UNIVERSITY OF BIRMINGHAM

Research at Birmingham

De novo transcriptome sequencing in a songbird, the dark-eyed junco (Junco hyemalis)

Peterson, Mark P; Whittaker, Danielle J; Ambreth, Shruthi; Sureshchandra, Suhas; Buechlein, Aaron; Podicheti, Ram; Choi, Jeong-Hyeon; Lai, Zhao; Mockatis, Keithanne; Colbourne, John; Tang, Haixu; Ketterson, Ellen D

Link to publication on Research at Birmingham portal

RESEARCH ARTICLE **Open Access**

De novo transcriptome sequencing in a songbird, the dark-eyed junco (Junco hyemalis): genomic tools for an ecological model system

Mark P Peterson[1*], Danielle J Whittaker[1,2], Shruthi Ambreth[3], Suhas Sureshchandra[3], Aaron Buechlein[3], Ram Podicheti[3], Jeong-Hyeon Choi[3,4], Zhao Lai[3,5], Keithanne Mockatis[3], John Colbourne[3], Haixu Tang[3] and Ellen D Ketterson[1]

Abstract

Background: Though genomic-level data are becoming widely available, many of the metazoan species sequenced are laboratory systems whose natural history is not well documented. In contrast, the wide array of species with very well-characterized natural history have, until recently, lacked genomics tools. It is now possible to address significant evolutionary genomics questions by applying high-throughput sequencing to discover the majority of genes for ecologically tractable species, and by subsequently developing microarray platforms from which to investigate gene regulatory networks that function in natural systems. We used GS-FLX Titanium Sequencing (Roche/454-Sequencing) of two normalized libraries of pooled RNA samples to characterize a transcriptome of the dark-eyed junco (Junco hyemalis), a North American sparrow that is a classically studied species in the fields of photoperiodism, speciation, and hormone-mediated behavior.

Results: From a broad pool of RNA sampled from tissues throughout the body of a male and a female junco, we sequenced a total of 434 million nucleotides from 1.17 million reads that were assembled de novo into 31,379 putative transcripts representing 22,765 gene sets covering 35.8 million nucleotides with 12-fold average depth of coverage. Annotation of roughly half of the putative genes was accomplished using sequence similarity, and expression was confirmed for the majority with a preliminary microarray analysis. Of 716 core bilaterian genes, 646 (90 %) were recovered within our characterized gene set. Gene Ontology, orthoDB orthology groups, and KEGG Pathway annotation provide further functional information about the sequences, and 25,781 potential SNPs were identified.

Conclusions: The extensive sequence information returned by this effort adds to the growing store of genomic data on diverse species. The extent of coverage and annotation achieved and confirmation of expression, show that transcriptome sequencing provides useful information for ecological model systems that have historically lacked genomic tools. The junco-specific microarray developed here is allowing investigations of gene expression responses to environmental and hormonal manipulations – extending the historic work on natural history and hormone-mediated phenotypes in this system.

Keywords: Transcriptome, Aves, pyrosequencing, microarray, Junco, 454 titanium cDNA sequencing, single nucleotide polymorphism.

* Correspondence: petersmp@indiana.edu
[1]Dept. of Biology, Center for Integrated Study of Animal Behavior, Indiana University, Bloomington, IN, USA
Full list of author information is available at the end of the article

Background

Studies of natural populations lie at the core of understanding the evolution of complex, ecologically relevant phenotypes. High-throughput approaches to the study of gene functions have accelerated discoveries of the genetic underpinnings of many traits in model organisms, but until recently organisms with well-understood ecology typically lacked sophisticated genomic tools.

Model laboratory systems have contributed enormously to the understanding of genetics, gene expression, and the functional interactions of genes; however, the ecological relevance of these findings must also be studied within outbred populations responding to natural environmental challenges. By combining the power of natural systems with similar sets of genomics tools developed in laboratory systems, a deeper understanding of the molecular basis of adaptive traits and the mechanisms of biodiversity is achieved in nature. For example, the repeated loss of armor in Alaskan stickleback populations is due to a parallel regulatory mutation [1], and the divergence of cichlid visual systems appears to be driven by changes in expression of opsins, rather than structural changes [2]. Neither of these findings would have been possible without combining the extensive genomic and natural history information available in these systems.

The introduction of high-throughput sequencing technologies has led to significant declines in the time and cost required to generate genomic tools for functional studies. For example, 347 new genomes were published in 2010 alone (NCBI Genome Database, accessed May 2011). However, these genomes are still dominated by invertebrates and prokaryotes. Only 6 of the 347 sequenced genomes were of vertebrates, reflecting the high cost of sequencing large and complex vertebrate genomes. In contrast, transcriptome sequencing – the sequence of all transcripts present in a single cell type, tissue type, or entire organism under defined conditions – has emerged as a cost-effective means of rapidly acquiring functional sequence information for non-model systems [3]. Comprehensive transcriptomes have recently been characterized for several well-studied natural animal species including the Glanville fritillary butterfly [4], staghorn coral [5], horned beetle [6], garter snake [7], great tit [8], and Asian tiger mosquito [9]. These transcriptomes provide species-specific genomic information needed to employ genomic approaches in natural systems where the tools were previously lacking, but without the prohibitive costs and time required for sequencing of a full genome. For example, species-specific microarrays can be developed from transcriptome sequence to assess gene expression in natural populations [6], an approach we employ here. The acquisition of genomic information regarding expressed sequences is a rapid method for identification of meaningful genetic

divergence between species (e.g. [10]) and the early gene-expression divergence [11] that is thought to play a major role in speciation [12,13]. Here, we seek to add to this growing store of genomic information.

The dark-eyed junco (*Junco hyemalis*) is a classic avian system that has been extensively studied for more than a century and will be made even more useful with functional genomics. This seasonally-breeding North American sparrow was the first vertebrate animal in which the effect of photoperiod on seasonality was demonstrated [14,15], and its behavior, ecology, and physiology have been extensively studied [16], as has its tendency to diverge phenotypically and genetically across its geographic range [17,18]. One population has been monitored for 30 years and has allowed for assessment of temporal and individual variation in hormone levels, parental behavior, extra-pair mating, and breeding phenology [19,20]. This population has also emerged as a model for phenotypic engineering [21-23]: over 20 studies involving the manipulation of the hormonal phenotypes of free-living individuals have made it possible to relate hormonal variation to variation in phenotype (e.g., [24,25]).

The addition of transcriptome information will allow for deeper understanding of the mechanisms behind variation in behavior and physiology as well as how natural selection acts on that variation. Further, the junco offers unique opportunities to study rapid evolutionary divergence at the level of populations. The dark-eyed junco consists of five morphologically distinct subspecific groups that are thought to have diverged over the past 10,000 years [18]. In addition, approximately 30 years ago a population of juncos colonized the city of San Diego and has undergone rapid phenotypic divergence from its ancestral population in physiology and behavior [26-31]. Previous attempts to reveal the phylogenetic relationships among the groups of juncos have been frustrated by the lack of genetic differentiation owing to recent divergence and a paucity of genetic markers [18]. Genomic-level information will open new lines of research in this natural system including allowing gene-expression analysis, targeted re-sequencing, and identification of genes recently under selection, each of which will allow greater insight into the evolution and mechanisms of hormone-mediated phenotypes and natural breeding biology.

We sequenced a transcriptome of this ecologically and evolutionarily well-characterized species, returning substantial sequence diversity and expression information from a relatively shallow sequence coverage depth of long reads. In this report, we describe our approach and demonstrate the utility of these data in defining genes, identifying potential sequence variants, and confirming expression with a custom microarray.

Results and Discussion

Sampling and sequencing

RNA was extracted from 14 tissues (see Methods for full list) from one male and one female adult dark-eyed junco (*Junco hyemalis*) and prepared into normalized sequencing libraries (see Methods for details). Two pools, one from each individual, were sequenced in parallel using GS-FLX Titanium pyrosequencing (Roche/454 Sequencing), yielding 1.17 million reads totaling 434 million nucleotides with a mean length of 372 base pairs (bp) after adaptor trimming. The quality-filtered (cleaned) reads have been deposited in the Sequence Reads Archive (SRA) Database under accession numbers SRX144177.1 and SRX144176.3.

Sampling decisions are a critical question for transcriptome projects as the genes recovered are directly related to the tissues, individuals, and states that are chosen for inclusion. Research in the junco has historically focused on adults and their breeding behavior. The use of only two adult individuals limits the identification of some classes of genes, such as the developmental genes expressed in growing young; however, the decision also improves our ability to confidently assemble sequences by reducing concerns about integrating allelic variation. Future interest in other conditions, or in the identification of more extensive polymorphic markers for comparing populations, will be able to use this assembly as a reference to improve and focus effort on the condition of interest. In other systems, the relative importance of these questions have led to some projects sequencing only developing individuals [32], across multiple ages (e.g. [6]), across multiple eco-types (e.g. [7]), only specific tissues(e. g. [33,34]), or simply more individuals (e.g. [8]). In the junco system, this initial approach provides, in our opinion, the most widely valuable research tools for currently anticipated applications and the strongest base from which to launch future projects.

Assembly

Sequencing reads were assembled using GS *de novo* Assembler (NEWBLER v2.3; Roche), resulting in 40,564 contigs assembled from 828,612 reads (71% of total, 83% of cleaned reads) covering 35.8 million bases of sequence with an average length of 884 bp and 12-fold average coverage with 166,177 singletons remaining unassembled (Table 1). The assembled contigs have been deposited in the NCBI Transcriptome Shotgun Assembly (TSA) Database under accession numbers JV157086-JV188856. This assembly is similar to other *de novo* transcriptome assemblies, which have been characterized by 40,000 to 50,000 contigs with 63% to 90% of reads assembled [4-8,34]. Variation between individuals and alleles can artificially break contigs, as can alternative splicing [7], so further assembly was required to more

Table 1 Sequencing and assembly statistics

	Number	Length	Average Length
Reads	1.17 million	434 million	372
Contigs	40,564	35.8 million	884
Singletons	166,177	57.1 million	344
Isotigs	31,739	NA	1,248
Isogroups	22,765	NA	NA

Sequencing reads were assembled into contigs which were further grouped into isotigs and isogroups (see text for details).

accurately estimate the number of unique genes in this transcriptome.

We further combined contigs into groups based on shared broken reads in the initial assembly following manufacturer's directions (Genome Sequencer FLX System Software Manual, version 2.3, Roche; see Methods for more detail). Briefly, many contigs are broken apart by the assembly software due to variability that can be induced by gene duplications, splice variants or even allelic variation [7]. The GS mapper assembly software stores information about these breaks and then pools contigs that shared broken reads into clusters called isogroups; the contigs within a cluster that are joined by broken reads are called isotigs. In total, the junco assembly yielded 31,739 isotigs (average length of 1248 bases; Figure 1) in 22,765 isogroups (only 4,288 isogroups contain multiple isotigs).

Each isogroup likely represents a gene, while unique isotigs typically represent alternative splice forms; however, the isotigs may also represent divergent alleles, or
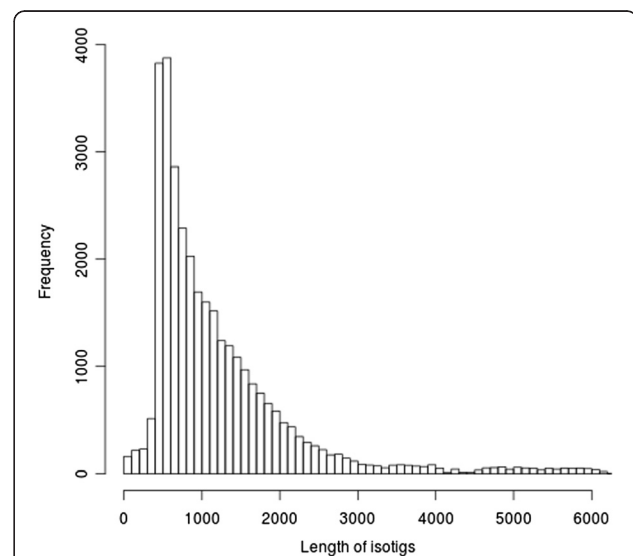

Figure 1 Distribution of isotig lengths. The frequency of isotigs of various lengths. Note that isotigs longer than 6,000 bases are omitted from the figure for clarity. The frequency continues to decline for greater lengths.

members of a gene family [7]. The current depth of sequencing precludes an accurate distinction between these possibilities, but future re-sequencing projects and gene expression studies could potentially distinguish between splice variants and recent duplicates. The Ensembl release 63 [35] for the zebra finch (*Taeniopygia guttata;* tae.Gut3.2.4), accessed via BioMart [36,37] contains 19,484 predicted or sequenced genes for the zebra finch, suggesting that our assembly of 22,765 isogroups may be incomplete (multiple isogroups represent a single true gene) or that there are more expressed regions in bird genomes than predicted by current zebra finch sequencing, annotation, and gene models.

Reference assembly

In order to confirm the validity of our sequences and to test the *de novo* assembly against a reference assembly, we also assembled the junco transcriptome using GMAP [38] with standard parameters against the closest available draft genome assembly: the zebra finch the first passerine genome sequencing project [39], and a species that diverged from juncos approximately 25 Ma ago [40]. Of the 1,180,500 reads from the junco transcriptome used for this assembly, 1,031,427 (87%) had a significant alignment to the zebra finch genome identified by the software, with an average of 92.9% identity. However, due to the limitation imposed by introns, the lengths of the assembled regions from the reference assembly are substantially shorter (331 bp vs 872 bp), and split into many more groups, than our *de novo* assembly (see Additional file 1 for further information on statistics and approach). For this reason, we chose to use the *de novo* assembly for all further analyses.

Annotation

The isotigs and singletons were queried against the NCBI non-redundant protein database[41,42] using BlastX to identify homologous, annotated genes. From this search, 17,884 of the isotigs and 15,871 of the singletons (49% and 11% of total respectively) returned a significant homolog (e value $< 10^{-5}$; Table 2; should refer to Additional file 2 for full annotation from all approaches). The proportion of assignment for isotigs is slightly higher than that of other *de novo* transcriptomes (range 23% - 35%), possibly because our isotigs are longer than the contigs used for annotation in other transcriptome annotations (1,248 bp vs 197 to 871 bp) [4-8,34]. Isotigs covered an average of 56.4% of the protein to which they aligned; singletons covered an average of only 18.4% of the protein to which they aligned, consistent with their shorter length. Isotigs were, on average, 82.4% identical across the full length of their alignment, while singletons were 79.4% identical.

Table 2 Number of isogroups and singletons annotated by different approaches, confirmed expressed, and containing SNPs

Assignment	Isogroup	Singleton
Total Assembly	22,765	166,177
NCBI nr-protein	10,276	15,871
Ensembl - ZF	9,863	14,019
Gene Ontology	9,120	737
KEGG	3,827	3,984
OrthoDB	10,062	14,019
Expressed	16,781	16,096
SNP present	6,992	NA

Sequences were annotated by sequence similarity to multiple databases, yielding the above number of annotations.

Because isogroups represent the full genetic unit, we sought to combine the isotigs into a single annotation for each isogroup, resulting in an annotation for 48% of isogroups (11,015 of 22,765). For nearly all annotated isogroups (10,276; 93%), only a single annotation was identified for all of its member isotigs. For a subset of isogroups (716; 7%), multiple annotations were identified but these were confirmed synonymous by literature or alignment searches to confidently assign an isogroup annotation (see Methods for details). Some isogroups (23; <1%) could not be reduced to a single annotation and might represent errors in assembly or transcripts with no currently identifiable homology. Our ability to assign a single annotation to most (99%) of annotated isogroups suggests that our assembly accurately characterizes many of the genes of the junco.

Of the 7,918 unique annotations identified among the isogroups reduced to a single annotation, 6,722 (85%) are represented by a single isogroup; the remaining 1,196 are represented by 4,268 isogroups (average of 3.57 isogroups per annotation). Multiple isogroups assigned to a single annotation suggests that some true genes in our dataset may be incompletely assembled, or represent recently diverged gene duplicates. For example, 6 isogroups are annotated as "nebulin," which is a conserved, single-copy gene in vertebrates [43,44], coding for an actin binding protein with multiple isoforms [45] that can complicate transcriptome assembly. The human version of nebulin has 6,669 amino acids and contains 183 exons at least 43 of which are alternatively spliced [46]. Additionally, some annotations are not for specific gene products: there are 1,353 isogroups assigned to nine annotations named some variant of "unnamed" or "hypothetical" proteins. The presence of multiple isogroups annotated as a single gene suggests that further assembly, or additional sequencing, may be necessary to complete the assembly of some large or complex genes.

Functional Annotation

Gene Ontology (GO) is a hierarchical description of gene function that classifies genes based on known or predicted function in model organisms [47]. The use of GO terms allows a broader assessment of our annotation and permits the assignment of functional roles to individual genes. There are limitations to assigning GO terms by sequence similarity alone, which can result in over-assignment of GO terms to genes that have functionally diverged [48-50]. Consequently, we are not placing high levels of confidence in any particular GO assignment, but rather we are investigating the large-scale patterns revealed by these functional annotations. While there are still some concerns with this approach, preliminary microarray results reveal meaningful and expected GO terms (Peterson, Rosvall, Tang and Ketterson, unpublished data), suggesting that our functional annotation is sufficient for broad assessments with the caveat that confidence in any given assignment should be limited.

We assigned GO terms, based on sequence similarity, using Blast2GO [51] and set a strict threshold criterion of $e < 10^{-15}$ against the NCBI non-redundant protein database [41,42]. This process provided at least one GO term for 83% of annotated isogroups (9,120) and 5% of annotated singletons (737) at the "inferred from sequence similarity" level of evidence [47]. This reduced ability to functionally classify singletons is likely due to their shorter length (mean 344 bp vs. 1,248 bp), which reduces the length and quality of sequence alignments that are possible, especially considering the stringent criterion employed. We identified a total of 65,008 GO term annotations representing 3,515 unique GO terms distributed throughout the GO graph (should refer to Additional file 2 for full annotations).

Pathway annotation

We additionally annotated the junco transcriptome sequences by mapping to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotation [52-54]. Briefly, assembled isotigs and singletons were aligned to protein sequences from the Ref-Seq databases [41,42] for zebra finch, chicken, mouse, and human using tBlastX, resulting in 6,269 isotigs (from 3,827 isogroups) and 3,984 singletons being assigned across all 234 unique pathways available for annotation (should refer to Additional file 2). These pathways can, in the future, be targeted for manipulation or sequencing. In addition, gene expression studies will likely use this approach to analyze changes to whole pathways, an approach which may be more sensitive and interpretable than focusing on single genes [55], to identify small changes in expression level with phenotypic consequences.

Orthology group assignment

In addition to our other annotations, we assigned the junco sequences to OrthoDB [56] orthology groups. We aligned junco sequences to the Ensembl protein database for zebra finch [35] with BlastX with a criteria of $e < 10^{-10}$, and assigned the top alignment (by bit score) to each isotig and singleton. We then determined whether all isotigs in an isogroup were assigned to the same protein. This returned a single annotation for 9,863 isogroups (43%), and multiple annotations for 199 isogroups that we excluded from further portions of this analysis. In addition, 14,019 singletons (8%) were assigned to an Ensembl protein. Among the isogroups, 6,542 annotations were assigned to only a single isogroup, and an additional 1,442 were assigned to multiple isogroups. Including the singletons returned 2,830 additional unique annotations. We then used the OrthoDB database [56] to assign each junco sequence as the orthology group that the corresponding zebra finch protein. This process resulted in assignment to 9,633 of the 12,557 orthology groups previously identified in zebra finch [56]. These data will allow future studies to focus on the divergence of orthologous gene families between junco and closely related species and may aid in the identification of recently duplicated genes in the junco.

Completeness

In order to assess the completeness of this transcriptome based on our sequencing efforts, we searched for significant sequence alignments for the assembled isotigs and singletons against the eukaryotic clusters of orthologous group (KOG) database [57,58]. Specifically, we searched for a set of 716 genes identified to be present as single copies in all bilaterians (an animal clade, including deuterostomes and proterostomes, that diverged at least 555 Ma ago) studied to date [59]. This reciprocal Blast analysis identified homologs for 646 core genes (90.2%) in our dataset (should refer to Additional file 2 for full annotation). This high level of coverage is slightly lower (but still comparable) to that found for large-scale genome sequences, which generally identify 99% of these core genes [59]. The best junco alignment for each KOG group aligned with a mean of 82.4% identity and covered an average of 69.5% (301.8 amino acids) of the aligned protein. This alignment coverage is likely artificially reduced by the fact that only one isotig (rather than a full isogroup) was counted for this length, meaning that any alternative splicing (or allelic variation) could result in a low reported coverage. Other transcriptome projects have not, to our knowledge, used this analysis. These results provide strong support that the junco transcriptome has been sequenced to sufficient depth and from a sufficiently diverse pool of sampled RNA to uncover most of the expected genes.

Confirmation of expression

To confirm that our putative transcripts are expressed, we constructed a custom 12-plex microarray (Roche NimbleGen, Inc., Madison, WI; see Methods for details) and conducted an expression experiment involving four tissues (ventromedial telencephalon, hypothalamus, liver, and pectoralis muscle) in male and female juncos. The full results of this experiment will be described elsewhere; this initial analysis focuses simply on confirming the expression of genes in the tissues of wild-caught individuals. We confirmed the expression of 23,914 (71.3%) of the contigs (representing 16,871 (74%) isogroups) and 16,096 (46.8%) of the singletons on the array in at least one sex-tissue combination (should refer to Additional file 2 for list). The isogroups and singletons that did not show expression in these analyzed tissues may be tissue/condition-specific genes that are unexpressed in our sampled tissues, or could represent spurious sequences that do not accurately represent expressed transcripts. On the whole, this result suggests that our sequenced genes are largely accurate reflections of expressed genes in the junco.

Sequence variants

Allelic variants, such as single nucleotide polymorphisms (SNPs), are powerful tools for population genetic analysis and identification of population structure. The junco system currently only has nine microsatellite markers, which have been used extensively for paternity analysis in one population [20]. In addition, 243 SNPs were identified across multiple junco species, and this number was sufficient to identify species, but not subspecies, divergence [60]. The addition of more genetic markers will greatly improve the ability to do large population comparisons, including further refining the current junco subspecies phylogeny [18], and identifying variants that may play a role in the current rapid divergence of the genus *Junco*. To this end, we have conservatively identified 25,781 unique potential SNPs in 6,992 isogroups of this transcriptome (should refer to Additional file 3 for full list and methods for details) for an average rate of 0.72 potential variants per 1000 basepairs of assembled sequence. While this rate of sequence variant identification is low compared to transcriptomes sequencing a greater number of individuals [7], these potential sequence markers will provide a strong starting point for future studies on population genetics and divergence in the genus *Junco*.

Conclusions

We have successfully sequenced a transcriptome of an ecological model songbird, the dark-eyed junco, using pyrosequencing and de novo assembly. Through our assembly process, we identified 22,765 putative genes – half of which have been annotated and three-quarters of which were validated by microarray investigation. These genes, including over 25,000 potential sequence variants, will immediately begin to inform the study of the junco and will provide a valuable resource for the study of many songbird species. Based on the identification of putatively conserved genes, we have demonstrated 90% coverage.

The acquisition of these sequence data adds genomic resources to another non-model system. By combining these data (and the tools they produce) with the historical study of a natural population, a solid foundation has been laid to advance the study of ecology, evolution, and behavior. We have already begun to exploit these novel tools in gene expression studies, showing in this approach that the majority of our sequenced genes are expressed under at least some natural conditions. Further research will extend these tools to deeper study of the transcriptomic responses of juncos to environmental stimuli, as well as assist in guiding traditional sequencing projects.

Methods

Tissue collection and RNA extraction

We collected tissues from two adult dark-eyed juncos, one male and one female, for this transcriptome. Both individuals had been held in captivity from 2005 when they were captured as six-day old nestlings in the wild near Mountain Lake Biological Station in Giles County, VA (37° 22′ 31″N, 80° 31′ 24″W), and transported to Bloomington, IN (39° 09′ 02″N, 86° 23′ 46″W). Individuals were hand-reared on a standard diet and were not manipulated in any experiment prior to inclusion in the transcriptome. For six months prior to euthanasia, we housed the birds individually, but neither visually nor acoustically isolated, on a photoperiod of 16:8 (hours light: dark) to induce gonadal growth and to ensure that we sequenced genes expressed during long photoperiods.

Within 20 min of euthanization by an overdose of isoflurane, all tissues were removed from the individual, ground in TRIzol® (Invitrogen Life Sciences, Carlsbad, CA), and stored at –80 °C. We collected the following organs: whole brain, gonad, liver, pectoralis muscle, syrinx, beak, eye, gizzard, heart, kidney, lung, preen gland, skin, and tongue. These tissues were chosen to represent a wide swath of potential gene expression, while explicitly avoiding the tissues (e.g., stomach, spleen) that were most likely to contain other species. This procedure conformed to all animal care regulations and was approved by the Bloomington Institutional Animal Care and Use Committee at Indiana University (Protocol #09-037). We extracted RNA from each tissue separately for each individual following the TriReagent manufacturer's protocol (Invitrogen Life Sciences; [61]). Total RNA was resuspended in water, and we confirmed concentration and

quality with a Bioanalyzer nanochip (Agilent Technologies, Waldbronn, Germany).

Library preparation and sequencing

Total RNA from each individual was quantified by fluorimetry (Quant-iT™ RiboGreen®, Invitrogen) and prepared into equimolar pools of 800 ng, creating whole-body male and female pools. Sequencing libraries optimized for Roche/454 Titanium sequencing were prepared using IU CGB customized protocols as previously described in [7] modified from [5]. Briefly, cDNA was synthesized in a fashion similar to the Clontech™SMART system by PCR amplifying each RNA pool using primers optimized for 454 sequencing. The resulting double-stranded cDNAs was then normalized by treatment with duplex-specific nuclease to reduce representation of highly abundant transcripts. Male and female libraries were separately titrated by enrichment and prepared for sequencing by emulsion PCR, each on one region of a two-region GS-FLX Titanium PicoTitre™ plate. The reads were cleaned of all adaptor/primer and polyA sequence by a program developed in-house at the CGB, Indiana University (http://sourceforge.net/projects/est-clean/ website). After cleaning, sequences ≤30 bp were removed from the dataset.

Assembly

Reads from the male and female pools were combined to increase the accuracy and completeness of assembly. We assembled these pooled reads using NEWBLER (v2.3; Roche/454 Sequencing) with the default parameters (40 bp overlap; 90% identity) resulting in 40,564 contigs (and 166,177 remaining singletons), which were further assembled into isogroups. Schwartz et al. [7] previously referred to this approach as graph-clustering and contig-graphs, as it graphically combines clusters of contigs that appear to be transcribed together (Genome Sequencer FLX System Software Manual, version 2.3, October 2009). An isogroup is composed of contigs that were split during the initial assembly because some of the reads overlapped multiple, independent contigs. NEWBLER reports information about reads that were broken between contigs during assembly, and clusters the component contigs into a single isogroup representing a putative gene. Once an isogroup is formed, all potential paths through the cluster are traversed and those paths that are supported by broken reads are reported as isotigs – that is, putative transcripts. Isogroups can either represent alternatively spliced genes (with contigs indicating exons, and isotigs representing splice forms), or sets of recently duplicated genes (with contigs representing regions of divergence since duplication, and isotigs representing the divergent genes) either as gene families or multiple alleles of the same gene [7].

Annotation

After determining that the *de novo* assembly was superior to the reference assembly, we used isotigs and singletons as a query against the NCBI non-redundant protein database (Accessed in October 2010) using BlastX sequence alignment with a threshold e-value of 10^{-5}. We identified the top match for each isotig or singleton by bit-score, and the corresponding gene information was assigned to the junco sequence. Because a number of genes are referred to by multiple names and abbreviations, we manually curated those isogroups that contained multiple unique annotations to identify a single annotation for each isogroup. For the majority of multiply annotated isogroups (637), this involved simply collapsing multiple synonyms, but 102 isogroups contained annotations that could not be readily collapsed. The isotigs from these isogroups were queried against the zebra finch UniProt database using BlastX with strict threshold criterion (e value $< 10^{-10}$) to reduce spurious matches from divergent taxa. The top five sequence alignments (by bit score) for each isotig in an isogroup were compared and if a single gene appeared in all isotigs and was represented in the original annotation, it was assigned as the isogroup annotation. For cases in which two or more gene annotations were identified in all isotigs from an isogroup, the annotation with the highest cumulative bit score was assigned as the isogroup annotation if it matched an original annotation. This approach left 23 isogroups with multiple annotations; 7 did not match against anything in the zebra finch UniProt database, and 16 matched against genes other than those originally annotated and could not be fully collapsed – these isogroups were omitted from further isogroup analyses.

Additional Annotation

We also used isotigs and singletons to query several databases designed to provided additional information on the function, pathway, orthology group, and completeness of the assembled sequences. Table 3 details the databases and search parameters utilized for these annotations. The top alignment match from each search was assigned to the corresponding isogroup or singleton.

Completeness

We used the predicted open reading frames from the assembled junco sequences (isotigs and singletons) to search for a set of 716 conserved bilaterian genes [59] retrieved from the KOG protein database [57,58]. Reciprocal BlastP was performed with an e-value threshold of e $< 10^{-5}$. Junco sequences were identified as the homolog of a KOG protein if, and only if, they were reciprocal best alignments. The best junco alignment for each KOG group was then selected by bit score and alignment statistics were reported.

Table 3 Additional annotation approach details

Approach	Database	Blast Flavor	E-value	References	Notes
Functional Annotation	NCBI- NR	Blast2GO	10^{-15}	[41,42,47,51]	
Pathway Annotation	KEGG	tBlastX	10^{-5}	[52-54]	Limited to zebra finch, chicken, and mouse.
Orthologous Group	OrthoDB, Ensembl	BlastX	10^{-10}	[35-37,56]	Limited to zebra finch. Matched Ensembl hits to OrthoDB assignments

To obtain more detailed information about the function of the sequenced genes, searches beyond simple non-redundant proteins were conducted.

Confirmation of expression

A custom microarray was designed from the sequence of the junco transcriptome. For 33,545 contigs, three unique probes are present on the array, while another 61 contigs are represented by two probes and 65 contigs are represented by one probe, accounting for 100,822 probes on the array. An additional 34,365 probes were selected from the remaining singletons (one probe per chosen singleton). The array also contains control probes and 2,604 random probes designed to reflect the genome nucleotide composition by Markov modeling to experimentally determine the appropriate thresholds that measure significant hybridization signals over the background. Thus, each sub-array consists of over 137,000 long-oligonucleotide (60 bp) probes, and 12 such sub-arrays are placed on each glass slide (Roche NimbleGen Inc., Madison, WI). The microarray platform is deposited at NCBI Gene Expression Omnibus (GEO; accession number GPL14995).

We collected adult dark-eyed juncos from breeding grounds near Mountain Lake Biological Station (Pembroke, VA) in mist-nets between May 7 and 14, 2010 and held them individually in a semi-naturalistic outdoor aviary where they were neither acoustically nor visually isolated from other juncos, as part of a larger experiment. On June 9 and 10 individuals were euthanized by overdose of isoflurane. Tissues, including whole brains, were collected rapidly and stored on powdered dry ice within 20 min post-mortem to ensure negligible RNA degradation [62]. Brains were dissected into 14 distinct regions using anatomical landmarks, following previously established methods [63] based on the zebra finch brain atlas. These brain regions included the hypothalamus and the ventral medial telencephalon (VmT), which primarily consists of the nucleus taeniae, the avian homologue of the medial amygdala [64-66].

RNA from VmT, hypothalamus, liver, and pectoralis was extracted in TRIzol® following manufacturer's directions (Invitrogen, Carlsbad, CA). The microarray protocol follows previously published methods [67]. Briefly, total RNA was reverse-transcribed to ss-cDNA in the presence of oligodT primer and SuperScript II reverse transcriptase. This ss-cDNA was then converted to ds-RNA and labeled using CY-labeled random nonmer primer (either Cy3 or Cy5) and Klenow fragment

(following NimbleGen labeling protocols). We then hybridized 4 g of each of two labeled samples (one Cy3, one Cy5) to each sub-array and followed manufacturer's directions for post-hybridization washing and scanning (Roche NimbleGen, Inc., Madison, WI). Imaging was accomplished by Axon GenePix 4200A scanner (Molecular Devices, Sunnyvale CA) with GenePix 6.0 software and data were extracted with NimbleScan 2.4 (Roche NimbleGen, Inc., Madison WI). Raw microarray data were processed with the limma package [68] in R version 2.9.0 [69] to normalize expression scores.

To determine if a gene was expressed, we calculated the 97.5% quantile for expression score of random probes in each individual as the cutoff for calling expression. Thus, for each individual, a called expression is significant at a p-value of 0.025. For each contig, we tested the median probe value against this threshold, and for singletons we used the single expression value. Because our design employed biological replicates, we called a contig or singleton expressed only if at least three of the six individuals in a group were called as expressed, thus reducing the p-value further to 0.0006 (the probability of obtaining at least three of six individuals called for expression of a random probe). From this, we determined whether or not a gene had expression support in any of our tissues-sex pairings, and which genes were restricted to expression in one sex.

Sequence Variants

To identify sequence variants, we aligned cleaned sequence reads to the assembled transcriptome using Blastn with a threshold of $e < 10^{-5}$ and requiring 95% identity and alignment of all but the last five bases on each end of the read. The first and last five bases on each end of the read were trimmed before analysis (if they aligned) to prevent inaccurate SNP calls near the edges of reads. Reads that were assembled to multiple isogroups were omitted from further analysis. The aligned reads at each sequence position (base) were compared to the consensus sequence (the assembly).

Those positions that contained a minor allele represented with a frequency greater than 20%, and supported by at least three reads, were identified as potential SNPs. Alleles representing gaps were not included in this dataset because insertions and deletions are the most

common form of 454 sequencing error [70] and generally cause frameshift mutations in coding sequence (such as cDNA sequencing) and are therefore less likely to be true sequence variants. Sequencing errors for GS-FLX Titanium pyrosequencing (Roche/454 Sequencing) mismatches at a rate of 0.022% [70], suggesting that we should only expect 0.37 false SNPs (due to sequencing error) to be called under these stringent criteria (binomial distribution of at least three substitution errors at the same site with depth of coverage eight = 1.055×10^{-8}; times 35.8 million sites), providing strong support that any identified sequence variants are very likely to represent true SNP variation in these two individuals.

Because SNPs were identified in isotigs, there is the potential to identify the same SNP in more than one isotig of the same isogroup. To address this concern, we report the number of confidently unique SNPs along with the total number identified. Called SNPs that had the same major and minor allele were restricted to only those from the isogroup with the greatest number of that sequence variant. For example, in isogroup00018, 250 isotigs all had calls for SNPs with G as the major allele, and T as the minor allele at position 551 (likely because they all share the same first contig). Only the call for this polymorphism from a single isotig was included in the report of unique SNPs. This approach was conservative, as we likely omitted SNPs that were truly unique as well.

Additional files

Additional file 1: Contains statistics and a brief description of our attempt to develop a reference assembly using the zebra finch genome [71-75].

Additional file 2: List of isogroups and singletons (rows) with their corresponding annotations (from various sources of evidence) and expression support (columns).

Additional file 3: SNPs identified in the assembled transcriptome, including both unique and redundant (i.e., the same SNP call in multiple isotigs of the same isogroup) SNP calls.

Author details
[1]Dept. of Biology, Center for Integrated Study of Animal Behavior, Indiana University, Bloomington, IN, USA. [2]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI, USA. [3]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN, USA. [4]Cancer Center, Department of Biostatistics, Georgia Health Sciences University, Georgia, IN, USA. [5]Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA.

References
1. Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM, Petrov D, Jonsson B, Schluter D, Bell MA, Kingsley DM: **Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer.** *Science* 2010, **327:**302–305.
2. O'Quin KE, Hofmann CM, Hofmann HA, Carleton KL: **Parallel Evolution of Opsin Gene Expression in African Cichlid Fishes.** *Molecular Biolgoy and Evolution* 2010, **27:**2839–2854.
3. Ellegren H: **Sequencing goes 454 and takes large-scale genomics into the wild.** *Mol Ecol* 2008, **17:**1629–1631.
4. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17:**1636–1647.
5. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **219:**10.
6. Jeong-Hyeon C, Teiya K, Emilie S, Hongseok T, Youngik Y, Armin M, Justen A: **Gene discovery in the horned beetle Onthophagus taurus.** *BMC Genomics* 2010, **11:**703.
7. Schwartz TS, Tae H, Yang Y, Mockaitis K, Van Hemert JL, Proulx SR, Choi JH, Bronikowski AM: **A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences.** *BMC Genomics* 2010, **694:**11.
8. Santure AW, Gratten J, Mossman JA, Sheldon BC, Slate J: **Characterisation of the transcriptome of a wild great tit Parus major population by next generation sequencing.** *BMC Genomics* 2011, **283:**12.
9. Poelchau MF, Reynolds JA, Denlinger DL, Elsik CG, Armbruster PA: **A de novo transcriptome of the Asian tiger mosquito, Aedes albopictus, to identify candidate transcripts for diapause preparation.** *BMC Genomics* 2011, **12:**619.

10. Kunstner A, Wolf JBW, Backstrom N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, Jarvis E, Warren W, Ellegren H: **Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species.** *Mol Ecol* 2010, **19:**266–276.

11. Wolf JBW, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, Tautz D: **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* 2010, **19:**162–175.

12. Wray GA: **The evolutionary significance of cis-regulatory mutations.** *Nature Reviews Genetics* 2007, **8:**206–216.

13. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D: **Three Periods of Regulatory Innovation During Vertebrate Evolution.** *Science* 2011, **333(80):**1019–1024.

14. Rowan W: **Relation of light to Bird migration and developmental changes.** *Nature London* 1925, **115:**494–495.

15. Rowan W: **On photoperiod-ism, reproductive activity, and the annual migrations of birds and certain fishes.** *Proc. Boston Soc. Nat. Hist* 1926, **38:**147–189.

16. Nolan V Jr, Ketterson ED, Cristol DA, Rogers CM, Clotfelter ED, Titus RC, Schoech SJ, Snajdr E: **Dark-eyed Junco: Junco hyemalis.** *Birds of North America* 2002, **716:**1–44.

17. Miller AH: **Speciation in the Avian species Iunco.** *Univ. Calif. Pub. Zool* 1941, **44:**173–434.

18. Mila B, McCormack JE, Castaneda G, Wayne RK, Smith TB: **Recent postglacial range expansion drives the rapid diversification of a songbird lineage in the genus Junco.** *Proceedings of the Royal Society B-Biological Sciences* 2007, **274:**2653–2660.

19. Ketterson E, Atwell J, McGlothlin J: **Phenotypic integration and independence: Hormones, performance, and response to environmental change.** *Integr Comp Biol* 2009, **49:**365–379.

20. Gerlach N, McGlothlin J, Parker P, Ketterson E: **Promiscuous mating produces offspring with higher lifetime fitness.** *Proceeding of the Royal Society B* 2012, **279:**860–866.

21. Ketterson ED, Nolan VJ, Casto JM, Buerkle CA, Clotfelter E, Grindstaff JL, Jones KJ, Lipar JL, McNabb FMA, Neudorf DL, Parker-Renga I, Schoech SJ, Snajdr E, Dawson A, Chaturvedi CM: **Testosterone, phenotype and fitness: a research program in evolutionary behavioral endocrinology.** *Avian endocrinology* 2001, :19–40.

22. Ketterson E, Nolan V, Wolf L, Ziegenfus C, Dufty A, Ball G, Johnsen T: **Testosterone and avian life histories - the effect of experimentally elevated testosterone on corticosterone and body-mass in dark-eyed juncos.** *Horm Behav* 1991, **25:**489–503.

23. McGlothlin JW, Ketterson ED: **Hormone-mediated suites as adaptations and evolutionary constraints.** *Philosophical Transactions of the Royal Society B-Biological Sciences* 2008, **363:**1611–1620.

24. Reed WL, Clark ME, Parker PG, Raouf SA, Arguedas N, Monk DS, Snajdr E, Nolan V, Ketterson ED: **Physiological effects on demography: A long-term experimental study of testosterone's effects on fitness.** *American Naturalist* 2006, **167:**667–683.

25. McGlothlin JW, Whittaker DJ, Schrock SE, Gerlach NM, Jawor JM, Snajdr EA, Ketterson ED: **Natural Selection on Testosterone Production in a Wild Songbird Population.** *American Naturalist* 2010, **175:**687–701.

26. Newman MM, Yeh PJ, Price TD: **Reduced territorial responses in dark-eyed juncos following population establishment in a climatically mild environment.** *Anim Behav* 2006, **71:**893–899.

27. Rasner CA, Yeh P, Eggert LS, Hunt KE, Woodruff DS, Price TD: **Genetic and morphological evolution following a founder event in the dark-eyed junco, Junco hyemalis thurberi.** *Mol Ecol* 2004, **13:**671–681.

28. Yeh PJ: **Rapid evolution of a sexually selected trait following population establishment in a novel habitat.** *Evolution Int J Org Evolution* 2004, **58:**166–174.

29. Yeh PJ, Hauber ME, Price TD: **Alternative nesting behaviours following colonisation of a novel environment by a passerine bird.** *Oikos* 2007, **116:**1473–1480.

30. Yeh PJ, Price TD: **Adaptive phenotypic plasticity and the successful colonization of a novel environment.** *American Naturalist* 2004, **164:**531–542.

31. Atwell JW, DM O'Neal, Ketterson ED: **Animal migration as a moving target for conservation: intra-species variation and responses to environmental change, as illustrated in a sometimes migratory songbird.** *Lewis and Clark, Environmental Law Journal* 2011, **41(2):**289–316.

32. Subramanian S, Huynen L, Millar CD, Lambert DM: **Next generation sequencing and analysis of a conserved transcriptome of New Zealand's kiwi.** *BMC Evol Biol* 2010, **387:**10.

33. Casewell NR, Harrison RA, Wuester W, Wagstaff SC: **Comparative venom gland transcriptome surveys of the saw-scaled vipers (Viperidae: Echis) reveal substantial intra-family gene diversity and novel venom transcripts.** *BMC Genomics* 2009, **564:**10.

34. Wang W, Wang Y, Zhang Q, Qi Y, Guo D: **Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing.** *BMC Genomics* 2009, **465:**10.

35. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kaehaeri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Parker A, Proctor G, Vogel J, Searle SMJ: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39:**D800–D806.

36. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal-unified access to biological data.** *Nucleic Acids Res* 2009, **37:** W23–W27.

37. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart - biological queries made easy.** *BMC Genomics* 2009, **22.**

38. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21:**1859–1875.

39. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kuenstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TAF, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin YC, George J, Sweedler J, Southey B, Gunaratne P, Watson M, Nam K, Backstroem N, Smeds L, Nabholz B, Itoh Y, Whitney O, Pfenning AR, Howard J, Voelker M, Skinner BM, Griffin DK, Ye L, McLaren WM, Flicek P, Quesada V, Velasco G, Lopez-Otin C, Puente XS, Olender T, Lancet D, Smit AFA, Hubley R, Konkel MK, Walker JA, Batzer MA, Gu W, Pollock DD, Chen L, Cheng Z, Eichler EE, Stapley J, Slate J, Ekblom R, Birkhead T, Burke T, Burt D, Scharff C, Adam I, Richard H, Sultan M, Soldatov A, Lehrach H, Edwards SV, Yang SP, Li XC, Graves T, Fulton L, Nelson J, Chinwalla A, Hou S, Mardis ER, Wilson RK: **The genome of a songbird.** *Nature* 2010, **464:**757–762.

40. Barker F, Cibois A, Schikler P, Feinstein J, Cracraft J: **Phylogeny and diversification of the largest avian radiation.** *Proc Natl Acad Sci U S A* 2004, **101:**11040–11045.

41. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137–140.

42. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37:**D32–D36.

43. Hanashima A, Kubokawa K, Kimura S: **Characterization of amphioxus nebulin and its similarity to human nebulin.** *Journal of Experimnetal Biology* 2009, **212:**668–672.

44. Bjorklund AK, Light S, Sagit R, Elofsson A: **Nebulin: A Study of Protein Repeat Evolution.** *J Mol Biol* 2010, **402:**38–51.

45. McElhinny AS, Kazmierski ST, Labeit S, Gregorio CC: **Nebulin: The nebulous, multifunctional giant of striated muscle.** *Trends Cardiovasc Med* 2003, **13:**195–201.

46. Donner K, Sandbacka M, Lehtokari VL, Wallgren-Pettersson C, Pelin K: **Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts.** *European Journal of Human Genetics* 2004, **12:**744–751.

47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology Consortium: Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25–29.

48. Joshi T, Xu D: **Quantitative assessment of relationship between sequence similarity and function similarity.** *BMC Genomics* 2007, **222:**48.

49. Devos D, Valencia A: **Practical limits of function prediction.** *Proteins- Structure Function and Genetics* 2000, **41:**98–107.

50. Devos D, Valencia A: **Intrinsic errors in genome annotation.** *Trends in Genetics* 2001, **17:**429–431.

51. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**:3674–3676.

52. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28**:27–30.

53. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**:D355–D360.

54. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34**:D354–D357.

55. Tomfohr J, Lu J, Kepler TB: **Pathway level analysis of gene expression using singular value decomposition.** *BMC Bioinformatics* 2005, **225**:6.

56. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV: **OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011.** *Nucleic Acids Res* 2011, **39**:D283–D288.

57. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**(80):631–637.

58. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22–28.

59. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong QF, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang ZJ, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye YZ, Shaw JR, Andrews J, Crease TJ, Tang HX, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL: **The Ecoresponsive Genome of Daphnia pulex.** *Science* 2011, **331**(80):555–561.

60. McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT: **Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences.** *Mol Phylogenet Evol* 2012, **62**:397–406.

61. Chomczynki P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.** *Anal Biochem* 1987, **162**:156–159.

62. Cheviron ZA, Carling MD, Brumfield RT: **Effects of postmortem interval and preservation method on RNA isolated from field-preserved avian tissues.** *Condor* 2011, **113**:483–489.

63. Soma KK, Bindra RK, Gee J, Wingfield JC, Schlinger BA: **Androgen-metabolizing enzymes show region-specific changes across the breeding season in the brain of a wild songbird.** *J Neurobiol* 1999, **41**:176–188.

64. Thompson RR, Goodson JL, Ruscio MG, Adkins-Regan E: **Role of the archistriatal nucleus taeniae in the sexual behavior of male Japanese quail (Coturnix japonica): A comparison of function with the medial nucleus of the amygdala in mammals.** *Brain Behavior and Evolution* 1998, **51**:215–229.

65. Canoine V, Fusani L, Schlinger B, Hau M: **Low sex steroids, high steroid receptors: Increasing the sensitivity of the nonreproductive brain.** *Developmental Neurobiology* 2007, **67**:57–67.

66. Voigt C, Goymann W: **Sex-role reversal is reflected in the brain of African black coucals (Centropus grillii).** *Developmental Neurobiology* 2007, **67**:1560–1573.

67. Lopez J, Colbourne J: **Dual-Labeled Expression Microarray Protocol for High-Throughput Genomic Investigations.** *CGB Technical Report* 2011, **201**(1):2.

68. Smyth GK: **Limma: Linear models for microarray data.** *Bioinformatics and computational biology solution using R and Bioconductor* 2005, **420**:397–420.

69. R Development Core Team: R: *A Language and Environment for Statistical.* Vienna, Austria: R Foundation for Statistical Computing; 2010.

70. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF: **Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing.** *BMC Genomics* 2011, **12**:12.

71. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21**:1859–1875.

72. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kuenstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TAF, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin YC, George J, Sweedler J, Southey B, Gunaratne P, Watson M, Nam K, Backstroem N, Smeds L, Nabholz B, Itoh Y, Whitney O, Pfenning AR, Howard J, Voelker M, Skinner BM, Griffin DK, Ye L, McLaren WM, Flicek P, Quesada V, Velasco G, Lopez-Otin C, Puente XS, Olender T, Lancet D, Smit AFA, Hubley R, Konkel MK, Walker JA, Batzer MA, Gu W, Pollock DD, Chen L, Cheng Z, Eichler EE, Stapley J, Slate J, Ekblom R, Birkhead T, Burke T, Burt D, Scharff C, Adam I, Richard H, Sultan M, Soldatov A, Lehrach H, Edwards SV, Yang SP, Li XC, Graves T, Fulton L, Nelson J, Chinwalla A, Hou S, Mardis ER, Wilson RK: **The genome of a songbird.** *Nature* 2010, **464**:757–762.

73. Barker F, Cibois A, Schikler P, Feinstein J, Cracraft J: **Phylogeny and diversification of the largest avian radiation.** *Proc Natl Acad Sci U S A* 2004, **101**:11040–11045.

74. Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D: **Patterns of exon-intron architecture variation of genes in eukaryotic genomes.** *BMC Genomics* 2009, **47**:10.

75. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kaehaeri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Parker A, Proctor G, Vogel J, Searle SMJ: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**:D800–D806.