Taylor & Francis
Taylor & Francis Group

Mitochondrial DNA Part A

DNA Mapping, Sequencing, and Analysis

ISSN: 2470-1394 (Print) 2470-1408 (Online) Journal homepage: http://www.tandfonline.com/loi/imdn21

Mitogenome metadata: current trends and proposed standards

Jeff H. T. Strohm, Rodger A. Gwiazdowski & Robert Hanner

To cite this article: Jeff H. T. Strohm, Rodger A. Gwiazdowski & Robert Hanner (2016) Mitogenome metadata: current trends and proposed standards, Mitochondrial DNA Part A, 27:5, 3263-3269, DOI: 10.3109/19401736.2015.1015003

To link to this article: http://dx.doi.org/10.3109/19401736.2015.1015003

Mitochondrial DNA

http://informahealthcare.com/mdn
ISSN: 2470-1394 (print), 2470-1408 (electronic)

Mitochondrial DNA Part A, 2016; 27(5): 3263–3269
 DOI: 10.3109/19401736.2015.1015003

informa
healthcare

FULL LENGTH RESEARCH PAPER

Mitogenome metadata: current trends and proposed standards

Jeff H. T. Strohm[1], Rodger A. Gwiazdowski[2], and Robert Hanner[1]

[1]*Department of Integrative Biology, Centre for Biodiversity Genomics, University of Guelph, Ontario, Canada and* [2]*Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada*

Abstract

Mitogenome metadata are descriptive terms about the sequence, and its specimen description that allow both to be digitally discoverable and interoperable. Here, we review a sampling of mitogenome metadata published in the journal *Mitochondrial DNA* between 2005 and 2014. Specifically, we have focused on a subset of metadata fields that are available for GenBank records, and specified by the Genomics Standards Consortium (GSC) and other biodiversity metadata standards; and we assessed their presence across three main categories: collection, biological and taxonomic information. To do this we reviewed 146 mitogenome manuscripts, and their associated GenBank records, and scored them for 13 metadata fields. We also explored the potential for mitogenome misidentification using their sequence diversity, and taxonomic metadata on the Barcode of Life Datasystems (BOLD). For this, we focused on all Lepidoptera and Perciformes mitogenomes included in the review, along with additional mitogenome sequence data mined from Genbank. Overall, we found that none of 146 mitogenome projects provided all the metadata we looked for; and only 17 projects provided at least one category of metadata across the three main categories. Comparisons using mtDNA sequences from BOLD, suggest that some mitogenomes may be misidentified. Lastly, we appreciate the research potential of mitogenomes announced through this journal; and we conclude with a suggestion of 13 metadata fields, available on GenBank, that if provided in a mitogenomes's GenBank record, would increase their research value.

Keywords

Reference sequence, reproducibility, reproducible research, specimen metadata, voucher

History

Received 28 August 2013
Revised 28 January 2015
Accepted 31 January 2015
Published online 19 February 2015

Introduction

The number of publically available mitochondrial genome sequences (mitogenomes) is increasing dramatically, but their usefulness depends on the quality of their specimen metadata (Field et al., 2008; Foley et al., 2009). Mitogenome metadata that describe basic information such as collection, and voucher information lets these sequences be repurposed. Mitogenomes can serve as baseline comparisons, identified specimens, occurrence records, etc. but only when sequences and their basic metadata are associated (Goodman et al., 2014; Vasilevsky et al., 2013).

The need for specimen metadata associated with DNA sequences has resulted in several, complimentary, metadata standards including: the Minimum Information about a Marker Gene Sequence (MIMARKS) (Yilmaz et al., 2011), Minimum Information about a Genome Sequence (MIGS) (Field et al., 2008), the DNA barcode Data Standard (BARCODE) (Hanner, 2009), and the Darwin Core (here, we refer to the Simple Darwin Core format: SIMPLEDWC) (Wieczorek et al., 2012). MIMARKS, and MIGS are oriented towards prokaryotes, but all recommend a core level of collection, biological and taxonomic metadata (Field et al., 2008; Wieczorek et al., 2012; Yilmaz et al., 2011). These metadata include basic information such as: the collection date with precise location, the collector, the identifier, and a taxon name. We list the fields for these data standards in Table 1. Although these data standards can apply to mitogenomes (as sequences), no one standard encompasses a suite of metadata that exploits the unique biological and information properties of mitogenomes. For example, biological metadata for mitogenomes could include the number of specimens used for a finished genome, which is valuable because multiple specimens or even tissue types may harbor unique haplotypes resulting in a chimeric sequence (Carr & Marshall, 2008; Krjutškov et al., 2014). The sex of specimens or their life stage is also important to document because this variation can result in mis-identification at the species level, or higher. Also, reporting the location of a mitogenome's voucher establishes a re-accessible link between digital sequence data, and the source specimen (Agerer et al., 2000; Ruedas et al., 2000; Vink et al., 2012; Wheeler, 2003).

When available, mitogenome metadata allows for the integration of mitochondrial sequence data from multiple sources (Hobern et al., 2013), which enhances data discovery and can serve for quality control. As a striking example, a recent publication by Botero-Castro & colleagues (2014) suggests that a recently published mitogenome for the Leschenault's rousette bat (*Rousettus leschenaultii*) actually belongs to the Egyptian fruit bat (*R. aegyptiacus*). This conclusion was supported, in part, through the use of these mitogenome's metadata with the Barcode of Life Data Systems (BOLD; Ratnasingham & Hebert, 2007; www.boldsystems.org), highlighting the value of specimen metadata – within a comparative framework. Relatedly,

Correspondence: Jeff Strohm, Centre for Biodiversity Genomics, Department of Integrative Biology, University of Guelph, Ontario, Canada. Tel: +226 979 5489. E-mail: jstrohm@uoguelph.ca

Table 1. Metadata fields for collection, biological and taxonomic information categories that are generally congruent across existing DNA sequence metadata standards, as well as fields surveyed for this study, the percent of surveyed studies that report any information in those fields, and existing NCBI/GenBank metadata fields that correspond to the main categories.

Data Type	Metadata field	SIMPLEDWC	MIMARKS	MIGs	BARCODE Std	For mtDNA Survey	% in mtDNA survey	NCBI Fields
Collection	Specimen voucher number	■	■	■	■	■	16	Specimen-voucher
	Collector				■	■	15	Collected-by
	Collection date	■	■		■	■	12	Collection-date
	Latitude longitude	■	■		■		8	Lat-Lon
	Country	■	■		■		13	Country
	Host association	■	■					Host
	pcr primers used				■			PCR primers
	Type of study	■	■		■			
	Institution storing	■			■			
	Dataset/project name	■	■					
	Sequencing method		■					
	Sequence assembly			■				
	Sequence finishing strategy			■				
	Plant association		■					
	Environment (biome)		■					
	Environment (feature)		■					
	Environment (material)		■					
	Environmental package		■					
	Altitude (air)	■	■					
	Elevation*		■					
	Depth(water)	■	■					
	Depth*		■					
	Moderate location					■	40	
	Farm/zoo					■	10	
	Misc		■					
Biological	Sex	■				■	5	Sex
	Tissue					■	35	Tissue-type
	Life stage	■				■	4	Dev-stage
	# specimens					■	34	
	Estimated size			■				
	Ploidy			■				
	Number of replicons			■				
	Propagation			■				
	Growth conditions	■		■				
Taxonomic	Identifier	■			■	■	5	Identified-by
	Bio-material source			■				Bio-material
	Species name	■			■	■	100	Organism
	Identification method	■				■	6	

*(soil, sediment, microbial mat/biofilm) Shaded values represent metadata that are required by the data standard in question.

This table includes all fields specified by MIGs, MIMARKs and the DNA barcode Data Standard, but retains only SimpleDarwinCore fields that were congruent with the other standards. Additionally we introduced several fields as part of the criteria for this study: moderate location indicates a reference to a city or county; farm/zoo indicates farm, zoo, university or research center where population origin information was not clear; # of specimens indicates the number of specimens that contributed DNA to the final mitogenome sequence; identification method indicates the number of specimens that contributed DNA to the final mitogenome sequence. The NCBI fields are part of the submission format for GenBank sequence records; these fields are listed, and discussed further in Box 1.

mitogenomes with specimen metadata could be considered de-facto ''Ultra-barcodes'' (Kane et al., 2012) as they include the standard DNA barcode sequence region of COI, and can serve as an informatics link with other mitochondrial loci (e.g. Cytb, COII, ND3).

In the course of our own research on species diversity and identification, we encountered many mitogenomes that lacked specimen metadata useful for us, and we became curious to observe general patterns of mitogenome metadata content. To do this we took a fine-grained look at the metadata content of mitogenomes published in the Journal *Mitochondrial DNA,* and their associated Genbank records, compared against a core-level of specimen metadata from all the standards mentioned above. We then use these results to address three related questions: (1) What proportion of mitogenomes include metadata from DNA sequence metadata-standards regarding collection, taxonomic, and biological information (13 fields, defined below)? (2) How many mitogenomes may represent an incorrect or ambiguous assignment to a species? (3) Are mitogenomes with basic taxonomic metadata less likely to represent an incorrect species than those without these metadata?

Methods

We chose the Journal *Mitochondrial DNA* as it has historically published a large number of mitogenomes, and in 2009 introduced a fast-track manuscript category ''*Mitogenome Announcement*'' exclusively for the publication of mitogenomes (DeSalle & Kolokotronis, 2009) (from 2005–2008 the Journal was named *DNA sequence*). Our study set was a sub-sample of 146 mitogenome publications from *Mitochondrial DNA* between 2005–2014. We collected this set by reviewing all publications spanning 2005 to 2008, and from 2013 volume 24, numbers 4 through 6 and from 2014, volume 25, number 3, and extended our coverage by randomly selecting all mitogenomes from one volume-number per year, from each year not previously sampled (2009 through 2012). The sampling is displayed in Figure 1(A) and (B). R code used to generate Figures 1 and 2,

Box 1. Recommended GenBank metadata fields to accompany mitogenome submissions.

Metadata allow DNA sequences to be repurposed beyond their original study if they are associated with the sequences. The GenBank submission format includes many metadata fields that are congruent across DNA sequence metadata standards, including SimpleDarwinCore, MIGs, MINIMARKs and the BARCODE Data Standard. Below, we suggest 13 GenBank metadata fields that are generally congruent across these data standards. Inclusion of metadata from these collection, taxonomic and biological fields would increase the research value of mitogenomes. The field names below are verbatim GenBank format.

- Specimen-voucher: Identifier of the physical specimen from which the sequence was obtained.
- Collected-by: Name of person(s) or institute who collected sample.
- Collection-date: Date sample was collected.
- Lat-Lon: Latitude and longitude of location where sample was collected.
- Country: The country of origin of DNA samples used for epidemiological or population studies.
- Host: Natural or Laboratory host to the organism from which sequenced molecule was obtained.
- PCR primers, including direction, name and sequence
- Sex: Sex of the organism from which the sequence derives.
- Tissue-type: Type of tissue from which sequence derives.
- Dev-stage: Developmental stage of organism.
- Identified-by: Name of expert who identified specimen taxonomically.
- Bio-material: An identifier of the stored biological material from which the sequence was obtained.
- Organism: This may be manually entered, or guided by the NCBI taxonomy browser.

For further information for information about these fields, please see:
The GenBank Submissions Handbook: http://www.ncbi.nlm.nih.gov/books/NBK53701/
NCBI Sequin Help Documentation: http:/www.ncbi.nlm.nih.gov/Sequin/sequin.hlp.html

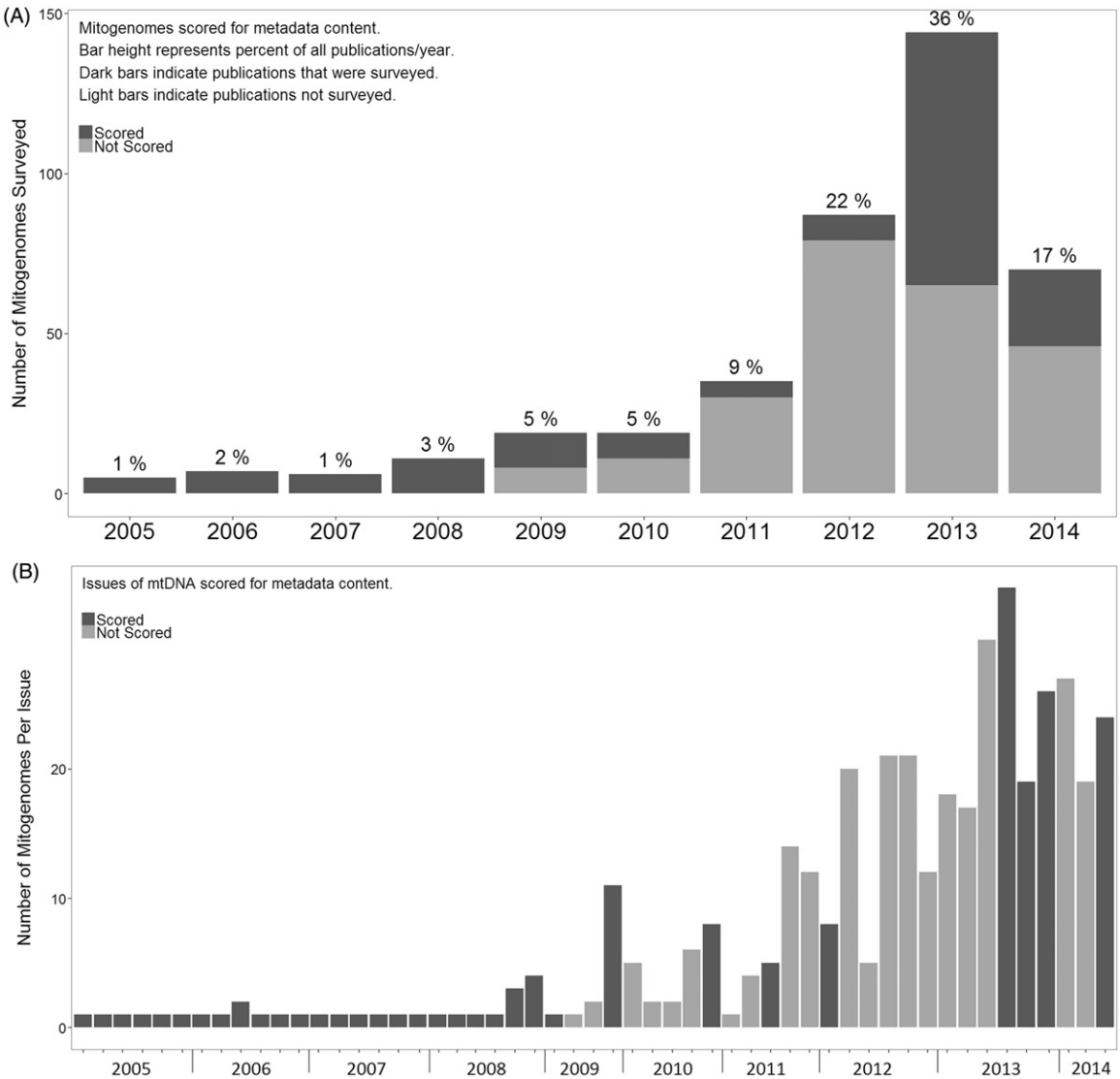

Figure 1. (A) Number of mitogenomes from *Mitochondrial DNA* scored for metadata content, per year (dark grey), combined with available mitogenomes that were not scored (light grey). Percentages per year listed above the bars are calculated from the sum of all mitogenomes published in mtDNA, to date. (B) Number of mitogenomes per volume-number, showing fine-scale coverage of scoring across volumes.

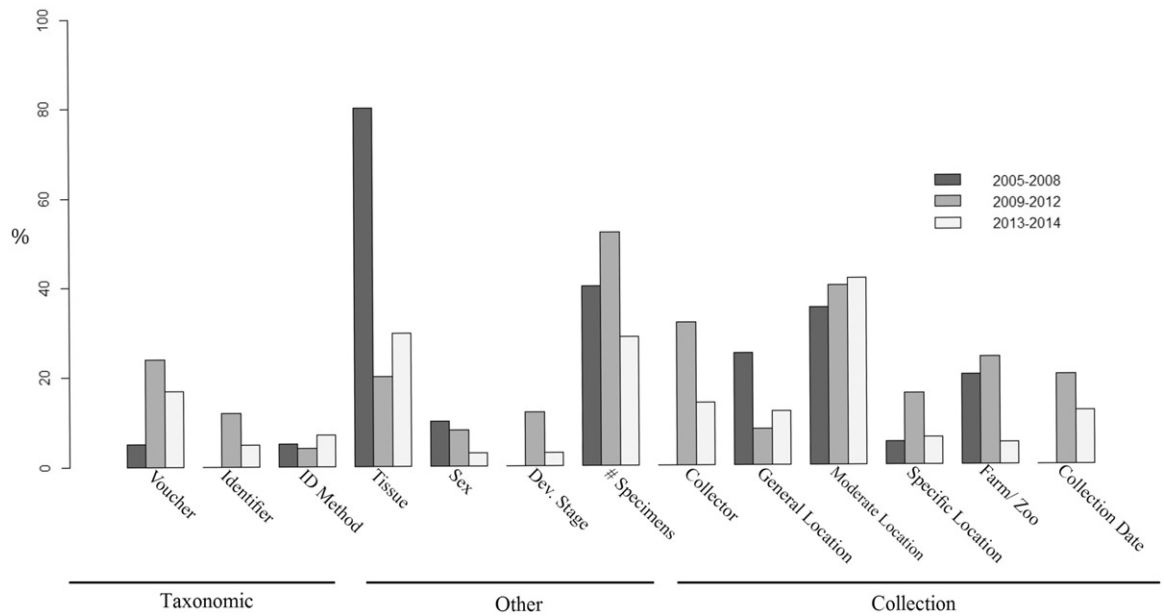

Figure 2. Percentages of projects (publications and Genbank records) that contained mitogenome specimen metadata. Project numbers for 2005–2008 ($n = 20$) 2009–2012 ($n = 25$), and 2013–2014 ($n = 101$).

including the sample data, is available as supplementary material (Supplementary R code (data + code), Supplementary Table 3 (data alone)).

Choice of metadata categories: collection, biological and taxonomic information

We grouped the metadata we reviewed into three categories: collection, taxonomic, and biological information, and selected metadata that overlapped across the MIGS, MIMARKS, SIMPLEDWC and BARCODE data standards as core fields for inclusion. Additionally, we also scored metadata for several fields we created, for a total of 13 metadata fields scored; we describe all categories, and fields below. Metadata fields for each of the 146 mitogenomes, and their Genbank records (collectively referred to as projects), were simply assigned a ''1'' if they included information for a metadata field and a ''0'' if they did not.

For collection metadata we recorded presence of a collection date, name of the collector, and collection location. Location metadata, if present, were further split into four fields: (1) general (specified to country or province), (2) moderate (city or county), (3) specific (Lat/Long/GPS coordinates or an address), (4) if samples were collected from a farm, zoo, university or research center then this was recorded as 'farm/zoo', because information about the original, natural population sampled was not clear. For taxonomic metadata, we recorded whether voucher specimens could be accessed through information about the institution storing, their collection, or accession numbers, as well as identification methods and the name of the identifier. Finally, for biological metadata we recorded information about: tissue type, sex, life stage and the number of specimens that contributed DNA to the final mitogenome sequence.

Accuracy of mitogenome specimen identification

To calibrate specimen identification of mitogenome sequences in a voucher-based framework, we used BOLD (Ratnasingham & Hebert 2007). Currently, BOLD holds 1,787,347 cytochrome c oxidase subunit 1 (*CO1*) sequences identified to species. BOLD also supports high resolution specimen images which can serve as e-vouchers (Monk & Baker, 2001; Steinke et al., 2008). In addition to sequence trace files, BOLD allows for the storage of: PCR primers, collection location and accuracy, sampling protocols, sex, life stage, taxonomic identification methods, and physical voucher status.

We focused our identifications on all mitogenomes for Lepidoptera ($n = 11$) (Arthropoda: Insecta) and Perciformes ($n = 29$) (Chordata: Actinopterygii) found in the literature review – as well as additional specimens from these taxa available from Genbank (randomly chosen, as described below). These taxa were chosen because they were abundantly represented in the literature review, and are particularly well curated on BOLD (Hausmann et al., 2011; Hebert et al., 2010; Ward et al., 2005, 2009). To expand beyond specimens from the literature review, all available Lepidoptera mitogenomes from Genbank were downloaded, for a total of 107. All Perciformes mitogenomes examined in the literature review were downloaded from Genbank, and included with an additional 78 Perciformes mitogenomes randomly subsampled from Genbank. This yielded a total number equal to Lepidoptera ($n = 107$). This subsampling was done by assigning each Perciformes Genbank accession number an ascending number and using a random number generator (www.random.org) for selection. This subsampling was done prior to the bony fish taxonomic revision, suggested by Betancur & colleagues (2013) that have been implemented in Genbank (Personal communication with Scott Federhen, head of the NCBI taxonomy group). As a result, Genbank's current accounting of the number of mitogenomes assigned to Perciformes has changed.

CO1 sequences from the Lepidoptera and Perciformes mitogenomes were extracted from the full mitogenome Genbank record, and aligned using Genbank-to-TNT (Goloboff & Catalano, 2012), and were subsequently queried against BOLD. Because BOLD mines sequence data from Genbank, all identifications were inspected manually to exclude self-hits. Results from BOLD were sorted into one of four kinds of results: (1) 'Match' – when the query matched a single species in BOLD; (2) 'Possible misidentification' – when the query matched one different species; (3) 'Closer examination required' – when the query matched more than one species; (4) 'Not available' – the specimens being queried did not to match any species in BOLD.

Lastly, we used R 3.0.0 (R Core Team, 2014) to perform a two sample test for equality of proportions, for relative comparisons between the BOLD results in the four results (just above) from Perciformes and Lepidoptera. All BOLD queries were made between January 14 – February 13, and on 8 June 2014 using the ''Species Level Barcode Records'' in the OpenIdEngine on BOLD (http://www.boldsystems.org/index.php/IDS_OpenIdEngine). All specimens examined, and their BOLD identification results are presented in Supplementary Table 2.

Results

From the overall sampling of the 146 projects reviewed here, we found the metadata content for mitogenomes is sparse. Thirty projects had zero specimen metadata beyond a species name. The highest reported category, at 40%, was our liberal ''moderate'' location where we accepted any information describing city, county or province-like locations, and the lowest category was ''Developmental Stage'' (larvae, juvenile, adult, etc.) at only 4%. Table 1 presents a summary of the proportional coverage by category, and we present the scored dataset for all 146 projects in Supplementary Table 1.

Collection metadata

We found 71% of projects provided either general, moderate, specific or farm/zoo collection locality data, although only 8% of these were for a specific location Figure 2 (Supplementary Table 1). Collection locations for 10% of all projects came from a farm, aquaculture facility, breeding center, zoo or university (i.e. not a natural population). Only one of these zoo/farm projects, the Asiatic black bear *Ursus thibetanus ussuricus* provided contextual information about the original wild population (Choi et al., 2010). Only 15% of projects provided the name of a collector. Lastly, only 12% of all projects provided a collection date.

Taxonomic metadata

None of the 146 projects provided all three of the taxonomic metadata categories we examined, presented in Supplementary Table 1. Only five projects provided metadata in more than one of these categories. Six percent of the projects provided the names of identifiers, 6% provided specific methods used to identify specimens, and only 16% specify the voucher specimens that were sequenced (Supplementary Figure 1, Supplementary Table 1).

Biological metadata

Tissue type was reported in 35% of all projects. A specimen's sex and developmental stage were reported in 5% and 4% of projects, respectively. Thirty four percent of all projects reported the number of specimens that contributed mitochondrial DNA to the finished mitogenome sequence.

Accuracy of mitogenome specimen identification

Thirty-four percent of Lepidoptera and 36% of Perciformes mitogenomes matched specimens sharing the same name in BOLD (Table 2). Thirty-five percent of Lepidoptera and 13% of Perciformes mitogenomes are possibly misidentified as their closest match was to a different species. In many of these cases there is scant opportunity to reexamine the original specimens since only 16% of projects deposited vouchers (Supplementary Figure 2). Thirty-four percent of Lepidoptera and 45% of Perciformes fell into the 'closer examination required' category where the query matched to more than one species. After the elimination of self hits, all remaining mitogenomes did not match to any specimens in BOLD. All four BOLD-related results ('match', 'closer examination required', 'possible mis-ID', and 'not in BOLD') did not differ significantly between Lepidoptera and Perciformes via a two sample test for equality of proportions (p values in Table 2). Brief notes regarding each mitogenome, and subsequent BOLD identification results are provided as supplementary material (Supplementary Notes 1).

Table 2. Results from querying *CO1* sequences from Lepidoptera and Perciformes mitogenomes into the BOLD species level identification engine. The two groups were compared via a two sample test for equality of proportions.

	Lepidoptera ($n = 107$)		Perciformes ($n = 107$)		
	Total	%	Total	%	p Value
Match	36	34	38	36	0.8971
Not in BOLD	21	20	10	9	0.0623
Closer Examination Required	34	32	45	42	0.1739
Possible Mis-ID	16	15	14	13	0.8593

Discussion

We were surprised to discover our answer to question 1, ''what proportions of mitogenomes are accompanied by core collection, taxonomic and biological metadata?'' was that none of 146 mitogenome projects included all 13 metadata fields we examined; and only 17 projects provided at least one field across these three categories. Many of the publications stated that their mitogenomes are valuable for future studies; for example several indicated potential uses for population genetics, but these same studies lacked information about collection information - particularly the locations (Bo et al., 2013; Hwang et al., 2013; Li & Zou 2013; Li et al., 2013; Shi et al., 2013; Xiang et al., 2013; Zhang et al., 2013). Several publications also suggested conservation applications from their work (He et al., 2013; Jia et al., 2013; Yang et al., 2013a, b; Zeng et al., 2013), but mitogenomes without basic metadata lack context for future applications (Brito, 2004; Vink et al., 2012). Also, publications frequently stated that their new sequence data could clarify species identifications or phylogenetic relationships (Gai et al., 2013; Gao et al., 2013; Li et al., 2013; Liu et al., 2013; Wan et al., 2013), yet only 5% of all publications refer to the identification methods they use. These results agree with previous literature reviews suggesting the legacy of many sequencing efforts (generally) is questionable; an early review by Ruedas et al. (2000) found only 27% of sequence-generating papers explicitly indicated the specimens examined and their voucher locations, while an equal number provided zero specimen information. A later, similar, review of 80 community ecology papers published between 2005 and 2007 found 62.5% of them made no reference to the taxonomic methods used to identify their study organisms, and 2.5% only deposited voucher specimens (Bortolus, 2008).

In discussing our preliminary results amongst ourselves, and other colleagues, we found these patterns of missing metadata prompted questions about their consequences and costs. But these appear difficult to calculate, in part perhaps, because missing metadata don't allow for revisionary studies. For example, Locke & Coates (2008) reviewed studies dating back to 1967 that published information about Caribbean corals that cite the name *Madracis mirabilis.* They found that in approximately half of the studies, the data referred to as *M. mirabilis* couldn't be linked with a single species. Consequently, they calculated this lack of attribution resulted in approximately 4 million USD spent for questionable results – based on data with limited use. Specifically, they indicate that results of these studies could have been reusable

if they had provided taxonomic metadata and consulted the International Code of Zoological Nomenclature (ICZN 1999; http://iczn.org/).

Taxonomic metadata, such as voucher specimens, are invaluable resources for re-confirming species identifications. For example, in a case where the phylogenetic placement of *Oxychloe* (a plant genus within the rush family Juncaceae) was under debate, Kristiansen et al., (2005) re-examined the voucher specimens and determined that chimeric and contaminated sequences were likely responsible. Similarly, in a phylogenetic case of two difficult-to-morphologically-distinguish species of Vultures (*Cathartes*) the presence of a voucher specimen allowed for a previous misidentification to be corrected (Griffiths & Bates, 2002). Such examples of re-evaluation appear to be relatively rare in the literature, and this may be due to lack of available metadata for re-accessing specimens, or sharing revisionary results. For an excellent review on the utility, and recommendations for specimen vouchering see Wheeler (2003).

When mitogenomes – even those without vouchers – are linked with accurate collection metadata, using standard formats such as Darwin Core (http://rs.tdwg.org/dwc/), or NCBI via GenBank, they can be connected to a wealth of species distribution information present in databases such as BOLD, and the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) (Guralnick et al., 2006). And this connectivity allows mitogenomes to contribute to, and be evaluated with, evergrowing global datasets.

Regarding our second and third questions involving identification accuracy, we found that these were expectedly difficult to answer because taxonomic identifications can be idiosyncratic, and the comparative material on BOLD may be questionable or not available (Meyer & Paulay, 2005). In particular, we found our third question ('Are mitogenomes with basic taxonomic metadata less likely to represent an incorrect species…'), was unaddressable because no studies provided metadata for all three taxonomic fields (voucher, ID method, identifier), and only five projects included any two. With respect to our second question, ''How many mitogenomes may represent an incorrect, or ambiguous assignment to a species?'' our results using BOLD suggest that 15% of Lepidoptera, and 13% of Perciformes mitogenomes have their closest match to different species (Table 2, Supplementary Notes 1). Incongruity within genomic databases is nothing new (Harris, 2003; Lis & Lis, 2011; Shen et al., 2013), although the potential to access vouchers on BOLD, facilitates re-examination. However, only 17% of mitogenome projects provided information about their vouchers. The number of mitogenomes requiring 'closer taxonomic examination' (34% of Lepidoptera and 33% of Perciformes), emphasizes the need for clarity of taxonomic practice given the amount of different organism names that share the same *CO1* sequences.The BOLD results from all four results ('match', 'closer examination required', 'possible mis-ID', and 'not in BOLD') did not differ significantly between Lepidoptera and Perciformes, and it may be interesting to see if this result is a pattern across other taxon groups.

Lastly, 21 mitogenomes were scored as 'not in BOLD' because there were no similar DNA barcodes on bold. If deposited in this database, the *CO1* regions from these mitogenomes could serve as DNA barcodes. This barcode's connection to the rest of the mitogenome could make this resource more discoverable, and possibly a high profile sequence of repurposeable value to many communities of molecular biologists.

In conclusion, we wish to highlight our observations about metadata in light of future mitogenome publications in the Journal *Mitochondrial DNA*. We call on the editors of *Mitochondrial DNA* to establish and enforce reasonable metadata standards, in their nucleotide deposition guidelines, for the publication of mitogenomes in this journal. We propose all publications submitting novel mitogenomes archived in GenBank could be accompanied by information in the 13 GenBank fields listed in Box 1 (and shown correlated with metadata standards in Table 1). Lastly, we gratefully recognize the difficult work, for all authors, to publically share sequence data; and we wish to encourage metadata standards that increase the value of these contributions.

Acknowledgements

We wish to thank James Robertson from the centre for Biodiversity Genomics for suggesting that Lepidoptera would make an interesting specimen identification case study. We would also like to thank Scott Federhen, head of the NCBI taxonomy group, for clarifying the changes made to the Perciformes taxonomy. We would like to thank Greg Singer for querying mitogenome sequence data in BOLD for an earlier version of this manuscript. Finally, we wish to thank the Hanner Lab graduate students for their valuable comments, on this work as it progressed.

Declaration of interest

The authors declare no conflict of interest and are entirely responsible for the writing of this paper. The authors received no specific funding for writing this manuscript.

References

Agerer R, Ammirati J, Baroni TJ, Blanz P, Courtecuisse R, Desjardin DE, Gams W, et al. (2000). Open letter to the scientific community of mycologists: ''Always deposit vouchers''. Mycorrhiza 10:95–7.

Betancur RR, Broughton RE, Wiley EO, Carpenter K, Lopez JA, Li C, Holcroft NI, et al. (2013). The tree of life and a new classification of bony fishes. PLoS Curr 5. doi: 10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.

Bo Z, Xu T, Wang R, Jin X, Sun Y. (2013). Complete mitochondrial genome of the Osbeck's grenadier anchovy *Coilia mystus* (Clupeiformes, Engraulidae). Mitochondrial DNA 24:657–9.

Bortolus A. (2008). Error Cascades in the Biological Sciences: The unwanted consequences of using bad taxonomy in ecology. AMBIO 37:114–18.

Botero-Castro F, Delsuc F, Douzery EJ. (2014). Thrice better than once: Quality control guidelines to validate new mitogenomes. Mitochondrial DNA. [Epub ahead of print]. DOI: 10.3109/19401736.2014.900666.

Brito D. (2004). Lack of adequate taxonomic knowledge may hinder endemic mammal conservation in the Brazilian Atlantic Forest. Biodivers Conserv 13:2135–44.

Carr SM, Marshall HD. (2008). Intraspecific phylogeographic genomics from multiple complete mtDNA genomes in Atlantic cod (*Gadus morhua*): Origins of the ''codmother,'' transatlantic vicariance and midglacial population expansion. Genetics 180:381–9.

Choi EH, Kim SK, Ryu SH, Jang KH, Hwang UW. (2010). Mitochondrial genome phylogeny among Asiatic black bear *Ursus thibetanus* subspecies and comprehensive analysis of their control regions. Mitochondrial DNA 21:105–14.

Desalle R, Kolokotronis S-O. (2009). Mitochondrial DNA announces new fast-track manuscript category: Mitogenome Announcements. Mitochondrial DNA 20:1–1.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, et al. (2008). The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26:541–7.

Foley DH, Wilkerson RC, Rueda LM. (2009). Importance of the ''what,'' ''when,'' and ''where'' of mosquito collection events. J Med Entomol 46:717–22.

Gai Y, Ma H, Sun X, Ma J, Li C, Yang Q. (2013). The complete mitochondrial genome of *Cermatobius longicornis* (Chilopoda: Lithobiomorpha: Henicopidae). Mitochondrial DNA 24:331–2.

Gao T, Li N, Zhang Y, Shi P. (2013). The complete mitochondrial genome of Japanese sandeel *Ammodytes personatus* (Perciformes, Ammodytidae): Rare structure in control region compared. Mitochondrial DNA 24:320–2.

Goloboff P, Catalano S. (2012). GB-to-TNT: Facilitating creation of matrices from GenBank and diagnosis of results in TNT. Cladistics 28: 503–13.

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, et al. (2014). Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol 10:e1003542.

Griffiths CS, Bates JM. (2002). Morphology, genetics and the value of voucher specimens: An example with *Cathartes* vultures. J Raptor Res 36:183–7.

Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ, Biogeomancer Working G. (2006). BioGeomancer: Automated georeferencing to map the world's biodiversity data. PLoS Biol 4:e381.

Hanner R. 2009. Data standards for BARCODE records in INSDC (BRIs). Available at: http://barcoding.si.edu/pdf/dwg_data_standards-final.pdf (Accessed 2 February 2015).

Harris JD. (2003). Can you bank on Genbank? Trends Ecol Evol 18: 315–17.

Hausmann A, Haszprunar G, Hebert PD. (2011). DNA barcoding the geometrid fauna of Bavaria (Lepidoptera): Successes, surprises, and questions. PLoS One 6:e17134.

He B, Lai T, Peng Z, Wang X, Pan L. (2013). Complete mitogenome of the Areolate grouper Epinephelus areolatus (Serranidae, Epinephelinae). Mitochondrial DNA 24:498–500.

Hebert PD, Dewaard JR, Landry JF. (2010). DNA barcodes for 1/1000 of the animal kingdom. Biol Lett 6:359–62.

Hobern DA, Apostolico E, Arnaud JC, Bello D, Canhos G, Dubois D, Field et al., (2013) Global Biodiversity Informatics Outlook. GBIF Secretariat (Copenhagen):1–41.

Hwang DS, Byeon HK, Lee JS. (2013). Complete mitochondrial genome of the freshwater sculpin *Cottus koreanu*s (Scorpaeniformes, Cottidae). Mitochondrial DNA 24:490–1.

International Commission on Zoological Nomenclature (ICZN). (1999). International Code of Zoological Nomenclature, 4th Edition. London, UK: International Commission on Zoological Nomenclature.

Jia XY, Li YW, Wang DQ, Tian HW, Xiong X, Li SH, Chen DQ. (2013). The complete mitochondrial genome of *Liobagrus kingi* (Teleostei, Siluriformes: Amblycipitidae). Mitochondrial DNA 24:323–5.

Kane NC, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk QCB. (2012). Ultra-barcoding in cacao (*Theobroma* spp.) using whole chloroplast genomes and nuclear ribosomal DNA. Am J Bot 99: 320–9.

Kristiansen K, Cilieborg M, Drábková L, Jørgensen T, Petersen G, Seberg O. (2005). DNA Taxonomy — The riddle of *Oxychloë* (Juncaceae). Syst Bot 30:284–9.

Krjutškov K, Koltsina M, Grand K, Vosa U, Sauk M, Tonisson N, Salumets A. (2014). Tissue-specific mitochondrial heteroplasmy at position 16,093 within the same individual. Curr Genet 60:11–6.

Li M, Zou K. (2013). Complete mitochondrial genome of *Anodontostoma chacunda* (Clupeiformes: Clupeidae): Genome characterization and phylogenetic consideration. Mitochondrial DNA 24:507–9.

Li Y, Pan X, Song N, Yanagimoto T, Gao T. (2013). Complete mitochondrial genome of Japanese *Konosirus punctatus* (Clupeiformes: Clupeidae). Mitochondrial DNA 24:481–3.

Lis JA, Lis B. (2011). Is accurate taxon identification important for molecular studies? Several cases of faux pas in pentatomoid bugs (Hemiptera: Heteroptera: Pentatomoidea) Zootaxa 50:47–50.

Liu X, Guo Y, Wang Z, Liu C. (2013). The complete mitochondrial genome sequence of *Trichiurus nanhaiensis* (Perciformes: Trichiuridae). Mitochondrial DNA 24:516–17.

Locke JM, Coates KA. (2008). What are the costs of bad taxonomic practices: And what is *Madracis mirabilis*? Proceedings of the 11th International Coral Reef Symposium; 2008 July 7-11; Ft. Lauderdale, Florida.

Meyer CP, Paulay G. (2005). DNA barcoding: Error rates based on comprehensive sampling. PLoS Biol 3:e422.

Monk RR, Baker RJ. (2001). e-Vouchers and the use of digital imagery in natural history collections. Museology 10:1–8.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Ratnasingham S, Hebert PDN. (2007). BOLD: The barcode of life data system (www.barcodinglife.org). Mol Ecol Notes 7:355–64.

Ruedas LA, Salazar-Bravo J, Dragoo JW, Yates TL. (2000). The importance of being earnest: What, if anything, constitutes a ''specimen examined''? Mol Phylogenet Evol 17:129–32.

Shen YY, Chen X, Murphy RW. (2013). Assessing DNA barcoding as a tool for species identification and data quality control. PLoS One 8:e57125.

Shi QH, Zhao F, Hao JS, Yang Q. (2013). Complete mitochondrial genome of the common evening brown, *Melanitis leda* Linnaeus (Lepidoptera: Nymphalidae: Satyrinae). Mitochondrial DNA 24: 492–4.

Steinke D, Hanner R, Hebert PDN. (2008). Rapid high-quality imaging of fishes using a flat-bed scanner. Ichthyol Res 56:210–11.

Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, Haendel MA. (2013). On the reproducibility of science: Unique identification of research resources in the biomedical literature. Peer J 1:e148.

Vink J, Paquin P, Cruickshank RH. (2012). Taxonomy and irreproducible biological science. BioSci 62:451–2.

Wan Q, Tao G, Cheng Q, Chen Y, Qiao H. (2013). The complete mitochondrial genome sequence of *Pseudobagrus ussuriensis* (Siluriformes: Bagridae). Mitochondrial DNA 24:333–5.

Ward RD, Hanner R, Hebert PD. (2009). The campaign to DNA barcode all fishes, FISH-BOL. J Fish Biol 74:329–56.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD. (2005). DNA barcoding Australia's fish species. Philos Trans R Soc Lond B Biol Sci 360:1847–57.

Wheeler T. A. (2003). The role of voucher specimens in validating faunistic and ecological research. Can J Arthropod Identif (Terrestrial Arthropods) 9:1–21.

Wieczorek J, Bloom D, Guralnick R, Blum, S, Döring M, Giovanni R, Robertson T, Vieglais D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. PLoS One 7:e29715.

Xiang T, Wang B, Liang X, Jiang J, Li C, Xie F. (2013). Complete mitochondrial genome of *Paramegophrys oshanensis* (Amphibia, Anura, Megophryidae). Mitochondrial DNA 24:472–4.

Yang B, Zhang J, Yamaguchi A, Zhang B. (2013a). Mitochondrial genome of *Dasyatis bennettii* (Chondrichthyes: Dasyatidae). Mitochondrial DNA 24:344–6.

Yang C, Xiang C, Zhang X, Yue B. (2013b). The complete mitochondrial genome of the Alpine musk deer (*Moschus chrysogaster*). Mitochondrial DNA 24:501–3.

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29:415–20.

Zeng T, Tu F, Ma L, Yan C, Yang N, Zhang X, Yue B, Ran J. (2013). Complete mitochondrial genome of blood pheasant (Ithaginis cruentus). Mitochondrial DNA 24:484–6.

Zhang Z, Zhao L, Song N, Gao T. (2013). The complete mitochondrial genome of Johnius grypotus (Perciformes: Sciaenidae). Mitochondrial DNA 24:504–6.

Supplementary material available online.

Supplementary Tables 1 and 2 & Figures 1 and 2.