Biodiversity hotspots house most undiscovered plant species

Lucas N. Joppa[a,b,c], David L. Roberts[b,c], Norman Myers[d,1], and Stuart L. Pimm[e]

[a]Microsoft Research, Cambridge CB3 0FB, United Kingdom; [b]Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, Canterbury CT2 7NR, United Kingdom; [c]Royal Botanic Gardens, Kew TW9 3AB, United Kingdom; [d]Green College, Oxford University, Oxford OX2 6HG, United Kingdom; and [e]Nicholas School of the Environment, Duke University, Durham, NC 27708

Contributed by Norman Myers, June 10, 2011 (sent for review April 6, 2011)

For most organisms, the number of described species considerably underestimates how many exist. This is itself a problem and causes secondary complications given present high rates of species extinction. Known numbers of flowering plants form the basis of biodiversity "hotspots"—places where high levels of endemism and habitat loss coincide to produce high extinction rates. How different would conservation priorities be if the catalog were complete? Approximately 15% more species of flowering plant are likely still undiscovered. They are almost certainly rare, and depending on where they live, suffer high risks of extinction from habitat loss and global climate disruption. By using a model that incorporates taxonomic effort over time, regions predicted to contain large numbers of undiscovered species are already conservation priorities. Our results leave global conservation priorities more or less intact, but suggest considerably higher levels of species imperilment than previously acknowledged.

global priorities | species discovery | angiosperm

Most species are not known to science (1–3); henceforth we call these "missing" species. Where they live is vital in setting international priorities for conservation, a conclusion that follows from basic generalities of biogeography (4). Sizes of species' geographical ranges are highly skewed, with many species having small ranges relative to the mean (5). Moreover, small-ranged species are geographically concentrated (5) and perversely mostly located in areas with disproportionately high levels of habitat destruction and human population growth (6, 7). Sufficiently small geographical range combined with an occurrence in areas of extensive habitat loss classifies a species as being "threatened" under the International Union for Conservation of Nature Red List criteria (8). These features yield the familiar idea of biodiversity hotspots (6): geographical concentrations of threatened species. Myers et al.'s (6) specific definition of them combines a measure of habitat destruction (<30% habitat remaining) and numbers of endemic flowering plant species (>1,500). Hotspots are international priorities for conservation, and significant financial resources have been directed toward them (9). Given the incompleteness of the taxonomic catalog, are there additional areas that we should consider hotspots if we knew more about the number of species within them? And, with complete information, would the rankings of hotspots change? Plant-based hotspots are also important for vertebrate taxa, so in time, we should assess those as well. However, given the global importance of the original identification of biodiversity hotspots, performing this exercise with plants is an essential first step.

Expert opinions and a model of the rates of plant description suggest that approximately 15% of flowering plant species are missing (10, 11). They are surely like almost all recently discovered species, in being locally rare and geographically restricted (10). (Ref. 12 explores a rare exception, while ref. 13 suggests that many of these species have already been collected and are waiting in herbaria for formal taxonomic description.) These facts lead to the obvious questions we address: will knowing where the missing species reside change the way we set conservation priorities? Will relative priorities change as taxonomists complete the catalog? Will new priorities become apparent? Are the missing species in places where they are likely to be threatened, and indeed, will we discover them before they become extinct?

Estimating Missing Species

The original hotspots of Myers et al. (6) were based on the number of vascular plants endemic to a region and the extent of regional habitat destruction. Currently, there are estimated to be ~350,000 species of vascular plants, of which 96% are flowering plants (14). Working with only flowering plants, which includes the vast majority of vascular plants, therefore does not bias our analysis in regard to the original implementation of the hotspots idea.

Estimates of the numbers of missing species encounter two large problems. First, taxonomists inadvertently give different names to the same species. We avoid this issue by using the World Checklist of Selected Plant Families (WCSP; http://www.kew.org/wcsp) (14), a unique and continuously updated synonymised world list of plants. The taxonomy has been reviewed, but for only some plant families: all monocots (approximately 60,000 species, excluding grasses) and approximately 50,000 species of nonmonocots. For these species, the database provides location data in the form of 368 Taxonomic Database Working Group (TDWG; http://www.nhm.ac.uk/hosted_sites/tdwg/geogrphy.html) (15) regions in which the species occur. These range from individual states within the United States, to regions within countries (especially large ones, such as China), to countries themselves. An additional approximately 10,000 grass species are taxonomically revised (16), but unlike the WCSP, the database does not provide immediately accessible location data. We do not include them in our analysis.

We acknowledge that the species concept has, for some taxa, changed considerably through time. We work only with species as recognized within the WCSP, a point to which we will return later in the *Discussion*.

The second problem involves the rates of species descriptions. Earlier studies attempted to extrapolate the number of species described over time, with the expectation that the numbers of new species per time interval will decrease as the pool of unknown species diminishes (17–20). Generally, they do not; indeed, the rates of increase are often exponential (10, 11), as examples in Fig. 1, *Left*, demonstrate. The underlying cause is a broadly exponential increase in the numbers of taxonomists over time. (We define taxonomists simply as those who describe species.) Indeed,

Author contributions: L.N.J., D.L.R., N.M., and S.L.P. designed research; L.N.J., D.L.R., and S.L.P. performed research; L.N.J., D.L.R., and S.L.P. contributed new reagents/analytic tools; L.N.J., D.L.R., N.M., and S.L.P. analyzed data; and L.N.J., D.L.R., N.M., and S.L.P. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

See Commentary on page 12971.

[1]To whom correspondence should be addressed. E-mail: myers1n@aol.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1109389108/-/DCSupplemental.

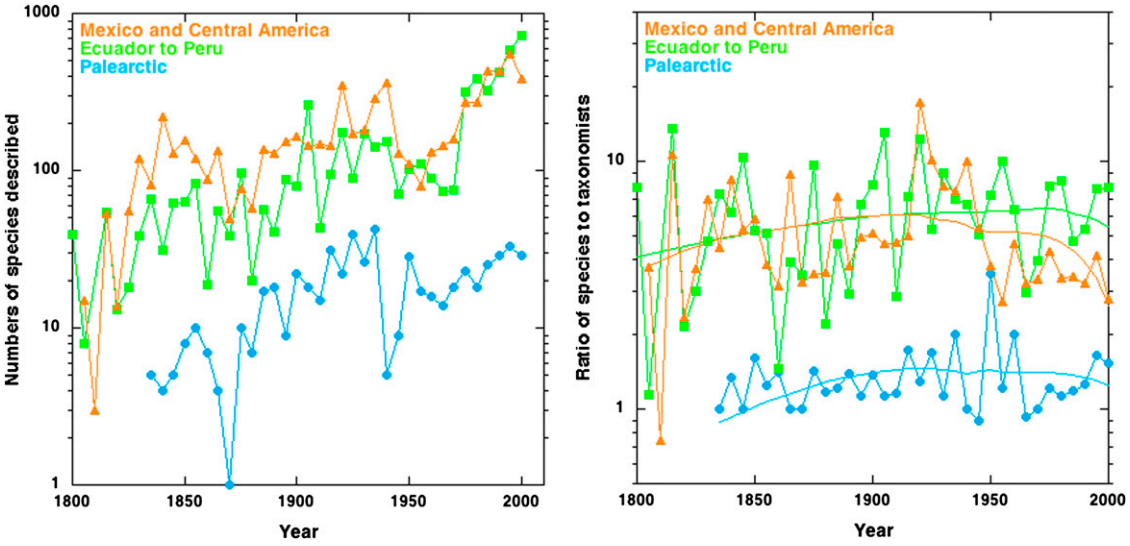

Fig. 1. *Left*: The numbers of species described per 5-y interval for three selected regions increase broadly exponentially over time. *Right*: Corrected for the number of taxonomists involved in their description, most regions are like Mexico and Central America in showing a pronounced decrease in the rate of species described per taxonomist, leading to a modeled increase that averages 15% more species than at present. For both the region from Ecuador to Peru and the Northern Palearctic, however, there is no decrease, suggesting that considerable numbers of species are still missing from the taxonomic catalog. In both figures, plotted data start when the region has accumulated at least 40 endemic species.

the statistical best predictor of numbers of species described in an interval is the number of taxonomists describing them (10). We can imagine several other ways to count taxonomists, which we shall consider in the *Discussion*.

When divided by the number of taxonomists involved in description, the numbers of species described over time generally decreases during the past century. Initially, however, these corrected numbers usually increase, as do all three examples in Fig. 1, *Right*. We attribute this to increasing taxonomic efficiency (10, 11). That could involve many factors, including (*i*) better access to the sometimes remote places where many species occur, (*ii*) increasing access to the literature and specimens housed in herbaria globally, and (*iii*) easier circumscription of taxa as a result of increasing numbers of species with which to compare, i.e., being able "to see more of the puzzle." We assume efficiency to increase linearly over time, and independently verify that assumption in the *SI Text*. (About this general increase in efficiency, there will surely be fluctuations driven by many factors, including the changing fashions of splitting or lumping species.)

Combined, these effects allow a statistical estimate of the total number of species within broad taxonomic groups and individual families (10) or, as here, for geographical subsets of species. Thus:

$$S_i = (a + b\ Y_i)*(T_i)*(S_T - \Sigma S_i), \qquad [1]$$

where S_i is the number of species described per unit time, Y_i the time interval, T_i the number of taxonomists involved in the description, and ΣS_i the total number of species described to that time; a, b, and S_T are constants to be estimated, the last one being the predicted total number of species (10). The model component $(a + b\ Y_i)$ estimates taxonomic efficiency, with the b parameter the slope of the temporally increasing expectation of this relationship. Further details of the model are provided in *Materials and Methods*.

We applied this model to estimate numbers of species remaining to be described in geographical regions (Fig. 1 provides selected examples). With a few exceptions, we modeled species discovery within as many of the original TDWG regions (15) as possible, with a minimum of 500 species endemic to them. Many of the regions met this criterion, and they appear in Table 1 as the names of countries (e.g., Cuba) or regions within countries (e.g., Brazil-north). Elsewhere, we had to group regions. The northern Palaearctic, from Iceland eastward to Siberia and Mongolia, is an example: only by combining such a huge region does it include a net of more than 500 endemic species. One exception to this criterion is the Caucasus, which has 267 endemic species in our selection of species. For a more complete set of species, Myers et al. (6) found it to contain 1,600 endemic species and thus qualify as a hotspot. The other exceptions to our 500-species rule are the regions termed "Guyanas," "West Africa: Senegal to Benin," and "Cambodia, Laos" (Table 1 and Dataset S1). These regions contain fewer than 500 endemic species each, but had floras with relatively little overlap with species elsewhere, so we retained them as separate regions.

Finally, in a few cases, we combined two regions if both regions had unusually large predictions of missing species, to improve the estimates. The details of the composition of the regions used, and how we clustered them, are provided in *Materials and Methods* and Datasets S1 and S2.

In sum, we retained as many regions as possible consistent with obtaining credible estimates of the numbers of missing species. In doing so, we classified approximately 79,000 of the more than 108,000 species into one of the 50 regions in Table 1, with the remainder occurring in two or more regions.

Results

Table 1 summarizes our results, and Fig. 2 maps them. Details of the statistical fits, along with maximum-likelihood confidence intervals (CIs) of the estimates, are provided in *Materials and Methods* and Dataset S1. Overall, in the analysis we present here, we predict 21% of species are missing from our sample of approximately 108,000 species. This estimate is similar to estimates produced for all the monocots (17%) and for a combination of the taxonomically revised nonmonocots (13%) (10).

Forty of 50 regions show a marked reduction over time in the number of species described per taxonomist per 5-y interval. This leads to low ratios of numbers of predicted species over numbers of presently known species (Fig. 1 provides the example for Mexico and Central America, where the ratio is 1.17, meaning we predict an additional 17%, a percentage close to the numbers

Table 1. Summary statistics for regions modeled

Region	Current total, %	Predicted total, %	Ratio	Missing species, %	Rank change
North and Central America, Caribbean					
North America	2.47	2.10	1.09	0.77	↓ 1
Cuba	1.53	1.20	1.00	0.02	↓ 5
Caribbean islands other than Cuba	1.90	1.74	1.17	1.14	↓ 2
Mexico to Panama	9.00	8.24	1.17	5.50	↓ 1
South America					
Colombia*	3.62	4.15	1.46	6.05	↑ 2
Venezuela	1.90	1.78	1.20	1.33	→ 0
Guyanas	0.57	0.52	1.18	0.36	→ 0
Ecuador to Peru*	7.22	11.94	2.11	28.92	↑ 1
Bolivia	1.26	1.19	1.21	0.94	→ 0
Brazil, north	0.94	0.85	1.16	0.54	↓ 2
Brazil, northeast*	1.29	1.65	1.63	2.94	↑ 6
Brazil, south	0.72	0.62	1.10	0.25	→ 0
Brazil, west-central	0.72	0.65	1.15	0.39	↑ 2
Brazil, southeast	4.42	3.61	1.04	0.68	↓ 1
Paraguay, Argentina, Uruguay, Chile*	1.69	2.50	1.89	5.40	↑ 7
Temperate Eurasia					
Palearctic: Iceland to Western Siberia, south to Mongolia*	0.81	1.09	1.72	2.11	↑ 10
European Mediterranean	1.60	1.59	1.27	1.56	→ 0
North Africa and Middle East Mediterranean	0.89	0.90	1.29	0.92	↑ 1
Turkey	0.82	0.74	1.15	0.44	↓ 2
Caucasus	0.34	0.30	1.13	0.16	→ 0
Iraq, Iran, Afghanistan, Pakistan	0.73	0.66	1.15	0.41	↑ 1
Turkmenistan and Kazakhstan east through Tibet to Xinjiang, China	1.16	1.05	1.16	0.68	→ 0
China, south-central	1.94	1.98	1.30	2.13	↓ 1
China, southeast	1.50	1.47	1.25	1.37	→ 0
North-central China, Japan, Korea	0.77	0.65	1.08	0.21	↓ 2
Africa					
North East (Burundi, Ruanda, Uganda, Kenya, Somalia, Ethiopia, Eritrea, Sudan, Djibouti)	1.21	1.13	1.19	0.83	↑ 2
West Africa: Senegal to Benin*	0.46	0.47	1.31	0.52	→ 0
West and Central Africa: Nigeria to Central African Republic and Congo	1.27	1.08	1.08	0.37	↓ 6
Democratic Republic of Congo	0.73	0.67	1.18	0.47	↑ 4
Tanzania*	0.79	1.08	1.76	2.14	↑ 11
Angola, Malawi, Mozambique, Zambia, Zimbabwe*	1.16	1.25	1.37	1.55	↑ 7
Botswana and Namibia southward*	5.47	7.76	1.81	15.96	↑ 2
Madagascar	4.21	3.39	1.03	0.45	↓ 1
Tropical Mainland Asia					
Assam, Bangladesh, Himalayas, Burma, Nepal	1.46	1.26	1.11	0.55	→ 0
Indian, Sri Lanka	1.84	1.51	1.05	0.31	↓ 3
Thailand	0.96	0.78	1.03	0.12	↓ 5
Mainland Malaya	1.23	0.99	1.02	0.10	↓ 5
Cambodia, Laos	0.39	0.33	1.07	0.10	→ 0
Vietnam	1.15	1.11	1.24	0.98	↑ 5
Tropical Asia Islands					
Sumatra	1.25	0.98	1.00	0.00	↓ 7
Java, lesser Sundas	1.21	0.96	1.01	0.03	↓ 5
Borneo	3.30	2.65	1.03	0.31	→ 0
Sulawesi	0.74	0.61	1.05	0.14	↓ 4
Bismarck, Caroline, Marianas, Solomon Islands	0.80	0.67	1.06	0.18	↓ 2
Philippines	2.79	2.18	1.00	0.00	↓ 1
New Caledonia	1.50	1.29	1.10	0.52	→ 0
Fiji, Samoa, Society Islands, Tonga, Vanuatu	0.96	0.87	1.16	0.57	↓ 2
Australasia					
Australia*	7.09	7.29	1.31	8.02	↓ 1
New Guinea	5.74	4.52	1.01	0.14	↓ 1
Miscellaneous oceanic Islands, worldwide	2.45	2.00	1.05	0.40	↓ 1
Total	100.00	100.00	—	100.00	—
Sample size	78,799	100,719	—	21,920	—
Species occurring across multiple regions	29,917	30,784	1.03	867	NA

Regions, fractions of currently known species from a sample of taxonomically revised families of flowering plants, and predictions of those fractions once missing species are included, the ratio of predicted to currently known species numbers, the fractions of missing species in different regions, and the change in rank order of diversity are shown. The main differences are discussed in the text. A full version of the table, which shows raw species numbers, along with parameter estimates and maximum-likelihood CIs on all model parameters, is included as Dataset S1.

*Regions with high ratios, meaning at least 30% of the species remaining to be discovered.

for all species combined). Our model suggests other regions with floras that are relatively well known. These include some biodiversity hotspots including Cuba, southeastern Brazil, India and Sri Lanka, and much of mainland tropical Asia.

Do our estimates change our understanding of the rankings of global biodiversity? Table 1 shows the percentage distribution of known species, as well as for when we include the predicted numbers of species. For ease, we have also added the change in rank. Arrows indicate the direction of this change.

Fig. 2 maps the current (Fig. 2, *Left*) and predicted (Fig. 2, *Right*) overall percentages of endemic species around the world. The similarities between panels in Fig. 2, along with the rank change column in Table 1, indicate that existing conservation priorities would not be changed significantly by correcting for missing species ($r_s = 0.97, P < 0.001$). For example, the northern Andes and Atlantic Coast forests of South America, Southern Africa, Australasia, and the islands of tropical Asia will remain major centers of plant endemism, as well as areas under threat.

That said, there are geographic differences between what is currently known and what we predict. The region including Ecuador and Peru houses more than 5% of the species in our sample and is currently ranked as the second richest single area. However, this region is projected to contain 29% of the world's missing species, and by incorporating this datum, we expect this region to overtake the area currently ranked first (Mexico to Panama). Further, we predict that Tanzania and the southern cone of South American countries will become relatively much richer in species when missing species are included.

Additionally, it is of conservation interest to note that South American regions will, on average, increase their ranking by 16% (based on the percentage of total possible moves in rank) and African regions by 10%. In contrast, the tropical mainland Asian regions will fall in rankings on average by 6% and the tropical Asian islands by more than 10%. We note that, although increasing taxonomic knowledge generally implies increasing numbers of known species, there has been at least one exception to the rule. In the Eastern Arc Mountains, better information on species ranges and synonymies caused a downward revision of regional endemic plant species, and necessitated a regrouping of the hotspot region (21). Generally, correcting for missing species, as we do here, will increase the number of species endemic to a region.

Finally, there are other species not endemic to the regions we describe, although, given the rather large ranges of these taxa, it is likely that the pool of unknown species is quite small. Approximately 25% of the species we analyzed occur in two or more of the regions we define, and, as one would expect, such generally widespread species are well known and the rate of taxonomic description is now low (Table 1). Species endemic to oceanic islands also show only slow rates of discovery, perhaps because such places are well explored and have been for a long time (Table 1).

Discussion

Sources of Uncertainty. As highlighted in the results section, the majority of regions provide sensible estimates of species remaining to be discovered. We are far less certain of the total numbers for the 10 regions with large percentages of missing species (i.e., where the models predict a >30% increase in numbers of species). Fig. 1, *Right*, shows results for two of these regions. For both the northern Palaearctic and the region that includes Ecuador and Peru, not only are the numbers of species described increasing roughly exponentially over time, but also there is only a slight decline in the numbers of species described per taxonomist over time. This lack of an obvious decline leads to predictions that there are many missing species, but how many more is statistically much less certain. This uncertainty is borne out in the estimated CIs for total species in each of these ten regions, which are significantly wider than for the other regions (t test, $P < 0.001$; Dataset S1).

The model itself need not always be appropriate. It is possible that, for regions rich in missing species, the rate-limiting factor is not the number of missing species, but the number of taxonomists available to describe them (9). Those taxonomists may work through the species in a region (or family), sometimes genus by genus, resulting in a more or less constant rate of description until abruptly the supply of missing species is exhausted (10).

Ecuador provides an example of this. In that country alone, there are more than 3,544 species of orchids, including 1,706 in the subtribe Pleurothallidinae. The majority of these epiphytic orchids are tiny, including some of the smallest flowers in the world. They are often restricted in distribution to three or fewer populations within a limited range (indeed, 1,125 of the 1,706 Pleurothallid species in Ecuador are endemic to the country). These characteristics make them difficult to find, whereupon our model of diminishing returns should apply. However, they are also a challenge to identify and describe. One individual, Carl Luer, has been involved in the description of more than 1,000 of them. This example strongly suggests the working habitats of even an individual taxonomist must play the key role in the rate of species description. In sum, we may interpret these high ratios as

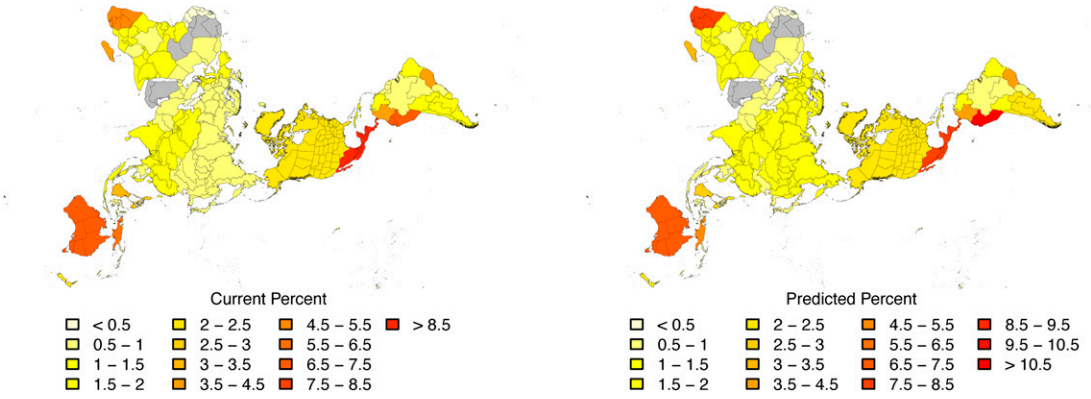

Fig. 2. Currently known patterns of flowering plant species richness as a percent of all species (*Left*) and patterns when corrected for species predicted to be missing from the taxonomic record (*Right*). They are broadly similar; differences are noted in the text. Although our results do not take area into account, we plot the results in an equal-area projection to help visualize the sometimes confusing relationship between undiscovered species and region. We excluded gray regions (Saharan Africa and the Arabian Peninsula) from the analysis as a result of low numbers of endemic species. Original TDWG regions are shown in Fig. S1. Fig. S3 shows raw estimates of numbers of missing species in each region.

representing areas of continued taxonomic activity, even if not precise indications of richness.

Certainly, access to the different areas likely affects these estimates. For example, decades of social unrest have limited exploration of key areas in Colombia, surely distorting the rates of species description. This may explain why the percentage of missing species is much smaller there than in countries immediately to the south. Our predictions of relatively few missing species for New Guinea may reflect the remoteness of much of its forests—it is one of the few endemic-rich areas of the world that is not a hotspot because its forests were still relatively intact when the hotspots were originally created (as detailed later).

Finally, the high predicted numbers for the Palearctic may be a consequence of splitting of very similar species, something likely in only very well known floras and a different process from exploration-driven discovery elsewhere.

There are broader concerns with our results. First, one can readily imagine why counting all the authors involved in a scientific description should not count equally toward a measure of taxonomic effort. There may be temporally changing fashions to count students, assistants in the laboratory or in the field, and those that provide specialized help, for example with genetic analyses. To what extent would these trends change our measure of taxonomic effort? Fig. 3, *Left*, shows, over time, the numbers of all taxonomic authors, the first and last authors, and the numbers of just first authors. The second of these measures is plausible because it counts both authors when there are only two, but always includes the last author (who may often be the head of the research group that organized the descriptions). Fig. 3, *Right*, shows the relationship of the first measure with the other two. All three measures are strongly correlated, and the first two are very similar indeed. We chose not to proliferate estimates by using all three measures, but to note that, in the equation used to estimate species, replacing the number of all taxonomists (T_i) by the very similar number of first and last named taxonomists would not alter the results. Neither would replacing that number with the proportional number of senior authored taxonomists (kT_i, where k is a constant), although the slope of the taxonomic efficiency would be proportionately reduced.

A second broad concern is whether the WCSP dataset is representative of all seed plants in terms of discovery rates and taxonomic effort. Although the catalog is complete for monocots, it is not for other species. However, the set of more than 108,000 species we do analyze is both large and contains such families as the orchids that have high levels of endemism in the tropics. Simply, the set of species we use is not obviously biased in favor of or against the conclusions we draw.

The third broad concern involves the nature of the definition of "species." This certainly changes over time. Were this study done in the future, or had it been done in the past, the different definitions might produce different numbers of missing species. Our point in using data with consistent and recent revisions of species names, however, is to ensure that all the data herein are broadly consistent in the definition of species.

Confirming Conservation Priorities. Our first simple conclusion is that, even accounting for the uncertainty described here, our results leave untouched the broad idea of biodiversity hotspots. We predict the great majority of the endemic missing species to be in hotspot regions where, by definition, habitat loss is extensive: Mexico to Panama (6% of all predicted missing species), Colombia (6%), Ecuador to Peru (29%), Paraguay and Chile southward (5%), southern Africa (16%), and Australia (8%) combined have 70% of all the species we predict missing—and these are not all of the hotspots. Of the nine regions with high estimates of missing species, only two [the northern Palaearctic (2%) and Angola to Zimbabwe (2%)] do not contain hotspots (Table 1). It is reassuring to note that, in the absence of any large-scale changes in the destruction of natural habitat, we do not predict the appearance of additional hotspots when we incorporate numbers of plant species yet to be described.

The definition of hotspot has two parts: one for endemism, the other for habitat loss. So, there are some places where recent habitat loss might cause the region to now qualify under the criteria of Myers et al. (6). New Guinea, for example, has far greater than 1,500 endemic plant species, but was not originally included as a hotspot because of its relatively intact natural vegetation. Although our model predicts New Guinea's catalog of plant species to be nearly complete, ongoing loss of natural habitat through logging, mining, and road construction (22) could make the region a likely candidate for inclusion in the future.

There is, however, a matter of urgency. When—or if—taxonomists describe these missing species, they are likely to be classified as threatened with extinction, given their geographic

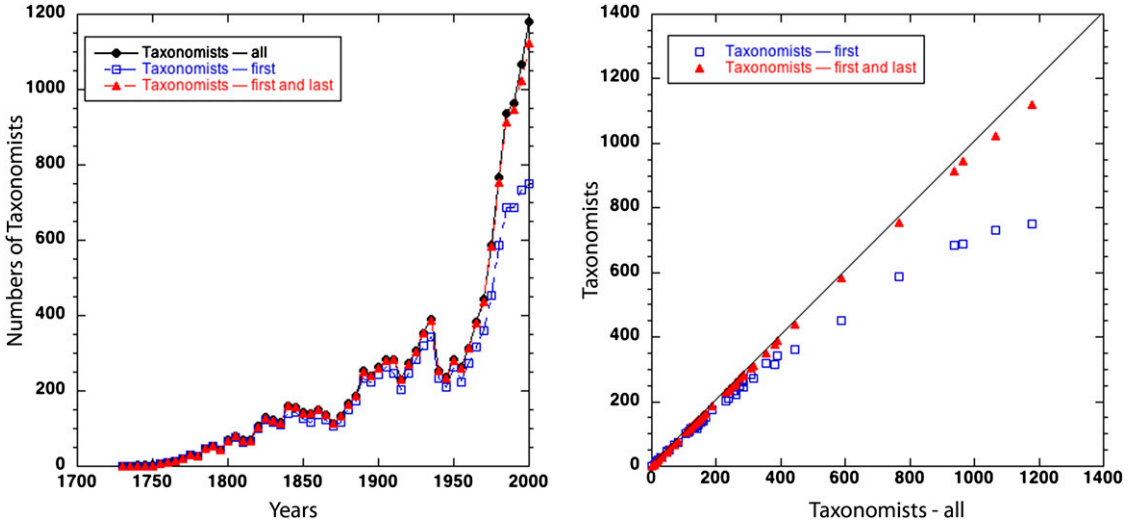

Fig. 3. *Left*: The number of all of taxonomists involved in species description in a given 5-y interval, the first and last named, and just the first named. *Right*: The last two measures plotted against the first. The very strong correlations mean that using other counts of taxonomic effort than the one we use (all taxonomists) would not seriously alter our estimates of species numbers.

locations and biological traits. Thus, estimates of the fractions of plants that are endangered (based on residing within biodiversity hotspots) are clearly far too low.

We conclude that conservation efforts can be broadly targeted at currently accepted hotspots without the fear of missing cryptic, and thus far unappreciated, biodiversity hotspots as a result of the incompleteness in our checklist of species on Earth. However, we also show that most missing species are in select hotspots, and thus many more species are at risk for extinction than previously recognized.

Materials and Methods

Creating Regions. The component TDWG regions that comprise any clustered regions are listed in Dataset S2. All TDWG regions were included in our analysis (Dataset S2), with the exception of Saharan Africa (TDWG regions: Burkina, Chad, Mali, Mauritania, Niger, and Western Sahara) and the Arabian Peninsula (TDWG regions: Gulf States, Kuwait, Oman, Saudi Arabia, and Yemen). These two clustered regions were excluded because of the extremely low number of endemic species they contain (Saharan Africa, $n = 23$; Arabian Peninsula, $n = 131$). Original TDWG regions are shown in Fig. S1, and the clustered regions that were analyzed are shown in Fig. S2.

Summarizing a Region's Data. For each region analyzed, we combined taxonomic summary statistics for each 5-y period, starting in 1730. These statistics included the number of species described in that 5-y time interval (S_i) and the number of unique taxonomists that described species within that interval (T_i). From the S_i statistic, we also calculated the cumulative number of species described up to that time interval (ΣS_i). Following this procedure, we then removed all years before 1770 from the dataset [to remove the enormous and cumulative influence of Linneaus (23)]. The species described before 1770 were still represented with the ΣS_i statistic), as well as those years when no species were described (i.e., S_i of 0). We then truncated each region's dataset to start only when the region had accumulated at least 40 endemic species. Finally, we calculated the number of years each time interval was from the first year the region first accumulated 40 species (Y_i).

Model and Model Fitting. These data enter a model to predict the number of species described per 5-y interval within a region:

$$S_i = (a + b\,Y_i)^*\left(T_i^*\left(S_T - \sum S_i\right)\right) \quad [2]$$

S_i, Y_i, T_i, ΣS_i have been described here, and a, b, and S_T are parameters to be estimated, the last one being the predicted total number of species.

The model was fit by using the bespoke software package Filzbach, which is a set of C++ libraries that allows for robust and rapid maximum-likelihood estimation. The Filzbach libraries themselves, along with the code for our analyses, are available from the authors upon request. We assumed a normal distribution for the errors between the observed (S_i) and estimated (S_{iest}) number of species per time interval, whereby the SD of the normal distribution was given to be the square root of the mean. This approximates a Poisson distribution, and was necessary because of the large expectations in certain years of our analysis. Additionally, we introduced a fourth parameter, z, as a method for handling data dispersion, which modified the width of the normal distribution. The likelihood calculation was thus:

$$L(X \mid a,\ b,\ S_T) = \sum \ln\left[\text{Normal\_Density}\left(S_i,\ S_{ies}, z^*\left(\sqrt{S_{iest}}\right)\right)\right], \quad [3]$$

where z was a shape-modifying parameter of the normal distribution.

The use of maximum-likelihood methods allowed us to obtain 95% CIs on all parameters, with S_T being the parameter of principal interest. Dataset S1 reports these CI estimates for S_i, and the maximum-likelihood estimate (MLE) only for the parameters a, b, and z.

The parameter we take most interest in here is S_T and the ratio we derive from it (current number of species/S_T) that lends insight into regions with potentially high numbers of missing species. Fig. S3 maps the numbers of missing species we estimate remain to be found in each region. Here it is useful to note that, across all regions, the width of the CI around this ratio is positively and significantly related to the MLE of the ratio itself ($r^2 = 0.63$, $P < 0.0001$), although this result is strongly influenced by two outlier regions in this relationship (New Caledonia and West Africa: Senegal to Benin), which have, respectively, predicted ratios of 1.10 and 1.31, but which range from almost no new species to a 300% increase (New Caledonia) and a 10% to 300% increase (West Africa: Senegal to Benin). Excluding these two regions changes the relationship between CI and actual prediction to an even stronger one ($r^2 = 0.79$, $P < 0.0001$).

Taxonomic Efficiency. Given an estimate of the total number of species in a taxon or region and thus the number of presently missing species, the model allows calculation of the taxonomic efficiency. This measures how many species are described per time interval per taxonomist, corrected for the number of species remaining to be discovered. Algebraically: the efficiency is as follows:

$$S_i / [T_i (S_T - \Sigma S_i)] \quad [4]$$

Fig. S4 shows the plot of this measure for approximately 50,000 species of flowering plants that are not monocots. This is the largest set of systematically revised, monophyletic species available to us. Using this large number of species offers the best chance to uncover any trends or deviations from a trend that would invalidate our assumption of a linear increase in efficiency. The consistent overall increase in this measure is apparent. Quite clearly, given the number of species remaining to be discovered and given the number of taxonomists involved in describing species, more are described now than in the past. There is no compelling reason to model this relationship with anything other than a linear function: ($a + b\,Y_i$).

ACKNOWLEDGMENTS. The authors thank the Royal Botanic Gardens, Kew, United Kingdom, for providing access to and assistance with the WCSP; R. Govaerts and A. Paton for access to the World Checklist of Selected Plant Families; D. Simpson for helpful discussions on GrassBase; G. Russell for statistical advice; and D. Purves for statistical advice and access to the software used to produce our results. P. Raven provided the initial challenge to estimate the number of missing species (and where they are), plus extensive comments on this manuscript.

1. May RM (1988) How many species are there on Earth? *Science* 241:1441–1449.
2. May R (1990) How many species? *Philosophical Transactions of the Royal Society B* 330:293–304.
3. Wilson EO (1992) *The Diversity of Life* (Harvard Univ Press, Cambridge, MA).
4. Brooks TM, et al. (2006) Global biodiversity conservation priorities. *Science* 313:58–61.
5. Pimm S, Jenkins C (2010) Chapter 10. In *Conservation Biology for All*, eds Ehrlich P, Sodhi N (Oxford Univ Press, Oxford), pp 181–196.
6. Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature* 403:853–858.
7. Cincotta RP, Wisnewski J, Engelman R (2000) Human population in the biodiversity hotspots. *Nature* 404:990–992.
8. Baillie J, Hilton-Taylor C, Stuart S (2004) *IUCN Red List of Threatened Species. A Global Species Assessment* (IUCN, Gland, Switzerland).
9. Dalton R (2000) Biodiversity cash aimed at hotspots. *Nature* 406:818.
10. Joppa LN, Roberts DL, Pimm SL (2010) How many species of flowering plants are there? *Proceedings of the Royal Society B* 278:554–559.
11. Pimm SL, Jenkins C, Joppa LN, Roberts DL, Russell G (2010) How many endangered species remain to be discovered in Brazil? *Natureza & Conservação* 1:71–77.
12. Mabberley DJ (2009) Plant science. Exploring terra incognita. *Science* 324:472.
13. Bebber DP, et al. (2010) Herbaria are a major frontier for species discovery. *Proc Natl Acad Sci USA* 107:22169–22171.
14. Paton A, et al. (2008) Towards target 1 of the global strategy for plant conservation: A working list of all known plant species – progress and prospects. *Taxon* 57:602–611.
15. Brummitt R (2001) *World Geographical Scheme for Recording Plant Distributions* (Hunt Institute for Botanical Documentation, Pittsburgh), 2nd Ed.
16. Clayton WD, Vorontsova MS, Harman KT, Williamson H (2009) *GrassBase—The Online World Grass Flora*. Available at http://www.kew.org/data/grasses-db/. Accessed February 20, 2010.
17. Wilson SP, Costello MJ (2005) Predicting future discoveries of European marine species by using a non-homogeneous renewal process. *Appl Stat* 54:897–918.
18. Solow AR, Smith WK (2005) On estimating the number of species from the discovery record. *Proceedings of the Royal Society B* 272:285–287.
19. Cotello M, Wilson S (2010) Predicting the number of known and unknown species in European seas using rates of description. *Glob Ecol Biogeogr* 20:319–330.
20. 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 B* 275:149–155.
21. Ahrends A, et al. (2011) Conservation and the botanist effect. *Biol Conserv* 144: 131–140.
22. Shearman P, Ash J, Mackey B, Bryan J, Lokes B (2009) Forest conversion and degradation in Papua New Guinea 1972-2002. *Biotropica* 41:379–390.
23. Linné C (1753) Species plantarum. *Imprensis Laurentii Salvii Stockholm* 2:970–971.