*Molecular Ecology*

WILEY

MOLECULAR ECOLOGY

MINI REVIEW OPEN ACCESS

# In Vouchers We (Hope to) Trust: Unveiling Hidden Errors in GenBank's Tetrapod Taxonomic Foundations

Albert Carné[1,2,3] | David R. Vieites[4] | Matthijs P. van den Burg[2]

[1]Science and Business S.L., Edificio CITEXVI, Campus Universitario de Vigo, Vigo, Galicia, Spain | [2]Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales (MNCN), CSIC, Madrid, Spain | [3]Department of Biodiversity Ecology and Evolution, Faculty of Biological Sciences, Complutense University of Madrid, Madrid, Spain | [4]Integrative Marine Ecology Group. Department of Ecology and Marine Resources, Instituto de Investigaciones Marinas (IIM), CSIC, Vigo, Galicia, Spain

**Correspondence:** David R. Vieites (david.vieites@csic.es) | Matthijs P. van den Burg (thijs.burg@gmail.com)

**Received:** 11 February 2025 | **Revised:** 12 May 2025 | **Accepted:** 19 May 2025

**Handling Editor:** Joanna Freeland

**Funding:** This work was supported by Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación 10.13039/501100011033.

**Keywords:** data quality | DNA sequencing | GenBank | genetic data | intraspecific diversity | museum specimens | repositories | taxonomy

## ABSTRACT

Genetic repositories are invaluable resources foundational to various biological disciplines. While their data and metadata reliability are essential for robust research outcomes, numerous studies have highlighted data quality and consistency issues. Here, we detect and quantify errors at the most fundamental level by analysing the congruence of sequences derived from the same genetic marker and specimen voucher across tetrapods. Our analysis reveals that 32% of re-sequenced vouchers (with identical field or museum numbers) yield unequal sequences, ranging from a few mutations to significant divergences (0.06%–33.95%). These divergences may result from sample misidentification, labelling errors, fidelity disparities between sequencing methods, or contamination at various stages of the research process. Our findings demonstrate errors within GenBank at its most basal level and suggest that, although undetectable, a similar error rate likely exists in non-re-sequenced data. These previously overlooked errors are concerning because they arise from replicated experiments, which are uncommon, and raise serious questions about the reliability of non-re-sequenced specimens. Such errors can compromise the accuracy of biodiversity assessments (e.g., taxonomic assessment, eDNA and barcoding), phylogenetic analyses and conservation planning by artificially inflating the intraspecific divergence or misidentifying (to-be-described) species. Additionally, the accuracy of large-scale biological studies that rely on such data can be compromised. Our concerning results call for protocols ensuring sample traceability to the specimens or tissues during the whole process of data generation, analysis and deposition in a database. We propose a third-party annotation system for individual GenBank records that would allow flagging common errors and alert both the original submitter and all users to potential problems without modifying the original records.

## 1 | Introduction

Intraspecific diversity is increasingly perceived as one of the most important ecological facets of biodiversity and conservation (Des Roches et al. 2018, 2021). It also plays a central role in taxonomy (Vences et al. 2021; Ramirez et al. 2023), where understanding the boundaries between intra- and interspecific variation allows for formulating taxonomic hypotheses (e.g., Puillandre et al. 2012). Concerningly, levels of intraspecific diversity remain severely under-evaluated by global surveys

(Moran et al. 2016; Laikre et al. 2020), and intraspecific diversity has long been neglected in conservation assessments and management practices (Leigh et al. 2021; Schmidt et al. 2023). Given the consequences of global change and the bleak future forecasted for biodiversity, genetic diversity must play a central role in conservation to preserve the adaptive potential and adequately represent species diversity (Pauls et al. 2013). Recently, taking advantage of genetic repositories, a new discipline has emerged to address previously intractable global hypotheses: Macrogenetics. This field explores the patterns and predictors of intraspecific genetic variation across thousands of taxa at broad taxonomic, spatial and/or temporal scales (Blanchet et al. 2017). By repurposing thousands of sequences, macrogenetics has tackled ground-breaking questions and drawn global conclusions regarding the spatial patterns of genetic diversity, its predictors and how it is impacted by anthropogenic causes such as habitat destruction and climate change (Miraldo et al. 2016; Pelletier and Carstens 2018; Leigh et al. 2019; Barrow et al. 2021; Manel et al. 2020; Millette et al. 2020; Theodoridis et al. 2020, 2021). Occasionally, macrogenetic results have been used to guide genetic diversity conservation planning (Hu et al. 2021). However, variable and contradictory patterns have already been detected across macrogenetic studies. Conclusions drawn from big data provided by third parties may be influenced by undetectable errors that can affect datasets, analyses, results and conclusions. All 'macro' disciplines must acknowledge that a certain degree of error will be included within a global framework. Although Leigh et al. (2021) reviewed the challenges and limitations within macrogenetics, they did not address a fundamental question: Does the genetic diversity contained in the repositories reflect real diversity?

Molecular data – easier, faster and cheaper to obtain than ever (Wetterstrand 2020) – have gained increasing importance across many biological fields. Within the field of taxonomy, molecular data are frequently being used (1) to resolve questions about the taxonomic identity of specimens (e.g., Rancilhac et al. 2020), (2) as a first step to characterise candidate species before their formal description (e.g., Carné and Vieites 2024) and (3) for assigning newly discovered populations to a taxonomic unit (e.g., Mullin et al. 2022). Formal species descriptions are time-consuming; ideally, they should be supported through multiple lines of evidence following an integrative taxonomic approach (Padial et al. 2009, 2010). Those can be challenging to compile, especially for cryptic, narrowly distributed species or species with low population densities. Ultimately, exploring intraspecific genetic diversity can increase the speed at which taxonomic units are identified, delimited and characterised, consequently speeding up the description of taxa. In some contexts (e.g., eDNA, DNA [meta]barcoding and turbo-taxonomy), molecular data have completely replaced morphological identification due to the ease and speed with which these data can be obtained, the potential of reproducibility at any time, the ease of having a digital format and the removed need for taxonomical expertise (Tautz et al. 2003). Generally, a standard set of preselected genetic markers (most commonly mitochondrial genetic markers using 'universal primers') have been sequenced to obtain a taxonomic assignment following comparisons with previous sequence data, e.g., to describe large numbers of species (e.g., Sharkey et al. 2021) or to rapidly assess the richness of aquatic species (see review by Rees et al. 2014). These DNA-based approaches have been criticised given multiple pitfalls (Hofstetter et al. 2019), including uncertainties about species boundaries, such as hybridisation (Fitzpatrick et al. 2015) and mito-nuclear discordance (Toews and Brelsford 2012; Després 2019).

DNA-oriented studies that assign taxonomic status and/or compare to (a-prior) well-identified specimens are highly dependent on previously published genetic data. Therefore, online repositories, in which molecular data are massively stored and publicly available, are crucial (e.g., GenBank, Sayers et al. 2025; BOLD, Ratnasingham and Hebert 2007). GenBank has seen exponential growth in the number of specimens for which sequence data are archived (Sayers et al. 2025), offering a valuable, practical and easily retrievable tool for a wide array of future DNA-oriented studies (Renner et al. 2024). As databases like GenBank are a community effort in which data from multiple researchers and sources are stored, they facilitate research advancements across multiple disciplines. Generating new genetic datasets without leveraging existing data is uncommon due to budget, resources and time constraints. Therefore, users are often dependent on third-party genetic data and rely on the quality of the deposited sequence and its associated metadata. However, errors from deposited data are common and known to cause erroneous data-driven conclusions (Fietz et al. 2013; Beaz-Hidalgo et al. 2015).

The responsibility for uploading curated sequences, with authenticated metadata and taxonomic verification, to public repositories lies with the submitting party (usually the researchers from the associated study). Only the original submitter (or GenBank staff under written agreement from the submitter) can modify or correct GenBank entries once submitted (Federhen 2015; Phillips et al. 2022). Since GenBank began to accept direct sequence submissions in 1993 (Choudhuri 2014), regular near-annual updates have included improvements and quality control mechanisms (e.g., amino acid translation, phylogenetic approaches, intraspecific divergence) that are utilised by GenBank staff to detect major errors after submission (Leray et al. 2019; Schoch et al. 2020; Sayers et al. 2024 and references therein); nevertheless, errors in GenBank sequences have been identified by multiple users (Harris 2003; Nilsson et al. 2006; Buhay 2009; Li et al. 2018; Kappel and Schröder 2020; Steinegger and Salzberg 2020; van den Burg et al. 2020; Phillips et al. 2022; van den Burg and Vieites 2023). These errors can occur at various stages of the research process, including sample misidentification, laboratory contamination, cross-sample contamination, sequencing and sequence editing. They may also arise during sample labelling, cross-labelling and even data entry errors when uploading information to repositories. Recently, Leray et al. (2019) analysed the Metazoan mitochondrial data in GenBank to assess the reliability of taxonomic labels. They concluded that GenBank is generally reliable and has low mislabelling rates at higher taxonomic levels. However, they investigated incongruences at all taxonomic levels except for species and acknowledged that the percentage of incorrect assignments increased at lower taxonomic ranks, reaching up to ca. 3.5% at the genus level. Errors at lower taxonomic levels are expected to be more prevalent, as mistakes at these levels tend to be subtle and, therefore, easily overlooked.

Vouchers are permanently preserved partial or complete specimens, usually deposited in collections, that are accessible to

other researchers and serve as essential references in biological research, including the foundation of taxonomy (Pleijel et al. 2008; Funk et al. 2018). In DNA-based analyses, vouchers play an important role as a quality control check. Sequences obtained from a voucher specimen are (in principle) inherently tied to an identified and re-examinable specimen, creating a reliable link between genetic data and the individual from which it was derived (Buckner et al. 2021). In consequence, except in cases of genetic introgression, hybridisation, ancestral polymorphism events or misidentification of voucher specimens, newly generated sequences that match the voucher-derived sequence (within an acceptable range or mismatch) will belong to the same species (Mulcahy et al. 2022). However, cases of singleton species (those with only a single sequence in GenBank) can be problematic since the identification of errors is flawed, and caution is required when using those as quality checks. Singleton species can occur in high numbers in some taxa and genetic markers (i.e., 35%–41%; van den Burg et al. 2020; van den Burg and Vieites 2023), further emphasising the need for careful evaluation. GenBank introduced the specimen_voucher qualifier in 1998, allowing authors to reference the voucher institution number (or field numbers) from which the sequence was derived. Despite the fact that GenBank routinely sends a form requesting this information when it is absent from sequence submissions, it is not a requirement, and therefore, many sequences lack those data (Federhen et al., 2009; Buckner et al. 2021). Moreover, although a standardised format for uploading voucher data is required, adherence to these guidelines is inconsistent and leads to high variability in data entries that hamper efforts to link records from the same voucher.

Here, we highlight and discuss an additional level of errors in GenBank that we discovered during the comparative analyses of sequences purportedly originating from the same voucher specimens and genetic markers. Re-sequencing a specific voucher specimen for the same gene is uncommon, leads to unnecessary costs and, in some cases, highlights a lack of prior literature review, communication between colleagues and/or scrutiny of repository data. It can happen when researchers are unaware that the specimen was already sequenced, when multi-gene datasets are created de novo for the same specimen, or when there is doubt about the identity or correctness of the available sequence. These re-sequenced vouchers allowed us to evaluate the rate of potential sequencing or labelling errors in GenBank tied to the very same specimens and genetic markers. Despite their expected congruence, these sequences exhibit notable disparities, underscoring a previously undocumented level of error within the database.

## 2 | Methods

We downloaded 16S and COI records from GenBank for Amphibians, Birds, Mammals and Reptiles. To do so, we conducted independent GenBank searches based on taxonomic group and gene, using the following queries: [Amphibia[Orgn] AND ("large subunit ribosomal" OR "16S") NOT (sp OR aff OR cf. OR unverified OR hybrid OR numt)] and [Amphibia[Orgn] AND ("cytochrome oxidase subunit 1" OR "cytochrome oxidase subunit I" OR "cytochrome c oxidase subunit I" OR "cytochrome c oxidase subunit 1" OR "COI" OR "cox1" OR "MT-CO1") AND (mitochondrion[Filter]) NOT (sp OR aff OR cf. OR unverified OR hybrid OR numt)], including all available accessions up to 1 October 2021. These search queries were repeated for each taxonomic group.

We then imported the downloaded sequences into Geneious (Prime 2021.1.1; Kearse et al. 2012), where we excluded duplicated 'NC_' accessions and extracted the specimen_voucher field from each record. To identify duplicate vouchers, we manually reviewed all voucher tags. Although GenBank promotes using the Darwin Core triplet identifier naming system (Federhen 2015), most data submitters do not follow this standard. Therefore, when comparing voucher tags, we ignored certain characters (e.g., ':', '-', and '_') to account for formatting inconsistencies and identify duplicate voucher tags.

We isolated the corresponding FASTA files for vouchers with at least two 16S- or COI-tagged accession records. We then aligned the sequences from each voucher with reference mitochondrial genomes (Amphibians, *Xenopus tropicalis* AY789013; Birds, *Gallus gallus* NC_040970; Mammals, *Canis lupus* NC_008092; Reptiles, *Anolis carolinensis* EU747728) to identify and extract only those nucleotides from each target genetic marker that were represented by at least two sequences, with a minimum length requirement of 25 base pairs (i.e., the standard length of a primer). Alignments were performed using the MUSCLE algorithm implemented in Geneious. We chose MUSCLE due to its strong performance and significantly lower computational time compared to MAFFT. We calculated the percentage of identity between sequence pairs/groups in Geneious and computed its complement to obtain dissimilarity values. Finally, we used BLAST to identify the taxonomic identity of voucher sequences exhibiting more than 5% dissimilarity. We acknowledge that divergence thresholds for mtDNA markers vary across tetrapods and genetic markers (e.g., Vences et al. 2005; Nicolas et al. 2012; Winker 2021). We recorded the scientific name and percent identity of the closest matching GenBank record. We categorised these sequences either as resulting from sequencing (or editing) errors or as cases of voucher-sequence misassignment. We considered as singletons (or potential NUMTs, see below) those sequences with a percent identity difference of more than 5% from their closest match.

## 3 | Results

Our GenBank search queries retrieved 320,307 records (116,882 16S-tagged and 203,425 COI-tagged), of which 51.6% (59.6% in 16S and 47.0% in COI) included a specimen voucher tag. Comparison of these tags indicated the presence of vouchers with multiple 16S (2365)/COI (805) records, with sequence alignments showing that 45.4% and 86.2% of these had sequence overlap, respectively (Table 1, Table S1). All sequence alignments are available as Data S1.

Among the overlapping sequences (i.e., sequences from the same voucher that were submitted separately) across tetrapods, approximately 32% (410 dissimilar sequences/1074 overlapping sequences = 38.2% in 16S and 154/694 = 22.2% in COI) yielded unequal sequences with varying degrees of sequence dissimilarity. Dissimilarity values range from 0.06% to 33.9%, with an

**TABLE 1** | Overview of retrieved 16S/COI-tagged GenBank records and stepwise filtering. Numbers in parentheses represent the percentage relative to the row above.

| 16S | Amphibians | Birds | Mammals | Reptiles | Total |
|---|---|---|---|---|---|
| Queried 16S records | 55,801 | 5632 | 24,612 | 30,837 | 116,882 |
| Records with specimen_voucher tag | 40,811 (73.1) | 1835 (32.6) | 8528 (34.7) | 18,506 (60.0) | 69,680 (59.6) |
| Vouchers with > 1 16S sequence | 2045 | 4 | 118 | 198 | 2365 |
| Vouchers with overlapping 16S seqs. | 898 (43.9) | 4 (100) | 59 (50) | 113 (57.1) | 1074 (45.4) |
| Vouchers with divergent overlapping seqs. | 336 (37.4) | 3 (75.0) | 27 (45.8) | 44 (38.9) | 410 (38.2) |
| Min divergence (%) | 0.07 | 2.54 | 0.06 | 0.12 | |
| Quartile 1 (%) | 0.27 | 4.24 | 0.21 | 0.29 | |
| Mean divergence (%) | 3.28 | 5.37 | 5.05 | 3.92 | |
| Median divergence (%) | 0.71 | 5.95 | 0.40 | 1.02 | |
| Quartile 3 (%) | 3.13 | 6.78 | 7.15 | 3.60 | |
| Max divergence (%) | 30.68 | 7.62 | 20.45 | 33.95 | |
| COI | | | | | |
| Queried COI records | 28,642 | 38,845 | 114,199 | 21,739 | 203,425 |
| Records with specimen_voucher tag | 18,732 (65.4) | 27,477 (70.7) | 35,640 (31.2) | 13,720 (63.1) | 95,569 (47.0) |
| Vouchers with > 1 COI sequence | 319 | 165 | 191 | 130 | 805 |
| Vouchers with overlapping COI seqs. | 295 (92.5) | 146 (88.5) | 154 (80.6) | 100 (76.9) | 694 (86.2) |
| Vouchers with divergent overlapping seqs. | 44 (14.9) | 43 (29.5) | 38 (24.7) | 30 (30.0) | 154 (22.2) |
| Min divergence (%) | 0.16 | 0.08 | 0.07 | 0.15 | |
| Quartile 1 (%) | 0.48 | 0.16 | 0.15 | 1.07 | |
| Mean divergence (%) | 3.26 | 4.43 | 5.64 | 11.48 | |
| Median divergence (%) | 1.45 | 0.56 | 0.46 | 14.38 | |
| Quartile 3 (%) | 3.04 | 5.81 | 11.56 | 18.68 | |
| Max divergence (%) | 17.48 | 19.78 | 27.84 | 31.01 | |

average dissimilarity of 3.5% for 16S and 5.8% for COI (Figure 1, Table 1). Overlap length differed between 25 and 1604 bp, with an average of 687 bp and a median of 570 bp. A total of 139 alignments (89 in 16S and 50 in COI) showed more than 5% dissimilarity. These include 122 voucher-sequence misassignments, 12 pairs with apparent sequencing or sequence editing problems and five pairs that were too short to allow categorisation. Of the 122 voucher-sequence misassignments, 100 had < 5% difference to the closest BLAST match (mean of 99.35%) and the remaining 22 were identified as singletons with an 89.9% mean difference to the closest BLAST match.

## 4 | Discussion

The value and utility of a database are proportionately related to the accuracy of the data it stores. Genetic and genomic databases like GenBank harbour a gigantic amount of data used across many disciplines but are inherently prone to different types of errors. If the same individual is sequenced twice at the same locus, the results should be identical. However, our assessment demonstrates that 32% of re-sequenced voucher specimens yielded different sequences across tetrapods. Our review provides unequivocal evidence of a high error rate in GenBank data, raises questions about the intraspecific diversity in the repository and highlights that studies utilising these data may be jeopardised.

The degree of genetic dissimilarity among sequences from the same voucher varied considerably. We identified cases with low divergences likely explained by sequencing or sequence editing errors, or by cross-labelling between taxonomically closely related species. Contrarily, other cases showed much higher levels of dissimilarity (> 5% and up to 33.95%; Table S1), which are likely explained by cross-sampling contamination during the laboratory work or by label mix-ups between non-closely related species during subsequent analyses or sequence submission. Identifying the exact error is not possible in hindsight, though

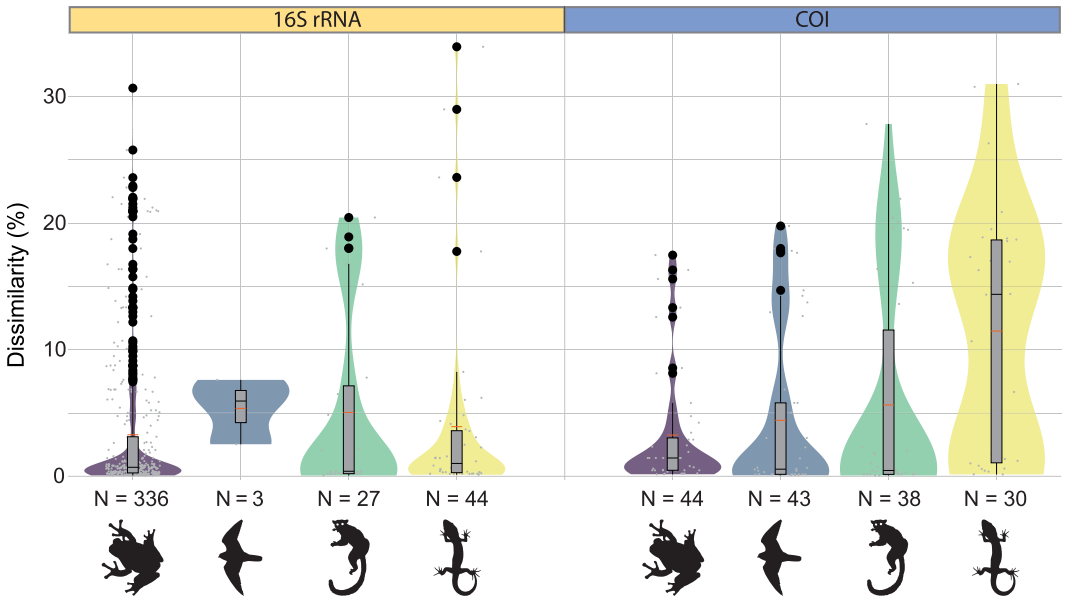


**FIGURE 1** | Genetic dissimilarity between 16S/COI-tagged GenBank accession records from identical voucher specimens across Amphibians, Birds, Mammals and Reptiles. Only pairs of sequences that exhibit divergence (i.e., non-identical sequences) are included in the plot. Orange lines represent average dissimilarity values.

our BLAST analyses reveal that voucher misattribution is likely to be the primary cause of high dissimilarity values. However, some highly dissimilar sequences lacked close matches in BLAST. This can result from two scenarios: When the sequenced individual belongs to a species that has never been sequenced before (i.e., singletons; Table S1), or when the sequence derives from a nuclear-mitochondrial pseudogene (NUMT). NUMTs are copies of mitochondrial genes integrated into the nuclear genome, evolving independently and which may diverge substantially from their mitochondrial counterpart (e.g., Song et al. 2008; Dayama et al. 2020). If the NUMT is identical or presents slight divergence, distinguishing it from intraspecific variation becomes impossible, and herein, such cases would be attributed to sequencing or sequence editing errors. Conversely, if the NUMT has diverged significantly, our analysis would flag the sequence as not matching the expected voucher, leading us to assume that a different species was sequenced. Only when a highly divergent NUMT appears as a singleton in GenBank can we suspect a pseudogenic origin. However, distinguishing it from singleton records remains impossible without voucher examination or re-sequencing.

Our results suggest that large datasets assembled from GenBank, as commonly used in high-impact macroecological and macrogenetic studies, can be compromised. We have identified artificial genetic diversity among re-sequenced voucher specimens, revealing an otherwise undetectable layer of false divergence. This raises the question of how much of the intraspecific diversity found in the database reflects true divergence (i.e., real mutations) or stems from unidentified methodological errors, generating concerns about the reliability of the biological diversity observed in non-re-sequenced specimens. All authors that utilise GenBank records rely on the assumption that the observed intraspecific variation is true variation, not artefactual. Our findings challenge this assumption. If the observed genetic variability is an artefact, it could falsely add millions of years to evolutionary phylogenetic trees, distorting the divergence age estimates, phylogeographic hypotheses and phylogenetic relationships. Similarly, if protected areas are proposed or expeditions to find candidate species are undertaken based on population units that appear unique due to methodological errors, valuable resources and conservation efforts may be misallocated. Particularly concerning is the increasing prevalence of next-generation sequencing (NGS) data in public repositories, which have higher error rates than traditional Sanger methodologies (Fox et al. 2014; Pfeiffer et al. 2018). If the errors detected so far are common, we can expect their frequency and severity to rise.

Our findings may underestimate the number of re-sequenced vouchers in GenBank due to inconsistent voucher-coding practices, including the absence of such data for 48.4% of the records (though the percentage differs among taxonomic groups, Table 1). Our effort highlights that the absence of voucher tags or their inconsistent coding undermines error filtering and detection, artificially inflating genetic and intraspecific diversity and making it difficult to determine whether a specific voucher has already been sequenced. Hence, a fair concern arises when considering genomic, metagenomic, or transcriptomic data, as these fields lack extensive comparative datasets for detecting errors, unlike mtDNA. Errors such as incorrect voucher assignments or labelling mistakes in 'omics' research are considerably harder to identify than those in mtDNA analyses. To mitigate these risks, genomic initiatives and consortia (e.g., European Reference Genome Atlas [ERGA], Vertebrate Genomes Project [VGP], Darwin Initiative, etc.) require depositing physical vouchers in collections as a reference standard. Additionally, robust traceability protocols for samples should be implemented in laboratories to minimise the potential for labelling or procedural errors. Our results highlight the urgent need for a standardised method to upload specimen data, facilitating easier consultation and retrieval. This should include strict curation of metadata

that allows only correctly formatted data when uploading sequences to ensure sample-traceability (see the template and proposal in Miralles et al. 2020); e.g., using Laboratory Information Management Systems. Standardised methods and traceable workflows are essential when sequencing type specimens. In current 'Museomics' practices (e.g., Rancilhac et al. 2020), historical type specimens are sequenced to resolve taxonomic issues that cannot be addressed without genetic data from the specimens used in the species' descriptions (i.e., the type series). These experiments are time-consuming, costly and require tissue from invaluable type specimens that should be preserved as best as possible. Therefore, strict protocols to ensure proper labelling of the resulting sequences are essential, as replicating these experiments to detect errors is unlikely.

Although human error cannot be entirely eliminated, we suggest taking advantage of the existing extensive infrastructure of GenBank to improve it further and make it more reliable for all utilising parties. Even though numerous authors have suggested and created parallel databases (e.g., Peña and Malm 2012; Bell et al. 2020; Lendemer et al. 2020), GenBank is the premier database for genetic data and is likely to remain a key source in the future. Wikifying GenBank, which would allow the community to facilitate the detection and curation of existing errors, has previously been suggested (Bidartondo 2008), although this was rejected by GenBank, who cited the potential 'chaos' it could cause (Pennisi 2008). We agree with GenBank staff that one of the reasons this database is widely used is that nobody can edit the original records. However, hundreds of published papers have identified taxon-specific errors in GenBank, many of which were found unintentionally. Recently, we found that only a minor fraction of identified erroneous GenBank records are removed or corrected (van den Burg and Vieites 2023), which means that only those users who stay up-to-date with the literature can avoid previously identified errors. GenBank members should not be responsible for identifying errors and should rely on and utilise the scientific community.

We propose developing and adding a new feature to GenBank, allowing interactive commentaries in a manner similar to iNaturalist (https://www.inaturalist.org/). Such a feature could allow identified users (based on ORCID ID) to flag and publicly comment on potential issues related to a particular GenBank accession number without modifying the original record. This could be implemented by introducing a predefined error-checking system, such as a set of checkboxes, with clearly defined labels for commonly found issues like incorrect taxonomic identification, mislabelling, sequencing errors, wrong locality, sequencing gaps, low sequencing resolution and voucher conflicts. When clicking these checkboxes, a word would automatically be added preceding the sequence description (similar to what currently happens with 'UNVERIFIED' sequences; Benson et al. 2012), and a new parameter, such as 'error: taxonomic_mismatch', would be created. This parameter, integrated into the list of parameters requested in metabarcoding queries (e.g., BLASTn; Chen et al. 2015; see BLASTn output options: https://www.metagenomics.wiki/tools/blast/) would allow users to identify the issues visually. This checkbox system would allow users who process thousands of sequences in a single search (e.g., metabarcoding), to easily filter out problematic sequences by incorporating these terms into their scripts, either to exclude these accessions or to identify them by viewing the associated issue. Flags should also be removable or a 'problem-solved' flag should be available, though making such changes should be limited to the original submitter or GenBank personnel. In addition to the checkboxes, allowing users to provide more detailed explanations by writing brief descriptions of the specific issues would offer further clarity to others interested in the same accession number or taxonomic group. This would help users better understand the nature of the problem and avoid duplicating the work already done by previous users. This solution would preserve the original sequence and metadata while adding valuable information about potential issues for future users. Additionally, a system could be implemented in which any record commented on or flagged would trigger an automatic email to the original submission author (assuming the author is still active and uses the same email address as when the sequence was submitted). This would inform submitters of any identified issues with their sequence, allowing them to review and correct the problem. While we acknowledge that the feasibility of this approach depends on the specific digital architecture of GenBank, it would allow subsequent users to identify potential issues and prevent potential errors.

### Author Contributions

Conceptualisation: Albert Carné, Matthijs P. van den Burg. Data Curation: Albert Carné, Matthijs P. van den Burg. Formal Analysis: Albert Carné, Matthijs P. van den Burg. Investigation: Albert Carné, Matthijs P. van den Burg. Methodology: Albert Carné, Matthijs P. van den Burg. Project Administration: Matthijs P. van den Burg. Supervision: Matthijs P. van den Burg. Validation: Albert Carné, Matthijs P. van den Burg, David R. Vieites. Visualisation: Albert Carné, Matthijs P. van den Burg. Writing – Original Draft Preparation: Albert Carné. Writing – Review and Editing: Albert Carné, Matthijs P. van den Burg, David R. Vieites.

### Acknowledgements

We acknowledge the colleagues who have contributed through fruitful discussions on this issue, mainly Annie Machordom and Andrea Corral-Lou. A.C. was funded by MCIN/AEI/10.13039/501100011033, contract for industrial doctorates aid DIN2021-011964. We thank the subject editor, Mark D. Scherz, and two anonymous reviewers for their valuable comments and suggestions, which have significantly improved the quality of this manuscript.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

All data used in the study are available in GenBank. Accession numbers with specimen voucher tags are provided in the Supporting Information section Table S1. All alignments used to calculate dissimilarities are available in the Supporting Information. No data with mandated deposition are included in the manuscript.

### References

Barrow, L. N., E. Masiero da Fonseca, C. E. Thompson, and B. C. Carstens. 2021. "Predicting Amphibian Intraspecific Diversity With Machine Learning: Challenges and Prospects for Integrating Traits, Geography, and Genetic Data." *Molecular Ecology Resources* 21, no. 8: 2818–2831.

Beaz-Hidalgo, R., M. J. Hossain, M. R. Liles, and M. J. Figueras. 2015. "Strategies to Avoid Wrongly Labelled Genomes Using as Example the Detected Wrong Taxonomic Affiliation for *Aeromonas* Genomes in the GenBank Database." *PLoS One* 10, no. 1: e0115813.

Bell, R. C., D. G. Mulcahy, S. W. Gotte, et al. 2020. "The Type Locality Project: Collecting Genomic-Quality, Topotypic Vouchers and Training the Next Generation of Specimen-Based Researchers." *Systematics and Biodiversity* 18, no. 6: 557–572.

Benson, D. A., I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers. 2012. "GenBank." *Nucleic Acids Research* 40: D48–D53.

Bidartondo, M. I. 2008. "Preserving Accuracy in GenBank." *Science* 319, no. 5870: 1616.

Blanchet, S., J. G. Prunier, and H. De Kort. 2017. "Time to Go Bigger: Emerging Patterns in Macrogenetics." *Trends in Genetics* 33, no. 9: 579–580.

Buckner, J. C., R. C. Sanders, B. C. Faircloth, and P. Chakrabarty. 2021. "The Critical Importance of Vouchers in Genomics." *eLife* 10: e68264.

Buhay, J. E. 2009. ""COI-Like" Sequences Are Becoming Problematic in Molecular Systematic and DNA Barcoding Studies." *Journal of Crustacean Biology* 29, no. 1: 96–110.

van den Burg, M. P., S. Herrando-Pérez, and D. R. Vieites. 2020. "ACDC, a Global Database of Amphibian Cytochrome-b Sequences Using Reproducible Curation for GenBank Records." *Scientific Data* 7, no. 1: 268.

van den Burg, M. P., and D. R. Vieites. 2023. "Bird Genetic Databases Need Improved Curation and Error Reporting to NCBI." *Ibis* 165, no. 2: 472–481.

Carné, A., and D. R. Vieites. 2024. "A Race Against Extinction: The Challenge to Overcome the Linnean Amphibian Shortfall in Tropical Biodiversity Hotspots." *Diversity and Distributions* 30: e13912.

Chen, Y., W. Ye, Y. Zhang, and Y. Xu. 2015. "High Speed BLASTN: An Accelerated MegaBLAST Search Tool." *Nucleic Acids Research* 43, no. 16: 7762–7768.

Choudhuri, S. 2014. "Chapter 5—Data, Databases, Data Format, Database Search, Data Retrieval Systems, and Genome Browsers." In *Bioinformatics for Beginners*, edited by S. Choudhuri, 77–131. Academic Press. https://doi.org/10.1016/B978-0-12-410471-6.00005-0.

Dayama, G., W. Zhou, J. Prado-Martinez, T. Marques-Bonet, and R. E. Mills. 2020. "Characterization of Nuclear Mitochondrial Insertions in the Whole Genomes of primates." *NAR Genomics and Bioinformatics* 2, no. 4: lqaa089.

Des Roches, S., L. H. Pendleton, B. Shapiro, and E. P. Palkovacs. 2021. "Conserving Intraspecific Variation for Nature's Contributions to People." *Nature Ecology & Evolution* 5, no. 5: 574–575.

Des Roches, S., D. M. Post, N. E. Turley, et al. 2018. "The Ecological Importance of Intraspecific Variation." *Nature Ecology & Evolution* 2, no. 1: 57–64.

Després, L. 2019. "One, Two or More Species? Mitonuclear Discordance and Species Delimitation." *Molecular Ecology* 28, no. 17: 3845–3847.

Federhen, S. 2015. "Type Material in the NCBI Taxonomy Database." *Nucleic Acids Research* 43, no. D1: D1086–D1098.

Fietz, K., J. A. Graves, and M. T. Olsen. 2013. "Control Control Control: A Reassessment and Comparison of GenBank and Chromatogram mtDNA Sequence Variation in Baltic Grey Seals (*Halichoerus grypus*)." *PLoS One* 8, no. 8: e72853.

Fitzpatrick, B. M., M. E. Ryan, J. R. Johnson, J. Corush, and E. T. Carter. 2015. "Hybridization and the Species Problem in Conservation." *Current Zoology* 61, no. 1: 206–216.

Fox, E. J., K. S. Reid-Bayliss, M. J. Emond, and L. A. Loeb. 2014. "Accuracy of Next Generation Sequencing Platforms." *Journal of Next Generation, Sequencing & Applications* 1, no. 1: 1–14.

Funk, V. A., R. Edwards, and S. Keeley. 2018. "The Problem With (Out) Vouchers." *Taxon* 67, no. 1: 3–5.

Harris, D. 2003. "Can You Bank on GenBank?" *Trends in Ecology & Evolution* 18: 317–319.

Hofstetter, V., B. Buyck, G. Eyssartier, S. Schnee, and K. Gindro. 2019. "The Unbearable Lightness of Sequenced-Based Identification." *Fungal Diversity* 96, no. 1: 243–284.

Hu, Y., H. Fan, Y. Chen, et al. 2021. "Spatial Patterns and Conservation of Genetic and Phylogenetic Diversity of Wildlife in China." *Science Advances* 7, no. 4: eabd5725.

Kappel, K., and U. Schröder. 2020. "Difficulties in DNA Barcoding-Based Authentication of Snapper Products due to Ambiguous Nucleotide Sequences in Public Databases." *Food Control* 118: 107–348.

Kearse, M., R. Moir, A. Wilson, et al. 2012. "Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data." *Bioinformatics* 28: 1647–1649.

Laikre, L., S. Hoban, M. W. Bruford, et al. 2020. "Post-2020 Goals Overlook Genetic Diversity." *Science* 367, no. 6482: 1083–1085.

Leigh, D. M., A. P. Hendry, E. Vázquez-Domínguez, and V. L. Friesen. 2019. "Estimated 6% Loss of Genetic Variation in Wild Populations Since the Industrial Revolution." *Evolutionary Applications* 12, no. 8: 1505–1512.

Leigh, D. M., C. B. van Rees, K. L. Millette, et al. 2021. "Opportunities and Challenges of Macrogenetic Studies." *Nature Reviews Genetics* 22, no. 12: 791–807.

Lendemer, J., B. Thiers, A. K. Monfils, et al. 2020. "The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education." *Bioscience* 70, no. 1: 23–30.

Leray, M., N. Knowlton, S. L. Ho, B. N. Nguyen, and R. J. Machida. 2019. "GenBank Is a Reliable Resource for 21st Century Biodiversity Research." *Proceedings of the National Academy of Sciences* 116, no. 45: 22,651–22,656.

Li, X., X. Shen, X. Chen, D. Xiang, R. W. Murphy, and Y. Shen. 2018. "Detection of Potential Problematic Cytb Gene Sequences of Fishes in GenBank." *Frontiers in Genetics* 9: 30.

Manel, S., P. E. Guerin, D. Mouillot, et al. 2020. "Global Determinants of Freshwater and Marine Fish Genetic Diversity." *Nature Communications* 11, no. 1: 692.

Millette, K. L., V. Fugere, C. Debyser, A. Greiner, F. J. Chain, and A. Gonzalez. 2020. "No Consistent Effects of Humans on Animal Genetic Diversity Worldwide." *Ecology Letters* 23, no. 1: 55–67.

Miraldo, A., S. Li, M. K. Borregaard, et al. 2016. "An Anthropocene Map of Genetic Diversity." *Science* 353, no. 6307: 1532–1535.

Miralles, A., T. Bruy, K. Wolcott, et al. 2020. "Repositories for Taxonomic Data: Where We Are and What Is Missing." *Systematic Biology* 69, no. 6: 1231–1253.

Moran, E. V., F. Hartig, and D. M. Bell. 2016. "Intraspecific Trait Variation Across Scales: Implications for Understanding Global Change Responses." *Global Change Biology* 22, no. 1: 137–150.

Mulcahy, D. G., R. Ibáñez, C. A. Jaramillo, et al. 2022. "DNA Barcoding of the National Museum of Natural History Reptile Tissue Holdings Raises Concerns About the Use of Natural History Collections and the Responsibilities of Scientists in the Molecular Age." *PLoS One* 17, no. 3: e0264930.

Mullin, K. E., I. M. Barata, J. Dawson, and P. Orozco-terWengel. 2022. "First Extraction of eDNA From Tree Hole Water to Detect Tree Frogs:

A Simple Field Method Piloted in Madagascar." *Conservation Genetics Resources* 14, no. 1: 99–107.

Nicolas, V., B. Schaeffer, A. D. Missoup, et al. 2012. "Assessment of Three Mitochondrial Genes (16S, Cytb, CO1) for Identifying Species in the Praomyini Tribe (Rodentia: Muridae)." *PLoS One* 7, no. 5: e36586.

Nilsson, R. H., M. Ryberg, E. Kristiansson, K. Abarenkov, K. H. Larsson, and U. Kõljalg. 2006. "Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective." *PLoS One* 1, no. 1: e59.

Padial, J. M., S. Castroviejo-Fisher, J. Koehler, C. Vila, J. C. Chaparro, and I. De la Riva. 2009. "Deciphering the Products of Evolution at the Species Level: The Need for an Integrative Taxonomy." *Zoologica Scripta* 38, no. 4: 431–447.

Padial, J. M., A. Miralles, I. De la Riva, and M. Vences. 2010. "The Integrative Future of Taxonomy." *Frontiers in Zoology* 7, no. 1: 1–14.

Pauls, S. U., C. Nowak, M. Bálint, and M. Pfenninger. 2013. "The Impact of Global Climate Change on Genetic Diversity Within Populations and Species." *Molecular Ecology* 22, no. 4: 925–946.

Pelletier, T. A., and B. C. Carstens. 2018. "Geographical Range Size and Latitude Predict Population Genetic Structure in a Global Survey." *Biology Letters* 14, no. 1: 20170566.

Peña, C., and T. Malm. 2012. "VoSeq: A Voucher and DNA Sequence Web Application." *PLoS One* 7, no. 6: e39071.

Pennisi, E. 2008. "Proposal to 'Wikify' GenBank Meets Stiff Resistance." *Science* 319: 1598–1599. https://doi.org/10.1126/science.319.5870.1598.

Pfeiffer, F., C. Gröber, M. Blank, et al. 2018. "Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing." *Scientific Reports* 8, no. 1: 10–950.

Phillips, M. J., M. Westerman, and M. Cascini. 2022. "The Value of Updating GenBank Accessions for Supermatrix Phylogeny: The Case of the New Guinean Marsupial Carnivore Genus *Myoictis*." *Molecular Phylogenetics and Evolution* 166: 107–328.

Pleijel, F., U. Jondelius, E. Norlinder, et al. 2008. "Phylogenies Without Roots? A plea for the Use of Vouchers in Molecular Phylogenetic Studies." *Molecular Phylogenetics and Evolution* 48, no. 1: 369–371.

Puillandre, N., A. Lambert, S. Brouillet, and G. J. M. E. Achaz. 2012. "ABGD, Automatic Barcode Gap Discovery for Primary Species Delimitation." *Molecular Ecology* 21, no. 8: 1864–1877.

Ramirez, J. L., P. Valdivia, U. Rosas-Puchuri, and N. L. Valdivia. 2023. "SPdel: A Pipeline to Compare and Visualize Species Delimitation Methods for Single-Locus Datasets." *Molecular Ecology Resources* 23, no. 8: 1959–1965.

Rancilhac, L., T. Bruy, M. D. Scherz, et al. 2020. "Target-Enriched DNA Sequencing From Historical Type Material Enables a Partial Revision of the Madagascar Giant Stream Frogs (Genus *Mantidactylus*)." *Journal of Natural History* 54, no. 1–4: 87–118.

Ratnasingham, S., and P. D. Hebert. 2007. "BOLD: The Barcode of Life Data System." *Molecular Ecology Notes* 7, no. 3: 355–364.

Rees, H. C., B. C. Maddison, D. J. Middleditch, J. R. Patmore, and K. C. Gough. 2014. "The Detection of Aquatic Animal Species Using Environmental DNA–a Review of eDNA as a Survey Tool in Ecology." *Journal of Applied Ecology* 51, no. 5: 1450–1459.

Renner, S. S., M. D. Scherz, C. L. Schoch, M. Gottschling, and M. Vences. 2024. "Improving the Gold Standard in NCBI GenBank and Related Databases: DNA Sequences From Type Specimens and Type Strains." *Systematic Biology* 73, no. 2: 486–494.

Sayers, E. W., M. Cavanaugh, K. Clark, et al. 2024. "GenBank 2024 update." *Nucleic Acids Research* 52, no. D1: D134–D137.

Sayers, E. W., M. Cavanaugh, L. Frisse, et al. 2025. "GenBank 2025 update." *Nucleic Acids Research* 53, no. D1: D56–D61.

Schmidt, C., S. Hoban, M. Hunter, I. Paz-Vinas, and C. J. Garroway. 2023. "Genetic Diversity and IUCN Red List Status." *Conservation Biology* 37, no. 4: e14064.

Schoch, C. L., S. Ciufo, M. Domrachev, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database* 2020: baaa062.

Sharkey, M. J., D. H. Janzen, W. Hallwachs, et al. 2021. "Minimalist Revision and Description of 403 New Species in 11 Subfamilies of Costa Rican Braconid Parasitoid Wasps, Including Host Records for 219 Species." *ZooKeys* 1013: 1–665.

Song, H., J. E. Buhay, M. F. Whiting, and K. A. Crandall. 2008. "Many Species in One: DNA Barcoding Overestimates the Number of Species When Nuclear Mitochondrial Pseudogenes Are Coamplified." *Proceedings of the National Academy of Sciences of the United States of America* 105, no. 36: 13,486–13,491.

Steinegger, M., and S. L. Salzberg. 2020. "Terminating Contamination: Large-Scale Search Identifies More Than 2,000,000 Contaminated Entries in GenBank." *Genome Biology* 21: 1–12.

Tautz, D., P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler. 2003. "A plea for DNA Taxonomy." *Trends in Ecology & Evolution* 18, no. 2: 70–74.

Theodoridis, S., D. A. Fordham, S. C. Brown, S. Li, C. Rahbek, and D. Nogues-Bravo. 2020. "Evolutionary History and Past Climate Change Shape the Distribution of Genetic Diversity in Terrestrial Mammals." *Nature Communications* 11, no. 1: 2557.

Theodoridis, S., C. Rahbek, and D. Nogues-Bravo. 2021. "Exposure of Mammal Genetic Diversity to Mid-21st Century Global Change." *Ecography* 44, no. 6: 817–831.

Toews, D. P., and A. Brelsford. 2012. "The Biogeography of Mitochondrial and Nuclear Discordance in Animals." *Molecular Ecology* 21, no. 16: 3907–3930.

Vences, M., A. Miralles, S. Brouillet, et al. 2021. "iTaxoTools 0.1: Kickstarting a Specimen-Based Software Toolkit for Taxonomists." *Megataxa* 6: 77–92.

Vences, M., M. Thomas, A. Van der Meijden, Y. Chiari, and D. R. Vieites. 2005. "Comparative Performance of the 16S rRNA Gene in DNA Barcoding of Amphibians." *Frontiers in Zoology* 2: 1–12.

Wetterstrand, K. 2020. *DNA sequencing costs*. National Human Genome Research Institute. www.genome.gov/sequencingcostsdata.

Winker, K. 2021. "An Overview of Speciation and Species Limits in Birds." *Ornithology* 138, no. 2: ukab006.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.