DNA Research **19**, 325–333, (2012)
Advance Access publication on 29 May 2012
doi:10.1093/dnares/dss015

# Transcriptome Analysis of a North American Songbird, *Melospiza melodia*

Anuj Srivastava[1,*,†,‡], Kevin Winker[2,†], Timothy I. Shaw[1], Kenneth L. Jones[3,4,5], and Travis C. Glenn[1,3,4]

*Institute of Bioinformatics, Davidson Life Sciences, University of Georgia, Athens, GA 30602, USA[1]; University of Alaska Museum, 907 Yukon Drive, Fairbanks, AK 99775, USA[2]; Georgia Genomics Facility, University of Georgia, 110 Riverbend Road, Athens, GA 30602, USA[3]; Department of Environmental Health Science, University of Georgia, EHS Building, Athens, GA 30602, USA[4]; and Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA[5]*

*To whom correspondence should be addressed. Tel. +1 207-288-6000. Fax. +1 907-474-5469.
E-mail: anuj.srivastava@jax.org

Edited by Minoru Yoshida
(Received 8 January 2012; accepted 19 April 2012)

**Abstract**

**An effective way to understand the genomics of divergence in non-model organisms is to use the transcriptome to identify genes associated with divergence. We examine the transcriptome of the song sparrow (*Melospiza melodia*) and contrast it with the avian models zebra finch (*Taeniopygia guttata*) and chicken (*Gallus gallus*). We aimed to (i) obtain a functional annotation of a substantial portion of the song sparrow transcriptome; (ii) compare transcript divergence; (iii) efficiently characterize single nucleotide polymorphism/indel markers possibly fixed between song sparrow subspecies; and (iv) identify the most common set of transcripts in birds using the zebra finch as a reference. Using two individuals from each of three populations, whole-body mRNA was normalized and sequenced (110 Mb total). The assembly yielded 38 539 contigs [N50 (the length–weighted median) = 482 bp]; 4574 were orthologous to both model genomes and 3680 are functionally annotated. This low-coverage scan of the song sparrow transcriptome revealed 29 982 SNPs/indels, 1402 fixed between populations and subspecies. Referencing zebra finch and chicken, we identified 43 and 5 fast-evolving genes, respectively. We also identified the most common set of transcripts present in birds with respect to zebra finch. This study provides new insight into songbird transcriptomes, and candidate markers identified here may help research in songbirds (oscine Passeriformes), a frequently studied group.**

**Key words:** EST; genetic markers; next generation sequencing; songbird speciation; SNP characterization

## 1. Introduction

Determining the genetic underpinnings of organismal divergence and speciation will provide insight into the evolutionary generation of biodiversity, and next-generation sequencing is propelling such studies in non-model organisms.[1,2] An effective way to initiate genomic-wide data sets in non-model organisms is to focus on the transcriptome, or expressed sequence, which, unlike a whole-genome approach, increases the data's focus on functional genomic attributes.[3,4] As these data become available, evolutionary biologists will be able to make contrasts within and among lineages to identify genes associated with divergence.[5–8] To gain insight into the genes associated with avian diversification, we examine the transcriptome of the song sparrow (*Melospiza melodia*) and contrast it with the model birds zebra finch (*Taeniopygia guttata*) and chicken (*G. gallus*).

† These authors contributed equally.
‡ Present address: The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA.

© The Author 2012. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

The song sparrow is broadly distributed across North America and exhibits pronounced morphological variation, with 25 subspecies recognized (of 52 described[9]). It has been extensively studied over the past 70 yrs; it is considered a model vertebrate species for field research; and it will continue to be a focus for questions about the causes of population variation in behaviour, demographics, and morphology.[10] Our goals in this study were to (i) obtain a functional annotation of a substantial portion of the song sparrow transcriptome; (ii) compare transcript divergence between the song sparrow and the two bird genomes sequenced and assembled to the highest quality thus far, zebra finch (*T. guttata*) and chicken (*G. gallus*); (iii) efficiently characterize a set of single nucleotide polymorphism (SNP)/indel markers that may be fixed between song sparrow subspecies; and (iv) identify the most common set of transcripts present in bird species using the zebra finch as a reference. Achieving these goals will establish important baseline data for a non-model organism in a speciose group (passerines or songbirds) frequently studied.

## 2. Materials and methods

### 2.1. Samples, cDNA library, and sequencing

Two song sparrows still undergoing growth (from embryo to just-fledged) were sampled from each of three Alaska populations (the northwestern most distribution of the species), chosen because they span some of the most pronounced morphological diversity that occurs in the species (Fig. 1): two island populations of *M. m. maxima* (from Attu and Adak islands; an egg and a very young nestling from Attu Island, unvouchered; and vouchers UAM 27831 and 27832 from Adak Island) and one mainland population of *M. m. caurina* (from Cordova, vouchers UAM 27829 and 27830). The Attu and Adak populations of *Melospiza m. maxima* are the largest in the species and also have different plumage coloration; in addition, they are non-migratory, unlike the population from Cordova, which is also smaller and darker (Fig. 1).

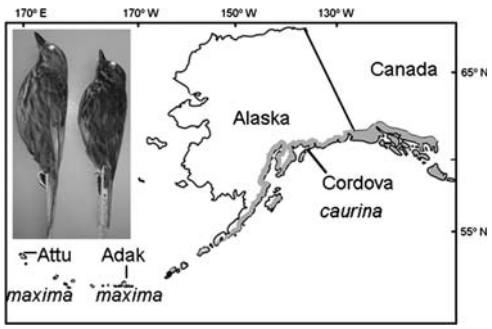


**Figure 1.** Samples in this study came from Cordova (*Melospiza melodia caurina*, right in inset) and Adak and Attu islands (*M. m. maxima*, left in inset); grey shading indicates the species' range.

All samples were obtained in June (spring) at a very young age and only two were sexable (both females, one each from Cordova and Adak). The egg was homogenized, whereas from the others six tissues (brain, liver, heart, muscle, bone, and pancreas) were taken, minced and placed in RNAlater (Qiagen, Valencia, CA) within minutes of death and then frozen. In the laboratory, tissues were homogenized and total RNA was isolated using Trizol (Invitrogen, Carlsbad, CA) and subsequently cleaned using a Qiagen RNeasy column.

Equal amounts of RNA from individuals of each population were pooled and an MINT universal cDNA kit (Evrogen, Moscow, Russia) with primers modified specifically for 454 procedures[11] was used to create cDNA libraries enriched for full-length transcripts. We then normalized the three cDNA libraries using the TRIMMER cDNA normalization kit (Evrogen) to substantially decrease the relative abundance of common transcripts. The normalized cDNA was fragmented and prepared for sequencing using standard 454 procedures, including independent molecular identifiers [MID tags: Cordova (MID 13), Attu (MID 18) and Adak (MID 19)] for each of the three populations. As each library contained a unique MID tag, libraries were pooled and sequenced as a single sample. Sequencing was performed at the University of Georgia's Georgia Genomics Facility on a Roche 454 FLX using Titanium chemistry.

### 2.2. Assembly, polymorphism, and ortholog identification

Bases were called from the 454-generated sff file using Pyrobayes,[12] which provides improved accuracy in the estimation of base qualities for pyrosequences. We removed MINT primer sequences, short sequences, and other contaminatants using SeqClean (http://compbio.dfci.harvard.edu), and reads from all three populations were combined. We performed a combined assembly of reads using MIRA,[13] and then used GigaBayes,[14] a short-read SNP and short indel discovery program, to detect polymorphisms. To make the SNP/indel predictions more reliable, we used the more stringent criteria that the minor allele must occur at least three times and be present at $\geq$10% relative to the major allele frequency when $>$30 reads per locus were obtained (after combining all the reads for particular alleles among different subspecies; sequences with fewer reads are considered the minor allele and sequences with more reads are considered the major allele). We identified orthologous contigs (against the

zebra finch and chicken genomes) using the reciprocal blast approach, because it has been found to be superior to sophisticated orthology detection algorithms.[15] A stringent cutoff of 1e−20 was used to separate paralogues from orthologues. The cDNA sequences from the zebra finch (taeGut3.2.4.60.cdna.all.fa) and chicken (WASHUC2.60.cdna.all.fa) were obtained from the Biomart database (www.biomart.org). Although the zebra finch is a passerine and thus more closely related to the song sparrow, the chicken database contains sequences from whole growing chicks, whereas that of the zebra finch emphasizes neural transcripts.

To identify likely genomic positions of the song sparrow contigs, we mapped them against genomic sequences of the zebra finch (taeGut3.2.4.60.dna_rm.toplevel.fa) and chicken (WASHUC2.60.dna_rm.toplevel.fa) using BLAT[16] with default criteria. We obtained feature information for protein-coding genes and ncRNA using the Ensemble (http://uswest.ensembl.org/index.html/) Xenoref and gtf files, respectively.

### 2.3. Most common set of transcripts in birds

To find the most common set of transcripts in birds with respect to zebra finch, we collected and assembled (454 GS assembler version 2.5) the transcriptome sequence of 12 bird species (publicly available sequence[5,7,8,17]). The orthologous sequence with respect to zebra finch was determined using the bidirectional blast best hit method (1e−20). Only contigs >200 bp were used in the analysis. After determining the orthologous sequences, we sorted them in decreasing order and added orthologous sequences from other species sequentially to find the most common set.

### 2.4. Functional annotation of contigs

We used Blast2GO[18] (B2G) to functionally annotate the contigs. A combined graph was generated for each gene ontology (GO) category. For the molecular function division, a graph was obtained using default criteria and for the other two divisions (cellular component and biological process), seq/node filter values were changed to 4/10 to prevent overloading the graphs.

### 2.5. Estimation of substitution rates

Substitution rates were estimated for contigs that were orthologous to both zebra finch and chicken. Reading frames for these contigs were identified using BLASTX[19] against protein sequences of zebra finch (taeGut3.2.4.60.pep.all.fa) and chicken (WASHUC2.60.pep.all.fa) obtained from Biomart (www.biomart.org). Sequences that produced significant alignments were extracted (using their coordinates), translated, and aligned using CLUSTALW.[20] Sequences that contained frame shifts were excluded from the analysis. Corresponding codon alignments were produced using PAL2NAL,[21] and, finally, rates were estimated using a maximum likelihood method implemented in the CODEML program of the PAML package Version 4.1.[22] Pairwise maximum likelihood analyses were performed in runmode-2. The estimated rates of nonsynonymous to synonymous substitutions ($K_a/K_s$ values) were plotted as a scatter plot in the range of 0−2.0.

## 3. Results and discussion

### 3.1. Sequence assembly

The pooled reads from all three populations yielded 131 Mb (458 808 sequences) of raw data, which was reduced to 110 Mb (381 474 sequences) after the use of SeqClean (Table 1). The mean raw and cleaned read lengths were 286 and 290 bp, respectively. Poor-quality reads were often very short and were purged entirely prior to assembly. Without a reference genome for the song sparrow, *de novo* assembly was required. Cleaned sequences were assembled into 38 539 contigs with N50 and N90 values of 482 and 317 bp, respectively (Supplementary data). There were 1417 singletons. The mean coverage per contig was 3.93 X and the mean GC content per contig was 43.6%.

We acknowledge that the amount of sequencing presented is insufficient to allow a high-quality assembly of the extremely diverse transcriptome that we have sampled. A large number of tissues were sampled, and these clearly contain a large and diverse set of transcripts (see Section 3.2). Simulations indicate that transcriptomes sequenced with 454 Titanium

**Table 1.** Number of reads and assembly statistics for three song sparrow populations (SRA 048516)

| Subspecies | Locality | $n$[a] | MID | Raw reads | Cleaned reads | Cleaned bases (MB) |
|---|---|---|---|---|---|---|
| *M. m. caurina* | Cordova | 2 | 13 | 138 439 | 114 098 | 32.5 |
| *M. m. maxima* | Adak | 2 | 19 | 135 588 | 117 166 | 34.7 |
| *M. m. maxima* | Attu | 2 | 18 | 184 781 | 150 210 | 42.8 |
| *Combined* | — | 6 | — | 458 808 | 381 474 | 110 |

[a]Number of individuals pooled prior to sequencing.

chemistry will quickly lead to about twice as many contigs as transcripts, and additional sequences only gradually cause the number of contigs to reach the number of transcripts (i.e. the point when contigs = transcripts; data not shown). Thus, quite large numbers of additional sequences will be necessary to fully assemble the transcripts contained in these cDNA libraries. Given the relatively high cost of 454 sequencing, it would be more economical to obtain the additional sequences as paired-end reads on Illumina or Ion Torrent platforms.

### 3.2. *Functional annotation*

B2G, which we used to functionally annotate the contigs, has three annotation steps involving (i) a blast against databases, (ii) mapping against GO resources, and (iii) annotation to generate reliable functional assignments. In our data, 12 880 of the contigs (33.46% overall, of which 8540 were unique hits) had significant matches to currently known proteins in the NCBI non-redundant protein database. Because one-third of the contigs hit the same proteins as other contigs in our data, this indicates that large transcripts were often split among multiple contigs in our assembly. Although it is possible to use the zebra finch or chicken proteins as a reference to scaffold the song sparrow contigs, we did not do this because it could make chimeras, and assembly of full-length genes was not a major goal of this work.

As expected, zebra finch and chicken were identified as the top two species with the best blast hits for our song sparrow contigs (Table 2). Contigs with significant blast matches were functionally annotated. GO resource assignment was found for 3949 (10.2%) of the total contigs (with 24 363 GO terms; there can be multiple terms per contig), of which 3367 (8.7% of all contigs) were functionally annotated (Supplementary Sheet 1).

In the first GO division, 'biological process',[23] 22 categories were identified. Most contigs (3578 = 53.1%) were involved in 'cellular and metabolic processes'. The second most abundant category was 'biological regulation and localization' (1253 = 18.6%; Supplementary Fig. S1A). Within the second division, 'molecular function',[23] nine major categories were identified. Most of the contigs were functionally related to 'nucleotide binding' (1966 = 43.9%) and 'catalytic activity' (1266 28.2%; Supplementary Fig. S1B). Finally, the last division, 'cellular component',[23] also had nine categories. Gene products were primarily expressed intracellularly (2322 = 41.9%) or in the membrane bound/non-membrane bound organelle (1787 = 32.3%; Supplementary Fig. S1C).

All of the GO results should be viewed with caution because the depth of the available sequences ensures that most highly expressed transcripts will have been sequenced but many low-expression transcripts will not have been detected. The normalization techniques used substantially increased the number of low-expression transcripts sequenced, but the number of sequences obtained is insufficient to overcome the bias toward highly expressed transcripts.

### 3.3. *Polymorphism detection*

We detected a total of 29 982 SNPs/indels that were spread relatively evenly within, between, and among all three populations (Fig. 2, Supplementary Sheet 2). A total of 1402 SNPs/indels were fixed between populations and subspecies (Fig. 3; the sum of all pairwise comparisons is 1635 because some pairwise SNPs are found in more than one pair). Out of the 1402, there were 392 and 410 SNPs/indels between subspecies and within-subspecies, respectively. This provides many SNPs/indels for further study (Supplementary Sheet 2), although given our limited sampling of individuals within populations ($n = 2$) many will not be true fixed differences (i.e. they are false positives, other individuals contain these variants). We also note that we have used quite stringent criteria for SNP/indel assignment.

**Table 2.** Species with ≥100 top hits from B2G

| Species | Hits |
|---|---|
| *T. guttata* | 7820 |
| *G. gallus* | 2222 |
| *Homo sapiens* | 235 |
| *Monodelphis domestica* | 193 |
| *Mus musculus* | 187 |
| *Ailuropoda melanoleuca* | 177 |
| *Ornithorhynchus anatinus* | 149 |
| *Canis familiaris* | 119 |
| *M. melodia* | 113 |
| *Rattus norvegicus* | 100 |

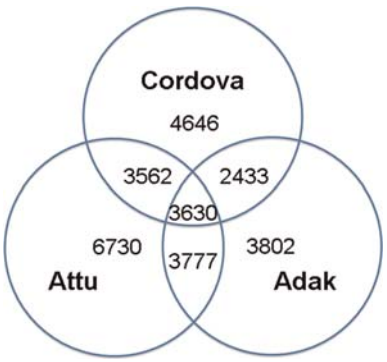


**Figure 2.** Numbers of SNPs and indels that are within and shared between and among three populations of song sparrows.

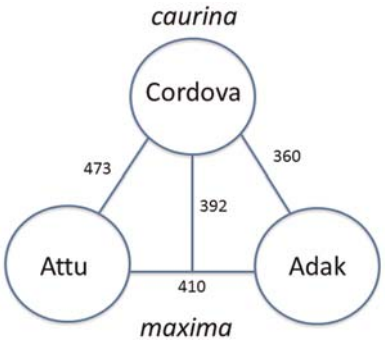


**Figure 3.** SNPs and indels that are fixed between and among three populations of song sparrows. There are 392 SNPs/indels that are identical in Attu and Adak, but different from Cordova. Because sample sizes are small, these figures include false positives.

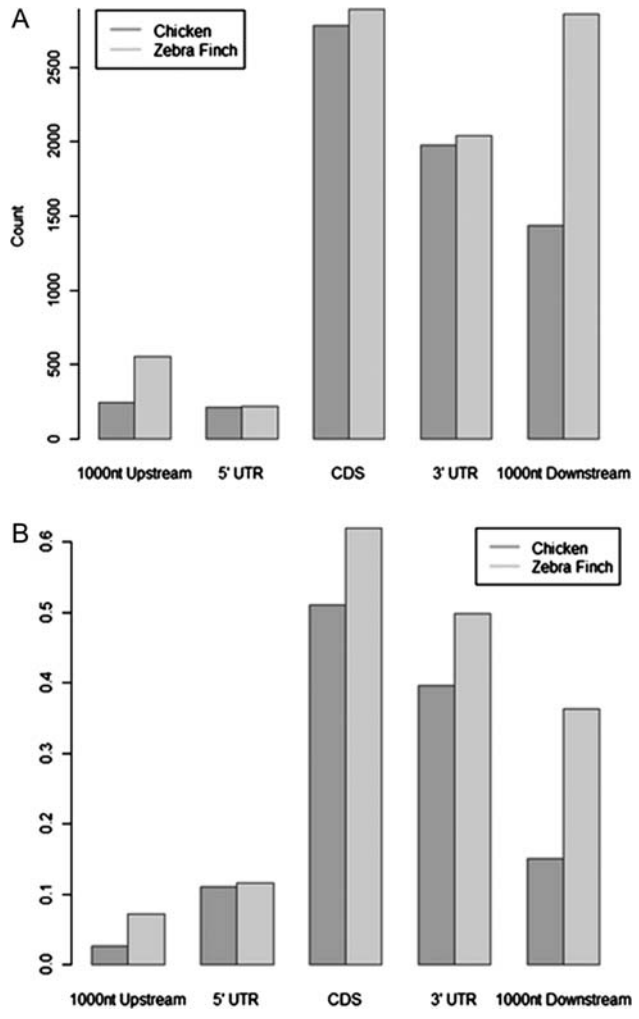


**Figure 4.** Histogram displaying the proportion of contigs mapped to particular features of protein coding genes of zebra finch and chicken (UTR is the untranslated region, and CDS is the coding sequence). The upper panel displays the raw count and the lower panel normalized values (the proportion discovered relative to how many could be discovered within each category).

By requiring at least three reads for the minor allele, a minimum of six times coverage is required to call a SNP. Because our average assembly depth is only about four times, most polymorphic nucleotides in our contigs will not pass our criteria for SNP discovery. Because of this, we have biased the SNPs to be from the relatively highly expressed transcripts. Many additional SNPs/indels occur in song sparrows, we describe only those with a high probability of being real, not sequencing artefacts. None of these issues limits our ability to achieve our stated goals, but we note them so that it is understood that we have made appropriately cautious interpretations of our results.

### 3.4. *Orthology with zebra finch and chicken*

The reciprocal blast approach identified 4574 contigs as orthologous to both zebra finch and chicken. As expected because of phylogenetic relationships, more contigs were identified as orthologous to the zebra finch than the chicken: the set [unique song sparrow (orthologues) unique zebra finch] was [32 435 (6104) 12 493], whereas the set [unique song sparrow (orthologues) unique chicken] was [32 767 (5772) 16 518]. A substantial number of orthologous contigs (3894) were found to have the same chromosome location in the zebra finch and chicken (Supplementary Sheet 1).

### 3.5. *Localization of contigs*

The zebra finch and chicken genomes were used as references to locate the contigs. BLAT mapping of our assemblies against these genomes showed sequences that uniquely mapped to particular features of the reference genomes [5′UTR (untranslated region), 3′UTR, CDS (coding sequence), 1 kb upstream, 1 kb downstream; Fig. 4A]. Based on the zebra finch genome annotation, nearly 34% of mapped contigs (2890 of 8561) were found to be in CDS regions. Even with the use of the MINT cDNA construction kit, which is meant only to allow amplification of full-length transcripts, we still observed a substantial bias toward contigs mapping to 3′UTR and 1 kb downstream _relative to 5′UTR and 1 kb upstream. The normalized distributions clearly indicate that our libraries contain relatively few transcripts that are full length (Fig. 4B). Similar patterns, although with slightly fewer hits, were obtained from mapping to the chicken genome. The localization of contigs containing SNPs/indels mapped against the zebra finch and chicken genomes showed that a major proportion of polymorphisms belongs to coding sequences (Supplementary Fig. S2A and B). Contigs with SNPs/indels had more blast hits to the zebra finch than to the chicken, reflecting the overall pattern of all contigs. Few RNA genes were also

found by BLAT mapping (Supplementary Fig. S3A and B).

### 3.6. *Common set transcripts in birds*

We determined the orthologous transcripts with respect to zebra finch using the bidirectional blast best hit method in 12 bird species. From the orthologous sequences, we determined the most common set of transcripts of zebra finch which is present in all species or most of the species. The first big set of transcripts (1004 zebra finch sequences) was present in seven bird species. The second largest set comprised 219 and 126 sequences present in 10 and 12 bird species, respectively, and, finally, 19 sequences were present in all 13 species. Detailed information regarding species used and orthologous sequences is given in the Supplementary Sheet 3. Further, we checked the pathways in which these common transcripts might be involved using DAVID[24,25] and found that they mainly related to oxidative phosphorylation, ribosome biogenesis, and cardiac muscle contraction. These are housekeeping genes[26,27] which explains the frequent occurrence of these in all avian species. With respect to the chromosomal location of common transcripts, we did not find any significant bias related to any particular chromosome.

### 3.7. *Estimation of $K_a/K_s$*

Substitution rates were estimated for the 4574 contigs orthologous to both zebra finch and chicken. After filtering (based on the length of alignment and removing frame shifts), the number of contigs was reduced to 3821. We excluded contigs that were either identical or which had $K_s = 0$ (which made $K_a/K_s$ incalculable). Thus, $K_a/K_s$ was estimated for 3252 (zebra finch) and 3127 (chicken) contigs. Rate estimation with zebra finch identified 43 contigs with $K_a/K_s \geq 1$ and 283 with values of 0.5–1.0 (Fig. 5A). Rate estimations with chicken yielded 5 and 58 contigs with $K_a/K_s \geq 1$ and between 0.5 and 1.0, respectively (Fig. 5B). Afterwards, assuming the song sparrow contigs have the same chromosome organization as zebra finch and chicken, the calculated ratios were organized into chromosomes (Table 3); this is not an unrealistic assumption considering the high degree of chromosomal conservation among avian genomes[28,29] and the fact that such a high proportion (85.1%) of our orthologous contigs was found to have shared chromosomal locations with zebra finch and chicken.

Although $K_a/K_s$ (sometimes calculated as $d_N/d_S$ or $\omega$) is commonly misinterpreted,[30] this ratio of rates of non-synonymous to synonymous substitutions can give some context to candidate genes and allows for subsequent hypothesis testing.[31,32] Data organized into chromosomes suggest that contigs may have undergone more selection with respect to the zebra finch than the chicken (as high $K_a/K_s$ values are typically interpreted, though see ref. 30).

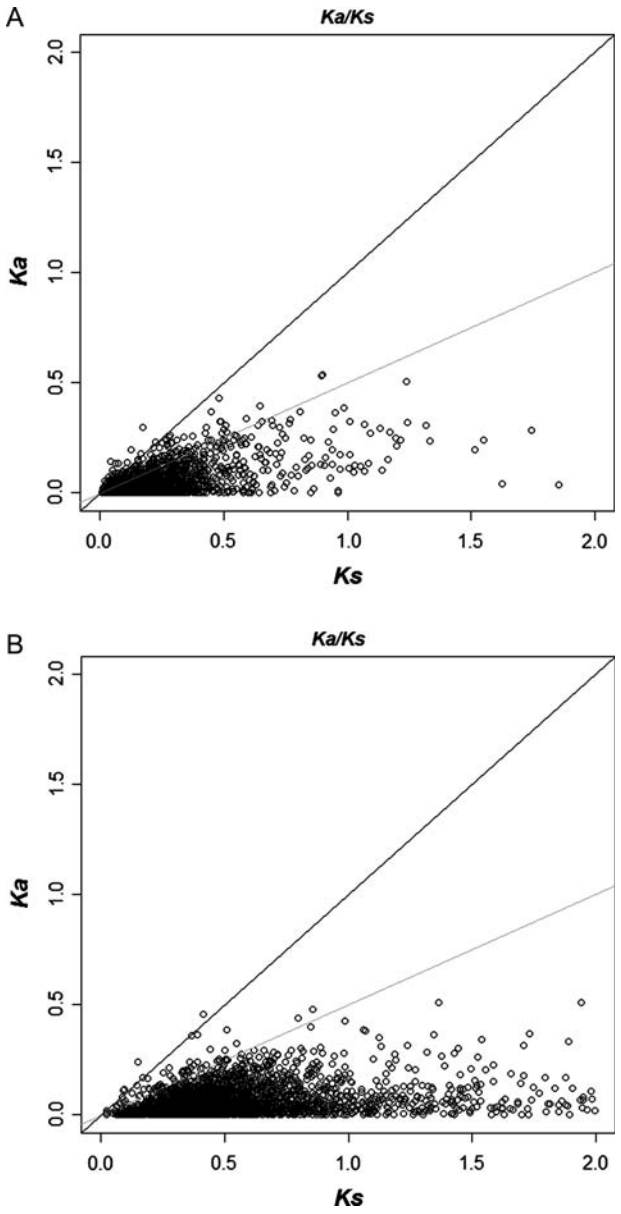


**Figure 5.** The distribution of $K_a/K_s$ ratio for the contigs orthologous to both zebra finch (A) and chicken (B). Contigs with $K_a/K_s$ values of 0.5–1.0 fall above the grey line and values >1.0 fall above the black line.

The fact that $K_a/K_s$ values were higher on average for the zebra finch than for the chicken (Table 3) is likely a methodological artefact. The zebra finch is in the same taxonomic order as the song sparrow (Passeriformes), whereas the chicken is taxonomically distant (Galliformes). Estimates of $\omega$ necessarily classify sites with differences as non-synonymous or synonymous, and errors in the estimation of either can profoundly affect the outcome of these analyses.[33]

**Table 3.** Number of contgis orthologous to particular zebra finch and chicken chromosomes, and mean $K_a/K_s$ ratio for each chromosome, assuming the orthologous contigs have the same chromosomal location as zebra finch and chicken

| Chr | Contigs orthologous to particular zebra finch chromosome | Total number of transcripts from particular zebra finch chromosome in Biomart file | $K_a/K_s$ (mean ± SD) | Contigs orthologous to particular chicken chromosome | Total number of transcripts from particular chicken chromosome in Biomart file | $K_a/K_s$ (mean ± SD) |
|---|---|---|---|---|---|---|
| 1 | 261 | 1124 | 0.2552 ± 0.2733 | 492 | 2994 | 0.1528 ± 0.1694 |
| 2 | 338 | 1345 | 0.2434 ± 0.2465 | 339 | 1995 | 0.1457 ± 0.1326 |
| 3 | 309 | 1169 | 0.2434 ± 0.2807 | 314 | 1672 | 0.1565 ± 0.1497 |
| 4 | 188 | 741 | 0.2258 ± 0.3347 | 252 | 1516 | 0.1374 ± 0.1274 |
| 5 | 229 | 936 | 0.2103 ± 0.2184 | 234 | 1299 | 0.1280 ± 0.1219 |
| 6 | 107 | 562 | 0.2447 ± 0.2112 | 106 | 781 | 0.1486 ± 0.1187 |
| 7 | 124 | 521 | 0.2220 ± 0.2103 | 120 | 767 | 0.1361 ± 0.1235 |
| 8 | 111 | 416 | 0.2581 ± 0.2196 | 127 | 723 | 0.1436 ± 0.1251 |
| 9 | 90 | 458 | 0.2286 ± 0.3839 | 86 | 598 | 0.1045 ± 0.1087 |
| 10 | 86 | 394 | 0.1784 ± 0.1738 | 90 | 599 | 0.1220 ± 0.1890 |
| 11 | 68 | 371 | 0.2330 ± 0.2978 | 61 | 499 | 0.1429 ± 0.1439 |
| 12 | 73 | 349 | 0.1799 ± 0.2206 | 68 | 427 | 0.1076 ± 0.1122 |
| 13 | 77 | 321 | 0.1845 ± 0.2319 | 83 | 499 | 0.0994 ± 0.1225 |
| 14 | 80 | 390 | 0.2541 ± 0.3448 | 79 | 578 | 0.1333 ± 0.1288 |
| 15 | 76 | 350 | 0.1817 ± 0.2299 | 73 | 531 | 0.0925 ± 0.1207 |
| 17 | 49 | 300 | 0.1705 ± 0.1597 | 46 | 432 | 0.0967 ± 0.0861 |
| 18 | 54 | 309 | 0.2230 ± 0.1950 | 55 | 428 | 0.1085 ± 0.0907 |
| 19 | 68 | 313 | 0.2004 ± 0.2982 | 66 | 443 | 0.0858 ± 0.0952 |
| 20 | 50 | 329 | 0.2419 ± 0.2444 | 51 | 476 | 0.1336 ± 0.1277 |
| 21 | 34 | 192 | 0.1470 ± 0.1569 | 44 | 346 | 0.0847 ± 0.1058 |
| 22 | 16 | 98 | 0.1000 ± 0.0976 | 11 | 160 | 0.0441 ± 0.0593 |
| 23 | 34 | 205 | 0.1783 ± 0.1828 | 33 | 288 | 0.0782 ± 0.0920 |
| 24 | 27 | 181 | 0.1961 ± 0.1906 | 24 | 270 | 0.1000 ± 0.0982 |
| 25 | 7 | 92 | 0.1161 ± 0.1069 | 6 | 169 | 0.0711 ± 0.1017 |
| 26 | 31 | 176 | 0.1148 ± 0.1081 | 29 | 341 | 0.0824 ± 0.0927 |
| 27 | 31 | 252 | 0.1471 ± 0.1438 | 28 | 345 | 0.0698 ± 0.0727 |
| 28 | 27 | 227 | 0.1102 ± 0.1256 | 23 | 284 | 0.0476 ± 0.0414 |
| Z | 149 | 745 | 0.2321 ± 0.2293 | 146 | 990 | 0.1381 ± 0.1174 |

Taxonomic or lineage distance (longer branches) will affect the reconstruction of synonymous substitution rates especially (through an expected increase in repeated mutations, or multiple hits), and we consider this to be a likely source of the consistent differences in apparent molecular selection between our song-sparrow-to-zebra-finch and song-sparrow-to-chicken contrasts (Table 3; see also ref. 34). Nevertheless, these contrasts are valuable in highlighting the chromosomal distributions (assuming chromosomal stability[28]) and relative values of $\omega$ between closer and more distant relatives of the song sparrow, providing insights into attributes of selection in the coding genome across these scales. Unfortunately, this approach is not valid within species.[35–37]

Chromosomes 22 and 26 showed the greatest differences between the zebra finch and the chicken in the percentage of song sparrow contigs mapped (relative to the number of genes available in the Biomart database for the zebra finch and chicken). Both of these chromosomes had significantly different frequencies of mapped-song-sparrow versus Biomart data-available genes between the zebra finch and the chicken ($G_{adj} = 4.4$, $P < 0.05$, and $G_{adj} = 6.9$, $P < 0.01$, respectively at 1 d.f., *G*-test with Williams' correction; Table 3). In both cases, proportionally more contigs were mapped to the zebra finch than to the

chicken given the sizes of the respective databases (Table 3).

### 3.8. Chromosomal distributions of between-subspecies SNPs/indels

Two findings emerged in comparing the among-chromosome locations (mapped against the zebra finch) of the between-subspecies SNPs/indels that were mapped to chromosomes (218 SNP/indel-bearing, between-subspecies song sparrow contigs; Supplementary Sheet 2) versus all orthologous song sparrow contigs (Table 3). First, the chromosomal distribution of the candidate loci was significantly different from the distribution of all orthologous contigs ($G_{adj} = 51.5$, 27 d.f., $P < 0.005$), indicative of a non-random process (e.g. selection). Importantly, the chromosomal distribution of the 199 unique, mappable SNP/indel-bearing contigs between Attu and Adak islands (within the subspecies *maxima*), where we expected drift rather than selection to be more pronounced, was not significantly different from the chromosomal distribution of all orthologous contigs ($G_{adj} = 35.1$, 27 d.f., $P > 0.1$). Secondly, the greatest differences in the distribution of between-subspecies candidate loci from the distribution of all contigs occurred among chromosomes 2, 5, and Z (where proportionally fewer SNP/indel-bearing contigs occurred than expected) and chromosomes 3 and 11 (where relatively more SNP/indel-bearing contigs occurred than expected).

Finally, in contrasting our between-subspecies results with those of our between-species comparisons above, we found that seven of the SNP/indel-bearing contigs between subspecies were also contigs that exhibited evidence suggestive of selection (high $K_a/K_s$ values) when compared with the zebra finch and the chicken. Each contig has one between-subspecies SNP, and the functions of these loci are variable (Supplementary Sheet 4). Three of these seven occurred on chromosome 3 and one on chromosome 11, where the between-subspecies contrasts suggested elevated levels of SNPs/indels. These contigs and their chromosomal locations may thus be important in songbird divergence, but we do not yet know why.

### 3.9. Summary

In summary, our analysis identified the major categories of song sparrow genes and orthologous loci between song sparrow/zebra finch and song sparrow/chicken. Substitution rate estimation yielded the fastest evolving loci, and some of the loci that were fixed between subspecies were also highlighted as possibly under selection between the song sparrow and the zebra finch. Although additional sequencing of these libraries and validation of within-species SNPs/indels in multiple populations and lineages is required, we consider that the loci described here will include some of broad utility for studying the genomics of songbird divergence.

**Acknowledgements:** We thank NSF Alaska EPSCoR (EPS-0701898) and the University of Alaska Museum for supporting this research. Jack Withrow assisted in fieldwork, Roger Nilsen made the normalized cDNA libraries, Jeff Wagner sequenced the libraries, and Christin Pruett and Erik Postma provided helpful comments. We are also grateful for support from the UGA ARCS foundation.

**Supplementary data:** Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

## Funding

This study was supported in part by resources and technical expertise from the University of Georgia, Georgia Advanced Computing Resource Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Office.

## References

1. Nadeau, N.J. and Jiggins, C.D. 2010, A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations, *Trends Genet.*, **26**, 484–92.
2. Stapley, J., Reger, J., Feulner, P.G., et al. 2010, Adaptation genomics: the next generation, *Trends Ecol. Evol.*, **25**, 705–12.
3. Bouck, A. and Vision, T. 2007, The molecular ecologist's guide to expressed sequence tags, *Mol. Ecol.*, **16**, 907–24.
4. Wheat, C.W. 2010, Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing, *Genetica*, **138**, 433–51.
5. Kunstner, A., Wolf, J.B., Backstrom, N., et al. 2010, Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species, *Mol. Ecol.*, **19** (Suppl. 1), 266–76.
6. Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W. and Buerkle, C.A. 2010, Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery, *BMC Genomics*, **11**, 180.
7. Santure, A.W., Gratten, J., Mossman, J.A., Sheldon, B.C. and Slate, J. 2011, Characterisation of the transcriptome of a wild great tit *Parus major* population by next generation sequencing, *BMC Genomics*, **12**, 283.
8. Wolf, J.B., Bayer, T., Haubold, B., Schilhabel, M., Rosenstiel, P. and Tautz, D. 2010, Nucleotide divergence vs. gene expression differentiation: comparative

transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow, *Mol. Ecol.*, **19** (Suppl. 1), 162–75.
9. Patten, M.A. and Pruett, C.L. 2009, The song sparrow, *Melospiza melodia*, as a ring species: patterns of geographic variation, a revision of subspecies, and implications for speciation, *Syst. Biodivers.*, **7**, 33–62.
10. Arcese, P., Sogge, M.K., Marr, A.B. and Patten, M.A. 2002, Song sparrow (*Melospiza melodia*), In: Poole, A. and Gill, F. (eds), *The Birds of North America*. The Birds of North America, Inc.: Philadelphia, PA, No. 704.
11. Beldade, P., Rudd, S., Gruber, J.D. and Long, A.D. 2006, A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model, *BMC Genomics*, **7**, 130.
12. Quinlan, A.R., Stewart, D.A., Stromberg, M.P. and Marth, G.T. 2008, Pyrobayes: an improved base caller for SNP discovery in pyrosequences, *Nat. Methods*, **5**, 179–81.
13. Chevreux, B., Pfisterer, T., Drescher, B., et al. 2004, Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs, *Genome Res.*, **14**, 1147–59.
14. Marth, G.T., Korf, I., Yandell, M.D., et al. 1999, A general approach to single-nucleotide polymorphism discovery, *Nat. Genet.*, **23**, 452–6.
15. Altenhoff, A.M. and Dessimoz, C. 2009, Phylogenetic and functional assessment of orthologs inference projects and methods, *Plos Comput. Biol.*, **5**, e1000262.
16. Kent, W.J. 2002, BLAT–the BLAST-like alignment tool, *Genome Res.*, **12**, 656–64.
17. Subramanian, S., Huynen, L., Millar, C.D. and Lambert, D.M. 2010, Next generation sequencing and analysis of a conserved transcriptome of New Zealand's kiwi, *BMC Evol. Biol.*, **10**, 387.
18. Conesa, A., Götz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.
19. Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
20. Higgins, D.G., Larkin, M.A., Blackshields, G., et al. 2007, Clustal W and clustal X version 2.0, *Bioinformatics*, **23**, 2947–8.
21. Suyama, M., Torrents, D. and Bork, P. 2006, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Res.*, **34**, W609–12.
22. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.
23. The Gene Ontology Consortium. 2000, Gene Ontology: tool for the unification of biology, *Nat. Genet.*, **25**, 25–9.
24. Huang, D.W., Sherman, B.T. and Lempicki, R.A. 2009, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, **4**, 44–57.
25. Lempicki, R.A., Huang, D.W. and Sherman, B.T. 2009, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res.*, **37**, 1–13.
26. De Boever, S., Vangestel, C., De Backer, P., Croubels, S. and Sys, S.U. 2008, Identification and validation of housekeeping genes as internal control for gene expression in an intravenous LPS inflammation model in chickens, *Vet. Immunol. Immunopathol.*, **122**, 312–7.
27. Ekblom, R., Balakrishnan, C.N., Burke, T. and Slate, J. 2010, Digital gene expression analysis of the zebra finch genome, *BMC Genomics*, **11**, 219.
28. Ellegren, H. 2010, Evolutionary stasis: the stable chromosomes of birds, *Trends Ecol. Evol.*, **25**, 283–91.
29. Griffin, D.K., Robertson, L.B.W., Tempest, H.G. and Skinner, B.M. 2007, The evolution of the avian genome as revealed by comparative molecular cytogenetics, *Cytogenet Genome. Res.*, **117**, 64–77.
30. Hughes, A.L. 2007, Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level, *Heredity*, **99**, 364–73.
31. Barreto, F.S., Moy, G.W. and Burton, R.S. 2011, Interpopulation patterns of divergence and selection across the transcriptome of the copepod *Tigriopus californicus*, *Mol. Ecol.*, **20**, 560–72.
32. Elmer, K.R., Fan, S., Gunter, H.M., et al. 2010, Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes, *Mol. Ecol.*, **19** (Suppl. 1), 197–211.
33. Yang, Z.H. and Bielawski, J.P. 2000, Statistical methods for detecting molecular adaptation, *Trends Ecol. Evol.*, **15**, 496–503.
34. Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G.H. and Graur, D. 2009, Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment, *Genome Biol. Evol.*, **1**, 114–18.
35. Kryazhimskiy, S. and Plotkin, J.B. 2008, The population genetics of dN/dS, *PLoS Genet.*, **4**, e1000304.
36. Rocha, E.P., Smith, J.M., Hurst, L.D., et al. 2006, Comparisons of dN/dS are time dependent for closely related bacterial genomes, *J. Theor. Biol.*, **239**, 226–35.
37. Wolf, J.B., Kunstner, A., Nam, K., Jakobsson, M. and Ellegren, H. 2009, Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection, *Genome Biol. Evol.*, **1**, 308–19.