Marine Mammal Science

MARINE MAMMAL SCIENCE, 33(SPECIAL ISSUE): 132–155 (June 2017)

DOI: 10.1111/mms.12411

Guidelines and quantitative standards to improve consistency in cetacean subspecies and species delimitation relying on molecular genetic data

This is the sixth of six papers forming a special issue of Marine Mammal Science (Vol. 33, Special Issue) on delimiting cetacean subspecies using primarily genetic data. An introduction to the special issue and brief summaries of all papers it contains is presented in Taylor et al. (2017). Together, these papers lead to a proposed set of guidelines that identify informational needs and quantitative standards (this paper) intended to promote consistency, objectivity, and transparency in the classification of cetaceans. The guidelines are broadly applicable across data types. The quantitative standards are based on the marker currently available across a sufficiently broad number of cetacean taxa: mitochondrial DNA control region sequence data. They are intended as "living" standards that should be revised as new types of data (particularly nuclear data) become available.

Barbara L. Taylor,[1] **Frederick I.Archer, Karen K. Martien,** Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla Shores Drive, La Jolla, California 92037, U.S.A.; **Patricia E. Rosel,** Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, 646 Cajundome Boulevard, Lafayette, Louisiana 70506, U.S.A., **Brittany L. Hancock-Hanser, Aimee R. Lang, Matthew S. Leslie, Sarah L. Mesnick, Phillip A. Morin, Victoria L. Pease, William F. Perrin, Kelly M. Robertson,** Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla Shores Drive, La Jolla, California 92037, U.S.A.; **Kim M. Parsons,** National Oceanic and Atmospheric Administration, National Marine Fisheries Service, National Marine Mammal Laboratory, Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, Washington 98115, U.S.A.; **Amélia Viricel,** Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, 646 Cajundome Boulevard, Lafayette, Louisiana 70506, U.S.A. and Environnement et Sociétés, UMR 7266 CNRS – Université de La Rochelle, 2 rue Olympe de Gouges, 17000 La Rochelle, France; **Nicole L. Vollmer,** Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, 646 Cajundome Boulevard, Lafayette, Louisiana 70506, U.S.A. and National Systematics Laboratory, National Marine Fisheries Service, NOAA, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, U.S.A.; **Frank Cipriano,** Conservation Genetics Laboratory, San Francisco State University, San Francisco, California 94132, U.S.A.; **Randall R. Reeves,** Okapi Wildlife Associates, 27 Chandler Lane, Hudson, Quebec J0P 1H0, Canada; **Michael Krützen,** Anthropological Institute and Museum, University of Zurich, Winterhurerstrasse 190, CH-8057 Zurich, Switzerland; **C. Scott Baker,** Marine Mammal Institute and Department of Fisheries and Wildlife, Oregon State University, 2030 SE Marine Science Drive, Newport, Oregon 97365, U.S.A.

[1]Corresponding author (e-mail: barbara.taylor@noaa.gov).

Abstract

Taxonomy is an imprecise science that delimits the evolutionary continuum into discrete categories. For marine mammals, this science is complicated by the relative lack of morphological data for taxa that inhabit remote and often vast ranges. We provide guidelines to promote consistency in studies relying primarily on molecular genetic data to delimit cetacean subspecies from both populations and species. These guidelines identify informational needs: basis for the taxonomic hypothesis being tested, description of current taxonomy, description of relevant life history, sample distribution, sample size, number and sequence length of genetic markers, description of measures taken to ensure data quality, summary statistics for the genetic markers, and analytical methods used to evaluate the genetic data. We propose an initial set of quantitative and qualitative standards based on the types of data and analytical methods most readily available at present. These standards are not expected to be rigidly applied. Rather, they are meant to encourage taxonomic arguments that are consistent and transparent. We hope professional societies, such as the Society for Marine Mammalogy, will adopt quantitative standards that evolve as new data types and analytical methods become widely available.

Key words: cetacean taxonomy, genetic data, guidelines, quantitative standards, species definition, subspecies definition, taxonomy.

Species evolve through unique pathways and usually diverge along a continuum, making delimitation of species boundaries a fascinating and difficult process. Delimiting subspecies is even more difficult because ongoing gene flow is expected, and therefore without a clear biological threshold, their definition has some obviously subjective qualities. While we cannot reduce the wonderful complexity of speciation to a simple formula, we can try to delimit clearly and consistently the units that we use to better understand how species evolve, and as conservation scientists, to determine the magnitude and timing of management actions.

Taylor *et al.* (2017) show that cetacean taxonomy tends to reflect underclassification errors, *i.e.*, a failure to recognize subspecies or species-level classification when such a distinction is justified. This underclassification results from the difficulties of amassing sufficient morphological data for traditional taxonomic studies when species have large geographical ranges and are sometimes rare in abundance or inaccessible in their distribution. Given these difficulties, genetic data have become increasingly valuable in studies of cetacean taxonomy (Reeves *et al.* 2004; Rosel *et al.* 2017*a*, *b*). Rosel *et al.* (2017*b*) reviewed recent papers that relied on molecular genetic data to make taxonomic cases for marine mammals and found little consistency in approach. Often, key information was absent, such as maps of species distribution and sampling locations, descriptions of life history, and information on population dynamics. These gaps often made it impossible to judge the validity or strength of the case.

In a workshop on subspecies delimitation held at the 2009 biennial conference of the Society for Marine Mammalogy, there was consensus that guidelines would help molecular geneticists make more consistent taxonomic cases. Here, we begin by developing such guidelines and then extend the effort by proposing an initial set of quantitative and qualitative standards. The guidelines provide a list of items to be included in any taxonomic argument and that should enable researchers to judge whether their data are generally adequate. The quantitative standards specify a level of evidence sufficient to justify subspecies delimitation. The qualitative standards, like providing evidence that rules out male-mediated gene flow, suggest the type of

evidence needed but do not further quantify the standard. The uniqueness of every study, both in terms of the evolutionary history of the focal taxon or taxa and the data available to address the taxonomic questions, means that no set of standards will be perfect. Nevertheless, the existence of standards will allow some cases to be made easily and for others, provide researchers with an impetus to hone their arguments against a consistent set of standards. We expect that the guidelines and standards presented here will evolve as new studies reveal the strengths and weaknesses of our proposals and as new types of data and analytical methods become available that improve our ability to understand evolution and grasp the essence of biodiversity. We hope the process developed here can serve as a template for professional societies to build upon.

Definitions of "Species" and "Subspecies" Concepts

Our guidelines and standards are based on the subspecies and species definitions set out by Taylor *et al*. 2017 and de Queiroz 2007, respectively. We summarize those definitions here and refer the reader to Taylor *et al*. 2017 for detailed discussion. Taylor *et al.* 2017 developed the following subspecies definition: "A *subspecies* is a population, or collection of populations, that *appears to be* a separately evolving lineage with discontinuities resulting from geography, ecological specialization, or other forces that restrict gene flow to the point that the population or collection of populations is diagnosably[2] distinct." This definition is consistent with the subspecies concept discussed in Reeves *et al.* (2004), but is more explicit in requiring diagnosability. Here, we use the definition of diagnosability as proposed by Archer *et al.* (2017): "diagnosability is a measure of the ability to correctly determine the taxon of a specimen of unknown origin based on a set of distinguishing characteristics." In agreement with Patten and Unitt (2002), subspecies cannot be clinal[3] and subspecies, unlike species, do not have to be reproductively isolated from other subspecies. Taylor *et al.* (2017) also emphasize that diagnosability must be based on a heritable character.

The units flanking subspecies are populations at the lower boundary and species at the upper boundary. For "population" we use the Taylor *et al*. 2017 definition: "A *population* refers collectively to a series of Units to Conserve, ranging from Demographically Independent Populations (DIP)[4] to Evolutionarily Significant Units (as defined in Waples *1995*[5])." We base our species definition on the unified species concept of de Queiroz 2007, which is "A *species* is a separately evolving metapopulation lineage." To make our definition consistent with our population and subspecies

[2]Diagnosability implies a high probability (but not necessarily a 100% probability) of identifying an individual as belonging to the taxon.

[3]In social species like cetaceans, clinal refers to a series of populations that differ from one another (for example, frequency differences in mitochondrial DNA) but where no strong discontinuities are apparent. Many coastal species, like porpoises, are found in a stepping-stone pattern with positive correlation between genetic and geographic distance (*i.e.,* isolation by distance but no clear discontinuities resulting from restrictions to gene flow).

[4]A Demographically Independent Population is a sympatric group of individuals whose dynamics are more a consequence of births and deaths within the group (internal dynamics) than of immigration or emigration (external dynamics) (Taylor 2005, Taylor *et al.* 2010).

[5]The ESU is defined as a DIP or a collection of DIPs that is substantially reproductively isolated from other conspecific population units and represents an important component in the evolutionary potential of the species (Waples 1995).

definitions and because several cetacean species are single populations our definition is: "A *species* is a separately evolving lineage comprised of a population or collection of populations."

The main difference between species and subspecies is that species *are* a separately evolving lineage and subspecies *appear to be* headed in that direction. By considering that subspecies *appear to be* separately evolving lineages, we are capturing within the definition several different types of uncertainty encountered by practitioners. When the degree of divergence is small, an apparently diverging lineage may reconverge as conditions change (see fig. 1 in Taylor *et al.* 2017). For example, a barrier to gene flow, such as water of an unsuitable temperature, may be removed, allowing gene flow to resume and divergence to cease. There may be cases where there is only partial divergence with some ongoing gene flow, as seems to be the case for spinner dolphin (*Stenella longirostris*) subspecies in the Eastern Tropical Pacific. Finally, there may be cases where the evidence for divergence is too weak to make a case for a new species and more evidence is needed. An example would be the case of fin whales (*Balaenoptera physalus*) in the North Pacific and North Atlantic. Mitochondrial DNA (mtDNA) suggests divergence between these ocean basins (Archer *et al.* 2013). Although allopatric distributions in the North Pacific, North Atlantic, and Southern Hemisphere suggest that these lineages have diverged, the case for lineage divergence would be much stronger with nuclear DNA data to support it. Hey *et al.* (2001) would probably consider this case as an hypothesis for a species where sufficient evidence has not yet been acquired for testing.

With these definitions in mind, we developed the following set of principles to steer the process of developing guidelines and quantitative standards to allow use of molecular genetic data to improve cetacean alpha and gamma[6] taxonomy:

- minimize overall taxonomic errors, with consideration of the balance between the consequences of over- and underclassification,
- apply to both data-poor and data-rich taxa,
- treat taxa equivalently regardless of effective population size,[7]
- promote consistency and transparency,
- allow use of all available relevant lines of evidence, and
- strive to be pragmatic, such that appropriate data could be collected within a 10 yr period.

[6]Alpha taxonomy is the discipline of finding, describing, and naming species, which relates here to efforts to distinguish the upper boundary between subspecies and species. Gamma taxonomy is the study of intraspecific variation, which relates here to delimiting the boundary between subspecies and populations. This effort is not concerned with the relationships among taxa (phylogeny, or beta taxonomy). For example, use of a genetic marker that led to an incorrect topology in a tree but correctly delimited species would not constitute an error in the alpha taxonomy, and therefore is not of concern within our context.

[7]Application of the standards to populations with small effective population size (N_e) could result in overclassification errors (because of lineage sorting or social structure resulting in neighboring populations being completely diagnosable but not because they meet the spirit of the subspecies definition), whereas application to populations with very large N_e could result in underclassification errors (due to incomplete sampling and/or the potential for shared haplotypes in large, recently diverged populations where effects of genetic drift take longer to manifest). We, nevertheless, recommend application of the guiding principles to all taxa, regardless of N_e. Authors must then carefully consider the potential impacts of N_e and social structure on observed divergence and/or diagnosability results in making their final recommendation for classification.

To present a persuasive argument for delimiting new cetacean subspecies or species, a publication should contain the following:

1. Review of Focal Taxon or Taxa
   - ☐ Review of current taxonomy and species concept used
   - ☐ Review of relevant life history characteristics (movement patterns, social structure, habitat or behavioral constraints, effective population size)
   - ☐ Basis for taxonomic hypothesis, including both primary cues and other independent lines of evidence

2. Sampling Considerations
   - ☐ Distribution map or description including (if applicable) feeding and breeding areas, migration route(s), historical distribution, and areas of sympatry and parapatry
   - ☐ Map of all sampling locations
   - ☐ Description of sampling procedures/selection to demonstrate random sampling was conducted
   - ☐ Analyses to identify and remove closely related individuals, as appropriate
   - ☐ Summary of the total number of samples and number of samples per stratum, for each marker, in the final data set
   - ☐ Evidence that sampling is adequate (*e.g.*, discovery curve, power analysis)
   - ☐ Sex of sampled individuals

3. Laboratory Analyses
   - ☐ For mtDNA sequence data, adequate sequence length (minimum 300 bp)
   - ☐ For nuDNA, adequate sequence length or number of loci (10–20 for microsatellites)
   - ☐ Description of laboratory procedures
   - ☐ Description of QA/QC protocols

4. Data Analyses
   - ☐ Descriptive summary statistics, including
     - o For mtDNA – sequence length (excluding primer sequences), haplotypic diversity, nucleotide diversity, and proportion of haplotypes represented by a single individual
     - o For nuDNA – heterozygosity, allelic richness, number of alleles per locus, and tests for linkage and Hardy-Weinberg equilibrium
   - ☐ Estimate of percent diagnosable (not necessary if groups are clearly 100% diagnosable, as in the case of fixed differences)
   - ☐ Explanation to rule out rapid lineage sorting or strong social structure-induced pattern
   - ☐ Estimate of divergence, such as net nucleotide divergence (d_A), percent fixed differences, or divergence time
   - ☐ If using mtDNA and male-mediated gene flow is plausible, an estimate of differentiation based on other data, such as nuDNA
   - ☐ Concordance with any morphological or behavioral characters or discontinuities

5. Taxonomic inference
   - ☐ Synthesis of lines of evidence relative to taxonomic hypothesis
   - ☐ Statement of taxonomic conclusion
   - ☐ Meet requirements of International Code of Zoological Nomenclature

Figure 1. Guidelines for studies of cetacean taxonomy based on genetic data.

Guidelines for Taxonomic Studies Relying on Molecular Genetic Data

Building a case to describe a new taxonomic unit requires marshaling evidence and presenting it in a way that enables other scientists to fully evaluate the validity of the proposed unit. Uncertainties should be described in all steps taken to build a case. We structure our guidelines into five categories of information that must be provided

so that the taxonomic argument based primarily on genetic data can be evaluated: (1) review of the focal taxon or taxa, (2) sampling considerations, (3) laboratory analyses, (4) data analyses, and (5) taxonomic inferences. We discuss each of these categories below, and summarize the guidelines in a checklist for easy reference (Fig. 1). We intend these guidelines to be used both in the study design and data analysis phases of a taxonomic study, and by other scientists when evaluating the strength of a study. We improved the guidelines and standards by having a subset of coauthors use them to evaluate five case studies and we provide those examples to assist readers (Appendix S1).

Category 1. Review of Focal Taxa

Authors should provide the reader with all background information necessary to evaluate the strength of evidence supporting or refuting the delimitation of a new subspecies or species. Authors should operationally define the species or subspecies concept they are using in their work. They should include a review of the current taxonomy summarizing the grounds for previous decisions about separation or nonseparation and identifying strengths and/or weaknesses of the arguments. The review ensures that the reader has the proper context to interpret the case being made given past taxonomic work. In many cases, it may also be necessary to describe life history characteristics relevant to subspeciation/speciation, including seasonal changes in distribution, interannual changes in distribution, breeding range, social structure (specifically the likelihood of female site fidelity coupled with male-mediated gene flow), habitat constraints that contribute to allopatry, historical changes in distribution (likely effects of cooling or warming periods), and types of behavior conducive to constrictions in gene flow (mating system, acoustics, dietary specialization, *etc*.). Authors should also describe population dynamics relevant to speciation, including effective population size (N_e) for different strata (even if this must be approximate, such as "numbers at least in the tens of thousands") and equilibrium status (changes in abundance or connectivity resulting from events like whaling or habitat fragmentation).

Authors also need to point out the cues that formed the basis of the taxonomic hypothesis being tested and those used to stratify the data. These often include observed discontinuities in geographical distribution or differences in morphology, ecology, acoustics, behavior, or other features. In addition to the primary cue used to formulate the hypothesis, authors should review all pertinent independent lines of evidence. For example, samples from resident and transient killer whales (*Orcinus orca*) may be stratified based on morphological cues (saddle patch category, dorsal fin shape), but the two ecotypes are also separable based on acoustics, dietary specialization, group size, and social organization. Highlighting which lines of evidence are likely to have adaptive value will help the reader develop context for the taxonomic argument being made.

Category 2. Sampling Considerations

Sample distribution—To assess the adequacy of sampling, publications must describe the full geographic distribution of the focal taxa and the collection locations of all samples. In most cases, the distribution of the animals and the samples will be communicated most clearly with maps. Distribution maps for migratory species should indicate range during the breeding and feeding seasons and show the

migratory route, particularly if samples have been obtained during the feeding season or along the migratory route. Maps should indicate areas of known or potential sympatry or parapatry between or among the taxa under consideration. Readers should be able to assess the "purity" of strata (*i.e.*, whether failure to reveal diagnostic differences could result from having sampled mixed strata; see Cicero *et al.* 2006). Use of different symbols may be helpful for indicating the stratum to which each sample belongs or the likelihood that stratum origin is unknown (*e.g.*, migratory species sampled in the nonbreeding season). The distribution section should convince the reader that a clinal distribution of animals is not plausible as clinal distributions do not meet our subspecies definition.

If current distribution differs from historical distribution, details of the changes should be provided. The cause(s) for distributional changes will also aid the reader in assessing how they might affect interpretation of the genetic data. For example, a recent expansion in distribution may make founder effects plausible and raise the amount of evidence needed to draw inferences regarding whether observed divergence is a result of recent random genetic effects like founder effects and lineage sorting for social species with low N_e. Reoccupation of portions of the historical distributions during recovery from depletion by commercial whaling may require extra caution in interpreting data given the violation of the assumptions of genetic equilibrium and stable age distribution that are common to many analytical methods.

Sample size—There is no simple answer to the question of how many samples are adequate to address subspecies delimitation. Here, we briefly summarize the importance of adequate sampling and state some general principles for ensuring adequate sample size. Because of the critical importance of sample size when evaluating a taxonomic study, we have included a more detailed discussion of sample size considerations in the Appendix.

The adequacy of a particular set of samples depends on a complex interplay between the effective population size (N_e), social structure, and demographic history within each stratum, as well as the degree of differentiation (effect size) among strata. These characteristics are difficult to estimate and are rarely completely known (Harris and Allendorf 1989, Frankham 1995, Bossart and Prowell 1998, Schwartz *et al.* 1998, Leberg 2005). Inadequate sample size can bias metrics of differentiation and mislead interpretations (Roff and Bentzen 1989, Leberg 2002, Pruett and Winker 2008). Small sample sizes can also lead to an inability to detect differentiation due to poor statistical power (Björklund and Bergek 2009, Morin *et al.* 2009, Landguth *et al.* 2012). If the effect size (the degree of genetic differentiation) can be roughly estimated, researchers should attempt to estimate statistical power to gain some insight into minimum required sample sizes (Ryman and Palm 2006, Ryman *et al.* 2006, Morin *et al.* 2009). Nonetheless, the author must convince the reader that the sample size is sufficient to draw solid inferences for the question at hand and the analytical method used to address it. For some rare cetaceans (including most of the beaked whales), obtaining an ideal sample size may simply not be feasible and special arguments for basing taxonomic decisions on the best available data will need to be made such as providing details on why data accrual is difficult and giving readers an idea of how much time might be required to obtain better data. Such cases may require more judgment (Lim *et al.* 2012).

When evaluating the adequacy of a set of samples, it helps to keep in mind that the underlying goal is to ensure that samples accurately reflect the genetic diversity of the stratum they represent. The implications of this seemingly trivial statement are rarely fully considered or directly examined in practice (Tajima 1995, Waples

1998). Properly summarizing the genetic diversity of a stratum entails ensuring that both a majority of the haplotypes (or alleles in the case of diploid data) are present, and that the haplotypic frequency distribution is similar to that of the entire stratum. Consideration of the potential influence of the relative representation of males and females in the data set is important, particularly for species where there may be differences between the sexes in social philopatry and dispersal. Although one cannot know how representative the observed frequencies are of the actual frequencies without sampling all individuals within a stratum, a good sampling design aims towards this ideal.

Category 3. Laboratory Analyses

Marker choice—As summarized by Martien *et al.* (2017), mtDNA has advantageous attributes for purposes of classification. The ease of generating sequences has resulted in the data available for the pairwise comparisons used by Rosel *et al.* (2017*a*). Results from Rosel *et al.* (2017*a*) show that for most cetacean taxonomic arguments, the mtDNA control region is a good marker choice. A minimum of 300 base pairs (bp) of control region data is often adequate to capture sufficient variation to evaluate divergence and diagnosability (Rosel *et al.* 2017*a*). Longer sequences are preferred and increasingly easy to generate. Shorter sequences tend to result in reduced ability to diagnose, and hence, their use potentially leads to underclassification errors. For subspecies delineation, a region with a high mutation rate, like the control region, is preferred because differentiation of populations into subspecies can occur over a relatively short evolutionary timescale. For strata with low haplotypic diversity, researchers should strive for longer sequences to strengthen inferences (*e.g.*, Morin *et al.* 2010*a*). Note, however, that use of longer sequences will no longer be comparable to the metrics used in the quantitative standards based on control region sequences analyzed by Rosel *et al.* (2017*a*). For taxa with high haplotypic diversity due to high historical abundance (*e.g.*, pelagic subspecies of spinner dolphins, *Stenella longirostris*), the mtDNA control region may not be an adequate marker for subspecies delineation (see details in simulations done by Archer *et al.* 2017).

Rosel *et al.* (2017*b*) found no instances where nuclear sequence data (nuDNA) were used to evaluate the lower subspecies boundary (however, see Andrews *et al.* 2013 with confirmatory data for already-named subspecies), and only a few examples of nuDNA sequences being used to evaluate the upper boundary, though this will change with technological advances. While nuclear loci such as microsatellites and Single Nucleotide Polymorphisms (SNPs) can be used effectively for assignment tests and characterization of social structure and population differentiation, they should be used with caution in evaluation of higher taxonomic units (Martien *et al.* 2017). In general, nuclear loci evolve more slowly and have a larger effective population size than mtDNA, so they drift to fixation at a slower rate. As a result, more loci and/or longer sequences need to be screened to find variation. As next-generation sequencing methods allow rapid and cost-effective screening of many nuclear sequences (*e.g.,* Hancock-Hanser *et al.* 2013), these markers are certain to become more useful in divergence and diagnosability analyses, and in phylogenetic analyses as well (*e.g.*, Bryant *et al.* 2012, Viricel *et al.* 2014). However, given the difficulties in genotyping a sufficient number of loci for enough samples across a wide range of populations, subspecies, and species, it is unlikely that a comparative study of differentiation at multiple nuclear loci, similar to that conducted for mtDNA in Rosel *et al.* (2017*a*), will be available in the near future.

Nonetheless, nuclear loci may be adequate for use in diagnosability as long as other evidence is marshaled that the putative taxa are on independent evolutionary trajectories. An advantage of using nuDNA is that both diagnosability and male-mediated gene flow can be addressed with the same data. Y-markers (see Greminger *et al.* 2010 for methods to generate markers) are particularly relevant because they directly measure male gene flow. It should be noted that the standard assignment tests commonly used with microsatellite and SNP data are not equivalent to traditional classification methods for diagnosability in that they are based on a combination of population allele frequencies and an underlying evolutionary model rather than on identifying diagnostic characteristics. In a simple case, private alleles can be used as diagnostic characters, but a probabilistic assignment test may not be regarded by traditional taxonomists as a valid means of assessing diagnosability. Further effort should be devoted to develop classification algorithms specific to bi- and multiallelic loci that are comparable to traditional methods.

Quality descriptions—All laboratory procedures, both chemistry and equipment, should be described in sufficient detail for future reproducibility. Readers should be convinced of the quality of the data and that the laboratory Quality Assurance and Quality Control (QA/QC) protocols followed were sufficient to identify influential errors and/or quantify the error rate (Taberlet *et al.* 1996, Pompanon *et al.* 2005, Morin *et al.* 2010*b*)

Category 4. Data Analyses

One of the first sections of "results" for genetics papers typically consists of summary statistics to describe characteristics of the markers themselves. These metrics are important because they can affect the performance of different analytical methods and they provide context for the strength of inferences resulting from the analyses. For all markers the final sample size used must be given, and for sequence data the length of each alignment used, excluding the primer sequences, should be specified. Haplotypic and nucleotide diversities (mtDNA) and observed and expected heterozygosities (nuDNA) are strongly correlated with N_e and can aid the reader in inferring long-term N_e if it is not directly known. The percent of singleton haplotypes (haplotypes found in only one sampled individual) characterizes the adequacy of the sampling and provides context for assessing the quality of percent diagnosed and frequency statistics.

Similar summary statistics should also be given for nuDNA, and additional tests should be conducted for both linkage and Hardy-Weinberg equilibrium. These tests inform the reader as to whether multiple strata may have been inadvertently included within a single stratum and suggest whether the assumption of genetic equilibrium has been met.

The exact analyses needed to make a compelling case for delimiting subspecies or species will depend in part on specific aspects of the given case, such as whether the species is known to exhibit female philopatry with potential for male-mediated gene flow. The magnitude of expected differences for the lower *vs*. upper boundary will be important factors in analysis choice. Diagnosability is an important feature to delimit subspecies from populations, both under the subspecies definition we have adopted and in traditional taxonomic literature (Amadon 1949, Helbig *et al.* 2002, Patten and Unitt 2002, Patten 2010). Analytical methods for estimating diagnosability include assignment tests and multivariate methods (see Martien *et al.* 2017 for a detailed review).

Genetic data are sometimes sufficiently powerful to diagnose social groups and recent isolates with high accuracy. Consequently, the taxonomic argument should include evidence that the putative unit is not a social group or a result of rapid lineage sorting in a recent isolate. Analyses to demonstrate genealogical cohesion may be useful for this purpose (*e.g.*, estimating GSI; Cummings *et al.* 2008). Likewise, an estimate of divergence among strata will allow one to evaluate whether they are evolving separately. Martien *et al.* 2017 reviewed numerous methods and metrics that can be applied to the subspecies/species boundary, including net nucleotide divergence (d_A, Nei 1987), percent fixed differences, divergence time estimation, and other phylogeny-based methods. Rosel *et al.* (2017*a*) considered many of these metrics and found d_A (net nucleotide divergence) to be the most informative for cetaceans. If subspecies delimitation is based on mtDNA data and male-mediated gene flow is plausible, then differentiation or diagnosability should also be evaluated using nuDNA or other nongenetic data.

When evaluating both the population/subspecies and the subspecies/species boundaries, researchers should state why they chose to use particular methods and not other methods. They should scrutinize the case in question to determine whether it meets assumptions of the analytical methods employed and whether congruent inferences can be drawn from multiple methods (Carstens *et al.* 2013).

Category 5. Taxonomic Inference

Categories 1–4 present the lines of evidence including new analyses. The taxonomic inferences synthesize the lines of evidence into an argument for the taxonomic status warranted given the data. A clear statement should be made by the authors, such as, "Bryde's whales in the Gulf of Mexico *likely* belong to at least an undescribed subspecies of what is currently recognized as *Balaenoptera edeni*." This statement will cue the Committee on Taxonomy that authors are proposing a new taxon and the Committee can consider the merits of the argument that follows. A good example of synthesizing data for the taxonomic inference is Robineau *et al.* 2007. The best evaluation will be based not only on what evidence is available, but what evidence can be obtained with some estimate of time frame. For example, low sample sizes and gaps in sample distribution are expected for rare pelagic species like beaked whales. Thus, some lenience may be warranted with respect to having an ideal sample size and distribution.

In describing a new subspecies, the requirements of the International Code of Zoological Nomenclature (ICZN 1999) should be met. The name must be clearly indicated as new by use of a term such as "ssp. nov." or "nom. nov." A name-bearing type (a holotype specimen or expressly indicated syntypes) must be designated. The collection in which the name-bearing type is or will be deposited must be named. If these conditions are not met, the name is not considered "available" under the ICZN. It is also recommended in the Code that all the information that appears on the label(s) accompanying the specimen(s) should be published with the subspecies description.

The International Commission on Zoological Nomenclature maintains its Official Register of Zoological Nomenclature with Zoobank as its online version. By amendment of the Code in 2012, for electronic publication (in a journal published online, with suitable hardcopy archives) the name of the new subspecies must be registered in Zoobank before it is published (ICZN 2012), with evidence of such registration in the publication.

Quantitative and Qualitative Standards for Subspecies Delimitation

The primary purpose of quantitative standards is to promote consistency and transparency in taxonomic decisions. Reviewing recent publications that made taxonomic arguments primarily using genetic data revealed no standard approach to decide how much difference warrants a subspecies or species designation (Rosel *et al.* 2017*b*). A comparative approach was sometimes used for species (*e.g.*, Caballero *et al.* 2007), but for subspecies there are few named subspecies of cetaceans that can be used as a basis for comparisons. In the few subspecies comparisons that have been made, different metrics were calculated, which makes direct comparisons difficult.

The use of quantitative standards has progressed furthest in bird taxonomy, where rules were developed by the British Ornithological Union (Helbig *et al.* 2002). Tobias *et al.* (2010) capitalized on the wealth of data on accepted subspecies and species of birds to develop a point-based rule system that uses morphological and acoustic data, but not genetic data. Patten and Unitt (2002) made a persuasive case for using diagnosability as the standard for describing subspecies, but again relying on morphological data. Because morphology has been the primary basis for describing many thousands of subspecies and species of birds, this initial reliance on morphology is sensible. However, as stated earlier, morphological data are difficult to obtain for most cetaceans (Taylor *et al.* 2017).

Tobias *et al.* (2010) also had the benefit of a large number of bird species, allowing them to examine what types of data and weighting should be used to create a classification point system consistent with current taxonomy. A similar approach cannot be taken with cetaceans because there are an insufficient number of recognized taxa for a cross-validation approach. The best available data for developing quantitative standards for taxonomic studies of cetaceans come from Rosel *et al.* (2017*a*). Those authors carefully chose undisputed pairs of populations, subspecies, and species. They then used mtDNA control region sequences to calculate a range of metrics commonly used in taxonomic studies. Examination of their results reveals some fairly simple metric standards that result in relatively little overlap between populations, subspecies, and species (Fig. 2). They also chose cases that were likely to be difficult to classify using a single neutral marker, namely those with very low effective population sizes (killer whales and false killer whales, *Pseudorca crassidens*) and very high effective population sizes (pelagic dolphins). The metrics estimated by Rosel *et al.* 2017*a* form the backbone of our quantitative and qualitative standards. Comparisons in Rosel *et al*. (2017*a*) are limited to the control region of mtDNA because this is the only marker with sufficient empirical data currently available from cetaceans to make such comparisons. For that reason, the standards described below rely heavily on the use of mtDNA control region data.

Proposed Quantitative Standards

Of the metrics examined by Rosel *et al.* (2017*a*), the two that performed best at distinguishing pairs of populations, subspecies, and species were net nucleotide divergence (d_A) and percent diagnosable (PD) based on Random Forests (Archer *et al.* 2017). An algorithm for diagnosability, like Random Forests, is not always required, as cases with fixed basepair differences are 100% diagnosable. Based on these results (Rosel *et al.* 2017*a*), we chose $d_A > 0.02$ as our quantitative standard for the upper subspecies boundary and $d_A > 0.004$ for the lower subspecies boundary. These values are clearly defined and empirically based to minimize classification errors for this set

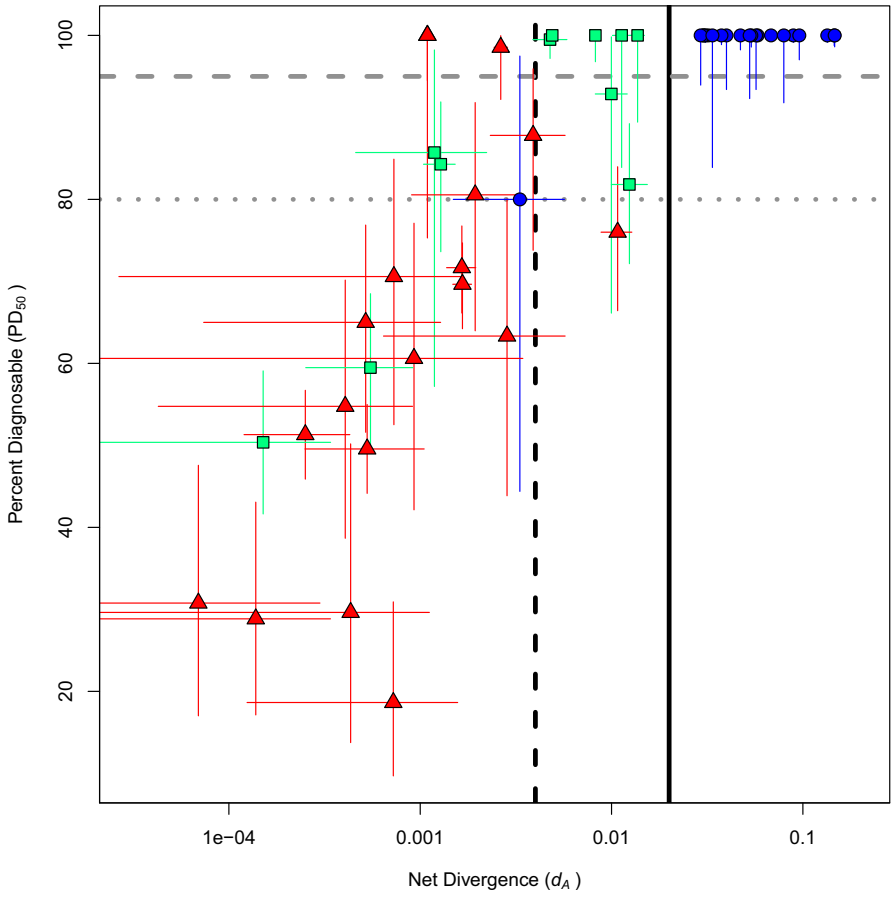

Figure 2. A comparison of the pairs of populations (red triangles), subspecies (green squares) and species (blue circles) estimated by Rosel *et al*. (2017*a*). Net nucleotide divergence (d_A) is shown on a natural log scale to better illustrate differences between the pairwise comparisons at low levels of divergence. Bars show the central 95th-pecentile of the estimate distributions. The solid vertical line at d_A = 0.020 delimits all but one species and correctly excludes all subspecies pairs. The vertical dashed line at d_A = 0.004 delimits all populations from the higher taxonomic levels and correctly delimits seven of eleven subspecies. The horizontal dashed lines are two potential thresholds for percent diagnosable (80% and 95%) that are discussed in the text.

of comparisons and this marker. The definition we chose for subspecies (Taylor *et al.* 2017) requires evidence of both high diagnosability (PD) *and* an independent evolutionary trajectory (d_A).

The choice of a value for the PD threshold is not obvious as either of two pathways could be taken: (1) choose a value consistent with PD used for morphology (95% as suggested by Patten and Unitt 2002) or (2) choose a value that minimizes errors for the empirical data set as was done for d_A. Using a threshold of 80% would decrease the number of underclassification errors for subspecies (Fig. 2) by two and would add

no overclassification errors using the best estimates for PD. The choice of how different groups must be to be considered subspecies is an arbitrary one. Examination of the magnitude of differentiation between different pairs of populations, subspecies, and species that were chosen as pairs where classification was noncontroversial reveals the level of mtDNA control region difference that is consistent with past practice. It could be argued that since past taxonomic designations were based on morphology, the level of diagnosability should be consistent with what has been used for morphology (95%). Alternatively, it could be argued that the PD for this marker should be chosen, just as d_A was, to minimize errors for agreed classifications, given this set of pairwise comparisons and this marker, and thus set at 80%.

We find merit in both of these perspectives, but have chosen to propose the use of 95% as the threshold value for this initial set of standards. The slight edge given to a 95% over an 80% threshold results from the perception that most practitioners would consider a 1 in 5 chance of misidentifying an individual to subspecies to be too high. On the other hand, there is a very small window between 95% PD for subspecies and near 100% PD for species and therefore the 95% bar could be set too high. The choice of a threshold value for PD should be the subject of ongoing debate that is increasingly informed and refined with additional data.

These quantitative standards are meant to provide a starting point for making an argument as to whether a given group of cetaceans merits subspecies classification. In making any particular argument, the researcher is encouraged to argue why even though the standards may not be met, the group of animals in question still merits subspecies status. For example, Figure 2 shows not only the best estimates used in the standards but also the 95% confidence limits. In some cases these limits span a large range including the standard threshold. An argument could be made that the uncertainty arises from small sample size, something that cannot be easily remedied, and that therefore there is only weak evidence that the case meets the standard. However, that weak evidence, when considered together with other lines of evidence, may tip the balance and make the argument convincing.

A somewhat surprising result from the pairwise comparisons was that using only a single metric and single female-inherited marker there were no cases of mistakenly classifying a population as a subspecies (an overclassification error). We intentionally included two cases (false killer whale populations in Hawaii compared with offshore populations and resident killer whale populations) where there is virtually no movement of females between groups and hence some strong potential to mistake social structure for evolutionary structure. In both cases, using percent diagnosable alone would have been misleading as both meet the standard for subspecies or species. However, requiring both percent diagnosable and d_A resulted in a correct classification. Nevertheless, it will be important to see whether any cases emerge where the d_A standard is met with a nontrivial level of male-mediated gene flow suggesting that a subspecies designation is not warranted.

Though these standards do an excellent job at the subspecies/species boundary, they result in several subspecies underclassifications (as well as classifying one accepted species as merely a population; Fig. 2). These misclassifications provide some important lessons and make it clear that with currently available data, there will be cases where the results are inconclusive. For example, two of the underclassification errors that would result from blind application of the standards involve subspecies of spinner (*Stenella longirostris* spp.) and spotted (*S. attenuata* spp.) dolphins in the Eastern Tropical Pacific (the two green subspecies pairs in the lower left of Fig. 2). These are taxa that have very large effective population sizes (and hence *a*

priori a high chance of low diagnosability) and are recently diverged (and hence *a priori* a low d_A). These highly abundant groups are also more likely to be poorly sampled as it could take hundreds or even thousands of samples to fully characterize haplotypic variability. The designations of these subspecies were based on morphology and distribution (Perrin 1975, 1990). Use of the guidelines provided here should aid authors in describing sample size in a way that enables the reader to better assess the conclusiveness of results.

To address some of the issues exemplified by such misclassifications, we combined our quantitative standards with some qualitative standards to create a single flow diagram that can be used when evaluating the strength of the taxonomic case. The combined quantitative and qualitative standards are depicted in Figure 3.

One qualitative judgment required in the flow diagram is "Is it plausible that N_e is large or that the split is recent?" As seen in Rosel *et al.* (2017*a*), many of the subspecies pairs did not have $d_A > 0.004$ and all but one of those had a percent diagnosable less than 95%. Many of these were populations with large (more than hundreds of thousands) N_e strata where drift would be expected to be very slow even in the absence of gene flow. Clearly, for such cases, the mtDNA control region alone is not sufficient to describe how separate the strata are. Archer *et al.* (2017) used simulations

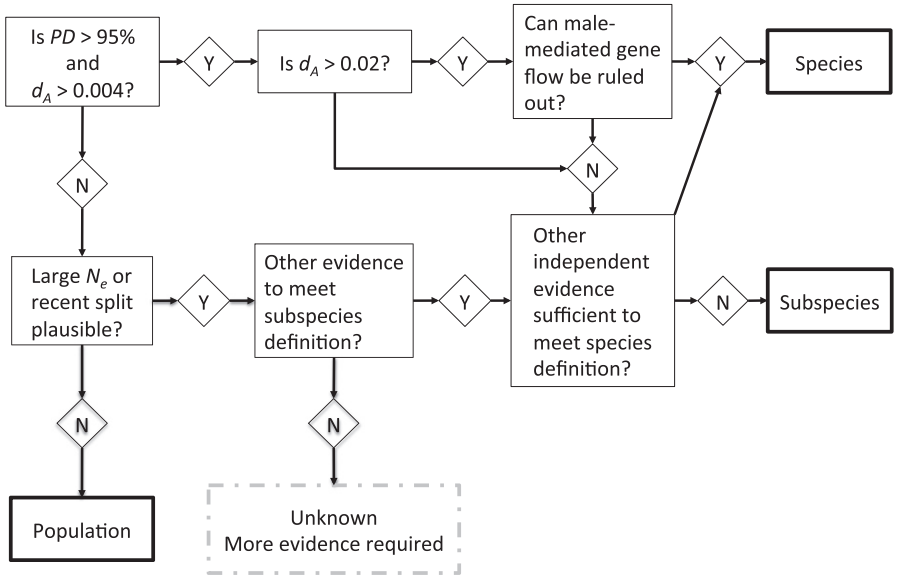

Figure 3. Flow diagram for subspecies delineation using combined quantitative and qualitative standards. The threshold values assume the user is evaluating a case relying on mtDNA control region data. Percent Diagnosable (PD) is the smallest strata-specific correct classification score in a given comparison (*e.g.*, PD_{50} in two-strata comparisons in Archer *et al.* 2017). The second box in the second row (other evidence to meet subspecies definition) allows for subspecies delineation when both conditions are not met using mtDNA. This box could be used either for the case when one condition is met and one unmet or when both just barely miss meeting the standards. For example, consider the case with PD < 95% and $d_A > 0.004$. Diagnosability could be achieved with morphological data or nuclear data that are sufficient for subspecies but not for full species.

to confirm that these high N_e cases are characterized by low diagnosability for this short sequence of mtDNA. Similarly, in the case of finless porpoises (*Neophocoena* spp.) the mtDNA control region was found to be uninformative, with the two putative species sharing a common haplotype but having 100% diagnosability using morphology and nuDNA (Wang *et al.* 2008). The authors suggested that speciation had been very recent. In both the tropical dolphin case and the finless porpoise case additional lines of evidence would be needed beyond mtDNA control region sequences to make a compelling taxonomic argument. Additional lines of evidence could include morphology, nuDNA, distribution, and behavior (including acoustics, mating system, and ecological specialization). We emphasize that when populations are large, neither a lack of evidence that $d_A > 0.004$ nor lack of evidence that the split was recent is sufficient to demonstrate that the unit in question is *not* a subspecies. All that can be concluded in such cases is that the control region data are inconclusive.

For species delimitation we are seeking multiple independent lines of evidence to address whether the unit in question is a separately evolving metapopulation lineage. The need for additional evidence to convince readers that the unit in question is a separately evolving lineage prompts the question: "Is there other independent evidence sufficient to support species delimitation?" There are three routes to reach this decision point: (1) the case is not warranted for subspecies based on mtDNA but is using other evidence, (2) the mtDNA evidence is sufficient for subspecies status but not for species status ($d_A < 0.02$), or (3) the case warrants species status based on mtDNA but male-mediated gene flow cannot be ruled out. In the latter case, sufficient evidence would consist of another single line of evidence that the putative taxon is a separately evolving metapopulation lineage. Cases 1 and 2 would require at least 2 such lines of evidence. Such evidence would preferably be nuDNA or morphology, but congruence of several other lines of evidence (acoustics, diet, distribution) may suffice if these nonheritable lines of evidence are put in the context of how they contribute to the evolution of separate species.

Cases of hybridization between well-differentiated cetacean species are known (*e.g.*, blue, *Balaenoptera musculus*, and fin whales), so a negligible amount of gene flow does not necessarily negate the validity of a species-level split. Nevertheless, the inferred amount of gene flow should be extremely small. The most common approach will be to use nuDNA to address this question. Diagnosably different morphology is another suitable line of evidence. Several more indirect lines of evidence (*e.g.*, a large distributional hiatus of unsuitable habitat together with acoustics or behavior) may also suffice.

Guidelines and Quantitative Standards in Practice

Examples of applying the guidelines and standards are given in Appendix S1, including detailed assessments of four of the papers reviewed by Rosel *et al.* (2017*b*). Those who conducted the assessments (six coauthors of the present paper) had difficulty determining whether the sampling was sufficient. This difficulty was in part because sampling adequacy is generally a challenging subject (and is the reason we supply more detail in Appendix S1) and in part due to the lack of sufficient detail in the published papers. Three categories (see Fig. 1) were notably poor across these examples: (1) description of life history (Category 1), (2) description of quality control in laboratory procedures (Category 3), and (3) descriptive statistics (Category 4). These shortcomings of the examples are surprising and would have been remedied by use of the guidelines. Both the life history information and the descriptive statistics

would have been helpful in evaluating the adequacy of sampling (Category 2), which is an area where the assessors struggled and had differing opinions. In spite of the difficulties, there was complete agreement between the "taxonomic" levels chosen by authors of the example papers (populations, subspecies, or species) and those chosen by the six assessors using the flow diagram.

Discussion

We recognize that given the uniqueness of evolutionary pathways, standards could be viewed as inimical to genuine understanding. In other words, no two paths in evolution are exactly alike and therefore no single, rigid set of standards should be expected to elucidate relationships accurately in every case. However, our intention is that the standards be used as starting points for taxonomic arguments rather than as rigid rules. For cases where the evidence far exceeds the standards, researchers can make easy arguments with little controversy. We anticipate that many if not most subspecies cases for cetaceans, particularly those with allopatric distributions in different ocean basins, will fit this profile. For borderline cases, the availability of standards should promote the development of good arguments, which in turn will make it possible to evaluate the standards' usefulness and robustness. We anticipate that the standards will evolve to accommodate new lines of evidence and new analytical tools and approaches.

One new type of evidence will be increased quantities of genetic data. Next-generation sequencing (NGS) methods allow for the rapid collection of high volumes of data (several orders of magnitude more than in the past) at a lower cost than traditional Sanger sequencing or microsatellite genotyping (Hancock-Hanser *et al.* 2013, Stolle and Moritz 2013). Complete mitochondrial genomes yield greater phylogenetic resolution than use of partial sequences in many cases (*e.g.*, Morin *et al.* 2010*a*, Vilstrup *et al.* 2011, Archer *et al.* 2013) and greater precision for estimates of divergence and proportion of fixed differences (Duchene *et al.* 2011). For nuclear loci, single nucleotide polymorphism (SNP) genotypes overcome some of the data quality and analysis issues of microsatellites. For instance they can be generated efficiently, from even poor-quality samples (Morin and McCarthy 2007; Seeb *et al.* 2009, 2011), using SNP-assays, or directly from NGS data (Lemmon *et al.* 2012, Narum *et al.* 2013). SNPs often prove to be equivalent or better than microsatellites for population genetic and phylogenetic analyses, especially when larger numbers of SNPs can be genotyped (Morin *et al.* 2009, 2012; Willing *et al.* 2012). Finally, with the ability to survey many genes comes the opportunity to detect genes that are under divergent selection between strata, in both nuclear and mitochondrial loci (Hohenlohe *et al.* 2010, Foote *et al.* 2011). These may provide direct evidence of the genetic drivers of divergence rather than simply reflecting the demographic history of populations as neutral markers do. We expect that the detection of genes under selection will become another line of evidence for subspecies or species delimitation, though more studies will have to be conducted to determine how to interpret such loci for determining taxonomic status. The ongoing "genomic revolution" will change both the types and amounts of data available for marine mammal population genetics and phylogenetics, and although we have tried to keep these changes in mind while compiling the guidelines, we fully expect that changes will need to be made to accommodate advances in molecular methods and markers.

Other lines of evidence will likely also improve our ability to delimit subspecies as they become available across a wide spectrum of populations, subspecies, and species. Variation in acoustic signals and in color patterns from digital photographs are particularly promising because large quantities of data can be obtained from wild populations using noninvasive methods. Acoustic signals have proven effective for addressing taxonomic questions regarding birds, mammals, amphibians, and insects (*e.g.*, Anderson *et al.* 2000, Gray and Cade 2000, Irwin *et al.* 2001, Ryan *et al.* 2007). To evaluate the validity of using acoustic signals for subspecies delimitation in cetaceans, studies are needed to determine which components of vocalizations are taxonomically informative and to examine the concordance of morphology, genetics, acoustics, and taxonomy across representative temporal and spatial scales (see May-Collado *et al.* 2007 for a phylogenetic analysis of tonal sound production in cetaceans; McDonald *et al.* 2006 and Baumann-Pickering *et al.* 2014 for studies of acoustic diagnosability in cetacean taxa; and Rendell *et al.* 2011 for studies of acoustic and genetic concordance).

We make three recommendations to help ensure quality and consistency as taxonomy integrates classical taxonomic approaches with genetic approaches. First, marine mammalogists should strive for a holistic approach to taxonomy. Future researchers will likely be experts in genomics and bioinformatics, but they should not become so by sacrificing knowledge of organismal biology: anatomy, behavior, and ecology. Working in natural history museum collections, stranding networks, and boat-based surveys are among the ways of acquiring such knowledge. In this spirit, morphology-oriented museums need to be open to alternative uses of their holdings, encouraging students and staff to blend morphology, genetics, isotopes, *etc*. into their taxonomic projects and allow "destructive" sampling for well-conceived taxonomic investigations. Institutions that offer traditional courses in mammalogy should continue to incorporate material on the latest approaches to molecular genetics, taxonomy, morphology, behavior, and ecology.

Second, a lot can be learned from the past; hence, a familiarity with the history of marine mammal taxonomy will benefit future researchers. An eye on the past will add context to the research of modern taxonomists while providing warnings of the recurring pitfalls.

Third, researchers need to know the process of taxonomy and nomenclature and participate in discussions of best practices. Stronger ties to systematists (regardless of the taxon) will greatly benefit marine mammal taxonomists. Involvement in a professional organization like the Society of Systematic Biology will bring exposure to novel approaches to taxonomy and nomenclature. Marine mammal taxonomists should actively and thoughtfully participate in the development of taxonomic standards and best practices for novel data types and philosophies. An example is the ongoing movement for a phylogenetic nomenclature (http://www.phylocode.org) that attempts to modernize the traditional biological nomenclature.

The guidelines and standards given here will become outdated and therefore will fall out of use if practitioners do not update them as new lines of evidence and new methods become available. Much of what we propose here is a template for a process aimed at operationalizing the interpretation of disparate lines of evidence into subspecies and species concepts. The guidelines should promote consistent and thorough taxonomic arguments, and the standards will allow progress toward correcting the large number of underclassification errors in a timely manner, something that is critical for marine mammal conservation.

Literature Cited

Amadon, D. 1949. The seventy-five percent rule for subspecies. Condor 51:250–258.

Anderson, M. J., L. Ambrose, S. K. Bearder, A. F. Dixson and S. Pullen. 2000. Intraspecific variation in the vocalizations and hand pad morphology of Southern lesser bush babies (*Galago moholi*): A comparison with *G. senegalensis*. International Journal of Primatology 21:537–555.

Andrews, K. R., W. F. Perrin, M. Oremus, L. Karczmarski, B. W. Bowen, J. B. Puritz and R. J. Toonen. 2013. The evolving males: Spinner dolphin (*Stenella longirostris*) ecotypes are divergent at Y chromosome but not mtDNA or autosomal markers. Molecular Ecology 22:2408–2423.

Archer, F. I., P. A. Morin, B. L. Hancock-Hanser, *et al*. 2013. Mitogenomic phylogenetics of fin whales (*Balaenoptera physalus* spp.): Genetic evidence for revision of subspecies. PLoS ONE 8:e63396.

Archer, F. I., K. K. Martien and B. L. Taylor. 2017. Diagnosability of mtDNA with Random Forests: Using sequence data to delimit subspecies. Marine Mammal Science 33(Special Issue):101–131.

Baumann-Pickering, S., M. A. Roch, R. L. Brownell, Jr., *et al*. 2014. Spatio-temporal patterns of beaked whale echolocation signals in the North Pacific. PLoS ONE 9(1):e86072.

Björklund, M., and S. Bergek. 2009. On the relationship between population differentiation and sampling effort: Is more always better? Oikos 118:1127–1129.

Bossart, J., and D. Prowell. 1998. Genetic estimates of population structure and gene flow: Limitatons, lessons and new directions. Trends in Ecology & Evolution 13:202–206.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg and A. R. Choudhury. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution 29:1917–1932.

Caballero, S., F. Trujillo, J. A. Vianna, *et al*. 2007. Taxonomic status of the genus *Sotalia*: Species level ranking for "tucuxi" (*Sotalia fluviatilis*) and "costero" (*Sotalia guianensis*) dolphins. Marine Mammal Science 23:358–386.

Carstens, B. C., T. A. Pelletier, N. M. Reid and J. D. Satler. 2013. How to fail at species delimitation. Molecular Ecology 22:4369–4383.

Cicero, C., and N. K. Johnson. 2006. Diagnosability of subspecies: Lessons from sage sparrow (*Amphispiza belli*) for analysis of geographic variation in birds. The Auk 123:266–274.

Cummings, M. P., M. C. Neel and K. L. Shaw. 2008. A genealogical approach to quantifying lineage divergence. Evolution 62:2411–2422.

de Queiroz, K. 2007. Species concepts and species delimitation. Systematic Biology 56:879–886.

Duchene, S., F. I. Archer, S. Caballero and P. A. Morin. 2011. Mitogenome phylogenetics: The impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. PLoS ONE 6(11):e27138.

Efron, B., and R. Thisted. 1976. Estimating the number of unseen species: How many words did Shakespeare know? Biometrika 63:435–447.

Foote, A. D., P. A. Morin, J. W. Durban, E. Willerslev, L. Orlando and M. T. P. Gilbert. 2011. Out of the Pacific and back again: Insights into the matrilineal history of Pacific killer whale ecotypes. PLoS ONE 6(9):e24980. doi:10.1371/journal.pone.0024980.

Frankham, R. 1995. Conservation genetics. Annual Review of Genetics 29:305–327.

Gray, D. A., and W. H. Cade. 2000. Sexual selection and speciation in field crickets. Proceedings of the National Academy of Science of the United States of America 97:14449–14454.

Greminger, M. P., M. Krützen, C. Schelling, A. Pienkowska-Schelling and P. Wandeler. 2010. The quest for Y-chromosomal markers – methodological strategies for mammalian non-model organisms. Molecular Ecology Resources 10:409–420.

Hale, M., T. Burg and T. Steeves. 2012. Sampling for microsatellite-based population genetic studes: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. PLoS ONE 7:e45170.

Hancock-Hanser, B. L., A. Frey, M. S. Leslie, P. H. Dutton, F. I. Archer and P. A. Morin. 2013. Targeted multiplex next-generation sequencing: Advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. Molecular Ecology Resources 13:254–268.

Hansen, M. M., E. E. Nielsen and K. L. D. Mensberg. 1997. The problem of sampling families rather than populations: Relatedness among individuals in samples of juvenile brown trout *Salmo trutta* L. Molecular Ecology 6:469–474.

Harris, R., and F. Allendorf. 1989. Genetically effective population size of large mammals: An assessment of estimators. Conservation Biology 3:181–191.

Hey, J. 2001. Genes, categories and species. Oxford University Press, Oxford, U.K..

Helbig, A. J., A. G. Know, D. T. Parkin, G. Sangster and M. Collinson. 2002. Guidelines for assigning species rank. Ibis 144:518–525.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics 6:e1000862.

ICZN (International Commission on Zoological Nomenclature). 1999. International Code of Zoological Nomenclature. Fourth edition. International Trust for Zoological Nomenclature, % The Natural History Museum, Cromwell Road, London SW7 5BD, U.K.

ICZN (International Commission on Zoological Nomenclature). 2012. Amendment of Articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. ZooKeys 219:1–10.

Irwin, D. E., S. Bensch and T. D. Price. 2001. Speciation in a ring. Nature 409:333–337.

Jackson, J., N. Patenaude, E. Carroll and C. Baker. 2008. How few whales were there after whaling? Inference from contemporary mtDNA diversity. Molecular Ecology 17:236–251.

Jiménez-Valverde, A., and J. Lobo. 2007. Determinants of local spider (*Araneidae* and *Thomisidae*) species richness on a regional scale: climate and altitude vs. habitat structure. Ecological Entomology 32:113–122.

Kalinowski, S. 2005. Do polymorphic loci require large sample sizes to estimate genetic distances? Heredity 94:33–36.

Kohn, M., E. York, D. Kamradt, B. Haught, R. Sauvajot and R. Wayne. 1999. Estimating population size by genotyping faeces. Proceedings of the Royal Society of London, Series B: Biological Sciences 266:657–663.

Landguth, E. L., B. C. Fedy, S. J. Oyler-McCance, *et al*. 2012. Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. Molecular Ecology Resources 12:276–284.

Leberg, P. 2002. Estimating allelic richness: Effects of sample size and bottlenecks. Molecular Ecology 11:2445–2449.

Leberg, P. 2005. Genetic approaches for estimating the effective size of populations. Journal of Wildlife Management 69:1385–1399.

Lemmon, A. R., S. A. Emme and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Systematic Biology 61:727–744.

Lim, G. S., M. Balke and R. Meier. 2012. Determining species boundaries in a world full of rarity: Singletons, species delimitation methods. Systematic Biology 61:165–169.

Lynch, M., and K. Ritland. 1999. Estimation of pairwise relatedness with molecular markers. Genetics 4:1753–1766.

Martien, K. K., S. J. Chivers, R. W. Baird, *et al*. 2014. Genetic differentiation of Hawaiian false killer whales (*Pseudorca crassidens*): Nuclear and mitochondrial markers provide insight into complex evolutionary history. Journal of Heredity 105:611–626.

Martien, K. K., M. S. Leslie, B. L. Taylor, *et al*. 2017. Analytical approaches to subspecies delimitation with genetic data. Marine Mammal Science 33(Special Issue):27–55.

May-Collado, L. J., I. Agnarsson and D. Wartzok. 2007. Phylogenetic review of tonal sound production in whales in relation to sociality. BMC Evolutionary Biology 7:136.

McDonald, M. A., S. L. Mesnick and J. A. Hildebrand. 2006. Biogeographic characterization of blue whale song worldwide: Using song to identify populations. Journal of Cetacean Research and Management 8:55–65.

Morin, P. A., and M. McCarthy. 2007. Highly accurate SNP genotyping from historical and low-quality samples. Molecular Ecology Notes 7:937–946.

Morin, P. A., K. K. Martien and B. L. Taylor. 2009. Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology Resources 9:66–73.

Morin, P. A., F. I. Archer, A. D. Foote, *et al*. 2010*a*. Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. Genome Research 20:908–916.

Morin, P. A., K. K. Martien, F. I. Archer, *et al*. 2010*b*. Applied conservation genetics and the need for quality control and reporting of genetic data used in fisheries and wildlife management. Journal of Heredity 101:1–10.

Morin, P. A., F. I. Archer, V. L. Pease, *et al*. 2012. Empirical comparison of single nucleotide polymorphisms and microsatellites for population and demographic analyses of bowhead whales. Endangered Species Research 19:129–147.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller and P. A. Hohenlohe. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology 22:2841–2847.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York, NY.

Patten, M. A. 2010. Null expectations in subspecies diagnosis. Ornithological Monographs 67:170–173.

Patten, M. A., and P. Unitt. 2002. Diagnosability versus mean differences of sage sparrow subspecies. The Auk 119:26–35.

Periera, L., and A. Amorim. 2004. How much more should the Y-STR haplotype reference database increase to reach a pragmatic saturation level? International Congress Series 1261:88–90.

Perrin, W. F. 1975. Variation of spotted and spinner porpoise in the eastern tropical Pacific and Hawaii. Bulletin of the Scripps Institution of Oceanography 21. xv + 206 pp.

Perrin, W. F. 1990. Subspecies of *Stenella longirostris* (Mammalia: Cetacea: Delphinidae). Proceedings of the Biological Society of Washington 103:453–463.

Pompanon, F., A. Bonin, E. Bellemain and P. Taberlet. 2005. Genotyping errors: Causes, consequences and solutions. Nature Reviews. Genetics 6:847–859.

Pruett, C., and K. Winker. 2008. The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. Journal of Avian Biology 39:252–256.

Queller, D. C., and K. F. Goodnight. 1989. Estimating relatedness using genetic markers. Evolution 43:258–275.

Reeves, R. R., W. F. Perrin, B. L. Taylor, C. S. Baker and S. L. Mesnick. 2004. Report of the workshop on the shortcomings of cetacean taxonomy in relation to needs of conservation and management, April 30–May 2, 2004 La Jolla, California. U.S. Department of Commerce, NOAA Technical Memorandum NOAA-TM-NMFS-SWFSC-363. 94 pp.

Rendell, L., S. L. Mesnick, M. L. Dalebout, J. Burtenshaw and H. Whitehead. 2011. Can genetic differences explain vocal dialect variation in sperm whales, *Physeter macrocephalus?* Behavior Genetics 42:332–343.

Robineau, D., R. Goodall, F. Pichler and C. S. Baker. 2007. Description of a new subspecies of Commerson's dolphin, *Cephalorhynchus commersonii* (Lacépède, 1804), inhabiting the coastal waters of the Kerguelen Islands. Mammalia 71:172–180.

Roff, D., and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: $X2$ and the problem of small samples. Molecular Biology and Evolution 6:539–545.

Rosel, P. E., B. L. Hancock-Hanser, F. I. Archer, *et al*. 2017*a*. Examining metrics and magnitudes of genetic differentiation used to delimit cetacean subspecies based on mitochondrial DNA control region sequences. Marine Mammal Science 33(Special Issue):76–100.

Rosel, P. E., B. L. Taylor, B. L. Hancock-Hanser, *et al*. 2017*b*. A review of genetic markers and analytical approaches that have been used for delimiting marine mammal subspecies and species. Marine Mammal Science 33(Special Issue):56–75.

Ryan, J. J., X. E. Bernal and A. S. Rand. 2007. Patterns of mating call preferences in túngara frogs, *Physalaemus pustulosus*. Journal of Evolutionary Biology 20:2235–2247.

Ryman, N., and S. Palm. 2006. POWSIM: A computer program for assessing statistical power when testing for genetic differentiation. Molecular Ecology Resources 6:600–602.

Ryman, N., S. Palm, C. André, *et al*. 2006. Power for detecting genetic divergence: Differences between statistical methods and marker loci. Molecular Ecology 15:2031–2045.

Schwartz, M., D. Tallmon and G. Luikart. 1998. Review of DNA-based census and effective population size estimators. Animal Conservation 1:293–299.

Seeb, J. E., C. E. Pascal, R. Ramakrishnan and L. W. Seeb. 2009. SNP genotyping by the 5′-nuclease reaction: Advances in high-throughput genotyping with nonmodel organisms. Methods in Molecular Biology 578:277–292.

Seeb, J. E., G. Carvalho, L. Hauser, K. Naish, S. Roberts and L. W. Seeb. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources 11(Supplement 1):1–8.

Soberón-M., J., and J. Llorente-B. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology 7:480–488.

Solow, A., and W. Smith. 2005. On estimating the number of species from the discovery record. Proceedings of the Royal Society of London., Series B: Biological Science 272:285–287.

Stolle, E., and R. F. Moritz. 2013. RESTseq - Efficient benchtop population genomics with RESTriction fragment SEQuencing. PLoS ONE 8:e63960.

Taberlet, P., S. Griffin, B. Goossens, *et al*. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24:3189–3194.

Tajima, F. 1995. Effect of non-random sampling on the estimaton of parameters in population genetics. Genetics Research 66:267–276.

Taylor, B. L. 2005. Identifying units to conserve. Pages 149–164 *in* J. E. Reynolds III, W. F. Perrin, R. R. Reeves, S. Montgomery and T. J. Ragen, eds. Marine mammal research: Conservation beyond crisis. The John Hopkins University Press, Baltimore, MD.

Taylor, B. L., K. Martien and P. Morin. 2010. Identifying units to conserve using genetic data. Pages 306–344 *in* I. L. Boyd, W. D. Bowen and S. J. Iverson, eds. Marine mammal ecology and conservation—a handbook of techniques. Oxford University Press, Oxford, U.K.

Taylor, B. L., W. F. Perrin, R. R. Reeves, *et al*. 2017. Why we should develop guidelines and quantitative standards for using genetic data to delimit subspecies for data-poor organisms like cetaceans. Marine Mammal Science 33(Special Issue):12–26.

Thompson, G., P. Withers, E. Pianka and S. Thompson. 2003. Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in Western Australia. Austral Ecology 28:361–383.

Tobias, J. A., N. Seddon, C. N. Spottiswoode, J. D. Pilgrim, L. D. C. Fishpool and N. J. Collar. 2010. Quantitative criteria for species delimitation. Ibis 152:724–746.

Vilstrup, J. T., S. Y. W. Ho, A. D. Foote, *et al*. 2011. Mitogenomic phylogenetic analyses of the Delphinidae with an emphasis on the Globicephalinae. BMC Evolutionary Biology 11:65.

Viricel, A., E. Pante, W. Dabin and B. Simon-Bouhet. 2014. Applicability of RAD-tag genotyping for interfamilial comparisons: Empirical data from two cetaceans. Molecular Ecology Resources 14:597–605.

Wang, J. 2002. An estimator for pairwise relatedness using molecular markers. Genetics 160:1203–1215.

Wang, J. Y., T. R. Frasier, S. C. Yang and B. N. White. 2008. Detecting recent speciation events: The case of the finless porpoise (genus *Neophocaena*). Heredity 101:145–155.

Waples, R. S. 1995. Evolutionarily significant units and the conservation of biological diversity under the Endangered Species Act. American Fisheries Society Symposium 17:8–27.
Waples, R. 1998. Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. Journal of Heredity 89:438–450.
Willing, E. M., C. Dreyer and C. Van Oosterhout. 2012. Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. PLoS ONE 7:e42649.
Woodley, M., D. Naish and H. Shanahan. 2008. How many extant pinniped species remain to be described? Historical Biology 20:225–235.
Zhang, A., L. He, R. Crozier, C. Muster and C.-D. Zhu. 2009. Estimating sample sizes for DNA barcoding. Molecular Phylogenetics and Evolution 2010:1035–1039.

Appendix

Sample Size Evaluation

In assessing the adequacy of sample size, a first step is to ensure that the samples have been randomly collected. In the context of most studies, this means that the collection should be random with respect to the underlying frequency distribution of haplotypes or alleles. Deviation from this ideal can occur when some samples are taken from family groups or closely related individuals. When this is the case, haplotypes or alleles may be overrepresented, thus biasing the frequency distributions. A strategy that can be employed to minimize this effect in a data set is to estimate relatedness (Queller and Goodnight 1989, Lynch and Ritland 1999, Wang 2002) between all pairs of individuals and remove one from each pair whose relatedness exceeds some threshold value. However, it is a given that even with truly random sampling, there is some probability, albeit small for strata of any appreciable abundance, that samples will be collected from closely related individuals. Thus, it is not good practice to conduct wholesale censoring without one or more additional rationales. In fact, removing related individuals can introduce a bias toward over-estimation of diversity if removal is indiscriminate. When screening related individuals, one should consider the manner in which they came to be sampled. For example, if two individuals were collected in the same sampling bout, or in relatively close time or space, then removal of one individual from the data set might be warranted as it could be argued that the two samples are not independent draws from the same distribution. In highly social organisms, like killer whales or pilot whales, care should be taken to ensure that the spatial and temporal scale of the sampling is large enough to ensure that the data set is representative of the stratum or strata under study rather than a potentially biased subset of social groups (Hansen *et al.* 1997). Examination of photo-id data and construction of social networks where data are sufficient can aid in evaluating how representative samples are relative to the question (Martien *et al.* 2014).

With sequence data, there is the secondary issue of ensuring that the data set contains a majority of the haplotypes present in the population, rather than just the most frequent ones. This is because methods like phylogenetic reconstruction and assessments of diagnosability are more influenced by haplotypic variability and similarities among haplotypes rather than haplotype frequencies (see Archer *et al.* 2017 and Martien *et al.* 2017 for a review). Evaluating that the complete haplotypic diversity is

represented in a data set is best accomplished by estimating the shape of the haplotypic "discovery curve." The concept of discovery curves draws heavily from the ecological literature on assessing biodiversity to produce measures of species richness (Efron and Thisted 1976, Soberón and Llorente 1993, Thompson *et al.* 2003, Solow and Smith 2005, Jiménez-Valverde and Lobo 2007, Woodley *et al.* 2008), and is frequently used in deciding when forensic databases are sufficiently near to being complete (Periera and Amorim 2004, Zhang *et al.* 2009). As more samples are collected, the expectation is that new haplotypes will become less frequent until all haplotypes in the population have been observed. Early in this process, the rate of discovery of new haplotypes will be relatively rapid, leveling off as more sampling effort yields fewer new haplotypes.

Because samples for genetic studies often are collected opportunistically and nonsystematically, estimating the shape of a given haplotype discovery curve is usually done by resampling the observed data for a series of smaller sample sizes, effectively generating multiple hypothetical sampling histories (Kohn *et al.* 1999, Jackson *et al.* 2008). Any of a series of asymptotic functions can then be fit to the replicate data, and from the fitted curve one can then estimate what percentage of the total haplotypes is represented in the current data, and thus how many more samples would be required to have sufficient representation. If haplotypic diversity is high, or one has simply not collected enough samples, the asymptote of the discovery curve may be difficult to estimate, as the increase in the number of haplotypes with increased sampling will be relatively linear.

As can be seen from above, it is difficult to prescribe a set value for what would be considered "acceptable" sample sizes, but the following rules of thumb may be helpful. For nuclear microsatellite data, empirical and simulation studies have shown an increase in variability around estimates of genetic diversity with fewer than 20 samples (Pruett and Winker 2008, Hale *et al.* 2012). It is also unlikely that with fewer samples one can either appropriately characterize a haplotype frequency distribution or assess where one is along a haplotypic discovery curve. Thus, 20 samples should be considered the barest minimum. Additionally, researchers should interpret cautiously inferences based on data sets where a large fraction (>30%) of the samples is of individuals with unique haplotypes (*i.e.*, haplotypes found in only one sample). This is a strong indication that they are relatively low on the discovery curve and do not have a fully representative sample of the haplotypic or nucleotide variability.

However, in some cases smaller sample sizes could be justified. For example, when there is a large distributional hiatus together with large divergence differences (large effect size) larger sample sizes may not be required to demonstrate discreteness (Kalinowski 2005). The large effect size could be due in part to relatively small effective population sizes, in which case, fewer than 20 individuals could still represent a relatively large percentage of a population (Hale *et al.* 2012). Nonetheless, in such cases the authors should also explain why obtaining additional samples is difficult. An example from Rosel *et al.* (2017*a*) is the comparison between southern right whales (*Eubalaena australis*, $n = 23$) and North Pacific right whales (*Eubalaena japonica*, $n = 7$), the latter now numbering only a few dozen animals in the eastern North Pacific where samples can be obtained. The sample size for the North Pacific is low, but it is unlikely to increase substantially (without greater international cooperation and sharing of samples), there is a large hiatus between the distributions of the southern and North Pacific species, and the effect size is large ($d_A = 0.03$). Another example is the comparison between the two subspecies of Commerson's dolphin (Robineau *et al.* 2007) where there are only 11 samples from the relatively inaccessible subspecies in

the Kerguelen Islands, but there are no unique haplotypes, there are shared fixed differences between two of the three Kerguelen haplotypes, there is high diagnosability, a moderate d_A, and a large distributional hiatus between this stratum and the larger South American coastal stratum ($n = 196$) (Rosel *et al.* 2017*a*).

Received: 23 September 2015
Accepted: 10 January 2017

Supporting Information

The following supporting information is available for this article online at http://onlinelibrary.wiley.com/doi/10.1111/mms.12411/suppinfo.

Appendix S1. Examples of applying the guidelines and standards.

Table S1. Application of the Guidelines to the argument presented by Caballero *et al.* (2007) for a new species within the genus *Sotalia*. X means inadequate; ✔ means adequate.

Table S2. Application of the guidelines for five examples where information is adequate (✔) or inadequate (X).

Table S3. Application of the standards in Figure 3 to the examples.