



## Investigations

# The Evidence for New Species Across the Tree of Life: Morphology Still Rules the Largest Kingdoms

Adam J. Ziegler<sup>1</sup>, Xin Li<sup>2</sup>, John Wiens<sup>3</sup>

<sup>1</sup> Genetics Program, University of Arizona, <sup>2</sup> Institute of Applied Biology, Shanxi University, <sup>3</sup> Department of Ecology and Evolutionary Biology, University of Arizona

Keywords: biodiversity, cryptic species, species delimitation, taxonomy

<https://doi.org/10.18061/bssb.v4i1.10466>

---

## Bulletin of the Society of Systematic Biologists

---

### Abstract

Over the past ~20 years, several forces have converged to potentially create a seismic shift in how new species are described. These forces include: (1) pleas for DNA-based taxonomy, (2) large-scale genomic datasets for species delimitation, (3) new statistical methods for molecular species delimitation, (4) the discovery of hundreds of cryptic species hiding within morphology-based species, (5) the possibility that most morphologically distinct species are already described, and (6) the putative decline of morphology-based taxonomy. But has a major shift towards molecular-based taxonomy actually happened? Here, we examined newly described species from 9 major groups across the Tree of Life and the evidence used to delimit them. We found five major results. First, in the largest groups, most new species were still described based on morphological data alone, including arthropods, mollusks, and plants (groups collectively including ~90% of all known species). Second, in other groups, most new species were described based on both molecular and morphological evidence, including chordates, fungi, bacteria, and archaeans. Third, species described based only on molecular data remain rare. Fourth, within animals, the majority of species descriptions that incorporated molecular data included only mitochondrial sequences. Fifth, molecular data were typically used to build a tree and generate genetic distances, rather than being used for statistical delimitation methods. Our results suggest that many new developments in species delimitation are underutilized by taxonomists (e.g. genomics), likely because these developments do not offer the fastest way to describe new species before they become extinct. Our results also suggest that many morphologically distinct species (and cryptic species) remain to be described.

## 1 Introduction

Biologists are in a race to discover and describe Earth's species before they are lost forever. For many years now, potential changes have been developing that could revolutionize how most new species are discovered and described, relative to traditional, morphology-based taxonomy. First, several authors have advocated for DNA-based taxonomy as a way to accelerate the pace of species discovery and description (Blaxter, 2004; Tautz et al., 2003). Second, new genomic tools now allow researchers to obtain and analyze thousands of markers to delimit species (Leaché et al., 2014). Third, statistical methods have been developed that can analyze these molecular markers to estimate how many species are present in a given sample of individuals (e.g. Jones et al., 2015; Leaché et al., 2014; Pons et al., 2006; Puillandre et al., 2012; Smith & Carstens, 2020; Yang &

Rannala, 2010). Fourth, many molecular analyses within species suggest that morphological data might often underestimate the number of species present, including studies in animals (e.g. Adams et al., 2014; Bickford et al., 2007; Cahill et al., 2023; Pérez-Ponce de León & Poulin, 2016; Pfenninger & Schwenk, 2007) and plants (e.g. Ji et al., 2020; Kinosian et al., 2020). For example, an analysis across insects suggested that each species initially delimited by morphological data might hide (on average) three cryptic species (Li & Wiens, 2023). Fifth, some authors suggested (>10 years ago) that most new species across life had already been described (Costello et al., 2012; Costello, Wilson, et al., 2013), at least among those species distinguishable based on morphological evidence. Finally, there is the perception that traditional morphology-based taxonomy is disappearing (Löbl et al., 2023; Orr et al., 2020). Given these six factors, one might reasonably expect that many (if not



most) new species are now delimited and described based on molecular data, including many species that are morphologically cryptic.

On the other hand, many new species might continue to be based on morphological data instead. For example, many independent studies have agreed that there are ~4–5 million undescribed insect species (review in Stork, 2018), without including cryptic species. These would be in addition to the ~1 million insect species that are already described, with insects making up almost half of all 2.2 million known species across life (Bánki et al., 2025). Thus, these projections suggest that the most species-rich group of organisms will not be running out of new morphologically distinct species to describe in the near future. Furthermore, describing new cryptic species may not be as straightforward as describing new species based on morphological data. Discovering cryptic species may often require large-scale sampling and molecular analyses across the geographic range of a described, morphology-based species. Such analyses may not be feasible for many species, for a variety of reasons (e.g. time, expense, species distributed across many countries with variable access). There may also be pressure against describing species that are not morphologically diagnosable (Cook et al., 2010). Note that here and throughout, we use the word “morphology” to mean phenotypic variables related to form, including size, shape, color, and the number of specific features (e.g. limbs, scales, color patches), but not necessarily ecology, behavior, or geography.

We speculate that the factors that determine the data used in species delimitation might also vary among clades. For example, some clades might be sufficiently well studied taxonomically that they are running out of new morphologically distinct species, whereas in others, there might still be large numbers of new, morphologically distinct species to describe (e.g. birds vs. amphibians; Moura & Jetz, 2021). Moreover, in some groups, morphology alone might be inadequate to reliably distinguish species (e.g. some microscopic organisms).

What is needed is to quantitatively evaluate the evidence that is currently used to discover and diagnose species across major groups of organisms. Few previous studies have attempted this, although there have been valuable studies within specific groups, such as amphibians (Streich et al., 2020) and reptiles (Guedes et al., 2024). An invaluable study by Miralles et al. (2020) examined major groups of eukaryotes and documented whether molecular data were used in new species descriptions from 1990 to 2018. They found that from 2012–2018, molecular data were used infrequently in insects and plants (<50% of new species) and more frequently (>50%) in protists, fungi, and vertebrates. However, they did not address whether studies that used molecular data used only molecular data or a combination of molecular and morphological data. Therefore, they did not address how many newly described species might be morphologically cryptic (or at least based only on molecular data). Furthermore, for studies that incorporated molecular data, they did not address the types of markers used (e.g. mitochondrial vs. nuclear sequences)

or the methods used for species delimitation. More broadly, they excluded two of the three domains of life (missing Bacteria and Archaea). These are areas that remain in need of further study.

Here, we take a snapshot of recent species descriptions of living taxa to address the evidence used to reveal new species across the Tree of Life. We compile information on new species descriptions and randomly sample a set number of newly described species from each major group (bacteria, archaeans, plants, fungi, protists, and the three most species-rich animal phyla [Arthropoda, Chordata, Mollusca]). For each selected species, we then review the original description to determine what evidence (i.e. morphological, molecular) was used to delimit the species and conclude that it was new, and the type of markers and delimitation methods used for molecular data. Using these data, we compare the evidence used to discover new species across groups.

## 2 Methods

### 2.1 Species Selection

We first generated a list of newly described species for the year 2020 from the Catalogue of Life (CoL; Bánki et al., 2023). We initially focused on 2020 because preliminary analyses suggested that there can be delays in the addition of newly described species to the CoL. Based on these preliminary analyses, 2020 appeared to be the most recent year with relatively complete data. We did not expect taxonomic practices to vary extensively between 2020 and 2025. We describe how we obtained the list of new species from 2020 in Supplementary Appendix S1. These data (for 2020) were obtained on 15 August 2024.

We focused on sampling new species from nine major groups. These included the three most species-rich phyla of animals (Arthropoda, Chordata, Mollusca; Datasets S1–S3, respectively; all datasets available at: <https://figshare.com/s/2d5cff510d0a493a7ead>; Ziegler et al., 2025) and six major groups ranked as kingdoms by the CoL (Archaea, Bacteria, Chromista, Fungi, Plantae, Protozoa; Datasets S4–S9, respectively). We emphasized sampling major groups of animals because animals make up ~73% of all described species across life (1.6 of 2.2 million species), whereas arthropods make up 75% of animals (1.2 million), and more than half of all species across life (Bánki et al., 2025). We excluded viruses because they are often not considered to be alive (Moreira & López-García, 2009).

We generated a list of newly described species for each major group (Supplementary Appendix S1; at the end of this file). We then randomly sampled 50 species from each group. Species were randomly selected by listing the species alphabetically, assigning numbers to each species, and then randomly selecting a number from a random number generator ([random.org](https://random.org)). Upon examination, some taxa represented new combinations of genus and species names rather than newly described species. In these cases, another species was randomly selected until the desired number of new species was reached.

For some groups, it was necessary to modify our protocols somewhat. Chromista and Protozoa had a limited number of novel species described for 2020 (16 and 9, respectively). For these two groups we used all new species from 2020 and then randomly selected new species from 2015 to 2019 until a total of 50 new species was reached. Moreover, for Chordata all new species listed for 2020 in the CoL belonged to Squamata. Therefore, we included new chordate species described from 2015 to 2020 and we randomly sampled 50 species from among them.

For Bacteria and Archaea, there were relatively few newly described species in the CoL from 2015 to 2020. Instead, we used data from the Global Biodiversity Information Facility (GBIF, 2024) to generate lists of new species for each group (Supplementary Appendix S2; at the end of this file). For Archaea we randomly selected 50 species described from among 75 new species described from 2015–2020. We also obtained a list of 242 “new” bacterial species from 2020 from GBIF (2024). However, most represented new combinations of genus and species names (not new species) and one was a fossil taxon. This left 35 new bacterial species from 2020. We then randomly selected 15 species from 2019 to make up the full set of 50 species.

Once we obtained a set of 50 unique species for each group, we examined the original species description for each species. Based on these descriptions, we determined whether the new species was delimited from other (described) species based on morphological data, molecular data, or both. We note that some studies mentioned other types of phenotypic variables in their species descriptions besides morphology (e.g. physiology and biochemistry in bacteria and archaeans) but across most groups, most taxonomic studies focused on molecular and/or morphological markers.

We also examined the type of molecular data that were used. Specifically, we determined whether nuclear, mitochondrial, and/or chloroplast sequences were used. For bacteria and archaeans, we treated their molecular data as nuclear (although they lack a nucleus and plastids). We were primarily interested in how often new animal species were delimited based on mitochondrial data alone (which has long been controversial; Ballard & Whitlock, 2004; Galtier et al., 2009; Rubinoff & Holland, 2005; Yuan et al., 2023; Zink & Barrowclough, 2008). Similarly, we were interested in how often new plant species were inferred using chloroplast data alone, given the potential problem of chloroplast capture and incongruence between chloroplast and nuclear genomes (e.g. Acosta & Premoli, 2010; Liu et al., 2020; Rieseberg & Soltis, 1991).

For species delimited using molecular data (either alone or with morphological data), we also determined what species delimitation method was used (if any). In many cases, the authors used molecular phylogenies as part of the species description but did not describe a specific method for species delimitation. This does not mean that the molecular data were not important: for example, the phylogeny can convincingly show that newly found populations are not closely related to putatively conspecific populations of a morphologically similar described species,

without utilizing a specific species delimitation method. Many authors used various measures of genetic distance and identity to inform their decisions about species limits (e.g. DNA-DNA hybridization, DNA G+C content, genome-to-genome distance analysis, pairwise average nucleotide identity, pairwise differences among polymorphic sites, corrected and uncorrected  $p$ -distances). We summarized these simply as genetic distances.

Many studies used both molecular and morphological data to delimit species. In these cases, it was generally unclear which data were first used to identify the species as potentially distinct or which were considered more important in determining species status. Therefore, we merely noted that both were used.

We also briefly summarized the types of data and methods used for morphological characters. Most studies did not describe a specific methodology for how they used the morphological data to determine that an undescribed species was present. However, they generally focused on morphological characters that could be used to distinguish among species. We recorded how often these characters were quantitative (e.g. measurements, counts), were qualitative (e.g. presence/absence, texture, color), whether they were analyzed statistically (specifically, differences between species), and whether there was a phylogenetic analysis of the morphological data included.

## 2.2 Statistical Analyses

We tested whether different groups showed significant differences in their proportions of sampled species that were described based on molecular data, morphological data, or both. We used  $\chi^2$  tests to evaluate significant differences in the proportions of species, testing the hypothesis that different groups had identical frequencies of each data type among species. We used the function “chisq.test” in the *stats* library of R version 4.4.1 (R Core Team, 2024). We used “simulate.p.value” set to “true” with 10,000 replicates and the seed set to 6000. Because Archaea and Bacteria had identical frequencies, we did not perform separate tests for each. This resulted in 26 independent tests. An analysis of false discovery rates (FDR) was performed on the 26 tests using the function “p.adjust” in *stats*, with the method set to “BH” (Benjamini & Hochberg, 1995). All  $p$ -values reported are adjusted. Based on initial analyses, only groups that differed by 15% or more in the use of morphological data (Table 1) were potentially significantly different, and so only those were tested. All R code used is provided in Dataset S10.

We recognize that 50 species might be considered a limited sample size. Therefore, we performed resampling analyses to evaluate how often a random sample of 10 species (from the full set of 50) supported the conclusions from the full dataset of 50 species (regarding which data type was used most frequently within the group). We assumed that if sampling 10 species is a reasonable approximation of the results from 50 species, then our sample of 50 species should approximate the results from the hundreds or thousands of species described from the larger groups each year. We generated 20 replicates of 10 species

each and evaluated how often the most frequent data type for each group (morphological, molecular, or both) in each replicate differed from the majority data type for the full set of 50 species (Dataset S11). We did not perform statistical analyses on the subsampled data because we assumed a sample size of 10 was generally too small to generate significant results.

We also assessed whether our sampling of 50 species in each group was representative of the full sample of new species in that group. Therefore, we compared the distribution of the 50 species among higher taxa in each group to the distribution of all new species among these higher taxa. For example, for arthropods we compared the distribution of the new species among those classes with new species in 2020 to the distribution of the 50 sampled species among these same classes. We performed least-squares regression between the proportion of all new species in each class (independent variable) and the proportion of the 50 new species in each class (dependent variable). We assigned a value of 0 if no new species were sampled in the set of 50. If the sampling of 50 species broadly reflected the overall pool of new species in a group, then we expected strong, positive relationships. Alternatively, if our sampling of 50 species was biased within a group, these relationships should not be strong or significant. We performed these analyses among arthropod classes ( $n = 11$ ), chordate classes ( $n = 14$ ), mollusk classes ( $n = 7$ ), and phyla of fungi ( $n = 9$ ). For plants, there were only four phyla, so we used classes instead ( $n = 11$ ). For Archaea there were only 3 phyla, so we again used classes ( $n = 8$ ). For Bacteria, Chromista, or Protozoa, our sampling of 50 species represented a mixture of species from 2020 and a random sample of species from earlier years. For Bacteria, most species (70%) were from 2020 and the rest were from 2019, and so we did not compare our sampling to earlier years. For Chromista and Protozoa the majority of new species were randomly sampled from earlier years (68% and 82%, respectively). We therefore included these two groups, but we expected weaker relationships between the sample of 50 species and the distribution of all new species within these groups. Prior to performing these analyses, we eliminated potential new combinations (indicated with parentheses around the author names) from the overall list of new species in each group, since new combinations were not treated as new species.

Finally, for the most species rich group overall (arthropods) we compared our initial results to another random sample of 50 species, for all of our main questions. We expected that if a sample of 50 species was adequate, then another random sample of 50 species (with no overlap in species) should give very similar results.

### 3 Results

We found dramatic differences among groups in the percentage of new species delimited based on morphological versus molecular data (Fig. 1; Table 1). In animals, most (84%) new species of arthropods were delimited based on morphological data alone, whereas other new arthropods

were delimited based on both morphological and molecular data (16%). Mollusks showed a similar pattern, with most (68%) new species based only on morphological data, and some new species based on both types of data (32%). By contrast, most new chordate species (68%) were delimited based on both morphological and molecular data (Table 1). New plant species were also based primarily on morphological data alone (86%). By contrast, most new fungus species were based on both molecular and morphological data (86%), as were most new Chromista (68%). All new archaeans and bacteria sampled were also based on both molecular and morphological data (100%). In Protozoa, half of the new species were based on molecular and morphological data (50%) and almost half were based on morphological data alone (48%). We found that most of these differences among groups were statistically significant (Table 2).

Across all groups, the number of new species that were delimited based on molecular data alone was very limited (Table 1; Fig. 1). No groups had >2% of their new species based on molecular data alone, and these were only in fungi and protozoa.

We used subsampling to address the robustness of these conclusions to finite sample sizes (Table 3). For each major group we randomly sampled 10 species from the full set of 50 species and repeated this for a total of 20 replicates. For arthropods, fungi, and plants, the overall pattern (i.e., majority data type) estimated from the full set of 50 species was recovered in 100% of the replicates. Furthermore, for bacteria and archaeans, the results were the same across the full set of 50 species for each group, guaranteeing that subsampling yields the same result as the full sample. For chordates, mollusks, and chromistans, the majority pattern was recovered in 75–90% of replicates: these are groups in which the majority pattern was more ambiguous (68%). For protozoans, the majority pattern was recovered in only 50% of replicates but in this group 48% of the species were morphology-based and 50% were based on both data types (i.e. frequencies were almost identical and so the inferred majority is expected to be highly sensitive to sampling). Overall, these results suggest that it should generally be possible to estimate the most frequent data type for a group with a limited sample size (i.e. 10 species), especially when the frequency of that data type is high (>80% as in arthropods, fungi, plants, bacteria, and archaeans). We also found that our sampling of species within each group appeared to be unbiased (see below).

We then examined the type of molecular data that was used most frequently for species delimitation in each group (Table 4). Among animal phyla, mitochondrial data alone were used most frequently (Arthropoda = 62%, Chordata = 59%, Mollusca = 69%), followed by studies that used both nuclear and mitochondrial data (38%, 41%, 31%, respectively). In plants, the majority of molecular studies used chloroplast data alone (57%) whereas others used both chloroplast and nuclear data (43%). In fungi, most studies used nuclear data alone (80%), and the remainder used both nuclear and mitochondrial data (20%). In Chromista, most studies used both nuclear and mitochondrial data (56%), and in Protozoa, most used nuclear data alone (69%). All



studies of bacteria and archaeans used “nuclear” data. Generally, these studies each examined only a limited number of markers, especially those that only included mitochondrial or chloroplast DNA.

For those studies that used molecular data, we also recorded the most commonly used species delimitation method in each group (Table 5). We found that in most groups, multiple approaches were often used for each species. Across most groups, the most frequently used approach involved utilizing a phylogeny, but with no specific species delimitation method (>90% in all groups except animals, 44–88% among the three animal phyla). This approach was often accompanied by genetic distance analyses to compare the divergence of the new species to described species (Table 5), especially in chordates, archaeans, and bacteria. One exception to the overall pattern was Mollusca, in which 75% of the relevant studies used genetic distances and 50% used the ABGD method (Automatic Barcode Gap Discovery; Puillandre et al., 2012).

For those studies that used morphological data, we recorded the broad types of morphological data and analyses used (Table 6). We found that almost every study in every group considered both quantitative and qualitative traits. However, despite the widespread inclusion of quantitative characters, very few studies performed statistical comparisons to show that their new species was significantly different from others. Studies that included statistical comparisons were most common in chordates (26%) and plants (6%). Studies that included a phylogenetic analysis of their morphological data were even more rare (<3% across all groups).

We tested whether our sampling of species within each group reflected the distribution of new species among higher taxa within that group or was instead biased by limited sampling. Within each group, we found strong, significant, positive relationships between our sampling and the distribution of all new species ( $r^2 = 0.95$ – $1.00$  for most groups; Table 7), suggesting that our sampling was broadly representative and not taxonomically biased within each group. Relationships were somewhat weaker in Chromista and Protozoa ( $r^2 = 0.75$ – $0.79$ ) but these were the two groups in which the set of 50 species was based only partially on random sampling (i.e. we used all new species from 2020 and then random sampling from 2015–2019 for the rest).

Finally, we compared our initial results for arthropods to those from a second random sample of 50 species (Dataset S15). We found that the major results were very similar including the percentage of morphology-only studies (initial = 84% vs. second = 80%), the percentage of studies incorporating both molecular and morphological data (16% vs. 20%), and studies with molecular data using mitochondrial data alone (62% vs. 60%), and those using combined mitochondrial and nuclear data (38% vs. 40%). We again found that the most widely used molecular approach for molecular species delimitation was a phylogeny with no specific delimitation method (88% vs. 90%) and genetic distances (38% vs. 50%), with several studies using both of these approaches in both the first set (38%) and the second set (40%). There were no studies using specific mol-

ecular methods for species delimitation in the second set for arthropods (and only one in the first). There were also strong positive relationships between the distribution of species among arthropod classes for all new species and for those in the second sample ( $r^2 = 0.98$ ,  $P < 0.0001$ ), and between the first and second samples ( $r^2 = 0.99$ ,  $P < 0.0001$ ). Overall, the second set of data from arthropods reinforced our overall conclusions from the initial analysis and suggested that sampling 50 species should adequately represent other groups as well.

## 4 Discussion

Systematists are in a race to describe all of Earth’s species before they are erased by human impacts (Costello, May, et al., 2013). Given this race, there have been pleas to accelerate the pace of species discovery and description by using molecular-based species delimitation and taxonomy, along with other approaches (Blaxter, 2004; Godfray, 2002; Maddison et al., 2012; Tautz et al., 2003). Large-scale genomic datasets have been developed for application to species delimitation, along with sophisticated methods for estimating species limits from molecular data. There has also been the molecular-based discovery of hundreds of cryptic species hiding within morphology-based species (Bickford et al., 2007; Kinoshita et al., 2020; Li & Wiens, 2023; Pérez-Ponce de León & Poulin, 2016). Further, some authors have suggested that most morphologically distinct species have already been described (Costello, Wilson, et al., 2013). Given these developments, we found a surprising result: in the largest groups of animals (arthropods) and in plants (which together include most known species; 2.0 of 2.2 million, 91%; Bánki et al., 2025), most species continue to be delimited and described based on morphological data alone. Thus, a molecular revolution in taxonomy has failed to materialize. A similar pattern was also found by Miralles et al. (2020) for major eukaryotic groups, with morphology still being frequently used in descriptions of new plants and insects.

Why do so many new species continue to be delimited based only on morphology? We suggest two main reasons. First, because there are still many morphologically distinct species to be described. For example, in the two largest groups (arthropods, plants), most new species were described based on morphology, and there were still thousands of new species described in 2020 (arthropods, Dataset S1; plants, Dataset S6). Moreover, in other groups, most new species are delimited based on both morphological and molecular evidence, not molecular evidence alone. If Earth were running out of morphologically distinct species, we would anticipate that most new species described would be morphologically cryptic.

The second reason is that molecular datasets may not offer the fastest and cheapest way to describe large numbers of new species. Many morphology-based species descriptions are only a page or two long. By contrast, species delimitation analyses based on thousands of loci and statistical species delimitation methods may take years to complete and large amounts of money. The latter analyses

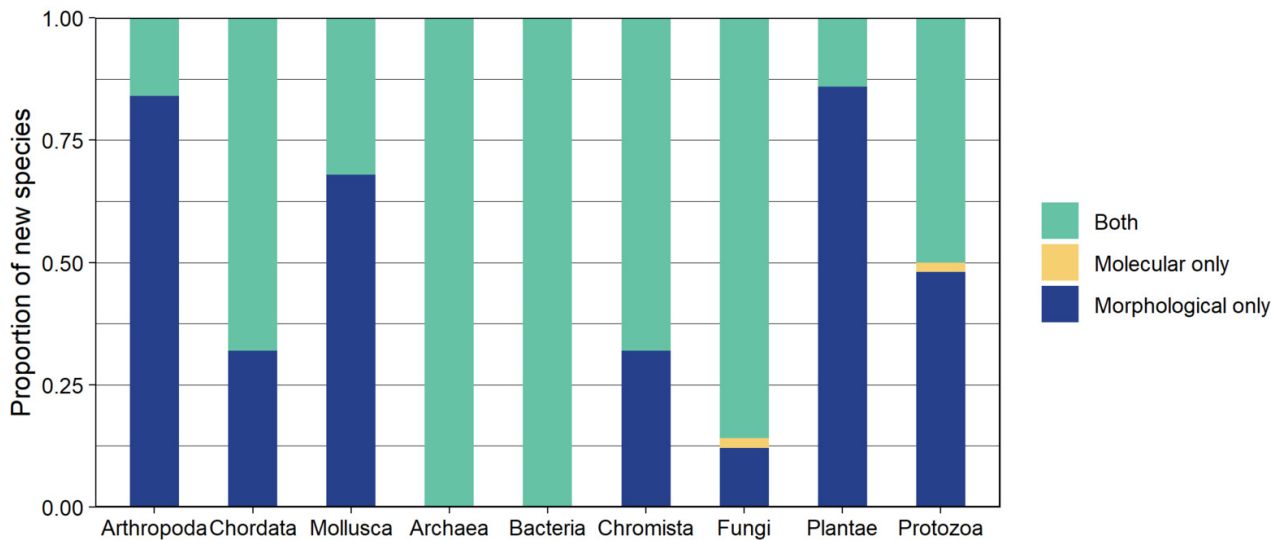


Figure 1. The proportion of species delimited based on morphological, molecular, or both data types in each major group of living organisms. Data are also summarized in [Table 1](#).

Table 1. The proportion of new species delimited with morphological data, molecular data, or both data types across major groups. For each major group we give the number of species (out of 50) and the percentage in each data-type category.

Group	Morphological only	Molecular only	Both
Animals			
Arthropoda	42 (84%)	0 (0%)	8 (16%)
Chordata	16 (32%)	0 (0%)	34 (68%)
Mollusca	34 (68%)	0 (0%)	16 (32%)
Archaea	0 (0%)	0 (0%)	50 (100%)
Bacteria	0 (0%)	0 (0%)	50 (100%)
Chromista	16 (32%)	0 (0%)	34 (68%)
Fungi	6 (12%)	1 (2%)	43 (86%)
Plantae	43 (86%)	0 (0%)	7 (14%)
Protozoa	24 (48%)	1 (2%)	25 (50%)

might be more accurate (e.g. new species inferred from thousands of markers may be less likely to be synonymized than those based on just a few markers, regardless of whether the markers are molecular or morphological). However, Earth is in an extinction crisis (Díaz et al., 2019; Pimm et al., 2014), and there is an urgent need to rapidly discover and describe new species before they are lost forever. Therefore, an important challenge for the study of species delimitation is to not only make new approaches that require more and more time and money. Instead, more approaches are needed that can rival morphological data for being rapid and inexpensive (along with being accurate). DNA barcoding is potentially one such approach (Hebert et al., 2003, 2004). However, DNA barcoding is traditionally based on a short segment of mitochondrial data alone, and its accuracy has been repeatedly questioned (Hickerson et al., 2006; Meier et al., 2006; Will et al., 2005). The problem of obtaining high-quality molecular data from suboptimally

preserved morphology-based type material might also be a limiting factor for a fully molecular taxonomy (although this might not be necessary in many cases).

In a similar vein, we also found that many studies did not use multi-locus genomic data ([Table 4](#)), nor did they use explicit molecular methods designed for species delimitation ([Table 5](#)). We found that when studies did incorporate molecular data ([Table 4](#)), the majority of studies in animals used only mitochondrial data (which are often considered controversial; Ballard & Whitlock, 2004; Galtier et al., 2009; Rubinoff & Holland, 2005; Zink & Barrowclough, 2008). Similarly, in plants, the majority of studies including molecular data used only chloroplast data ([Table 4](#)): this is potentially problematic because of the problem of plastid capture and incongruence between the chloroplast and nuclear genomes (e.g. Acosta & Premoli, 2010; Liu et al., 2020; Rieseberg & Soltis, 1991). Studies of fungi tended to use multi-locus nuclear data. Moreover, across all groups,

Table 2. Chi-squared analyses testing for differences among groups in the frequencies of different data types used in new species descriptions. Chi-squared values are shown below the diagonal, and *p*-values (adjusted for false-discovery rates) are shown above. Statistically significant *p*-values are shown in bold. Cells with “NA” are comparisons of groups in which the frequency of the use of morphological data alone differed by < 15%, and thus no tests were done. Archaea and Bacteria had identical frequencies (100% both) and so are not shown separately.

	Arthropoda	Chordata	Mollusca	Archaea/ Bacteria	Chromista	Fungi	Plantae	Protozoa
Arthropoda	-	<b>&lt;0.001</b>	0.109	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	NA	<b>&lt;0.001</b>
Chordata	27.75	-	<b>0.002</b>	<b>&lt;0.001</b>	NA	<b>0.038</b>	<b>&lt;0.001</b>	0.110
Mollusca	3.51	12.96	-	<b>&lt;0.001</b>	<b>0.002</b>	<b>&lt;0.001</b>	0.067	0.081
Archaea/ Bacteria	72.41	19.05	51.52	-	<b>&lt;0.001</b>	<b>0.019</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
Chromista	27.75	NA	12.96	19.05	-	<b>0.038</b>	<b>&lt;0.001</b>	0.110
Fungi	52.02	6.60	32.96	7.53	6.60	-	<b>&lt;0.001</b>	<b>&lt;0.001</b>
Plantae	NA	30.13	4.57	75.44	30.14	54.86	-	<b>&lt;0.001</b>
Protozoa	14.67	3.97	4.70	33.33	3.97	15.57	16.51	-

Table 3. Subsampling experiments to address the impacts of limited sampling of species. We show the data type (majority data type) that was most frequent among the 50 sampled new species in each group (frequency in parentheses). Then we show the percentage of the 20 subsampled replicates (10 species per replicate) in which the majority data type (for the replicate) matches the overall majority data type from 50 samples. For each data type, we then give the mean number of species (out of 10) delimited by that data type among the 20 replicates for that group, followed by the minimum and maximum values among replicates in parentheses (see Dataset S11 for full results for each replicate).

Group	Majority data type	Replicates matching majority	Morphology only	Molecular only	Both
Arthropoda	Morph. (84%)	100%	8.40 (7–10)	0 (0)	1.60 (0–3)
Chordata	Both (68%)	85%	3.00 (0–5)	0 (0)	7.00 (5–10)
Mollusca	Morph. (68%)	90%	6.75 (4–10)	0 (0)	3.25 (0–6)
Archaea	Both (100%)	100%	-	-	-
Bacteria	Both (100%)	100%	-	-	-
Chromista	Both (68%)	75%	3.20 (0–6)	0 (0)	6.80 (4–10)
Fungi	Both (86%)	100%	1.20 (0–3)	0.20 (0–1)	8.60 (6–10)
Plantae	Morph. (86%)	100%	8.40 (7–10)	0 (0)	1.60 (0–3)
Protozoa	Both (50%)	50%	4.80 (1–9)	0.20 (0–1)	5.00 (1–9)

most studies that utilized molecular data used them to estimate a species-level phylogeny and genetic distance values (Table 5), rather than using methods explicitly designed for species delimitation. Importantly, we found that almost all studies that utilized molecular data also included morphological data, which makes a strong case for the distinctness of these species. Therefore, we are not suggesting that these new species are questionable because they did not use multi-locus data or statistical delimitation methods. Instead, we reiterate that an important challenge for future development of molecular species delimitation data and methods is to make sure that there are methods that

are broadly applicable to studies that formally describe new species.

Another surprising result of our study is that newly described cryptic species were rare (Table 1). Instead, most studies used morphological data alone to delimit species, or both morphological and molecular data. Only 2 of 500 new species were based on molecular data alone. This is surprising because cryptic species have been found in most major animal groups (Cahill et al., 2023; Pérez-Ponce de León & Poulin, 2016; Pfenninger & Schwenk, 2007) and plant groups (e.g. Ji et al., 2020; Kinoshita et al., 2020) and might outnumber morphologically distinct species in some (e.g. insects; Li & Wiens, 2023). One potential explanation for

Table 4. Types of molecular data used for species delimitation in each group. The columns “Nuclear”, “Mitochondrial”, and “Chloroplast” give the number of studies in which only those markers were used. “Nuc-Mito” corresponds to the use of both nuclear and mitochondrial DNA data whereas “Nuc-Chlo” indicates the use of both nuclear and chloroplast DNA.

Group	Nuclear	Mitochondrial	Chloroplast	Nuc-Mito	Nuc-Chlo
Animals					
Arthropoda	-	5 (62%)	-	3 (38%)	-
Chordata	-	20 (59%)	-	14 (41%)	-
Mollusca	-	11 (69%)	-	5 (31%)	-
Archaea	50 (100%)	-	-	-	-
Bacteria	50 (100%)	-	-	-	-
Chromista	14 (41%)	1 (3%)	-	19 (56%)	-
Fungi	35 (80%)	-	-	9 (20%)	-
Plantae	-	-	4 (57%)	-	3 (43%)
Protozoa	18 (69%)	-	-	8 (31%)	-

Table 5. Species delimitation methods used for molecular data. For each group, we examined the species descriptions that incorporated molecular data, tallied the methods that were used to infer species in each case, and estimated the proportion of relevant studies that used each method. Most studies used phylogenies but with no specific method listed (PNM) and most also used genetic distances (see Methods for those approaches considered under this category). Some new species had more than one delimitation method used in their description. Therefore, the combined percentages among methods within a group often exceed 100%. Abbreviations are as follows: 2D: consensus secondary structure models; ABGD: automatic barcoding gap discovery; BPP: Bayesian phylogenetics and phylogeography; GMYC: generalized mixed Yule coalescent; PNM: phylogeny used but no method given; PTP: Poisson tree process (bPTP: Bayesian implementation of PTP); unclear: it was unclear what species delimitation method the authors used to describe the species.

Group	Most common method	Second most common	Third most common
Arthropoda	PNM ( $n=7$ , 88%)	genetic distance ( $n=3$ , 38%)	ABGD/bPTP/GMYC ( $n=1$ each, 12% each)
Chordata	PNM ( $n=25$ , 74%)	genetic distance ( $n=24$ , 71%)	BPP ( $n=3$ , 8%)
Mollusca	genetic distance ( $n=12$ , 75%)	ABGD ( $n=8$ , 50%)	PNM ( $n=7$ , 44%)
Archaea	PNM ( $n=50$ , 100%)	genetic distance ( $n=50$ , 100%)	-
Bacteria	PNM ( $n=50$ , 100%)	genetic distance ( $n=49$ , 98%)	-
Chromista	PNM ( $n=34$ , 100%)	genetic distance ( $n=8$ , 24%)	-
Fungi	PNM ( $n=43$ , 98%)	genetic distance ( $n=4$ , 9%)	2D/PTP (1 each, 2% each)
Plantae	PNM ( $n=7$ , 100%)	genetic distance ( $n=2$ , 29%)	-
Protozoa	PNM ( $n=24$ , 92%)	genetic distance ( $n=5$ , 19%)	unclear ( $n=2$ , 8%)

this pattern is that those researchers who are discovering cryptic species are not primarily taxonomists, but are instead more interested in phylogeography, speciation, and other topics. Thus, they may not be focused on describing new species. Furthermore, large-scale phylogeographic analyses of cryptic species may require considerably more time and effort than many morphology-based species descriptions, leading to a much slower pace for the description of cryptic species. Another possible explanation is that cryptic species are not actually as widespread as estimated from studies that focused on cryptic species. However, such a pattern cannot be inferred solely from the high frequency of morphology-only species descriptions: it can only be inferred from relevant molecular studies that fail to find cryptic species. There might also be some pressure against publishing new descriptions of species based only on molecular

data (Cook et al., 2010), or a lack of confidence in describing species inferred from molecular data alone. It is also possible that many species that were initially thought to be cryptic proved to be morphologically distinct upon closer study.

We also need to make an important caveat about the time and expense of morphological studies. Although morphology-based taxonomy can appear fast and cheap at the scale of individual studies, there are also important long-term investments of time and money associated with it. These include the training of individuals with relevant taxonomic and morphological expertise, fieldwork, and the utilization, growth, and maintenance of museum collections (Engel et al., 2021).

The results here also included a survey of methods used in morphology-based species descriptions (Table 6). Most studies did not perform statistical or phylogenetic analyses



Table 6. Morphological data types and methods used in recent species descriptions. We recorded the number of species that were described using morphological data (alone or in combination with other data) and whether the traits examined were quantitative (e.g. measurements, counts) or qualitatively described (presence/absence, color), whether the differences between species were analyzed statistically, or whether there was a phylogenetic analysis of the morphological data.

Group	Species with morphological data	Quantitative traits	Qualitative traits	Statistical analysis	Phylogenetic analysis
Arthropoda	50 (100%)	50 (100%)	50 (100%)	1 (2%)	1 (2%)
Chordata	50 (100%)	50 (100%)	50 (100%)	13 (26%)	1 (2%)
Mollusca	50 (100%)	50 (100%)	50 (100%)	0 (0%)	0 (0%)
Archaea	50 (100%)	50 (100%)	50 (100%)	0 (0%)	0 (0%)
Bacteria	50 (100%)	45 (90%)	49 (98%)	0 (0%)	0 (0%)
Chromista	50 (100%)	50 (100%)	50 (100%)	2 (4%)	1 (2%)
Fungi	49 (98%)	49 (100%)	49 (100%)	0 (0%)	1 (2%)
Plantae	50 (100%)	50 (100%)	50 (100%)	3 (6%)	0 (0%)
Protozoa	49 (98%)	48 (98%)	49 (100%)	0 (0%)	0 (0%)

Table 7. Testing for biased sampling of new species in each group. Within each group, we used linear regression to test the relationship between the proportion of all new species in each higher taxon (independent variable) and the proportion of the 50 sampled species in each higher taxon (dependent variable). The strong relationships suggest that our sampling of 50 species in each group is unbiased. Note that species selection for Chromista and Protozoa was based only partially on random sampling, and they showed weaker relationships. Bacteria were not included because most species (70%) were from 2020, and so there was very little random sampling.

Group	Taxa	<i>n</i>	<i>r</i> <sup>2</sup>	<i>P</i>
Arthropoda	classes	11	0.952	<0.0001
Chordata	classes	14	0.986	<0.0001
Mollusca	classes	7	1.000	<0.0001
Fungi	phyla	9	1.000	<0.0001
Plantae	classes	11	0.992	<0.0001
Archaea	classes	8	0.995	<0.0001
Chromista	classes	7	0.754	0.0112
Protozoa	classes	9	0.790	0.0014

of their morphological data. Instead, they described selected morphological traits of their new species (both quantitative and qualitative), including potential diagnostic characters that can potentially distinguish the new species from previously described species. We note that the support for new morphology-based species could be further reinforced by statistical analyses of quantitative traits between species, by analyzing whether sample sizes are sufficient to distinguish the new species with qualitative traits (e.g. Wiens & Servedio, 2000), and by phylogenetic analyses of morphology that can show that the new species is phylogenetically distinct from related and/or similar species.

Our study offers the first survey of the data and methods used in species descriptions across living organisms (not just eukaryotes). We note several limitations of our study. However, these should not overturn our major conclusions. First, our sample sizes of species within each group were finite. For example, we could (hypothetically) have obtained information from each of the thousands of new species described in 2020, given infinite time. But we see no reason

why a sample of 50 species should be insufficient. Our subsampling experiments (Table 3) suggest that a sample of only 10 species would generally yield similar patterns. Importantly, our goal was to estimate the overall pattern within each major group. Sampling 500 arthropod species instead of 50 might show that the frequency of morphology-only species was (for example) 87% instead of 84%, but we do not think that such a difference is important: our point is that most arthropod species were described based on morphology. Some readers might reasonably be concerned that our sample of species is biased, by focusing on only one year. We see no reason why 2020 should be problematic, especially since any issue would have to extend across dozens of different studies within each group. On the other hand, 2020 included few new species for some groups, and thus we included earlier years for those groups (e.g. Chordata, Chromista, Bacteria, Archaea). We also note that Miralles et al. (2020) generally found similar patterns for major eukaryote groups (regarding the frequency of studies without molecular data) for 2012, 2014, 2016, and 2018,

strongly suggesting that our results are not an artifact of limited sampling or of focusing on 2020.

We also showed (Table 7) that our sampling of 50 species within each group was closely related to the overall patterns of new species in each group, in terms of their distribution among higher taxa (e.g. fungal phyla and classes of arthropods, chordates, mollusks, and plants). This pattern strongly suggests that our sampling within each group was broadly representative and not biased. Furthermore, our sampling of new species within each group was generally concordant with large-scale richness patterns within them (Bánki et al., 2025). For example, new arthropods included 60% insects and 30% arachnids whereas these groups include 83% and 8% of arthropods, and new arthropod species are 66% insects and 19% arachnids (Dataset S1). New insects in our sample were dominated by the largest orders, including Coleoptera, Diptera, Lepidoptera, and Orthoptera. Sampled mollusks (Dataset S3) consisted mostly of gastropods (98%) which are 72% of described mollusk species (and 94% of new species). Sampled plants were primarily angiosperms (88%) which are 91% of described species and 92% of new species (Dataset S6). Fungal species sampled (Dataset S5) belonged to Ascomycota (70%) and Basidiomycota (30%), which contain 64% and 34% of fungal richness (and 69% and 29% of new species). In other groups, our subsamples did not necessarily reflect overall richness patterns, but they did reflect the distribution of new species among subgroups (Table 7). In Chromista (Dataset S4), our 50 sampled species belonged mostly to Foraminifera (54%) and Oomycota (46%). Most described chromistans instead belong to Foraminifera (80%) not Oomycota (3%), but most newly described species ( $n=164$ ) were Oomycota (44%) and Foraminifera (33%). In Protozoa (Dataset S7), almost all sampled species belonged to Mycetozoa (98%) and mycetozoans contain only 50% of all described protozoans, but newly described species ( $n=84$ ) were mostly mycetozoans (96%). In chordates (Dataset S2), most sampled species belonged to ray-finned fishes (22%), amphibians (32%), and squamates (42%), whereas these groups make up 44%, 11%, and 16% of all described species, respectively. However, among new species ( $n=2415$ ; Dataset S2), most are squamates (41%), amphibians (25%), and ray-finned fishes (22%). In archaeans and bacteria, our sampling included most new species in each group. Overall, these results suggest that our sampling of new species within each group generally reflected the largest higher taxa within each group (arthropods, mollusks, fungi, plants), or those higher taxa that were growing the most quickly (chromistans, protozoans, chordates). Thus, our sampling of species within each group did not appear to be biased by limited sample sizes.

## 5 Conclusions

Species delimitation has become a major topic in systematic biology, as systematists race to discover life's diversity before it disappears. But how do systematists actually de-

scribe new species across life? We reviewed recent species descriptions to estimate the data and methods used to delimit new species across living organisms. We show that a molecular revolution in taxonomy has yet to materialize (at least by 2020): most new species in the largest groups of organisms were still described based on morphological data alone (arthropods, plants). Nevertheless, integration between molecular and morphological data has become standard in many important groups (archaeans, bacteria, chordates, fungi). However, use of multi-locus data and use of explicit methods for species delimitation remain uncommon, even when molecular data are used. Overall, we show a disconnect between the development of genomic datasets and statistical methods for species delimitation and how most new species are actually described by systematists. We speculate that morphology-based taxonomy remains dominant because it can be relatively fast and cheap, and because vast numbers of morphologically distinct species remain to be described. Thus, the field of species delimitation should also focus on how molecular species delimitation can also become faster and cheaper (and accurate), to help win the race between species discovery and species extinction. Furthermore, given the demonstrated importance of morphology-based taxonomy, winning this race will also depend on overcoming the rate-limiting factors of morphological taxonomy (i.e. taxonomic expertise, new fieldwork and specimen collection; Engel et al., 2021). Finally, cryptic species might equal or greatly outnumber morphologically distinct species, but we show that they are only rarely described formally. Therefore, the potentially monumental task of describing Earth's cryptic diversity has barely begun.

## Acknowledgments

We thank Bryan Carstens, Ryan Garrick, and Richard Leschen for helpful comments on the manuscript.

## Data Availability Statement

All data and code are currently available on figshare, and can be accessed via a private link (<https://figshare.com/s/2d5cff510d0a493a7ead>). These data have a permanent doi (10.6084/m9.figshare.28673615) which will be made public upon publication.

## Author Contributions

A.J.Z., X.L. and J.J.W. conceptualized and designed the research, collected and analyzed the data, and wrote the paper.

Submitted: April 10, 2025 EDT. Accepted: July 02, 2025 EDT.

Published: July 24, 2025 EDT.

## References

- Acosta, M. C., & Premoli, A. C. (2010). Evidence of chloroplast capture in South American *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). *Molecular Phylogenetics and Evolution*, 54, 235–242. <https://doi.org/10.1016/j.ympev.2009.08.008>
- Adams, M., Raadik, T. A., Burridge, C. P., & Georges, A. (2014). Global biodiversity assessment and hypercryptic species complexes: more than one species of elephant in the room? *Systematic Biology*, 63, 518–533. <https://doi.org/10.1093/sysbio/syu017>
- Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, 13, 729–744. <https://doi.org/10.1046/j.1365-294x.2003.02063.x>
- Bánki, O., Roskov, Y., Vandepitte, L., DeWalt, R. E., Remsen, D., ... Schalk, P. (2023). *Catalogue of life checklist (Version 2023-09-14)*. Catalogue of Life. <http://www.catalogueoflife.org>
- Bánki, O., Roskov, Y., Vandepitte, L., DeWalt, R. E., Remsen, D., ... Schalk, P. (2025). *Catalogue of life checklist (Version 2025-03-14)*. Catalogue of Life. <http://www.catalogueoflife.org>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57, 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., Ingram, K. K., & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, 22, 148–155. <https://doi.org/10.1016/j.tree.2006.11.004>
- Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 359, 669–679. <https://doi.org/10.1098/rstb.2003.1447>
- Cahill, A. E., Megléc, E., & Chenuil, A. (2023). Scientific history, biogeography, and biological traits predict presence of cryptic or overlooked species. *Biological Reviews*, 99, 546–561. <https://doi.org/10.1111/brv.13034>
- Cook, L. G., Edwards, R. D., Crisp, M. D., & Hardy, N. B. (2010). Need morphology always be required for new species descriptions? *Invertebrate Systematics*, 24, 322–326. <https://doi.org/10.1071/IS10011>
- Costello, M. J., May, R. M., & Stork, N. E. (2013). Can we name Earth's species before they go extinct? *Science*, 339, 413–416. <https://doi.org/10.1126/science.1230318>
- Costello, M. J., Wilson, S., & Houlding, B. (2012). Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology*, 61, 871–833. <https://doi.org/10.1093/sysbio/syr080>
- Costello, M. J., Wilson, S., & Houlding, B. (2013). More taxonomists describing significantly fewer species per unit effort may indicate that most species have been discovered. *Systematic Biology*, 62, 616–624. <https://doi.org/10.1093/sysbio/syt024>
- Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., ... Arneeth, A. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science*, 366, eaax3100. <https://doi.org/10.1126/science.aax3100>
- Engel, M. S., Ceriaco, L. M. P., Daniel, G. M., Dellapé, P. M., Löbl, I., ... Marinov, M. (2021). The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society*, 193, 381–387. <https://doi.org/10.1093/zoolinlean/zlab072>
- Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, 18, 4541–4550. <https://doi.org/10.1111/j.1365-294X.2009.04380.x>
- GBIF Secretariat. (2023). GBIF Backbone Taxonomy [Dataset]. In *Checklist dataset*. <https://doi.org/10.15468/39omei>
- Godfray, H. C. J. (2002). Challenges for taxonomy. *Nature*, 417, 17–19. <https://doi.org/10.1038/417017a>
- Guedes, J. J. M., Gomes de Lima, H. V., Mendonça, L. R., Chen-Zhao, R., Diniz-Filho, J. A. F., & Moura, M. R. (2024). Temporal trends in global reptile species descriptions over three decades. *Systematics and Biodiversity*, 22, 2419832. <https://doi.org/10.1080/14772000.2024.2419832>
- Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270, 313–321. <https://doi.org/10.1098/rspb.2002.2218>

- Hebert, P. D. N., Penton, E. H., Burns, J., Janzen, D. J., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly, *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences, U.S.A.*, 101, 14812–14817. <https://doi.org/10.1073/pnas.0406166101>
- Hickerson, M. J., Meyer, C., & Moritz, C. (2006). DNA barcoding will fail to discover new animal species over broad parameter space. *Systematic Biology*, 55, 729–739. <https://doi.org/10.1080/10635150600969898>
- Ji, Y., Liu, C., Yang, J., Jin, L., Yang, Z., & Yang, J.-B. (2020). Ultra-barcoding discovers a cryptic species in *Paris yunnanensis* (Melanthiaceae), a medicinally important plant. *Frontiers in Plant Science*, 11, 411. <https://doi.org/10.3389/fpls.2020.00411>
- Jones, G., Aydin, Z., & Oxelman, B. (2015). DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*, 31, 991–998. <https://doi.org/10.1093/bioinformatics/btu770>
- Kinosian, S. P., Pearse, W. D., & Wolf, P. G. (2020). Cryptic diversity in the model fern genus *Ceratopteris* (Pteridaceae). *Molecular Phylogenetics and Evolution*, 152, 11. <https://doi.org/10.1016/j.ympev.2020.106938>
- Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic Biology*, 63, 534–542. <https://doi.org/10.1016/j.ympev.2020.106938>
- Li, X., & Wiens, J. J. (2023). Estimating global biodiversity: the role of cryptic insect species. *Systematic Biology*, 72, 391–403. <https://doi.org/10.1093/sysbio/syaa069>
- Liu, L.-X., Du, Y.-X., Folk, R. A., Wang, S.-Y., Soltis, D. E., Shang, F.-D., & Li, P. (2020). Plastome evolution in Saxifragaceae and multiple plastid capture events involving *Heuchera* and *Tiarella*. *Frontiers in Plant Science*, 11, 361. <https://doi.org/10.3389/fpls.2020.00361>
- Löbl, I., Klausnitzer, B., Hartmann, M., & Krell, F.-T. (2023). The silent extinction of species and taxonomists—an appeal to science policymakers and legislators. *Diversity*, 15, 1053. <https://doi.org/10.3390/d15101053>
- Maddison, D. R., Guralnick, R., Hill, A., Reyesenbach, A. L., & McDade, L. A. (2012). Ramping up biodiversity discovery via online quantum contributions. *Trends in Ecology and Evolution*, 27, 72–77. <https://doi.org/10.1016/j.tree.2011.10.010>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Meier, R., Shiyang, R. K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, 55, 715–728. <https://doi.org/10.1080/10635150600969864>
- Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., ... Beszteri, B. (2020). Repositories of taxonomic data: where we are and what is missing. *Systematic Biology*, 69, 1231–1253. <https://doi.org/10.1093/sysbio/syaa026>
- Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on Earth and in the Ocean? *PLoS Biology*, 9, e1001127. <https://doi.org/10.1371/journal.pbio.1001127>
- Moreira, D., & López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, 7, 306–311. <https://doi.org/10.1038/nrmicro2108>
- Moura, M. R., & Jetz, W. (2021). Shortfalls and opportunities in terrestrial vertebrate species discovery. *Nature Ecology and Evolution*, 5, 631–639. <https://doi.org/10.1038/s41559-021-01411-5>
- Orr, M. C. C., Ascher, J. S., Bai, M., Chesters, D., & Zhu, C.-D. (2020). Three questions: How can taxonomists survive and thrive worldwide? *Megataxa*, 1, 19–27. <https://doi.org/10.11646/megataxa.1.1.4>
- Pérez-Ponce de León, G., & Poulin, R. (2016). Taxonomic distribution of cryptic diversity among metazoans: not so homogeneous after all. *Biology Letters*, 12, 20160371. <https://doi.org/10.1098/rsbl.2016.0371>
- Pfenninger, M., & Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology*, 7, 1211–1216. <https://doi.org/10.1186/1471-2148-7-121>
- Pimm, S. L., Jenkins, C. L., Abell, R., Brooks, T. M., Gittleman, J. L., ... Joppa, L. N. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344, 1–11. <https://doi.org/10.1186/1471-2148-7-121>



- Pons, J., Barraclough, T. G., Gomes-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., & Vogler, A. P. (2006). Sequence based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55, 595–609. <https://doi.org/10.1080/10635150600852011>
- Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, automatic barcode gap discovery for primary species delimitation. *Molecular Ecology*, 21, 1864–1877. <https://doi.org/10.1111/j.1365-294X.2011.05239.x>
- R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rieseberg, L. H., & Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *American Journal of Botany*, 5, 65–84.
- Rubinoff, D., & Holland, B. S. (2005). Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology*, 54, 952–961. <https://doi.org/10.1080/10635150500234674>
- Shen, W., & Ren, H. (2021). TaxonKit: a practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics*, 48, 844–850. <https://doi.org/10.1016/j.jgg.2021.03.006>
- Smith, M. L., & Carstens, B. C. (2020). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, 74, 216–229. <https://doi.org/10.1111/evo.13878>
- Stork, N. E. (2018). How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology*, 63, 31–45. <https://doi.org/10.1146/annurev-ento-020117-043348>
- Streicher, J. W., Sadler, R., & Loader, S. P. (2020). Amphibian taxonomy: Early 21st century case studies. Editorial. *Journal of Natural History*, 54, 1–13. <https://doi.org/10.1080/00222933.2020.1777339>
- Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology and Evolution*, 18, 70–74. [https://doi.org/10.1016/S0169-5347\(02\)00041-1](https://doi.org/10.1016/S0169-5347(02)00041-1)
- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE*, 7, e29715. <https://doi.org/10.1371/journal.pone.0029715>
- Wiens, J. J., & Servedio, M. R. (2000). Species delimitation in systematics: inferring diagnostic differences between species. *Proceedings of the Royal Society of London, Series B*, 267, 631–636. <https://doi.org/10.1098/rspb.2000.1049>
- Will, K. W., Mishler, B. D., & Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, 54, 844–851. <https://doi.org/10.1080/10635150500354878>
- Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences, U.S.A.*, 107, 9264–9269. <https://doi.org/10.1073/pnas.0913022107>
- Yuan, Z., Wu, D., Xu, W., Gao, W., Dahn, H. A., Liu, X., Jin, J., Yu, C., Xiao, H., & Che, J. (2023). Historical mitochondrial genome introgression confounds species delimitation: evidence from phylogenetic inference in the *Odonotricha grahami* species complex. *Current Zoology*, 69, 82–90. <https://doi.org/10.1093/cz/zoac010>
- Ziegler, A. J., Li, X., & Wiens, J. J. (2025). *Data from: The evidence for new species across the Tree of Life: morphology still rules the largest kingdoms*. <https://doi.org/10.6084/m9.figshare.28673615>
- Zink, R. M., & Barrowclough, G. F. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, 17, 2107–2121. <https://doi.org/10.1111/j.1365-294X.2008.03737.x>

## Appendices

### Appendix S1. Obtaining Data from the Catalogue of Life

#### Obtaining Data

We downloaded all species records from the Catalogue of Life website (CoL) using the Catalogue of Life Data Package (ColDP) on 13 December, 2023 (Bánki et al., 2023, database version: 2023-11-24). The ColDP includes several files with different types of information (see <https://www.catalogue-oflife.org/about/colusage#data-formats>). In this study, we focused on the “NameUsage.tsv” file, which includes the species name and the date of description.

To prepare the dataset for further analysis, we followed several steps. The 2,045,078 accepted species records were selected using the “rank” of species and were listed as “accepted” under the column for status (i.e. we included only accepted species). There were 3,892 duplicate species records with identical scientific names. The duplicated species were dropped and only the first species was kept, using the function `drop_duplicates(["col:scientificName"], keep="first")` in the *pandas* data frame (McKinney, 2010). The 134,818 extinct species were excluded (species listed as “True” under the column “extinct”). Species were retained as extant when there was no value in this column.

To determine the higher taxa to which each species belonged, we recursively searched the “NameUsage.tsv” file using the species’ “parentID” column. We then assigned each species to different taxonomic ranks above the genus level (i.e. kingdom, phylum, class, order, family). Each of these taxonomic ranks was given a different column in our datasheet. Missing higher taxon names (above the family) were filled by using the species’ parent higher taxon name. For example, we filled in the order name of a species by searching for the parent taxon of the family to which the species was assigned in the previous recursive search step (and then repeated this for class, etc.). This procedure was executed using a Python module *pytaxonkit* (<https://github.com/bioforensics/pytaxonkit>) which provides a Python binding for the program *taxonkit* (Shen & Ren, 2021). The program *taxonkit* searches for the parent lineage of a given taxon name in the NCBI taxonomy database (<https://www.ncbi.nlm.nih.gov/taxonomy/>, downloaded on 19 December 2023).

We excluded viruses from our dataset given that many scientists do not consider viruses to be living organisms (Moreira & López-García, 2009). The names of virus kingdoms were used to identify virus species in our dataset and exclude them. We obtained virus-related kingdom names from the Global Biodiversity Information Facility (GBIF) website, since GBIF gives more kingdom names than the CoL. Viruses were excluded by matching a species’ kingdom name (including “Unranked-Viruses”) to the viruses kingdom names from GBIF.

#### Determining the Year of Publication

The year that the species description was published was parsed from the column “authorship”. A regular expression pattern (“()”; four consecutive numbers) was used to parse the year. This was done using a custom script in Python after the dataset was read as a *pandas* data frame (McKinney, 2010). However, not all records included the publication year within the authorship field. In these cases, we generally retrieved this information from the reference field “issued” with the same regular expression (e.g. extracting only the year from a date given as “2000-01-01”). We also obtained this information from raw text from the columns “nameReferenceID” and “referenceID”. For this we used a new regular expression (“\(\); four consecutive numbers enclosed in parentheses) in the file “reference.tsv”. We also created a new column “x\_citation\_fix” to designate reference texts that needed subsequent checking (see below). The column “nameReferenceID” corresponds to the nomenclatural reference. The column “referenceID” corresponds to the taxonomic reference(s), and “issued” is the date the work was issued or published.

Some species records in CoL had more than one reference associated with the species name. In these cases, the earliest year was chosen as the year of publication (assuming that the scientific name was published first in that year). Some journals corrected the year in which the paper was published (e.g. “2007 publ. 2008”). Following the International Code of Zoological Nomenclature Fourth Edition (ICZN), the formal published year is the later year (see the Article 21.2, 21.9, <https://www.iczn.org/the-code/the-code-online/>). The choice between adjacent years should have very limited impact on our results.

Some species still lacked a publication year, even after our initial screening directly from authorship, from the issued year of publication, and from the raw references (a total of 18,155 species). We manually checked by eye and corrected 45 species with ambiguous dates (e.g. unrealistically early or later than 2023). We did the same for 1,367 species that had a reference text but lacked a year of publication or had only a range of publication dates (e.g. publication year was written in the format of 1834–35 or 1834–1835, which cannot be recognized by the previous regular expression pattern). For references in which the year of publication was missing or unrealistic, we searched for the scientific names on [bing.com](https://www.bing.com) to find the year of publication in other databases or citations. We then validated the publication year from those external sources based on the reference in our dataset. Some references used a range of dates as the date of publication, in which case the latest year was chosen as the final year of publication (following Article 21.6 of the ICZN). There were 2,589 species published before 1758. These were removed since they were clearly not relevant to recent species descriptions.. Mora et al. (2011) also removed records before 1758. After these manual corrections, we were able to correct and fill in the publication year for an additional 1,292 species (45+1,247). However, among the 1,367 that were checked by hand, we were still unable to find the year of publication for 120 species. There

remained 16,908 species that lacked the necessary reference information to determine their date of publication.

After the data cleaning, the final dataset had 1,896,615 unique species records. The year of publication was obtained directly from CoL for 91.14% of the species. The publication year was parsed from the references for 7.97% of the species (including manual correction for 0.07%). The year of publication remained unclear for 0.89% of the species.

## Data Filtering

We filtered the data to find only those species that were described in 2020. We generated a list of the species described in 2020 for nine major groups. These included the three largest animal phyla (Arthropoda, Mollusca, Chordata) and six major groups ranked as kingdoms by the CoL (Archaea, Bacteria, Chromista, Fungi, Plantae, Protozoa). These nine groups spanned most animal species and included all non-animal groups across the tree of life. The related code was supplied in Dataset S12 and S13.

## Appendix S2. Obtaining Data from GBIF

### Obtaining Data

We used data from GBIF for Archaea and Bacteria. We downloaded data from the Global Biodiversity Information Facility website (GBIF), on 29 October, 2024. We used the database version: 28 August, 2023 (GBIF Secretariat, 2023). We obtained taxon information in the Darwin Core Archive (DwC-A) format (Wieczorek et al., 2012). We focused on the “Taxon.tsv” file, which includes the species name and the date of description.

As described in Appendix S1, we selected the accepted species and excluded the virus species in a *pandas* data frame (McKinney, 2010). We excluded all species except for

those in Archaea and Bacteria. A total of 50,370 species of Archaea and Bacteria were initially included.

## Determining the Year of Publication

The year of publication for each species in the GBIF dataset was obtained from three columns: “originalNameUsageID”, “scientificNameAuthorship” and “namePublishedIn”. We first obtained the original species records (if available) by following the taxon ID in “Taxon.tsv” file and then extracted the primary published year. For species lacking an “originalNameUsageID”, we used different regular expression patterns to extract the primary published year, depending on whether the authorship indicated a new combination. The `r"()\` was used to extract four consecutive numbers followed by a right parenthesis for a new combination of species and genus, and `r"()"` for truly new described species. If all previous steps failed to yield the primary published year, we resorted to the “namePublishedIn” column, which contained the raw text for the reference. Following these data processing steps, we obtained the primary publication year for 14,069 archaeal and bacterial species. Some of these were fossil species, but these were removed when encountered.

### Data Filtering

We initially included only Archaea and Bacteria species published in 2020 from the GBIF dataset. However, we found no new archaean species for 2020 and relatively few Bacteria. Therefore, we broadened the published year and included species described from 2015 to 2020 for both groups. The related code is supplied in Dataset S14.