Molecular Phylogenetics and Evolution 66 (2013) 69–79

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Empirical evaluation of partitioning schemes for phylogenetic analyses of mitogenomic data: An avian case study

Alexis F.L.A. Powell *, F. Keith Barker, Scott M. Lanyon

*Department of Ecology, Evolution and Behavior, Bell Museum of Natural History, 100 Ecology Building, 1987 Upper Buford Circle, University of Minnesota, St. Paul, MN 55108, USA*

ARTICLE INFO

*Article history:*
Received 29 May 2012
Revised 8 September 2012
Accepted 8 September 2012
Available online 18 September 2012

*Keywords:*
Avian RNA structure
Dataset partitioning
Mitochondrial genome
Phylogenetic analysis

ABSTRACT

Whole mitochondrial genome sequences have been used in studies of animal phylogeny for two decades, and current technologies make them ever more available, but methods for their analysis are lagging and best practices have not been established. Most studies ignore variation in base composition and evolutionary rate within the mitogenome that can bias phylogenetic inference, or attempt to avoid it by excluding parts of the mitogenome from analysis. In contrast, partitioned analyses accommodate heterogeneity, without discarding data, by applying separate evolutionary models to differing portions of the mitogenome. To facilitate use of complete mitogenomic sequences in phylogenetics, we (1) suggest a set of categories for dividing mitogenomic datasets into subsets, (2) explore differences in evolutionary dynamics among those subsets, and (3) apply a method for combining data subsets with similar properties to produce effective and efficient partitioning schemes. We demonstrate these procedures with a case study, using the mitogenomes of species in the grackles and allies clade of New World blackbirds (Icteridae). We found that the most useful categories for partitioning were codon position, RNA secondary structure pairing, and the coding/noncoding distinction, and that a scheme with nine data groups outperformed all of the more complex alternatives (up to 44 data groups) that we tested. As hoped, we found that analyses using whole mitogenomic sequences yielded much better-resolved and more strongly-supported hypotheses of the phylogenetic history of that locus than did a conventional 2-kilobase sample (i.e. sequences of the cytochrome *b* and ND2 genes). Mitogenomes have much untapped potential for phylogenetics, especially of birds, a taxon for which they have been little exploited except in investigations of ordinal-level relationships.



## 1. Introduction

Mitochondrial genomes (mitogenomes) are an attractive source of data for molecular phylogenetic studies of animal taxa. Because of their rapid time to coalescence, relatively high substitution rates, and large size (~17,000 bp), mitogenomes are more likely than other loci to evolve in concert with, and harbor evidence of, the population histories of species (Moore, 1995). Moreover, their high copy number, haploidy, and lack of recombination make mitogenomes especially easy to obtain, sequence, and analyze (Avise, 1998; Berlin et al., 2004). Given their merits, we contend that mitochondrial DNA (mtDNA) sequences should be included as one marker among many (Fisher-Reid and Wiens, 2011) in coalescent-based "species tree" and other multilocus analyses, rather than being abandoned for use in phylogeny construction, as some have advocated (e.g. Ballard and Whitlock, 2004; Galtier et al., 2009; reviewed by Rubinoff and Holland (2005)). Even as technological advances reduce the cost and difficulty of sequencing large numbers of nuclear loci, so should there be a concomitant increase in the use of mitogenomes, as they too are more readily acquired, whether intentionally or as by-products of genomic sequencing (e.g. Nabholz et al., 2010). Consequently, we argue that the routine practice of utilizing only 1–2 kilobases of mtDNA sequence in phylogenetic analyses should be replaced by the use of whole mitogenomes so as to take full advantage of the potential resolving power of the locus, especially with groups of closely-related organisms in which genetic distances are small. Although mitogenomic data have great potential, standards for their rigorous and objective use in phylogenetic analyses are currently lacking.

Of particular relevance to developing best methods for phylogenetic analyses of mitogenomes is that they exhibit heterogeneity in base composition and evolutionary rates at various scales across the molecule (Anderson et al., 1982; Cummings et al., 1995), which suggests that such analyses should benefit from data partitioning (Yang, 1996; Nylander et al., 2004). Partitioning improves model fit by dividing alignments into relatively homogeneous sets of sites before selecting and optimizing a substitution model for each set independently. Nevertheless, data partitioning is not widely used with mitogenomes. To survey current practice, we examined 71

* Corresponding author. Fax: +1 612 624 6777.
*E-mail address:* alveypowell@yahoo.com (A.F.L.A. Powell).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.09.006

papers with phylogenies (Appendix S1-A), published in association with recent submissions of metazoan whole mitogenome sequences to Genbank, and found that only about a third employed a partitioning scheme. To further review practices of the researchers most likely to employ exemplary methods, we reviewed an additional 40 papers (Appendix S1-B), most published in the past 5 years, which we selected for their focus on recovering phylogeny (rather than describing novel mitogenomes). We found that while 78% used some sort of partitioning—still a remarkably low proportion in our view—there was little uniformity of approach. Generally, protein-coding sites were sorted by gene and/or by codon position, and RNA sites by type, gene, and/or transcript structure (e.g. stems versus loops), but other criteria (e.g. template strand, evolutionary rate) were used in some cases. Of greater concern, discussion of partitioning options, consideration of their evaluation, or references to model studies were generally lacking. Consequently, we found much unexplained variation among partitioned analyses; for example, the number of data subsets utilized ranged from two to 42, with a mode of five groups.

One pervasive feature of mitogenomic studies is exclusion of data. Although some researchers have argued against that practice (Cameron et al., 2007; Kjer and Honeycutt, 2007), of the 111 studies that we reviewed, under 10% made use of all alignable sequence positions. Most studies did not justify data omission, but those that did gave reasons of intragenomic heterogeneity, substitutional saturation, and unreliability of the signal from certain portions of the mitogenome—all problems that can be ameliorated with data partitioning (Cameron et al., 2007; Kjer and Honeycutt, 2007). We were surprised to find no clear relationship between use of data partitioning and data exclusion: over half the studies that used full alignments used no partitioning, and many partitioned analyses discarded data, especially noncoding regions and RNA genes. In vertebrate studies, even those employing partitioning, it was common practice to exclude the ND6 gene (the only L-strand protein template) because of its markedly different base composition from other protein-coding genes.

In our view, partitioning should be favored over data exclusion as a strategy for dealing with heterogeneity within the mitogenome. However, the systematics community lacks a general strategy for selecting partitioning schemes for such datasets. Therefore, our primary goal for this study was to develop a procedure for partitioning mitogenomes that would make use of all alignable positions, accommodate among-site heterogeneity in base composition and evolutionary rates, and avoid overparameterization. In fulfilling this objective, we found inspiration in a method proposed by Li et al. (2008) for partitioning datasets composed of multiple nuclear protein-coding genes. The first steps in this approach include finely partitioning the data according to *a priori* categories, estimating evolutionary model parameter values for each subset, and grouping subsets with a clustering algorithm. The resulting clustering hierarchy is then used to define a nested set of alternative partitioning schemes, which are evaluated using tools derived from information theory. We believe that the procedure presented here, though more involved than current practices in mitogenomic phylogenetics, offers more explicit criteria for selecting a partitioning scheme and is efficient and flexible enough to serve as a model for future studies.

Another objective of this paper was to test the utility of whole mitogenome datasets for avian phylogenetics, particularly for resolving relationships among species within families that have undergone rapid diversification. Use of mitogenomes had a troubled early history in ornithology. The first studies, which examined interordinal relationships (Härlid et al., 1999; Mindell et al., 1999), generated results that were so at odds with other evidence and traditional views that they garnered considerable skepticism, both toward their findings and the general value of mitogenomic data (e.g. Johnson, 2001). Whereas mitogenomes are often used within other vertebrate classes (e.g. Teleostei), they have remained unpopular in avian phylogenetics, even though the spurious results of early studies were later explained as artifacts of inadequate evolutionary models and taxon sampling (Braun and Kimball, 2002; Slack et al., 2007). Moreover, avian mitogenomic studies have focused on higher-level relationships (e.g. Paton et al., 2002; Harrison et al., 2004; Gibb et al., 2007; Morgan-Richards et al., 2008; Pratt et al., 2009; Pacheco et al., 2011) even though such data likely have more promise for resolving recent divergences, which present fewer issues with signal saturation and inter-taxon base compositional heterogeneity. A new era of avian mitogenomic phylogenetics may now be emerging as evidenced by the very recent publication of the first species-level mitogenomic phylogenies of select clades—15 crane species (Krajewski et al., 2010), 19 Hawaiian honeycreepers and 28 related species (Lerner et al., 2011), and 9 swallows, (Cerasale et al., 2012). Incredibly, despite the fact that the order Passeriformes accounts for over half of extant bird species diversity, the mitogenomes of only ten passerines had been published prior to the last two studies. Our study adds to this roster by inferring relationships within a passerine subfamily, the grackles and allies clade of New World blackbirds (Icteridae), using newly sequenced mitogenomes of 23 species.

## 2. Methods

### 2.1. Taxon sampling

Most ingroup sampling (Table 1) was designed to infer relationships within a clade of New World blackbirds (Icteridae) that is endemic to South America ("group 1" of Johnson and Lanyon, 1999). We included one individual from 16 of 19 recognized species, thereby representing 12 of 13 genera (Gill and Donsker, 2011; Remsen et al., 2011). Missing were *Curaeus forbesi*, *Macroagelaius subalaris*, and the monotypic *Hypopyrrhus pyrohypogaster*. To further examine generic relationships within the grackles and allies—the subfamily to which the South American clade belongs (Lanyon and Omland, 1999)—we included one representative of each of the remaining six genera in that taxon. Recent molecular analyses of nine-primaried oscine phylogeny (Barker et al. submitted) found New World orioles sister to the grackles and allies, so we used *Icterus mesomelas* as an outgroup.

### 2.2. Laboratory procedures

We obtained *Agelaius phoeniceus* and *Molothrus aeneus* specimens as purified mitochondrial DNA extracts, prepared as described by Lansman et al. (1981; samples provided by J.C. Avise and D. Walker). For all other species, we extracted genomic DNA from frozen tissue samples using a DNeasy Tissue Kit (Qiagen, Valencia, CA) following manufacturer's instructions. We aimed first to amplify the complete mitochondrial genome of *Agelaius phoeniceus* in 11 overlapping fragments, each ~2000 bp in length, and then sequence the products with 21 complementary pairs of overlapping ~1000 bp reads. Some primers were unreliable or unsuccessful, and so were modified to achieve a better match. The resulting primer set was used for amplification and sequencing of the remaining blackbird taxa, but obtaining some fragments from some species required additional primers (Table S1).

All reactions were performed in 12.5-μL aqueous solution with working concentrations of reagents as follows: 0.4 μM for each primer, 0.4 μM dNTP, 3 mM $MgCl_2$, 1× Green GoTaq Flexi Buffer (Promega Corporation, Madison, WI), 0.025 units/μL GoTaq Hot Start DNA polymerase, and, when we suspected problems due to secondary structure formation in some rRNA-coding fragments, 1 M glycine betaine. Cycling parameters for initial PCR were usually