Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species

AXEL KÜNSTNER,* JOCHEN B. W. WOLF,* NICLAS BACKSTRÖM,* OSCEOLA WHITNEY,‡ CHRISTOPHER N. BALAKRISHNAN,§ SCOTT V. EDWARDS,** DANIEL E. JANES,** BARNEY A. SCHLINGER,†† RICHARD K. WILSON,† ERICH D. JARVIS,‡ WESLEY C. WARREN† and HANS ELLEGREN*

*Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, 75236 Uppsala, Sweden, ‡Genome Sequencing Center, Washington University School of Medicine, 63110 St Louis, MO, USA, ‡Department of Neurobiology, Duke University Medical Center, 27710 Durham, NC, USA, §Institute of Genomic Biology, University of Illinois, 61801 Urbana, IL, USA, ‡Department of Biology, University of Mississippi, MS 38677, USA, **Department of Organismic and Evolutionary Biology, Harvard University, 02138 Cambridge, MA, USA, ††Department of Physiological Science, University of California Los Angeles, CA 90095, USA

Abstract

Next-generation sequencing technology provides an attractive means to obtain large-scale sequence data necessary for comparative genomic analysis. To analyse the patterns of mutation rate variation and selection intensity across the avian genome, we performed brain transcriptome sequencing using Roche 454 technology of 10 different non-model avian species. Contigs from de novo assemblies were aligned to the two available avian reference genomes, chicken and zebra finch. In total, we identified 6499 different genes across all 10 species, with ~1000 genes found in each full run per species. We found evidence for a higher mutation rate of the Z chromosome than of autosomes (male-biased mutation) and a negative correlation between the neutral substitution rate (dS) and chromosome size. Analyses of the mean dN/dS ratio (ω) of genes across chromosomes supported the Hill–Robertson effect (the effect of selection at linked loci) and point at stochastic problems with ω as an independent measure of selection. Overall, this study demonstrates the usefulness of next-generation sequencing for obtaining genomic resources for comparative genomic analysis of non-model organisms.

Keywords: Next generation sequencing, 454, Avian genomics, Male-mutation bias, Selection, Hill-Robertson effect

Received 10 August 2009; revision received 14 October 2009; accepted 19 October 2009

Introduction

A major goal in evolutionary biology is to identify and subsequently study those loci that contribute to key phenotypes involved in adaptation, reproduction and survival (Benfey & Mitchell-Olds 2008; Ellegren & Sheldon 2008; Stinchcombe & Hoekstra 2008). Not long ago, this seemed to be a far-fetched goal for biologists working with natural populations of non-model organisms.

Large-scale genomic analysis was mainly restricted to model organisms such as Drosophila, Arabidopsis, Saccharomyces, and domestic animals and plants. However, the last few years have seen a change, in that genomic tools have become integral parts of molecular ecology. To mention but a few examples, gene expression (transcriptome) profiling has helped to reveal how variation in gene regulation can be tied to phenotypic variation (Rockman & Kruglyak 2006), the use of genome-wide sets of genetic markers have greatly facilitated mapping of trait loci in pedigrees (linkage mapping) (Beraldi et al. 2007) or populations (association mapping) (Wood...
et al. 2008), and the focus on candidate genes has offered a direct link to phenotypes in natural populations (Abzhanov et al. 2004).

Ultimately, if we are to fully be able to dissect the relationship between genotypes and phenotypes, genome sequence information will be needed from ecologically relevant organisms. The recent introduction of ‘next-generation’ or massive parallel sequencing technologies (Rothberg & Leamon 2008) in molecular ecology (Vera et al. 2008) holds promise that this will eventually come about (Bonneaud et al. 2008; Ellegren 2008b; Hudson 2008). Importantly, the generation of large amounts of DNA sequence data from related species will allow comparative genomic approaches for the identification of trait loci, and this is particularly so with transcriptome sequencing (‘RNA-seq’; Wang et al. 2009). Transcriptome sequencing reflects the subset of genes from the genome that are functionally active in a selected tissue and species of interest.

Together with expression profiling, genetic mapping and candidate gene approaches, comparative genomics represents one of the main routes for dissecting the genetic basis of phenotypic variation (Ellegren & Sheldon 2008). One common application in comparative genomics is to analyse sequences of orthologous genes from two or more species for rates of divergence and to test whether this divergence deviates from rates expected under a neutral scenario (Ellegren 2008a). Accelerated divergence beyond neutral expectations would indicate evidence of adaptive evolution where positive selection has increased the fixation rate of beneficial alleles (Nielsen 2005; Wright & Andolfatto 2008). In contrast, sequence conservation beyond neutral expectations would indicate evidence of purifying selection due to functional constraints (Ponting 2008). Assessing these rates typically involves analysing the rates of accumulation of non-synonymous substitution \( (d_N) \) and synonymous substitution \( (d_S) \) in protein-coding sequence.

So far, there are only a few studies that have utilized genome or transcriptome sequence data from multiple species to make general interferences about natural selection across species. This includes analyses of whole-genome sequences from drosophilids (Begun et al. 2007; Clark et al. 2007) and mammals (Kosiol et al. 2008). The purpose of the present study was to use a next-generation sequencing approach for comparative genomic analyses of mutation and selection processes in avian genomes. Specifically, we performed brain transcriptome sequencing of 10 bird species using Roche 454 technology. This generated a total of more than 1 Gb of raw sequence data, yielding an assembly of 7.27 Mb that could be unambiguously identified as protein-coding sequences.

### Materials and methods

#### Bird species, cDNA library preparation, normalization and transcriptome sequencing

Ten bird species were included in the study (Table 1). They were chosen based on a combination of traits, representing ecologically well-studied species (blue tit, pied flycatcher and crow) and including a few divergent bird lineages (suboscine, hummingbirds, doves, ratites). For nine of the species (not including European crow, see below), RNA was extracted from adult brain tissue preserved in RNA later (QIAGEN). Whole brains were homogenized with a tissue ruptor (Promega or QIAGEN) and total RNA isolation was performed on aliquots of the homogenate following the manufacturer’s instructions for the RNeasy Kit (QIAGEN) or the SV Total RNA Isolation System (Promega). To determine RNA quality, RNA samples were run on 1% agarose gels or a RNA 6000 Nano LabChip using the 2100 Bioanalyzer (Agilent). RNA was prepared from between 1 and 10 individuals of each species, as a means for the identification of single nucleotide polymorphisms (data not shown). When multiple individuals were used, equal amounts of RNA from each individual, as measured by quantification with spectrophotometry (Nanodrop), were pooled prior to sequencing. Individual samples were not tagged.

Library construction and normalization was performed using a variation of the Clontech SMART system as described previously (Warren et al. 2008). The features of this library are that the synthesis starts from the 3’ end of the mRNA using an oligo dT primer, and the cDNA are expected to be full-length. The normalized cDNA was sequenced according to the standard 454-FLX library protocol (Roche). Sequences were processed to remove adaptors and short reads prior to assembly. Raw read data have been deposited on the NCBI short read archive. Accession numbers can be found in Table 1.

European crow (Corvus corone) sequencing was carried out at a different 454 FLX platform, and was based on non-normalized cDNA library (Wolf et al. 2010).

#### Sequence assembly, contig annotation and estimation of substitution rates

Reads from the different species were assembled de novo separately using the Newbler assembler (Roche). Reads that could not be assembled and thus remained as singletons were disregarded from further analyses, which thus were based on sequence contigs only. By excluding singletons we sought to reduce the impact of sequencing errors (Huse et al. 2007).
Table 1  Summary statistics of brain transcriptome sequencing in 10 bird species

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Ind*</th>
<th>Plates†</th>
<th>Bases (Mb)</th>
<th>Reads</th>
<th>Contigs</th>
<th>Coverage‡</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archilochus colubris</td>
<td>Ruby-throated hummingbird</td>
<td>1</td>
<td>1</td>
<td>106.1</td>
<td>496</td>
<td>627</td>
<td>15 204</td>
<td>11.43</td>
</tr>
<tr>
<td>Calypte anna</td>
<td>Anna’s hummingbird</td>
<td>1</td>
<td>1</td>
<td>96.0</td>
<td>464</td>
<td>910</td>
<td>16 355</td>
<td>10.65</td>
</tr>
<tr>
<td>Corvus brachyrhynchos</td>
<td>American crow</td>
<td>1</td>
<td>1</td>
<td>73.4</td>
<td>352</td>
<td>032</td>
<td>14 814</td>
<td>8.83</td>
</tr>
<tr>
<td>Corvus corone</td>
<td>European crow</td>
<td>12</td>
<td>2</td>
<td>149.5</td>
<td>856</td>
<td>675</td>
<td>18 133</td>
<td>8.11</td>
</tr>
<tr>
<td>Dromaius novaehollandiae</td>
<td>Emu</td>
<td>1</td>
<td>2</td>
<td>86.0</td>
<td>398</td>
<td>710</td>
<td>20 603</td>
<td>7.60</td>
</tr>
<tr>
<td>Ficedula hypoleuca</td>
<td>Pied flycatcher</td>
<td>10</td>
<td>2.5</td>
<td>230.0</td>
<td>1 056</td>
<td>316</td>
<td>45 229</td>
<td>8.15</td>
</tr>
<tr>
<td>Manacus vitellinus</td>
<td>Golden collared manakin</td>
<td>1</td>
<td>1</td>
<td>66.8</td>
<td>312</td>
<td>014</td>
<td>15 782</td>
<td>7.70</td>
</tr>
<tr>
<td>Melopsittacus undulatus</td>
<td>Budgerigar</td>
<td>2</td>
<td>2</td>
<td>89.2</td>
<td>467</td>
<td>567</td>
<td>19 198</td>
<td>8.28</td>
</tr>
<tr>
<td>Parus caeruleus</td>
<td>Blue tit</td>
<td>7</td>
<td>1</td>
<td>99.3</td>
<td>404</td>
<td>049</td>
<td>19 205</td>
<td>7.89</td>
</tr>
<tr>
<td>Streptopelia risoria</td>
<td>Ring-necked dove</td>
<td>1</td>
<td>1</td>
<td>94.8</td>
<td>448</td>
<td>621</td>
<td>21 789</td>
<td>9.22</td>
</tr>
</tbody>
</table>

*Number of individuals that were pooled prior to sequencing.
†One plate represent a full Roche 454 run.
‡Mean number of reads per site.

All contigs were first mapped onto the zebra finch genome using BLAT (Kent 2002) (minimum score = 40, step Size = 5) and the best hit for each contig was considered for annotating the genomic region that was covered. To specifically detect protein-coding genes in the transcriptome data, we downloaded zebra finch coding sequences from the BioMart database (ENSEMBL 54) using BLASTX and FASTY for mapping. We followed a reciprocal blast approach to minimize false positive orthologous sequences. All contigs with a blast hit >10^-20 were discarded, a stringent criterion that leads to a higher identification of orthologous as opposed to paralogous sequences. Quality scores for the contigs were not taken into account. Sequences were assumed to be orthologous if the zebra finch protein sequence had the best hit to a contig and that same contig had the best hit to that particular zebra finch sequence (reciprocal blast criterion). Additionally, sequences that contained frame shifts were discarded. Pair-wise and multiple alignments were generated for all species based on protein sequences using MAFFT Version 6.704b (Katoh & Toh 2008) and back-translated to DNA sequences for subsequent analysis. Alignments are available upon request. Substitution rates were estimated separately for synonymous (ds) and non-synonymous substitutions (dn) using a maximum likelihood method implemented in the CODEML program of the PAML package Version 4.1 (Yang 2007). Pair-wise maximum likelihood analysis were performed in runmode-2. We excluded all alignments that were shorter than 150 bp or that had ds larger than 2 to minimize statistical artefacts from short sequences and saturation effects in ds. To calculate the mean ratio between dn and ds (denoted mean ω) per chromosome or for two species, we divided the mean non-synonymous rate by the mean synonymous rate.

Data on chromosomal location of zebra finch genes was taken from ENSEMBL. For the purpose of this study, we assumed the same location of genes in the other species. Although we know this assumption can not be correct in every single case, we believe it is justified by the high overall degree of chromosome conservation across birds (Griffin et al. 2007). To estimate the male mutation bias (zm), we used the following equation (see Miyata et al. 1987):

\[
zm = \frac{3Z/A - 2}{4 - 3Z/A}
\]

where Z represents mean ds of the Z chromosome and A the mean ds of autosomes.

Statistical analyses

We used weighted multiple regression analysis based on linear mixed effect modelling to decipher the relationship between mean ω per chromosome, mean ds per chromosome and chromosome length. We defined mean ω as the response variable, mean ds and chromosome length as fixed explanatory variables. Visual inspection of the relationship between mean ω and chromosome length suggested adding a quadratic term to the model equation. All variables were normalized by logarithmic transformation prior to analysis. Data were further centred (z-transformed), as parameter estimates of the fixed effects then represent standardized semi-partial regression coefficients that allow direct comparisons of effect sizes across variables. To appropriately account for the study design and pseudo-replication, species identity and differences in slopes of all variables were included as random factors. Model selection was based on a backward selection and Akaike’s information criteria that qualitatively yielded the same results. Chromosomes that harbour only a few genes were given less weight by weighting data points by the reciprocal standard error thus taking into account both
variance of data points and samples sizes. R 2.9.1 (R
DCT 2006) was used to perform all statistical analyses.

Results

Sequencing and assembly

We prepared cDNA from adult brain tissue of 10 differ-
ent bird species (Table 1, Fig. 1). Prior to sequencing
on a Roche 454 GS-FLX platform, the cDNAs were nor-
malized to in theory increase the number of different
transcripts expected to be detected by random sequenc-
ing of templates (cf. Hale et al. 2009). About 1–2.5 full
GS-FLX plate runs were performed for each species and
the total amount of raw sequence data obtained varied
from 53 Mb for Manacus vitellinus (1 plate) to 188 Mb
for Ficedula hypoleuca (2.5 plate runs were performed for
the latter species). The mean read length varied
between 211 and 247 bp.

Before further data analyses, the reads were assem-
bled into larger contigs. Due to the fact that a reference
genome is not available for any of the species, it was
necessary to perform de novo assembly. The average
contig length across species was 453 bp (range 78–
5911 bp). The number of contigs obtained from the
assembly step varied between 15 204 in Archilochus colu-
bris to 45 229 in F. hypoleuca. Mean coverage per base
varied between 7.60 in Dromais novaehollandiae and
11.43 in A. colubris (Table 2). Somewhat surprisingly,
we did not find a significant correlation between cover-
age per base and the total amount of raw sequence data
(R^2 adj = 0.1153, P = 0.80). However, the number of con-
tigs correlated well with the total amount of raw data
(R^2 adj = 0.9006, P = 1.7 \times 10^{-5}).

Contig annotation

With the recent completion of draft zebra finch genome
sequencing (Warren et al., unpublished), there are now
two bird genomes available, also including chicken
(Hillier et al. 2004). We aligned all contigs against the
assembled chicken and zebra finch genomes as to anno-
tate the transcriptome sequences from non-model avian
species. As expected, the proportion of contigs that
aligned to the reference genomes was inversely corre-
lated with genetic distance (Fig. 1, Table 2). For exam-
ple, when zebra finch (which belongs to Passeriformes)
was used as a reference, 91–96% of the contigs from the
other oscine passerines, 90% of the suboscine passerines
(most basal clade within Passeriformes), 68–76% of the
parrot and 46–57% of the dove and emu contigs
aligned. For emu, which is equally distant to chicken
and zebra finch, a very similar proportion of contigs
align to the two reference genomes (Table 2). In the fol-
lowing analyses, we focused on the zebra finch as refer-
ence given that most species sequenced are more
closely related to it than to chicken (Fig. 1).

Fig. 1 A schematic phylogenetic tree indicating the approxi-
mate relationships between species included in the study, fol-
lowing Hackett et al. (2008). As the precise topology is not
critical for the purpose of this study, controversial nodes are
not shown resolved. Branch lengths are not drawn according
to scale.

<table>
<thead>
<tr>
<th>Species</th>
<th>Chicken Contigs</th>
<th>Chicken Proportion</th>
<th>Zebra finch Contigs</th>
<th>Zebra finch Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archilochus colubris</td>
<td>8819</td>
<td>0.58</td>
<td>10 330</td>
<td>0.68</td>
</tr>
<tr>
<td>Calypte anna</td>
<td>9907</td>
<td>0.61</td>
<td>11 423</td>
<td>0.70</td>
</tr>
<tr>
<td>Corvus brachyrhynchos</td>
<td>9252</td>
<td>0.62</td>
<td>13 434</td>
<td>0.91</td>
</tr>
<tr>
<td>Corvus corone</td>
<td>12 647</td>
<td>0.70</td>
<td>17 168</td>
<td>0.95</td>
</tr>
<tr>
<td>Dromais novaehollandiae</td>
<td>11 964</td>
<td>0.58</td>
<td>11 838</td>
<td>0.57</td>
</tr>
<tr>
<td>Ficedula hypoleuca</td>
<td>24 071</td>
<td>0.53</td>
<td>43 049</td>
<td>0.95</td>
</tr>
<tr>
<td>Manacus vitellinus</td>
<td>10 426</td>
<td>0.66</td>
<td>14 188</td>
<td>0.90</td>
</tr>
<tr>
<td>Melopsittacus undulatus</td>
<td>12 360</td>
<td>0.64</td>
<td>14 585</td>
<td>0.76</td>
</tr>
<tr>
<td>Parus caeruleus</td>
<td>10 927</td>
<td>0.57</td>
<td>18 516</td>
<td>0.96</td>
</tr>
<tr>
<td>Streptopelia risoria</td>
<td>8605</td>
<td>0.39</td>
<td>10 122</td>
<td>0.46</td>
</tr>
</tbody>
</table>
represented other ENSEMBL-annotated RNA-types and pseudogenes (Fig. 3). We found that there is a propensity for reads to come from the 3′ end of the coding regions (CDS) of genes (Fig. 4). This is also evident as an excess of 3′ UTR reads over 5′ UTR reads (two-sample t-test, \( P = 3.5 \times 10^{-5} \)) and an excess of the 1 kb downstream 3′ UTR region over the 1 kb upstream 5′ UTR region (two-sample t-test, \( P = 0.0015 \)).

We focused our subsequent analysis on contigs that represent CDS regions. Between 635 (\( A. \) colubris) and 3662 (\( F. \) hypoleuca) ENSEMBL modelled protein coding genes matched those in the zebra finch (Table 3). The number of genes obtained is highly correlated with the amount of raw sequence data (\( R_{adj}^2 = 0.7001, P = 0.0016 \)). On average, ~27% of the CDS of zebra finch protein-coding genes are covered by one or more contigs per species. This value ranges from 20.0% in \( M. \) undulatus to 37.1% in \( C. \) corone.

The ability to perform many molecular evolutionary analyses is dependent on that sequence information is available from the same gene in different species. In total, we were able to retrieve sequence data for 6499 protein-coding genes combined from all species. However, due to the random sampling process of 454 sequencing, there was limited overlap of genes across species. Of the 6499 genes, only 55 were sequenced across all 10 species plus zebra finch. Moreover, even if there is sequence data from the same gene in two or more species, it does not necessarily mean that the same region of the gene has been sequenced in all spe-

### Table 3

<table>
<thead>
<tr>
<th>Species</th>
<th>Aligning contigs (%)</th>
<th>Genes</th>
<th>Coverage*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archilochus colubris</td>
<td>68</td>
<td>635</td>
<td>23.57</td>
</tr>
<tr>
<td>Calypte anna</td>
<td>70</td>
<td>898</td>
<td>23.72</td>
</tr>
<tr>
<td>Corvus brachyrhynchos</td>
<td>91</td>
<td>869</td>
<td>25.41</td>
</tr>
<tr>
<td>Corvus corone</td>
<td>95</td>
<td>2208</td>
<td>37.11</td>
</tr>
<tr>
<td>Dromaius novaehollandiae</td>
<td>57</td>
<td>1395</td>
<td>24.43</td>
</tr>
<tr>
<td>Ficedula hypoleuca</td>
<td>95</td>
<td>3661</td>
<td>31.10</td>
</tr>
<tr>
<td>Manacus vitellinus</td>
<td>90</td>
<td>1206</td>
<td>29.37</td>
</tr>
<tr>
<td>Melopsittacus undulatus</td>
<td>76</td>
<td>1125</td>
<td>20.00</td>
</tr>
<tr>
<td>Parus caeruleus</td>
<td>96</td>
<td>2291</td>
<td>27.94</td>
</tr>
<tr>
<td>Streptopelia risoria</td>
<td>46</td>
<td>941</td>
<td>25.36</td>
</tr>
</tbody>
</table>

*Mean sequence coverage of each gene.
cies. When we exclude all genes that do not have overlapping sequence data in all species, of the 55 genes, we obtain only 15 genes (ATP5A1, B5KFQ7, C1orf151, DYNLRB1, GPM6A, ITGB1BP3, KRAS, LSM3, PVALB-2, RANP1, RPL5, RPL7, RPL23A, SOD2, UQCRB; i.e. three rRNA genes) that are sequenced in all 10 species.

Substitution rate variation

Based on the alignment of orthologous gene sequences, we made pair-wise estimates of the synonymous substitution rate (\(d_S\)) in comparisons between zebra finch and each of the 10 species (Table 4). Mean \(d_S\) increased with genetic distance between species, which is expected under the assumption of a molecular clock. When mean \(d_S\) is estimated for each chromosome in the zebra finch karyotype, we find a clear negative correlation between chromosome size and \(d_S\) in all species comparisons (Fig. 5). Although we do not have information on the chromosomal location of genes or karyotype organization in other species than zebra finch and chicken, using data on chromosome size from zebra finch can be justified on basis of the overall high degree of chromosomal conservation in avian genomes (Griffin et al. 2007). Thus, small chromosomes appear to have elevated mutation rates.

We also find that the Z chromosome tends to have higher \(d_S\) than autosomes. The male mutation bias (\(x_m\)) can be estimated by contrasting the rate of mutation in autosomes and sex chromosomes. As there were no W-linked genes in the sets of genes identified by transcriptome sequencing in different species, we estimated \(x_m\) by comparisons of autosomal and Z-linked rates (see Table 4). Estimates of the mean rates of synonymous (\(d_S\)) and non-synonymous substitution rate (\(d_N\)) and their ratio (\(\omega\)) in comparisons with zebra finch

Table 4 Estimates of the mean rates of synonymous (\(d_S\)) and non-synonymous substitution rate (\(d_N\)) and their ratio (\(\omega\)) in comparisons with zebra finch

<table>
<thead>
<tr>
<th>Species</th>
<th>No. genes</th>
<th>Mean length</th>
<th>(d_N)</th>
<th>(d_S)</th>
<th>(\omega)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archilochus colubris</td>
<td>617</td>
<td>390</td>
<td>0.0286</td>
<td>0.3793</td>
<td>0.0753</td>
</tr>
<tr>
<td>Calypte anna</td>
<td>867</td>
<td>400</td>
<td>0.0283</td>
<td>0.3795</td>
<td>0.0746</td>
</tr>
<tr>
<td>Corvus brachyrhynchos</td>
<td>837</td>
<td>377</td>
<td>0.0173</td>
<td>0.1665</td>
<td>0.1037</td>
</tr>
<tr>
<td>Corvus corone</td>
<td>2168</td>
<td>565</td>
<td>0.0160</td>
<td>0.1400</td>
<td>0.1142</td>
</tr>
<tr>
<td>Dromaius novaehollandiae</td>
<td>1315</td>
<td>400</td>
<td>0.0378</td>
<td>0.4658</td>
<td>0.0811</td>
</tr>
<tr>
<td>Ficedula hypoleuca</td>
<td>3575</td>
<td>523</td>
<td>0.0191</td>
<td>0.1520</td>
<td>0.1254</td>
</tr>
<tr>
<td>Manacus vitellinus</td>
<td>1180</td>
<td>410</td>
<td>0.0217</td>
<td>0.2616</td>
<td>0.0828</td>
</tr>
<tr>
<td>Melopsittacus undulatus</td>
<td>1094</td>
<td>336</td>
<td>0.0337</td>
<td>0.3740</td>
<td>0.0900</td>
</tr>
<tr>
<td>Pardus caeruleus</td>
<td>2249</td>
<td>495</td>
<td>0.0146</td>
<td>0.1238</td>
<td>0.1179</td>
</tr>
<tr>
<td>Streptopelia risoria</td>
<td>764</td>
<td>397</td>
<td>0.0356</td>
<td>0.4233</td>
<td>0.0841</td>
</tr>
</tbody>
</table>

Alignments shorter than 150 bp or with \(d_S > 2\) were excluded from the estimates.

Materials and methods. For eight of the 10 species comparisons \(x_m\) was higher than one (Table 5). Altogether, mean \(x_m\) was 2.4, suggesting that the male mutation rate is at least twice as high the female mutation rate across broad avian comparisons.

Selection

We sought to study the overall patterns of natural selection in avian genomes by estimating the \(d_{NS}/d_S\) ratio (\(\omega\)) of orthologous genes in each comparison with zebra finch and one of the species subject to transcriptome sequencing. In birds, there is evidence for striking differences in recombination rate among chromosomes, with a strong negative correlation between recombination rate and chromosomes size (Hillier et al. 2004; Groenen et al. 2009). Theory predicts that differences in recombination rates should translate into differences in effective population size (Hill & Robertson 1966). We

Table 5 Estimates of the male mutation bias (\(x_m\)) based on comparisons of synonymous substitution rates (\(d_S\)) in autosomes and the Z chromosome according to the zebra finch genome mapping

<table>
<thead>
<tr>
<th>Species</th>
<th>(d_S\ Z)</th>
<th>(d_S\ auto)</th>
<th>(x_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archilochus colubris</td>
<td>0.449</td>
<td>0.368</td>
<td>4.888</td>
</tr>
<tr>
<td>Calypte anna</td>
<td>0.354</td>
<td>0.380</td>
<td>0.0283</td>
</tr>
<tr>
<td>Corvus brachyrhynchos</td>
<td>0.198</td>
<td>0.162</td>
<td>5.000</td>
</tr>
<tr>
<td>Corvus corone</td>
<td>0.158</td>
<td>0.137</td>
<td>2.703</td>
</tr>
<tr>
<td>Dromaius novaehollandiae</td>
<td>0.476</td>
<td>0.464</td>
<td>1.168</td>
</tr>
<tr>
<td>Ficedula hypoleuca</td>
<td>0.163</td>
<td>0.147</td>
<td>1.970</td>
</tr>
<tr>
<td>Manacus vitellinus</td>
<td>0.290</td>
<td>0.259</td>
<td>2.210</td>
</tr>
<tr>
<td>Melopsittacus undulatus</td>
<td>0.415</td>
<td>0.368</td>
<td>2.242</td>
</tr>
<tr>
<td>Pardus caeruleus</td>
<td>0.145</td>
<td>0.122</td>
<td>3.604</td>
</tr>
<tr>
<td>Streptopelia risoria</td>
<td>0.376</td>
<td>0.423</td>
<td>0.500</td>
</tr>
</tbody>
</table>
therefore expect that selection is more efficient in small than in large chromosomes. Under the assumption that most mutations that affect fitness are slightly deleterious and that selection is more efficient in chromosomes with high recombination rates, $d_N/d_S$ ratios should be smaller in small chromosomes. Indeed, we found that within each of the 10 species comparisons with zebra finch, mean $\omega$ per chromosome was positively correlated with chromosome size, supporting the idea that genes in small chromosomes evolve under stronger constraint (Fig. 6). An alternative explanation would be that adaptive evolution is more important for genes in larger chromosomes.

However, as stated before mean $d_S$ per chromosome was also found to be negatively correlated with chromosome size. Recently, there have been theoretical and empirical claims that mean $d_S/d_S$ may not be independent of mean $d_S$ (Rocha et al. 2006; Kryazhimskiy & Plotkin 2008; Wolf et al. 2009). This notion is reflected in our observation that mean $\omega$ of each species comparison (0.075–0.125) is negatively correlated with genetic distance between zebra finch and the compared species, at least when $d_S$ is taken as a proxy for this distance (Fig. 7). The relationship between mean $d_N/d_S$ and chromosome size (and hence recombination rate) may thus not reflect differences in selection efficiency but be simply explained by differences in mutation rate (measured by mean $d_S$) among chromosomes. To account for this possibility, we included both mean $d_N$ and chromosome size as explanatory variables of mean $d_N/d_S$ in the statistical models. Closer inspection of the models suggests that mean $\omega$ is indeed negatively correlated with mean $d_N$ (Fig. 8). The two best statistical models include both chromosome length (linear and quadratic terms) and mean $d_S$ as explanatory variables (Table 6). Inspection of standardized semi-partial regression coefficients suggests that the effects of mean $d_S$ and chromosome size are similar (Table 6). Despite considerable co-variation in mean $d_S$ and chromosome length ($r = −0.35$), we thus conclude that both $d_S$ and chromosome size are correlated with mean $\omega$.

**Discussion**

*Retrieving gene sequence data from non-model organisms*

There has recently been a burst of studies, in a range of species, reporting on transcriptome sequencing efforts...
using next-generation sequencing technology (Cheung et al. 2006, 2008; Novaes et al. 2008; Vera et al. 2008; Barakat et al. 2009; Hahn et al. 2009; Hale et al. 2009; Meyer et al. 2009). Although these studies have generally focused on a single species, our approach was to perform transcriptome sequencing across several divergent bird lineages, ultimately with the purpose of being able to study the role of natural selection in avian gene sequence evolution. There was no reference genome available for any of the species sequenced so we made use of the genome of the closest relative, the zebra finch, for compiling and annotating genes.

We identified roughly 1000 genes per full GS-FLX run and there was no obvious saturation in gene discovery in those cases where 1.0–1.5 additional runs were performed. This finding supports the common belief that brain tissue represents a rich source of different transcripts and that even larger-scale sequencing should identify more genes; more sequencing should also yield increased sequence depth and gene coverage.

About three-quarters of the contigs that aligned to the zebra finch genome are from protein-coding genes. This indicates either such transcripts are rarely expressed in avian brains or that the ENSEMBL models have grossly underestimated the occurrence of such transcripts, given the recent realization of the high abundance of transcribed non-coding RNAs in vertebrate genomes (Bertone et al. 2004; Fantom et al. 2005; Sultan et al. 2008). In turn, this has bearing to the fraction of contigs that do not align to the zebra finch genome. Although only 5–10% of contigs from the species most closely related to zebra finch failed to align, this proportion increased to ~50% in the more distantly related species. This result is not unexpected, as increasing genetic distance will lead to decreasing levels of homologous alignments and thus difficult to detect orthologous sequence relationship of regions evolving under limited constraints (Ponting 2008). We also note that the decreasing proportion of contigs aligning to the reference genome with increasing genetic distance argues against that non-aligning sequences would consist of a significant amount of xenobiotic contamination or represent technical artefacts.

**Nucleotide substitution and chromosome size**

Two general conclusions about substitution rate variation in birds can be drawn from our study. These are evidence for male-biased mutation and a negative relationship between chromosome size and mutation rate. As there were so few genes for which we could retrieve sequence data in multiple species, all substitution rate

<table>
<thead>
<tr>
<th>Model selection criteria</th>
<th>Parameter estimates of fixed effects</th>
<th>Standard deviation of random effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>(CL)²</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>(CL)²</td>
</tr>
<tr>
<td>1 (15)</td>
<td>2409</td>
<td>2432</td>
</tr>
<tr>
<td>2 (6)</td>
<td>2412</td>
<td>2468</td>
</tr>
<tr>
<td>3 (5)</td>
<td>2442</td>
<td>2461</td>
</tr>
<tr>
<td>4 (4)</td>
<td>2454</td>
<td>2469</td>
</tr>
</tbody>
</table>

Listed are model names and their degrees of freedom (d.f.), information criteria used for model selection (AIC, Akaike's information criterion; BIC, Bayesian information criterion; LRT, likelihood ratio test), parameter estimates of fixed effects and variance components of random effects. As data has been z-transformed prior to analysis, parameter estimates represent standardized semipartial regression coefficients that can be directly compared between variables and give an impression on relative effect sizes

**a** < 0.001, **a** < 0.01, **a** < 0.05.

**Table 6** Comparison of linear mixed effect models fitting the relationship of mean α per chromosome vs. chromosome length (CL and CL²) and mean dₛ per chromosome (dₛ) as explanatory variables.
estimates come from pair-wise comparisons of zebra finch and one of the species from the transcriptome sequencing effort. As zebra finch is included in every such comparison, this means that the observed substitution rates are not independent. Moreover, as several of the species subject to transcriptome sequencing belong to the same family or order of birds, they share substitutions from internal branches. This should be kept in mind when the results are examined.

If synonymous substitutions are considered effectively neutral and thus accepted to reflect the underlying mutation rate, it is clear for all species comparisons that the incidence of mutation is higher in small chromosomes (Fig. 5). This mirrors what was observed in chicken-human (Hillier et al. 2004) and chicken-turkey comparisons (Axelsson et al. 2005), and therefore seems to represent a general trend in avian genomes. It is often observed, across taxonomic groups, that several genomic parameters like base composition, recombination rate and substitution rate covary, although it may be difficult to reveal the causality of such relationships (e.g. Hardison et al. 2003). GC content and recombination rate correlate negatively with chromosome size in birds (Hillier et al. 2004) and both offer plausible explanations to the elevated rate of synonymous substitutions in small chromosomes. The incidence of the highly mutable CpG dinucleotide is positively correlated with GC content and there is significant correlation between GC and substitution rate in birds (Axelsson et al. 2005; Webster et al. 2006). We surmise that recombination might indirectly be responsible for the correlation between \(d_S\) and chromosome size as high recombination tends to give high GC content, likely due to biased gene conversion (Pozzoli et al. 2008). Alternatively, there might be a more direct link due to recombination being mutagenic, i.e. that recombination and mutation rates correlate positively (Hellmann et al. 2003). There are several important implications to the within-genome heterogeneity seen in avian mutation rates. For example, attempts towards molecular dating of divergence times need to make sure that rates are calibrated using appropriately selected sequence data.

**Male-biased mutation**

The number of mitotic cell divisions in germ line is typically much higher in males than in females. If germ line mutations are somehow replication-associated, due for example to replication-errors, the majority of mutations can be expected to have a paternal origin (Ellegren 2007). Previous work in birds, in most cases based on data from a limited number of species, have revealed a moderate male mutation bias, most often in the range of two to three times higher mutation rate in males than in females (e.g. Axelsson et al. 2004). Our study confirms this level of male excess in avian germ line mutation, at least broadly speaking (mean \(x_m = 2.4\)). However, the estimates of \(x_m\) in individual species comparisons vary considerably (0.03–5.00). This can most likely be attributed to sampling error as estimates of \(x_m\) from autosome–Z chromosome comparisons are sensitive to parameter estimation of substitution rates in individual chromosome categories (more so than in autosome–W or Z–W chromosome comparisons; Ellegren 2007). Alternatively, as it has been demonstrated that the magnitude of the male mutation bias can be affected by life history (Bartosch-Harlid et al. 2003), the variation might at least in part be biologically meaningful. Specifically, increased sperm production from sexual selection would, in theory, increase the male mutation bias. However, we find no obvious reasons to believe that sexual selection would be more intense in those lineages represented by particularly high \(x_m\) estimates (American crow, Ruby-throated hummingbird, Blue tit) compared to those with very low estimates (Anna’s hummingbird, Ring-necked dove).

**Selection**

Theory predicts that slightly deleterious mutations should be more effectively purged in regions of high recombination, the so-called Hill-Robertson effect (Hill & Robertson 1966). As slightly deleterious mutations are likely to contribute to the accumulation of non-synonymous substitutions in populations with low-moderate effective population size \(N_e\) (Hahn 2008; Ellegren 2009), \(\omega\) is expected to be lower in regions of high recombination and higher in regions of low recombination (see Bullaughey et al. 2008). The rate of recombination varies considerably among avian chromosomes and is strongly negatively correlated with chromosome size (Hillier et al. 2004; Groenen et al. 2009). Previously, \(\omega\) was found to be smaller in microchromosomes than in macrochromosomes in the chicken-turkey comparison, and it was hypothesized that this was the result of Hill–Robertson inference (Axelsson et al. 2005). Given that chromosome size is a good predictor for recombination rates, our observation, seen across all species comparisons, of decreasing \(\omega\) with decreasing chromosome size would be in accordance with this interpretation.

The situation is complicated by there being substantial uncertainties associated with the basic properties of the apparent \(d_S/d_K\) ratio, including time dependency (Rocha et al. 2006), effects of within-population variation (Kryazhimskiy & Plotkin 2008), gene conversion (Berglund et al. 2009), estimation procedure, sequencing and annotation errors (Schneider et al. 2009). Using simulations and empirical genomic data, we have...
recently demonstrated that \( \omega \) taken as phase value can show a strong negative relationship with \( d_S \) (Wolf et al. 2009). Most of this variation is explained by the correlated sampling variance in \( d_S \), which affects the denominator of \( \omega \) and \( d_S \) likewise. This leads to a negative correlation between \( \omega \) and \( d_S \), which will disappear when taking the correct mean of \( \omega \) (sum of \( d_S \))/sum of \( d_S \) for a large number of genes. The influence of correlated sampling variances will persist, however, if the mean is based on only a few genes, as is the case for most small chromosomes in our data set. This suggests that at least part of the correlation between mean \( \omega \) and mean \( d_S \) reflects a stochastic artefact. Due to co-linearity between \( d_S \) and chromosome size, this artefact will also be visible in the relationship between mean \( \omega \) and chromosome size and can lead to incorrect inferences regarding the role of selection. We here treated this problem with a semi-partial regression, finding that the effects of mean \( d_S \) and chromosome size on \( \omega \) are similar in size, and that chromosome size therefore is an important factor in explaining the role of natural selection on avian genes. The most obvious explanation to this would be through an effect of recombination, i.e. Hill–Robertson inference. To formally test this hypothesis, recombination rate data from the species under investigation would be needed and better ways of controlling for the inter-dependency of mean \( \omega \) and mean \( d_S \) must be found (see, e.g. Figane & Eyre-Walker 2009).

**Acknowledgements**

We thank Diethard Tautz for support and Garth Spellman for help with samples. H.E. acknowledges funding from the Swedish Research Council and the Knut and Alice Wallenberg Foundation, and B.A.S. acknowledges NSF IBN-0646459 Grant. Postdoctoral support for D.E.J. was provided by National Science Foundation, and B.A.S. acknowledges NSF IBN-0646459 Grant.

**Conflicts of interest**

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**References**


Groenen MAM, Wahlberg P, Foglio M et al. (2009) A high-density SNP-based linkage map of the chicken genome.


