Obtaining mtDNA genomes from next-generation transcriptome sequencing: A case study on the basal Passerida (Aves: Passeriformes) phylogeny

Benoit Nabholza, Erich D. Jarvis, Hans Ellegren

Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Department of Neurobiology, Howard Hughes Medical Institute, Duke University Medical Center, Durham, USA

Article history:
Received 15 March 2010
Revised 8 June 2010
Accepted 15 June 2010
Available online 25 June 2010

Keywords:
Songbirds
Passeriformes
Passerida
Mitochondrial DNA
Paridae
Fatiroidea
Sylvoidea
Muscicapoidae
Molecular phylogeny
Next-generation sequencing

Abstract
Classically, the mitochondrial genome is sequenced by a series of amplicons using conserved PCR primers. Here we show how shot-gun transcriptome sequencing can be used to obtain the complete set of protein-coding genes from the mtDNA of four passerine bird species. With these sequences, we address the still unresolved basal Passerida relationships (Aves: Passeriformes). Our analysis suggests a new hypothesis for the basal relationships of Passerida, namely a clade grouping Sylvioidea and Passeroidea, with Paridae and Muscicapidae as successive sister groups to this clade. This study demonstrates the usefulness of next-generation sequencing transcriptome sequencing for obtaining new mtDNA genomes.

1. Introduction
Passeriformes is by far the most speciose avian order, comprising nearly 60% of all living bird species (Clements, 2007). Since the pioneering DNA–DNA hybridization work of Sibley and Ahlquist (1990), significant progress has been made in understanding the relationships among passerine birds (Ericson et al., 2002a,b; Barker et al., 2002, 2004; Ericson and Johansson, 2003; Johansson et al., 2008). For example, a consensus has emerged concerning the deepest relationships including the basal position of the enigmatic and species-poor New-Zealand wren family (Acanthisittidae) and the early dichotomy between Oscine (or songbirds) and Suboscine passerines (Barker et al., 2002; Ericson et al., 2002b). Within songbirds, several groups of Australo-papuan affinities such as lyrebirds (Menuridae) or bowerbirds (Ptilonorhynchidae), previously placed within the Corvoidea (i.e., crows and allied, sensu Sibley and Ahlquist, 1990), have a basal position within the Oscine clade (Barker et al., 2002, 2004). Finally, the huge clade of Passerida appears to be monophyletic with rockfowls (Picathartidae) and Australasian robins (Petroicidae) being sister groups to this clade (Ericson et al., 2002a,b; Ericson and Johansson, 2003; Barker et al., 2004).

Contrasting with these robust basal relationships, several more recent relationships are still ambiguous. Amongst the major uncertainties, the basal relationships within Passerida seem to be some of the most difficult. The use of several combinations of nuclear markers has lead to weakly supported alternative topologies (summarized in Fig. 1A–C; Barker et al., 2004; Treplin et al., 2008; Johansson et al., 2008). Johansson et al. (2008) reviewed the current knowledge of the early relationships among Passerida and concluded that they are best depicted by a polytomy of nine clades: Bombycillidae, African Hyliotas, Muscicapoidae, Certhioidae, Sylvoidea, Paridae (including Parinae and Remizinae), Regulidae, Stenostiridae, and Passeroidea. Recently, complete mitochondrial genomes (mtDNA) have been used to infer basal Neoavian relationships with various degree of success and in some disagreement with comprehensive studies using nuclear markers (e.g., Hackett et al. (2008) versus Pratt et al. (2009)). For the shallower divergences, the mitochondrial genome has been, and remains to be, a valuable marker. The standard method of retrieving new mtDNA genomes has been to use conserved PCR primers to amplify a suite of overlapping fragments that correspond to the whole mtDNA molecule, which are then sequenced and assembled (e.g., Slack et al., 2006). Commonly, long-
range PCR is first employed to generate templates for subsequent amplification of shorter fragments suitably sized for DNA sequencing. This two-step amplification procedure is thought to select for amplification products of true mitochondrial origin, as nuclear copies of mtDNA (numts) that have been transposed to the genomic DNA are usually short.

DNA sequencing is now gradually moving from Sanger-based technology, coupled with capillary electrophoresis, to next-generation sequencing (NGS) approaches based on sequencing-by-synthesis, or other concepts, coupled with massively parallel processing of templates (Meyer et al., 2008). While Sanger technology may yield up to 100 kb of sequence data per run on standard sequencing instruments, the new high-throughput platforms yield several hundreds of Mb up to several tens of Gb of DNA sequence per run (although run times are typically much longer in the latter than in the former approach). Recently, NGS approaches for retrieving mtDNA genomes based on overlapping amplicons – obtained through long-PCR from genomic DNA – have been reported (Jex et al., 2008).

The exceptional amount of DNA sequence data generated by NGS methodology should imply realistic possibilities to obtain more or less complete mtDNA genomes from shot-gun sequencing without prior amplification of specific regions of the mtDNA molecule. In this study we evaluated the feasibility of this approach by analyzing brain transcriptome sequences from four passerine species obtained through Roche 454 sequencing as recently reported by Künstner et al. (2010). We then used these data together with 11 previously published avian mtDNA genomes to study Passerida relationships.

2. Materials and methods

2.1. Passerine mitochondrial data set

We built a mitochondrial DNA data set composed of all 12 heavy-strand protein-coding genes: COX1–3, ND1–5, ND4L, ATP6, ATP8, and CYTB. As in prior studies (Paton et al., 2002; Slack et al., 2006), the NADH dehydrogenase 6 (ND6) gene was excluded because it is encoded on the opposite strand and has a very different base composition compared to the other mitochondrial protein-coding genes. We focused on protein-coding genes because the RNA isolation method used in Künstner et al. (2010) was designed to select for transcripts with terminal poly-A (see below). We first downloaded all complete passerine mitochondrial genomes available (May 2010) in GENBANK from NCBI (http://www.ncbi.nlm.nih.gov/), including gray-headed broadbill (Smithornis sharpei, NC_000879), fuscous flycatcher (Cnemotriccus fuscatus, NC_007975), superb lyrebird (Menura novaehollandiae, NC_000880), rook (Corvus frugilegus, NC_002069), eastern-orphean warbler (Sylvia crassirostris, NC_010229), village indigobird (Vidua chalybeata, NC_007883), zebra finch (Taeniopygia guttata, NC_007897), Taiwan (Syan’s) Bulbul (Pycnonotus taivanus, NC_013483), blackcap (Sylvia atricapilla, NC_010228), and Eurasian reed warbler (Acrocephalus scirpaceus, NC_011027). We also downloaded the near complete mitochondrial genome of the rifleman (Acant thisilla chloris, AY325307), considered to belong to the sister group of all other passerines (Acant thisitidae, New-Zealand wrens) (Barker et al., 2002, 2004; Ericson et al., 2002b). We also performed an analysis using the mitochondrial genomes of two non-passerine species, the great roadrunner (Geococcyx californianus, NC_017111) and the kelp gull (Larus dominicanus, NC_007006).

Four passerine species were subject to brain transcriptome sequencing: golden-collared manakin (Manacus vitellinus, Pipridae, SRR029477–78, available in Sequence Read Archive, http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi), American crow (Corvus brachyrhynchos, Corvidae, SRR029463–64), blue tit (Cyanistes [Parus] caeruleus, Paridae, SRR029162), and pied flycatcher (Ficedula hypoleuca, Muscicapidae, SRR029159–61). The source of brain samples, methods for RNA isolation, cDNA preparation and next-generation sequencing are given in Künstner et al. (2010). For each species, the reads, which averaged 250 bp in length, were de novo assembled into contigs using the NEWBLER software (version 2.0) distributed with the ROCHE 454-instrument. Then we chose a closely related species for which the complete mitochondrial genome was available and performed a reciprocal best-hit BLAST analysis between contigs from the transcriptome data and the reference species, using BLATZ (Schwartz et al., 2003). We used the rook as a reference for the American crow, the eastern-orphean warbler for the pied flycatcher and the blue tit, and the fuscous flycatcher for the golden-collared manakin. We removed the last 10 bp of ATP8 and the last 7 bp of ND4L that overlap with ATP6 and ND4, respectively. Amino acid sequence alignments were performed using the MAFFT software with high accuracy parameters (Katoh et al., 2002) and verified by visual inspection.

2.2. Phylogenetic analyses

We used both maximum likelihood (ML) and a Bayesian method to estimate the topology of the passerine tree. We performed the ML analyses with PHYML (version 3.0, Guindon and Gascuel, 2003).
We found that the expression level of individual mitochondrial genes varied within an order of magnitude; for the American crow they ranged from 12 to 104 reads per base per gene (excluding ND4L which was not detected), for the pied flycatcher from 19 to 64 (excluding ATP8) and for the golden-collared manakin from 9 to 132. There was no correlation in levels of gene expression among these species (American crow–pied flycatcher, $r = 0.254$, $p = 0.42$; American crow–golden-collared manakin, $r = 0.057$, $p = 0.86$; pied flycatcher–golden-collared manakin, $r = 0.55$, $p = 0.064$). The blue tit was an outlier relative to the other species; three genes were not detected at all (ATP8, ND1, and ND4L) and coverage for the other genes varied extensively, from 9 reads per base (ND3) to 3080 (ATP6). Such variation indicates that although all 12 genes are thought to be transcribed from one common heavy strand promoter (Scarpulla, 2008), their expression levels differ, perhaps due to either technical, species–specific, or brain state differences in the expression of mitochondrial genes controlled by other DNA regulatory elements.

### 3.2. Phylogenetic analyses

In addition to the four new species for which we retrieved mtDNA genome sequence data by transcriptome sequencing, complete mtDNA data was obtained from GenBank for another 11 passerines (see Section 2). After alignment, this resulted in a matrix of 15 species and a maximum of 10,830 bp across the coding sequences of the 12 genes. We computed the GC content at the third codon position (GC3) for all the species. GC3 varies significantly across species, from 38% in the fuscous flycatcher to 55% in the blue tit.

### Table 2

Statistical supports obtain for the interesting nodes (as labeled in Fig. 2).

<table>
<thead>
<tr>
<th>Node No.</th>
<th>PHYML</th>
<th>GARLI</th>
<th>PHYML with partitioning</th>
<th>Bootstrap support (ML analyses)</th>
<th>Posterior probability (Bayesian analyses)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Genes partitioning</td>
</tr>
<tr>
<td>1</td>
<td>85.6</td>
<td>74.1</td>
<td>76.4</td>
<td>87.5</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>97.8</td>
<td>99.8</td>
<td>100</td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>60.6</td>
<td>60.6</td>
<td>73.9</td>
<td>90.0</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>42.8</td>
<td>38.5</td>
<td>36.1</td>
<td>0.26</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>99.6</td>
<td>100</td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>99.0</td>
<td>91.3</td>
<td>100</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Table 1

Size and coverage of the sequences obtained through 454 transcriptome sequencing. Coverage is computed by mapping the reads onto the assembled gene-specific contigs in each species and quantifying the mean number of reads per base per gene.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Total size (Mean coverage)</th>
<th>Blue tit (Mean coverage)</th>
<th>Pied flycatcher (Mean coverage)</th>
<th>Golden-collared manakin (Mean coverage)</th>
<th>American crow (Mean coverage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP8</td>
<td>681 (270)</td>
<td>156 (0)</td>
<td>1559 (25)</td>
<td>1464 (55)</td>
<td>156 (26)</td>
</tr>
<tr>
<td>ATP8</td>
<td>156 (0)</td>
<td>156 (26)</td>
<td>681 (74)</td>
<td>681 (104)</td>
<td>144 (12)</td>
</tr>
<tr>
<td>COX1</td>
<td>1548 (1448)</td>
<td>1472 (478)</td>
<td>1539 (25)</td>
<td>1464 (55)</td>
<td>1464 (93)</td>
</tr>
<tr>
<td>COX2</td>
<td>681 (672)</td>
<td>675 (31)</td>
<td>681 (74)</td>
<td>681 (104)</td>
<td>762 (98)</td>
</tr>
<tr>
<td>COX3</td>
<td>783 (783)</td>
<td>756 (24)</td>
<td>783 (99)</td>
<td>762 (98)</td>
<td>1134 (92)</td>
</tr>
<tr>
<td>CVTB</td>
<td>1140 (930)</td>
<td>1107 (29)</td>
<td>1119 (43)</td>
<td>1134 (92)</td>
<td>1334 (92)</td>
</tr>
<tr>
<td>ND1</td>
<td>975 (0)</td>
<td>927 (19)</td>
<td>909 (132)</td>
<td>480 (93)</td>
<td>480 (93)</td>
</tr>
<tr>
<td>ND2</td>
<td>1038 (1017)</td>
<td>1026 (21)</td>
<td>1038 (68)</td>
<td>1035 (72)</td>
<td>1368 (39)</td>
</tr>
<tr>
<td>ND3</td>
<td>348 (348)</td>
<td>345 (64)</td>
<td>348 (108)</td>
<td>348 (44)</td>
<td>348 (44)</td>
</tr>
<tr>
<td>ND4</td>
<td>1377 (851)</td>
<td>1377 (27)</td>
<td>1362 (53)</td>
<td>1368 (39)</td>
<td>1368 (39)</td>
</tr>
<tr>
<td>ND4L</td>
<td>288 (0)</td>
<td>273 (33)</td>
<td>283 (9)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ND5</td>
<td>1815 (889)</td>
<td>192 (1794)</td>
<td>1692 (43)</td>
<td>1800 (45)</td>
<td>1800 (45)</td>
</tr>
<tr>
<td>Total sequence size</td>
<td>10,830 (0)</td>
<td>7208 (34)</td>
<td>10,461 (3)</td>
<td>10,428 (4)</td>
<td>9888 (9)</td>
</tr>
<tr>
<td>Mean coverage per gene (SD)</td>
<td>1130 (1007)</td>
<td>30 (13)</td>
<td>63 (36)</td>
<td>71 (30)</td>
<td>71 (30)</td>
</tr>
</tbody>
</table>
blue tit (52% for the European reed warbler). This variation could be problematic for phylogenetic reconstruction if a substitution model assuming constant GC content is used. RY coding provides an empirical test to the influence of GC content on phylogenetic reconstruction (Delsuc et al., 2003).

All the phylogenetic analyses performed (ML and Bayesian analyses with and without RY-recoding of the third codon position) provided the same topology with most of the nodes obtaining systematically good statistical support (Fig. 2 and Table 2). The mitochondrial dataset recovered without ambiguity the basal Oscine–Suboscine split. Within Suboscines, the golden-collared manakin was grouped with the fuscous flycatcher (Cnemotriccus), reflecting the well-known Old World–New World Suboscine split (Barker et al., 2002, 2004; Ericson et al., 2002a,b). Within Oscines, we obtained strong statistical support for the basal position of the lyrebird (Menura) (Ericson et al., 2002a).

The extremely short internodes at the base of the Passerida group (Fig. 2) indicate a rapid radiation and probably explain why previous work have yielded conflicting results (e.g., Ericson and Johansson, 2003; Barker et al., 2004; Treplin et al., 2008; Johansson et al., 2008; summarized in Fig. 1A–C). The fast evolving mitochondrial genome could potentially provide some phylogenetic signal for these nodes. In this regard, our analysis suggests a new hypothesis for the Passerida relationship (summarized in Fig. 1D). Specifically, we obtain moderate to strong support for grouping the Sylvoidea (represented by Acrocephalidae, the Eurasian reed warbler, two Sylviidae, the blackcap and eastern-orphean warbler and one Pycnonotidae, the Taiwan bubul) and Passeroidea (represented by an Estrildidae, the zebra finch, and Viduidae, the village indigobird) (Fig. 2 and Table 2, node No. 3). Within Sylvioidae, we obtain support for grouping the Pycnonotidae and the Acrocephalidae (Fig. 2 and Table 2, node No. 1). Finally, the Paridae, represented by the blue tit, appeared to be an immediate sister group of this clade, followed by the Muscicapidae, represented by the pied flycatcher. However, the branching pattern of Paridae and Muscicapidae obtained a low bootstrap support (Fig. 2 and Table 2, node No. 4) and thus remains unresolved.

We performed two control analyses to check the robustness of our topology. Firstly, we assessed the influence of the outgroup on topology reconstruction. We added two non-passerine species to the data set, namely the great roadrunner (Geococcyx californianus) and the kelp gull (Larus dominicanus). We replicated the ML analysis using PHYML with RY coding and obtained a similar topology, except for that the blue tit and pied flycatcher position was swapped. The Passeroida–Sylvioida relationship obtained a comparable support (68%, Supplementary Fig. S1). Secondly, given that the blue tit had the highest proportion of missing data, we removed it from the alignment, and replicated the ML analysis using PHYML. The Sylvioida + Passeroida clade was confirmed with a higher bootstrap support than before (90% compared to 60–70% in the analysis including blue tit) (Table 2).

3.3. Nuclear copies of mtDNA (numts)

Numts can create problems for phylogenetic reconstruction due to their mistaken identity as mitochondrial sequences (Sorenson and Quinn, 1998). However, a previous study suggested that numts are rare in the avian genome (Pereira and Baker, 2004). Numts are generally considered pseudogenes, although some are transcribed and mutations in them can cause human disease (Hazkani-Covo et al., 2010). In theory, if numts are transcribed in birds, our shot-gun cDNA sequencing should generate numerous numt sequences. To test for the presence of numts in the shot-gun transcriptome data, we were in a few cases able to compare our assembled sequences with available mitochondrial gene sequences of the same species obtained by PCR amplification from genomic DNA: for blue tit COX1 (GQ481677) and CYTB (DQ474043), and for pied flycatcher COX1 (GQ481901) and ND2 (DQ146345). For the blue tit, this revealed in each case four mismatches out of 618 and 809 sites compared, respectively. For the pied flycatcher, we found no mismatch out of 692 sites for COX1 and two out of 1011 sites for ND2. For all comparisons, the amount of observed divergence (<1%) is within the expected level of polymorphism of bird species which is around 3% mean pairwise synonymous divergence (Berlin et al., 2007; Nabholz et al., 2009). These findings indicate that numts may have not significantly influenced our phylogenetic analyses.

---

Fig. 2. Maximum likelihood (ML) tree with a GTR + I + Gamma6 model. Numbers close to the nodes indicate statistical support as recorded in Table 2. All the other nodes are supported by bootstrap or posterior probabilities equal to 100%. Scale indicates substitution/site.
Finally, we performed phylogenetic reconstruction of all species for each of the 12 genes independently using PHYML with 200 bootstrap replications (Supplementary Fig. S2). Comparisons between the genes trees and the topology presented in Fig. 2 did not reveal any strongly supported incongruence (bootstrap >70%). The only exception was for CYTB, where the monophyly of Suboscines is not recovered. To further explore the congruence between the single genes and full matrix analyses, we applied the supertree method PHYSIC_JST (Scornavacca et al., 2008), considering only the node with the bootstrap support higher than 70% (PHYSIC_JST option -b 70). The supertree topology was identical to the one obtained with the full matrix analyses with Paridae/Muscicapidae node being the only node unresolved.

4. Conclusions

Our finding that more or less the complete protein-coding part of the mitochondrial genome can be obtained as a byproduct in NGS transcriptome sequencing is encouraging for further phylogenetic work because the amount of sequence data generated by NGS methodologies is expected to increase significantly in the near future. Here we found support for a novel hypothesis concerning basal relationships within Passerida in the form of a Passeroidea + Sialiaevia clade. Increased taxonomical sampling to include Bombolichididae, Hyliola, Certhioidae, Regulus or Stenostornidae would be necessary to further support or refute this hypothesis and would provide additional information on the relationships within the highly diversified Passerida group (>2000 species).

Acknowledgments

Computational work was performed at the Uppsala University UPPMAX platform for high-performance computing. H.E. acknowledges funding from the Swedish Research Council, The Knut and Alice Wallenberg Foundation and a European Research Council Advanced Investigator Grant. E.D.J. acknowledges funding from NIH and HHMI. The authors thank the editor and the three anonymous reviewers for their valuable suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2010.06.009.

References
