ELSEVIER

# The incidence of species-level paraphyly in animals: A re-assessment

CrossMark

Howard A. Ross *

School of Biological Sciences, Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand

A B S T R A C T

Species-level paraphyly was found by Funk and Omland (2003) to occur in 23% of animal species on the basis of a meta-analysis of published mitochondrial gene trees. Given the potential for bias in the selection of study organisms and the subsequent publication of their gene trees, I re-estimated the incidence of paraphyly in an independent dataset of publicly accessible COI sequences from the Barcode of Life Data System. Among 7368 animal species represented by two or more sequences, 19% were paraphyletic, slightly less than in the previous study. Rates within major taxonomic groups mirrored, but were slightly lower than, that observed earlier. Tests were made for operational factors that could inflate, and sampling effects that could underestimate, the rate of paraphyly. Overall the previous findings are confirmed. The observed incidence suggests that on average animal species diverged 2–3 $N_e$ generations in the past, far short of the predicted 5 $N_e$ generations required for complete monophyly.

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

The dynamic nature of species has formed the core of our understanding of Nature since Darwin's (1859) *Origin of Species*. Early forays into molecular phylogenetics used exemplar sequences to represent species because of cost and technical issues, and because of the lack of appreciation of the rate of speciation (Barraclough and Nee, 2001). The growing awareness of genetic variation within species translated, in the study of interspecies relationships, to an understanding that speciation involved a measurable genetic transition at marker genes (Avise et al., 1987; Maddison, 1997; Nichols, 2001). The magnitude of this transience was highlighted when Funk and Omland (2003) published the results of a survey of the incidence of paraphyly and polyphyly among animal species. This study revealed that a significant proportion (>20%) of species were still entangled with related species, that is they shared ancestral polymorphisms in mitochondrial (mtDNA) genomes.

Funk and Omland's (2003) review brought important insights to investigators; it has garnered hundreds of citations since its publication. The review revealed the pervasiveness of paraphyly; its existence had been beyond doubt. Gene tree paraphyly and polyphyly, whether caused by biological processes or operational issues, raised questions about the reliability of using mtDNA markers to infer species phylogenies and boundaries (reviewed by Grechko (2013)). Hybridization and introgression, ancient

retained polymorphisms, deviations from neutral evolution and clonal inheritance, heteroplasmy and recombination all compromise the reliability of the evolutionary history inferred from mtDNA markers. The absence of monophyly was also seen by some (Meyer and Paulay, 2005) as compromising the identification of species by single gene markers (Hebert et al., 2003; Ratnasingham and Hebert, 2007).

The study by Funk and Omland (2003) was effectively a meta-analysis of the phylogenetic literature. For the results of a meta-analysis to describe a biological phenomenon accurately, the published literature must provide a representative sample of the instances of that phenomenon. A non-representative literature can arise from publication bias as a result of geographical and taxonomic biases in the choice of study organisms, the reluctance of authors to submit all manuscripts for publication, the criteria used by editors and reviewers, and the prestige of the publication or author's institution (Cassey et al., 2004; Fanelli, 2010; Møller and Jennions, 2001). There are strong forces promoting the publication of papers that find positive, as opposed to negative, support for hypotheses or other large treatment effects. There is the possibility that the literature available to Funk and Omland was subtly biased, in a way that would inflate the estimate of paraphyletic and polyphyletic species. There may have been a greater inclination to study taxonomic groups in which relationships were uncertain in preference to those where non-genetic methods had identified clearly resolved and well-defined species relationships.

My goal is to re-estimate the incidence of species-level paraphyly using a completely different dataset of mtDNA sequences to test the validity of Funk and Omland's (2003) result. The data

* Fax: +64 (9) 367 7136.
  E-mail address: h.ross@auckland.ac.nz

are derived from the database of cytochrome *c* oxidase subunit 1 (COI) sequences that has been assembled in the Barcode of Life Data System (BOLD) for the purpose of making species identifications (Ratnasingham and Hebert, 2007). As the goal of the International Barcode of Life (iBOL, ibol.org) initiative is to establish a reference collection of sequences for all species, then it is unlikely that its inherent sampling biases will be the same as those of the studies forming the basis of Funk and Omland's (2003) study.

The distinctions among monophyly, paraphyly and polyphyly are based on the phylogenetic relationships of haplotypes, sampled from different species. A monophyletic species is one in which all of its haplotypes form a clade on a phylogenetic tree, and this clade contains only haplotypes of this species. All of the haplotypes of a paraphyletic species are in a single clade, but nested within it are haplotypes of one or more additional species. When haplotypes belonging to a species are scattered among other species on a phylogenetic tree then this represents narrow-sense polyphyly.

Funk and Omland (2003) chose to use the term polyphyly to include both paraphyly and narrow-sense polyphyly so as to avoid the awkwardness of non-monophyly and to maintain a continuity with the older systematic literature. Recent keyword searches of the Scopus (www.scopus.com) bibliographic database indicate that neither term (paraphyly or polyphyly) has gained ascendency. Nevertheless, here the term paraphyly will be used to represent all situations where a species is not monophyletic because it is frequently associated with mitochondrial gene trees. In fact, the running title of Funk and Omland's paper was "Species-level Paraphyly" (not Polyphyly).

Here I find that an independent assessment of paraphyly in animals returns similar patterns and frequencies of species-level paraphyly to that described earlier. I identify several cases where incorrect taxonomy or other operational issues appear to inflate the incidence of paraphyly. Conversely an analysis of the effect of sampling on apparent paraphyly indicates that additional reference haplotypes are likely to increase the observation of paraphyly.

## 2. Methods

### 2.1. Sequence selection

The COI sequences used in this study are derived from the publicly available data in the Barcode of Life Database (BOLD) on 1 July 2011. The GenBank and BOLD accessions are available in Supplementary data file 3. Each animal genus that was represented by two or more species, of which at least one species was represented by two or more sequences, was selected for inclusion. All sequences from these chosen genera were used, including species represented by singletons. This corresponded to 21,337 species

**Table 1**
Number of species (sequences) obtained for each phylum. Columns labelled "1" contain the number of species represented by a single sequence. Columns labelled "2+" contain the number of species (sequences) represented by two or more sequences. Species with Linnaean binomials are subdivided into those in genera lacking OTUs (Group 1) and those containing OTUs (Group 2). Group 3 comprised OTUs.

| Phylum | Total | Group 1 | | Group 2 | | Group 3 | |
|---|---|---|---|---|---|---|---|
| | | 2+ | 1 | 2+ | 1 | 2+ | 1 |
| Annelida | 228 (2359) | 55 (428) | 26 | 60 (1246) | 40 | 33 (605) | 14 |
| Arthropoda | 13,926 (112,942) | 3722 (38,581) | 1769 | 3281 (35,436) | 1700 | 2493 (35,495) | 961 |
| Brachiopoda | 2 (4) | 1 (3) | 1 | | | | |
| Bryozoa | 6 (18) | | | 2 (14) | 3 | | 1 |
| Chaetognatha | 7 (24) | 6 (23) | 1 | | | | |
| Chordata | 4709 (31,649) | 2911 (23,375) | 822 | 647 (6047) | 194 | 88 (1164) | 47 |
| Cnidaria | 121 (427) | 36 (179) | 20 | 41 (167) | 14 | 6 (43) | 4 |
| Echinodermata | 169 (2761) | 69 (1529) | 18 | 44 (1119) | 25 | 8 (65) | 5 |
| Mollusca | 2003 (18,466) | 522 (7355) | 250 | 708 (8917) | 417 | 80 (1501) | 26 |
| Nematoda | 60 (264) | 12 (46) | 8 | 15 (154) | 20 | 3 (34) | 2 |
| Nemertina | 15 (100) | 7 (92) | 8 | | | | |
| Onychophora | 6 (136) | 6 (136) | | | | | |
| Platyhelminthes | 53 (309) | 18 (223) | 8 | 14 (60) | 9 | 3 (9) | |
| Porifera | 7 (62) | | | 3 (9) | 2 | 2 (51) | |
| Rotifera | 22 (765) | 3 (15) | | 11 (527) | 2 | 6 (221) | |
| Tardigrada | 4 (37) | | | 2 (34) | 1 | 1 (2) | |
| Grand total | 21,337 (170,323) | 7368 (71,985) | 2931 | 4828 (53,730) | 2427 | 2723 (38,190) | 1060 |