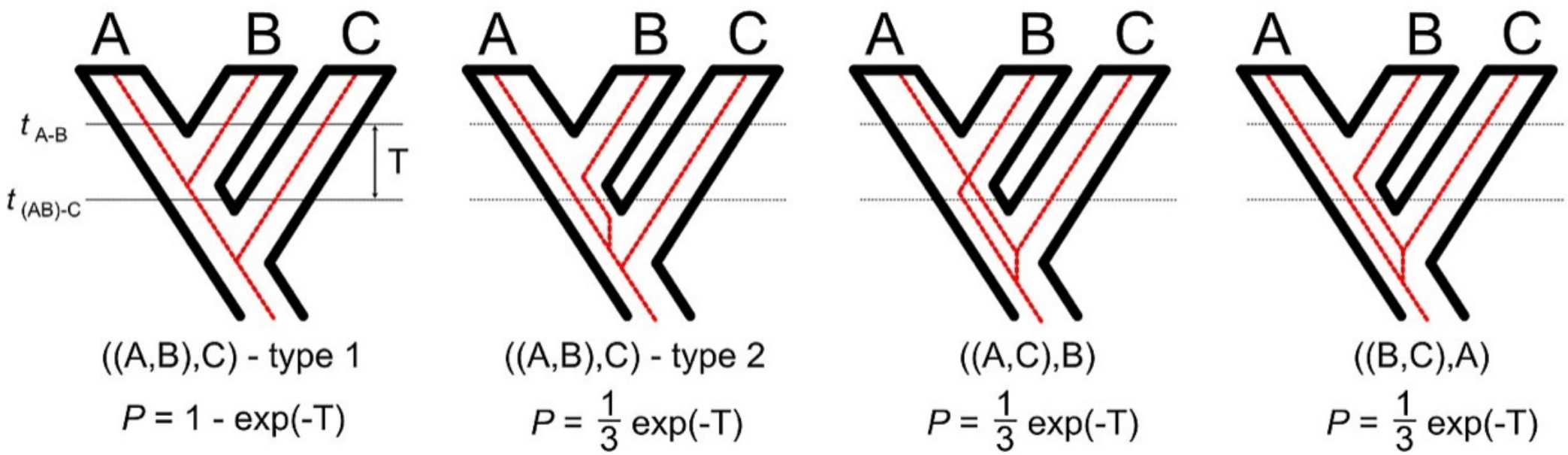
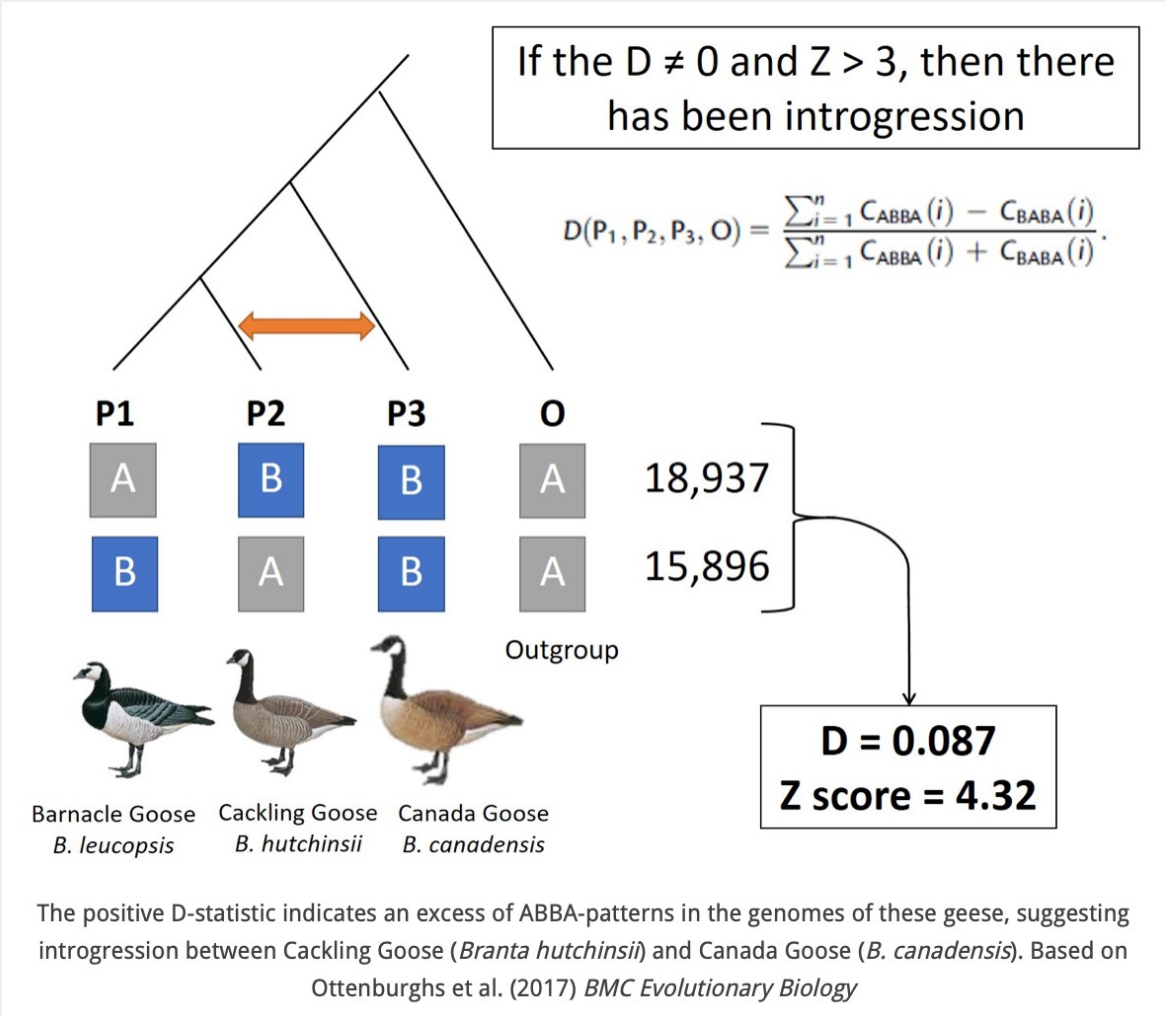


Illustration of the multispecies coalescent showing the relationship between the *species tree* (black outline) and *gene trees* (dashed red lines embedded in the species tree). The time between the two speciation events (T, measured in coalescent units) can be used to calculate the probability of the four possible gene trees (using the equations shown). Note that two of the gene trees are topologically identical but they differ in the times at which lineages coalesce.

Wikipedia Sept 12, 2025: Multispecies Coalescent

The figure below illustrates the D-statistic with an example from my own work (see Ottenburghs et al. (2017) for more details). Comparing the genomes of four goose species reveals that Cackling Goose (*Branta hutchinsii*) and Canada Goose (*B. canadensis*) share more derived alleles than expected by chance. The resulting positive D-statistic suggests introgression between these species, which is not that surprising because there is a hybrid zone between these geese.



The positive D-statistic indicates an excess of ABBA-patterns in the genomes of these geese, suggesting introgression between Cackling Goose (*Branta hutchinsii*) and Canada Goose (*B. canadensis*). Based on Ottenburghs et al. (2017) *BMC Evolutionary Biology*

Jente Ottenburghs: https://avianhybrids.wordpress.com/2019/11/09/d-statistics-for-dummies-a-simple-test-for-introgression/

For these reasons, discrete model selection alone may not always provide the best estimate of lineage independence when both gene flow and genetic drift are modeled. When inferring species limits using PHRAPL, it is thus important to also consider parameter estimates derived from a grid search, as these values are informative of the species boundaries we aim to infer. To facilitate this, we developed the genealogical divergence index (*gdi*), which can be calculated from estimates of migration rate and coalescence time obtained from a PHRAPL analysis. This index provides an estimate of the overall degree of genetic divergence between two taxa due to the combined effects of genetic isolation and gene flow and is useful in the interpretation of the results from model selection and parameter estimation. If one samples two gene copies from species 1 and one gene copy from species 2, then let G_1 be the resulting genealogy in which the two gene copies from species 1 are sister to each other. We define the unscaled GDI_u to be

$$GDI_u = \mathbb{P}(G_1 | M_1, M_2, t)$$

where M_1 and M_2 are bi-directional migration rates because the divergence of the species at time t. Rather than analytically calculating GDI_u, we approximate it using ms (Hudson 2002) such that

$$gdi_u = \text{observed}(GDI_u)$$

For a given M_1, M_2, and t, we iteratively simulate coalescent trees with three gene copies under the two-taxon species tree model described above, and then calculate the proportion of simulated trees in which the two gene copies originating from species 1 are sisters. The index is then scaled to be between 0 (panmixia) and 1 (strong divergence) using

$$gdi = [\text{observed}(gdi_u) - \min(gdi_u)] / [\max(gdi_u) - \min(gdi_u)]$$

where $\min(gdi_u) \approx 1/3$ (with three tips, under panmixia, species 1 monophyly is expected $\sim 1/3$ of the time) and $\max(gdi_u) = 1$ (with extreme isolation, species 1 will always be monophyletic). The *gdi*, along with confidence intervals, can be calculated using the CalculateGdi function within PHRAPL.

The *gdi* is similar to the genealogical sorting index (*gsi*; Cummings et al. 2008) in that it calculates the degree of nonmonophyly in a set of gene trees, and in fact these two indexes are highly correlated ($R^2 = 0.9$) and perform similarly if used to delimit species based on a range of theoretical cutoffs (see Supplementary Figs S3a, b). However, the two indexes differ in two important ways. First, PHRAPL aims to delimit species while simultaneously understanding those aspects of demographic history that have given rise to these groups, and the *gdi* explicitly incorporates these inferred processes (in the form of estimated parameter values). The *gsi*, in contrast to this, measures divergence directly from genetic data, and thus does not presuppose any information about the underlying causes of divergence. Secondly, the *gdi* measures divergence between two focal sister populations or groups, which is the level at which delimitation questions arise. However, the *gsi* measures the exclusivity of a focal taxon relative to the entire tree, and thus the degree of divergence inferred for that taxon can depend on patterns of genetic structure within other parts of the tree (Winter et al. 2016).

Jackson et al. 2017

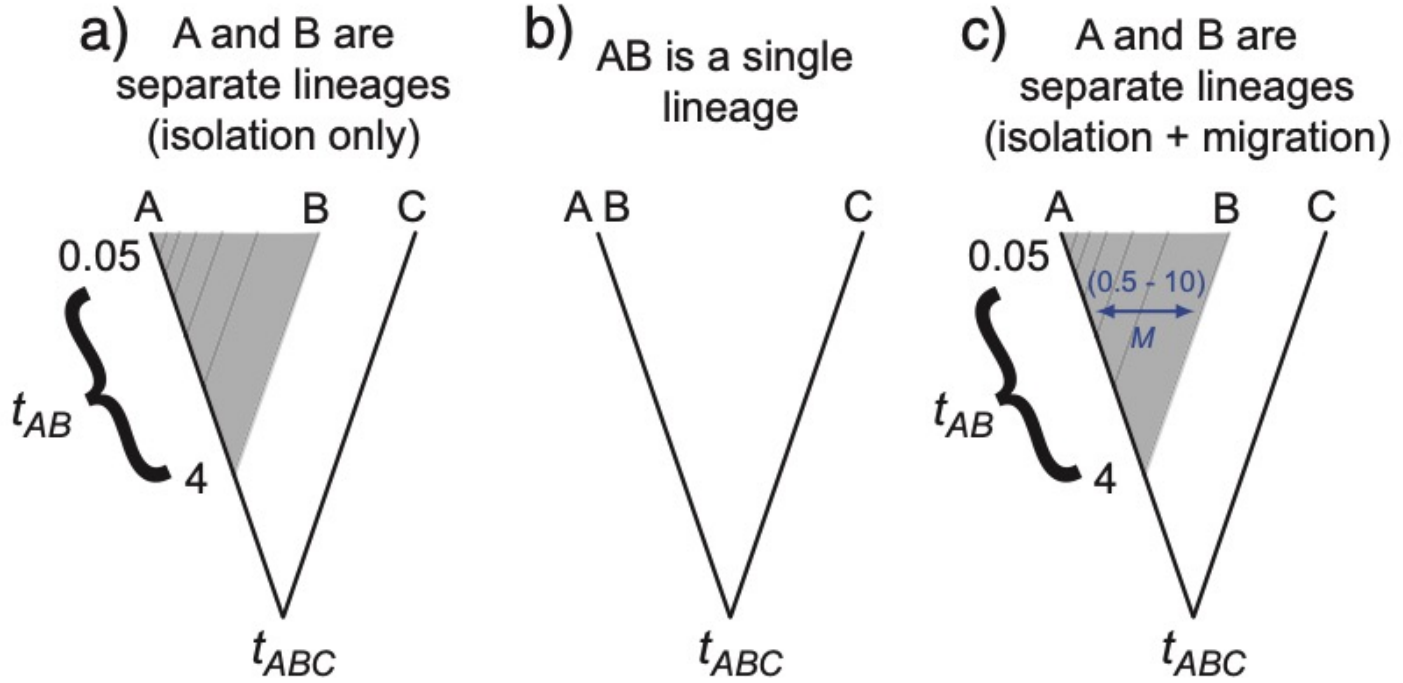


FIGURE 1. Three histories underlying simulated datasets that were analyzed using PHRAPL: a) taxa A and B diverged at time t_{AB} in the past; b) taxa A and B are a single panmictic lineage; and c) taxa A and B diverged at time t_{AB} in the past, but continued to share migrants at rate M. In histories (a) and (c), species coalescent times t_{AB} were varied between 0.05 and 4 (shown in the shaded region), where times t are in units of $4N$. In history (c), M was varied between 0.5 and 10, where $M=4Nm$. Note that the branch lengths preceding t_{ABC} in these trees were adjusted according to the t_{AB}simulated (see text); however, this scaling is not shown here.

Jackson et al. 2017

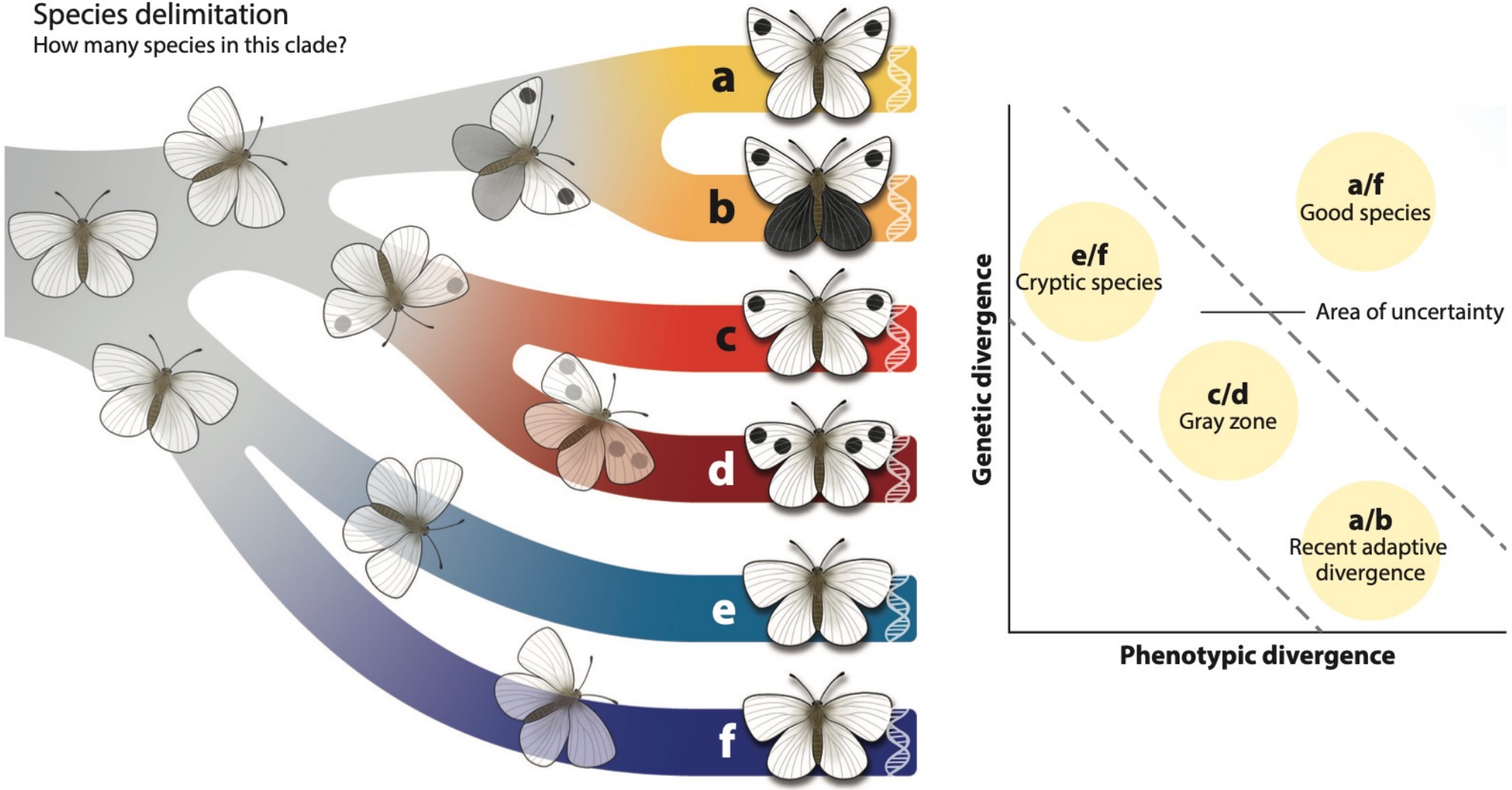
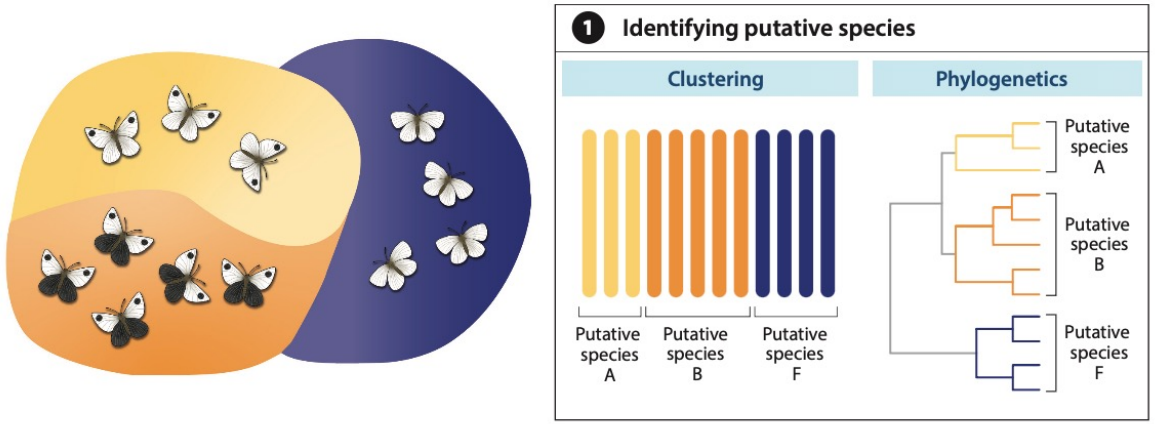


Figure 1

The goal of species delimitation is to define the number of species in a clade and the boundaries between them. Here we focus on how genetic and phenotypic divergence can be used to define species. When species are markedly different genetically and phenotypically (e.g., butterflies *a* vs. *f*), then they are clearly good species. However, when levels of genetic and phenotypic divergence are uncorrelated (e.g., butterflies *e* vs. *f* and *a* vs. *b*) or when divergence is intermediate (e.g., butterflies *c* vs. *d*), species status can be ambiguous.


1 Identifying putative species
Clustering
Phylogenetics
Putative species A
Putative species B
Putative species F
Putative species A
Putative species B
Putative species F
2 Evaluating the validity of putative species
Identifying breaks in isolation by distance
Using the multispecies coalescent
Putative species A vs. B
Putative species B vs. F
Species A
Species B
Species F
Between species
Genetic difference
Geographic distance
<<1.0
1.0
Measuring introgression using geographic clines
Estimating genealogical divergence
Putative species A vs. B
Putative species B vs. F
% Species B
Population
Ambiguous
Species
0
0.2
0.7
1
gdi

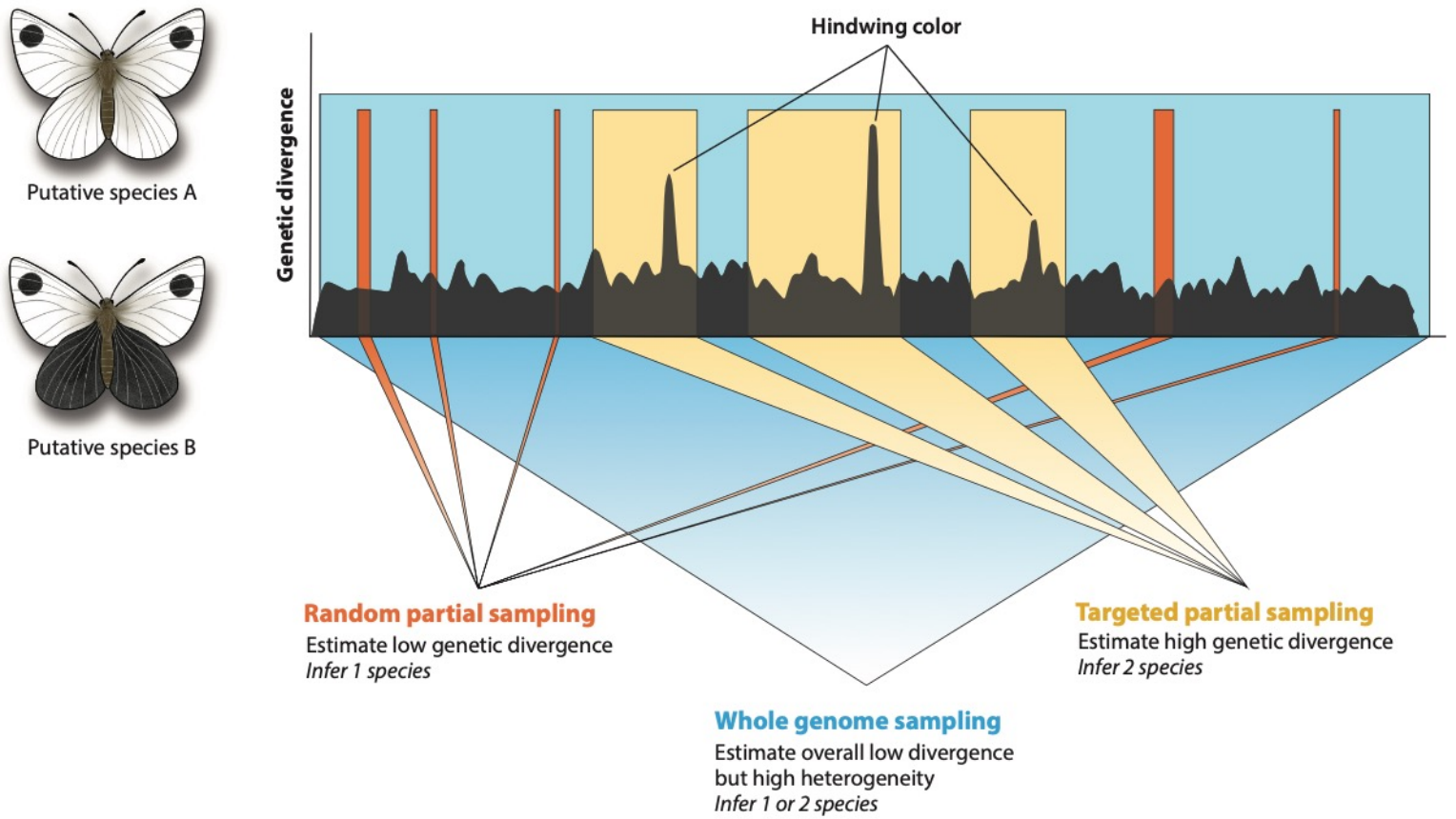


Figure 4

Heterogeneous genome divergence can affect estimates of species limits. In the case of recently diverged, phenotypically distinct species, random partial sampling throughout the genome results in low estimates of genetic divergence, suggesting the two species should be collapsed. If instead we sequence those parts of the genome that underpin phenotypic differences, we infer high genetic divergence, suggesting the two species are distinct. However, if we sequence the whole genome, we uncover high levels of heterogeneity and overall low divergence, and the status of these two species remains ambiguous.

Despite concerns that genetic approaches are overly zealous in splitting species, 36% of studies identified either the same number or fewer species than the initial taxonomy. In those cases where genomic species delimitation increased the number of species, the median increase in species number was twofold. In many cases, the genomic species delimitation aligned closely with species delimitations inferred from smaller mitochondrial or multilocus datasets, suggesting that smaller genetic datasets can delimit species as accurately as genomic datasets. Many studies refrained from revising taxonomy, commenting that they would enact taxonomic change after they further validated findings through additional sampling or phenotypic data. In the end, only 36% of the studies proposed formal taxonomic changes; notably, studies that revised taxonomy were 2.2 times more likely to include phenotypic data.

While the costs of sequencing have dropped six orders of magnitude since 2000, collecting genomic data remains prohibitively expensive and inaccessible for many researchers. Further, genomic datasets require access to high-speed networks to download and share the data, multicore computers with high memory footprints and reliable power sources to process the data, and redundant hard drive storage to preserve the data long-term (but see Handika & Esselstyn 2024). These resources are not universally available, nor is the training to collect and analyze these data. Expecting or requiring genomic data for species-delimitation studies could inadvertently exacerbate existing inequities in the field and further bias taxonomic shortfalls (Linck & Cadena 2024). Indeed, the regions predicted to have the greatest shortfalls (Freeman & Pennell 2021) are some of the same regions with less developed research infrastructure (**Figure 3*a***). We thus caution against setting genomic data as the gold standard for the field; many of the most pressing, practical concerns of species delimitation can be effectively and efficiently addressed using smaller datasets (Bertola et al. 2024).

FUTURE ISSUES

1. What other types of genomic data could we collect to aid in genomic species delimitation?
2. How can genomic species delimitation methods include other sources of data—e.g., morphological differentiation, reproductive barriers—to create an integrative delimitation approach?
3. How can the extent of heterogeneity in genomic divergence be used as a metric for capturing species divergence, and thus, in delimiting species?
4. How can genomic species delimitation help address the taxonomic impediment?
5. How can genomic species delimitation help unite the fields of taxonomy, systematics, and speciation biology?
6. How can artificial intelligence help delimit species more efficiently and effectively?