Estimating vertebrate biodiversity using the tempo of taxonomy – a view from Hubbert's peak

BRUCE H. WILKINSON[1,2,*], LINDA C. IVANY[2] and CARL N. DRUMMOND[3]

[1]*Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI 48109-1005, USA*
[2]*Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY 13244, USA*
[3]*Department of Physics, Purdue University Fort Wayne, Fort Wayne, IN 46805, USA*

Reservoirs of natural resources are finite and, with increasing exploitation, production typically increases, reaches a maximum (Hubbert's peak) and then declines. Similarly, species are the currency of biodiversity, and recognized numbers are dependent upon successful discovery. Since 1758, taxonomists have exploited a shrinking reservoir of as-yet-unnamed vertebrate taxa such that rates of species description at first rose, reached a peak and then declined. Since about 1950, increases in research funding and technological advances have fostered a renewed increase in rates of discovery that continues today. Many attempts to estimate global biodiversity are forecasts from data on past rates of description. Here we show that rates of discovery of new vertebrate taxa have been dependent upon the size (richness) of the taxonomic pool under consideration and the intensity of 'sampling' effected by taxonomists in their efforts to discover new forms. Because neither the current number of as-yet-to-be-described taxa nor future amounts of taxonomic efforts can be known a priori, attempts to produce an accurate estimate of total global biodiversity based on past rates of discovery are largely unconstrained.

ADDITIONAL KEYWORDS: biodiversity – Hubbert – Linnaeus – taxonomy – vertebrates.

'We will only get a good answer to the age-old question of "how many species are there?" when we understand the population biology and social behavior of taxonomists' (Pimm & Joppa, 2015).

INTRODUCTION

Taxonomy is one of the oldest areas of biological investigation. The finding, naming and grouping of taxonomic entities are generally understood to be critical pursuits in the quest for a fuller understanding of the current state of Earth's biodiversity and its ongoing loss. As recently discussed by Moura & Jetz (2021), an assortment of biological, environmental and sociological circumstances serve to influence the probability that a new species will be discovered and described. Although past approaches to estimating group richness have often relied on knowledge of numbers of described taxa, the translation of these metrics to total or as-yet-undiscovered richness is fraught with assumptions. Coming from our perspective as geoscientists, we posit that the realization of a more complete taxonomy, and thus a better appreciation of Earth's total biodiversity, is similar to attempts to infer the sizes of global reserves of many natural resources currently under development. Species of vertebrates, barrels of oil and tons of copper, for example, can be thought of as the units that comprise reservoirs to be produced (discovered). They are distributed within local reservoirs (higher taxa, oil fields, ore deposits) that exhibit fractal-like (many small – few large) size frequency distributions, and the probability of their discovery depends on the sizes of the units in question and on the intensity of efforts devoted to their development. This comparison offers insight into our ability to project the total biodiversity of a group from the history of taxonomic discovery of the fraction currently known.

In the following, we first examine the numbers of taxonomic units comprising the biological reservoirs

---

*Corresponding author. E-mail: eustasy@umich.edu

that have been 'exploited' during the history of vertebrate taxonomy, and then discuss several metrics that might serve to quantify the scientific effort expended to produce said taxonomies. For the former, we describe the numbers and size frequencies of taxonomic units at different levels in the Linnaean hierarchy and how they have changed over the past ~260 years of discovery; in the latter, we consider change in the numbers of papers and authors and their association with the pace of description of new taxa over that same window of time. Based on observed patterns of change at different taxonomic levels, we then develop models that reproduce observed tempos of description. These confirm the supposition that observed rates of resource extraction (rates of taxonomic description) are a function of both reservoir size (total biological diversity) and intensity of exploitation (taxonomic effort). While past rates of description are known from observation, past and future expenditures of taxonomic effort reflect a complex concatenation of factors that cannot be forecast a priori. As a result, we are left with one known – the success rate of past efforts at description – and two unknowns – the actual amount of biological diversity that exists on the Earth's surface and the total effort that must be expended to effect a complete census of that diversity. We therefore suggest that, as has been maintained for reserves of natural resources (e.g. Lynch, 2010), estimates of total biodiversity are largely unconstrained and cannot be calculated with confidence from information on the history of taxonomic description to date. As much as one might like to have 'the number', species richness of the planet will remain an elusive target.

AVAILABILITY, SOURCES AND TREATMENT OF DATA

A variety of sources of taxonomic information are generally available for vertebrate organisms (data for mammals, for example, are accessible from The Mammal Diversity Database, the Mammal Species of the World, and the Integrated Taxonomic Information System). We have examined several databases for each of the groups considered here and find no important differences with respect to results of analyses performed; those selected for analysis herein afforded the most complete and/or current tabulations. Data for fishes, amphibians, reptiles, birds and mammals come, respectively, from FishBase (Froese and Pauly, 2019; www.fishbase.se), the Integrated Taxonomic Information System online database (www.itis.gov), the Reptile Database (Uetz *et al.*, 2020; www.reptile-database.org) the International Community of Ornithologists (IOC) World Bird List (Gill *et al.*, 2020; www.worldbirdnames.org) and the Mammal Diversity

Database, 2020 (www.mammaldiversity.org). Details about dates of access and numbers of taxonomic units, authors and publications are listed as the Supporting Information (Table S1).

Current names, taxonomic classification and age of discovery were tabulated for all valid species; many are now considered to have a different classification than originally described, and we used the most current. Dates of establishment for all higher taxonomic units are taken as those of the earliest identified species within that group. We use parameters of best-fit skewed normal and exponential distributions in order to derive semi-quantitative descriptors of the general trends apparent in these somewhat noisy data on taxonomic histories. Parameters of these best-fit functions are appended in the Supporting Tables. Many aspects of taxonomic histories are similar among different major groups. For brevity, we have elected to show and/or discuss various relationships with respect to one particular group as an example, and to illustrate that same relationship among other vertebrate classes in the Supporting Figures.

THE OBSERVED PACE OF TAXONOMY

The tempo of description of vertebrate taxa with time shows clear patterns related to hierarchical level (Fig. 1). Higher taxonomic levels (e.g. orders and families) were differentiated by and soon after Linnaeus in 1758, while numbers at lower taxonomic levels (e.g. genera and species) gradually increased, reached a peak and then decreased. Around 1950, numbers of species then underwent an abrupt and generally exponential increase in description that continues today. Of the 27 orders of modern mammals, 16 (59.3%) were recognized by Linnaeus (1758), and all were defined by 1894, a span of 137 years. In contrast, only 157 (2.0%) of the 6485 currently recognized mammal species were known to Linnaeus in 1758; about 40 new species are currently documented each year, and this rate of discovery has increased by about 2% per year since 1950. This 'ontogeny of taxonomy' embodied in the history of classification of mammals (Fig. 1) is typical of other vertebrate classes (Supporting Information, Fig. S1; Tables S1 and S2).

The history of exploration of oil fields and ore deposits reflects the fact that the largest reservoirs are more likely to be discovered early on. So too is the case with higher taxa: vertebrate groups described chronologically earlier are also those with the highest memberships. As above, more than half the current orders of mammals (16 of 27) were established by Linnaeus in 1758 based on his initial description of only 157 mammal species. Those original 16 mammal orders currently contain 95% of the 6485 now-recognized
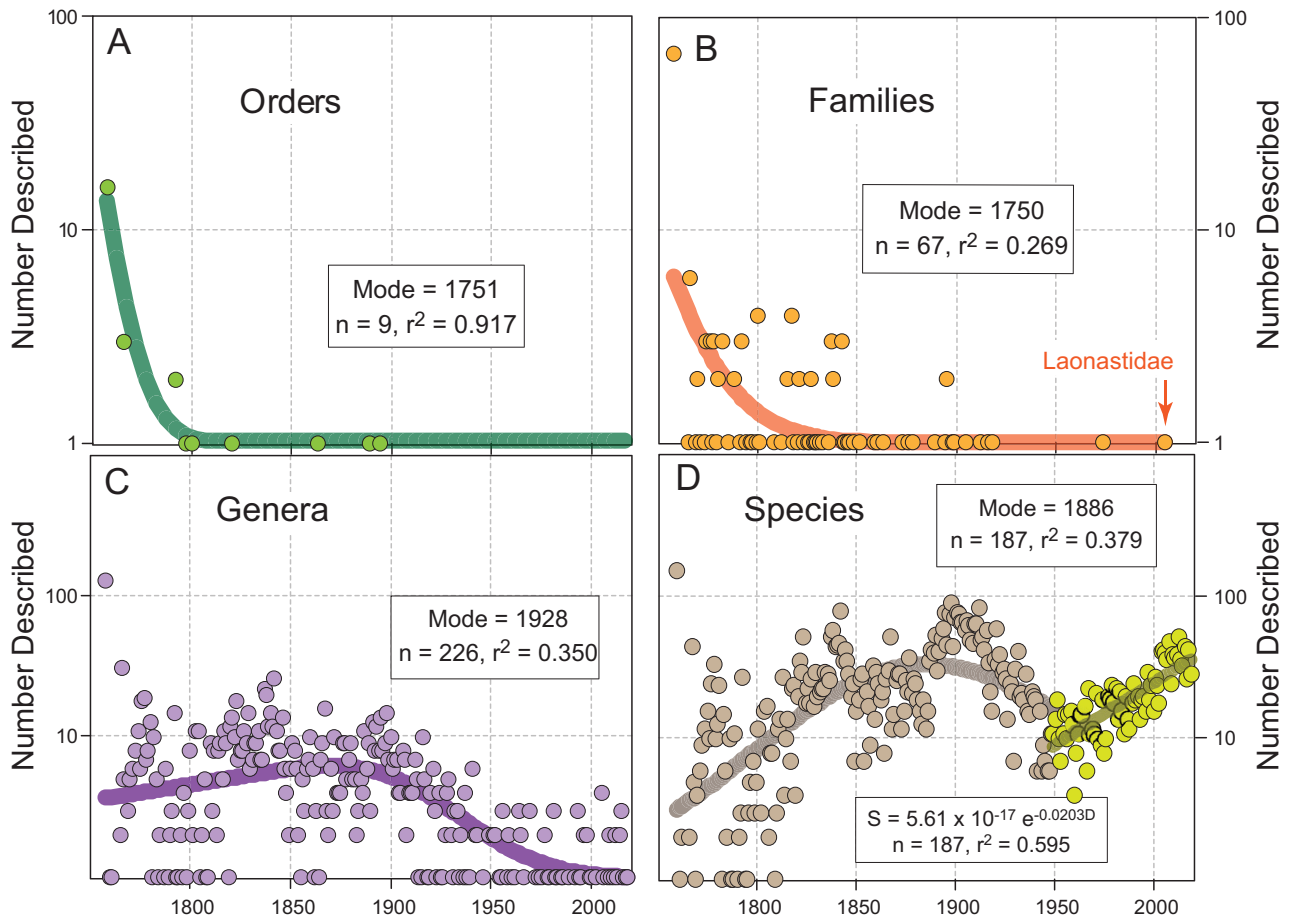
Figure 1. Rates (number/year) of named groups of living mammals. Data as filled circles; general trends as best-fit skewed normal distributions except post-1950 species (D, yellow) which is a best-fit exponential. With decreasing taxonomic rank, rates change from decreasing (orders and families) to modal (genera) to presently increasing (species). Note logarithmic *y*-axes. Similar plots for fish, amphibians, reptiles and birds as Figure S1A–M.

species (Fig. 2). Similarly, the 68 mammal families recognized by Linnaeus (1758) represent 41% of (the 167) currently recognized families and 74% of currently recognized species. Conversely, the 'youngest family' of living mammals (Laonastidae) was erected with the description of *Laonastes aenigmamus* by Jenkins *et al.* (2005) (though this family might actually represent a 'Lazarus' group that was thought to have become extinct in the Miocene; Dawson *et al.*, 2006). As currently defined, the family contains one genus and one species (Fig. 2B). Likewise, among finfishes, the 47 orders erected by Linnaeus (1758) (60% of the 78 modern orders) contain 33 913 (94%) of now-recognized species. The most recently defined order, Stylephoriformes (Miya *et al.*, 2007), contains one living species (*Stylephorus chordates*, the deep-sea tube-eye or thread-tail). Similar relationships between numbers of species and higher group (order, family,

genus) taxonomic–ontogenetic age are apparent among data for other classes of vertebrates (Supporting Information, Fig. S2).

These observations – of early recognition of higher taxa, continuing increases in description at the species level, and earlier recognition of the most speciose groups at any level – are requisite features of any effort to model the history of taxonomy. Embedded in these metrics are the need to understand the rules of taxonomic membership (the numbers of subtaxa within supertaxa) as well as the pace of description of the taxa themselves. We investigate both of these in subsequent sections before constructing numerical models to reproduce the major features of the histories of taxonomic endeavour, and ultimately to evaluate the potential for using such a model to predict future discoveries of new species.
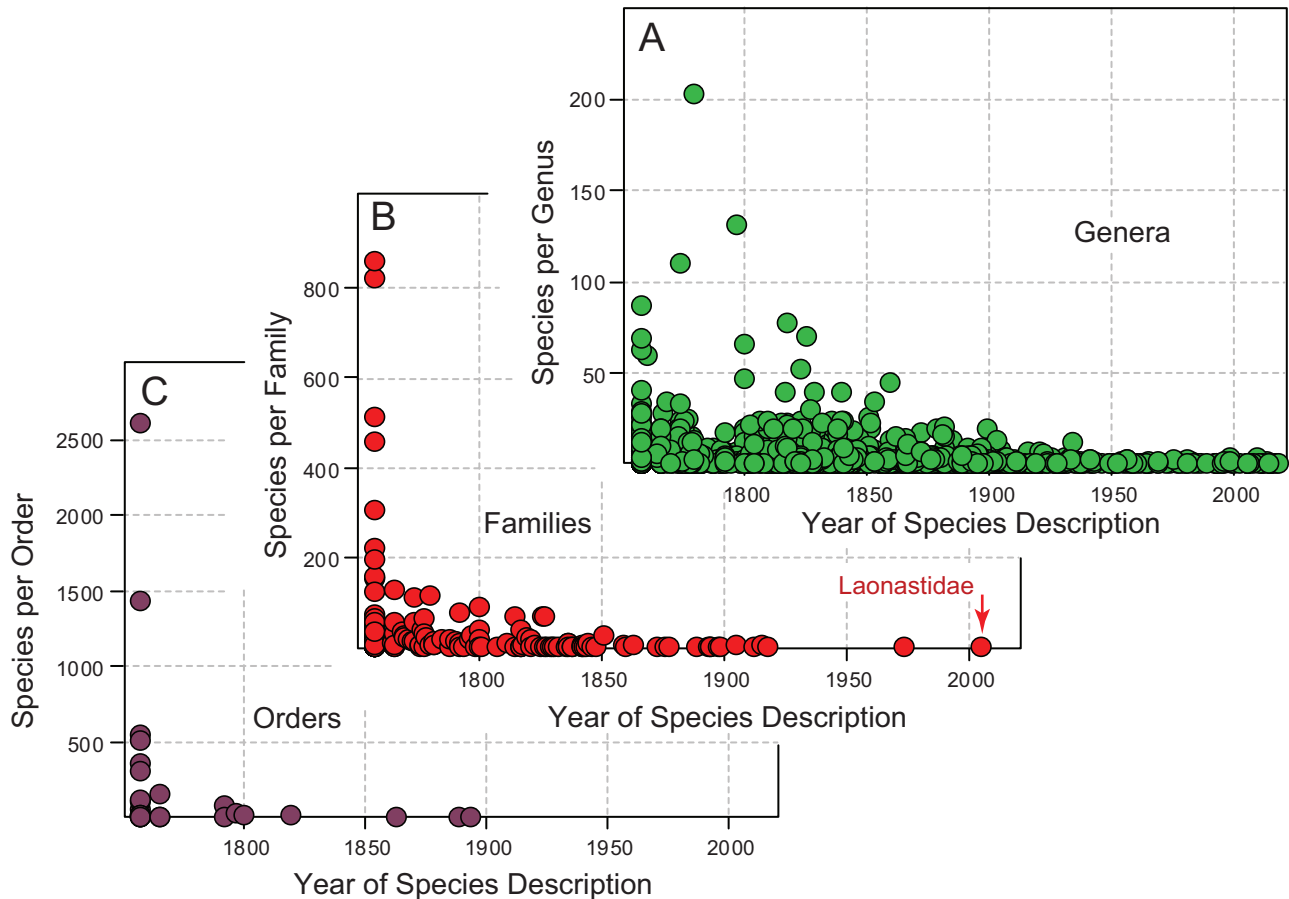
Figure 2. Relationships between the first year of description of some currently-recognized mammal species (*x*-axes) and the numbers of species in that group (*y*-axes). Number of members (species) largely predicates probability (the date) of group (order, family, genus) recognition. Similar plots for fish, amphibians, reptiles and birds as Figure S2A–L.

SIZES AND MEMBERSHIPS OF TAXONOMIC GROUPS

Understanding the numbers of subtaxa within any higher level of taxonomic consideration has been a focus of scholarship since Willis (1922) described the 'hollow-curve' nature of membership frequencies. Within any supertaxon, relatively few groups (taxa) contain most of the subtaxa, while many are monotypic (Supporting Information, Table S1). Different authors have variably interpreted these distributions as representing hyperbolic, logarithmic, log-normal, exponential, geometric and/or power law functions; Anderson (1974) provides an excellent review. Moreover, such 'hollow curves' have been ascribed to both deterministic and stochastic processes of biological diversification, and/or historical artefacts of taxonomic classification (e.g. Walters, 1961; Reddingius, 1971; Chu & Adami 1999; Scotland & Sanderson, 2004). An understanding of membership frequencies of subtaxa within supertaxa is critical to understanding differences in taxonomic histories at different levels of Linnean classification

(e.g. Fig. 1) and is requisite for the construction of model taxonomies, which provide insight into the utility of observed taxonomic data for estimating the numbers of remaining taxa to be described.

A higher taxonomic unit, such as an order, can be visualized as encompassing some amount of *n*-dimensional morphospace whose volume is occupied (and defined) by some number of non-contiguous subspaces, each representing a group (e.g. a family) that is a member of that order. The amount of morphospace occupied by that order is proportional to the number of families it contains, and we know from the frequency distribution of family sizes that a few are large (comprising a significant fraction of the genera or species within that order) and many are small. A two-dimensional (2-D) plane through this visualized morphospace transects the volumes of the contained families such that each is represented by an area on that plane. The diameters of those areas describe an exponential density function that is the same as that which describes the variation in the sizes of taxonomic

subgroups (like the families comprising an order) (Wilkinson, 2011). Because here we presuppose that taxonomic 'size' equates to numbers of subgroups, 'size diameter' approximately corresponds to the square root of the numbers of members comprising that group (e.g. genera or species included in a family).

Consider the 11242 species partitioned among the 92 families that comprise the class Reptilia. The frequency distribution of memberships (species) within families defines a curvilinear trend in log membership vs. log exceedance space (Fig. 3A) that reflects the variation in areas of families intersected by any 2-D transect through reptile morphospace. The diameters of those areas exhibit an exponential distribution (many small, few large) (Fig. 3B).

Given this relationship, we can define a membership inclusion parameter **p**, the probability that the addition of 1 unit of 'size' (in this case a species) will result in its inclusion in a different family. This inclusion parameter is dependent only upon the number of families (F) and species (S) in this class:

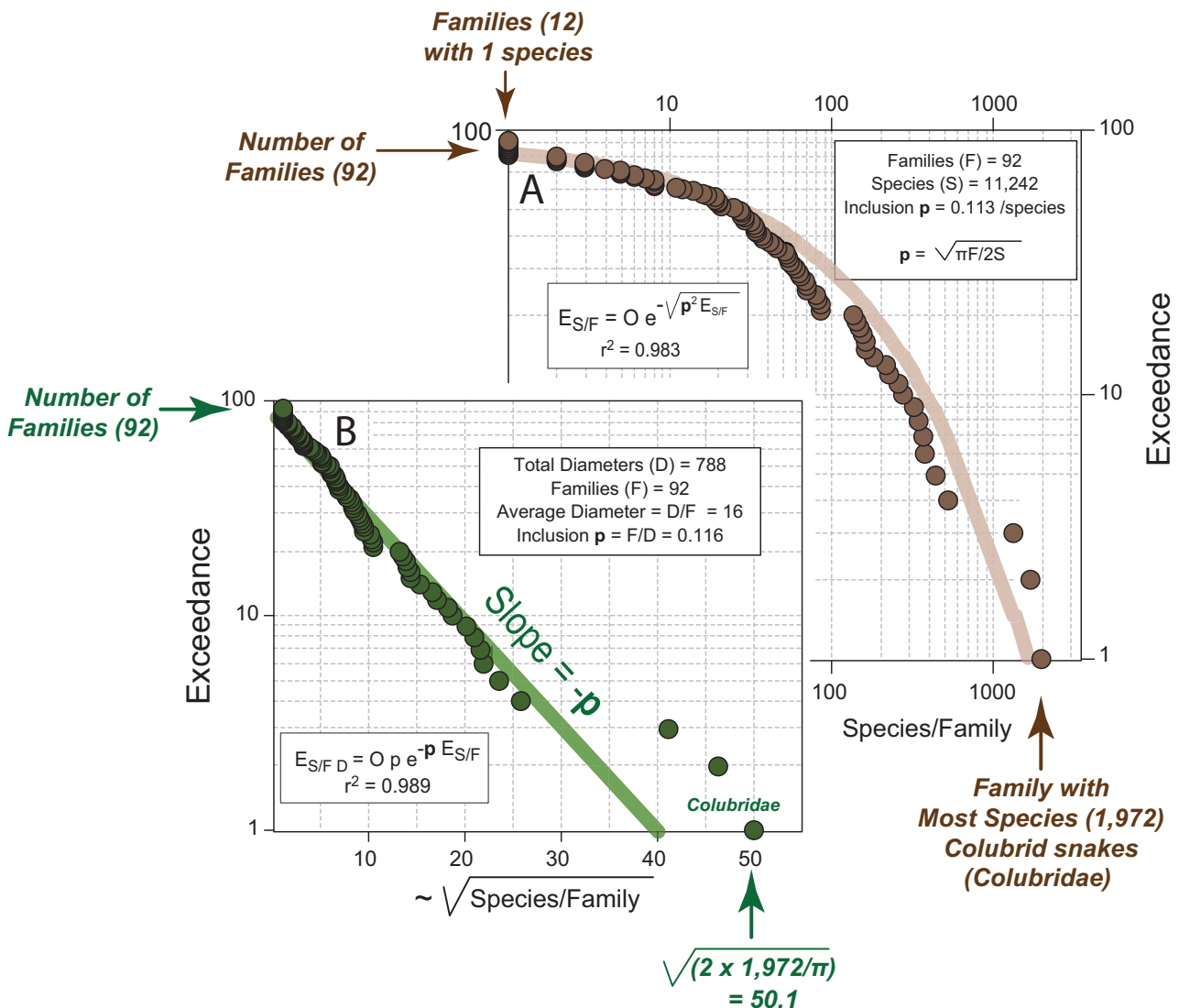$$\mathbf{p} = \sqrt{\frac{\pi \mathbf{F}}{2\mathbf{S}}}$$

Figure 3. Sizes of families of reptiles as measured by numbers of species per family. A, number of species per family among the 92 families of reptiles. B, frequencies of distances across 2-D transects of 'shape-space' represented by each of the families. Functions describing numbers of species (A) and 'diameters' of numbers of (B) both suggest that lateral boundaries of taxonomic units occur independently (randomly). In both cases, the inclusion parameter, **p** (~0.113), is the likelihood of incrementing the number of families with the addition of one species.

The cumulative percentage of families with memberships equal to or greater than some number of species (s) is the exceedance (E_{Fs}), and is given as:

$$\mathbf{E_{Fs} = Fe}^{-\sqrt{\mathbf{p^2 Fs}}}$$

This function yields good agreement with the distribution of species memberships within families of reptiles (Fig. 3A), and suggests that the distribution of taxon sizes (measured as the square root of numbers of subtaxa within that group) is the same as that of the diameters across morphospaces whose areas are exponentially distributed. Agreement between this theoretical distribution and observed data on taxon sizes suggests that taxonomic memberships largely represent the random subdivision of the n-dimensional morphospace occupied by some larger taxon (e.g. a class) where intermediate taxon sizes (e.g. families) are represented by the numbers of contained subtaxa (e.g. species). Memberships of taxa within a larger group are therefore dependent only on their number and the total number of subtaxa of which they are comprised.

Observed taxonomic memberships of vertebrate groups at all levels of Linnaean classification are closely approximated when assuming such a stochastic pattern of division. When considering the six possible taxonomic subgroupings (species per genus, per family, per order; genera per family, per order; and families per order) within the class Mammalia, for example, all model memberships are in good agreement with the data (Fig. 4). Data on group memberships for finfishes, amphibians, reptiles and birds exhibit similar correspondence (Supporting Information, Fig. S3, Table S3). Agreement between a stochastic function representing the chance division of morphospace and taxonomic membership data observed in the real world, for these and other animal groups, suggests that current taxonomic classification largely serves to randomly subdivide a morphospace continuum (Wilkinson, 2011).

The relationship describing the distribution of subtaxa among taxa in a larger group is consistent across taxonomic levels; that is, **p** is similar between levels with similar amounts of taxonomic separation (TS; Table S3), and generally decreases with increasing degrees of TS (Fig. 5) as:

$$\mathbf{p = e}^{-1.12}\mathbf{TS}$$

The exponent of this decrease averaged across vertebrate classes, −1.12, is the natural logarithm of 32.5%, which represents the average decrease in **p** with increasing separation among Linnaean levels (i.e., **p** = 0.325^{TS}). Taxonomic memberships – the partitioning of subtaxa among higher taxa – are largely the same, regardless of taxonomic levels of consideration.

ASSESSING EFFORT – WHAT CONTROLS RATES OF TAXONOMIC DESCRIPTION?

Forecasts of biodiversity also depend on attaining some aggregate measure of the intensity of efforts exerted by taxonomists in order to reveal said diversity. The issue of 'taxonomic effort' is important (e.g. Cribb & Bray, 2011). If the collective sampling efforts needed to discover one new taxon were invariant, then the rate of description of new taxa should decrease with time because the pool of remaining (undescribed) taxa will decrease as well; a classic case of rarefaction. The fact that rates of new descriptions at lower taxonomic levels (genera and species) have increased during of much of the last 200 years (Fig. 1; Supporting Information, Fig. S1) requires that the intensity of sampling – the amount of effort devoted to measuring biodiversity – has increased as well.

How does one measure aggregate taxonomic effort? Rates of discovery must directly or indirectly correspond to the number of opportunities that arise to recognize a new taxon each year, where 'opportunity' is the observed occurrence of a species in some place by some individual; some subset of these observations will be that of a new taxon. Data of this nature, on occurrences regardless of novelty, are recorded in some datasets (e.g. the Paleobiology Database, paleobio.org), but not consistently for modern taxa of vertebrates. A more tractable measure of taxonomic effort is perhaps the numbers of biologists engaging in descriptive research, and the number of publications in which new taxa are described per year. A number of studies suggest a relationship between rates of taxonomic description and the aggregate scientific exertions of systematists, who together examine many individuals from many populations spanning many biogeographical regions in search of something different enough to be considered a new taxon. Of particular interest has been the nature of relationships between rates of species description, and numbers of taxonomists and papers describing new species (e.g. Joppa et al., 2011a, b; Mora et al., 2011; Bebber et al, 2014; Costello et al., 2014; Gómez-Daglio & Dawson, 2019). We explore these below with respect to data on the major classes of vertebrates.

Rates of new species described and rates of taxonomic papers published (in which a new species is described) exhibit similar patterns of temporal variation. For example, beginning with the description of 330 species of finfish by Linnaeus in 1758, 33 913 species (Supporting Information, Table S1) have been described to the end of 2019 (254 years) in 6055 publications (3.0 species per paper). Over this time interval, rates of species description and rates of publication have both increased more or less exponentially. Since 1950, the rate of naming of new finfish species has increased
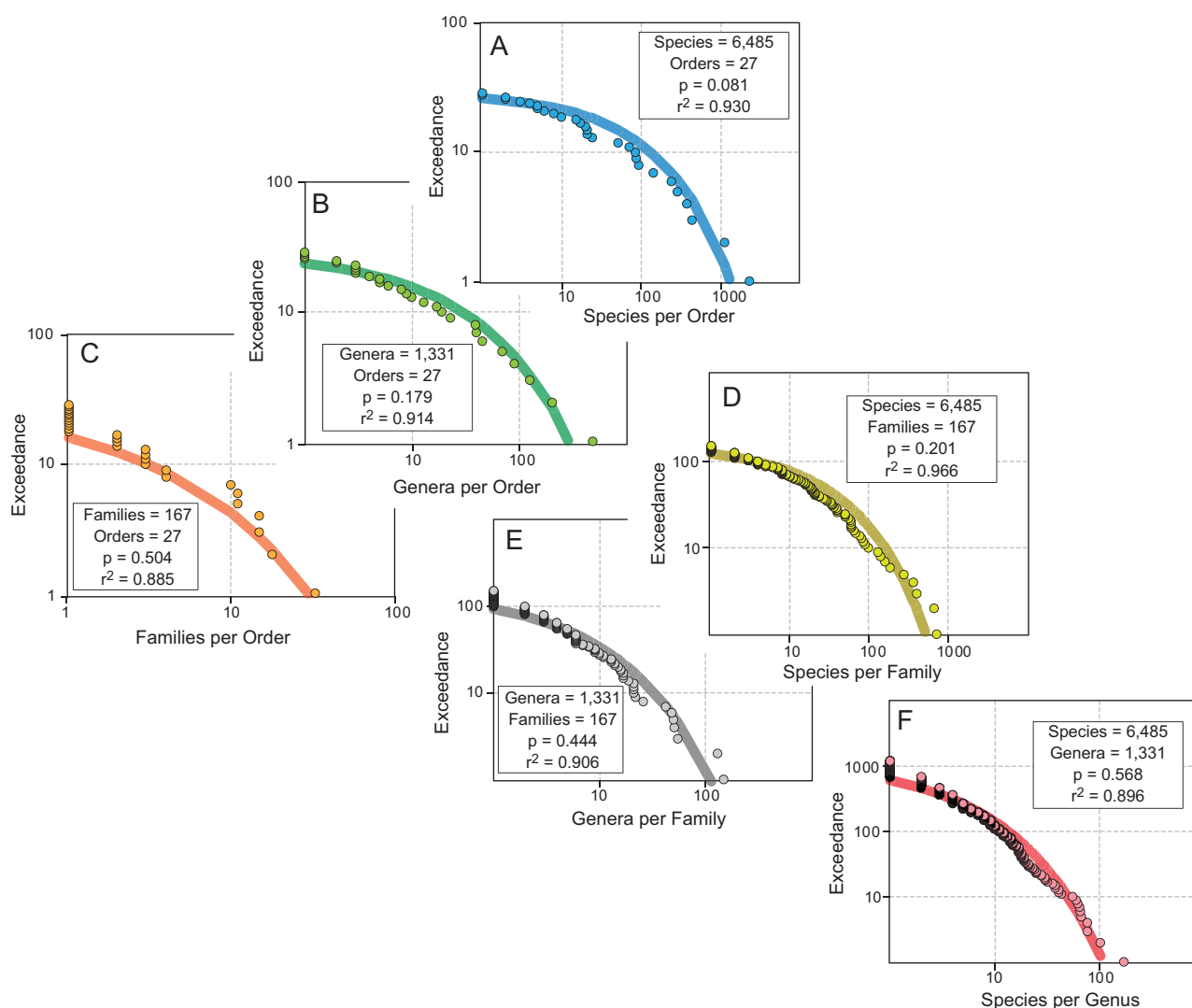
Figure 4. Random division functions fit to taxonomic data on mammals plotted as number of subtaxa among each hierarchically higher level of supertaxa (*x*-axes) relative to numbers of subtaxa that are equal to or are greater than some *x*-axis membership number (*y*-axes). Metrics for each membership curve are listed in Table S2. Similar plots for fish, amphibians, reptiles and birds as Figure S3.

by ~2.1% per year to a current pace of ~400 species per year (Fig. 6A). Similarly, since 1950, the number of publications containing species descriptions has increased by ~2.6% per year (Table S4) to a current rate of ~250 papers per year (Fig. 6B). Data on rates of description and numbers of papers related to species of amphibians, reptiles, birds and mammals exhibit nearly identical relationships (Fig. S4).

A decrease in the variance of the rate of species description is also apparent over time in all vertebrate taxonomies (Fig. 7), and most probably reflects the decreasing importance of monographs as an outlet for the formal recognition of new groups (e.g. Joppa *et al.*, 2011a). Residuals of best-fit trends through data on the numbers of species described per publication

define a trend of exponentially decreasing importance of monographs toward the present (Fig. 7B for amphibians; see Supporting Information, Fig. S5, Table S5 for other classes). These changes are just one ramification of the increasing specialization of modern scientific inquiry.

Trends in numbers of described species and numbers of papers designating new taxa exhibit a rather obvious change around the early 1950s (Fig. 6; Supporting Information, Fig. S4), when rates of description and publication began renewed and exponential increases that continue to the present. Rate metrics for all vertebrate classes are similar except for birds (an order of magnitude lower; Fig. S4), and equate to the naming of about 500, 160,
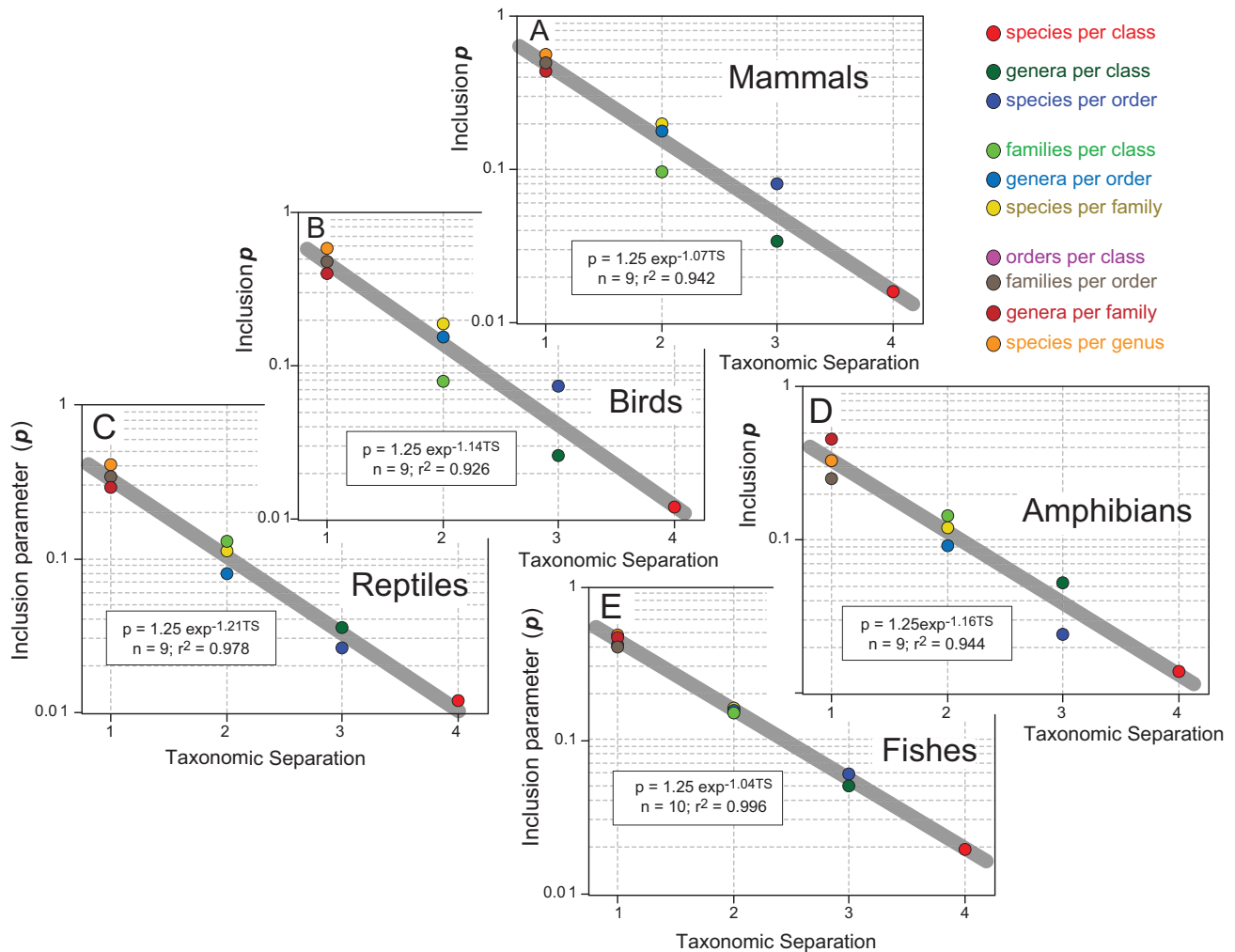
Figure 5. Values of the inclusion parameter **p** as a function of degree of taxonomic separation among classes of vertebrates. Here, the slope (on average, −1.12) is the natural logarithm of 0.325, and represents the rate of decrease in **p** for each increase in level of separation (0 = 100%; 1 = 32.5%, 2 = 10.6%, 3 = 3.4%, 4 = 1.1%). This rate of change corresponds to about a three-fold increase in membership with each increase in Linnaean level of taxonomy.

150, five and 40 new species of fishes, amphibians, reptiles, birds and mammals, respectively, in the year 2020. Interestingly, current rates of discovery and description are some several hundred times higher than estimated coeval rates of extinction of vertebrate species (Ceballos *et al.*, 2017, 2020).

Such a profound change – one of both sign (negative to positive) and acceleration – in species identification and publication rates seen across all vertebrate classes since *c.* 1950 argues for a common cause (e.g. Joppa *et al.*, 2011b). Publication norms have changed since 1950 (Ioannidis *et al.*, 2018; Gómez-Daglio & Dawson, 2019), particularly with respect to the proportion of multi-authored papers per year; might this contribute to the inflection? Single and dual authored papers account for ~85% of all new species descriptions over the past 250 years, but the number of multi-authored

papers began to increase (by ~10% per year) around 1950 (Supporting Information, Fig. S6). Papers with three or more authors account for ~65% of species descriptions since 2000. With more authors overall, the number of papers published has grown (Fig. 8A), and papers published since 1950 are also more likely to describe fewer (or one) new species per paper (Fig. 8B). The shift to multi-authored and monospecific papers alone, however, cannot account for the increase in both papers *and* species descriptions per year. An increase in the latter must be associated with increasing rates of discovery of new taxa.

Several factors in particular probably account for the 1950s' inflection in the rate at which new species are described. The establishment of national agencies dedicated to the public support of university-based research, including taxonomic studies, surely
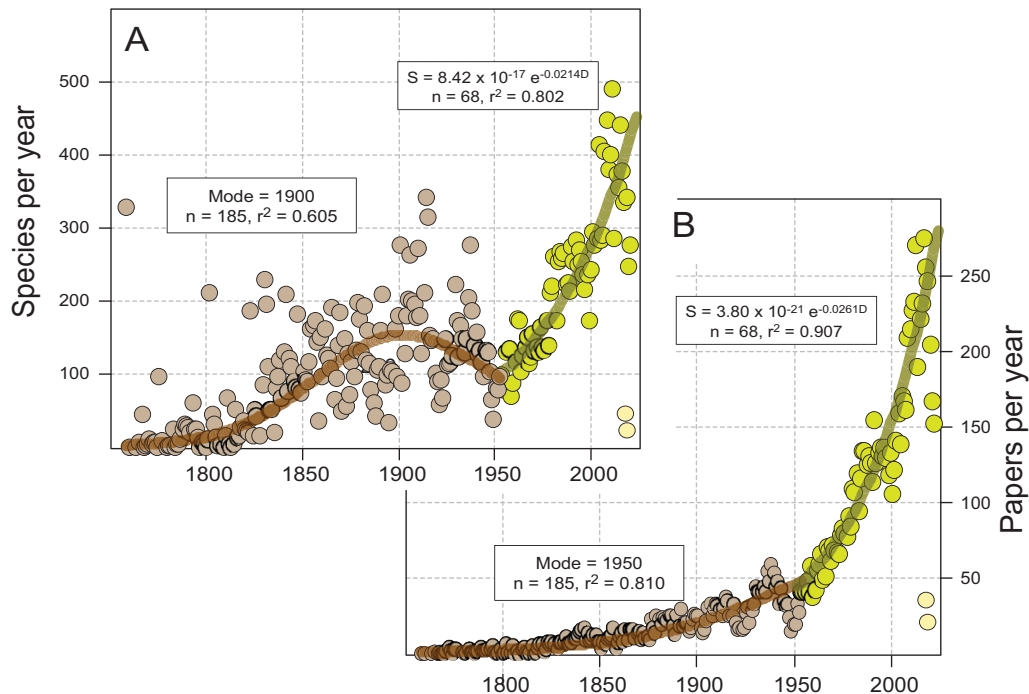
Figure 6. Relationships between naming of 33 913 new species of finfishes (A) and rates of their publication as papers per year by unique authors of sets of authors (B). Brown curves are best-fit skewed normal distributions to data before 1950; yellow curves are best-fit exponentials. Low numbers for 2018 and 2019 in both plots (lighter points) represent incomplete tabulations. Since 1950, rates of newly named species and associated papers have increased at about the same rate: by ~2.14 and 2.62% per year, respectively. Similar plots for amphibians, reptiles, birds and mammals as Figure S4A–H.

contributes to the renewed increase in description rate. In the United States, for example, the National Science Foundation was created in 1950. In Canada, while the National Research Council was created in 1916, war-related and medical research were handed off to newly formed organizations around 1950, and research funding in the natural sciences has been handled through the Natural Sciences and Engineering Research Council since 1978. Federal funding for scientific endeavour promoted growth in the number of taxonomic researchers and hence in the number of papers and new species described per year.

Access to new habitats afforded by technological innovation has also allowed researchers to explore biodiversity in places that had been difficult or impossible to reach before (Donoghue & Alverson, 2000). As an example, the ratio of marine to terrestrial species described each year began to increase at about this time (Costello *et al.*, 2012), reflecting a progressive expansion in the exploration of marine habitats, including deep-sea settings. Similarly, the increasing globalization of science has allowed more attention to formerly difficult-to-access or otherwise understudied biogeographical regions through the growing number of international collaborations and the increasing ease

with which scholars in the developing world can share their work with the broader international community (e.g. Grieneisen *et al.*, 2012).

Finally, the advent of molecular techniques with which to recognize and distinguish among genetically distinct populations must contribute to the renewed increase in the rate of description of new species. Data in Bouchet *et al.* (2016), for example, suggest that the description of marine molluscs based on molecular criteria has increased by ~35% per year over the past decade. In addition, the recognition of cryptic species has resulted in a subdivision of existing taxa into multiple new taxa based on genetic information (e.g. Bickford *et al.*, 2007; Pfenninger & Schwenk, 2007; Ladner & Palumbi, 2012); molecular methods may add tens of thousands of cryptic marine species (Appeltans *et al.*, 2012). Genetically distinct populations recognized using molecular data, however, are not always the same as the species that would have originally been defined morphologically, so the inclusion of both types of 'species' in the same dataset largely equates to the development of new reservoirs that were not considered during earlier research.

An intriguing correlate to the naming of vertebrate species can be found in the nearly identical changes
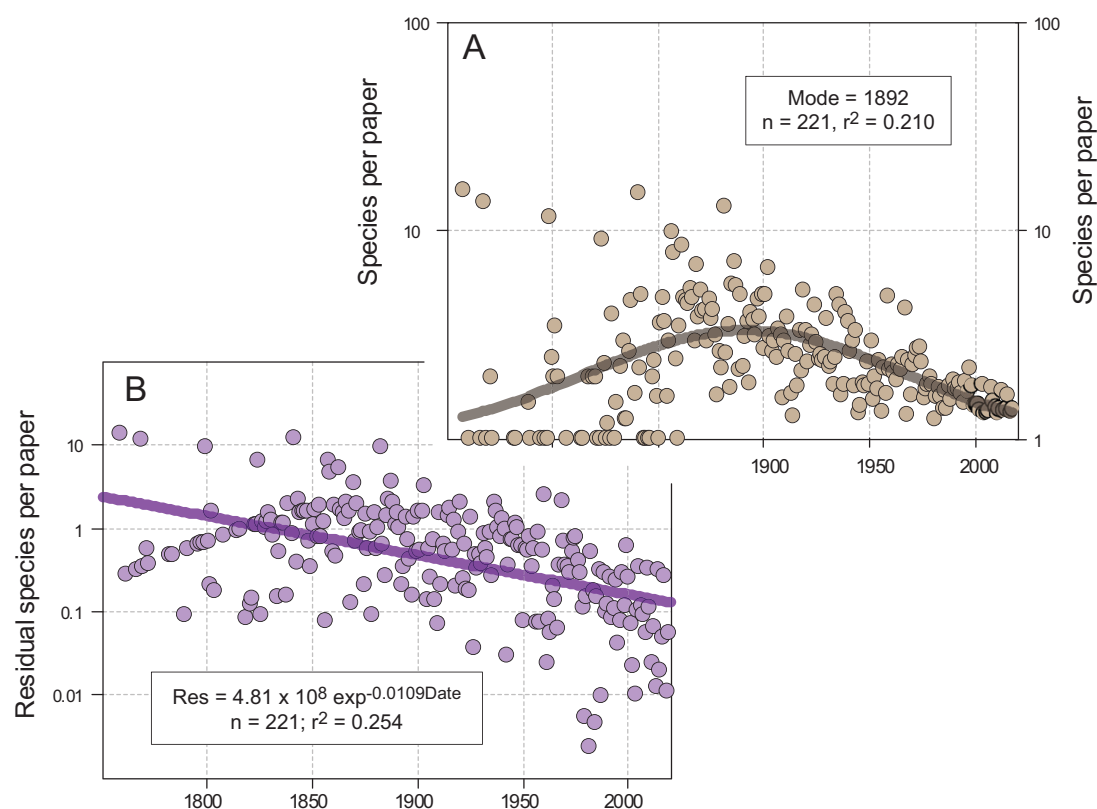
Figure 7. A, historical (*x*-axis) change in rate of description of new species of amphibians (*y*-axis, log scale). Best-fit skewed normal distribution (brown line) to log rates (brown dots) suggests a maximum taxonomic 'per paper productivity' *c.* 1890. Note decreasing scatter of rates with decreasing age. B, exponential decrease in the 'monograph effect' among amphibians manifested as decreasing differences between observed rates of description of new species and that described by a longer-term average. 'Spikiness' in the description of new species has decreased by ~1.1% per year since the end of the 17th century. Similar data for fishes, reptiles, bids and mammals as Figure S5.

in the rate of discovery of new minerals and the proportion of multi-authored papers describing new mineral discoveries that occurred at the same time. Barton (2019) reports that the rate of discovery of new minerals was relatively constant (~15 per year) from 1917 to about 1950, but then increased exponentially (by ~1.9% per year) to a current (2020) rate of ~100 per year. Over the same time interval, the average number of authors on mineral discovery papers also underwent an exponential (~2% per year) increase from abour two in 1950 to over six in 2020. She ascribes these changes to a variety of interrelated factors, including the greater availability of instrumentation and an exponential growth in the funding and focus of mineralogical research at universities and museums. The striking coherence in pattern between such disparate fields argues for similar drivers, and the combination of federal research funding, globalization and technological advance would affect both in the same way.

All this being said, do numbers of authors or publications or described taxa per year serve as valid proxies for 'taxonomic effort'? Probably not. Linnaean species are the coin of the biodiversity realm; the first description of any vertebrate species necessitates the identification of some genus, some family, and some order within the class to which it belongs. Numbers of species descriptions, published papers and/or contributing authors are all measures of the success of taxonomic efforts, rather than a direct measure of the aggregate scientific exertion expended by the community in deriving this classification. The difficulty in assessing taxonomic effort is that there is no easily accessible record of 'unsuccessful' research; there is no adequate metric for measuring the cost and/or effort expended during those expeditions that failed to locate and/or identify new forms.

Even if a satisfactory metric were available with which to track the history of taxonomic effort, several factors would complicate a straightforward forecast of future industry and its impact on probabilities of discovering new forms. First, there are a number of factors intrinsic to specific taxa that predicate
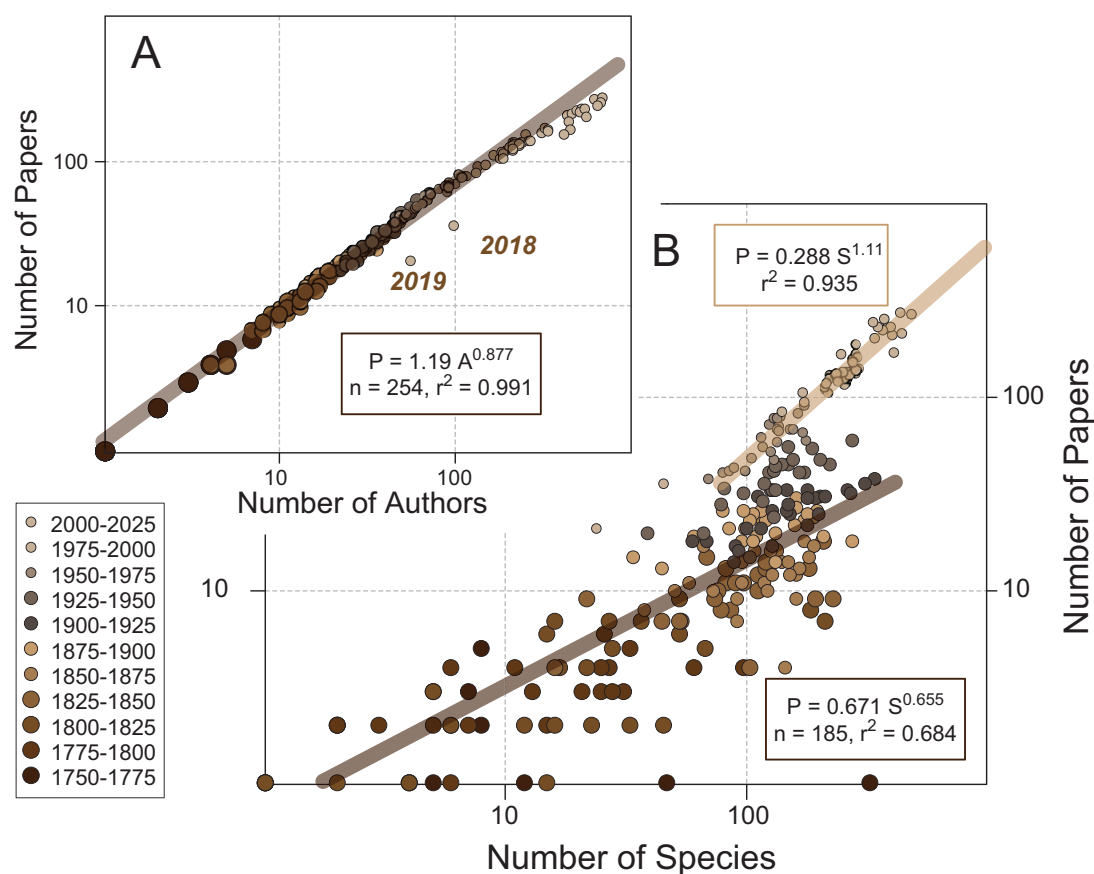
Figure 8. Relationships between numbers of papers and participating authors describing new vertebrate species (A), and numbers of species described (B) since 1758. A, number of authors and resulting papers are well correlated, an increasing proportion of multi-authorship giving rise to a slight decrease in slope. The two youngest points (2018 and 2019) are thought to represent incomplete database tabulations. B, correlations between numbers of papers and described species are also increasingly well correlated, with greater noise during earlier years reflecting a greater impact of monographs.

probabilities of discovery. These include things related to appearance such as flamboyance of colour, openness of biome and mode of mobility, as well as abundance of members (e.g. Cribb & Bray, 2011), range size (e.g. Collen *et al.*, 2004; Krasnov *et al.*, 2005) and body size (e.g. Gaston, 1991). Second, the spatial distributions of new (undescribed) taxa are geographically heterogeneous, and in order to attain taxonomic 'success' through their description, even relatively abundant species will not be found until exploration expands into their range; the discovery of hydrothermal vent communities in the late 1970s is a classic example. Geographical ranges and densities of taxa are also often variable and uncorrelated. More recently discovered species, for example, have typically been found in biodiversity hotspots of high endemism (e.g. Thompson *et al.*, 2021); Myers *et al.* (2000) estimated that 35% of all recently described amphibian, reptile, bird and mammal species are restricted to some 25 hotspots that collectively make

up only about 1.4% of the Earth's land surface. Sociopolitical infrastructures are also frequently limited in high-biodiversity regions, and scientists from countries with more robust taxonomic infrastructures typically require significant resources to explore in other areas (e.g. Grieneisen *et al.*, 2012).

The concatenation of all these and other factors, as illustrated so dramatically by the 1950s' inflection, collectively makes projections about the success rate of future taxonomic efforts based on past trajectory highly uncertain and subject to happenstance. If, as noted by Pimm & Joppa (2015), we can only achieve an accurate estimate of the size of the global biodiversity reservoir when we can also accurately forecast numbers, practices and collective efforts of systematists, we are left in a precarious position. The equation has one known – the success rate of past collective efforts – and two unknowns – the actual amount of biological diversity that exists on the Earth's surface, and the total effort necessary to effect a census of that diversity.

A better understanding of how different scenarios of future effort might affect estimates of total diversity is afforded from simple numerical models

MODELLING TEMPOS OF TAXONOMIES

Thus far, we have proposed that: (1) the tempo at which new Linnaean categories have been established is similar for equivalent ranks across different vertebrate classes; (2) rates of definition at the family level and higher have deceased rapidly since 1758, when a large number of groups were first erected; (3) rates of generic description increased until *c.* 1870 before falling off; (4) rates of description of new species exhibit a similar peak in the mid-to-late 1800s (1889 ± 20), declined until about 1950, and have exponentially increased since then, now by ~2.5% per year; (5) the most speciose orders, families and genera are historically the earliest ones defined; (6) Linnaean group memberships measured as numbers of contained subtaxa exhibit frequency distributions similar to those expected from the random division of taxonomic morphospace; and (7) numbers of described taxa, numbers of publications and numbers of contributing authors are all correlated manifestations of taxonomic successes. Although their rates of appearance are directly related to scientific effort expended, no readily available data exist by which to adequately quantify increases in total taxonomic effort.

To better understand the processes underlying these observations, and hence to evaluate the potential for predicting them in the future, we construct several numerical models that incorporate parameters reflecting both the inherent nature of the groups being sampled and described (numbers and memberships) as well as some model estimate of the effort expended in the process by systematists over time. As demonstrated in the prior section, quantifying the effort involved in discovering a new taxon is difficult, but one can bring some clarity to the impact of different degrees of effort on resulting biodiversity trajectories and tallies through the use of such simple numerical models. To this end, we combine the understanding of taxonomic memberships at different Linnaean levels with three prescribed scenarios of taxonomic effort in order to: (1) replicate taxonomic tempos at different Linnaean levels; (2) assess the impact of differing effort scenarios on taxonomic history at a single Linnaean level; and (3) examine how changes in taxonomic effort affect the tempo of taxonomic description.

Conceptually, we might imagine a very large population of marbles that represents some high taxonomic level (like a class) of vertebrates. Marbles each represent some unit at a lower taxonomic level, like species, and marbles comprise different colours,

each of which represents some intermediate level of Linnaean taxonomy, like families. The population is structured such that that most of the marbles (e.g. species) represent only a few colours (e.g. families), while many other colours are represented by a few or only one marble(s). The distribution of marbles (subtaxa) among colours (taxa) is determined by the stochastic function defining memberships for a given taxon provided by the inclusion parameter **p** (Section IV, [Figs 4](#), [5](#)). The population is randomly sampled with replacement over time, with each draw reflecting the taxonomic observation of a (new or previously recognized) species. Because the same marble (species) can be drawn more than once, the number of marbles drawn per time step reflects some measure of effort, as some observations turn out to be taxa already described. The time step in which new marbles (subtaxa) and new colours (taxa) are first sampled is the recorded year of discovery.

If sampling intensity – the number of marbles drawn in each time step – were constant over time, the number of new taxa discovered at any level per time step (per year) would have to decline from the inception, as speciose clades are discovered early and then repeatedly resampled, and progressively less speciose groups are eventually encountered by chance over greater intervals of time. This simple scenario, however, is inconsistent with the observation of modes in the number of new descriptions per year, which also occur later in time at progressively lower taxonomic levels (e.g. [Fig. 1](#)). These observed trends require that sampling intensity (taxonomic effort) must be increasing over time.

To more accurately capture the observed structure in taxonomic histories, we specify various changes in the rate of sampling (the number of marbles drawn, or taxonomic observations made) over time. The first model assumes a constant rate of increase in sampling effort and considers how membership histories unfold as a function of taxonomic rank. A second set of models explores the impact of each of three different rates of increase in sampling effort on the shapes of resulting membership histories at a single taxonomic rank. Both of these models span 350 years (1758–2108) of taxonomic history.

The first scenario ([Fig. 9](#)) is parameterized so as to loosely approximate the taxonomic histories of a clade of vertebrates (e.g. mammals, with 6485 described species divided among 1331 genera and 176 families; [Supporting Information, Table S1](#)). A pool of taxonomic observations comprising 6500 species is randomly sampled with replacement each year for 350 years, for a total of 30 000 observations. We presuppose that sampling effort increases exponentially by ~1% per year, from a rate of six observations made per year in 1758 to 154 per year in 2108 ([Fig. 9A](#)). Drawn
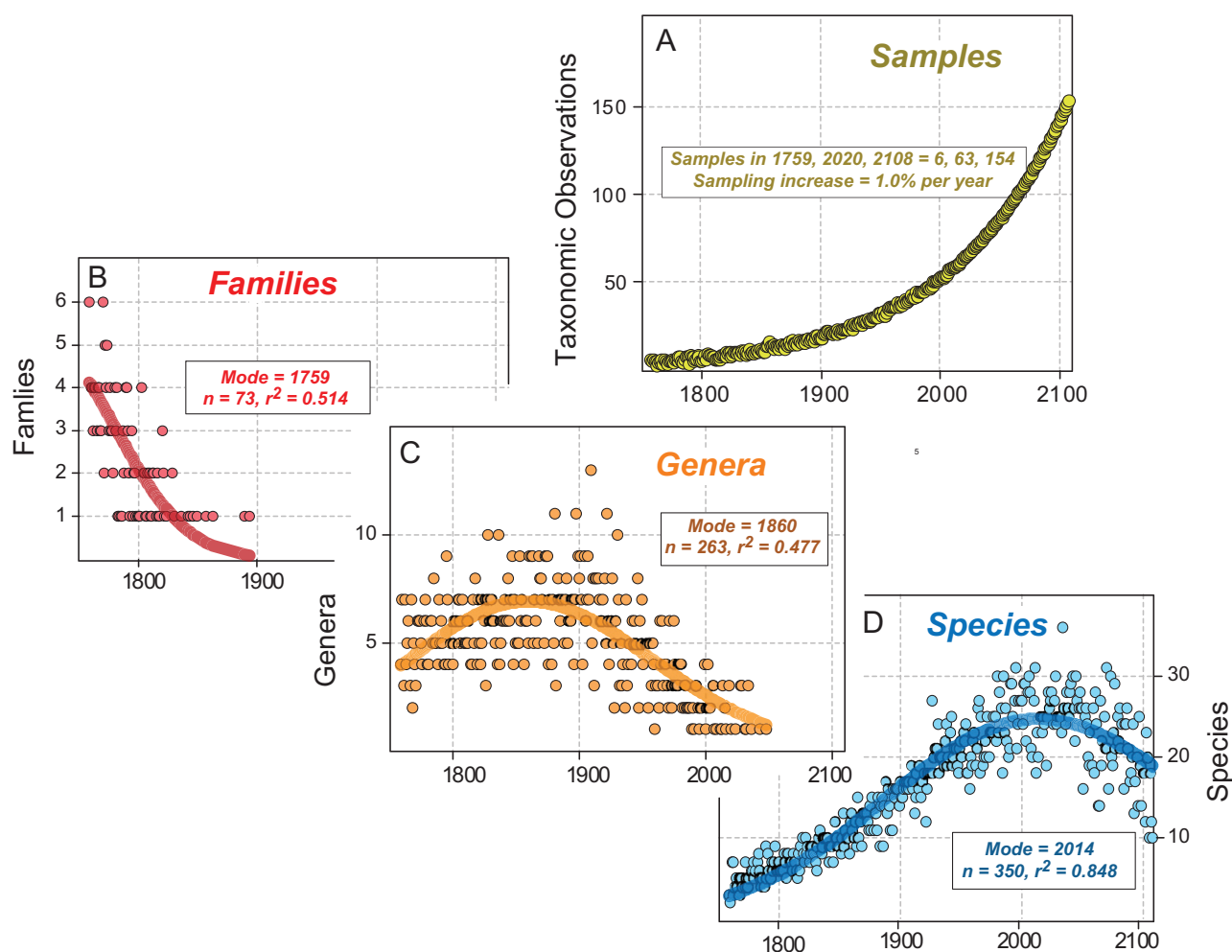
Figure 9. Modelling differences in taxonomic histories as a function of Linnaean level of classification employing the same sampling regime. A, a scenario of taxonomic effort modelled as an exponential change in per-year 'attempts' at taxon description, increasing from six per year in 1758 to 154 per year in 2108. B–D, rates of 'discovery' of new taxa as a function of their memberships at different taxonomic levels.

observations constituting new species are apportioned into genera and families based on the inclusion parameters (**p**) for modern mammals (Table S3). We plot the first model year in which each species, genus and family is first 'discovered', yielding histories of discovery over 350 years at each of those taxonomic levels (Fig. 9B–D).

Such modelled taxonomic histories are qualitatively similar to those observed among all vertebrate groups: higher levels (e.g. families) are completely censused within a few decades, while rates of 'discovery' at lower levels (e.g. genera and species) initially increase because rates of sampling exceed rates of pool depletion, reach a peak rate of discovery (in model year 1860 for genera, year 2014 for species), and then decline as rates of pool depletion overtake rates of sampling. In this example, 100% (170) of families and 99% (1355) of genera, but

only 76% (4965) of species, are discovered by the year 2020. By model year 2108, a time span of 350 years, 98% (6393) of the prescribed 6500 species have been discovered. The lower the Linnaean taxonomic rank, the later the date of 'peak taxonomy'.

The second model compares the effect of varying the rate of increase in sampling on the recovered taxonomic history at a single taxonomic level, in this case genera. A total of 15 000 taxonomic observations are drawn with replacement from a pool comprising 6500 species, and the membership inclusion parameter (**p**; 0.57 per species) comes from modern mammals. We calculate the number of new genera discovered/described each year given each of three different rates of increase of sampling intensity (Fig. 10). The results demonstrate that taxonomic effort has a first-order influence on resultant taxonomic histories for a given
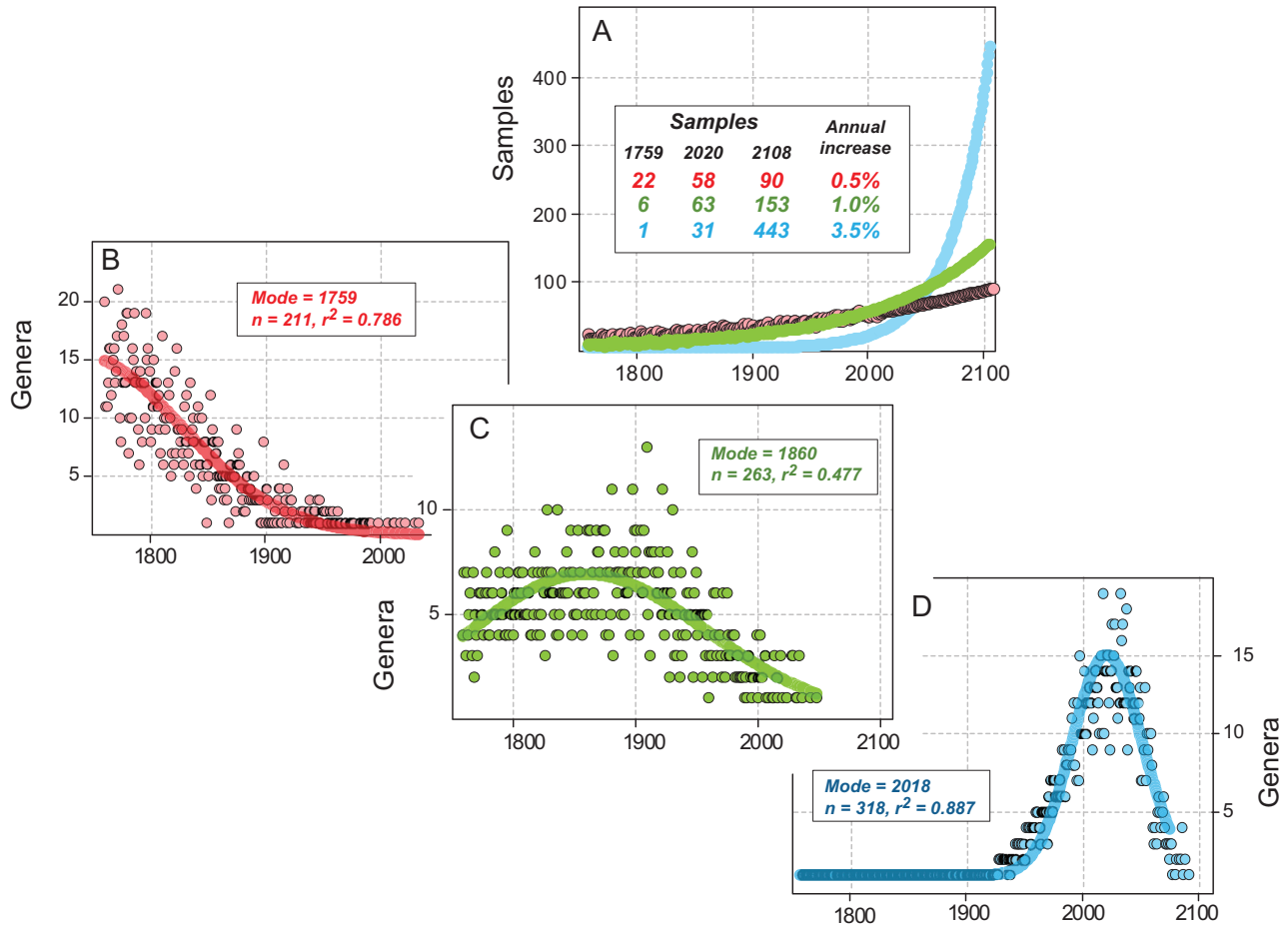
Figure 10. Differences in simulated taxon discovery histories as a function of sampling intensity of subtaxa. A, three different scenarios of taxonomic effort modelled as exponential increases in number of subtaxon descriptions/year resulting from sampling a pool of 6500 subtaxa. B–D, rates of 'discovery' of new taxa as a function of the three subtaxon sampling scenarios in A.

hierarchical level. Low rates of increase in sampling intensity (i.e. closer to the initial scenario of constant sampling intensity over time) result in progressive, approximately exponential depletion of the pool of as-yet undescribed genera, while higher rates of increase in taxonomic effort yield progressively more bell-shaped 'discovery' histories. The model year of peak rate of discovery marks the time that the impact of increasing sampling intensity is overtaken by the influence of depletion of the pool of undiscovered genera. With higher rates of increase in taxonomic effort, 'peak taxonomy' occurs at progressively later dates.

Finally, recall that the rate of description of new species in nearly all vertebrate groups experienced a renewed and exponential rise following *c.* 1950 (Fig. 1D; Supporting Information, S1). Because (presumably) finite pools of vertebrate species are increasingly depleted with each new discovery, changes in sign of the slope of rates of description

must reflect synchronous and significant changes in the amount of taxonomic effort. We examine this supposition with a model of taxonomic history similar to that described above but now imposing two changes in sampling effort over the 350 model years. Both cases comprise a net taxonomic 'effort' of 30 000 observations. In the first instance (Fig. 11A, B), the pool contains 6500 species. We make observations at an initial rate of four per year in 1758 increasing to 35 per year in 1840, holding steady at that rate until 1950, and then increasing up to ~280 per year in 2108 (Fig. 11A). This scenario results in the 'discovery' of ~5000 species (~76%) by 2020 and ~6400 (~98%, nearly all) by 2108 (Fig. 11B). In the second instance, we again make 30 000 total observations, but now drawn from a pool containing 20 000 species. We make observations at an initial rate of ten per year in 1758 increasing to 40 per year in 1840, then decreasing in rate to one per year in
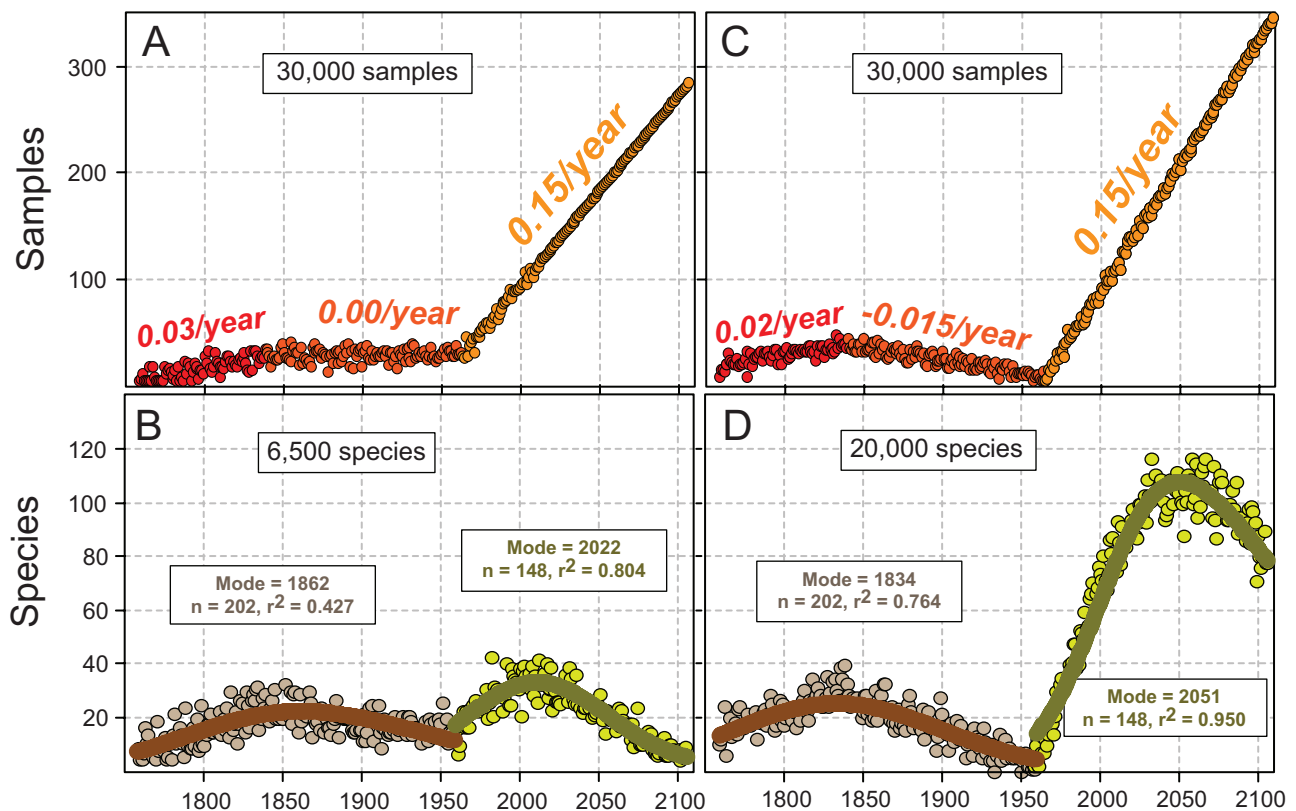
Figure 11. Similarities and differences in 350-year taxonomic histories resulting from slight differences in taxonomic effort. A, hypothetical tempo of effort as two changes in rate of sampling (+0.03/year to 1840; 0.00/year to 1965; +0.15/year to 2107). B, rates of recognition when drawing samples from a pool of 6500 hypothetical species; 76% are sampled by 2020, 98% by 2107. C, hypothetical effort as two changes in rates of sampling (+0.02/year to 1840; −0.015/year to 1965; +0.15/year to 2107). D, rates of recognition of when drawing samples from a pool of 20 000 hypothetical species; 35% are sampled by 2020, 77% by 2107. Although the latter scenario (C and D) represents about three times the biodiversity, rates of recognition before ~2020 (brown circles and lines in B and D) are nearly the same.

1950, and then increasing to ~350 per year in 2108 (Fig. 11C). This scenario results in the 'discovery' of ~7000 species (~35%) by 2020 and ~16 000 (only ~78%) by 2108 (Fig. 11D).

The rate of species 'discovery' in either case comprises two cycles in which rates of sampling first exceed rates of pool depletion, then are overtaken by it. More importantly, despite the fact that total numbers of species in the latter case is about three times that in the former, histories of taxonomic discovery are very similar up to ~2020, and comparable to those actually observed for classes of vertebrates. That small differences in taxonomic effort can yield a high degree of similarity in realized histories of taxonomy, yet might represent grossly different numbers of total extant species, casts doubt on our ability to accurately estimate total standing biodiversity on the basis of currently known tabulations.

TAXONOMY AND HUBBERT'S PEAK – PARALLELS WITH RESOURCE PRODUCTION

The histories of discovery and description of new taxa embody philosophic and practical similarities to processes involved in the discovery and exploitation of natural resources such as petroleum or mineral deposits. In both instances, rate of recovery is dependent on the inherently finite sizes of reservoirs to be exploited and the effort devoted to exploitation. The production of a barrel of petroleum, extraction of a tonne of ore and the description of a new taxon are all exercises in the successful discovery/sampling of some resource reservoir that is being depleted, and histories of taxonomic description and resource exploitation therefore share a number of similarities. Memberships of taxonomic units, volumes of petroleum reservoirs and tonnages of ore deposits exhibit similar 'many small – few large' size frequency distributions

(e.g. Laherrere, 2000). A greater likelihood of sampling from the larger entities results in their earlier dates of discovery in comparison to small/monospecific entities.

Exponential decreases in the rate of description of higher Linnaean levels over time reflect the early depletion of resource pools; any subsequent increase in discovery or production requires substantial renewed effort. In each case, rates of exploitation/discovery have been a manifestation of the 'successful sampling' of these reservoirs; increases occurred during intervals when rates of sampling exceeded rates of reservoir depletion, while times of decrease represent intervals when rates of reservoir depletion exceed effort at exploitation. Following the first oil well drilled near Titusville, Pennsylvania, in 1859, ever-increasing efforts at exploitation led to a more-or-less exponential increase in production until the mid-1970s, when production peaked, followed by a subsequent drop-off (Fig. 12A), much like the history of species descriptions. 'Hubbert's peaks' in both records demark the apices of exploitation (Hubbert, 1956).

Because resource exploitation curves can be described mathematically, one is then tempted to extrapolate them out to their presumed time of depletion in the future, thereby forecasting the amount remaining in the reservoir at any given time and the total cumulative reservoir size. All else being equal, this is not unreasonable. However, following decades of decline, both histories exhibit an abrupt change in slope: oil production has increased rapidly since ~2010 and description of vertebrate species has increased rapidly since ~1950 (Fig. 12). In both cases, the central reasons for the turnaround include an abrupt increase in effort at exploration, the application of new technologies and the subsequent 'discovery' of new large reservoirs that were not available for earlier exploitation. In the case of petroleum production, growth since *c.* 2010 has been almost entirely due to the increasing exploitation of 'unconventional oil' from coal, tar sands, biofuels and primarily shale, made possible by the technological advances of horizontal drilling and hydraulic fracturing, as opposed to 'conventional oil' produced by drilling a vertical hole in the ground. The striking similarities in these otherwise unrelated histories of reservoir exploration suggest that lessons learned from one will apply to the other. In both cases, attempted forecasts for when the reservoir would be fully explored, and hence how big it is in total, based on the documented histories of sampling would have been very different prior to the renewed sampling enabled by the – largely unpredictable – respective developments in each field. We consider such efforts in more detail below.

TAXONOMIC TEMPO AND THE LIMITS OF MEASURING BIODIVERSITY

Earth's biodiversity is currently decreasing at rates that many have compared to the mass extinctions in the geological past (e.g. Pimm *et al.*, 1995). Our collective understanding of the scope and magnitude of depletion would therefore be improved by the knowledge of how many species actually exist within groups (e.g. Brito, 2010; Braje & Erlandson, 2013). The phrase 'how many species' appears in the titles
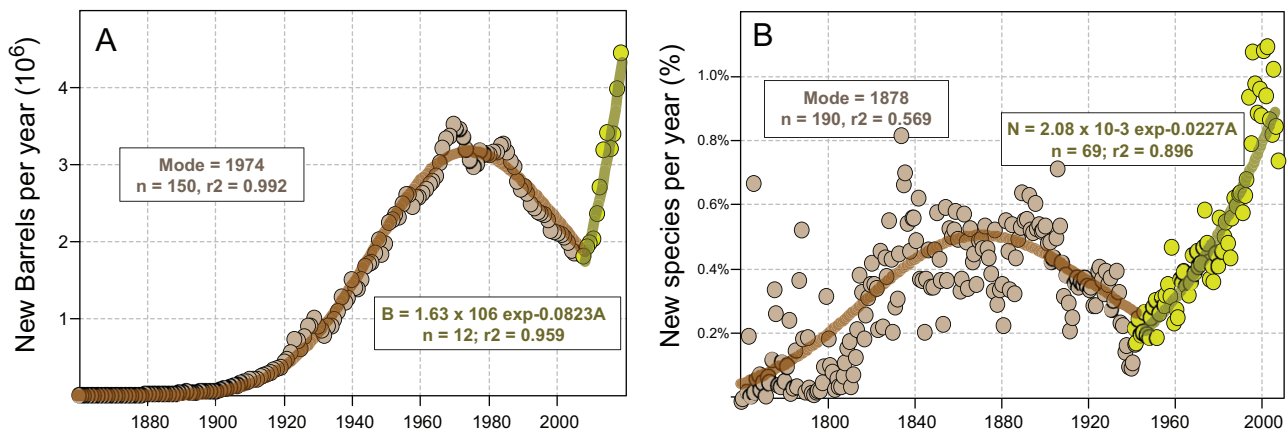

Figure 12. Rates of production of crude oil and naming of vertebrate organisms. A, US field production of crude oil from 1858 to 2019. Skewed normal distribution to interval up to 2007 peaks at *c.* 1974; exponential distribution between 2007 and 2019 reflects an increase of 8.2% per year. B, proportional naming of all new species of vertebrate organisms between 1760 and 2020. Skewed normal distribution between 1760 and 1950 peaks at *c.* 1878; exponential distribution between 1950 and 2020 defined by an annual increase of 2.23%. Abrupt changes in slope in both oil and taxonomic curves in part reflect changes in the nature of resource reservoirs being sampled.

of nearly 200 refereed publications over the past 50 years, including classics from Robert May (1988, 1990). Despite this interest and increasing attention (the appearance of this phrase in titles is currently increasing by ~7%/year), a wide range of opinion exists with respect to extant species richness among different groups of organisms.

Direct observation at the species level for all taxa is impractical to impossible, and so approaches to estimating biodiversity must rely to varying degrees on the taxonomic literature (e.g. Wang & Dodson, 2006; Bebber *et al.*, 2007; Mora *et al.*, 2008, 2011; Nabout *et al.*, 2013; Hopkins, 2019). The history of taxonomic practice and research effort therefore bear directly on efforts to understand the richness of Earth's biota. The aforementioned exercises, however, provide four cautionary notes: (1) while Linnaean species are the currency of biodiversity, the practical means by which to recognize and discriminate among biological species have changed over time; (2) memberships of Linnaean groups exhibit heavy-tailed, fractal-like size frequency distributions, suggesting that a complete census is intractable; (3) ratios of memberships at different Linnaean ranks change over time, making predictions about one level from another unstable; and (4) accurate estimation of as-yet undiscovered biodiversity requires a full understanding of the intensity of scientific industry that produced current tallies – without adequate metrics to quantify past taxonomic effort and project it forward, forecasts of biological diversity based solely on the record of past successes are unconstrained.

Estimating biodiversity is much the same exercise as the tallying of numbers of component members of any existing reservoir. The validity of any forecasting of total amount (or amount remaining) based on past history of exploitation depends upon consistency in the entities under consideration and the mechanism(s) through which they are being discovered. The resurgence in rates of description of vertebrate species and production of oil signify a significant change in the nature of the reservoir and how it is exploited. Although such innovation enhances our collective understanding of the systems under consideration and our ability to exploit them, it greatly compromises the use of past history as a tool for forecasting future rates of description and estimation of total reservoir size, be it measured in barrels of oil or species of vertebrates.

Likewise, estimates of reservoir size based on past rates of production are also dependent upon the richness and evenness of component units included in that reservoir. As we have seen, size distributions of oil fields, ore deposits and taxa are 'heavy tailed' in that many components are very small and few are very large, and large units are likely to be discovered earlier. The high proportion of monotypic genera (91%)

early in the history of taxonomy has decreased to about 47% today after two and a half centuries of discovery (Supporting Information, Figs S7, S8, Table S6), but even if the number of currently described species doubles in the future (e.g. birds; Barrowclough *et al.*, 2016), the degree of monotypy among all vertebrate genera will still be on the order of 40% (Fig. S7). Most of the species remaining to be discovered belong to small genera. Such a high degree of monotypy, combined with the realization that many will also be rare and/or geographically restricted in their distributions, portends ever-decreasing probabilities of future discovery. We will never get them all.

Mora *et al.* (2011) argue, as have others, that because memberships of lower and higher taxonomic units are highly correlated within a group, an estimation of the total number of species is possible from the more complete censuses available at higher taxonomic levels. However, we show above that the relatively predictable relationships between the memberships of different taxonomic levels observed today are themselves time variables (Supporting Information, Fig. S8). Because those ratios evolve over time with continued discovery, forecasts of total numbers of lower level taxonomic units based on them are time dependent as well.

Despite these challenges, estimating total numbers of species by extrapolating from past tempos of species discovery appears tantalizingly feasible, and approaches often employ plots of cumulative number of taxa vs. date of description (e.g. Bebber *et al.*, 2007; Mora *et al.*, 2008; Woolhouse *et al.*, 2012). Such 'discovery' or 'accumulation' curves are typically sigmoidal in form, and are well approximated by several numerical functions (e.g. logistic, binomial Gompertz, extreme value) that allow for an estimate of the maximum membership value as cumulative discoveries approach some asymptotic level. This approach, however, is only applicable for well-sampled groups where the shape of the curve is more fully realized. In addition, the derivatives of such curves are no more than rates of taxonomic description, and therefore afford no additional leverage to the estimation of biodiversity than some cumulative of rate data themselves. As noted above, rates of discovery are dependent on both the biodiversity that exists to be measured as well as the effort of taxonomists to discover and describe new forms, and both of these are unconstrained.

Taking fishes as an example among the vertebrates, we might consider what extrapolation of a cumulative discovery curve would look like if we were to have carried out such an analysis in 1950 vs. in 2020. As of 1950, a little more than 18 000 species of living fishes had been documented, and rates of description had peaked in the early 1900s (Fig. 13A). A best-fit discovery curve to data available at that time yields an asymptote
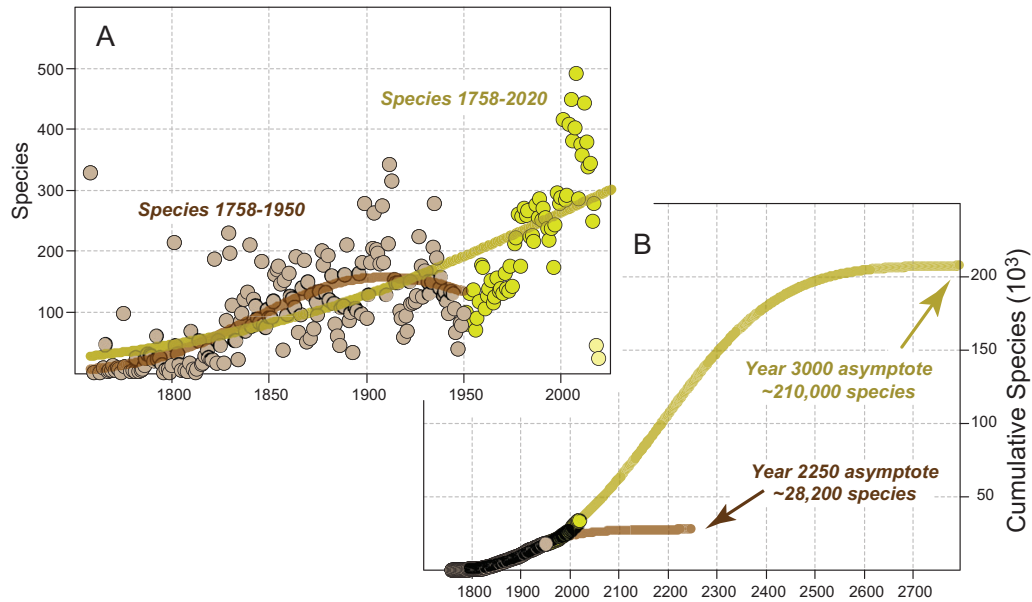
Figure 13. Two forecasts of fish diversity based on an abbreviated (1758–1950) and current (1758–2018) census. A, best-fit skewed normal distributions to rates of species descriptions based on data available up to 1950. Low numbers for 2018 and 2019 in both plots (lighter points) represent incomplete tabulations. B, best-fit cumulative species numbers (collector's curves) based on current data. Greater taxonomic effort as a result of better support of taxonomic research beginning in about 1950, in conjunction with the subsequent increased exploitation of molecular studies as a result of amplified of DNA-based technologies, results in a ten-fold increase in the forecasted size of reservoir of fish species.

suggesting a cumulative biodiversity of about 28 200 species, nearly fully realized by around 2100 (Fig. 13B). Rates of description then surged; ~34 000 fish species are presently recognized (already more than the total predicted pre-1950), and the rate of discovery is still increasing. A single, best-fit discovery curve spanning the entire history of this taxonomy (the shape of which is far more tenuous because less of the sigmoidal form is constrained by data) yields an asymptote at about 210 000 species, roughly six times the current total (Fig. 13B), and this census is not achieved until around the year 2700. Such analyses of the other classes of vertebrates yield similar results. How realistic those predicted biodiversities are is unknowable, as they rest entirely on the risky assumption that effort remains consistent over time (and that enough of the cumulative curve is constrained by the observation to correctly approximate it).

The main lesson to be learned is that estimates of total biodiversity from taxonomic discovery curves (like total reservoir size from oil production curves) are only valid if rates can somehow be forecast into the future with confidence. However, the tempos of taxonomy are based on both the total biodiversity of the reservoir being sampled and the intensity of past and future sampling efforts. The 1950s' inflection

demonstrates that predictions about future sampling intensity can vary widely and be sorely incorrect. While the number of species that exist on Earth remains an alluring statistic for biologists and the public alike, in the absence of constraints on effort, data on past rates of discovery alone are insufficient for estimating total biodiversity. As articulated by Pimm & Joppa (2015), we can only achieve an accurate estimate of the size of the global biodiversity reservoir when we can also accurately forecast numbers, practices and collective efforts of systematists.

REFERENCES

Anderson S. 1974. Patterns of faunal evolution. *Quarterly Review of Biology* **49:** 311–332.

Appeltans W, **et al**. **2012.** The magnitude of global marine species diversity. *Current Biology* **22:** 2189–2202.

Barrowclough GVF, **Cracraft J**, **Klicka K**, **Zink RM. 2016.** How many kinds of birds are there and why does it matter? *PLoS ONE* **11:** e0166307.

Barton IF. 2019. Trends in the discovery of new minerals over the last century. *American Mineralogist* **104:** 641–651.

Bebber DP, **Marriott FHC**, **Gaston KJ**, **Harris SA**, **Scotland RW. 2007.** Predicting unknown species numbers using discovery curves. *Proceedings of the Royal Society B* **274:** 1651–1658.

Bebber DP, **Wood JR**, **Barker C**, **Scotland RW. 2014.** Author inflation masks global capacity for species discovery in flowering plants. *New Phytologist* **201:** 700–706.

Bickford D, **Lohman DJ**, **Sodhi NS**, **Ng PK**, **Meier R**, **Winker K**, **Ingram KK**, **Das I. 2007.** Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution* **22:** 148–155.

Bouchet P, **Bary S**, **Héros V**, **Marani G. 2016.** How many species of molluscs are there in the world's oceans, and who is going to describe them? In: Héros V, Strong E, Bouchet P, eds. *Tropical deep-sea benthos 29.* Paris: Muséum National d'Histoire Naturelle, 9–24 (Mémoires du Muséum national d'Histoire naturelle; 208).

Braje TJ, **Erlandson JM. 2013.** Human acceleration of animal and plant extinctions: a late Pleistocene, Holocene, and Anthropocene continuum. *Anthropocene* **4:** 14–23.

Brito D. 2010. Overcoming the Linnaean shortfall: data deficiency and biological survey priorities. *Basic and Applied Ecology* **11:** 709–713.

Ceballos G, **Ehrlich PR**, **Dirzo R. 2017.** Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences USA* **114:** E6089–E6096.

Ceballos G, **Ehrlich PR**, **Raven PH. 2020.** Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proceedings of the National Academy of Sciences USA* **117:** 13596–13602.

Chu J, **Adami C. 1999.** A simple explanation for taxon abundance patterns. *Proceedings of the National Academy of Sciences USA* **96:** 15017–15019.

Collen B, **Purvis A**, **Gittleman JL. 2004.** Biological correlates of description date in carnivores and primates. *Global Ecology Biogeography* **13:** 459–467.

Costello MJ, **Wilson S**, **Houlding B. 2012.** Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology* **61:** 871–883.

Costello MJ, **Houlding B**, **Joppa LN. 2014.** Further evidence of more taxonomists discovering new species, and that most species have been named: response to Bebber *et al*. (2014). *New Phytologist* **202:** 739–740.

Cribb TH, **Bray RA. 2011.** Trematode families and genera: have we found them all? *Trends in Parasitology* **27:** 149–154.

Dawson MR, **Marivaux L**, **Li C**, **Beard KC**, **Métais G. 2006.** *Laonastes* and the 'Lazarus effect' in Recent mammals. *Science* **311:** 1456–1458.

Donoghue MJ, **Alverson WS. 2000.** A new age of discovery. *Annals of the Missouri Botanical Garden* **87:** 110–126.

Froese R, **Pauly D. 2019.** *FishBase, World Wide Web electronic publication*. Available at: www.fishbase.org.

Gaston KJ. 1991. Body size and probability of description – the beetle fauna of Britain. *Ecological Entomology* **16:** 505–508.

Gill F, **Donsker D**, **Rasmussen P. 2020.** *IOC world bird list (v10.2)*. Available at: http://www.worldbirdnames.org/new/.

Gómez-Daglio L, **Dawson MN. 2019.** Integrative taxonomy: ghosts of past, present and future. *Journal of the Marine Biological Association of the United Kingdom* **99:** 237–46.

Grieneisen ML, **Zhang Y**, **Potter D**, **Zhang M. 2012.** Biodiversity, taxonomic infrastructure, international collaboration, and new species discovery. *BioScience* **64:** 322–332.

Hopkins JG. 2019. Are we close to knowing the plant diversity of the Amazon? *Anais da Academia Brasileira de Ciências* **91:** 3.

Hubbert MK. 1956. Nuclear Energy and the Fossil Fuels. Publication No. 95, Shell Development Company, Houston, June 1–57. Available at: http://www.hubbertpeale.chubbert/1956/1956.pdf

Ioannidis JPA, **Klavans R**, **Boyac KW. 2018.** Thousands of scientists publish a paper every five days. *Nature* **561:** 167–169.

Jenkins PD, **Kilpatrick CW**, **Robinson MF**, **Timmins RJ. 2005.** Morphological and molecular investigations of a new family, genus and species of rodent (Mammalia: Rodentia: Hystricognatha) from Lao PDR. *Systematics and Biodiversity* **2:** 419–454.

Joppa LN, **Roberts DL**, **Pimm SL. 2011a.** How many species of flowering plants are there? *Proceedings of the Royal Society B* **278:** 554–559.

Joppa LN, **Roberts DL**, **Pimm SL. 2011b.** The population ecology and social behavior of taxonomists. *Trends in Ecology and Evolution* **26:** 551–553,

Krasnov BR, **Shenbrot GI**, **Mouillot D**, **Khokhlova IS**, **Poulinet R. 2005.** What are the factors determining the probability of discovering a flea species (Siphonaptera)? *Parasitology Research* **97:** 228–237.

Ladner JT, **Palumbi SR. 2012.** Extensive sympatry, cryptic diversity and introgression throughout the geographic distribution of two coral species complexes. *Molecular Ecology* **21:** 2224–2238.

Laherrere J. 2000. Distribution of field sizes in a petroleum system: parabolic fractal, lognormal or stretched exponential? *Marine and Petroleum Geology* **17:** 539–546.

Linnaeus C. 1758. *Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. 1, 10th edn*. 1–824.

Lynch MC. 2010. The new pessimism about petroleum resources: debunking the Hubbert model (and Hubbert Modelers). *Minerals & Energy – Raw Materials Report* **18:** 21–32.

Mammal Diversity Database. 2020. *Mammal Diversity Database (Version 1.2).* Zenodo. Available at: http://doi.org/10.5281/zenodo.4139818.

May RM. 1988. How many species are there on earth? *Science* **241:** 1441–1449.

May RM. 1990. How many species? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **330:** 293–304.

Myers N, **Mittermeier RA**, **Mittermeier CG**, **daFonseca GAB**, **Kent J. 2000.** Biodiversity hotspots for conservation priorities. *Nature* **403:** 853–858.

Miya M, **Holcroft NI**, **Satoh TP**, **Yamaguchi M**, **Nishida M**, **Wiley EO. 2007.** Mitochondrial genome and a nuclear gene indicate a novel phylogenetic position of deep-sea tube-eye fish (Stylephoridae). *Ichthyological Research* **54:** 323–332.

Mora C, **Tittensor DP**, **Myers RA. 2008.** The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **275:** 149–155.

Mora C, **Tittensor DP**, **Adl S**, **Simpson AGN**, **Worm B. 2011.** How many species are there on earth and in the ocean? *PLoS Biology* **9:** e1001127.

Moura MR, **Jetz W. 2021.** Shortfalls and opportunities in terrestrial vertebrate species discovery. *Nature Ecology & Evolution* **5:** 631–639.

Nabout CJ, **da Silva Rocha B**, **Carneiro FM**, **Sant'Anna CL. 2013.** How many species of Cyanobacteria are there? Using a discovery curve to predict the species number. *Biodiversity and Conservation* **22:** 2907–2918.

Pfenninger M, **Schwenk K. 2007.** Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology* **7:** 1–6.

Pimm SL, **Joppa LN. 2015.** How many plant species are there, where are they, and at what rate are they going extinct? *Annals of the Missouri Botanical Garden* **100:** 170–176.

Pimm SL, **Russell GJ**, **Gittleman JL**, **Brooks TM. 1995.** The future of biodiversity. *Science* **269:** 347–350.

Reddingius J. 1971. Gambling for existence: a discussion of some theoretical problems in animal population ecology. *Acta Biotheoretica* **20:** 3–208.

Scotland RW, **Sanderson MJ. 2004.** The significance of few versus many in the tree of life. *Science* **303:** 643.

Thompson MSA, **Couce E**, **Webb TJ**, **Grace M**, **Cooper KM**, **Schratzberger M. 2021.** What's hot and what's not: making sense of biodiversity 'hotspots'. *Global Change Biology* **27:** 521–535.

Uetz P, **Freed P**, **Hošek J. 2020.** *The reptile database.* Available at: http://www.reptile-database.org.

Wang SC, **Dodson P. 2006.** Estimating the diversity of dinosaurs. *Proceedings of the National Academy of Sciences USA* **37:** 13601–13605.

Walters SM. 1961. The shaping of angiosperm taxonomy. *New Phytologist* **60:** 74–84.

Wilkinson BH. 2011. On taxonomic membership. *Palaeobiology* **37:** 519–536.

Willis JC. 1922. *Age and area: a study in geographical distribution and origin of species.* Cambridge: Cambridge University Press.

Woolhouse MEJ, **Scott F**, **Hudson Z**, **Howey R**, **Chase-Topping M. 2012.** Human viruses: discovery and emergence. *Proceedings of the Royal Society B* **367:** 2864–2871.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Rates of naming of groups of living finfishes (A–D), amphibians (E–G), reptiles (H–J), and birds (K–M). Data as filled circles; general trends are best-fit skewed normal distributions except post-1950 species abundances (D, G, J, M) which are a best-fit exponentials (data for mammals as Figure 1). See Table S3 for a list of best-fit parameters.
Figure S2. Relationships between the first year of a currently recognized species (*x*-axes) and the numbers of species in that group (*y*-axes) for taxonomic data on living finfishes (A–D), amphibians (E–F), reptiles (G–H) and birds (L–L). Number of members (species) largely predicates the probability of group (order, family, genus) recognition. Data for mammals as Figure 2.
Figure S3. Stochastic model of group memberships for fishes (A–F), amphibians (G–I), reptiles (J–L), and birds (M-O) plotted as number of subtaxa (*y*-axes) among each hierarchically higher level of supertaxa (*x*-axes). Metrics for each membership curve are listed in Figure S4. Data for mammals as Figure 4.
Figure S4. Relationships between rates of naming of new species and the numbers of papers with taxonomic data on living amphibians (A, B), reptiles (C, D), birds (E, F), and mammals (G, H). Number of new species and associated papers are largely correlated. Data for finfishes as Figure 6.
Figure S5. Rates of description of new species of finfishes (A), amphibians (C), reptiles (E), birds (G) and mammals (I). Best-fit skewed normal distributions (brown lines) to per-year rates (brown dots) suggest a maximum taxonomic 'efficiency' in the early 1800s. In all cases, note decreasing scatter of rates with decreasing age. Exponential decreases (B, D, F, and H) in this 'monograph effect' manifest as diminishing differences between observed rates

of description of new species and that described by a longer term average. Average 'spikiness' in description of new species has decreased by ~1.2% per year since the mid-16$_{th}$ century. Data for amphibians as Figure 7.

Figure S6. Dates of publication vs. numbers of authors with respect to species of living finfishes (A, B), amphibians (C, D), reptiles (E, F), birds (G, H) and mammals (I, J). Upper left panels plot proportions of single and dual authored papers (red points and lines); lower right panels plot paper data by number of contributing authors; heavier blue points and connecting lines are median ages. For all groups, note an abrupt increase in multi-authored papers about 1950.

Figure S7. Relationships between percentage of monotypic genera and numbers of described species of vertebrate organisms over the past 262 years (1758–2020). All classes exhibit decreasing degrees of monotypy as numbers of described species have increased. Percentages of monotypic genera are closely approximated with logarithmic law functions (green lines) throughout their histories (Table S7). Reference years as lighter green labelled circles.

Figure S8. Historical changes in ratios of genera per family (orange points and lines) and species per genus (green points and lines) among vertebrate species over the past 262 years (1758–2020). Reference years labelled dark blue.

Table S1. Sources of taxonomic data and numbers of taxa, papers and authors.

Table S2. Parameters for best-fit trends to taxonomic group dates and numbers of members.

Table S3. Parameters of taxonomic memberships.

Table S4. Parameters for best-fit trends to numbers of publications.

Table S5. Parameters for best-fit trends to dates of species descriptions and monograph effects.

Table S6. Parameters for best-fit logarithmic trends to monotypes per genus.