Ornithology AmericanOrnithology.org

Volume 139, 2022, pp. 1–15
https://doi.org/10.1093/ornithology/ukac045

RESEARCH ARTICLE

Updating splits, lumps, and shuffles: Reconciling GenBank names with standardized avian taxonomies

Peter A. Hosner,[1,2,*] Min Zhao,[3] Rebecca T. Kimball,[3] Edward L. Braun,[3] and J. Gordon Burleigh[3]

[1] Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[2] Center for Global Mountain Biodiversity, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[3] Department of Biology, University of Florida, Gainesville, Florida, USA
* Corresponding author: peter.hosner@snm.ku.dk

Submission Date: February 7, 2022; Editorial Acceptance Date: July 27, 2022; Published August 26, 2022

ABSTRACT

Biodiversity research has advanced by testing expectations of ecological and evolutionary hypotheses through the linking of large-scale genetic, distributional, and trait datasets. The rise of molecular systematics over the past 30 years has resulted in a wealth of DNA sequences from around the globe. Yet, advances in molecular systematics also have created taxonomic instability, as new estimates of evolutionary relationships and interpretations of species limits have required widespread scientific name changes. Taxonomic instability, colloquially "splits, lumps, and shuffles," presents logistical challenges to large-scale biodiversity research because (1) the same species or sets of populations may be listed under different names in different data sources, or (2) the same name may apply to different sets of populations representing different taxonomic concepts. Consequently, distributional and trait data are often difficult to link directly to primary DNA sequence data without extensive and time-consuming curation. Here, we present RANT: Reconciliation of Avian NCBI Taxonomy. RANT applies taxonomic reconciliation to standardize avian taxon names in use in NCBI GenBank, a primary source of genetic data, to a widely used and regularly updated avian taxonomy: eBird/Clements. Of 14,341 avian species/subspecies names in GenBank, 11,031 directly matched an eBird/Clements; these link to more than 6 million nucleotide sequences. For the remaining unmatched avian names in GenBank, we used Avibase's system of taxonomic concepts, taxonomic descriptions in Cornell's Birds of the World, and DNA sequence metadata to identify corresponding eBird/Clements names. Reconciled names linked to more than 600,000 nucleotide sequences, ~9% of all avian sequences on GenBank. Nearly 10% of eBird/Clements names had nucleotide sequences listed under 2 or more GenBank names. Our taxonomic reconciliation is a first step towards rigorous and open-source curation of avian GenBank sequences and is available at GitHub, where it can be updated to correspond to future annual eBird/Clements taxonomic updates.

Keywords: big data, DNA sequence data, genomics, NCBI, nomenclature

LAY SUMMARY

- 23% of avian names on GenBank do not match eBird/Clements, a widely used standardized avian taxonomy.
- More than 600,000 nucleotide sequences on GenBank are associated with names that do not match eBird/Clements.
- 10% of eBird/Clements names have nucleotide sequences listed under multiple GenBank names.
- We provide an open-source taxonomic reconciliation to mitigate difficulties associated with non-standardized name use for GenBank sequences.

Actualizando divisiones, agrupamientos y reorganizaciones: reconciliando nombres de GenBank con taxonomías aviares estandarizadas

RESUMEN

La investigación sobre biodiversidad ha avanzado al evaluar las expectativas de las hipótesis ecológicas y evolutivas a través de la vinculación de bases de datos genéticos, de distribución y de rasgos a gran escala. El auge de la sistemática molecular en los últimos 30 años ha dado como resultado una gran cantidad de secuencias de ADN de todo el mundo. Sin embargo, los avances en la sistemática molecular también han creado inestabilidad taxonómica, ya que las nuevas estimaciones de las relaciones evolutivas y las interpretaciones de los límites de las especies han requerido cambios generalizados en los nombres científicos. La inestabilidad taxonómica, coloquialmente llamada "divisiones, agrupamientos y reorganizaciones," presenta desafíos logísticos para la investigación de la biodiversidad a gran escala porque (1) las mismas especies o conjuntos de poblaciones pueden estar listados con diferentes nombres en diferentes fuentes de datos, o (2) el mismo nombre puede aplicarse a diferentes conjuntos de poblaciones representando diferentes conceptos taxonómicos. En consecuencia, los datos de distribución y rasgos a menudo son difíciles de

vincular directamente a los datos primarios de secuencias de ADN sin una curación extensa y demandante de tiempo. Aquí, presentamos RANT (por sus siglas en inglés): reconciliación de la taxonomía aviar del Centro Nacional para la Información Biotecnológica (CNIB). RANT aplica la reconciliación taxonómica para estandarizar los nombres de taxones aviares en uso en el GenBank de CNIB, una fuente principal de datos genéticos, con la taxonomía aviar ampliamente utilizada y actualizada periódicamente de eBird/Clements. De los 14.341 nombres de especies/subespecies de aves en GenBank, 11.031 coincidieron directamente con eBird/Clements; estos se vinculan a más de 6 millones de secuencias de nucleótidos. Para los restantes nombres de aves no coincidentes en GenBank, utilizamos el sistema de conceptos taxonómicos de Avibase, descripciones taxonómicas en Aves del Mundo de Cornell y metadatos de secuencias de ADN para identificar los nombres correspondientes de eBird/Clements. Los nombres reconciliados vincularon a más de 600.000 secuencias de nucleótidos, ~9% de todas las secuencias de aves en GenBank. Casi el 10% de los nombres de eBird/Clements tuvieron secuencias de nucleótidos enumeradas bajo dos o más nombres en GenBank. Nuestra reconciliación taxonómica es un primer paso hacia la curación rigurosa y de código abierto de las secuencias aviares de GenBank y está disponible en GitHub, donde se puede actualizar para que corresponda con las futuras actualizaciones taxonómicas anuales de eBird/Clements.

Palabras clave: Centro Nacional para la Información Biotecnológica, datos de secuencias de ADN, genómica, macrodatos, nomenclatura

INTRODUCTION

Public data repositories are rich information sources constituting vital infrastructure for integrative research in organismal biology. As a taxonomic group, birds are well suited to these endeavors. Their global ubiquity, relative ease of observance and identification, and charismatic appearances lend to their enduring popularity among professional and recreational scientists alike. The quantity and extent of avian data have proliferated in recent years, a direct result of efforts to grow and share these data (Table 1). The information available documenting and describing avian genetics, population dynamics, distributions, behaviors, and physical traits has become truly staggering.

To leverage the vast wealth of avian information and effectively implement phylogenetic comparative methods (Felsenstein 1985) and other evolutionary analyses, it is crucial to have clear, one-to-one linkage between data records and the populations of organisms from which they are derived. Over the past 30 years, molecular systematics has wholly transformed avian taxonomy and nomenclature (Sangster 2009, Gill 2014, Barrowclough et al. 2016). Its insights have reorganized the avian tree of life (Beresford et al. 2005, Hackett et al. 2008, Lovette et al. 2010, Moyle et al. 2012, Jarvis et al. 2014, Braun et al. 2019, Oliveros et al. 2019, Harvey et al. 2020) and reformed practical applications of species limits (e.g., Andersen et al. 2014, Hosner et al. 2018). An unfortunate consequence of these much-needed reorganizations is that they often require changes to organisms' scientific names. In modern implementations of the Linnaean system of nomenclature, higher taxa (e.g., genus, family, order) are required to be monophyletic. Hence, any move of a species to a different genus or an update of species limits requires scientific name changes for at least some populations.

Identifying and tracking avian nomenclatural changes over time is itself a difficult task. As an example, we compared two major taxonomic works completed before DNA sequencing technology was widely available in ornithology, the Peters checklist series (Peters et al. 1931–1987) and Sibley and Monroe (1993), to the eBird/Clements v2019 (Clements et al. 2019) list (Data Repository D1). Only 6,288 of the 9,204 (69%) Peters checklist species names, and 7,470 of 9,702 (77%) of the Sibley and Monroe (1993) species names matched exactly to the 10,721 eBird/Clements v2019 (Clements et al. 2019) names. Allowing the last 2 letters of the species epithet to mismatch, to account for minor differences in spelling, only improved name matching slightly (27 more matches for the Peters checklist, 20 more matches for Sibley and Monroe 1993). Although the details of broad list comparisons will vary depending on exactly which taxonomies are compared, all modern avian taxonomies differ substantially from corresponding works produced only decades ago.

In addition to instability stemming from name changes through time, another contributor to scientific name instability is the existence of multiple competing standardized avian taxonomies. Currently, there are 4 main global choices: (1) eBird/Clements v2021 (Clements et al. 2021), (2) IOC World Bird List (Gill et al. 2022), (3) *Howard and Moore Complete Checklist of Birds of the World* (Dickinson and Remsen 2013, Dickinson and Christidis 2014), and (4) HBW/BirdLife Taxonomic Checklist (HBW and BirdLife International 2022). Although similar in many respects, each of these lists is governed differently, is updated at different intervals, and applies species recognition criteria differently (Garnett and Christidis 2017). For example, in raptors, a paraphyletic assemblage of predatory non-passerine landbirds, McClure et al. (2020) found that major world lists disagreed in species-level name application in 11–25% of cases. Beyond these most-referenced world lists, there are additional regional and country-specific avian taxonomies.

Different biodiversity databases use different underlying taxonomies, requiring users to reconcile names between sources (Boyle et al. 2013, Lepage et al. 2014) before downstream analyses are prudent. Some large avian data sources

TABLE 1. Examples, descriptions, links, and references to large global databases that contain avian data.

Name	Description	Website	Citation
INSDC	International Nucleotide Sequence Database Collaboration; Oversees sequence archives	https://www.insdc.org	(Cochrane et al. 2011)
iDigBio	Metadata derived from museum specimens; Integrated Digitized Biocollections	https://www.idigbio.org	
DiSSCo	Distributed System of Scientific Collection; Metadata from museum specimens	https://www.dissco.eu	(Nelson and Paul 2019)
ARCTOS	Metadata from museum specimens	https://arctos.database.museum	(Cicero et al. 2017)
Vertnet	Metadata from museum specimens	http://vertnet.org	(Constable et al. 2010)
GGBN	Global Genome Biodiversity Network, tissue resources for genomic sequencing	https://www.ggbn.org	(Droege et al. 2014)
Morphosource	Morphological data	https://www.morphosource.org	(Boyer et al. 2016)
Macaulay Library	Digital media, sound recordings, photos, videos	https://www.macaulaylibrary.org	
Xeno-Canto	Sound recordings	https://www.xeno-canto.org/	
AVoCet	Sound recordings	https://avocet.integrativebiology.natsci.msu.edu/	
MoveBank	Movement data	www.movebank.org	(Kranstauber et al. 2011)
eBird	Citizen Science locality data, linked to Macaulay Library	https://ebird.org	(Sullivan et al. 2009)
iNaturalist	Citizen Science locality data, photos, sound recordings	https://www.inaturalist.org	(Unger et al. 2020)
Birds of the World	Encyclopedic synthesis of avian life history information	https://birdsoftheworld.org	
AVONET	Synthesis of various data	https://figshare.com/s/b990722d72a26b5bfead	(Tobias et al. 2022)
EltonTraits	Synthesis of various data	https://opentraits.org/datasets/elton-traits.html	(Wilman et al. 2014)
BirdLife International Datazone	Synthesis of various data	http://datazone.birdlife.org/home	
GBIF	Global Biodiversity Information Facility. Spatial Data aggregator	https://www.gbif.org/	(Robertson et al. 2014)

employ standardized global avian taxonomies from the start. For example, eBird (https://www.eBird.org/) and the Macaulay Library (https://www.macaulaylibrary.org/) use the related eBird/Clements taxonomy, which is usually updated annually. On the other hand, Xeno-canto (https://www.xeno-canto.org/) uses the IOC World Bird List, which is updated twice a year—although it has been updated quadrennially in the past.

Using standardized taxonomies for databases vastly improves the ability for users to identify discrepancies between name usage and application, especially through the links between "taxonomic concepts" as implemented in Avibase (Lepage et al. 2014, McClure et al. 2020). Taxonomic concepts are circumscriptions of a scientific name applied by a specific taxonomic authority. Because scientific names are applied differently by authors through time, the same scientific name may refer to different sets of populations, or different scientific names may refer to the same sets of populations. Avibase organizes linkages between equivalent and partially matching taxonomic concepts, so that users can potentially match previously used names to those currently considered valid by various taxonomic authorities.

Name reconciliation can be trivial when working only with a few familiar taxa, but it requires extraordinary time and effort when managing mismatching taxonomic concepts, large numbers of taxa, and when working at global scales. Taxonomic reconciliation becomes even more difficult and time-consuming when data sources implement their own taxonomy *de novo* in lieu of a standardized list, or when data sources lack consistent name use. For databases where the taxonomic names are not readily traceable, it can be impossible to correctly link information from one database to another without Supplemental Information. Failure to link names correctly may cause available information to be ignored, excluded, or worse—attributed to the wrong population (McClure et al. 2020). This issue could be particularly problematic for bird groups in geographic regions disproportionately affected by taxonomic progress (Neate-Clegg et al. 2021), or for poorly known birds with limited data, such as rare or endangered taxa. In some cases, opportunities to better understand these regions and their birdlife could be lost simply because of taxonomic instability.

The National Center for Biotechnology Information (NCBI) GenBank (Benson et al. 2012), a partner of the International Nucleotide Sequence Database Collaboration (INSDC), is the major data repository and distributor for genetic data used in phylogenetic analyses. Accurate phylogenetic inference underpins most modern comparative studies, and hence it is necessary to confront naming issues in GenBank data before assembling large-scale, synthetic phylogenies (Jetz et al. 2012, Burleigh et al. 2015) and before linking such phylogenies to other comparative datasets (Pigot et al. 2018). Although GenBank implements policies to standardize names (Schoch et al. 2020), it does not rely on any single standardized avian taxonomy. GenBank policy states that taxonomic names must be published and valid, but in practice, names are user-submitted and sometimes informal. Furthermore, as names are updated and changed by some or all standardized avian taxonomies, GenBank largely relies on the original data uploader to curate and update records. This can lead to problems in light of taxonomic instability. For example, when the name of a species changes (e.g., it is moved to a new genus, or a different specific epithet is used), sequences may be organized under both former and present names. Hence, a researcher may obtain some sequences for a given taxon, but may not realize that other sequence data exist. Worse, a user may assume no data exist for a given taxon, as it could be listed under a former name without current acceptance in standardized lists. Additional uncertainties arise when species are split into two or more entities, or when species are lumped yet remain listed on GenBank under multiple names.

Ultimately, the only way to link GenBank sequences with other types of comparative data is to reconcile GenBank's avian names to standardized avian taxonomies. One strategy is producing an open source, parallel data structure, which can be curated and updated as avian taxonomy changes (Leray et al. 2020, Riginos et al. 2020). Each GenBank name has a unique numerical identifier (TaxID; Schoch et al. 2020), and each GenBank database record has a unique identifier. Using these identifiers, it is possible to link one or many names, corresponding to standardized lists. Here, we attempt such a reconciliation, linking GenBank taxon identifiers to the eBird/Clements v2019 (Clements et al. 2019) list for all avian GenBank TaxIDs. To further explore the extent to which taxonomic instability and its biases affect birds, we summarize avian data patterns related to taxonomic groups, geographical areas, and conservation status. Finally, we summarize the extent to which name-reconciled sequences apply to large comparative databases, namely the Macaulay Library and Xeno-canto sound archives, using the GenBank Nucleotide database, the GenBank product with the broadest taxonomic coverage.

Our goal was to reconcile the taxonomic names in GenBank (TaxIDs) to a major avian taxonomy in order to link GenBank sequences, and phylogenetic trees built from these sequences, to ancillary data sources. We did not attempt to equate Avibase taxonomic concepts with TaxIDs, because GenBank TaxIDs can be applied differently by different sequence authors, or at different times. Hence, the more challenging task of identifying taxonomic concepts for individual GenBank sequences is more prudently done following an initial reconciliation with TaxIDs. We selected eBird/Clements v2019 (Clements et al. 2019) as the focal

standardized bird list, because of its use in the world's largest bird observation dataset (eBird), its related media resources (Macaulay Library), and its linked Birds of the World information content. Existing tools can reconcile the eBird/Clements list with other standardized taxonomies (Lepage et al. 2014, Gill et al. 2022). Hence, once GenBank names are linked to a single standardized taxonomy, in this case, eBird/Clements, reconciling to other standardized taxonomies (IOC, BirdLife International, Howard and Moore) is straightforward.

METHODS

Taxonomic Reconciliation

We downloaded all names from the NCBI Taxonomy database (Schoch et al. 2020) that descended from "Aves" (TaxID: 8782) on May 3, 2020 (Data Repository D2). From this list, we extracted all species and subspecies names as well as their NCBI Taxonomy ID (TaxID) numbers. We then ran a custom Perl script (Data Repository D3) to exactly match binomial (genus, species) and trinomial (genus, species, subspecies) names from NCBI Taxonomy to the names recognized by eBird/Clements v2019 (Clements et al. 2019; Data Repository D4). For each mismatch with the NCBI Taxonomy name, we then identified the corresponding equivalent eBird/Clements species or subspecies. We first searched for names in Avibase (Lepage et al. 2014). However, Avibase's search function currently facilitates only exact matches to taxonomies it implements. For names that were not an exact match to an Avibase taxonomic concept, we implemented web searches (Google) which often identified minor spelling differences, consulted Cornell's Birds of the World Online (https://birdsoftheworld.org), and consulted relevant literature—often the papers that first published those sequence data.

We classified 9 categories of naming mismatches resulting from discrepancies between GenBank and eBird/Clements names: split, lump, shuffle, new, spelling, hybrid, extinct, domesticated, and unidentified (Table 2). "Split" is a name that corresponds to a subspecies rank in GenBank, but a species rank in eBird/Clements. For example, the GenBank subspecies name *Otus megalotis everetti* (taxiid: 56274) corresponds to the species name *Otus everetti* in eBird/Clements. "Lump" is a name that corresponds to species rank in GenBank, but a subspecies rank in eBird/Clements. For example, the GenBank name *Megascops colombianus* (TaxID: 1740167) corresponds to *Megascops ingens colombianus* in eBird/Clements. "Shuffle" is a taxon that has an equivalent rank in GenBank and eBird/Clements, but different name usage. Most often shuffles stem from changes in genera, but a few species epithets have changed because of new evidence regarding nomenclature priority. For example, the GenBank name *Mimizuku gurneyi* (id: 56287) corresponds to *Otus gurneyi* in eBird/Clements, reflecting a change in the generic name. "New" is a species or subspecies that was undescribed when its sequences were initially uploaded to GenBank. To preserve nomenclature priority, GenBank avoids unpublished or in press names of undescribed taxa, instead assigning an informal placeholder name. Typically, the placeholder name consists of the genus, the data uploaders' initials, and the year of first upload. For example, Megascops_sp._SMD-2015 (TaxID: 1740173) corresponds to the Santa Marta Screech-Owl *Megascops gilesi* Krabbe 2017. "Spelling" is a taxon that has an equivalent name in GenBank and eBird/Clements, but for which a slightly different spelling is implemented. For example, the GenBank name *Glaucidium nanum* (TaxID: 126809) corresponds to the eBird/Clements name *Glaucidium nana*. "Hybrid" is a hybrid individual and usually identified in GenBank by a name comprising the putative parental species separated by a cross "x." For example, the GenBank name *Strix occidentalis* x *Strix varia*. Hybrids were not reconciled to eBird/Clements names, although eBird taxonomy does include and organize names for some frequent avian hybrid parental combinations. "Extinct" is an extinct taxon that is not regulated by eBird/Clements because it was not documented in the modern era. For example, the elephant bird *Aepyornis maximus* (TaxID: 748142) is known from Holocene bones and eggshell materials that have yielded DNA sequences, but this name is not regulated by eBird/Clements. "Domesticated" is a domesticated breed or line. For example, GenBank has a listing for the domesticated "Society Finch" as *Lonchura striata domestica* (TaxID: 299123), but in eBird/Clements it refers to *Lonchura striata* because domesticated forms are not generally considered subspecies. Finally, "Unidentified" refers to TaxIDs where we were unable to assign a species name. These were generally samples not identified to species, or environmental DNA samples.

We summarized the total number and proportion of reconciled GenBank TaxIDs by bird orders, and within the largest bird order Passerformes, by families. We also summarized the number of GenBank nucleotide sequences and number of reconciliations for each IUCN conservation status category. For a taxon that did not have a direct match to an IUCN name, we placed it under "Not Assessed."

GenBank Sequences Associated with Avian Names

We tallied the number of core nucleotide sequences in GenBank associated with each taxonomic ID by downloading the "nucl_gb.accession2TaxID" file on November 2, 2020 (Data Repository D5). This file lists the accession number for each sequence in the GenBank nucleotide database and its corresponding taxonomic ID number. From this, we wrote a Perl script (Data Repository D6) to count the number of nucleotide sequences associated with each taxonomic ID corresponding to an avian

TABLE 2. Categories, descriptions, and examples of name mismatches between GenBank and eBird/Clements names.

Name	Description	Example
New	A species or subspecies that was undescribed when its sequences were uploaded to GenBank	Megascops_sp_SMD-2015 (TaxID: 1740173) corresponds to the Santa Marta Screech-Owl *Megascops gilesi* Krabbe, 2017
Lump	A name that corresponds to species rank in GenBank, but a subspecies rank in eBird/Clements	GenBank name *Megascops colombianus* (TaxID: 1740167) corresponds to *Megascops ingens colombianus* in eBird/Clements
Split	A name that corresponds to a subspecies rank in GenBank, but a species rank in eBird/Clements	GenBank subspecies name *Otus megalotis everetti* (taxiid: 56274) corresponds to the species name *Otus everetti* in eBird/Clements
Shuffle	A taxon that has an equivalent rank in GenBank and eBird/Clements, but different name usage	GenBank name *Mimizuku gurneyi* (id: 56287) corresponds to *Otus gurneyi* in eBird/Clements
Spelling	A taxon that has an equivalent name in GenBank and eBird/Clements, but a slightly different spelling is implemented	GenBank name *Glaucidium nanum* (TaxID: 126809) corresponds to the eBird/Clements name *Glaucidium nana*
Hybrid	A hybrid individual and usually identified in GenBank by a name comprising the putative parental species separated by a cross "x"	GenBank name *Strix occidentalis* x *Strix varia*, which was not reconciled to an eBird/Clements name, although eBird taxonomy does include and organize names for some frequent avian hybrid parental combinations
Extinct	An extinct taxon that is not regulated by eBird/Clements because it was not documented in the modern era	*Aepyornis maximus* (TaxID: 748142) is known from Holocene bones and eggshell materials that have yielded sequenceable DNA, but this name is not regulated by eBird/Clements
Domesticated	A domesticated breed or line	GenBank has a listing for the domesticated "Society Finch" as *Lonchura striata domestica* (TaxID: 299123), but in eBird/Clements it refers to *Lonchura striata* because domesticated forms are not generally considered valid subspecies
Unidentified	Refers to TaxIDs where we were unable to assign a species name	Generally samples not identified to species, or environmental DNA samples

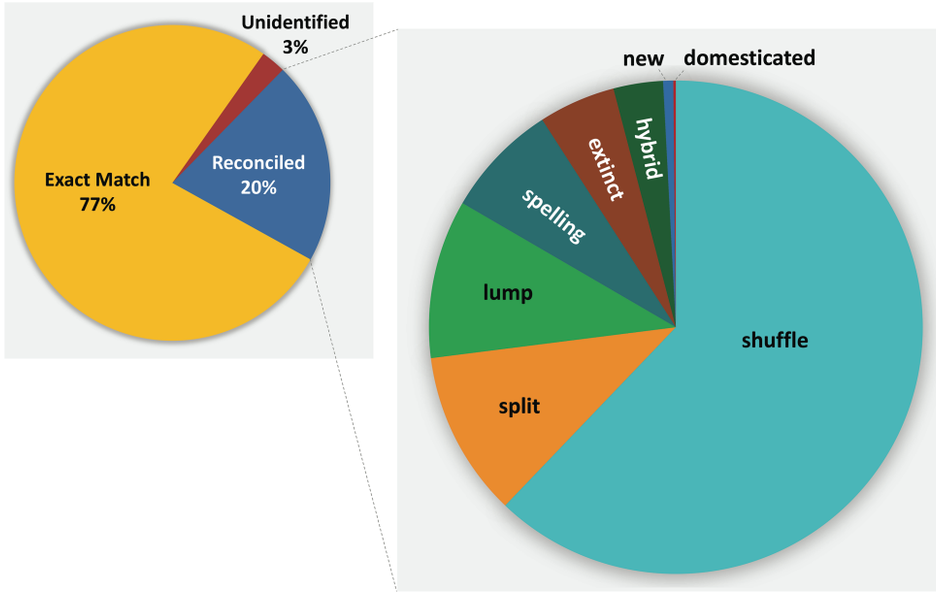

FIGURE 1. Proportions of GenBank species and subspecies names that are directly matched to eBird/Clements names (Exact Match), manually reconciled to eBird/Clements names (Reconciled), and unidentifiable that include taxa not identified to the species-level or erroneous taxa (Unidentified).

taxonomic IDs. To obtain counts of the number of runs in the NCBI Sequence Read Archive (SRA) associated with each bird species, we downloaded the *RunInfo* for the SRA runs (*SraRunInfo.csv*) within "Aves" on August 1, 2021 (Data Repository D7). To obtain counts of the number of genome sequences in GenBank associated with each name, we downloaded from NCBI on September 5, 2021 a summary of the NCBI Genome files ("genome_result.txt") within "Aves" (Data Repository D8).

Linking eBird/Clements Names to Geographic Realms

For TaxIDs that were successfully assigned to eBird/Clements species names (either by direct name match or taxonomic reconciliation), we delimited their geographic realms using the associated IOC breeding ranges (8 terrestrial realms and 4 oceanic realms). Here we implemented IOC, rather than eBird/Clements geographic information because eBird/Clements does not summarize species occurrence by geographic realm. We also manually assigned geographic realms for species without range information available in the IOC v10.1 checklist (master_ioc_list_v10.1.xlsx). We defined species that occur in only 1 realm as realm endemics, and species that occur in 2 or more realms as widespread. We then summarized the number of reconciliations and the number of GenBank nucleotide sequences for each realm, and widespread species.

Linking eBird/Clements Names to Other Databases

We used audio data as an example to examine the extent to which name-reconciled GenBank sequences apply to large avian comparative databases, such as the Macaulay Library and Xeno-canto. Because the Macaulay Library uses eBird/Clements taxonomy for its bird images, audios and videos, we can readily link these media resources to the GenBank nucleotide data under the same eBird/Clements names. We downloaded a summary of available audio data (April 2021) from the Macaulay Library (https://www.macaulaylibrary.org/resources/media-target-species/; Data Repository D9). We also examined Xeno-canto, a global avian vocalization database, which uses the IOC taxonomy. To match Xeno-canto's 10,909 avian names to eBird/Clements names, we filtered out the species with a direct name match and then reconciled the remaining using Avibase taxonomic concepts. Lastly, we summed up the number of Xeno-canto sound recordings (October 2020; https://www.xeno-canto.org/collection/species/all; Data Repository D10) under the same eBird/Clements name. For example, the Xeno-canto name *Colinus leucopogon* had 26 sound recordings and *Colinus cristatus* had 57, but the eBird/Clements name *C. cristatus* would have 83, because *C. leucopogon* is treated as a subspecies of *C. cristatus* by eBird/Clements.

RESULTS

Descriptive Statistics of Taxonomic Reconciliation

Of 14,341 GenBank species and subspecies TaxIDs within Aves, we were able to exactly match an eBird/Clements name for 11,031 (77%; Data Repository D11; Figure 1). Of the 3,310 GenBank names without an exact match, we

were able to reconcile 2917 to eBird/Clements names using Avibase taxonomic concepts and other sources. Thus, 23% of GenBank names needed reconciliation to match with eBird/Clements names. We were able to reconcile 88% of these non-exact-matching names, By far, the most frequent cause of discrepancy between GenBank and eBird/Clements names were shuffles (64%), most often because of a genus name change. Splits (11%) and lumps (11%), owing to classification differences at species/subspecies ranks, were nearly equally frequent. Spelling discrepancies (5%), names of extinct taxa not included in eBird/Clements (4%), and hybrids (3%) were relatively infrequent. Finally, only a few new species names (0.7%) or names used for domestic breeds (0.2%) contributed to naming discrepancies. In total, we were unable to assign 393 (3%) of GenBank names to eBird/Clements names.

Following reconciliation, we found that 9,361 eBird/Clements species names had at least one GenBank Nucleotide sequence attributed, whereas 1,832 species had no attributable sequences. We also found that 1,050 (10%) of eBird/Clements species names have sequences listed under 2 or more GenBank names. For the GenBank SRA (sets of DNA sequence reads derived from massively parallel sequencing runs), 24% of avian species and subspecies were associated with a record. Of the 3,375 species and subspecies with SRA data, only 316 (9%) required reconciliation. Among reconciled names, the most common reason was due to shuffles (59% of reconciliations). While many reconciliation categories showed similar proportions to the GenBank data, reconciled names associated with SRA data included a greater proportion of hybrids (7%), domestics (2%), and unidentified (9%), but a lower proportion of splits (3%). Fewer than 4% (20 out of 530) GenBank genome assemblies required reconciliation.

When organized by the number of sequences affected by taxonomic reconciliation, different patterns emerged. In the GenBank Nucleotide database, 6,302,287 (91%) of sequences were a direct match, 626,079 (9%) we reconciled to eBird/Clements, and 2,575 (0.02%) we failed to reconcile. Of the nucleotide sequences, we reconciled to eBird/Clements, 106,940 (17%) we attributed to shuffles, 16,129 (2.6%) we attributed to lumps, and 381,652 (61%) we attributed to splits. We attributed 1,952 sequences to extinct species names not regulated by eBird/Clements, 5,909 (0.9%) sequences to hybrids, 102 (0.016%) sequences to new species names, and 110,748 (17%) sequences to domestic breeds.

The total number and proportion of sequences reconciled varied substantially among bird orders and among families within Passeriformes (Figures 2 and 3). Orders with the largest numbers of reconciled taxa corresponded to those with the greatest species diversity, including the Passeriformes (Songbirds), Piciformes (Woodpeckers and allies), and Caprimulgiformes (nightjars & allies, swifts, and hummingbirds). However, the proportion of names reconciled was reasonably uniform across orders, with outliers in some very small orders where few taxonomic changes have a dramatic effect on proportion (Rheiformes, 2 species; Casuariiformes, 4 species; Suliformes, 10 species), and a few orders that have retained relative taxonomic stability over the past 30 years (e.g., Trogoniformes, Galbuliformes, Ciconiiformes). We also broke down taxonomic reconciliation by family in the large order Passeriformes, where similar patterns emerged. However, in passerines, a few large families exhibited high proportions of reconciled names. Species-rich passerine families with high proportions of reconciled names included: Phylloscopidae (52%), Leiothrichidae (50%), Sylviidae *sensu stricto* (46%), Scotocercidae (42%), Pellorneidae (41%), Locustellidae (37%), and Timaliidae (27%).

Taxonomic Reconciliation in Relation to IUCN Conservation Status and Geography

There was little relationship between IUCN status and the proportion of taxa reconciled (Table 3; Data Repository D12). The categories Least Concern (LC) Near-threatened (NT) Vulnerable, (V), and Endangered all had similar proportions of taxa reconciled. Critically Endangered taxa were more likely to have had exact matches between GenBank and eBird Clements. Taxa not assessed by IUCN were far less likely to have an exact match between GenBank and eBird/Clements.

There was marked geographic variation in the percentage of taxa that needed reconciliation. The percentages of widespread taxa (19%) vs. those that were endemic to one of the realms we considered (18%) were virtually identical (Data Repository D13). Antarctica had no reconciled names, no doubt reflecting the very limited number of taxa found there. The 3 New World realms and the Australasian realm had the lowest percentages of reconciled names (15% for the North American realm to 17% for the South American realm; Data Repository D13; Figure 4). Oceanic realms had the highest percentages (up to 37% for the Atlantic Ocean; Data Repository D13; Figure 4).

Descriptive Statistics Linking GenBank Names to Global Avian Data Sources

To assess benefits that reconciling NCBI names with a standardized taxonomy has for the linking of sequence data with phenotypic data, we examined a reconciliation between the Xeno-canto avian sound database (which uses the IOC World Birds list) and the eBird/Clements names. We matched all Xeno-canto avian taxa to eBird/Clements names, except for 13 undescribed and 3 extinct taxa that are not included in the eBird/Clements v2019 (Clements et al. 2019) list. 10,166 (93%) of the Xeno-canto names directly matched to eBird/Clements names, and 9,506 of those

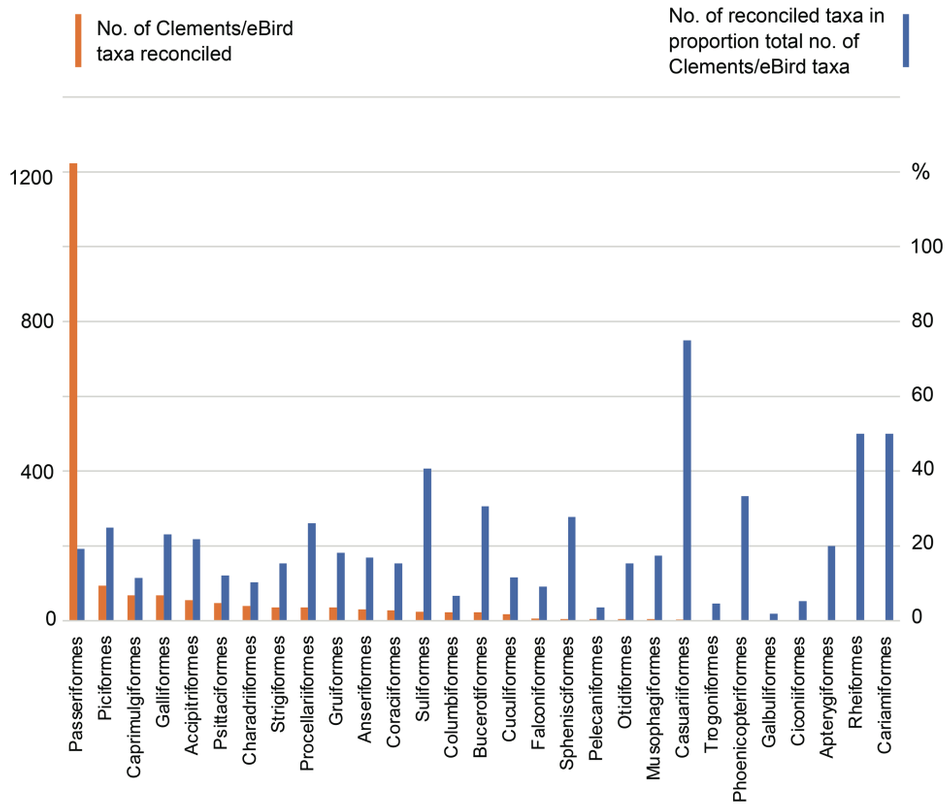

FIGURE 2. Number and proportion of taxonomic reconciliations applied to GenBank TaxIDs, by avian order.

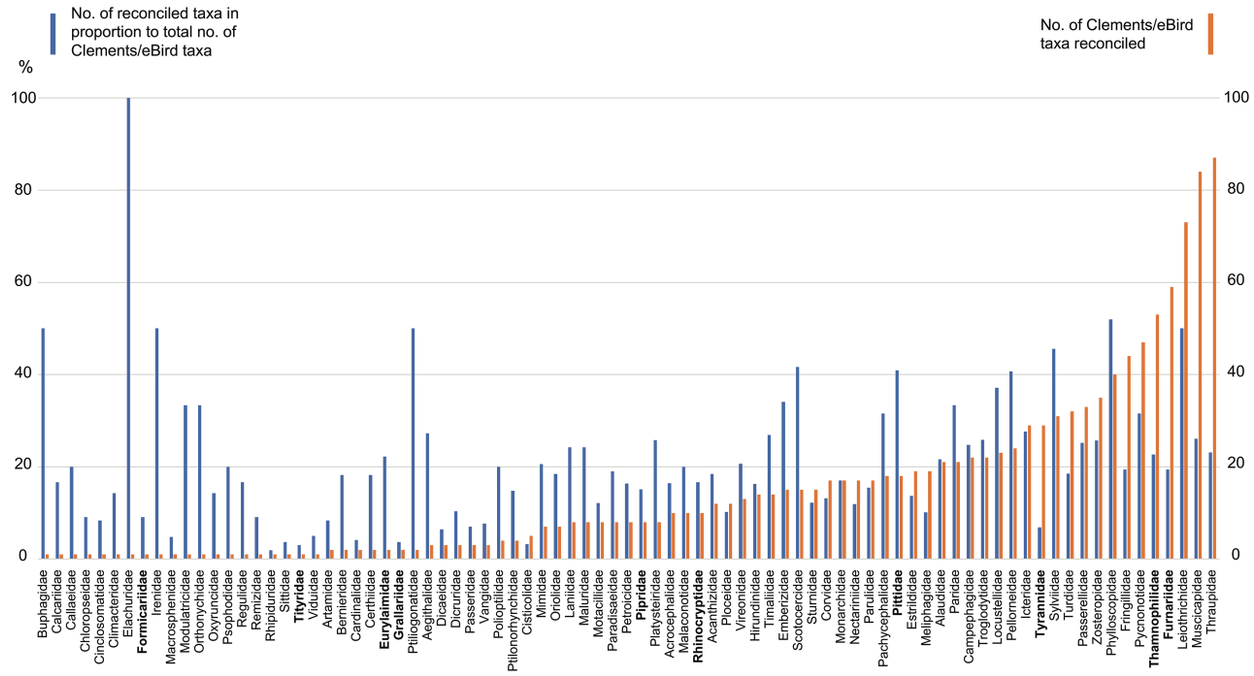

FIGURE 3. Number and proportion of taxonomic reconciliations applied to GenBank TaxIDs, by family within Passeriformes. Suboscine families are indicated by boldface text.

TABLE 3. Number and proportion of reconciliations by conservation status and their associated GenBank nucleotide data. The category Extinct includes both extinct taxa and the taxa that were extinct in the wild.

IUCN conservation status	Number of GenBank nucleotide sequences	Number of eBird/ Clements taxa	Number of eBird/Clements taxa reconciled	Proportion of eBird/ Clements taxa reconciled (%)
Least Concern	5,277,924	7,976	1,426	17.88
Vulnerable	782,643	750	113	15.07
Near-Threatened	397,032	932	150	16.09
Endangered	259,720	423	70	16.55
Critically Endangered	54,789	204	22	10.78
Extinct	482	160	4	2.50
Data Deficient	163	45	7	15.56
Not Assessed	150,901	286	80	27.97

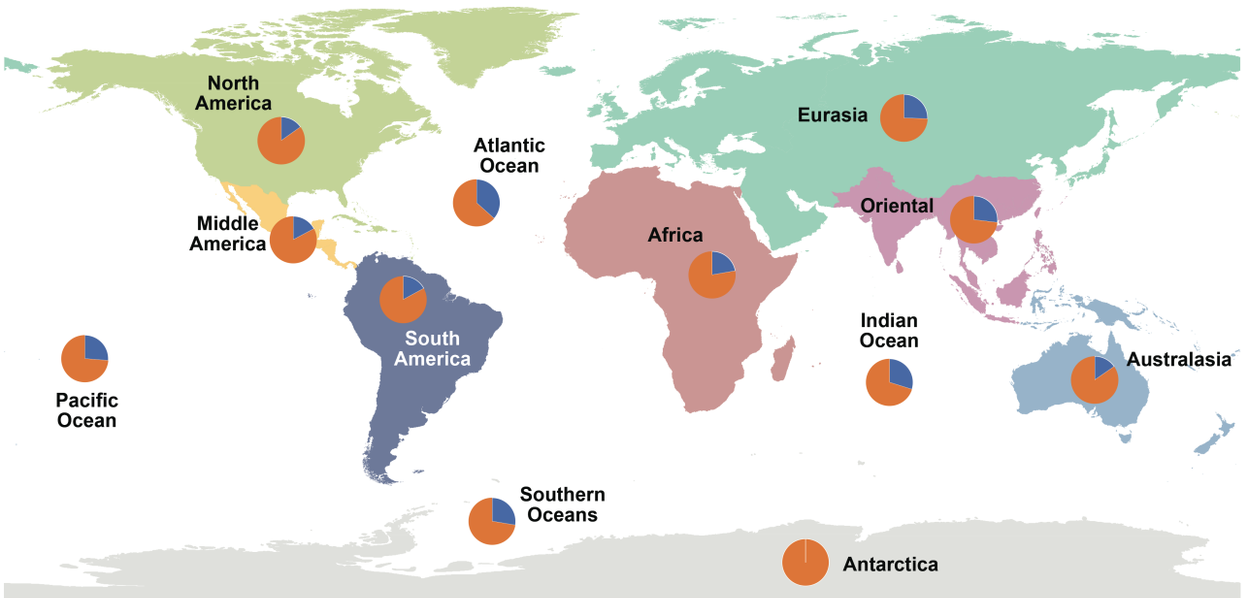

FIGURE 4. Geographic distribution of taxonomic reconciliations applied to GenBank TaxIDs. With the exception of the Antarctic realm, where there were no reconciliations between GenBank and eBird/ClementsClements/eBird, the proportion of reconciled names (blue) ranged from 15% (North America) to 37% (Atlantic Ocean). Widespread species may occur in multiple realms.

names have available sound recordings (Data Repository D14). The remaining 727 (7%) taxa were reconciled to eBird/Clements using Avibase taxonomic concepts. After reconciliation, we found 9,961 eBird/Clements species had sound recordings in Xeno-canto. In the Macaulay Library, there are 9,609 species with sound data, with an overlap of 9,399 species to Xeno-canto. By reconciling GenBank names with eBird/Clements taxonomy, we could easily link sequence data with the 2 largest avian sound databases that utilize standardized avian taxonomies (Table 4).

Open Source Access to Taxonomic Reconciliation

Our taxonomic reconciliation "RANT: reconciling avian NCBI taxonomy" is open source, and available at GitHub (https://github.com/ebraun68/RANT). Currently, the reconciliation is available for eBird/Clements version 2019 (Data Repository D11). Our intention is to update the reconciliation corresponding to eBird/Clements annual updates.

DISCUSSION

Successful Linkage of GenBank Names to Standardized Lists

Our reconciliation procedures have successfully linked GenBank taxonomic names (TaxIDs) with avian species and subspecies names regulated by eBird/Clements.

TABLE 4. Linking reconciled GenBank names with the two largest avian sound databases which utilize standardized avian taxonomies.

Database	eBird/ Clements species with sound data	Species with both sound and nucleotide data	Sound data only	Nucleotide data only	Neither sound nor nucleotide data
Xeno-canto	9,961 (92.9%)	8,693 (81.1%)	1,268	442	318
Macaulay Library	9,609 (89.6%)	8,409 (78.4%)	1,200	712	386
Combined	10,171 (94.9%)	8,837 (82.4%)	1,333	298	252

Nearly a tenth of all GenBank core nucleotide sequences had a name unrecognized in the eBird/Clements list, amounting to a total of more than 600,000 nucleotide sequences. Hence, it is now easier to link genetic data associated with GenBank TaxIDs to natural populations for comparative work, at least when comparative data have also been reconciled to the eBird/Clements taxonomic lists. If not, these GenBank TaxIDs can still be reconciled to other standardized lists (IOC, BirdLife International, Howard and Moore) through existing resources, namely Avibase and list comparisons freely available from the IOC World Bird List. If avian comparative data do not follow the names of one of these standardized global bird lists, then we advocate that database providers and curators reconcile their aggregated data to one of these standardized lists before its further use and publication.

Identifying Patterns and Biases in Naming Reconciliation

Reconciling GenBank TaxIDs to eBird/Clements names illustrates that naming problems are found throughout the avian tree of life yet they are concentrated in certain taxonomic groups. Unsurprisingly, these groups tend to have long histories of taxonomic instability. Reconciliation was especially frequent among members of the traditional "Old World warbler" (Sylviidae *sensu lato*) and "babbler" families (Timaliidae *sensu lato*). These groups have been split into a myriad of smaller families, each of which have undergone substantial revision (Cibois et al. 1999, 2002, Alström et al. 2011, 2018, Fregin et al. 2012, Moyle et al. 2012, Cai et al. 2019).

Outside the Old World warblers and babblers, several other passerine families had high proportions of reconciled names. Forty-one percent of Pittidae names required reconciliation. Perhaps this was because traditionally all pitta species were included in the genus *Pitta*. However, pitta diversity is now divided nearly equally among three genera: *Pitta*, *Hydronis*, and *Erythropitta* (Irestedt et al. 2006, Harvey et al. 2020). Additionally, the highly polytypic *Erythropitta erythrogaster* has been split into 12 species (Irestedt et al. 2013). Another group with a highly polytypic species is the Pachycephalidae. It contains the Golden Whistler (*Pachycephala pectoralis*), which was previously the world's most polytypic bird species (Andersen et al. 2014, Jønsson et al. 2014). Taxonomic revisions have since split the *P. pectoralis* complex into ~15 species. Reconciliations in Pittidae and Pachycephalidae illustrate how only a few major taxonomic revisions can create stark differences between names used on GenBank and those used in standardized avian bird lists.

One large family, Tyrannidae, had relatively few reconciliations. The eBird/Clements list currently considers 422 species of Tyrannidae, yet the proportion of reconciled names was low, only 7%. Small-scale molecular studies have revised parts of tyrannid nomenclature (Hosner and Moyle 2012, Rheindt et al. 2015). Yet until recently (Harvey et al. 2020, Ohlson et al. 2020), the Tyrannidae has lacked more comprehensive published molecular phylogenies and associated major taxonomic revisions. With the support of these recent publications, we expect the relative nomenclatural stability in Tyrannidae will prove short-lived, and a series of proposed changes will take effect in the coming years.

In addition to taxonomic biases, RANT identified large-scale geographic differences in GenBank name reconciliation. Widespread species—those found in more than one geographic realm—were only slightly more likely to have been subject to taxonomic reconciliation than those limited to a single geographic realm. North America and Australasia proportionally had the fewest reconciled names (Figure 4, Supplementary Material Table S1). Both of these realms are comparatively well studied, so a lack of taxonomic effort is not a viable explanation for their relative stability. One explanation for relative stability in North America and Australia could be the lack of highly problematic groups inhabiting those realms. Very few or no members of taxonomically problematic groups, such as Leiothrichidae, Phylloscopidae, or Sylviidae occur in North America or Australia. Although far more diverse than North America or Australia, Middle and South America had only slightly greater proportions of reconciled names (Figure 4), though several of the megadiverse Neotropical families, namely Thraupidae, Furnariidae, and Thamnophilidae; were among the families with the greatest total number of reconciliations. Proportionally, Old World realms had the most name changes of the terrestrial realms (Figure 4). We suspect the high proportion of reconciled names is related

to the concentration of taxonomically problematic groups in these regions, especially the Sylvioidea.

The oceanic realms all featured relatively large proportions of reconciled names. Seabirds in particular have undergone extensive taxonomic revisions, driven mostly by molecular genetic work which has revealed great levels of cryptic genetic diversity among ocean basins, breeding islands, and archipelagos (Pyle et al. 2011, Taylor et al. 2019). Among orders, the pelagic groups Procellariiformes, Suliformes, and Sphenisciformes all had large proportions of reconciled names.

Previous authors have raised alarms regarding how taxonomic instability can hamper conservation efforts (Garnett and Christidis 2017). However, we found that IUCN red-listed species were not more likely to have had name reconciliation compared to non-threatened taxa. Among IUCN conservation status categories, "Least Concern" had the greatest proportion of reconciled names whereas "Critically Endangered" had the lowest. Most critically endangered birds are highly range-restricted, and hence are not likely to have been subject to taxonomic splits into multiple species. Taxa not assessed by IUCN had a large proportion of reconciled names, probably driven by the fact that eBird/Clements names not assessed by IUCN are the result of nomenclatural differences between these sources.

The Problem of Name Application for GenBank Sequences

One glaring problem linking taxonomic names to DNA sequences remains, and that is far more insidious than the main problem addressed here. A GenBank TaxID associated with a eBird/Clements name does not necessarily mean that the DNA sequences ascribed to that name will apply correctly (Schoch et al. 2020). Before phylogenetic or population genetic analyses can commence, the correct application of eBird/Clements names to individual sequences must be verified, a process that is time-consuming and challenging to automate. Below is an example of how the verification process may proceed, drawn from an example of nucleotide data published on GenBank.

The *Robsonius* ground warblers (Cordillera Ground-Warbler [*Robsonius rabori*], Sierra Madre Ground-Wabler [*R. thompsoni*], Bicol Ground-Warbler [*R. sorsogonensis*]) have a complex taxonomic history which highlights many of the nomenclatural challenges inherent when working with GenBank data. Originally described in the wren-babbler genus *Napothera* (Rand and Rabor 1967), for most of their history they have been considered a single species. In 2006, they were split into 2 recognized species based on new morphological evidence and moved to the new genus *Robsonius* (Collar 2006). In 2013, a third species was described following the collection of the first adult specimen of true *R. rabori* (Hosner et al. 2013). All 4 standardized world lists currently recognize all 3 species: *R. rabori*, *R. thompsoni*, and *R. sorsogonensis*. However, GenBank nucleotide data are ascribed to only 2 TaxIDs: *R. rabori* (TaxID: 1149667, n = 76) and *R. thompsoni* (TaxID: 2162877, n = 3). Most of these data were uploaded under the name *Napothera rabori* prior to use of the *Robsonius* or the epithets *sorsogonensis/thompsoni* in GenBank taxonomy, but these sequences actually pertain to all 3 *Robsonius* species. After tracking down voucher numbers and metadata from publications and voucher specimens, the true taxonomic breakdown of nucleotide sequences is: *R. rabori*, n = 7; *R. sorsogonensis*, n = 24; and *R. thompsoni*, n = 48. Without confirming the application of names, several errors would hamper the use and interpretation of these data. A user might incorrectly conclude that no nucleotide data exist for *R. sorsogonensis*, because its sequences are labeled as *R. rabori*. A user might incorrectly conclude that *R. thompsoni* and *R. rabori* are not genetically distinct, because many *R. thompsoni* sequences are labeled as *R. rabori*. A user might incorrectly conclude that *R. rabori* has exceptional genetic diversity despite its tiny distribution because divergent *R. sorsogonensis* and *R. thompsoni* sequences are each labeled as *R. rabori*.

Resolving name application will be a far more difficult problem to solve than name reconciliation. Name reconciliation requires a set of non-standardized names, a standardized list, and tools or literature to match the non-standardized names to their standardized counterparts. Resolving name application, as in the *Robsonius* example above, requires individual sequence metadata, which is often not recorded in GenBank. Most name application issues arise from splits, when an inclusive former name is applied erroneously to one or more populations with which they were formerly considered conspecific. The most rigorous method to solve these taxonomic problems is to consult voucher specimens to confirm sequence identity. However, many GenBank sequences lack proper voucher specimen information (Peterson et al. 2007, Buckner et al. 2021), as we also noted. After filtering the "Aves" sequences in the nucleotide database to include only genomic DNA/RNA nucleotide sequences (excluding mRNA or rRNA sequences) from the INSDC (GenBank, not RefSeq) source database, we estimated that only 17% (484,232) of the 2,902,805 sequences included voucher information anywhere in the full GenBank record. While some other samples may have information included that could be used to trace the source of the sample, the majority of available sequence data lacks such information. Although there is a GenBank voucher field, it is not required, and is easy for sequence authors to omit. In some GenBank records voucher information is found in the sequence definition line instead. Checking vouchers one-by-one is not feasible for large-scale metadata correction of what, at present count, is more than 6 million avian nucleotide sequences, although it can still be used to resolve at least some problems.

Aside from vouchers, locality metadata is useful when resolving name application problems. Latitude and longitude can be included in GenBank metadata, but often are not. In some cases, they can be found in published papers or their Online Supplements, or in publicly shared museum databases if the samples were properly vouchered and digitized. However, this laborious task is not suited to large-scale applications without the development of automated tools such as georeferencing algorithms (Miraldo et al. 2016) to improve sequence attribution. When splits apply to allopatric populations, the latitude/longitude of the sample origin solves name application. However, when these splits do not apply cleanly to allopatric populations, or when migratory populations of split taxa overlap for part of the year, further information will be needed to resolve sequence identity with confidence.

A Call for Expert Curation of Avian GenBank Sequence Metadata

RANT is a first step towards active and decentralized management of metadata associated with avian sequence data. These standardized names provide a new benchmark for managing large-scale sequence meta-analyses, but many data problems remain—particularly the challenge of verifying names application to individual DNA sequences. Although GenBank provides a vast and important resource, large biodiversity datasets need constant management and expert curation to maximize their usefulness (Schoch et al. 2020, Sangster and Luksenburg 2021). One solution is to maintain a parallel database to update and store metadata related to GenBank sequences, but free of its restrictive updating policies (Riginos et al. 2020). With such a system, a team of expert curators could gather, review, proofread, and provide supplemental metadata associated with GenBank sequences (Marques et al. 2013), linked to the actual sequence data housed at GenBank through the accession number. In addition to validating metadata, curators can permanently flag or provide feedback on potentially problematic sequences (De Silva et al. 2019, Sangster and Luksenburg 2021). These strategies are effective for curating far larger sets of biodiversity data collected largely by non-professional scientists (Sullivan et al. 2009, Robertson et al. 2014).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Ornithology* online.

ACKNOWLEDGMENTS

We will thank anonymous reviewers who provided comments that improved the manuscript.

Funding statement: P.A.H. acknowledges the support of Villum Fonden grant # 25925. R.T.K. and E.L.B. were supported by the US National Science Foundation, grant number DEB-1655683.

Ethics statement: No declarations

Author contributions: P.A.H., R.T.K., E.L.B., and J.G.B. formulated questions; P.A.H., M.Z., E.L.B., and J.G.B. analyzed data; P.A.H., M.Z., and R.T.K. wrote the manuscript; M.Z. and E.L.B. drafted figures, all authors read and edited the manuscript.

Data availability: Analyses reported in this article can be reproduced using the data provided by Hosner et al. (2022).

LITERATURE CITED

Alström, P., A. Cibois, M. Irestedt, D. Zuccon, M. Gelang, J. Fjeldså, M. J. Andersen, R. G. Moyle, E. Pasquet, and U. Olsson (2018). Comprehensive molecular phylogeny of the grassbirds and allies (Locustellidae) reveals extensive non-monophyly of traditional genera, and a proposal for a new classification. Molecular Phylogenetics and Evolution 127:367–375.

Alström, P., S. Fregin, J. A. Norman, P. G. P. Ericson, L. Christidis, and U. Olsson (2011). Multilocus analysis of a taxonomically densely sampled dataset reveal extensive non-monophyly in the avian family Locustellidae. Molecular Phylogenetics and Evolution 58:513–526.

Andersen, M. J., Á. S. Nyári, I. Mason, L. Joseph, J. P. Dumbacher, C. E. Filardi, and R. G. Moyle (2014). Molecular systematics of the world's most polytypic bird: The *Pachycephala pectoralis/melanura* (Aves: Pachycephalidae) species complex. Zoological Journal of the Linnean Society 170:566–588.

Barrowclough, G. F., J. Cracraft, J. Klicka, and R. M. Zink (2016). How many kinds of birds are there and why does it matter? PLoS One 11:e0166307.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2012). GenBank. Nucleic Acids Research 41:D36–D42.

Beresford, P., F. K. Barker, P. G. Ryan, and T. M. Crowe (2005). African endemics span the tree of songbirds (Passeri): Molecular systematics of several evolutionary enigmas. . Proceedings of the Royal Society B: Biological Sciences 272:849–858.

Boyer, D. M., G. F. Gunnell, S. Kaufman, and T. M. McGeary (2016). MorphoSource: Archiving and sharing 3-D digital specimen data. Paleontological Society Papers 22:157–181.

Boyle, B., N. Hopkins, Z. Lu, J. A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, M. L. Narro, W. H. Piel, S. J. Mckay, et al. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. BMC Bioinformatics 14:16.

Braun, E. L., J. Cracraft, and P. Houde (2019). Resolving the avian Tree of Life from top to bottom: The promise and potential boundaries of the phylogenomic era. In Avian Genomics in Ecology and Evolution (R. H. S. Kraus, Editor). Springer International Publishing, Cham, Switzerland. pp. 151–210.

Buckner, J. C., R. C. Sanders, B. C. Faircloth, and P. Chakrabarty (2021). The critical importance of vouchers in genomics. eLife 10:e68264.

Burleigh, J. G., R. T. Kimball, and E. L. Braun (2015). Building the avian tree of life using a large-scale, sparse supermatrix. Molecular Phylogenetics and Evolution 84:53–63.

Cai, T., A. Cibois, P. Alström, R. G. Moyle, J. D. Kennedy, S. Shao, R. Zhang, M. Irestedt, P. G. P. Ericson, M. Gelang, et al. (2019). Near-complete phylogeny and taxonomic revision of the world's babblers (Aves: Passeriformes). Molecular Phylogenetics and Evolution 130:346–356.

Cibois, A., M. V. Kalyakin, H. Lian-Xian, and E. Pasquet (2002). Molecular phylogenetics of babblers (Timaliidae): Revaluation of the genera *Yuhina* and *Stachyris*. Journal of Avian Biology 33:380–390.

Cibois, A., E. Pasquet, and T. S. Schulenberg (1999). Molecular systematics of the Malagasy babblers (Passeriformes: Timaliidae) and warblers (Passeriformes: Sylviidae), based on cytochrome b and 16S rRNA sequences. Molecular Phylogenetics and Evolution 13:581–595.

Cicero, C., J. Cook, M. Campbell, K. Hildebrandt, T. Mayfield, and J. Wieczorek (2017). The Arctos community model for sustaining and enriching access to biodiversity data. Biodiversity Information Science and Standards 1:e20466.

Clements, J. F., T. S. Schulenberg, M. J. Iliff, S. M. Billerman, B. L. Fredericks, B. Sullivan, and C. L. Wood (2019). The eBird/Clements Checklist of Birds of the World: v2019. https://www.birds.cornell.edu/clementschecklist/download/

Clements, J. F., T. S. Schulenberg, M. J. Iliff, S. M. Billerman, T. A. Fredericks, J. A. Gerbracht, D. Lepage, B. L. Sullivan, and C. L. Wood (2021). The eBird/Clements checklist of Birds of the World: v2021. https://www.birds.cornell.edu/clementschecklist/download/

Cochrane, G., I. Karsch-Mizrachi, Y. Nakamura, and The International Nucleotide Sequence Database Collaboration (2011). The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research 39:D15–D18.

Collar, N. J. (2006). A partial revision of the Asian babblers (Timaliidae). Forktail 22:85–112.

Constable, H., R. Guralnick, J. Wieczorek, C. Spencer, A. T. Peterson, and The VertNet Steering Committee (2010). VertNet: A new model for biodiversity data sharing. PLoS Biology 8:e1000309.

De Silva, T. N., J. M. Bates, and A. T. Peterson (2019). Getting the Ploceidae tree right. Molecular Phylogenetics and Evolution 131:228.

Dickinson, E. C., and L. Christidis (2014). The Howard and Moore Complete Checklist of the Birds of the World. Fourth Edition, Volume 1: Non-passerines. Aves Press, Eastbourne, UK.

Dickinson, E. C., and J. V. Remsen (2013). The Howard and Moore Complete Checklist of the Birds of the World. Fourth Edition, Volume 2: Passerines. Aves Press, Eastbourne, UK.

Droege, G., K. Barker, J. J. Astrin, P. Bartels, C. Butler, D. Cantrill, J. Coddington, F. Forest, B. Gemeinholzer, D. Hobern, et al. (2014). The Global Genome Biodiversity Network (GGBN) Data Portal. Nucleic Acids Research 42:D607–D612.

Felsenstein, J. (1985). Phylogenies and the comparative method. The American Naturalist 125:1–15.

Fregin, S., M. Haase, U. Olsson, and P. Alström (2012). New insights into family relationships within the avian superfamily Sylvioidea (Passeriformes) based on seven molecular markers. BMC Evolutionary Biology 12:157.

Garnett, S. T., and L. Christidis (2017). Taxonomy anarchy hampers conservation. Nature 546:25–27.

Gill, F. B. (2014). Species taxonomy of birds: Which null hypothesis? The Auk 131:150–161.

Gill, F. B., D. Donsker, and P. C. Rasmussen (2022). IOC World Bird List (v.12.1). https://www.worldbirdnames.org/

Hackett, S. J., R. T. Kimball, S. Reddy, R. C. K. Bowie, E. L. Braun, M. J. Braun, J. L. Chojnowski, W. A. Cox, K. L. Han, J. Harshman, et al. (2008). A phylogenomic study of birds reveals their evolutionary history. Science 320:1763–1768.

Harvey, M. G., G. A. Bravo, S. Claramunt, A. M. Cuervo, G. E. Derryberry, J. Battilana, G. F. Seeholzer, J. S. McKay, B. C. Fairloth, S. V. Edwards, et al. (2020). The evolution of a tropical biodiversity hotspot. Science 370:1343–1348.

HBW and BirdLife International (2022). Handbook of the Birds of the World and BirdLife International digital checklist of the birds of the world. Version 6b. http://datazone.birdlife.org/userfiles/file/Species/Taxonomy/HBW-BirdLife_Checklist_v6b_Jul22.zip

Hosner, P. A., N. C. Boggess, P. Alviola, L. A. Sánchez-González, C. H. Oliveros, R. Urriza, and R. G. Moyle (2013). Phylogeography of the *Robsonius* Ground-Warblers (Passeriformes: Locustellidae) reveals an undescribed species from northeastern Luzon, Philippines. The Condor 115:630–639.

Hosner, P. A., L. C. Campillo, M. J. Andersen, L. A. Sánchez-González, C. H. Oliveros, R. C. Urriza, and R. G. Moyle (2018). An integrative species delimitation approach reveals fine-scale endemism and substantial unrecognized avian diversity in the Philippine Archipelago. Conservation Genetics 19:1153–1168.

Hosner, P. A., and R. G. Moyle (2012). A molecular phylogeny of black-tyrants (Tyrannidae: *Knipolegus*) reveals strong geographic patterns and homoplasy in plumage and display behavior. The Auk 129:156–167.

Hosner, P. A., M. Zhao, R. T. Kimball, E. L. Braun, and J. G. Burleigh (2022). Data from: Updating splits, lumps, and shuffles: Reconciling GenBank names with standardized avian taxonomies. Ornithology 139:ukac045. doi:10.5061/dryad.gtht76hqf.

Irestedt, M., P. H. Fabre, H. Batalha-Filho, K. A. Jønsson, C. S. Roselaar, G. Sangster, and P. G. P. Ericson (2013). The spatio-temporal colonization and diversification across the Indo-Pacific by a "great speciator" (Aves, *Erythropitta erythrogaster*). Proceedings of the Royal Society B: Biological Sciences 280:20130309.

Irestedt, M., J. I. Ohlson, D. Zuccon, M. Källersjö, and P. G. P. Ericson (2006). Nuclear DNA from old collections of avian study skins reveals the evolutionary history of the Old World suboscines (Aves, Passeriformes). Zoologica Scripta 35:567–580.

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers (2012). The global diversity of birds in space and time. Nature 491:444–448.

Jønsson, K. A., M. Irestedt, L. Christidis, S. M. Clegg, B. G. Holt, and J. Fjeldså (2014). Evidence of taxon cycles in an Indo-Pacific passerine bird radiation (Aves: *Pachycephala*). Proceedings of the Royal Society B: Biological Sciences 281:20131727.

Kranstauber, B., A. Cameron, R. Weinzerl, T. Fountain, S. Tilak, M. Wikelski, and R. Kays (2011). The Movebank data model for animal tracking. Environmental Modelling & Software 26:834–835.

Lepage, D., G. Vaidya, and G. Guralnick (2014). Avibase–a database system for managing and organizing taxonomic concepts. ZooKeys 117:117–135.

Leray, M., N. Knowlton, S. L. Ho, B. N. Nguyen, and R. J. Machida (2020). Reply to Locatelli *et al*.: Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group. Proceedings of the National Academy of Sciences USA 117:32213–32214.

Lovette, I. J., J. L. Pérez-Emán, J. P. Sullivan, R. C. Banks, I. Fiorentino, S. Córdoba-Córdoba, M. Echeverry-Galvis, F. K. Barker, K. J. Burns, J. Klicka, S. M. Lanyon, and E. Bermingham (2010). A comprehensive multilocus phylogeny for the wood-warblers and a revised classification of the Parulidae (Aves). Molecular Phylogenetics and Evolution 57:753–770.

Marques, A. C., M. M. Maronna, and A. G. Collins (2013). Putting GenBank data on the map. Science 341:1341–1341.

McClure, C. J. W., D. Lepage, L. Dunn, D. L. Anderson, S. E. Schulwitz, L. Camacho, B. W. Robinson, L. Christidis, T. S. Schulenberg, M. J. Iliff, P. C. Rasmussen, and J. Johnson (2020). Towards reconciliation of the four world bird lists: Hotspots of disagreement in taxonomy of raptors. Proceedings of the Royal Society B: Biological Sciences 287::20200683.

Miraldo, A., S. Li, M. K. Borregaard, A. Flórez-Rodríguez, S. Gopalakrishnan, M. Rizvanovic, Z. Wang, C. Rahbek, K. A. Marske, and D. Nogués-Bravo (2016). An Anthropocene map of genetic diversity. Science 353:1532–1535.

Moyle, R. G., M. J. Andersen, C. H. Oliveros, F. D. Steinheimer, and S. Reddy (2012). Phylogeny and biogeography of the core babblers (Aves: Timaliidae). Systematic Biology 61:631–651.

Neate-Clegg, M. H. C., J. D. Blount, and Ç. H. Şekercioğlu (2021). Ecological and biogeographical predictors of taxonomic discord across the world's birds. Global Ecology and Biogeography 30:1258–1270.

Nelson, G., and D. L. Paul (2019). DiSSCo, iDigBio and the Future of Global Collaboration. Biodiversity Information Science and Standards 3:e37896.

Ohlson, J. I., M. Irestedt, H. B. Filho, P. G. P. Ericson, and J. Fjeldså (2020). A revised classification of the fluvicoline tyrant flycatchers (Passeriformes, Tyrannidae, Fluvicolinae). Zootaxa 4747:167–176.

Oliveros, C. H., D. J. Field, D. T. Ksepka, F. K. Barker, A. Aleixo, M. J. Andersen, P. Alström, B. W. Benz, E. L. Braun, M. J. Braun, et al. (2019). Earth history and the passerine superradiation. Proceedings of the National Academy of Sciences USA 116:7916–7925.

Peters, J. L., E. Mayr, J. C. Greenway, R. A. Paynter, and M. E. Traylor, Editors (1931–1987). Check-list of Birds of the World, 16 volumes. Museum of Comparative Zoology, Cambridge, MA, USA.

Peterson, A. T., R. G. Moyle, Á. S. Nyári, M. B. Robbins, R. T. Brumfield, and J. V. Remsen (2007). The need for proper vouchering in phylogenetic studies of birds. Molecular Phylogenetics and Evolution 45:1042–1044.

Pigot, A. L., W. Jetz, C. Sheard, and J. A. Tobias (2018). The macroecological dynamics of species coexistence in birds. Nature Ecology and Evolution 2:1112–1119.

Pyle, P., A. J. Welch, and R. C. Fleischer (2011). A new species of shearwater (*Puffinus*) recorded from Midway Atoll, Northwestern Hawaiian Islands. The Condor 113:518–527.

Rand, A. L., and D. S. Rabor (1967). New birds from Luzon, Philippine Islands. Fieldiana Zoology 51:85–99.

Rheindt, F. E., N. Krabbe, A. K. S. Wee, and L. Christidis (2015). Cryptic speciation in the Lesser Elaenia *Elaenia chiriquensis* (Aves: Passeriformes: Tyrannidae). Zootaxa 4032:251.

Riginos, C., E. D. Crandall, L. Liggins, M. R. Gaither, R. B. Ewing, C. Meyer, K. R. Andrews, P. T. Euclide, B. M. Titus, N. O. Therkildsen, et al. (2020). Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. Molecular Ecology Resources 20:1458–1469.

Robertson, T., M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. PLoS One 9:e102623.

Sangster, G. (2009). Increasing numbers of bird species result from taxonomic progress, not taxonomic inflation. Proceedings of the Royal Society B: Biological Sciences 276:3185–3191.

Sangster, G., and J. A. Luksenburg (2021). Sharp increase of problematic mitogenomes of birds: Causes, consequences, and remedies. Genome Biology and Evolution 13:evab210.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database 2020:baaa062.

Sibley, C. G., and B. L. Monroe, Jr (1993). Distribution and Taxonomy of Birds of the World. Yale University Press, New Haven, CT, USA.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Tobias, J. A., C. Sheard, A. L. Pigot, A. J. Devenish, J. Yang, F. Sayol, M. H. Neate-Clegg, N. Alioravainen, T. L. Weeks, R. A. Barber, et al. (2022). AVONET: Morphological, ecological and geographical data for all birds. Ecology Letters 25:581–597.

Taylor, R. S., M. Bolton, A. Beard, T. Birt, P. Deane-Coe, A. F. Raine, J. González-Solís, S. C. Lougheed, and V. L. Friesen (2019). Cryptic species and independent origins of allochronic populations within a seabird species complex (*Hydrobates* spp.). Molecular Phylogenetics and Evolution 139:106552.

Unger, S., M. Rollins, A. Tietz, and H. Dumais (2020). iNaturalist as an engaging tool for identifying organisms in outdoor activities. Journal of Biological Education 55:537–547.

Wilman, H., J. Belmaker, J. Simpson, C. de la Rosa, M. M. Rivadeneira, and W. Jetz (2014). EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. Ecology 95:2027–2027.