A peer-reviewed version of this preprint was published in PeerJ on 22 May 2014.

Balakrishnan CN, Mukai M, Gonser RA, Wingfield JC, London SE, Tuttle EM, Clayton DF. 2014. Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior. PeerJ 2:e396 https://doi.org/10.7717/peerj.396

Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior

Christopher N Balakrishnan, Motoko Mukai, Rusty A Gonser, John C Wingfield, Sarah E London, Elaina M Tuttle, David F Clayton

Emberizid sparrows (emberizidae) have played a prominent role in the study of avian vocal communication and social behavior. We present here brain transcriptomes for three emberizid model systems, song sparrow *Melospiza melodia*, white-throated sparrow *Zonotrichia albicollis*, and Gambel's white-crowned sparrow *Zonotrichia leucophrys gambelii*. Each of the assemblies covered fully or in part, over 89% of the previously annotated protein coding genes in the zebra finch *Taeniopygia guttata*, with 16,846, 15,805, and 16,646 unique BLAST hits in song, white-throated and white-crowned sparrows, respectively. As in previous studies, we find tissue of origin (auditory forebrain versus hypothalamus and whole brain) as a primary determinant of overall expression profile. We also demonstrate the successful isolation of RNA and RNA-sequencing from *post-mortem* samples from building strikes and suggest that such an approach could be useful when traditional sampling opportunities are limited. These transcriptomes will be an important resource for the study of social behavior in birds and for data driven annotation of forthcoming whole genome sequences for these and other bird species.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.297v3 | CC-BY 4.0 Open Access | received: 25 Mar 2014, published: 25 Mar 2014

Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior

Christopher N. Balakrishnan[1, *], Motoko Mukai[2,3], Rusty A. Gonser,[4] John C. Wingfield[3], Sarah E. London[5], Elaina M. Tuttle[4], and David F. Clayton[6]

[1]Department of Biology, East Carolina University, Greenville, North Carolina, USA

[2] Department of Food Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, New York, USA

[3]Department of Neurobiology, Physiology and Behavior, University of California, Davis, California, USA

[4]Department of Biology and The Center for Genomic Advocacy (TCGA), Indiana State University, Terre Haute, Indiana, USA

[5]Department of Psychology, University of Chicago, Chicago, Illinois, USA

[6] Division of Biological & Experimental Psychology, School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

*Author for correspondence:

Christopher N. Balakrishnan

East Carolina University

Howell Science Complex

Greenville, NC 27858

balakrishnanc@ecu.edu

252 328 2910

Introduction

The comparative method, broadly speaking, is a powerful approach for understanding adaptations including behavior and central control of physiological responses to environmental change. Natural variation in behavior among species has been used in various taxonomic groups to begin to unravel the molecular underpinnings of animal social behavior. Among these comparative studies of behavior, different strategies and technologies have been deployed in order to gain an understanding of the proximate mechanisms at play. For example, experimental hormonal manipulations and gene sequence comparisons in different species of *Microtus* voles led to insights into the mechanisms of parental care (Young et al. 1999). Similarly, quantitative trait locus (QTL) mapping studies have recently revealed the genetic architecture of burrowing behavior in *Peromyscus* mice (Weber et al. 2013). Phylogenetic analyses of rates of molecular evolution based on transcriptomes in eusocial and solitary bees has also led to insights into potential underpinnings of social behavior variation (Woodard et al. 2011).

Songbirds, or oscine passerines, comprise roughly half of avian diversity and also serve as important models for the study of social behavior. Arguably the most prominent of the songbird species for behavioral research is the zebra finch *Taeniopygia guttata*, which now boasts a full suite of genomic and molecular tools including a complete genome sequence (Warren et al. 2010), RNA-seq based mRNA (Warren et al. 2010; Balakrishnan et al. 2012), microRNA data (Gunaratne et al. 2011; Luo et al. 2012), transgenics (Agate et al. 2009) and cell lines (Itoh & Arnold 2011; Balakrishnan et al. 2012). A key strength of songbirds as a model system, however, has always been the

behavioral complexity and diversity of songbirds as a group (Beecher & Brenowitz 2005; Brenowitz & Beecher 2005; Clayton et al. 2009).

Among songbirds, many comparative neurobiological studies have focused on three species of new world sparrows (emberizidae). Before the zebra finch assumed its role as a model system for vocal learning, Peter Marler and colleagues had demonstrated age-limited song learning and cultural transmission of song dialects in the white-crowned sparrow, *Zonotrichia leucophrys* (Marler & Tamura 1964). There is also a striking behavioral polymorphism in which some subspecies, such as Gambel's white-crowned sparrow *Z. l. gambelii*, are migratory, living in large non-territorial flocks during non-breeding seasons, whereas other subspecies are non-migratory and are territorial throughout the year (DeWolfe et al. 1989). White-throated sparrows *Zonotrichia albicollis* also show polymorphism in behavior but in this case, the polymorphism is known to be caused by a large chromosomal rearrangement on chromosome 2 (Thorneycroft 1966; Thorneycroft 1975). Tan morph individuals are homozygotic for the metacentric form of the chromosome whereas white morphs are almost always heterozygous. In addition to coloration, the two morphs differ in a suite of behaviors including increased aggression and promiscuity and decreased parental care in birds of the white morph (Knapton and Falls 1983, Collins & Houtman 1999; Tuttle 2003). Male song sparrows *Melospiza melodia* are distinctive in that they are territorial during both the breeding season (summer) and much of the non- breeding season (autumn and winter) (Wingfield & Hahn 1994; Mukai et al. 2009). Different hormonal mechanisms, however, appear to underlie this similar behavioral phenotype with increased plasma testosterone levels driving intensity and persistence of aggression during breeding, but not at other

times of year (Wingfield 1994; Wingfield & Soma 2002). With this comparative perspective in mind, we have generated brain transcriptomes for these three historically important emberizid songbird models for the study of social behavior: white-throated sparrow, Gambel's white-crowned sparrow, and song sparrow.

Methods

Sample Collection

Samples for each of the three species were collected for diverse research purposes of the laboratories involved, so sampling strategy for each species was unique. Animal procedures were approved by the Institutional Animal Care and Use Committees of the University of California, Davis (protocol 07-13208) and the University of Illinois (protocol 11062) and were conducted in accordance with the NIH Guide for the Principles of Animal Care.

White-throated Sparrow: During migration, white-throated sparrows and other birds are often killed in collisions with buildings. We took advantage of this unfortunate fact by collecting birds opportunistically following night migration and collision into McCormick Place, Chicago, IL. Birds that had been killed overnight were collected first thing in the morning beginning at dawn by David Willard, Collection Manager - Birds, Field Museum of Natural History, Chicago, IL. Specimens used in this study were collected during the spring migration in 2010. Each specimen was immediately vouchered at the Field Museum where measurements were taken and they were dissected to determine their sex. Whole brain tissue was stored in RNA-later (Ambion). Prior to analysis we determined the morph of each sampled bird using a modification of Michopoulous et al. (2007). For sequencing we used the brains

from 6 males, 3 white and 3 tan.

White-crowned sparrow: Gambelii's white-crowned sparrows (*Zonotrichia leucophrys gambelii*) were captured within the University of California, Davis campus in February 2008, using seed baited Potter traps, and their sexes were identified using published PCR methods (Griffiths et al. 1998). After two weeks of acclimation in captivity, males (n=12) were anesthetized with isoflurane, decapitated and whole hypothalamus was collected, and immediately frozen in liquid nitrogen. Fieldwork in California was conducted under US Fish and Wildlife permit (MB713321-0) and State of California permit (SC-004400).

Song sparrow: Seven male birds were captured in the field using song playbacks from behind a mist net. All the birds were captured between July and August 2011, from two locations in central Illinois: "Phillips Tract" (40 07' 54.74" N 88 08' 39.66" W) and Vermillion River Observatory (40 03' 50.79" N 87 33' 30.30" W). Immediately upon removal from the mist net, birds were decapitated. We then dissected auditory forebrain tissue (auditory lobule, or AL) which is a composite brain area including the caudomedial nidopallium (NCM), caudomedial mesopallium (CMM) and Field L and froze the specimens on dry ice. Flat skins of collected song sparrows have been accessioned in the Illinois Natural History Survey, Urbana Illinois. Fieldwork in Illinois was conducted under US Fish and Wildlife Service Permit SCCL-41077A.

RNA Extraction, Library Preparation and Sequencing

White-throated Sparrow and Song Sparrow: In order to broadly describe the brain-expressed transcriptome of the White-throated sparrow, we extracted RNA from whole

brain. We homogenized the entire brain in Tri-Reagent (Molecular Research Company) for RNA purification and extracted total RNA following manufacturers instructions. Total RNA was then DNase treated (Qiagen, Valencia CA) to remove any genomic DNA contamination, and the resulting RNA was further purified using Qiagen RNeasy columns. We assessed the purified total RNA for quality using an Agilent Bioanalyzer (Fig. 1). Library preparation and sequencing were done at the University of Illinois Roy J. Carver Biotechnology Center. Library preparation was done using Illumina TruSeq RNA Sample Prep Kit and manufacturer's protocols (Illumina, San Diego, CA). The six libraries were pooled in equimolar concentration and the pool was quantitated by qPCR. Sequencing was done in a single lane of an Illumina HiSeq 2000 using a TruSeq SBS sequencing kit version 3 and analyzed with Casava 1.8.2. The same basic procedure was used to sequence the song sparrow except for the fact that RNA was extracted from the dissected AL (rather than whole brain) tissue, and that samples from seven individuals were run in a single lane of paired end (rather than single end) sequencing.

White-crowned Sparrow: Total RNA was extracted from each hypothalamus using TRIzol reagent (Life Technologies, Carlsbad, CA) followed by RNA cleanup using Qiagen RNeasy Mini Kits. RNA samples were then pooled and run on Bioanalyzer for quality control (RIN = 8.5). This pooled RNA sample was used to generate a mRNA-seq library of 400 bp size with a mRNA-seq sample prep kit following manufacturer's protocol with slight modifications. Briefly, mRNA was isolated using oligo(dT), fragmented using divalent cations under elevated temperature, reverse transcribed into cDNA using random primers, modified and ligated with adapters. The resulting cDNA was run on an agarose gel, a band was excised at 400 bp and enriched with PCR. The

final library was validated using the Bioanalyzer and confirmed a distinct band at approximately 400 bp. Pair-end sequencing (100bp x 2) was performed by the Genome Center DNA Technologies Core at the University of California, Davis, using an Illumina HiSeq 2500.

Transcriptome Assembly, Annotation and Assessment

We checked overall sequence quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed reads using ConDeTriV2.2 (Smeds & Kunstner 2011). We used default settings for trimming except for the high quality (hq) threshold which was set to 20 and lfrac, the maximum fraction of reads with quality < 10, which was set to 0.2. The lfrac parameter allows for trimming, rather than complete removal, of reads with low quality ends.

We used the Trinity (version r20131110) assembler (Grabherr et al. 2011) to generate *de novo* assemblies for each species. For white throated sparrow we assembled the reads for the two color morphs both separately and combined. Assembling the reads separately was reasonable given evidence of sequence divergence within the inversion (Thomas et al. 2008) and assembling the reads together was reasonable to improve coverage outside such areas. We used default settings in Trinity besides those specific to our computing system (memory allocation, etc.). We used TransDecoder (included in the Trinity package) to identify open reading frames (ORFs) in our predicted transcripts.

We assessed the quality our assembly by estimating N50 and average transcript length. The shortcomings of such metrics for transcriptome assessment have been described (O'Neil and Emrich 2013) and we use it primarily to enable comparison with

previously published studies. Therefore, we also assessed 5' to 3' gene model coverage relative to annotated zebra finch genes (see details below) and quantified the number of transcripts containing both start and stop codons using the annotation information provided by TransDecoder ("type:complete" in the fastq header).

We used BLAST (Altschul et al. 1990) searches against a database of Ensembl (release 74) zebra finch transcripts to annotate our ORF-containing transcripts. Functional description of annotated transcripts was conducted using Gene Ontology, and statistical over and under representation was tested using CORNA software (Wu & Watson 2009) and Fisher's Exact Tests with *p* values adjusted for multiple testing (Benjamini & Hochberg 1995). For each assembly we tested our identified set of putative zebra finch orthologs relative to the full population of Ensembl transcripts.

Gene Expression and Read-Mapping Profiling

In order to compare read mapping and gene expression profiles across libraries, we mapped RNA-seq reads to the zebra finch whole genome assembly (2.3.4) using Stampy (Lunter & Goodson 2011) a read mapper tailored for divergent reads relative to the reference genome. We mapped reads for six individual white-throated sparrows, three song sparrows, and the pooled white-crowned sparrow using default settings but with the substitution rate set to 0.05 to accommodate sequence divergence. In addition, we mapped reads from previously published zebra finch auditory forebrain reads (Balakrishnan et al. 2012, GenBank Accession: SRX493920- SRX493922) using substitution rate = 0.01. The zebra finch data comprised three pools of 10 individuals each that had been collected on an Illumina Genome Analyser rather than HiSeq, and

processed with Illumina pipeline 1.6 rather than 1.8.

To quantify gene expression, we used htseq-count (Anders et al. 2014) and tallied reads relative to Ensembl gene models. Read counts were normalized using the regularized log transformation in DE-Seq2 (Anders & Huber 2010). Expression profiles were then visualized by Euclidean distance based clustering and principal components analysis (PCA) using heatmap.2 in the gplots R package, and the plotPCA function in DE-Seq2. We then also used the geneBody.py script within the RseqC package (Wang et al. 2012) to describe read coverage across gene models and to test specifically for a 3' bias in transcript coverage in *post-mortem* samples.

Results & Discussion

RNA extraction and sequencing

Despite collecting tissues for the white-throated sparrow opportunistically from building strikes, we were able to extract reasonably high quality RNA from all samples (Fig. 1). From a total of twelve samples, we selected a set of six (three per morph) with Bioanalyzer RNA integrity numbers (RIN) above 7 (10-083 (7.2), 10-092 (7.2), 10-093 (7.7) and 10-118 (8.5), 10-124 (8.0) and 10-308 (7.9). Samples for sequencing were also chosen such that tan and white morphs were collected at the same time of year (spring migration 2010). A consequence of this was that the chosen tan morph samples had higher average RINs than the white morph samples did. All of our RNA from the other two species were of good quality and met Illumina's standard QC benchmark of RIN > 8. All of our sequencing runs yielded high quality sequence data and after fairly stringent trimming, we retained over 89% of the initial nucleotides sequenced (Table 1). Raw RNA

seq reads have been deposited to the GenBank Short Read Archive under accession numbers SRX342288-SRX342293, SRX493875- SRX493882, and SRX493919.

Transcriptome Assembly and Annotation

In all of the assemblies we had a large number of transcripts (> 95,000) and open reading frame (ORF) containing transcripts (>54,000), exceeding the likely number of coding genes. These transcripts reflect a combination of partial transcripts, alternative isoforms, allelic variants, and noncoding transcripts. We were able to generate high quality transcriptomes based on N50 and average transcript length (Table 2). N50s for the assemblies ranged from 1,942 for the white morphed white-throated sparrow to 4,072 for the song sparrow. For the song sparrow, this is an improvement over a recent 454-based transcriptome (N50=482; Srivastava *et al.* 2012). As expected, N50 in general improved with increased sequencing depth (with paired end data sets benefitting from both the reads being paired and having more reads). One exception to this rule was in the white-throated sparrow, where combining reads from the two morphs actually generated a worse assembly in terms of N50 relative to the "Tan morph only" assembly. Tan morph individuals are homozygous for a large structural polymorphism spanning much of chromosome 2 whereas white morph individuals are heterozygous. Recombination within the inversion is suppressed, allowing genetic divergence in this region (Thomas et al. 2008), and potentially explaining the drop in N50. For the purposes of annotation of the white-throated sparrow we therefore used the two morph-specific assemblies, merging them after the assembly process.

Although N50s were generally high, the white-throated sparrow assemblies, which

were based on smaller, single-end datasets and *post-mortem* samples, had the lowest scores. This effect was even more dramatic when assemblies were assessed in terms of the number of complete transcripts possessing both a start and stop codon. White-crowned, song, and white-throated sparrow transcriptomes contained 115,515, 79,451, and 24,388 complete transcripts, respectively (Table 2).

For white-throated sparrow we were able to find predicted transcripts with significant blast hits to 15,805 zebra finch genes (89% of Ensembl annotated zebra finch genes), whereas for song sparrow we found 16,846 (94%) and White-crowned sparrow 16,646 (93%). Therefore, in terms of unique BLAST hits, the song and white-crowned assemblies were also better than that of the white-throated sparrows. All three assemblies, however, cover a large proportion of known genes and represent an improvement of over recent 454-based bird transcriptomes (e.g., violet-eared waxbill, 11,084 genes, Balakrishnan et al. 2013).

Gene Ontology (GO) representation in the three datasets overlapped greatly with eight GO categories significantly enriched and six categories underrepresented across all three species' datasets (Table 2). Gene Ontology categories "cytoplasm", "intracellular', "mitochondrion", "nucleic acid binding", "nucleolus", "protein binding", "protein phosphorylation" and "transferase activity" were enriched across all the three libraries. By contrast, "cytokine activity", "DNA integration", "extracellular region", "hormone activity", "immune response" and MCH Class I protein complex" were also all under-represented, reflecting in part the well-described pattern of limited immune activity, or "immune privilege" in the brain (Galea et al. 2007). As in previous studies of avian brain gene expression, however, we did see some evidence of expression of the MHC Class I

gene itself (Ekblom et al. 2010; Balakrishnan et al. 2013).

The white-throated sparrow yielded a larger number of statistically over- (29) and under-represented (47) GO categories in its transcriptome as compared to song sparrow (10 over- and 10 under-represented categories) and white-crowned sparrows (19 over- and 14 under-represented). All of the categories that were significantly enriched in white-throated sparrows trended in the same direction in all three species although some did not show statistical significance in other two species (often bordered on significance in all three). This set of GO terms included "olfactory receptor activity" (where observed/expected were 165/150 in white-throated sparrows, 165/156 in song sparrow, and 165/158 in white-crowned sparrow) out of a total of 168 annotated genes. This was notable as a previous 454-based whole brain transcriptome of another songbird did not detect any olfactory receptor genes at all (Balakrishnan *et al.* 2013). The detection of such genes here suggests that the increased sequencing depth provided by the Illumina platform has aided in this regard. Despite the generally tissue-restricted distribution of olfactory receptor expression, we were able to pick up these genes in all of our tissue samples irrespective of the brain region targeted. High depth RNA-sequencing data including those presented here will therefore be useful for annotating these diverse olfactory receptor transcripts.

Thirteen GO categories were significantly under-represented in white-throated sparrows but not in either of the other two sparrows (Table 3). Among these categories, there appeared to be a qualitative difference in gene expression and resultant GO representation. Gene ontology categories associated with brain function (visual function, G-protein coupled receptor activity, and neurotransmitter transport) were all under-

represented in white-throated sparrow but not the others. This difference in GO category representation likely reflects the fact that RNA was preserved *post-mortem.* Alternatively, the difference could be attributed either to differences in brain region (whole brain versus forebrain) or physiological condition (spring migration versus breeding condition versus captive/wintering).

Expression profiling relative to zebra finch gene models

Using the Stampy read mapper we were able to map between 82% and 94% of sparrow and zebra finch reads to the zebra finch genome. White-throated sparrow reads mapped at a lower rate (average = 84% of reads mapped) than the white-crowned sparrow (93%) and zebra finch (93%) data. Among the reads that did map to the genome, however, all of the species showed a similar profile, with a large proportion of reads (53.2 +/- 3.6%) mapping outside of currently defined zebra finch genes and suggesting extensive transcription outside of known genes.

As in previous analyses of coordinated microarray studies in songbirds (Replogle et al. 2008; Drnevich et al. 2012), we find a major effect of brain region on overall expression profile. Clustering of normalized expression profiles revealed that samples taken from the auditory forebrain, those from song sparrow and previously published zebra finch data, clustered closely together (Fig. 3). After the two auditory forebrain samples, the next most similar in profile was from the white-crowned sparrow hypothalamus, another forebrain region. Tissue of origin therefore appears to have a major effect of overall expression profile overriding the expected biological effects of phylogeny and the technical effects sequencing platform and lab-specific protocols (see

above). If phylogeny were the dominant contributor to expression profile, white-crowned and white-throated sparrows would be most similar, with zebra finch forming the most divergent lineage. The six whole brain white-throated sparrow libraries were the most divergent in profile, suggesting that inclusion of non-forebrain yielded altered expression for a large number of genes. We did not conduct statistical tests of differential gene expression due to multiple confounding variables, namely tissue, sequencing platform, and independent tissue collection and library preparation. Both euclidean distance-based clustering and PCA also highlight the fact that zebra finches, which were sacrificed in captivity and sequenced in pools of ten had much reduced variance in expression profile relative to our non-pooled, field-collected white-throated sparrow and song sparrow samples.

Based on read mapping to the zebra finch we were also able to assess coverage of annotated genes. This was important given our *post-mortem* sampling of white-throated sparrows. RNA quality as measured by RIN was only slightly lower in white-throated sparrow samples and we found that 3' bias was similar across all of our samples including those collected *post-mortem* (Fig. 2). This finding suggests that RNA degradation in these samples may not be the primary factor associated with the lower assembly metrics in the white-throated sparrow assembly. Rather, it's likely that the smaller single-end dataset used for the white-throated sparrow reduced our power for transcript assembly.

Cheviron et al. (2011) documented the time course of RNA degradation *post-mortem*, and also suggest that such samples can provide a useful source of RNA, even though such specimens are often overlooked. Similarly, a recent RNA-sequencing study

of pinnipeds successfully used *post-mortem* samples (Hoffman et al. 2013). Although clearly not an ideal strategy for studies aimed at quantifying gene expression, the use of recently *post-mortem* samples is viable strategy for initial transcriptome description, and in our study gave access to a large portion of the transcriptome. This approach could be particularly useful for rare species where collection of fresh specimens is impossible.

Conclusion

Transcriptome assemblies are a valuable resource, particularly for species without reference genomes, providing access to a large proportion of the coding and noncoding expressed genome. For taxa with genomes, or with genomes in progress, transcriptome data provides empirical (as opposed to model based) information on transcript structures including alternative isoforms that are not well-annotated in most species. We have presented here neuro-transcriptomic data for three important model species for the study of social behavior and neurobiology building on a growing body of such data (e.g., Balakrishnan et al. 2013, MacManes & Lacey 2012; Moghadam *et al.* 2013).

Acknowledgments

Thanks to Matt MacManes for helpful feedback on the preprint version of this paper. David Willard (Collection Manager – Birds, Field Museum of Natural History, Chicago, IL) collected and provided access to white-throated sparrow tissues used in this study. Antonio Celis Murillo provided invaluable assistance with fieldwork on song sparrows in Illinois.

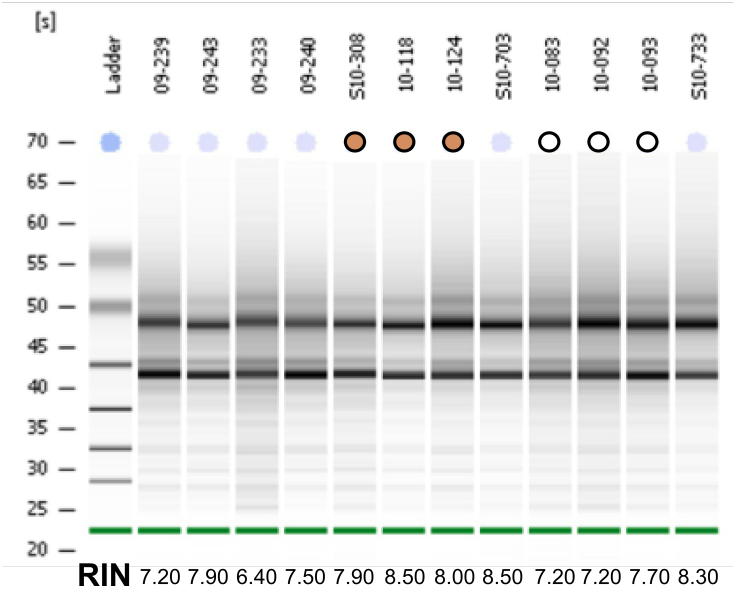
347 **Figure 1**. Bioanalyzer gel image showing RNA extracted from 12 white-throated

348 sparrows sampled *post-mortem*. RNA integrity numbers (RIN) are given at the bottom

349 and ranged from 6.4 to 8.5. Samples chosen for sequencing are indicated by tan and white

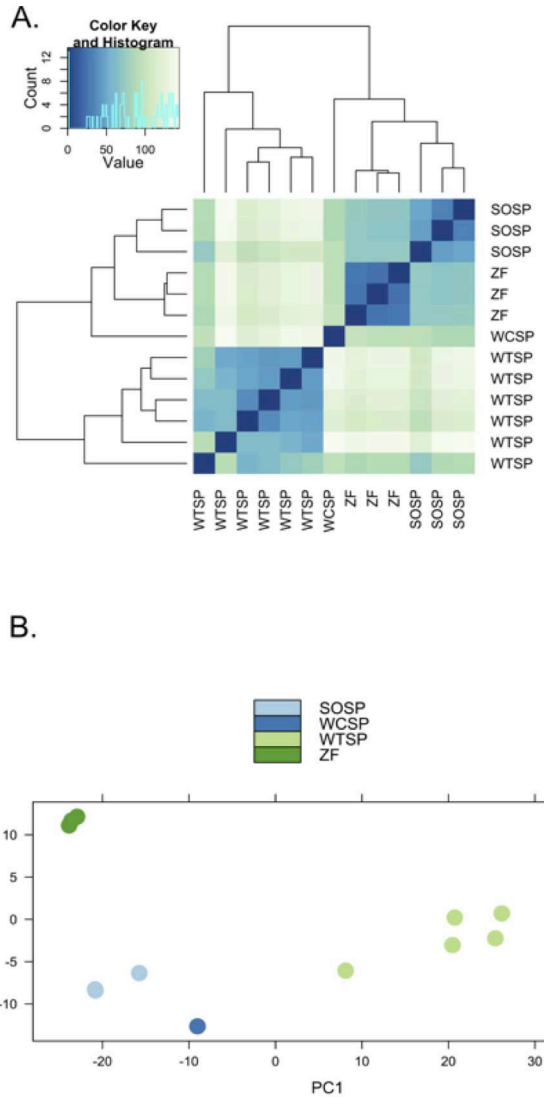
350 circles, representing tan and white morph sparrows, respectively.

Figure 1

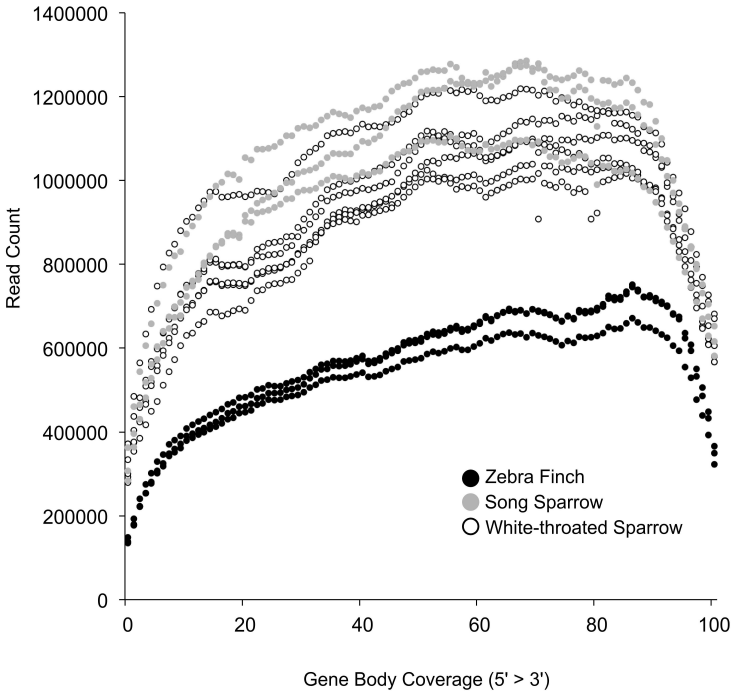


351

Figure 2. A) Hierarchical clustering and B) Principal components analysis of expression profiles for six white-throated sparrow (WTSP), three song sparrow (SOSP), three zebra finch (ZF) and one white-crowned sparrow libraries. Libraries derived from auditory lobule (AL) tissue cluster (SOSP and ZF) to the exclusion of the others. White-throated sparrow samples, taken from whole brain (rather than forebrain as the other samples are) show divergent and variable profiles. Zebra Finch (ZF) samples collected in captivity and generated from pools of 10 individuals, show much reduced sample variability.



370 **Figure 3.** Gene model coverage across all genes based on mapping of reads to the zebra
371 finch genome. Samples collected *post-mortem* from white-throated sparrow show a
372 similar gene coverage profile to freshly collected samples. Zebra finch data included
373 fewer total reads, explaining the lower depth across genes.



374

375

Table 1. Raw number of reads and bases before and after trimming with ConDeTri.

Species	Reads Before	Bases Before	Paired Reads After	Paired Read Bases After	Single Reads After	Single Read Bases After
WTSP-Tan	99,374,744	9,937,474,400	NA	NA	97,162,587	9,014,814,467
WTSP-White	97,605,312	9,760,531,200	NA	NA	95,347,015	8,779,352,471
SOSP-Paired	271,249,550	27,124,855,000	245,289,038	23,613,455,033	11,228,223	992,474,010
WCSP-Paired	160,229,712	16,022,971,200	153,636,836	14,171,465,431	2,871,235	213,815,184

376
377

Table 2. Tissue of origin, pool size, assembly statistics (N50, average transcript length, number of transcripts) and annotation description (number of zebra finch genes with significant BLAST hit) for whole assembly and open reading frame (ORF) containing transcripts. “Complete Transcripts” are those containing both a start and stop codon. We used the individual tan and white morph assemblies in the subsequent BLAST search and annotation which yielded 15,805 genes.

Species	Tissue	pool size	N50	Mean Length	# Transcripts	# ORF	Complete Transcripts	ZF genes
WTSP-Tan	Whole Brain	3	2,557	1,119	116,894	54,868	22,799	-
WTSP-White	Whole Brain	3	1,942	960	95,129	37,910	11,855	-
WTSP-Both	Whole Brain	6	2,284	982	149,184	58,284	24,388	15,805
SOSP	Auditory Forebrain	7	4,072	1,416	276,670	133,740	79,451	16,864
WCSP	Hypothalamus	12	3,415	1,591	307,617	206,926	115,515	16,646

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.297v3 | CC-BY 4.0 Open Access | received: 25 Mar 2014, published: 25 Mar 2014

Table 3. Gene Ontology categories significantly A) over- and B) under-represented in song (SOSP), white-crowned (WCSP) and white-throated (WTSP) sparrows (observed/ expected, FDR adjusted Fisher's Exact Test, $p < 0.05$).

A.

GO Category	SOSP	WCSP	WTSP
cytoplasm	1810/1739	1793/1718	1751/1650
intracellular	1629/1575	1632/1555	1577/1494
mitochondrion	790/753	788/744	781/715
nucleic acid binding	935/903	935/892	900/857
nucleolus	244/231	243/229	241/220
protein binding	5298/5218	5258/5154	5037/4951
protein phosphorylation	558/539	558/532	542/511
transferase activity, transferring phosphorous containing groups	538/519	538/513	522/493

B.

GO Category	SOSP	WCSP	WTSP
cytokine activity	43/58	40/58	37/55
DNA integration	8/13	7/13	4/12
extracellular region	263/320	264/316	238/303
hormone activity	31/43	32/43	26/41
immune response	68/88	61/87	57/84
MHC Class I protein complex	3/8	2/7	2/7

380
381

Table 4. GO terms underrepresented in *post-mortem* white-throated sparrow samples (observed/expected, adjusted $p < 0.01$), but not in song sparrow and white-crowned sparrow (adjusted $p > 0.05$).

GO Category	WTSP	WCSP	SOSP
photoreceptor activity	3/12	10/13	9/13
protein-chromophore linkage	3/12	10/13	9/13
visual perception	7/18	16/19	15/19
response to stimulus	7/17	14/18	13/18
G-protein coupled receptor activity	345/381	391/397	389/402
G-protein coupled purinergic nucleotide receptor activity	11/21	18/22	18/23
G-protein coupled purinergic nucleotide receptor signaling pathway	11/21	18/22	18/23
transporter activity	136/157	153/164	157/166
receptor activity	497/532	552/554	551/561
G-protein coupled receptor signaling pathway	463/496	513/517	514/523
integral to membrane	1564/1617	1683/1687	1692/1704
neurotransmitter transport	16/24	23/25	21/25

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.297v3 | CC-BY 4.0 Open Access | received: 25 Mar 2014, published: 25 Mar 2014

References

Agate RJ, Scott BB, Haripal B, Lois C, Nottebohm F (2009) Transgenic songbirds offer an opportunity to develop a genetic model for vocal learning. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 17963–17967.

Altschul SF, Gish W, Miller W, Myers EW, LIPMAN DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Anders S, Pyl PT, Huber W (2014) *HTSeq — A Python framework to work with high-throughput sequencing data*. *bioRxiv preprint*.

Balakrishnan CN, Lin Y-C, London SE, Clayton DF (2012) RNA-seq transcriptome analysis of male and female zebra finch cell lines. *Genomics,* **100**, 363–369.

Balakrishnan C, N., Chapus C, Brewer M, S., Clayton D, F. (2013) Brain transcriptome of the violet-eared waxbill *Uraeginthus granatina* and recent evolution in the songbird genome. *Open Biology*, **3**, 130063.

Beecher MD, Brenowitz EA (2005) Functional aspects of song learning in songbirds. *Trends in Ecology & Evolution*, **20**, 143–149.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289–300.

Brenowitz EA, Beecher MD (2005) Song learning in birds: diversity and plasticity, opportunities and challenges. *Trends in Neurosciences* **28**, 127–132.

406 Cheviron ZA, Carling MD, Brumfield RT (2011) Effects of postmortem interval and
407 preservation method on rna isolated from field-preserved avian tissues. *Condor*, **113**,
408 483–489.

409 Clayton D, F., Balakrishnan C, N., London S, E. (2009) Integrating Genomes, Brain and
410 Behavior in the Study of Songbirds. *Current Biology*, **19**, R865–R873.

411 Collins CE, Houtman AM (1999) Tan and white color morphs of White-throated
412 Sparrows differ in their non-song vocal responses to territorial intrusion. *Condor*, **101**,
413 842–845.

414 DeWolfe BB, Baptista LF, Petrinovich L (1989) Song development and territory
415 establishment in Nuttals White-Crowned Sparrows. *Condor*, **91**, 397–407.

416 Drnevich J, Replogle KL, Lovell P et al. (2012) Impact of experience-dependent and -
417 independent factors on gene expression in songbird brain. *Proceedings of the National*
418 *Academy of Sciences of the United States of America*, **109**, 17245–17252.

419 Ekblom R, Balakrishnan C, N., Burke T, Slate J (2010) Digital gene expression analysis
420 of the zebra finch genome. *BMC Genomics*, **11**, 219.

421 Galea I, Bechmann I, Perry VH (2007) What is immune privilege (not)? *Trends in*
422 *Immunology*, **28**, 12–18.

423 Goodson JL, Kelly AM, Kingsbury MA, Thompson RR (2012) An aggression-specific
424 cell type in the anterior hypothalamus of finches. *Proceedings of the National*
425 *Academy of Sciences of the United States of America*, **109**, 13847–13852.

426 Goodson JL, Wang YW (2006) Valence-sensitive neurons exhibit divergent functional
427 profiles in gregarious and asocial species. *Proceedings of the National Academy of*

428 *Sciences of the United States of America*, **103**, 17013–17017.

429 Grabherr MG, Haas BJ, Yassour M et al. (2011) Full-length transcriptome assembly from
430 RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–U130.

431 Griffiths R, Double MC, Orr K, Dawson RJG (1998) A DNA test to sex most birds.
432 *Molecular Ecology*, **7**, 1071–1075.

433 Gunaratne PH, Lin YC, Benham AL et al. (2011) Song exposure regulates known and
434 novel microRNAs in the zebra finch auditory forebrain. *BMC Genomics*, **12**, 277.

435 Hoffman JI, Thorne MAS, Trathan PN, Forcada J (2013) Transcriptome of the dead:
436 characterisation of immune genes and marker development from necropsy samples in
437 a free-ranging marine mammal. *BMC Genomics*, **14**, 52.

438 Itoh Y, Arnold AP (2011) Zebra finch cell lines from naturally occurring tumors. *In Vitro*
439 *Cellular & Developmental Biology-Animal*, **47**, 280–282.

440 Knapton, R.W. & Falls, J.B. 1983. Differences in parental contribution among pair types
441 in the polymorphic white-throated sparrow. *Canadian Journal of Zoology*. 61: 1288-
442 1292.

443 Lunter G, Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast
444 mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

445 Luo GZ, Hafner M, Shi ZM et al. (2012) Genome-wide annotation and analysis of zebra
446 finch microRNA repertoire reveal sex-biased expression. *BMC Genomics*, **13**, 727.

447 MacManes MD, Lacey EA (2012) The Social Brain: Transcriptome Assembly and
448 Characterization of the Hippocampus from a Social Subterranean Rodent, the Colonial
449 Tuco-Tuco (Ctenomys sociabilis). *PLoS One*, **7**, e45524.

450 Marler P, Tamura M (1964) Culturally transmitted patterns of vocal behavior in
451 sparrows. *Science*, **146**, 1483–148.

452 Michopoulos, V. Maney, D.L., Morehouse, C.B. & Thomas, J.W. 2007. A genotyping
453 assay to determine plumage morph in the White-throated Sparrow (*Zonotrichia*
454 *albicollis*). *The Auk* 124 No. 4 1330-1335.

455 Moghadam HK, Harrison PW, Zachar G, Szekely T, Mank JE (2013) The plover
456 neurotranscriptome assembly: transcriptomic analysis in an ecological model species
457 without a reference genome. *Molecular Ecology Resources*, **13**, 696–705.

458 Mukai M, Replogle K, Drnevich J et al. (2009) Seasonal Differences of Gene Expression
459 Profiles in Song Sparrow (*Melospiza melodia*) Hypothalamus in Relation to Territorial
460 Aggression. *PLoS One*, **4**, e8182.

461 O'Neil ST, Emrich SJ (2013) Assesing *De Novo* transcriptome assembly metrics for
462 consistency and utility. *BMC Genomics,* **14**, 465.

463 Replogle K, Arnold AP, Ball GF et al. (2008) The Songbird Neurogenomics (SoNG)
464 Initiative: Community-based tools and strategies for study of brain gene function and
465 evolution. *BMC Genomics*, **9**, 131.

466 Smeds L, Kunstner A (2011) CONDETRI - A Content Dependent Read Trimmer for
467 Illumina Data. *PLoS One*, **6**, e26314.

468 Srivastava A, Winker K, Shaw TI, Jones KL, Glenn TC (2012) Transcriptome Analysis
469 of a North American Songbird, *Melospiza melodia*. *DNA Research*, **19**, 325–333.

470 Thomas J, W., Caceres M, Lowman J, J. et al. (2008) The chromosomal polymorphism
471 linked to variation in social behavior in the white-throated sparrow (*Zonotrichia*

albicollis) is a complex rearrangement and suppressor of recombination. *GENETICS*, **179**, 1455–1468.

Thorneycroft HB (1966) Chromosomal polymorphism in white-throated sparrow *Zonotrichia albicollis* (Gmelin). *Science*, **154**, 1571–157.

Thorneycroft HB (1975) Cytogenetic study of white-throated sparrow, *Zonotrichia albicollis* (Gmelin). *Evolution*, **29**, 611–621.

Tuttle EM (2003) Alternative reproductive strategies in the white-throated sparrow: behavioral and genetic evidence. *Behavioral Ecology*, **14**, 425–432.

Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.

Warren W, C., Clayton D, F., Ellegren H et al. (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Weber JN, Peterson BK, Hoekstra HE (2013) Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice. *Nature*, **493**, 402–U145.

Wingfield JC (1994) Regulation of territorial behavior in the sedentary song sparrow, *Melospiza melodia morphna*. *Hormones and Behavior*, **28**, 1–15.

Wingfield JC, Hahn TP (1994) Testosterone and territorial behavior in sedentary and migratory sparrows. *Animal Behaviour*, **47**, 77–89.

Wingfield JC, Soma KK (2002) Spring and autumn territoriality in song sparrows: Same behavior, different mechanisms? *Integrative and Comparative Biology*, **42**, 11–20.

Woodard SH, Fischman BJ, Venkat A et al. (2011) Genes involved in convergent evolution of eusociality in bees. *Proceedings of the National Academy of Sciences of*

494 *the United States of America*, **108**, 7472–7477.

495 Wu X, Watson M (2009) CORNA: testing gene lists for regulation by microRNAs.

496 *Bioinformatics*, **25**, 832–833.

497 Young LJ, Nilsen R, Waymire KG, MacGregor GR, Insel TR (1999) Increased affiliative

498 response to vasopressin in mice expressing the V-1a receptor from a monogamous

499 vole. *Nature,* **400**, 766–768.