https://doi.org/10.22201/fc.25942158e.2021.02.281

INCLUSION IS NOT REPRESENTATIVENESS: THE CONTEXT OF COLOMBIAN SAMPLES IN THE TAXONOMIC AND SYSTEMATIC MID-LARGE HERPETOLOGICAL LITERATURE

LA INCLUSIÓN NO ES REPRESENTATIVIDAD: EL CONTEXTO DE LAS MUESTRAS COLOMBIANAS EN LA LITERATURA TAXONÓMICA Y SISTEMÁTICA HERPETOLÓGICA A MEDIANA-GRAN ESCALA

Juan D. Vásquez-Restrepo[1]*

[1]*Programa de Posgrado en Ciencias Biológicas, Instituto de Geología, Universidad Nacional Autónoma de México, Ciudad de México, México.*
**Correspondence: juanda037@outlook.com*

Received: 2021-03-07. **Accepted:** 2021-04-26.
Editor: Andrés Rymel Acosta Galvis, Colombia.

Colombia is considered as one of the most biodiverse countries in the world (Samper, 1998; Myers et al., 2000; Arbeláez-Cortés, 2013; Zachos & Habel, 2014), but at the same time, sampling efforts have been constrained historically by socio-political issues, limiting the possibility of understanding several aspects of that biodiversity. For instance, scientist in some neighboring countries (e.g., Brazil and Ecuador) discover and describe new taxa at a relatively more rapid rate than Colombia (Rivera-Correa, 2012), and at the same time conduct integrative studies of those groups. Therefore, it comes as no surprise that many comprehensive taxonomic and systematic studies of amphibians and reptiles incorporate little to no data from Colombia (see Appendix I).

While some regard taxonomy as a primarily descriptive and hampering science (Garnett & Christidis, 2017), others perceive it as part of the backbone of many areas such as ecology, evolution and systematics, and as a necessary activity in biology (Thomson et al., 2018). In addition, poorly developed taxonomy causes problems and bottlenecks in the biodiversity sciences (Kaiser et al., 2013; Vogel-Ely et al., 2017). The Neotropical biota has a complex evolutionary history closely related to geological history (Hoorn et al., 2010; Antonelli & Sanmartin, 2011; Antonelli et al., 2018). In these historical processes, the area where Colombia is currently located has played a significant role as a mid-point between Central and South America (Samper, 1998; Jaramillo & Oviedo, 2017). For this reason, studies with considerable sampling gaps address the evolutionary history of lineages partially only, and may be masking genetic structure or species threats (Hillis, 2019; Chambers & Hillis 2020, Cordier et al. 2021). Additionally, these sampling gaps may cause taxonomic instability, since the identity of the unsampled taxa or of taxa in unsampled regions are normally ignored or questioned but not resolved, therefore increasing the number of paraphyletic groups or species complexes. As well, taxonomic and systematic uncertainty generated affects the appropriation of knowledge by local communities and researchers, and also delay the generation of new knowledge.

Herein I will discuss two main topics: 1) the inclusion and representativeness concepts, focused on the herpetological context of Colombia; and 2) possible causes and consequences of systematic inclusion gaps.

In order to do so I performed a bibliographic analysis by mean of an advanced search through the Web of Science Core Collection (WoS, http://webofknowledge.com) for taxonomic and systematic studies of amphibians and reptiles in the Neotropical region during the last 20 years. I defined the beginning of the century as starting point for the bibliographic analysis, because from there, technological advances in computation and communication favored and expanded the accessibility to data and information. For the purposes of this study, I defined two criteria for considering a publication: the taxa studied must

occur in at least two countries (including Colombia), and studies must to be comprehensive, that is, they should to be focused on one or several traits (e.g., morphology, genetics, evolution, bioacoustics, distribution) of a taxon or taxa in a generalized way. I explicitly excluded Colombian only studies because they will include Colombian samples by definition, and this analysis aims to evaluate groups in which Colombian data are not used or limited, compared with the data used for the same group from neighboring countries.

Searches were conducted on the title, abstract and keywords, using as criteria three key elements: SUBJECT = Amphibia OR Reptilia + AREA = South America OR Neotrop* + TOPICS = Taxonom* OR Systematics OR Phylogen* OR Review OR Biogeograp* OR Distribution (* are text wildcards). Results were manually filtered in order to remove publications different to the main topics defined, extinct taxa, and explicitly delimited to a specific area out of the target. Subsequently some studies not recovered in the search were added manually.

I grouped the publications categorically according to whether they included samples from Colombia or not (i.e., yes, no, or partial when not for all lines of evidence used). The studies that included samples from Colombia were then grouped according to the source of those samples (national or foreign museums), and author's nationality or place of residency (natives or established). Author's nationality/residence was corroborated based on their names, historical institutional affiliations or education, personal information available in social networks, CvLAC, ResearchGate, Google Scholar, personal/institutional web sites, or by asking colleagues who possibly know or had worked with them. The idea behind using author's nationality/residence as classification criteria, lies in the fact that Colombian or Colombia resident authors may have a broader notion of the institutions housing biological specimens, their taxonomic and geographic coverage, or their accessibility. Therefore, an author's nationality/residence may affect the probability of choosing a source of data, either national, foreign or both. Subsequently, I calculated the conditional probability of using national and/or foreign museums P(A), given the nationality/rersidence of authors P(B).

$$P(A|B) = \frac{P(B|A)\, x\, P(A)}{\sum P(B|A) x\, P(A)}$$

As I discuss below, there are marked differences in data from national and foreign museums, a reason why the selection of the data sources may be related to the degree of inclusion and representativeness of a taxa in a particular study. To quantify and validate the data for Colombia in biological collections, I downloaded the list and associated information from the Registro Nacional de Colecciones Biológicas (RNC, http://rnc.humboldt.org.co), and a dataset for amphibians and reptiles with country "Colombia" from the Global Biodiversity Information Facility (GBIF, https://www.gbif.org), both to February 8, 2021 (Appendix II). For Colombian museums, the invisibility of the data was calculated as the proportion of data in GBIF divided by the total number of specimens reported in RNC, since this list contains the official number of housed specimens in each collection. Finally, I performed a descriptive analysis of the geographic coverage and data quality in national and foreign museums using the previously mentioned datasets. These data from museums coverage a 180 years span.

After cleaning the search results, a total of 160 studies (77 for amphibians and 83 for reptiles) meeting the inclusion criteria were selected (Appendix I). Most of them included at least one sample from Colombia for their interest groups, followed by those which did not, and a small proportion for which it was impossible to assess the origin of the sample since neither the methods nor the supplementary materials stated explicitly

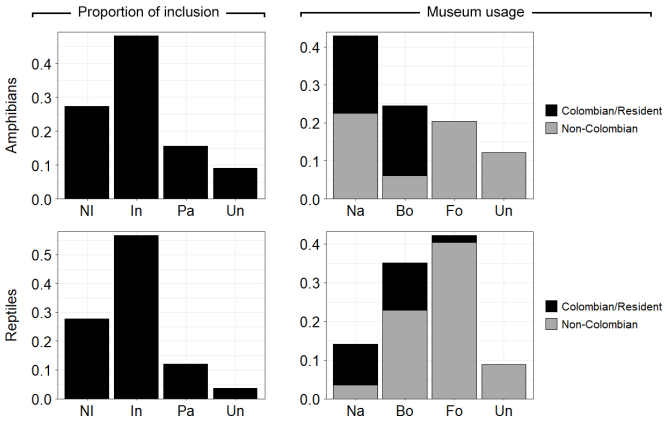

Figura 1. Proporción de inclusión de muestras colombianas y uso de museos en la literatura herpetológica analizada. NI = No incluido, In = Incluido (al menos una muestra), Pa = Parcial (al menos una muestra pero no para todas las líneas de evidencia), Un = Desconocido (no se indicó la localidad). Na = Museos nacionales, Bo = Ambos (nacionales + extranjeros), Fo = Museos extranjeros, Un = Desconocido (no se indicó la fuente de datos).

Figure 1. Proportion of inclusion of Colombian samples and museum usage in the herpetological literature analyzed. NI = Not included, In = Included (at least one sample), Pa Partial = (at least one sample but not for all lines of evidence), Un = Unknown (studies do not state localities). Na = National museums, Bo = Both (national + foreign), Fo = Foreign museums, Un = Unknown (studies do not state the source of data).

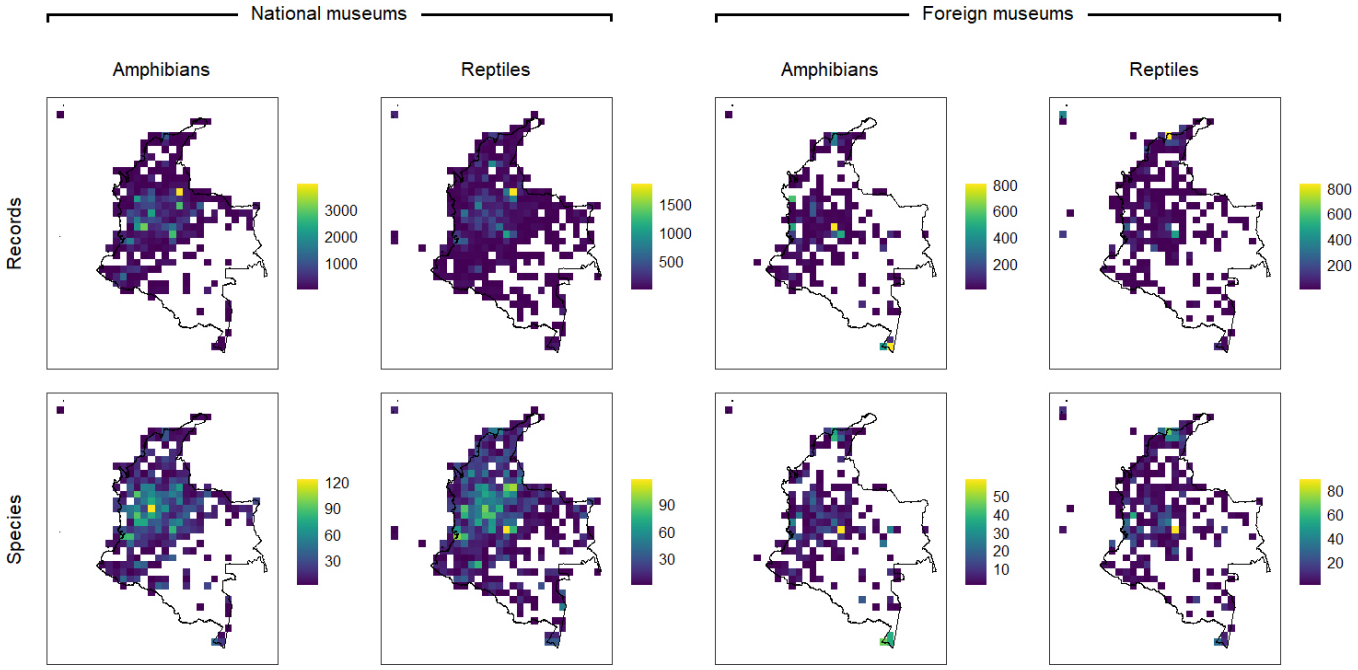

Figura 2. Panorama general de la representatividad geográfica y taxonómica de los anfibios y reptiles recolectados en Colombia, y depositados en museos nacionales y extranjeros. Los mapas están basados solo en registros georreferenciados. El periodo temporal de los museos extranjeros va desde ~ 1840 (probablemente 1800) a 2020, y desde 1875 para museos nacionales. Tamaño del pixel 0.5° x 0.5° en WGS84. Datos tomados del RNC y GBIF (al 8 de febrero, 2021). Las capas raster están disponibles en el Apéndice III.

Figure 2. Overview of the geographic and taxonomic representativeness of the amphibians and reptiles collected in Colombia, and held in national and foreign museums. Maps are based only on georeferenced records. Foreign museums temporal span goes from ~ 1840 (probably 1800) to 2020, and from 1875 in national museums. Pixel size 0.5° x 0.5° in WGS84. Data from RNC and GBIF (accessed February 8, 2021). Rasters are available in Appendix III.

the localities (Fig. 1). Regarding the museum usage, national collections were predominant for amphibians, while foreign were for reptiles (Fig. 1). Conditional probabilities (Table 1) showed that for amphibians it is more likely that in a randomly chosen paper, authors will include Colombian samples from national collections independent of their nationalities. However, for reptiles it is more likely that authors will include samples from foreign museums if they are foreigners, whereas for Colombian/Resident authors the probability of also including national collections increases substantially. Interestingly, data show that foreign museums have less taxonomic and geographic representativeness of Colombian herpetofauna (Fig. 2), and also specimens have less associated geographic information (Fig. 3). Both national and foreign museums possess a similar and significant proportion of specimens taxonomically identified to Family, Genus and Species levels. Georeferenced records also

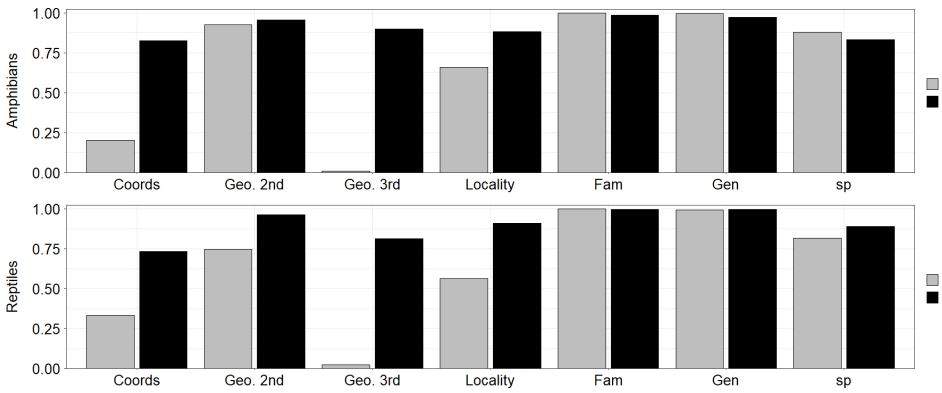

Figura 3. Proporción de datos asociados a anfibios y reptiles. N = Museos nacionales (barras negras), F = Museos extranjeros (barras grises). Geo. 2nd y 3rd hacen referencia al segundo y tercer nivel de división político administrativa (para Colombia departamentos y municipios, respectivamente). Datos tomados del RNC y GBIF (al 8 de febrero, 2021).

Figure 3. Proportion of associated data for amphibian and reptile specimens. N = national museums (black bars), F = foreign museums (gray bars). Geo. 2nd and 3rd refers to the second and third level of administrative political division (for Colombia department and municipality, respectively). Data from RNC and GBIF (accessed February 8, 2021).

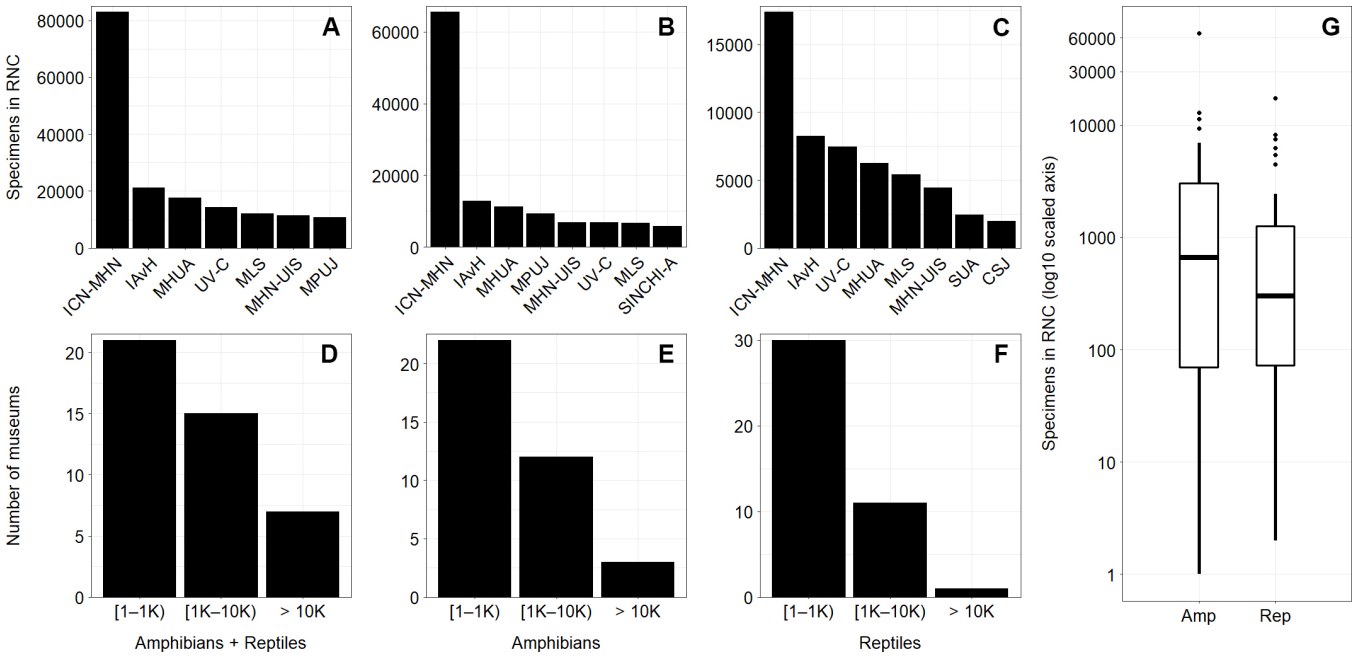

Figura 4. Colecciones herpetológicas colombianas más grandes por número total de especímenes de acuerdo al RNC, ambos grupos (A), anfibios (B), reptiles (C). Número de museos de acuerdo a la proporción de especímenes que albergan, ambos grupos (D), anfibios (E), reptiles (F). Distribución del número de anfibios y reptiles en museos colombianos de acuerdo al RNC. Para los acrónimos de los museos véase el Apéndice II.

Figure 4. Largest Colombian herpetological collections by total number of specimens according to RNC, both groups (A), amphibians (B), and reptiles (C). Number of museums according to the proportion of total number specimens, both groups (D), amphibians (E), and reptiles (F). Distribution of the number of specimens of amphibians and reptiles in Colombian museums according to RNC (G). For museum acronyms see Appendix II.

show that foreign museums have focused their sampling efforts in two areas, Sierra Nevada de Santa Marta in the Caribbean region, and Leticia and nearby areas of Amazonas, whereas national museums have focused on the central and north portion of the Andes of Colombia.

Excluding zoos and live collections, there are 43 herpetological museums formally registered in the RNC, of which 37 have both amphibian and reptile subdivisions, for a total of 80 collections. The approximate total number of preserved specimens (amphibians + reptiles) with country listed as "Colombia" is about 58 000 in foreign museums, and just over 214 000 in national collections (Table 2). Almost 80% of the latter are housed in seven museums, each of which has more than 10 000 specimens (Fig. 4A). Looking at each biological group independently, the ICN collection by far exceeds the number of specimens both for amphibians and reptiles when compared to the other largest museums (Fig. 4B-C). Most of the herpetological museums and collections in Colombia are small, having less than 1000 specimens, followed by those which have between 1000 and 10 000, and the largest ones with more than 10 000 (Fig. 4D–G). According to GBIF, 62.5% of the herpetological collections in Colombia do not have public data in that platform. For the available datasets, the invisibility of the specimens is quite variable, ranging from 0 to 98%, while some collections have more records in GBIF than those declared in the RNC (Fig. 5).

Tabla 1. Probabilidades del uso de museos P(A) en la literatura herpetológica, por nacionalidad/establecimiento de los autores P(B), expresado como una probabilidad condicional P(A|B). Los cálculos están basados en la proporción de artículos que incluyen al menos una muestra de Colombia. Ver Apéndice I.

Table 1. Probabilities of museums usage P(A) in the herpetological literature, by authors nationality/residence P(B), expressed as a conditional probability P(A|B). Calculations are based on the proportion of papers including at least one Colombian sample. See Appendix I.

Museum	Amphibians (n = 43)		Reptiles (n = 52)	
	Colombian/ Resident	Non-Colombian	Colombian/ Resident	Non-Colombian
National	0.6604	0.6294	0.2264	0.0193
Both	0.3396	0.0981	0.6604	0.3140
Foreign	0.0000	0.2725	0.1132	0.6667

According to the data of inclusion of Colombian samples in the literature analyzed, it seems there are more studies including samples (independently of their origin) than those do not. Here, I am considering the degree of inclusion in a given publication as the number of samples relative to the species' distribution, different from sampling proportion, which comprises the number of samples relative to the total, and representativeness, as the variation encompassed by those samples. However, it is important to keep in mind that both inclusion and representativeness are dependent on taxon-distribution and sample availability (Fig. 6). This means that they are not directly comparable among studies on a continuous or discrete scale, because even the same proportional sampling may be more or less inclusive and representative for certain taxa. Thus, in order to account for the overall inclusion, the use of categories is necessary (e.g., included vs. non-included). The primary limitation of this method is its trend to inflate the proportions, masking the true extent of the gaps, since just one single sample is sufficing to consider a study as inclusive. This is why in many taxonomic or systematic studies in herpetology the degree of inclusion and representativeness of Colombian samples is apparently high, when in reality is low to null, often limited to a few samples which do not reflect their actual availability in biological collections. Although it is easily explainable for rare taxa for which large samplings are not expectable, it is less understandable in the case of widespread or common groups.

This problem is a vicious circle, since the known distributions of many taxa represent theoretical approaches, normally reflecting sampling effort (Hortal et al., 2015), but they cannot be refined if data from specimens are not included in the studies. However, to discuss the degree of inclusion and representativeness, it is also important to consider that inclusion does not necessarily reflect representativeness. To better illustrate the inclusion-representativeness relationship, I have summarized some patterns of inclusion and representativeness based on hypothetical scenarios (Fig. 6). The first pattern is evidenced when there are few samples of a given taxon that has a small distribution range in site 1 relative to site 2; in this case it is expected for 1 to have a low proportional sampling, but the inclusion is high and representativeness should be considered good (Fig. 6A). The second pattern is identified when a taxon is widely distributed in two or more sites, but some of them are significantly more heavily sampled (Fig. 6B–C). In this case the inclusion is low and the representativeness poor to medium, because variation in the less sampled sites is uneven compared to variation in the more well-sampled sites, therefore inferences will be biased towards the latter. For the example B

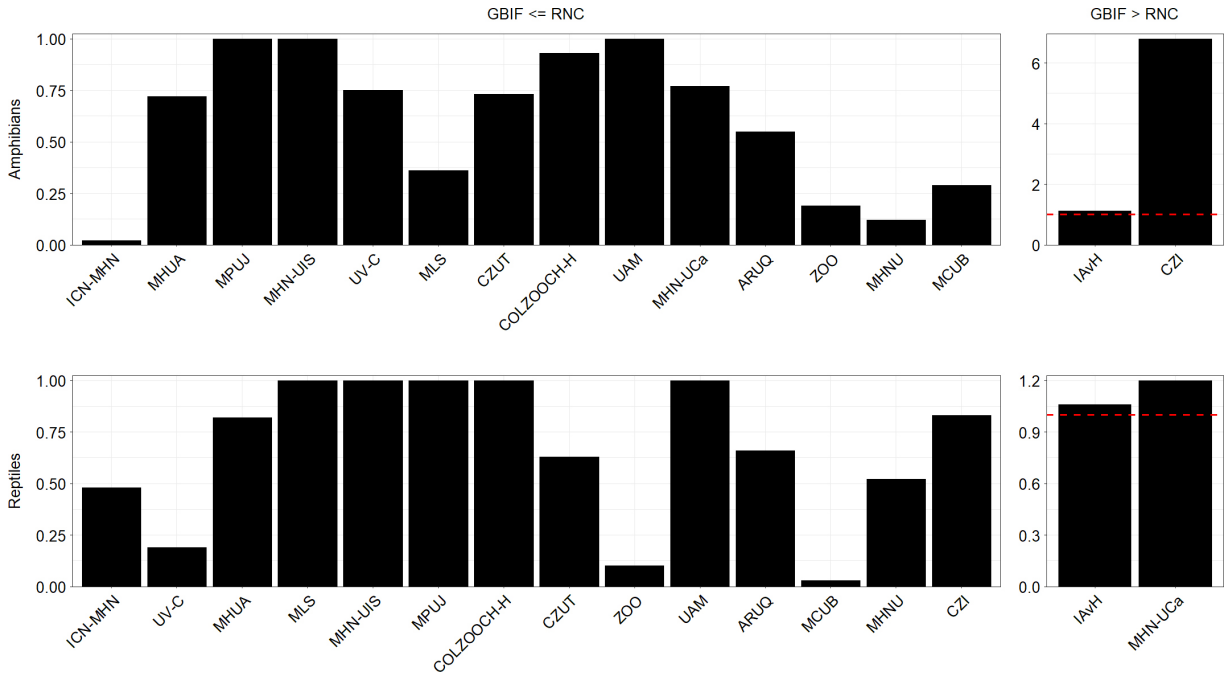

Figura 5. Visibilidad de los datos asociados a anfibios y reptiles en museos colombianos, como la proporción de datos publicados en GBIF sobre los del RNC (para GBIF ≤ RNC). La línea roja indica el número total de especímenes en el RNC (para GBIF > RNC). Para los acrónimos de los museos véase el Apéndice II.

Figure 5. Visibility of amphibians and reptile data in Colombian museums as the proportion of published records in GBIF datasets compared to of RNC (for GBIF ≤ RNC). The red dashed line represents the total number of specimens in RNC (for GBIF > RNC). For museum acronyms see Appendix II.

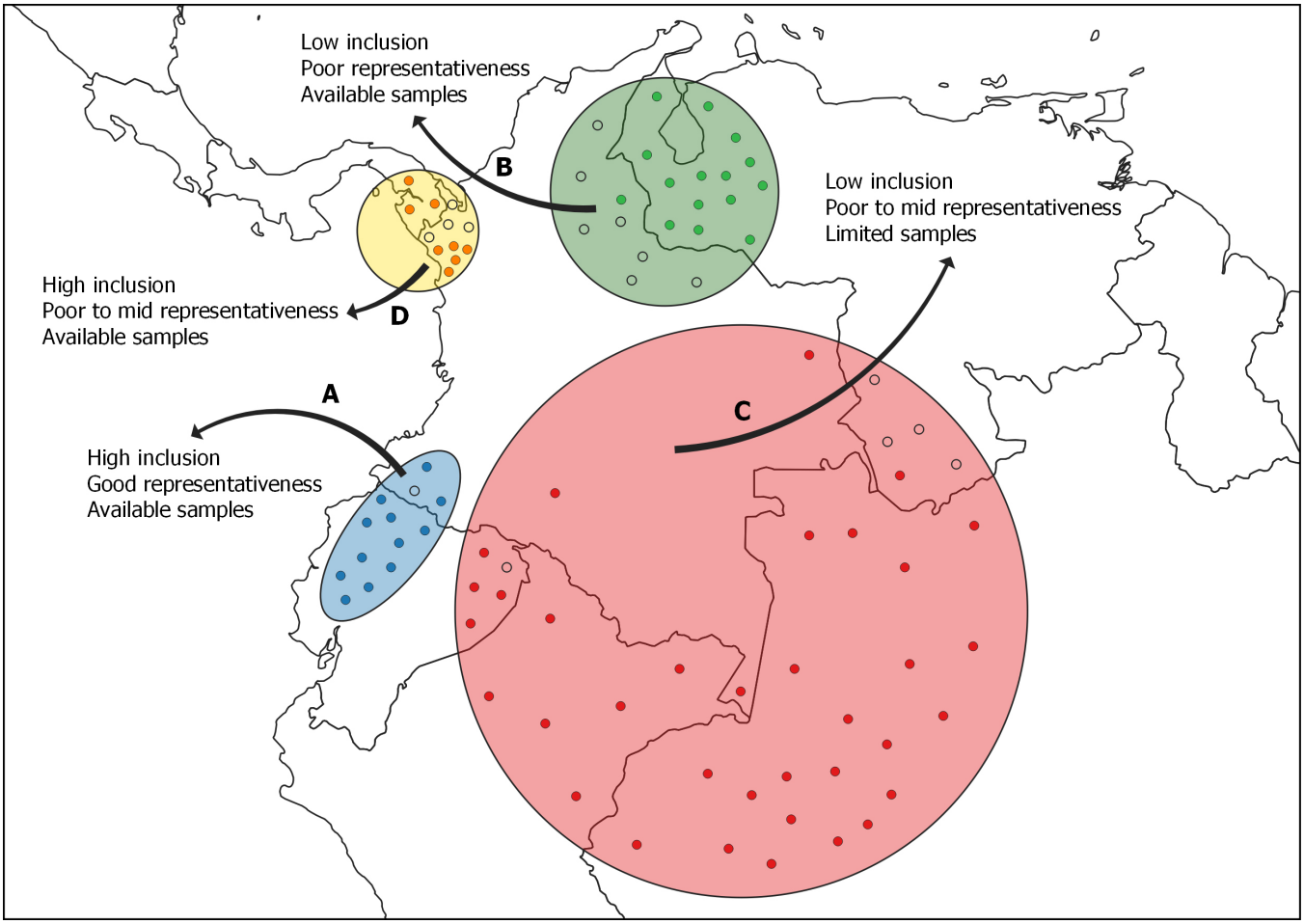

Figura 6. Patrones de inclusión y representatividad de acuerdo a la distribución de los taxones y disponibilidad de muestras en escenarios hipotéticos. Los puntos coloreados representan los registros incluidos en tres escenarios hipotéticos, mientras que los vacíos representan los no incluidos pero disponibles en colecciones biológicas. Para una descripción detallada véase el texto.

Figure 6. Patterns of taxonomic inclusion and representativeness according to taxa distributions and sampling availability in hypothetical scenarios. Solid circles represent included records in the three hypothetical scenarios, whereas open circles represent non-used but available specimens in biological collections. For detailed description see the text.

in the Fig. 6, low inclusion is non-justified given the availability of samples, contrary to the observed in the scenario C. It is worth noting that inclusion may increase without an increase in representativeness, when there is sampling but it does not imply more variation (Fig. 6D). There are not numerical criteria to define what is low or high, poor or good, the scenarios presented represent a referential conceptualization. In real life, the degree of inclusion and representativeness for a taxonomic group may be influenced by different reasons, that will be the subject of discussion later.

For instance, with Colombia as a reference point, the first pattern is commonly found in Amazonian taxa for many of the amphibians in the literature that was analyzed. Nonetheless, it is important to consider that the Colombian Amazon and Orinoquía are largerly unexplored, thus the real distribution of many groups may be underestimated (Wallacean shortfall). On the contrary, the second pattern is more frequent in reptiles' literature, for groups distributed in northwestern South America or extending from Central to South America. Historically, Colombian amphibians have attracted more attention than reptiles, this is the reason why they have been more thoroughly studied, and therefore their inclusion pattern is more accentuated.

As shown, at a large taxonomic scale inclusion and representativeness of Colombian samples may be masked by the difficulties of comparing among studies. For this reason,

Tabla 2. Especímenes de anfibios y reptiles en museos colombianos y extranjeros de acuerdo al RNC y GBIF. La invisibilidad está calculada como la proporción GBIF/RNC según los datos publicados (al 8 de febrero, 2021).

Table 2. Amphibian and reptile specimens in Colombian and foreign collections according to RNC and GBIF. Proportion of invisibility based on GBIF/RNC published data (accessed February 8, 2021).

Group	National				Foreign	
	Number of collections	Specimens RNC	GBIF/SiB	Invisibility proportion	Number of collections	GBIF
Amphibians	37	147 927	54 955	0.63	42	38 714
Reptiles	43	66 387	37994	0.43	59	19 814

it is necessary an approach based on an indirect measure. One criterion is to use the differences in geographic coverage and data associated with the specimens between national and foreign museums, which may limit the usage of some of them. Most specimens from foreign museums lack georeferenced data and/or information at the municipality level (3rd administrative level), hence generating a broad margin of error when spatializing them in if the locality does not supply that information. This is particularly true when records come from an orographically and ecologically heterogeneous zone like the Andes. For example, mainland Colombian departments range from ca. 1800 to 110 000 km², therefore, just with 2nd level information, locality uncertainty is larger than the area of Hong Kong (~ 1100 km²), El Salvador (~ 20 800 km²), Costa Rica (~ 51 000 km²) or Cuba (~ 109 800 km²). In this manner, if we consider foreign museums as the preferred source for foreign researchers, together with the amount of invisible data from Colombian museums, the number of spatializable specimens as a necessary condition for a comprehensive study in taxonomy and systematics is reduced. Another important problem arises when looking at the tables, appendices or supplementary data in the literature, because the traceability of some specimens because the traceability of some specimens (mainly genetic samples) is often messy because field numbers or provisional codes are normally provided, but not notes clarifying the ultimate disposition of the sample.

Regarding data quality for older specimens, it is also important that there is a historical bias for the capital cities to have many records and species, even if they do not occur there. For instance, despite the fact Bogotá is highly represented in foreign museums (the yellow pixel in central Colombia in Fig. 2), several records may be misattributed to this locality, a frequent mistake in specimens sent out from major cities during the XIX and XX centuries, which were assumed to be the locality of origin. This bias may induce errors by omission if researchers are unfamiliar with the history of collections in Colombia, or simply inhibit them from utilizing data if distributions are suspicious.

Another factor that may mask the sampling inclusion are the types of data used, which vary according to the objective of the study. Essentially, data may be of four types: geographic, morphologic, genetic and bioacoustic. Among them, geographical data are the easiest to obtain, since they merely require a GBIF download or a literature review, reason why studies employing this kind of information tend to have more inclusion and representativeness for Colombian samples. By comparison, morphology, genetics and bioacoustic data require the greatest effort to obtain, thus their inclusion and representativeness tend to be lower.

To conclude, I have explored five possible reasons that may explain the phenomena what have I been pointing out along the text. I do not consider these points to be isolated from each other, on the contrary, they may be synergetic among them and vary in importance and occurrence through time.

The true absence of sampling. It is well known that among living beings there are some groups/species that are particularly difficult to find. This difficulty may be due to: 1) the biology/habitat of the species (e.g., aquatic, fossorial or canopy-dwelling); 2) the spatial distribution or density of target populations; and 3) the inaccessibility of the sampling areas, because of conditions of both terrain or due to socio-political issues. This last is particularly true for Colombia, where internal armed conflict and its derivatives, have constrained the exploration of many territories during the last 60 years. The absence of sampling represents the most significant limitation, because it represents a real absence of data.

Within the absence of sampling, it is also necessary to highlight the absence of tissue samples for DNA extraction. In

Colombia taking tissue samples is an activity that dates back to 1998 (Arbeláez-Cortés et al., 2015), but it is limited given the resources and equipment required for storage, a reason why not all museums have a tissue collection, or if they do, it is relatively recent. Although molecular data do not represent a necessary condition for taxonomic studies, it is an important and comprehensive source of information in the -omics age (Will et al., 2005). This brings us to the problem that most collections from Colombia do not have associated genetic samples, either because they are old collections, or because researchers collect but do not extract tissues for DNA. At this point, it is important to highlight that the absence of genetic data for older collections does not make them useless, museum specimens are and have been traditionally used beyond molecular taxonomy (Castillo-Figueroa, 2018; Meineke et al., 2018; Guedes, 2021).

Now, the question is how to deal with the true absence of sampling? The most straightforward answer is to do more sampling. Nevertheless, this is easier on paper than in practice. Here, I am limiting myself to remark on some strategies that might help to address the problem of the true absence of sampling. First is the implementation of the integrative taxonomy approach when possible, by using various lines of evidence rather only one (Dayrat, 2005). This allows exploitation of old collections that may contain rare specimens or have accumulated a more complete geographic representation over time. Second, taking advantage of environmental consulting activities to fill in the sampling gaps in several taxa (i.e. taxonomic, geographical or genetic gaps). Environmental consulting activity is probably the chief source of new specimens to museums in Colombia, more than research itself, and normally companies that provide these services have budgets within which the collection activities are or may be included. Third, the implementation of novel techniques for recovering DNA from specimens that have been processed using formaldehyde (see Hykin et al. 2015 for an example in reptiles). I am aware that these last procedures may be expensive, however, they offer an option for very rare or even recently extinct taxa.

The bureaucracy. Bureaucracy depends on legislation, and laws are not static in time. Dealing with the bureaucracy is part of doing science, and to some extent a necessary one. Nevertheless, it is a critical factor to consider when conducting a research project, because it can generate delays, discourage researchers when processes are expensive or slow, and may even propitiate biopiracy (Fukushima et al., 2020).

For biological studies there are two principal ways to obtain data, museums and fieldwork, each has its limitations, although the former is among the most versatile (Meineke et al., 2018). In any case, in Colombia temporary or definitive collecting requires collection permission, and if definitive, the material must be deposited in a museum legally registered in the RNC.

The Colombian government issues permissions for commercial and non-commercial uses of biodiversity, and within these, for collecting specimens or accessing genetic resources (Decree 1076/2015). Collection permission without commercial purposes also include genetic research. These permissions are given by the environmental authorities to people or institutions interested in catching, removing or extracting, either temporarily or permanently, individuals from the wild. The request for permission may require several months for resolution, along with the requirements of submitting many forms and documents in support of the request (Supplementary Fig. 7A). For this reason, the most convenient solution is to conduct the research in association with a local institute that already has an umbrella permission, such as universities or scientific and governmental institutions.

Another issue arises when we want to sequence DNA. If the process is carried out in Colombia under an umbrella permission, there are no problems. But most institutions in Colombia do not have a sequencer, thus the most viable option is to send the samples outside the country for processing. Here is when bureaucracy may delay research, because additional permission is required to export biological specimens or samples (Supplementary Fig. 7B). As a result, researchers may avoid using genetic data from Colombia if it is not previously available in online repositories such as GenBank (https://www.ncbi.nlm.nih.gov/genbank), or if it cannot be generated locally. However, export permits are required for biological specimens, samples of them or their derivatives. This last point leaves an unsolved question, a loophole subject to interpretation, to what extent is an artificial product a derivative? Consider for example a PCR product, which is a synthetic soup of unordered nucleotides, an artificial model based on a mold, like a plaster cast of a mammal footprint, or like photographs and illustrations, which do not constitute the specimen itself or a sample of that specimen.

Finally, an additional consideration is the compensatory tax for wildlife hunting (Decree 1272/2016), which increases the economic cost of studies, given that the hunting concept includes temporal or definitive captures, and their subsequent processing, transportation and/or storage (Páez, 2016). This tax is calculated by specimen/sample, and employing a formula that considers the species and habitat conservation status, researcher's nationality, anthropic pressures, social

and environmental cost of the samples, commercial value of the species, and even how charismatic they are. Ironically, this creates a vicious circle, because the tax calculations require ecological data, but at the same time they are an impediment for the gathering of these data.

Then, how to deal with the bureaucracy? As I mentioned previously, one of the easiest solutions is to cooperate with local researchers and institutions. Furthermore, as I will discuss below, the use of online databases enhances the visibility of the existing specimens and facilitates their examination, since local researchers may help with data, reducing costs and time, at the same time building up the local researcher's visibility.

The invisibility of the data. One of the greatest advantages of the internet is that it brings the world just a click of distance away, and it is a powerful tool for museums and researchers. When a research project is planned, two of the first questions we need to think about are, what data are available? And, what data do I need to generate? There is an increasing trend towards making biodiversity data available through the internet, either through researchers or institutional web pages or by using repositories like GBIF.

But how this related to the problems of inclusion and representativeness? Fundamentally, the invisibility of data constitutes a false absence of sampling. This invisibility may be total if the museums do not utilize any electronic resources (or worse if they are secretive with their collection information), or partial if they utilize the resources but do not update information regularly, upload incomplete datasets, or if communication with the staff is not fluid and results in delays.

Biological collections in Colombia are regulated by Decree 1375/2013. According to this, museums assume the responsibility to keep updated the information associated with their specimens and upload it to SiB (Sistema de Información sobre Biodiversidad de Colombia, https://sibcolombia.net), a local repository of biodiversity similar to GBIF and also a source of information for the latter. This platform is centralized with a standardized format which offers an advantage over institutional museum webpages, which may not have advanced search tools, bulk download options, or may change their URL overtime or go offline.

Interestingly, the data from the biological collections in Colombia show that there are hundreds of thousands of amphibians and reptiles collected, covering a large spatial and temporal span, but it is important to note that spatial and temporal coverage does not imply taxonomic coverage, and some taxa may be oversampled while some others are undersampled. Then, why is there so little representativeness in the mid-large herpetological literature? It is important to keep in mind several considerations. First, literature reviewed herein encompasses a time period of 20 years, and probably the digitalization of many museum catalogues is more recent, as is the gradual increase in geographic coverage. In the same way, the extended use of the internet, digital resources and technology has increased enormously during the last decade, and some researchers have incorporated these changes faster than others. Therefore, it is impossible to expect that a researcher 15 years ago to have had the same access to the data that we have today. However, for some taxa with obtainable information (by whatever means), the invisibility of available samples has been, and will continue to be a problem.

Currently, the invisibility of specimens varies considerably among Colombian herpetological museums, being relatively high when grouped (Table 2), but smaller when individual collections are depicted (Fig. 5). This is explained by the large number of non-available records in some of the collections with the largest number of specimens, which shifts the general mean to higher values. On the other hand, some museums possess "extra" data when GBIF and RCN are compared. The latter reveals a phenomenon of inefficient handling of the platforms where the data can be uploaded or summarized. Given that not all the online platforms retrieve the data from a unique source, these differences are to be expected, for example, because curators may more frequently update the source they are more familiar with, or which its institution requires for most often. I am also aware that the entropy level (as a measure of disorder) in a biological collection will increase exponentially with the number of specimens they have, which makes easier the data curatorship and handling in smaller ones. Entropy is fine and expectable in a collection being used, but extreme entropy or entropy in unused collection are not desirable. As well, at least for the Colombian museums attached to university institutions, curators usually also play a role as professors, which also limit greatly the time and effort they can spend in curatorial activities.

To bypass this problem the implementation of good practices during data curatorship and collecting are required, not only inside individual museums but in the field at large. For example, taking the time to identify the specimens more than assign IDs based on localities or at bird's-eye view, using traceable unique identification vouchers, unifying databases in single or linked files, avoiding uncommon or non-universal abbreviations, georeferencing records, using copy functions instead of

transcribing, having data backups, and adding metadata when possible. It is also important to standardize the museum databases implementing tools or platforms specifically designed for that purpose, so that information can be ordered, available, and updated.

The market of local researchers. If you think scientific collections as a marketplace where researchers and specimens are the products, the limited inclusion of Colombian samples in the mid-large scale herpetological literature could be said due to by two essential factors: the poor offer for the actual demand, or the reduced visibility of the products. The decline or inertia in the taxonomic work has been much discussed in the literature, including in the Colombian context (e.g., Hopkins & Freckleton, 2002; Rivera-Correa, 2012; Tancoigne & Dubois, 2013). Among the causes there are the perceptions of taxonomy as an old and boring discipline, novel ways to perform research, and new research interest beyond classifying. This results in a lack of trained taxonomists to meet the demand when a comprehensive study in taxonomy, systematics or biogeography is planned or conducted. The limited visibility of local researchers is also a problem, and may be driven by several causes, such as the perception that they lack appropriate academic degrees, they have little connection with more experienced researchers, they are young researchers with low impact, or they do not use or regularly update scientific networks (e.g., ResearchGate, ORCID, Google Scholar, Publons).

These two factors may explain in part the low inclusion of Colombian samples in the herpetological literature. To correct this requires the teaching of taxonomy not as a discipline that merely describes new taxa, but as a valuable tool that must be integrated with whatever question we want to answer (e.g., in ecology, evolution, genetics, or biogeography). In the same way we choose a statistical analysis according to our data and what we want to answer, we should be equally cautious about the taxonomy, because the principle of "garbage-in garbage-out" is applicable.

The selfish problem. Science is in principle composed and made by a human community, and is not always a peaceful and ideal one. Within science, disagreement among researchers and institutions is not a novel problem (Sherwood, 2011), and historically it has been related not solely to conflicting interest with respect to ongoing research, but to personal, political, racial or gender reasons (Lemaitre, 2015; Lemaitre, 2017; Grosso et al., 2021). Addressing this problem is really hard since it does not imply a technical, methodological or budgetary impediment, but personal beliefs and mixed feelings. Of the five reasons discussed here, it is not the most prevalent, but without a doubt it is the most dangerous.

The problem of inclusion and representativeness is not a novel issue, and it is not unique to herpetology or restricted to Colombia, but in this particular case it exhibits an historical and systematic pattern. Data from other taxa are needed to validate whether the causes proposed herein represent a phenomenon extending to other fields. I am aware that it is not always possible to include the ideal number of samplings, but I encourage authors to be prudent with the titles, inferences and conclusions derived from comprehensive studies with considerable sampling gaps, since those gaps may be masking evolutionary, biogeographic and genetic patterns. To conclude, it is important to emphasize that I am not arguing that the low inclusion of Colombian samples be debt that must be paid by including Colombian authors, on the contrary, the main objective of this text is to expose the volume of underutilized data that are waiting to be used.

Acknowledgements.– I want to thank Daniela García-Cobos, Andrew J. Crawford and Fernando A. Cervantes, for their valuable comments on the different sections of this text which helped to improve the manuscript. To Marley Gómez, for her help getting the appendix of some turtle literature. To Carlos Jiménez-Rivillas, for his comments on the statistical procedures. To Aaron Bauer, for help me with the English review. And finally, to the editor and anonymous reviewers, who made valuable comments on the final version of this text.

SUPPLEMENTARY INFORMATION

The electronic version contains supplementary material available at https://drive.google.com/drive/folders/1_nyPjwH1P9BrGJfuc7KCVwtKPGbpnEm6?usp=sharing

and https://doi.org/10.6084/m9.figshare.14368742

Appendix I. Reviewed literature for amphibians and reptiles where the Colombian samples inclusion was evaluated.

Appendix II. Museums with Colombian samples (nationals + foreigners) and data summaries.

Appendix III. Figure 2 raster maps (in WGS84 datum, cell-size 0.5° x 0.5°).

Appendix IV. Spanish version of this manuscript.

Supplementary Fig. 7. Workflow diagram of the process to request a collection permission for non-commercial scientific research in Colombia (A), and to request permission to export and/or import non-CITES biological specimens (B). Adapted from Autoridad Nacional de Licencias Ambientales (ANLA, http://portal.anla.gov.co).

CITED LITERATURE

Antonelli, A. & I. Sanmartín. 2011. Why are there so many plant species in the Neotropics? Taxon 60:403-414.

Antonelli, A., A. Zizka, F.A. Carvalho, R. Scharn, C.D. Bacon, D. Silvestro & F.L. Condamine. 2018. Amazonia is the primary source of Neotropical biodiversity. Proceedings of the National Academy of Sciences of the United States of America 115:6034-6039.

Arbeláez-Cortés, E. 2013. Knowledge of Colombian biodiversity: published and indexed. Biodiversity and Conservation 22:2875-2906.

Arbeláez-Cortés, E., M.F. Torres, D. López-Álvarez, J.D. Palacio-Mejía, A.M. Mendoza & C.A. Medina. 2015. Colombian frozen biodiversity: 16 years of the tissue collection of the Humboldt Institute. Acta Biológica Colombiana 20:163-173.

Castillo-Figueroa, D. 2018. Beyond specimens: linking biological collections, functional ecology and biodiversity conservation. Revista Peruana de Biología 25: 343-348.

Chambers, E.A. & D.M. Hillis. 2020. The Multispecies Coalescent Over-Splits Species in the Case of Geographically Widespread Taxa. Systematic Biology 69:184-193.

Cordier, J.M., O. Rojas-Soto, R. Semhan, C.R. Abdala & J. Nori. 2021. Out of sight, out of mind: Phylogenetic and taxonomic gaps imply great underestimations of the species' vulnerability to global climate change. Perspectives in Ecology and Conservation. In Press.

Dayrat, B. 2005. Towards integrative taxonomy. Biological Journal of the Linnean Society 85:407-415.

Fukushima, C., R. West, T. Pape, L. Penev, L. Schulman & P. Cardoso. 2020. Wildlife collections for scientific purposes. Conservation Biology 35:5-11.

Garnett, S.T. & L. Christidis. 2017. Taxonomy anarchy hampers conservation. Nature 546:25-27.

Grosso, J., J. Fratani, G. Fontanarrosa, M. Chuliver, A.S. Duport-Bru, R.G. Schneider, M.D. Casagranda, D.P. Ferraro, N. Vicente, M.J. Salica, L. Pereyra, R.G. Medina, C. Bessa, R. Semhan & M.C. Vera. 2021. Male homophily in South American herpetology: one of the major processes underlying the gender gap in publications. Amphibia-Reptilia:1-12.

Guedes, T.B. 2021. A Matryoshka of scales: a single specimen reveals multiple new aspects of diet and distribution of snakes. Herpetology Notes 14:385-390.

Hillis, D.M. 2019. Species Delimitation in Herpetology. Journal of Herpetology 53:3-12.

Hoorn, C., F.P. Wesselingh, H. ter Steege, M.A. Bermudez, A. Mora, J.Sevink, I. Sanmartín, A. Sanchez-Meseguer, C.L. Anderson, J.P. Figueiredo, C. Jaramillo, D. Riff, F.R. Negri, H. Hooghiemstra, J. Lundberg, T. Stadler, T. Särkinen & A. Antonelli. 2010. Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. Science 330:927-931.

Hopkins, G.W. & R.P. Freckleton. 2002. Declines in the numbers of amateur and professional taxonomists: implications for conservation. Animal Conservation 5:245-249.

Hortal, J., F. De Bello, J.A. Diniz-Filho, T.M. Lewinsohn, J.M. Lobo & R.J. Ladle. 2015. Seven shortfalls that Beset Large-Scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics 46:523-549.

Hykin, S.M., K. Bi & J.A. McGuire. 2015. Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing. PLoS ONE 10:e0141579.

Jaramillo, C. & L.H. Oviedo. 2017. Hace tiempo: Un viaje paleontológico ilustrado por Colombia. Instituto Alexander von Humboldt e Instituto Smithsonian de Investigaciones Tropicales, Bogotá.

Kaiser, H., B.I. Crother, C.M.R. Kelly, L. Luiselli, M. O'Shea, H. Ota, P. Passos, W. Schleip & W. Wüster. 2013. Best practices: In the 21st century, taxonomic decisions in herpetology are acceptable only when supported by a body of evidence and published via peer review. Herpetological Review 44:8-23.

Lemaitre, B. 2015. An Essay on Science and Narcissism: How do high-ego personalities drive research? Bruno Lemaitre, France.

Lemaitre, B. 2017. Science, narcissism and the quest for visibility. The FEBS Journal 284:875-882.

Meineke, E.K., T.J. Davies, B.H. Daru & C.C. Davis. 2018. Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B: Biological Sciences 374:20170386.

Myers, N., R.A. Mittermeier, C.G. Mittermeier, G.A. da Fonseca & J. Kent. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853-858.

Páez, V.P. 2016. Colombia's tax on wildlife studies. Science 354:191.

Rivera-Correa, M. 2012. Colombian Amphibians: Cryptic diversity and cryptic taxonomy. FrogLog 100:36-37.

Samper, C. 1998. Biodiversity research in Colombia: what we know and what we need to know. Pp. 223-230. In: Tropenbos (Ed.), Seminar Proceedings: Research in Tropical Rain Forest: Its Challenges for the Future. Tropenbos, Wageningen.

Sherwood, S. 2011. Science controversies past and present. Physics Today 64:39-44.

Tancoigne, E. & A. Dubois. 2013. Taxonomy: no decline, but inertia. Cladistics 29:567-570.

Thomson, S.A., R.L Pyle, S.T Ahyong, M Alonso-Zarazaga, J Ammirati, J.F Araya, T.L Audisio, V.M Azevedo-Santos, N Bailly, W.J Baker, M Balke, M.V Barclay, R.L Barrett, R.C Benine, J.R Bickerstaff, P Bouchard, R Bour, T Bourgoin, C.B Boyko, A.S Breure, D.J Brothers, J.W Byng, C Campbell, L.M Ceríaco, I Cernák, P Cerretti, C.H Chang, S Cho, J.M Copus, M.J Costello, A Cseh, C Csuzdi, A Culham, G D'Elía, C d'Udekemd'Acoz, M.E Daneliya, R Dekker, E.C Dickinson, T.A Dickinson, P.P vanDijk, C.D.B Dijkstra, B Dima, D.A Dmitriev, L Duistermaat, J Dumbacher, W.L Eiserhardt, T Ekrem, N.L Evenhuis, A Faille, J.L Fernández-Triana, E Fiesler, M Fishbein, B.G Fordham, A.V Freitas, N.R Friol, U Fritz, T Frøslev, V.A Funk, S.D Gaimari, G.S Garbino, A.R Garraffoni, J Geml, A.C Gill, A Gray, F.G Grazziotin, P Greenslade, E.E Gutiérrez, M.S Harvey, C.J Hazevoet, K He, X He, S Helfer, K.M Helgen, A.H vanHeteren, F Hita-Garcia, N Holstein, M.K Horváth, P.H Hovenkamp, W.S Hwang, J Hyvönen, M.B Islam, J.B Iverson, M.A Ivie, Z Jaafar, M.D Jackson, J.P Jayat, N.F Johnson, H Kaiser, BB Klitgård, D.G Knapp, J Kojima, U Kõljalg, J Kontschán, F.T Krell, I Krisai-Greilhuber, S Kullander, L Latella, J.E Lattke, V Lencioni, G.P Lewis, M.G Lhano, N.K Lujan, J.A Luksenburg, J Mariaux, J Marinho-Filho, C.J Marshall, J.F Mate, M.M McDonough, E Michel, V.F Miranda, M.D Mitroiu, J Molinari, S Monks, A.J Moore, R Moratelli, D Murányi, T Nakano, S Nikolaeva, J Noyes, M Ohl, N.H Oleas, T Orrell, B Páll-Gergely, T Pape, V Papp, L.R Parenti, D Patterson, I.Y Pavlinov, R.H Pine, P Poczai, J Prado, D Prathapan, R.K Rabeler, J.E Randall, F.E Rheindt, A.G Rhodin, S.M Rodríguez, D.C Rogers, F Roque, K.C Rowe, .LA Ruedas, J Salazar-Bravo, R.B Salvador, G Sangster, C.E Sarmiento, D.S Schigel, S Schmidt, F.W Schueler, H Segers, N Snow, P.G Souza-Dias, R Stals, S Stenroos, R.D Stone, C.F Sturm, P Šty, P Teta, D.C Thomas, R.M Timm, B.J Tindall, J.A Todd, D Triebel, A.G Valdecasas, A Vizzini, M.S Vorontsova, J.M deVos, P Wagner, L Watling, A Weakley, F Welter-Schultes, D Whitmore, N Wilding, K Will, J Williams, K Wilson, J.E Winston, W Wüster, D Yanega, D.K Yeates, H Zaher, G Zhang, Z.Q Zhang, H.Z. Zhou. 2018. Taxonomy based on science is necessary for global conservation. PLoS Biology 16:e2005075.

Vogel-Ely, C., S.A. Bordignon, R. Trevisan & I.I. Boldrini. 2017. Implications of poor taxonomy in conservation. Journal for Nature Conservation 36:10-13.

Will, K.W., B.D. Mishler & Q.D. Wheeler. 2005. The Perils of DNA Barcoding and the Need for Integrative Taxonomy. Systematic Biology 54:844-851.

Zachos, F.E. & J.C. Habel. 2014. Biodiversity Hotspots Distribution and Protection of Conservation Priority Areas. Springer, Berlin.