

ELECTRONIC APPENDIX

This is the Electronic Appendix to the article

African endemics span the tree of songbirds (passeri): molecular systematics of several evolutionary 'enigmas'

by

P. Beresford, F. K. Barker, P. G. Ryan & T. M. Crowe

Proc. R. Soc. B (doi:10.1098/rspb.2004.2997)

Electronic appendices are refereed with the text; however, no attempt is made to impose a uniform editorial style on the electronic appendices.

Electronic Appendix part A. List of passerine taxa included in this study, classified to family level following Sibley and Monroe (1990). Species and higher taxa endemic to Africa and Madagascar are highlighted in boldface text. GenBank accession numbers for RAG-1 and RAG-2, along with specimen data (vouchers where available, tissue or collector's numbers are in brackets), are listed in parentheses after each species new to this study. Sequence accessions and voucher data for other samples are given in Barker et al. (2004).

Suborder Tyranni

Acanthisittidae: *Acanthisitta chloris*; Pittidae: *Pitta sordida*; Eurylaimidae: ***Smithornis rufolateralis***, *Psarisomus dalhousiae*; Philepittidae: ***Philepitta castanea***, ***Neodrepanis coruscans***; Tyrannidae: *Tyrannus tyrannus*; Thamnophilidae: *Thamnophilus nigrocinereus*; Furnariidae: *Furnarius rufus*

Suborder Passeri: Parvorder Corvida

Superfamily Menuroidea Climacteridae: *Climacteris picumnus*; Menuridae: *Menura novaehollandiae*; Ptilonorhynchidae: *Ptilonorhynchus violaceus*

Superfamily Meliphagoidea Maluridae: *Malurus melanocephalus*; Meliphagidae: *Meliphaga analoga*; Pardalotidae: *Pardalotus striatus*

Superfamily Corvoidea Petroicidae: *Eugerygone rubra*, *Tregellasia leucops*; Irenidae: *Irena cyanogaster*, *Chloropsis cochinchinensis*; Orthonychidae: *Orthonyx spaldingii*; Pomatostomidae: *Pomatostomus isidorei*; Laniidae: *Lanius excubitor*, ***Corvinella corvina***; Vireonidae: *Vireo philadelphia*; Corvidae: *Ptilorrhoa caerulescens*, *Pachycephala soror*, *Corvus corone*, *Cnemophilus loriae*, *Paradisaea raggiana*, *Cracticus quoyi*, *Artamus leucorhynchus*, ***Oriolus larvatus***, *Coracina lineata*, ***Dicrurus adsimilis***, ***Trochocercus [Elminia] nigromitratus***, *Monarcha axillaris*, *Aegithina tiphia*, ***Lanioturdus torquatus*** (AY799819, AY799838, PFI [Lt2.2 03/06/90]), ***Dryoscopus cubla***, ***Telophorus dohertyi***, ***Prionops plumatus***, ***Batis mixta***, ***Vanga curvirostris***; Callaeatidae: *Philesturnus [Creadion] carunculatus*

Suborder Passeri: Parvorder incertae sedis

Picathartidae: Picathartes gymnocephalus, Chaetops frenatus

Suborder Passeri: Parvorder Passerida

Superfamily Muscicapoidea Bombycillidae: *Bombycilla garrulus*; Cinclidae: *Cinclus cinclus*; Muscicapidae: *Catharus ustulatus*, *Turdus falcklandii*, *Muscicapa ferruginea*, *Culicicapa ceylonensis*, ***Namibornis herero*** (AY799820, RAG2 not sequenced, PFI [F2-6]) ; Sturnidae: *Sturnus vulgaris*, *Mimus patagonicus*

Superfamily Sylvioidea Sittidae: *Sitta carolinensis*; Certhiidae: *Certhia familiaris*, *Troglodytes aedon*; Paridae: *Remiz pendulinus*, *Parus inornatus*; Aegithalidae: *Aegithalos iouschensis*; Hirundinidae: *Hirundo pyrrhonota*; Regulidae: *Regulus calendula*; Pycnonotidae: ***Pycnonotus barbatus***, ***Bleda syndactyla*** (AY319976, AY799827, AMNH [PRS2036]), ***Xanthomixis (Phyllastrephus) zosterops*** (AY319996, AY799843, FMNH 345750 [SMG3444]), ***Nicator chloris*** (AY319991, AY799841, AMNH [PRS2029]); Cisticolidae: ***Cisticola anonymus***, ***Prinia bairdii*** (AY319998, AY799845, FMNH 355824 [TPG0092]), ***Apalis goslingi*** (AY799812, AY799826, AMNH [RWD23727]), ***Eminia lepida*** (AY799817, AY799834, FMNH 355931 [DW4294]), ***Camaroptera brachyura*** (AY799816, AY799831, ZMUC 117773 [O1984]), ***Euryptila subcinnamomea*** (AY799818, AY799836, PFI [C. Cohen Es2002]); Zosteropidae: ***Zosterops senegalensis***; Sylviidae: *Cettia brunnifrons* (AY319977, AY799832, AMNH [JGG1027]), ***Bradypterus baboecala*** (AY799813, AY799828 [partial], PFI AF47871), ***Bradypterus barratti*** (AY799814, AY799829, ZMUC 118831 [O3042]), ***Bradypterus victorini*** (AY799815, AY799830, PFI 190), ***Thamnornis chloropetoides*** (AY320004, AY799850, FMNH 356700 [JCR41]), ***Achaetops pycnopygius*** (AY799811, AY799824, PFI [Ap2.2 03/06/90]), ***Sphenoeacus afer*** (AY799822, AY799847, PFI [PGR0802]), ***Acrocephalus newtoni*** (AY319972, AY799825, FMNH 384719 [SMG8622]), ***Stenostira scita*** (AY799823, AY799848, PFI X73884 South Africa), *Orthotomus sutorius* (AY319992, AY799842, AMNH [PRS2473]), ***Sylvietta denti*** (AY320003, AY799849, AMNH [PRS2022]), ***Macrosphenus flavicans*** (AY319987, AY799839, AMNH [PRS2002]), ***Hylia prasina*** (AY319984, AY799837, AMNH [PRS2000]), *Phylloscopus collybita* (AY319997, AY799844, AMNH [PRS1935]), *Megalurus palustris* (AY319988, AY799840, FMNH 358387 [DW4528]), *Cincloramphus*

mathewsi (AY319978, AY799833, UWBM 57635), ***Schoenicola brevirostris*** (AY799821, AY799846, PFI [P. Leonard 4/1/02]), *Garrulax milleti*, ***Modulatrix stictigula***, ***Arcanator orostruthus***, *Yuhina nigrimenta* (AY320006, AY799851, AMNH [RTC596]), *Sylvia nana* (AY057033, AY443233)

Superfamily Passeroidea Alaudidae: *Alauda arvensis*, ***Eremopterix australis*** (AY319982, AY799835, PFI [*fide* P. Lloyd]); Nectariniidae: ***Promerops cafer***, *Dicaeum aeneum*, ***Nectarinia olivacea***; Melanocharitidae: *Melanocharis nigra*; Passeridae: *Passer montanus*, *Motacilla cinerea*, *Prunella collaris*, ***Ploceus cucullatus***; Fringillidae: *Fringilla montifringilla*, *Emberiza schoeniclus*

Electronic Appendix part B

MATERIALS AND METHODS.

(a) Sequence data collection

We gathered sequence data for sampled taxa from the nuclear genes RAG-1 and RAG-2. These genes are apparently derived from a transposon, acquired by the common ancestor of gnathostomes (Agrawal *et al.* 1998), and are essential to the functioning of the recombinatorial immune system (Oettinger *et al.* 1990, Agrawal *et al.* 1998). The two genes are closely linked in the *Gallus* genome (~10 kilobases apart, Carlson *et al.* 1991), and very similar in their evolutionary dynamics (see Results). Alone or in tandem these genes have been effective in resolving avian relationships across a broad hierarchical range (Groth and Barrowclough 1999; Barker *et al.*, 2004). We acquired sequences from approximately 3 kilobases of the single RAG-1 exon, and approximately 1 kilobase of the single RAG-2 exon, using primers and procedures previously described (Groth and Barrowclough 1999, Barker *et al.* 2002, Barker *et al.* 2004).

(b) Phylogenetic analyses

New sequences acquired during this study were aligned by eye with previously published alignments of these genes (EMBL Accessions ALIGN_000334, which replaces ALIGN_000206; ALIGN_000669; and ALIGN_000671). These alignments were analyzed using both parsimony (PAUP* v4.0b; Swofford 1998) and Bayesian (MrBayes v3.0a; Ronquist and

Huelsenbeck 2003) methods. Parsimony analysis of the concatenated RAG-1 and RAG-2 data was performed using equal weights for all characters and character state transitions, given evidence of relative evenness of base frequencies and lack of saturating substitutions (Groth and Barrowclough 1998). Prior to the combined analysis, significance of the ILD test (Farris *et al.* 1995) was evaluated, recognizing the two genes as partitions (100 replicates excluding uninformative characters, taxa added sequentially in random order, and branches swapped using nearest-neighbor interchange). Although interpretation of this test is problematic (Dolphin *et al.* 2000, Darlu and Lecointre 2002, Barker and Lutzoni 2002), we cautiously take non-significance of this test as *prima facie* evidence of congruence between two data sets. To identify the most-parsimonious trees, 50 heuristic searches were performed, with taxa added stepwise in random order, and branches swapped using tree bisection and reconnection (TBR; Swofford *et al.* 1996). Support for particular branches on the set of most-parsimonious trees was assessed by the non-parametric bootstrap (Felsenstein 1985), using two search strategies. First, we analyzed 100 pseudoreplicates with 5 heuristic searches each, adding taxa stepwise in random order, and swapping branches by TBR. Second, we analyzed 1000 pseudoreplicates with the same procedure, substituting subtree pruning and regrafting for TBR. Bayesian analysis of the data was performed using Metropolis-coupled Markov chain Monte Carlo (MC³), implemented in parallel with four incrementally heated chains (Altekar *et al.*, 2004). A general time-reversible model of sequence evolution was used (Yang 1994a), with a proportion of invariant sites, and rates at variable sites modelled using a discrete approximation to the Γ distribution (four rate categories, Yang 1994b; ff. Barker *et al.* 2002, Barker *et al.* 2004). Eight independent runs were performed, with random starting points, assuming uniform priors for all parameters save the nucleotide frequencies, which assumed a Dirichlet prior, and branch lengths, which assumed exponential priors. All runs allowed to proceed for 400,000 generations, sampling every 100. The approximate burn-in time for each run was determined graphically, and generations prior to that time discarded

(Huelsenbeck *et al.* 2002). The estimated posterior densities for parameter values and nodes were compared for consistency across runs, and comparable runs combined. An additional MC³ run was performed, allowing all parameters to vary between RAG-1 and RAG2, in order to derive parameter estimates for comparisons of evolutionary dynamics between the two genes.

(c) Calibration of molecular clock

Dates of divergences were estimated at several nodes based on the assumption that the New Zealand endemic *Acanthisitta* divergence from all other passeriforms was contemporaneous with the rifting of New Zealand from Antarctica approximately 82 Ma (Cracraft 2001; Barker *et al.* 2004). For the purposes of relaxed molecular clock analyses, a maximum likelihood tree was inferred from the molecular data using the methods of Guindon and Gascuel (2003), using a GTR+I+ Γ model of sequence evolution (see above). Although the robustness of this estimate was not evaluated via the bootstrap, the structure of the inferred tree was consistent with both the parsimony strict consensus and Bayesian majority rule consensus (not shown). Non-parametric rate smoothing (NPRS; Sanderson, 1997) was applied to the maximum likelihood tree and associated branch lengths in order to estimate the age of other nodes (see also Barker *et al.* 2004). Rate smoothing was performed using r8s (Sanderson 2002), invoking multiple random starting points and multiple perturbations per solution (5 of each). Sampling confidence intervals for inferred divergences were obtained by reanalysis of 100 bootstrap replicates of the complete dataset.

RESULTS

(a) Preliminary analyses

Complete sequences of both genes were obtained for all taxa except *Namibornis*, for which we lack RAG-2 sequence, and *Bradypterus baboecala*, for which we lack the 3' half of RAG-2. Alignment of these sequences with previously obtained sequences was straightforward. The RAG-1 sequence from *Namibornis* shared a one codon indel with that from *Muscicapa*, and the sequence from *Cettia* had an autapomorphic two codon deletion, both occurring at the 5' end of the gene. Also, the new RAG-2 sequence of

Eremopterix shared a single codon deletion at the 3' end of the gene with that obtained from *Alauda*. No other novel length variation was observed. The commonly-reported χ^2 contingency table tests for base composition stationarity were non-significant for all positions of both genes (RAG-1: $\chi^2=57.3$, $df=327$, $p>0.1$; RAG-2: $\chi^2=42.3$, $df=327$, $p>0.1$), or for third positions alone (RAG-1: $\chi^2=107.9$, $df=327$, $p>0.1$; RAG-2: $\chi^2=65.6$, $df=327$, $p>0.1$). However, this masked a deviation of *Sitta carolinensis* nucleotide frequencies at codon third positions, which is significant in taxon-specific goodness-of-fit tests for RAG-1 ($\chi^2=22.1$, $df=3$, $p<0.01$), and apparent if not significant for RAG-2 ($\chi^2=5.2$, $df=3$, $p>0.1$; both p-values adjusted for multiple comparisons). No other taxa exhibited similar deviations, and as the phylogenetic placement of *Sitta* was well supported (see below), and distant from the primary taxa of interest in this study, base composition is not considered further. The two genes had similar percentages of variable (49 and 57% for RAG-1 and RAG-2 respectively) and informative sites (37 and 41%), indicating similar overall rates of evolution, with RAG-2 slightly faster. Distances at the RAG-1 locus varied from 0.3-7.6% within oscines, to 9.5% between the basal passerine *Acanthisitta* and *Sitta*, and were similar for RAG-2 (0.1-8.4% within oscines, 10.3% between *Acanthisitta* and *Macrosphenus*). Finally, the ILD test we performed was not significant ($p=0.32$), consistent with similar evolutionary dynamics and underlying phylogenetic signal between the two data sets (Barker and Lutzoni 2002).

(b) *Phylogenetic analyses*

Bayesian analysis of the data yielded estimated nodal posterior probabilities which were completely consistent with the majority rule consensus of most-parsimonious trees. The estimates presented are based on eight MC³ runs of 400,000 generations, with the first 250,000 of each discarded as burn-in. With sampling of every 100th generation, this left 1500 trees per run, for a total of 12,000 sampled trees. The Bayesian analysis differed from the parsimony analysis only in significantly supporting four nodes which were not present in the parsimony consensus, but which were consistent with that consensus. Partitioned Bayesian analysis of RAG-1 and RAG-2 indicated that the two genes

are similar, though distinguishable, in evolutionary dynamics (Table 3). The nucleotide base composition of the two genes is very similar, although there is a small difference in the frequency of guanine residues, statistically significant as indicated by non-overlapping 95% credibility intervals. Likewise, several of the nucleotide substitution rate parameters (r_{ij}) differ significantly in value, as do the proportion of invariant sites (p_{iv}) and the relative rates of the two genes. However, all of these differences are small in magnitude, and in particular, the rank-order of nucleotide substitution parameters is nearly identical, with only the relative magnitude of r_{AC} and r_{GT} (two of the three least frequent substitution classes) differing between the two. Also reassuringly, the estimated nodal posterior probabilities from the partitioned analysis were indistinguishable from those derived from the eight unpartitioned analyses.

Electronic Appendix References

- Agrawal, A., Q. M. Eastman & D. G. Schatz. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394: 744-51.
- Altekar, G., S. Dwarkadas, J.P. Huelsenbeck & F. Ronquist. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20: 407-15.
- Barker, F.K., G.F. Barrowclough & J.G. Groth. 2002. A phylogenetic hypothesis for passerine birds: taxonomic and biogeographic implications of an analysis of nuclear DNA sequence data. *Proc. R. Soc. Lond. B* 269: 295-308.
- Barker, F. K. & F. M. Lutzoni. 2002. The utility of the incongruence length difference test. *Syst Biol* 51: 625-637.
- Barker, F.K., A. Cibois, P. Schikler, J. Feinstein & J. Cracraft. 2004. Phylogeny and diversification of the largest avian radiation. *Proc. Nat. Acad. Sci. USA* 101: 11040-11045.
- Carlson, L. M., M.A. Oettinger, D. G. Schatz, E. L. Masteller, E. A. Hurley, W. T. McCormack, D. Baltimore, & C. B. Thompson. 1991. Selective expression of RAG-2 in chicken B cells undergoing immunoglobulin gene conversion. *Cell* 64, 201-208.
- Darlu, P. & G. Lecointre. 2002. When does the incongruence length difference test fail? *Mol Biol Evol* 19: 432-437.
- Dolphin, K., R. Belshaw, C. D. Orme, & D. L. J. Quicke. 2000. Noise and incongruence: interpreting results of the incongruence length difference test. *Mol Phyl Evol* 17: 401-406.
- Farris, J. S., Källersjö, M., Kluge, A. G. & Bult, C. 1995. Constructing a significance test for incongruence. *Syst Biol* 44: 570-572.
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.
- Groth, J. G. & G. F. Barrowclough. 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. *Mol Phyl Evol.* 12: 115-123.
- Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52: 696-704.
- Huelsenbeck, J.P., B. Larget, R. E. Miller & F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol* 51: 673-88.
- Oettinger, M.A., D. G. Schatz, C. Gorka & D. Baltimore. 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* 248: 1517-23.
- Ronquist, F. & J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19:1572-4.

- Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14: 1218-1231.
- Sanderson, M.J. 2002. r8s, v1.50. University of California, Davis, CA: distributed by the author (<http://ginger.ucdavis.edu/r8s/>).
- Swofford, D. L. 1998. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-514 *in* D. M. Hillis, C. Moritz, and B. K. Mable (ed.), *Molecular Systematics*, 2nd edition. (Sinauer Associates: Sunderland, Massachusetts).
- Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. *J Mol Evol* 39: 105-111.
- Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39: 306-314.