## Title

Improved zebra finch brain transcriptome identifies novel proteins with sex differences



## Authors

He, Jingyan
Fu, Ting
Zhang, Ling
et al.

## Publication Date

2022-11-01

## DOI

10.1016/j.gene.2022.146803

Peer reviewed

Gene 843 (2022) 146803

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

# Improved zebra finch brain transcriptome identifies novel proteins with sex differences

Jingyan He [a], Ting Fu [b], Ling Zhang [a], Lucy Wanrong Gao [c], Michelle Rensel [d], Luke Remage-Healey [e], Stephanie A. White [a], Gregory Gedman [a], Julian Whitelegge [c], Xinshu Xiao [a], Barney A. Schlinger [a,*]

[a] Department of Integrative Biology and Physiology, University of California, Los Angeles 90095, United States
[b] Molecular, Cellular and Integrative Physiology Interdepartmental Program, University of California, Los Angeles 90095, United States
[c] The Pasarow Mass Spectrometry Laboratory, The Jane and Terry Semel Institute for Neuroscience and Human Behavior, Brain Research Institute, David Geffen School of Medicine, University of California, Los Angeles 90095, United States
[d] The Institute for Society and Genetics, University of California, Los Angeles 90095, United States
[e] Center for Neuroendocrine Studies, Neuroscience and Behavior, 639 N. Pleasant St, Morrill IVN Neuroscience, University of Massachusetts, Amherst, MA 01003, United States

ARTICLE INFO

Edited by: Dr. X. Carette

*Keywords:*
Zebra finch
RNA-seq
SMRT-seq
Song-system

ABSTRACT

The zebra finch (*Taeniopygia guttata*), a representative oscine songbird species, has been widely studied to investigate behavioral neuroscience, most notably the neurobiological basis of vocal learning, a rare trait shared in only a few animal groups including humans. In 2019, an updated zebra finch genome annotation (bTaeGut1_v1.p) was released from the Ensembl database and is substantially more comprehensive than the first version published in 2010. In this study, we utilized the publicly available RNA-seq data generated from Illumina-based short-reads and PacBio single-molecule real-time (SMRT) long-reads to assess the bird transcriptome. To analyze the high-throughput RNA-seq data, we adopted a hybrid bioinformatic approach combining short and long-read pipelines. From our analysis, we added 220 novel genes and 8,134 transcript variants to the Ensembl annotation, and predicted a new proteome based on the refined annotation. We further validated 18 different novel proteins by using mass-spectrometry data generated from zebra finch caudal telencephalon tissue. Our results provide additional resources for future studies of zebra finches utilizing this improved bird genome annotation and proteome.

## 1. Introduction

Songbirds (Order Passeriformes; Suborder Oscine) are well-established organisms for neurobiological studies especially those aimed at understanding the neural basis of vocal learning (Doupe and Kuhl, 1999; Clayton et al., 2009; Jarvis, 2019). This rare ability to acquire vocalizations through imitation of a model is observed in only a few mammalian and avian taxa. Various properties of birdsong acquisition and human speech parallel one another (Jarvis, 2004). For example, both species share corticostriatal circuits for vocalization production and demonstrate a direct projection from the motor cortex to brainstem vocal motor neurons, a connection unique to vocal learners (Jarvis, 2004; Jürgens, 2002; Bolhuis and Gahr, 2006; Petkov and Jarvis, 2012). In addition, brain regions involved in vocal learning pathways of songbirds and humans are functionally specialized and exhibit convergent transcriptional profiles, suggesting overlapped molecular mechanisms underlying complex vocal learning traits across the two evolutionarily distant species (Margoliash et al., 1994; Lovell et al., 2008; Lovell et al., 2013; Pfenning et al., 2014). Besides the behavioral, neuronal and molecular similarities shared with humans, zebra finches are amenable to captivity and experimental manipulation making them an outstanding model for investigation into the molecular basis of vocal

*Abbreviations:* SMRT, single-molecule real-time; G10K-VGP Project, Genome 10*K* Project; RNA-seq, RNA sequencing; NGS, next generation sequencing; TGS, third-generation sequencing; HVC, acronym used as proper name; LMAN, lateral magnocellular nucleus of the anterior nidopallium; RA, robust nucleus of the arcopallium.

* Corresponding author.
*E-mail address:* schlinge@lifesci.ucla.edu (B.A. Schlinger).

https://doi.org/10.1016/j.gene.2022.146803
Received 25 April 2022; Received in revised form 18 July 2022; Accepted 5 August 2022
Available online 9 August 2022

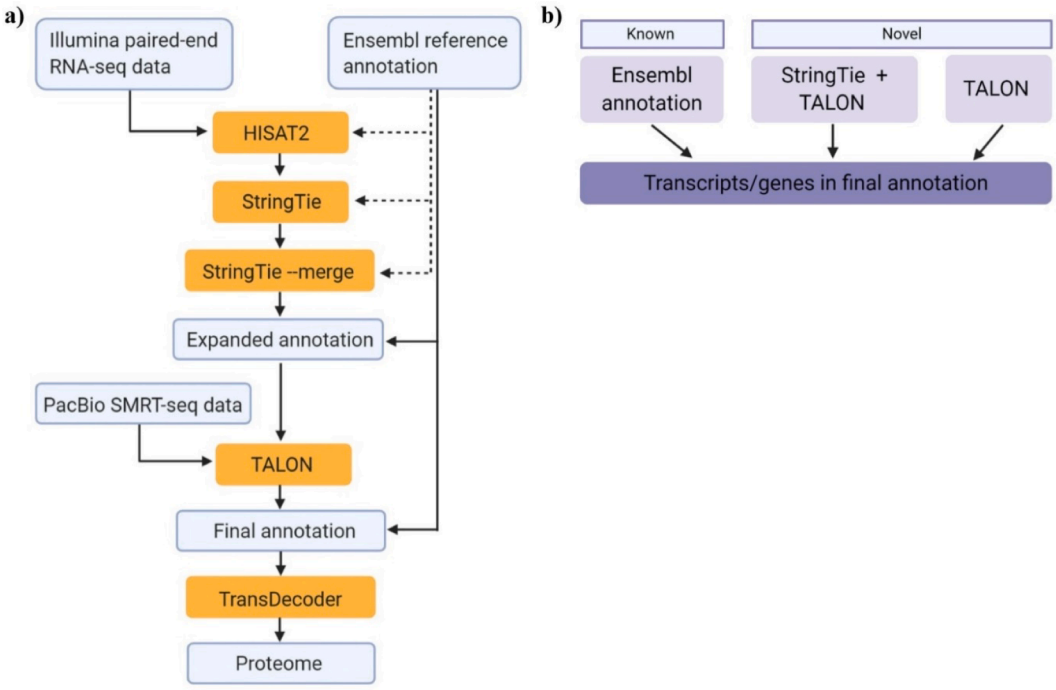


**Fig. 1. Overview of the analysis pipeline.** a) 1. Illumina paired-end RNA-seq data were aligned to Ensembl reference genome using HISAT2. The short-read transcriptome assembly was done by StringTie, and a unified transcriptome set containing novel transcripts were generated by StringTie merge mode. The StringTie novel transcripts were processed by a customized python script and merged with Ensembl annotation transcripts to obtain an expanded annotation. 2. The PacBio SMRT-seq generated FLnc reads were annotated against the expanded annotation through the TALON pipeline. The novel transcripts models supported by long read data were added to the Ensembl annotation to construct the final annotation. 3. A proteome was generated using TransDecoder based on the transcriptome in the final annotation. b) The transcripts and genes in the final annotation were either known ones from Ensembl annotation or novel ones supported by long-reads. The novel transcripts and genes were defined from two sources: 1) predicted by StringTie from short-read data and fall in TALON "Known" transcript novelty category during the long-read annotation; 2) predicted by TALON pipeline and were assigned to one of the TALON novel transcript categories.

learning (Heston and White, 2017).

The zebra finch genome was sequenced (Warren et al., 2010) as only the second avian species subject to whole-genome sequencing (Hillier et al., 2004). Whereas this zebra finch genome assembly, the nucleotide sequence of the genome, as well as the genome annotation have been widely used, several studies have pointed out that the annotation is incomplete (Balakrishnan et al., 2012; Fuxjager et al., 2016). In 2019, the Vertebrate Genome Project under The Genome 10 K Project (G10K-VGP Project) released an updated zebra finch genome assembly bTaeGut1_v1.p (INSDC Assembly GCA_003957565.2). The new genome assembly was generated using more advanced sequencing technologies and assembly methods. Notably, the new reference sequence was created from the same DNA sample that was used in the initial zebra finch genome assembly, a bird designated as Isolate: Black17. Based on the higher assembly quality, the Ensembl zebra finch genome annotation (bTaeGut1_v1.p, Genebuild released in December 2019) is substantially improved, being more comprehensive with nearly doubled transcript numbers compared to the first annotation.

A complete and accurate genome annotation, which identifies and records the information of functional elements along the sequence of a genome, lays the foundation of increased quality for genomic studies that address biological inquiries (Zhao and Zhang, 2015; Abril and Castellano, 2019). The remarkable enhancement in completion of the bird transcriptome, which is the total collection of RNA molecules transcribed from a genome, could largely advance the genomic studies of zebra finch in the context of RNA-seq analysis for various research purposes and beyond (Wu et al., 2012; Han et al., 2015; Srivastava et al., 2019). To our knowledge, no study has re-assessed the bird transcriptome to date.

High-throughput RNA sequencing (RNA-seq) is a promising approach to provide comprehensive investigation and insights into a transcriptome due to its capability of capturing expressed genes in the tissue samples (Ji and Sadreyev, 2018; Salzberg, 2019). The pervasive next generation sequencing (NGS) RNA-seq methods such as Illumina-based short-read RNA-seq have been used in numerous biomedical research applications including an unbiased survey of the entire transcriptome and gene expression quantification (Denoeud et al., 2008; Li et al., 2009). In addition to performing quantitative assessment, NGS RNA-seq analysis can also be exploratory with the capability of novel transcript discovery (Han et al., 2015). Nevertheless, the nature of short-read sequences limits the creation of an unambiguous assembly of NGS RNA-seq data, a complex and challenging bioinformatic task (Korf, 2013; Martin and Wang, 2011). Recently, the emerging third-generation sequencing (TGS) technologies such as PacBio single-molecule real-time (SMRT) sequencing present an alternative powerful method for transcriptome profiling. With the advantage of producing reads that are typically $>$ 10 Kbp long, the SMRT-seq method is able to reveal the complex structural variants of the expressed genes by sequencing the full-length transcripts (Roberts et al., 2013; Pollard et al., 2018). Further, the IsoSeq method, a part of SMRT Link analysis that was developed by PacBio, has enabled the production of high-quality full-length transcripts without the need of assembly (Wang et al., 2016; Chen et al., 2017). By integrating and combining NGS and TGS sequencing methods, many studies have successfully constructed complete transcriptomes for model and non-model organisms and discovered novel transcripts in well-annotated species (Zhang et al., 2019; Qiao et al., 2020; Deslattes Mays et al., 2019).

In this study, we used publicly available RNA-seq data generated from Illumina-based RNA-seq and PacBio SMRT-seq to investigate the current zebra finch annotation. Our analysis pipeline incorporated the advantages of short and long-read RNA-seq methods to discover high-confidence novel transcripts and genes. The novel discoveries from

our study uncovered additional transcripts laying outside the new zebra finch annotation, which implies the possibility and necessity of future improvement in zebra finch genome annotation. To assess the biological relevance and implications of our findings, we predicted open reading frames (ORF) and protein/peptide sequences for the novel transcript isoforms and genes. Interestingly, most of the predicted peptide sequences showed homologies to known proteins from the universal BlastP search against the Swissport protein database, which further suggests the existence of unannotated protein coding transcripts. To further validate the protein prediction from our analysis pipeline and to confirm the identity of novel proteins, we generated mass-spectrometry data from zebra finch caudal telencephalon, with some samples focused specifically on the NCM, a higher-level auditory cortex involved in acoustic processing (Bolhuis and Gahr, 2006; Remage-Healey et al., 2010). We captured 18 different proteins that were only present in either male or female tissue samples. The validation of sex-different novel proteins may provide additional clues for research investigating molecular mechanisms of birdsong, a sexually dimorphic trait in zebra finches (Nottebohm and Arnold, 1976). Overall, the novel findings from our analysis provide additional sources and information for the future studies of zebra finch brain and behavior.

## 2. Materials and methods

### 2.1. High-throughput RNA-seq datasets

The Illumina-based short-read RNA-seq data were obtained from Burkett et al., (Burkett et al., 2018). The study focused on the gene expression in Area X, a key vocal nucleus in zebra finch basal ganglia (Sohrabji et al., 1990). The RNA-seq data were generated from the Area X tissues of 7 juvenile male zebra finches (control birds with GFP expression) during the critical period of vocal learning and song development. cDNA libraries for each sample were sequenced twice by Illumina HiSeq 2500 platform, and 50 bp long paired-end short reads for each sample were obtained (Burkett et al., 2018). The data were retrieved from NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/, accession number: GSE96843).

The SMRT-seq long-read RNA-seq data were obtained from PacBio (https://downloads.pacbcloud.com/public/dataset/AvianBrainTranscriptome/). The data were generated from 6.76 μg total RNA of a zebra finch whole brain tissue, and the sample was sequenced using 4 SMRT Cells on the Sequel System. Raw sequences were processed through PacBio IsoSeq analysis in SMRT Link 5.0. The full-length non-chimeric (FLnc) CCS reads from the repository were used in our analysis (Vierra et al., 2017).

### 2.2. Overview of the RNA-seq analysis pipeline

Fig. 1 shows the overall flow of our analysis pipeline. Briefly, the Illumina-based short-read RNA-seq data were analyzed using StringTie (Pertea et al., 2015) for *de novo* transcript assembly. The StringTie results of the 7 brain samples were merged, and integrated with Ensembl annotation to obtain an expanded annotation. This expanded annotation then served as the reference annotation for the long-read analysis pipeline based on TALON (Wyman et al., 2019). The final annotation consists of two types of transcripts: Ensembl annotated transcripts and novel transcripts (that were supported by short and long reads data or long reads data alone).

In addition, to enable future biological studies of the novel transcripts in zebra finch transcriptome, we performed ORF and peptide sequence prediction to generate a predicted proteome based on the final annotation (Fig. 1).

### 2.3. Short-read RNA-seq data analysis

The Illumina-based RNA-seq data for each bird sample were first aligned to the Ensembl genome assembly bTaeGut1_v1.p, INSDC Assembly GCA_003957565.2 (https://uswest.ensembl.org/Taeniopygia_guttata/Info/Annotation) using HISAT2 v2.1.0 (Kim et al., 2015) with default parameters. Next, the uniquely mapped reads in the two technical replicates for each bird sample were merged into a single file using Samtools v1.9 (Li et al., 2009). The merged file for each biological replicate was passed to StringTie v2.1.3b (Pertea et al., 2015) for transcript assembly. We provided a reference annotation (Ensembl bTaeGut1_v1.p) to StringTie to guide the transcript assembly process with the -G option. In addition, we increased the minimum coverage to 5 with the -c option. To obtain a unified transcriptome, we used the StringTie merge mode with the -G option to assemble all the novel transcripts that were discovered across all the biological replicates with the reference transcript models in the Ensembl annotation (Pertea et al., 2016).

### 2.4. Construction of an expanded annotation based on short-read analysis

To accurately classify the StringTie predicted novel transcripts, we reassigned each novel transcript to a source gene based on the matched exon numbers and overlapping regions with known transcript models. To this end, each novel transcript was compared with all the known transcript variants to obtain the number of overlapped exons and the length of overlapping regions. Then, based on these results, each novel transcript was assigned to a parent gene with the most matched exons and the highest overlap in exonic regions. Transcripts that overlapped with none of the known transcript models were considered to be novel genes possibly reflecting previously unidentified genes. After the assignment process, the novel transcript variants and novel genes were merged with the original Ensembl annotation to generate an expanded reference annotation for the following long-read analysis pipeline. Lastly, to further polish the annotation, all novel transcripts and genes without strand specificity were removed.

### 2.5. Long-read RNA-seq data analysis

The FLnc reads from PacBio IsoSeq analyses were first aligned to the Ensembl genome assembly (bTaeGut1_v1.p, INSDC Assembly GCA_003957565.2) using Minimap2 v2.17 (Li and Birol, 2018), with default parameters. In addition, a specific output option –MD was used, as suggested by the TALON pipeline. The read alignment file was then passed to TranscriptClean, with default parameters, to correct mismatches, microindels, and noncanonical splice junctions in mapped long reads (Wyman et al., 2019). The resulted SAM file served as an input to the main TALON pipeline, along with the short reads-based expanded annotation from previous steps. To assign each long read to an annotated transcript, the parameter –cov and –identity were both set to 1.0 to increase the accuracy and reliability of long-read annotation. To further increase the reliability of novel transcript models, the TALON annotator results were filtered by the talon_filter_transcripts module with the specified parameter –minCount = 2 based on the average read coverage per unique transcript model. After this filtering step, only novel transcript models supported by at least two FLnc reads without evidence of internal priming were retained. Lastly, a read count matrix of the filtered transcriptome was extracted using the talon_abundance module. To organize the long-read annotation results, a GTF-formatted annotation for transcripts and genes supported by long reads was generated using the talon_create_GTF utility based on the filtered transcriptome and the reference annotation (Wyman et al., 2020).

### 2.6. Generation of the final annotation

The final annotation was created by combining novel transcripts/genes with Ensembl annotation (bTaeGut1_v1.p). Specifically, two types of novel transcripts/genes were included. The first type consists of those

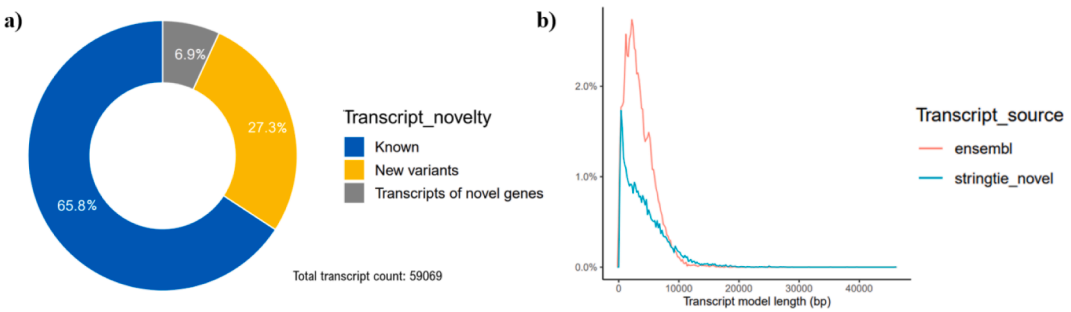


**Fig. 2. Short-read based expanded annotation.** a) Composition of StringTie assembled transcriptome based on transcript novelty classification after transcript-gene assignment. Among 59,069 StringTie assembled stranded transcript models, ~65.8 % are known transcripts from Ensembl annotation. 27.3 % of the transcripts are new transcript variants of known genes, and 6.9 % are novel transcripts that do not have an assignable source gene. b) Length distribution of transcript models from Ensembl annotation and StringTie prediction. The vast majority of the transcript models are less than 10 Kbp long.

supported by both short reads (StringTie) and long reads, whereas the second type includes those supported by long reads only.

### 2.7. Quantification of transcripts in the final annotation

Transcript quantification of Illumina-based RNA-seq data was performed using Kallisto v0.45.0 (Bray et al., 2016). To provide the FASTA formatted file for kallisto indexing, the nucleotide sequences of the transcripts in the final annotation were extracted from zebra finch genome assembly bTaeGut1_v1.p (GCA_003957565.2) using gffread v0.11.6 (Pertea and Pertea, 2020). A kallisto index was then built using the FASTA file with default parameters. The two technical replicates for each sample were merged for Kallisto transcript quantification.

### 2.8. ORF/peptide prediction

The ORF and peptide predictions were carried out using TransDecoder v5.5.0 (Haas et al., 2013). Only ORFs longer than 100 amino acids were retained. To obtain the optimal ORFs that may have functional significance, a universal BlastP (v 2.8.1 + ) search against the UniProtKB/Swiss-Prot database was carried out with suggested parameters in the TransDecoder manual (Camacho et al., 2009; The UniProt Consortium, 2019). The BlastP search output was used in the TransDecoder. Predict step with the –single_best_only option, which allowed one single best ORF for a likely coding region to be retained for each transcript. Lastly, a protein sequence was predicted based on the retained ORF.

### 2.9. Zebra finch brain tissue preparation

Brain samples were collected in two ways. Telencephalons of two males were flash-frozen in a fume hood at the University of Massachusetts following procedures followed for RNA isolation as in (Remage-Healey et al., 2009). Birds were sacrificed by rapid decapitation within 1 min of initial disturbance and capture. Brains were immediately dissected on ice-cold petri dish to isolate the left (odd-number samples) and right (even-number samples) posterior telencephalon (1.5 mm from caudal edge) that contains the NCM. Dissected brains were immediately frozen on dry ice (less than2 min from sacrifice to flash freezing of sample) and stored at −80 °C until assay. These samples were used to validate Mass-Spec procedures described below.

NCM samples for analysis were obtained from 3 male and 3 female adult zebra finches from a colony at UCLA across 3 separate days. Birds were obtained from large single-sex aviaries cages and no single cage was entered more than once on a given day to reduce potential stress to occupants. After capture in the dark, birds were euthanized via rapid decapitation within 35 s of lights out. We followed previously described methods to isolate the bilateral NCM (Rensel et al., 2018). Briefly, upon extraction from the skull the cerebellum was removed and the telencephalon situated on a petri dish on wet ice. Using a razor blade, we removed the rostral telencephalon, then made 2 parasagittal cuts ~ 1 mm from the midline to isolate the hippocampus, which was carefully peeled from the remaining caudal telencephalon. This exposed the caudal protuberances of the bilateral NCM, which were removed by taking ~ 1 $mm^2$ of tissue with forceps. Samples were immediately frozen on dry ice and stored at −80 °C until further processing.

### 2.10. Mass-spectrometry sample preparation and protein identification

Using previously described procedures (Capri and Whitelegge, 2017), brain tissue was homogenized in lysis buffer (200 uL, 12 mM sodium lauroyl sarcosine, 0.5 % sodium deoxy-cholate, 50 mM triethylammonium bicarbonate (TEAB), and HaltTM Protease Inhibitor Cocktail (Thermo Scientific, Waltham, MA)). Then, the samples were transferred to new tubes before being subjected to bath sonication (10 min) and heated (95 °C, 5 min). An aliquot of the resulting solution (9 uL) was taken for measurement of total protein concentration (bicinchoninic acid assay; Micro BCA Protein Assay Kit, Thermo Fisher Scientific, using BSA as a standard). The remainder of each sample was treated with tris (2-carboxyethyl) phosphine (20 uL, 55 mM in 50 mM TEAB, 30 min, 37 °C) followed by treatment with chloroacetamide (20 uL, 120 mM in 50 mM TEAB, 30 min, 25 °C in the dark). They were then diluted 5-fold with aqueous 50 mM TEAB, and incubated overnight and then 3 h with Sequencing Grade Modified Trypsin (2.0 ug in 20 uL of 50 mM TEAB; Promega, Madison, WI). An equal volume of ethyl acetate/trifluoroacetic acid (TFA, 100/1, v/v) was added before vigorous mixing (5 min) and centrifugation (13,000 × g, 5 min), in order to discard the supernatants and dry the lower phases in a centrifugal vacuum concentrator. The samples were then desalted using a modified version of Rappsilber's protocol in which the dried samples were reconstituted in acetonitrile/water/TFA (solvent A, 100 uL, 2/98/0.1, v/v/v) and then loaded onto a small portion of a C18-silica disk (3 M, Maplewood, MN) placed in a 200 uL pipette tip. Prior to sample loading the C18 disk was prepared by sequential treatment with methanol (20 uL), acetonitrile/water/TFA (solvent B, 20 uL, 80/20/0.1, v/v/v) and finally with solvent A (20 uL). After loading the sample, the disc was washed with solvent A (20 uL, eluent discarded) and eluted with solvent B (40 uL). The collected eluent was dried in a centrifugal vacuum concentrator and reconstituted in water/acetonitrile/FA (solvent E, 10 uL, 98/2/0.1, v/v/v), and aliquots (5 uL) were injected onto a reverse phase nanobore HPLC column (AcuTech Scientific, C18, 1.8um particle size, 360 um × 20 cm, 150 um ID), equilibrated in solvent E and eluted (500 nL/min) with an increasing concentration of solvent F (acetonitrile/water/FA, 98/2/0.1, v/v/v: min/% F; 0/0, 5/3, 18/7, 74/12, 144/24, 153/27, 162/40, 164/80, 174/80, 176/0, 180/0) using an Eksigent NanoLC-2D system (Sciex (Framingham, MA)). The effluent from the column was directed to a nanospray ionization source connected to a hybrid quadrupole-Orbitrap mass spectrometer (Q Exactive Plus, Thermo Fisher Scientific) acquiring mass spectra in a data-dependent mode

**Table 1**
**Summary of expanded annotation in comparison to reference Ensembl annotation.** 4,099 novel genes were added to expanded annotation resulting a total gene count of 26,249. The total number of transcripts has raised to 59,077 from 38,869 in the expanded annotation. The average transcript variants per gene has increased to 2.25 after annotation reconstruction based on short-read analysis.

| | Expanded annotation | Ensembl annotation (bTaeGut1_v1.p) |
|---|---|---|
| Total gene count | 26,249 | 22,150 |
| Total transcript count | 59,077 | 38,869 |
| Transcript per gene | 2.25 | 1.75 |

alternating between a full scan ($m/z$ 300–1700, automated gain control (AGC) target $3 \times 106$, 100 ms maximum injection time, FWHM resolution 70,000 at $m/z$ 200) and up to 20 MS/MS scans (quadrupole isolation of charge states $\geq 2$, isolation width 15 s) for a runtime of 180 min. The raw data were processed by Mascot 2.4, which identified proteins using tryptic peptides containing amino acid sequences unique to individual proteins in the zebra finch proteome predicted from our pipeline.

## 3. Results

### 3.1. Construction of expanded annotation based on novel discoveries from short-read data

We first carried out *de novo* transcriptome assembly using StringTie and the short-read RNA-seq data (Methods). Compared with Ensembl-

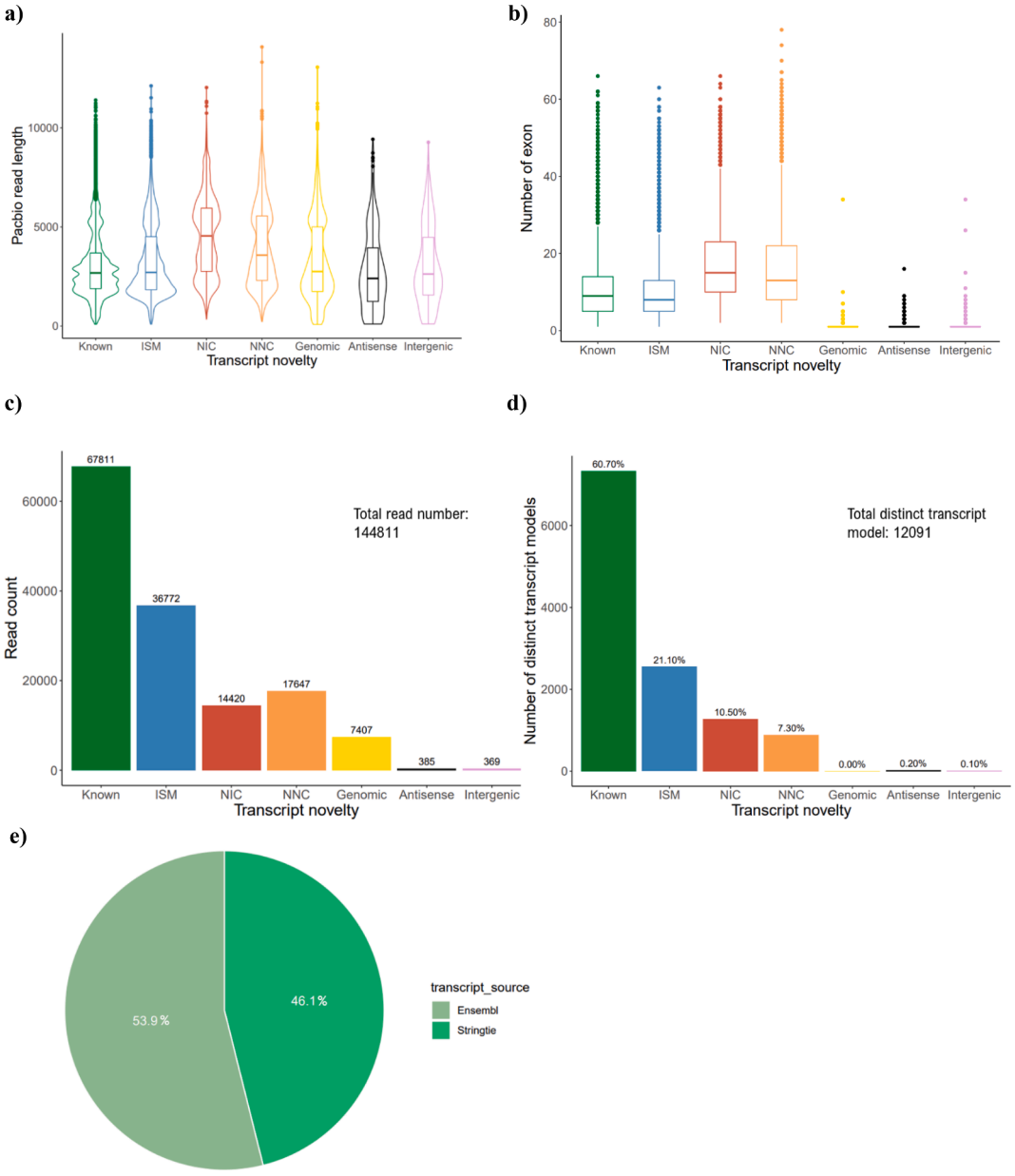


**Fig. 3. Talon annotation results using expanded annotation as reference.** a) Distribution of annotated PacBio FLnc read length in each TALON transcript novelty category; b) Exon num-bers of annotated reads in the TALON transcript novelty categories; c) Number of reads assigned to each transcript novelty category; d) Number of distinct transcript models in each transcript novelty category; e) Source composition of distinct transcript models assigned to "Known" transcript novelty category.

**Table 2**
**Content summary and comparison between final annotation and Ensembl annotation.** In comparison with the Ensembl annotation, the total gene count has increased to 22,370 and the total transcript count has increased to 47,003 in the final annotation. The average transcript number per gene has also raised to 2.1 from 1.75 per gene.

| | Final Annotation | Ensembl annotation (bTaeGut1_v1.p) |
|---|---|---|
| Gene count | 22,370 | 22,150 |
| Transcript count | 47,003 | 38,869 |
| Transcript per gene | 2.1 | 1.75 |

annotated transcripts, StringTie predicted a total of 20,614 novel transcripts. Among these, 16,109 were assigned to Ensembl-annotated genes, thus considered novel transcript variants of known genes. The remaining 4,505 novel transcripts did not match any known genes, thus representing potentially novel genes. The novel transcripts from both known and novel genes were merged with Ensembl annotation to create an expanded annotation (Fig. 2a). It should be noted that 406 novel genes were removed due to a lack of strand information.

This expanded annotation contains a total of 26,249 genes and 59,077 transcripts, with 2.25 transcripts per gene on average (compared to 1.75 in Ensembl annotation) (Table 1). The length distributions of transcript models from Ensembl annotations and StringTie predictions are shown in Fig. 2b. Although the majority of transcripts in both categories are less than 10 kb long, StringTie-predicted novel transcripts tend to be relatively shorter than Ensembl transcripts, which is likely due to incomplete transcript coverage in the RNA-seq. This observation indicates that additional scrutiny is needed to further examine the novel transcripts, as presented below.

### 3.2. TALON annotation of PacBio FLnc reads

In order to provide additional evidence to support the novel transcripts discovered from the short-read RNA-seq data, we analyzed PacBio FLnc reads using the TALON pipeline. The above expanded annotation was used as an input annotation file for TALON (including Ensembl known and StringTie-predicted novel transcripts). Among a total of 405,837 aligned FLnc reads, TALON successfully annotated 144,811 reads based on the given annotation file. Among the 144,811 reads, 67,811 were categorized as "Known" (defined as those that match transcripts in the expanded annotation). Note that some of the "Known" reads correspond to StringTie predicted novel transcripts relative to Ensembl. TALON also annotated other reads into the following categories: "Incomplete splice match (ISM)", "Novel in catalog (NIC)", "Novel not in catalog (NNC)", "Genomic", "Antisense", and "Intergenic" (Li and Birol, 2018).

The median length of the reads in the seven categories varied from ~ 2,500 bp to ~ 5,000 bp. In particular, NIC and NNC had higher medians than other categories (Fig. 3a). Consistent with the read length observation, the numbers of exons in the NIC and NNC categories were also the highest among all categories (Fig. 3b). About 47 % of the annotated reads corresponded to "Known" transcripts, which is the largest category in read numbers, followed by ISM, where the FLnc reads partially match known transcript isoforms (Fig. 3c). Next, we required a minimum of 2 FLnc reads for each novel transcript model to increase the stringency in novel transcript definition. As a result, 12,090 distinct transcript isoforms expressed from 6,520 genes were retained. Among these transcripts, 61 % fell into the "Known" transcript category (Fig. 3d). Strikingly, nearly half of the "Known" transcript models were defined by StringTie based on the short-read data, but not included in the Ensembl annotation (Fig. 3e).

### 3.3. Construction of final annotation and transcript quantification

Based on the short-read and long-read analysis results, we added 220 novel genes and 8,134 novel transcripts to the Ensembl annotation, resulting in a total of 22,370 genes and 47,003 transcripts in the final annotation file. Compared to the original Ensembl annotation, the average number of transcripts per gene increased from 1.75 to 2.1 (Table 2). Among all the novel genes and transcripts, 80 % genes and 41.6 % transcripts were supported by StringTie predictions (and by the long reads), whereas 20 % genes and 58.4 % transcripts were supported by long reads alone (Fig. 4).

To further characterize the predicted novel transcripts, we calculated their expression levels (transcript per million, TPM) in the short-read data via Kallisto. Note that we did not use TPM from long-read data as many transcripts had relatively low coverage following our stringent filters. For novel transcripts identified by StringTie, their TPM values in short-read RNA-seq were highly correlated across the 7 biological replicates (Fig. 5a). Next, we compared the expression levels of novel and known transcripts. As shown in Fig. 5b, novel transcripts from StringTie and TALON demonstrated higher expression levels compared to the Ensembl known transcript models. This observation is likely due to the existence of unexpressed genes among the Ensembl annotated genes. Nonetheless, the relatively high expression levels of the novel transcripts strongly support their validity.

### 3.4. Peptide prediction of transcripts in the final annotation

For each transcript in our final annotation, a single best protein peptide sequence was predicted by TransDecoder. Among the 48,003 transcript models in the final annotation, TransDecoder successfully predicted a likely coding region for 42,818 transcripts. In comparison

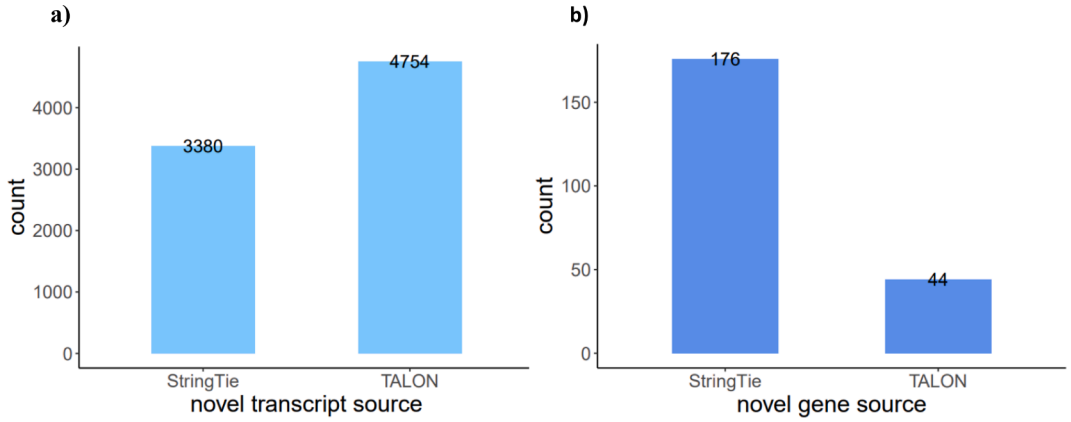


**Fig. 4. Source composition of novel transcripts and genes in the final annotation.** a) Among 8,134 novel transcripts, 3380 were from StringTie assembly, and 4754 from TALON annotation. b) Among 220 novel genes, 176 were identified from StringTie assembly, and 44 were based on TALON annotation.

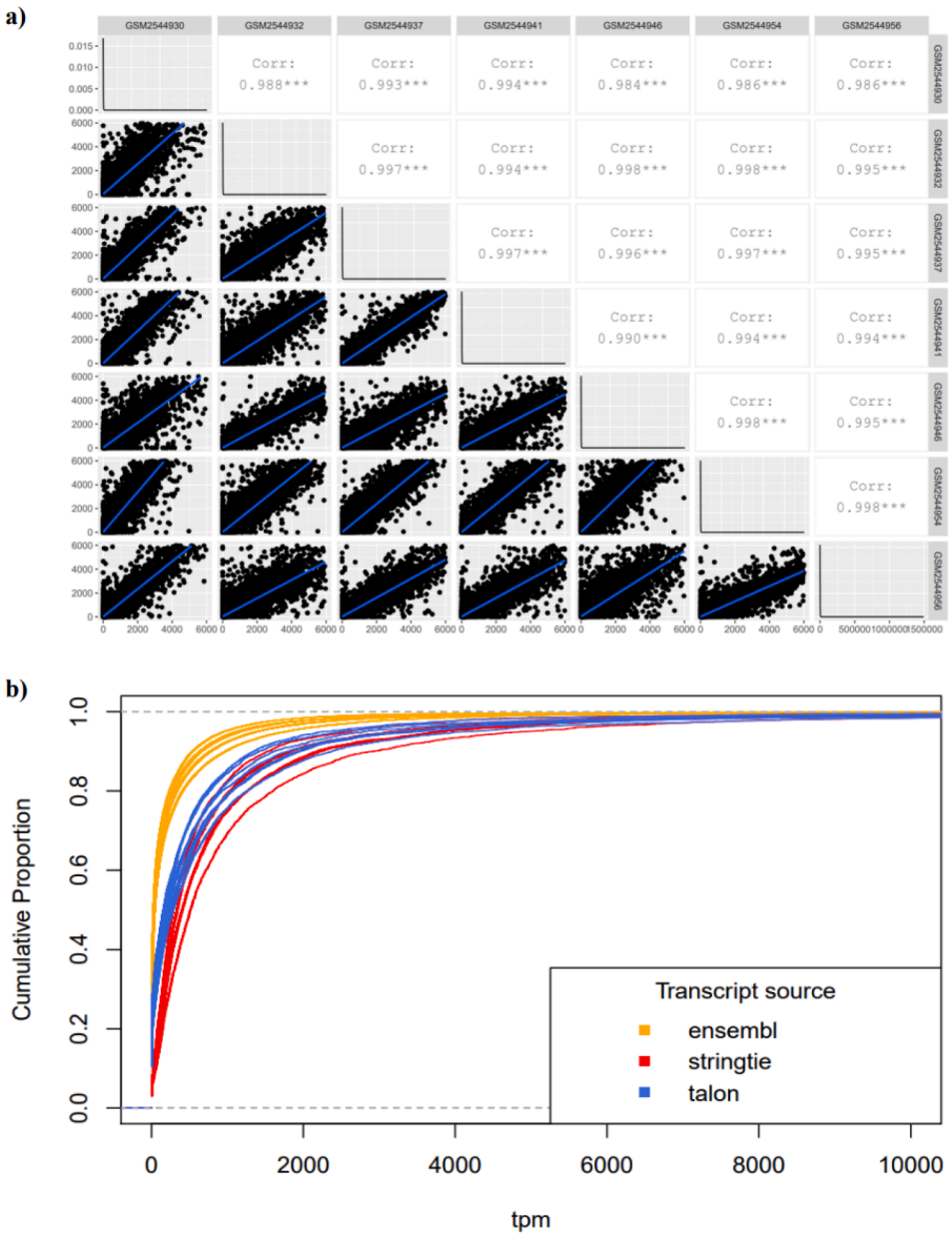


**Fig. 5. Transcript quantification based on final annotation.** a) Transcript expression correlation among 7 zebra finch shot-read samples. b) ECDF plot of transcript TPM of all transcript models in the final annotation.

with the UniProt zebra finch proteome (Proteome ID: UP000007754), the total number of proteins increased by 11,477 (~37 %) in the TransDecoder proteome. The distribution of protein peptide length is consistent in the two proteome files, where most of the peptides are within 1,000 amino acids long (Fig. 6).

### 3.5. Novel protein validation

To further access the reliability of the proteome prediction, we generated mass-spectrometry data from zebra finch NCM and performed data analysis. Following preliminary validation of the Mass-spec procedures on NCM samples, we specifically examined the 3 male and 3 female caudal telencephalon samples to assess the presence of novel peptides. Upon initial examination, 57 novel proteins were identified. Thirty-four of these were sex-specific of which 18 were identified from two independent peptides per sample with an ion score close to or larger than the ion score at a 95% confidence level. Because of previously identified sexually dimorphic features of NCM anatomy and function (Nottebohm and Arnold, 1976; Remage-Healey et al., 2009; Peterson et al., 2005; Remage-Healey et al., 2012), we focused our attention on these 18 that were fully sex-specific, that is 3 were present in all 3 female NCM samples and 15 were found in all 3 male NCM samples (Table 3). BlastP results with highest percent identity provided molecular and functional information of the predicted protein (Table 4).

### 3.6. Sex-specific transcripts exhibit specialized expression in male song control nuclei

To further explore the potential relevance of the novel proteins to the vocal learning trait, 14 of the 18 sex-specific novel proteins were compared to RNA-seq data from each of the four principal zebra finch song control nuclei (G. Gedman et al, unpublished observations in preparation). These are sexually-dimorphic cortical regions known as HVC (acronym used as proper name), the lateral magnocellular nucleus of the anterior nidopallium (LMAN) and the robust nucleus of the arcopallium (RA), in additional to basal ganglia region Area X. All data

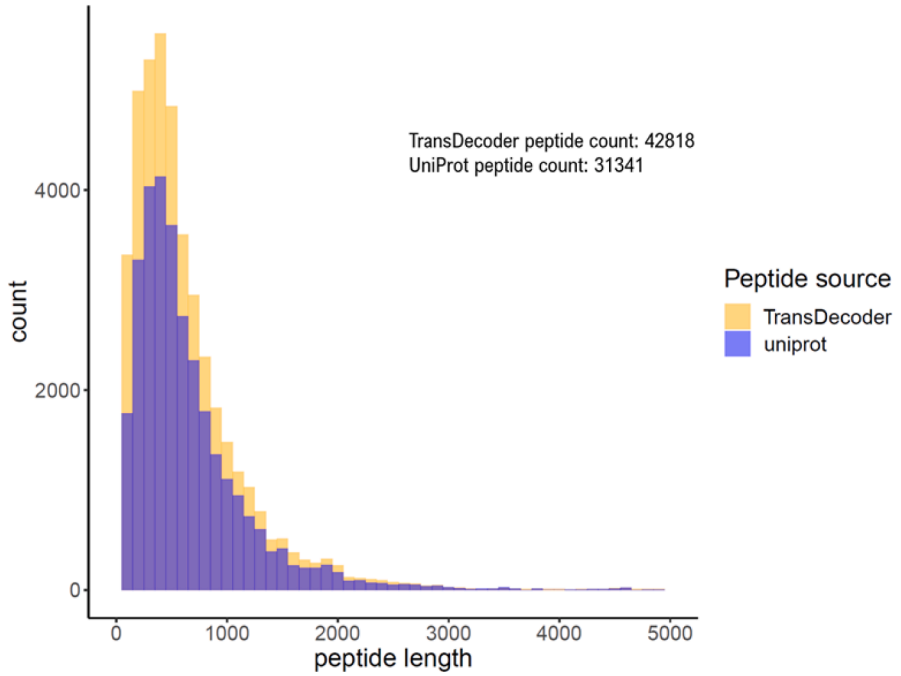


**Fig. 6. Peptide length distribution in TransDecoder predicted proteome and UniProt zebra finch proteome.** Histogram of the distribution of peptide length (bin width = 100).

were quantified using the recently improved reference annotation (bTaeGut1_v1.p) and tested for differential expression (DE, adjusted. pvalue less than 0.05) relative to the surrounding non-vocal motor brain tissue for each nucleus. Table 5 displays each gene coded by its DE status: 1 denotes significant upregulation, −1 denotes significant downregulation, and 0 denotes no significant difference. One gene (ATP6V0A1) was trending towards significant downregulation after multiple test corrections (p.adjust = 0.085) and is denoted with a −0.5. Ten out of 14 (70 %) novel transcripts examined exhibited specialized expression in at least one song control nucleus, highlighting their importance to this male-specific behavior in this species.

## 4. Discussion

The release of the zebra finch genome provided a powerful tool for complex avian genomic studies via high-throughput approaches. The revised zebra finch genome annotation released in 2019 from Ensembl provided a more comprehensive database. Our study utilized publicly available high-throughput RNA-seq data generated from both NGS and TGS sequencing technologies to investigate this new bird genome. Our analysis has uncovered the existence of previously undetected novel transcripts and genes which expand the songbird genome annotation.

PacBio SMRT sequencing technology is increasingly utilized for novel transcript discovery and transcriptome profiling. Meanwhile, the computational tools and algorithms to assemble short reads generated from widely used NGS sequencing technologies are more accurate and robust. With the availability of different sequencing methods and analysis tools, previous studies indicated the advantages of employing hybrid RNA-seq analysis approaches for transcriptome profiling (Sahraeian et al., 2017; Wang et al., 2019). Although many studies have presented effective analysis workflows that start with novel transcript discovery from PacBio FLnc reads, followed by validation using Illumina short-read data, our pipeline applied an alternative hybrid approach, which used long-read data to support short-read assembly discovery, to identify novel transcripts/genes.

Using StringTie on short-reads, we detected a large number (20,614) of novel transcripts, representing a 53 % increase relative to the reference annotation. Many of these transcripts were matched to known genes. However, false positives are expected to exist. Thus, we used PacBio FLnc reads to further support novel transcripts. Since we required precise matches at the exon boundaries between the FLnc reads and StringTie-predicted novel transcripts, the resulted novel transcript models are highly confident, supported by two sequencing modalities. Among a total of 12,090 long-read supported transcripts models, it is striking that 60.7 % belonged to the "Known" category (TALON definition), among which nearly half were novel transcripts predicted by StringTie (novel relative to Ensembl). This observation suggests the presence of a considerable number of unannotated genes or transcripts that are highly expressed in the bird brain. Besides the distinct transcript models in the "Known" category, ~39 % falls into "ISM", "NIC", and "NNC" novel transcript categories, and less than 0.5 % falls in "Antisense" and "Intergenic" categories. The ISM, NIC, and NNC categories were first defined by the SQANTI long read classification pipeline (Tardaguila et al., 2018), in which the transcript classification is based on their splice junctions compared to parent gene isoforms. The high percentage of long reads-based novel transcript models in NIC and NNC categories reflects a prominent advantage of full-length transcripts – clear definition of transcript structures. Interestingly, less than 0.1 % of the distinct transcript models fell into the intergenic category, an indication of improvement of our short reads-based expanded annotation.

In the final annotation, the novel transcripts supported by both long and short reads are highly confident, as supported by their relatively high expression levels. To illustrate the feasibility and validity of capturing and predicting proteins from our hybrid approach utilizing NGS and TGS RNA-seq data, we successfully identified 18 novel proteins in zebra finch brain tissues via Mass Spec. Among these, ten exhibited specialized expression in at least one of the zebra finch male song control nuclei. Interestingly, many of those male-specific proteins have involvement in neuroplasticity, a hallmark of the oscine brain or with neurotransmission (Rundstrom and Creanza, 2021). One gene, CaMKII-alpha, is an excellent marker for excitatory neurons in zebra finch NCM: AAVs targeting this promoter were able to selectively infect neurons with principal-neuron physiology and behavior (Spool et al., 2021). The identification of novel proteins provides additional information to investigate bird song learning and memory.

Despite the novel findings from our analysis, it is noteworthy that the

**Table 3**
Novel proteins validated by mass-spectrometry. 18 novel proteins were identified with each detected in three brain samples from at least two distinct peptides with an ion score close to or larger than the ion score at a 95% confidence level. BlastP search results with the highest percent identity for each protein were listed.

| Protein Accession | Sex | Tissue Sample | ion score (*p* less than 0.05) | Distinct peptide sequences with highest ion score | Ion score | BlastP (Accession/protein/species) |
|---|---|---|---|---|---|---|
| MSTRG.13722.1 | Female | 1 | 38 | K.LGMLDPDELDKDGMPLTAR.V | 43 | NP_001232302.1\| peroxiredoxin-6 [Taeniopygia guttata] |
| | | | | M.PGLLLGDEAPDFEADTTQGR.I | 29 | |
| | | 2 | 38 | K.LGMLDPDELDKDGMPLTAR.V | 60 | |
| | | | | M.PGLLLGDEAPDFEADTTQGR.I | 35 | |
| | | 3 | 39 | M.PGLLLGDEAPDFEADTTQGR.I | 49 | |
| | | | | K.LGMLDPDELDKDGMPLTAR.V | 45 | |
| MSTRG.2270.1 | Female | 1 | 38 | R.DLSAGIGLLAAATQSLNMPASLGR.M | 89 | XP_030139471.1\| matrin-3 isoform X1 [Taeniopygia guttata] |
| | | | | K.SVLEKPLVPDEFR.I | 31 | |
| | | 2 | 38 | R.DLSAGIGLLAAATQSLNMPASLGR.M | 111 | |
| | | | | K.SVLEKPLVPDEFR.I | 39 | |
| | | 3 | 39 | R.DLSAGIGLLAAATQSLNMPASLGR.M | 73 | |
| | | | | K.SVLEKPLVPDEFR.I | 32 | |
| MSTRG.2422.1 | Female | 1 | 38 | K.MCDPGMTAFEPEALGNLVEGLDFHR.F + Oxidation (M) | 82 | XP_002198859.1\| calcium/calmodulin-dependent protein kinase type II subunit alpha isoform X2 [Taeniopygia guttata] |
| | | | | R.ITQYVDSGGIPR.T | 59 | |
| | | 2 | 38 | K.MCDPGMTAFEPEALGNLVEGLDFHR.F | 80 | |
| | | | | R.DLKPENLLLASK.L | 30 | |
| | | 3 | 39 | K.MCDPGMTAFEPEALGNLVEGLDFHR.F | 80 | |
| | | | | K.VTEQLIEAISNGDFESYTK.M + Deamidated (NQ) | 62 | |
| MSTRG.11525.3 | Male | 1 | 38 | R.TAGPVLVSLR.Q | 58 | XP_030805072.1\|coronin-1B [Camarhynchus parvulus] |
| | | | | R.FSAPPVLGSGSATGGRLEEVLEEVAALR.A | 46 | |
| | | 2 | 38 | R.TAGPVLVSLR.Q | 49 | |
| | | | | R.LEEVLEEVAALR.A | 32 | |
| | | 3 | 38 | R.TAGPVLVSLR.Q | 62 | |
| | | | | R.FSAPPVLGSGSATGGRLEEVLEEVAALR.A | 46 | |
| MSTRG.13685.1 | Male | 1 | 38 | R.LTGFHETSNINEFSAGVANR.G | 77 | XP_030135503.2\| glutamine synthetase isoform X1 [Taeniopygia guttata] |
| | | | | R.RLTGFHETSNINEFSAGVANR.G | 66 | |
| | | 2 | 38 | R.LTGFHETSNINEFSAGVANR.G | 77 | |
| | | | | R.LTGFHETSNINEFSAGVANR.G + Deamidated (NQ) | 82 | |
| | | 3 | 38 | R.LTGFHETSNINEFSAGVANR.G + Deamidated (NQ) | 77 | |
| | | | | R.RLTGFHETSNINEFSAGVANR.G | 74 | |
| MSTRG.14674.1 | Male | 1 | 38 | K.VNVLDVAVLDQVEAR.L | 38 | XP_021406292.1\| dynactin subunit 2 isoform X6 [Lonchura striata domestica] |
| | | | | R.LALLEAAVR.C | 35 | |
| | | 2 | 38 | K.VNVLDVAVLDQVEAR.L | 49 | |
| | | | | K.TMKDNLAIVEDNFADIDAR.I | 30 | |
| | | 3 | 38 | K.VNVLDVAVLDQVEAR.L | 67 | |
| | | | | R.LALLEAAVR.C | 42 | |
| MSTRG.1982.2 | Male | 1 | 38 | -.MEAVDQLASAGTFR.V + Acetyl (Protein Nterm) | 98 | XP_002190706.2\| synaptoporin isoform X1 [Taeniopygia guttata] |
| | | | | K.VATDPDEVLLLMSACK.Q | 65 | |
| | | 2 | 38 | -.MEAVDQLASAGTFR.V + Acetyl (Protein Nterm) | 95 | |
| | | | | K.VATDPDEVLLLMSACK.Q | 60 | |
| | | 3 | 38 | -.MEAVDQLASAGTFR.V + Acetyl (Protein Nterm); Oxidation (M) | 84 | |
| | | | | K.VATDPDEVLLLMSACK.Q | 94 | |
| MSTRG.2070.2 | Male | 1 | 38 | R.LQVLEQDVVLQSIDR.A | 96 | XP_012428099.3\|NCK-interacting protein with SH3 domain isoform X1 [Taeniopygia guttata] |
| | | | | R.LTDLQGILQR.I | 44 | |
| | | 2 | 38 | R.LQVLEQDVVLQSIDR.A | 96 | |
| | | | | R.LTDLQGILQR.I | 55 | |
| | | 3 | 38 | R.LQVLEQDVVLQSIDR.A | 115 | |
| | | | | R.LTDLQGILQR.I | 49 | |
| MSTRG.2607.1 | Male | 1 | 38 | K.KLTPITYPQGLAMAK.E | 48 | 5FI0_B\| Crystal Structure Of Rac1 In Complex With The Guanine Nucleotide Exchange Region Of Tiam1 [Homo sapiens] |
| | | | | R.HHCPNTPIILVGTK.L | 35 | |
| | | 2 | 38 | K.YLECSALTQR.G | 38 | |

(*continued on next page*)

**Table 3** (*continued*)

| Protein Accession | Sex | Tissue Sample | ion score (*p* less than 0.05) | Distinct peptide sequences with highest ion score | Ion score | BlastP (Accession/protein/species) |
|---|---|---|---|---|---|---|
| | | | | K.KLTPITYPQGLAMAK.E | 38 | |
| | | 3 | 38 | K.KLTPITYPQGLAMAK.E | 41 | |
| | | | | K.YLECSALTQR.G | 37 | |
| MSTRG.2669.2 | Male | 1 | 38 | R.LVPAAEPGTAFTLEFR.C | 81 | XP_005523080.1\| PREDICTED: fascin [Pseudopodoces humilis] |
| | | | | K.VGKDELFALEQSCPQVVLR.A | 67 | |
| | | 2 | 38 | K.VGKDELFALEQSCPQVVLR.A | 81 | |
| | | | | R.LVPAAEPGTAFTLEFR.C | 65 | |
| | | 3 | 38 | K.VGKDELFALEQSCPQVVLR.A | 89 | |
| | | | | R.LVPAAEPGTAFTLEFR.C | 71 | |
| MSTRG.2698.1 | Male | 1 | 38 | K.GIEFPMADLDALSPIHTPQR.S | 47 | XP_012433002.1\|TOM1-like protein 2 isoform X2 [Taeniopygia guttata] |
| | | | | K.VMSEMLTEMVPGQEDSSDLELLQELNR.T | 38 | |
| | | 2 | 38 | K.GIEFPMADLDALSPIHTPQR.S | 40 | |
| | | | | R.VSNEEVTEELLHVNDDLNNVFLR.Y | 31 | |
| | | 3 | 38 | R.VSNEEVTEELLHVNDDLNNVFLR.Y | 61 | |
| | | | | K.VMSEMLTEMVPGQEDSSDLELLQELNR.T | 59 | |
| MSTRG.3482.2 | Male | 1 | 38 | K.IIFVVGGPGSGK.G | 53 | XP_030142137.1\| adenylate kinase isoenzyme 1 isoform X2 [Taeniopygia guttata] |
| | | | | K.LQAIMEKGELVPLDTVLDMLR.D | 51 | |
| | | 2 | 38 | K.IIFVVGGPGSGK.G | 42 | |
| | | | | K.LQAIMEKGELVPLDTVLDMLR.D | 38 | |
| | | 3 | 38 | K.IIFVVGGPGSGK.G | 48 | |
| | | | | K.LQAIMEKGELVPLDTVLDMLR.D | 48 | |
| MSTRG.5045.2 | Male | 1 | 38 | R.VNEAREELMR.M | 35 | NP_001019397.1\| ADP-ribosylation factor 1 [Homo sapiens] |
| | | | | K.KEMRILMVGLDAAGK.T | 30 | |
| | | 2 | 38 | R.ILMVGLDAAGK.T | 35 | |
| | | | | R.VNEAREELMR.M | 33 | |
| | | 3 | 38 | K.QDLPNAMNAAEITDKLGLHSLR.H | 42 | |
| | | | | R.ILMVGLDAAGK.T | 41 | |
| MSTRG.8035.4 | Male | 1 | 38 | R.DLNPDVNVFHR.K | 78 | XP_002197932.1\| V-type proton ATPase 116 kDa subunit a isoform X6 [Taeniopygia guttata] |
| | | | | K.KANIPIMDTGENPEVPFPR.D | 68 | |
| | | 2 | 38 | K.ANIPIMDTGENPEVPFPR.D | 88 | |
| | | | | R.DLNPDVNVFHR.K | 73 | |
| | | 3 | 38 | K.ANIPIMDTGENPEVPFPR.D | 84 | |
| | | | | R.DLNPDVNVFHR.K | 78 | |
| MSTRG.896.1 | Male | 1 | 38 | K.GPLVMELQTYR.Y | 61 | XP_002196835.1\| pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial isoform X2 [Taeniopygia guttata] |
| | | | | R.LEEGPGTTAVMTR.E | 54 | |
| | | 2 | 38 | R.LEEGPGTTAVMTR.E | 81 | |
| | | | | K.GPLVMELQTYR.Y | 42 | |
| | | 3 | 38 | R.LEEGPGTTAVMTR.E | 56 | |
| | | | | R.MVNNNLASVEELKEIDVAVR.K | 51 | |
| TALONT000068760 | Male | 1 | 38 | K.DLYANTVLSGGTTMYPGIADR.M | 113 | OPJ88443.1\| actin, cytoplasmic 2 [Patagioenas fasciata monilis] |
| | | | | K.SYELPDGQVITIGNER.F | 93 | |
| | | 2 | 38 | K.SYELPDGQVITIGNER.F | 98 | |
| | | | | R.KDLYANTVLSGGTTMYPGIADR.M | 104 | |
| | | 3 | 38 | R.KDLYANTVLSGGTTMYPGIADR.M | 102 | |
| | | | | K.DLYANTVLSGGTTMYPGIADR.M | 113 | |
| TALONT000074596 | Male | 1 | 38 | K.LASASTIDHAR.H | 56 | XP_030119865.1\|putative myelin basic protein variant 2 isoform X2 [Taeniopygia guttata] |
| | | | | R.VSHHVGSIPPR.S | 56 | |
| | | 2 | 38 | K.LASASTIDHAR.H | 57 | |
| | | | | R.HRDSGLLDSLGR.F | 47 | |
| | | 3 | 38 | K.LASASTIDHAR.H | 59 | |
| | | | | R.HRDSGLLDSLGR.F | 43 | |
| TALONT000093126 | Male | 1 | 38 | R.LNIISNLDCVNEVIGIR.Q | 101 | XP_030118040.1\| serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform [Taeniopygia guttata] |
| | | | | K.TDLVGAFQSLMKDCEAEVR.A | 69 | |
| | | 2 | 38 | U R.LNIISNLDCVNEVIGIR.Q | 98 | |
| | | | | K.TDLVGAFQSLMKDCEAEVR.A | 65 | |
| | | 3 | 38 | R.LNIISNLDCVNEVIGIR.Q | 92 | |
| | | | | K.TDLVGAFQSLMKDCEAEVR.A | 65 | |

**Table 4**
Novel proteins validated by mass-spectrometry.

| Protein Accession | BlastP (Accession/protein/species) |
|---|---|
| MSTRG.13722.1 | NP_001232302.1\| peroxiredoxin-6 [Taeniopygia guttata] |
| MSTRG.2270.1 | XP_030139471.1\| matrin-3 isoform X1 [Taeniopygia guttata] |
| MSTRG.2422.1 | XP_002198859.1\| calcium/calmodulin-dependent protein kinase type II subunit alpha isoform X2 [Taeniopygia guttata] |
| MSTRG.11525.3 | XP_030805072.1\|coronin-1B [Camarhynchus parvulus] |
| MSTRG.13685.1 | XP_030135503.2\| glutamine synthetase isoform X1 [Taeniopygia guttata] |
| MSTRG.14674.1 | XP_021406292.1\| dynactin subunit 2 isoform X6 [Lonchura striata domestica] |
| MSTRG.1982.2 | XP_002190706.2\| synaptoporin isoform X1 [Taeniopygia guttata] |
| MSTRG.2070.2 | XP_012428099.3\|NCK-interacting protein with SH3 domain isoform X1 [Taeniopygia guttata] |
| MSTRG.2607.1 | *1FOE\| Crystal Structure Of Rac1 In Complex With The Guanine Nucleotide Exchange Region Of Tiam1 [Homo sapiens] |
| MSTRG.2669.2 | XP_005523080.1\| PREDICTED: fascin [Pseudopodoces humilis] |
| MSTRG.2698.1 | XP_012433002.1\|TOM1-like protein 2 isoform X2 [Taeniopygia guttata] |
| MSTRG.3482.2 | XP_030142137.1\| adenylate kinase isoenzyme 1 isoform X2 [Taeniopygia guttata] |
| MSTRG.5045.2 | NP_001019397.1\| ADP-ribosylation factor 1 [Homo sapiens] |
| MSTRG.8035.4 | XP_002197932.1\| V-type proton ATPase 116 kDa subunit a isoform X6 [Taeniopygia guttata] |
| MSTRG.896.1 | XP_002196835.1\| pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial isoform X2 [Taeniopygia guttata] |
| TALONT000068760 | OPJ88443.1\| actin, cytoplasmic 2 [Patagioenas fasciata monilis] |
| TALONT000074596 | XP_030119865.1\|putative myelin basic protein variant 2 isoform X2 [Taeniopygia guttata] |
| TALONT000093126 | XP_030118040.1\| serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform [Taeniopygia guttata] |

**Table 5**
Sex specific transcripts exhibit specialized expression in male song control nuclei. Each gene coded by its Differential Expression status: 1 denotes significant upregulation, −1 denotes significant downregulation, and 0 denotes no significant difference. One gene was trending towards significant downregulation after multiple test corrections (p.adjust = 0.085) and is denoted with a −0.5.

| Gene | AX | LMAN | HVC | RA |
|---|---|---|---|---|
| **CORO1B** | 0 | 1 | 0 | 0 |
| **GLUL** | 0 | 0 | 0 | 0 |
| **DCTN2** | −1 | 0 | 0 | 0 |
| **SYNPR** | 0 | −1 | −1 | −1 |
| **NCKIPSD** | 0 | −1 | 0 | 0 |
| **FSCN1** | 0 | 0 | 0 | 1 |
| **TOM1L2** | 0 | 0 | 0 | 0 |
| **AK1** | 0 | 0 | 0 | 0 |
| **ARF1** | 0 | 0 | 0 | 0 |
| **AT6V0A1** | 0 | −1 | −0.5 | 0 |
| **PDHA1** | 0 | 1 | 0 | 0 |
| **ACTB** | −1 | 0 | 0 | 0 |
| **MBP** | 1 | 1 | 1 | 0 |
| **PPP2R1A** | 0 | 0 | 0 | 1 |

long-read data we used were generated from only one bird sample. Additional long-read data sets will likely allow a more comprehensive transcriptome annotation. Since the vocal learning trait of zebra finch is age-sensitive and sex-specific (Jarvis, 2004), it would be interesting to analyze and compare data sets derived from animals of varying age and sex. Furthermore, since we implemented stringent filters in the TALON annotation step to ensure the accuracy of predicted novel transcripts, we might have missed many true positives. Future efforts in leveraging the improved zebra finch annotation, including validation and functional annotation of the predicted novel transcripts, will be highly significant.

In summary, our work identified 8,134 novel transcript variants and 220 novel genes relative to the Ensembl zebra finch annotation (bTaeGut1_v1.p). Moreover, we have predicted a new proteome based on the transcriptome from the final annotation, expanding the current zebra finch proteome by ~ 37 %. These findings represent a substantial improvement in the songbird gene annotation. In addition, our results corroborated the effectiveness of the hybrid RNA-seq analysis approach adopting both NGS and TGS RNA-seq methods.

## Supplementary materials

The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

## Institutional review board statement

All animal use was in accordance with NIH guidelines for experiments involving vertebrate animals, consistent with guidelines of the American Veterinary Medical Association and approved by the University of California at Los Angeles Chancellor's Institutional Animal Care and Use Committee and University of Massachusetts, Amherst.

## Funding

This research was funded by NIH Grant No RO1MH070712 to SAW; UCLA Academic Senate to BAS.

## CRediT authorship contribution statement

**Jingyan He:** Methodology, Software, Investigation, Resources, Writing – original draft, Visualization. **Ting Fu:** Software, Data curation. **Ling Zhang:** Software. **Lucy Wanrong Gao:** Methodology, Data curation, Data curation. **Michelle Rensel:** Resources. **Luke Remage-Healey:** Resources, Writing – review & editing. **Stephanie A. White:** Supervision, Funding acquisition. **Gregory Gedman:** Formal analysis, Writing – review & editing. **Julian Whitelegge:** Methodology, Validation, Writing – review & editing, Supervision. **Xinshu Xiao:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration. **Barney A. Schlinger:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data is available at: https://github.com/gxiaolab/ZebraFinchTranscriptome.

## References

Abril, J.F., Castellano, S., 2019. Genome Annotation. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, pp. 195–209. https://doi.org/10.1016/B978-0-12-809633-8.20226-4.

Balakrishnan, C.N., Lin, Y.-C., London, S.E., Clayton, D.F., 2012. RNA-seq transcriptome analysis of male and female zebra finch cell lines. Genomics 100 (6), 363–369. https://doi.org/10.1016/j.ygeno.2012.08.002.

Bolhuis, J.J., Gahr, M., 2006. Neural mechanisms of birdsong memory. Nat. Rev. Neurosci. 7 (5), 347–357. https://doi.org/10.1038/nrn1904.

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34 (5), 525–527. https://doi.org/10.1038/nbt.3519.

Burkett, Z.D., Day, N.F., Kimball, T.H., Aamodt, C.M., Heston, J.B., Hilliard, A.T., Xiao, X., White, S.A., 2018. FoxP2 isoforms delineate spatiotemporal transcriptional

networks for vocal learning in the zebra finch. ELife 7, e30649. https://doi.org/10.7554/eLife.30649.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: Architecture and applications. BMC Bioinf. 10 (1), 421. https://doi.org/10.1186/1471-2105-10-421.
Capri, J., Whitelegge, J.P., 2017. Full Membrane Protein Coverage Digestion and Quantitative Bottom-Up Mass Spectrometry Proteomics. Methods in Molecular Biology (Clifton N.J.) 1550, 61–67. https://doi.org/10.1007/978-1-4939-6747-6_6.
Chen, S.Y., Deng, F., Jia, X., Li, C., Lai, S.J., 2017. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Sci. Rep. 7 (1), 7648. https://doi.org/10.1038/s41598-017-08138-z.
Clayton, D.F., Balakrishnan, C.N., London, S.E., 2009. Integrating Genomes, Brain and Behavior in the Study of Songbirds. Current Biology : CB 19 (18), R865–R873. https://doi.org/10.1016/j.cub.2009.07.006.
Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O., Artiguenave, F., 2008. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 9 (12), R175. https://doi.org/10.1186/gb-2008-9-12-r175.
Deslattes Mays, A., Schmidt, M., Graham, G., Tseng, E., Baybayan, P., Sebra, R., Sanda, M., Mazarati, J.-B., Riegel, A., Wellstein, A., 2019. Single-Molecule Real-Time (SMRT) Full-Length RNA-Sequencing Reveals Novel and Distinct mRNA Isoforms in Human Bone Marrow Cell Subpopulations. Genes 10 (4). https://doi.org/10.3390/genes10040253.
Doupe, A.J., Kuhl, P.K., 1999. Birdsong and human speech: Common themes and mechanisms. Annu. Rev. Neurosci. 22 (1), 567–631. https://doi.org/10.1146/annurev.neuro.22.1.567.
Fuxjager, M.J., Lee, J.-H., Chan, T.-M., Bahn, J.H., Chew, J.G., Xiao, X., Schlinger, B.A., 2016. Research Resource: Hormones, Genes, and Athleticism: Effect of Androgens on the Avian Muscular Transcriptome. Molecular Endocrinology (Baltimore, Md.) 30 (2), 254–271. https://doi.org/10.1210/me.2015-1270.
Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B.o., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8 (8), 1494–1512. https://doi.org/10.1038/nprot.2013.084.
Han, Y., Gao, S., Muegge, K., Zhang, W., Zhou, B., 2015. Advanced Applications of RNA Sequencing and Challenges. Bioinf. Biol. Insights 9 (Suppl 1), 29–46. https://doi.org/10.4137/BBI.S28991.
Heston, J.B., White, S.A., 2017. To transduce a zebra finch: Interrogating behavioral mechanisms in a model system for speech. J. Comp. Physiol. A. 203 (9), 691–706. https://doi.org/10.1007/s00359-017-1153-0.
Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A. M., Delany, M. E., Dodgson, J. B., Chinwalla, A. T., Cliften, P. F., Clifton, S. W., Delehaunty, K. D., Fronick, C., Fulton, R. S., Graves, T. A., Kremitzki, C., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, 432(7018), 695–716. 10.1038/nature03154.
Jarvis, E.D., 2004. Learned Birdsong and the Neurobiology of Human Language. Ann. N. Y. Acad. Sci. 1016, 749–777. https://doi.org/10.1196/annals.1298.038.
Jarvis, E.D., 2019. Evolution of vocal learning and spoken language. Science (New York, N.Y.) 366 (6461), 50–54. https://doi.org/10.1126/science.aax0287.
Ji, F., Sadreyev, R.I., 2018. RNA-seq: Basic Bioinformatics Analysis. Current protocols in molecular biology 124 (1), e68. https://doi.org/10.1002/cpmb.v124.110.1002/cpmb.68.
Jürgens, U., 2002. Neural pathways underlying vocal control. Neurosci. Biobehav. Rev. 26 (2), 235–258. https://doi.org/10.1016/s0149-7634(01)00068-96.6.
Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: A fast spliced aligner with low memory requirements. Nat. Methods 12 (4), 357–360. https://doi.org/10.1038/nmeth.3317.
Korf, I., 2013. Genomics: The state of the art in RNA-seq analysis. Nat. Methods 10 (12), 1165–1166. https://doi.org/10.1038/nmeth.2735.
Li, H., Birol, I., 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34 (18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
Lovell, P.V., Carleton, J.B., Mello, C.V., 2013. Genomics analysis of potassium channel genes in songbirds reveals molecular specializations of brain circuits for the maintenance and production of learned vocalizations. BMC Genomics 14 (1), 470. https://doi.org/10.1186/1471-2164-14-470.
Lovell, P. V., Clayton, D. F., Replogle, K. L., & Mello, C. V. 2008. Birdsong "Transcriptomics": Neurochemical Specializations of the Oscine Song System. PLOS ONE, 3(10), e3440. 10.1371/journal.pone.0003440.
Margoliash, D., Fortune, E.S., Sutter, M.L., Yu, A.C., Wren-Hardin, B.D., Dave, A., 1994. Distributed Representation in the Song System of Oscines: Evolutionary Implications and Functional Consequences (Part 1 of 2). Brain Behav. Evol. 44 (4–5), 247–255. https://doi.org/10.1159/000113580.
Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12 (10), 671–682. https://doi.org/10.1038/nrg3068.
Nottebohm, F., Arnold, A.P., 1976. Sexual dimorphism in vocal control areas of the songbird brain. Science (New York, N.Y.) 194 (4261), 211–213. https://doi.org/10.1126/science.959852.
Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33 (3), 290–295. https://doi.org/10.1038/nbt.3122.
Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., Salzberg, S.L., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11 (9), 1650–1667. https://doi.org/10.1038/nprot.2016.095.
Pertea, G., Pertea, M., 2020. GFF Utilities: GffRead and GffCompare. F1000Res 9, 304. https://doi.org/10.12688/f1000research10.12688/f1000research.23297.1.
Peterson, R.S., Yarram, L., Schlinger, B.A., Saldanha, C.J., 2005. Aromatase is presynaptic and sexually dimorphic in the adult zebra finch brain. Proceedings. Biological Sciences 272 (1576), 2089–2096. https://doi.org/10.1098/rspb.2005.3181.
Petkov, C.I., Jarvis, E., 2012. Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. Front. Evolut. Neurosci. 4 https://doi.org/10.3389/fnevo.2012.00012.
Pfenning, A.R., Hara, E., Whitney, O., Rivas, M.V., Wang, R., Roulhac, P.L., Howard, J.T., Wirthlin, M., Lovell, P.V., Ganapathy, G., Mountcastle, J., Moseley, M.A., Thompson, J.W., Soderblom, E.J., Iriki, A., Kato, M., Gilbert, M.T.P., Zhang, G., Bakken, T., Jarvis, E.D., 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. Science 346 (6215). https://doi.org/10.1126/science.1256846.
Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., Sandhu, M.S., 2018. Long reads: Their purpose and place. Hum. Mol. Genet. 27 (R2), R234–R241. https://doi.org/10.1093/hmg/ddy177.
Qiao, Y., Ren, C., Huang, S., Yuan, J., Liu, X., Fan, J., Lin, J., Wu, S., Chen, Q., Bo, X., Li, X., Huang, X., Liu, Z., Shu, W., 2020. High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. Nat. Commun. 11 (1), 2653. https://doi.org/10.1038/s41467-020-16444-w.
Remage-Healey, L., Oyama, R.K., Schlinger, B.A., 2009. Elevated aromatase activity in forebrain synaptic terminals during song. J. Neuroendocrinol. 21 (3), 191–199. https://doi.org/10.1111/j.1365-2826.2009.01820.x.
Remage-Healey, L., Coleman, M.J., Oyama, R.K., Schlinger, B.A., 2010. Brain estrogens rapidly strengthen auditory encoding and guide song preference in a songbird. PNAS 107 (8), 3852–3857. https://doi.org/10.1073/pnas.0906572107.
Remage-Healey, L., Dong, S.M., Chao, A., Schlinger, B.A., 2012. Sex-specific, rapid neuroestrogen fluctuations and neurophysiological actions in the songbird auditory forebrain. J. Neurophysiol. 107 (6), 1621–1631. https://doi.org/10.1152/jn.00749.2011.
Rensel, M.A., Ding, J.A., Pradhan, D.S., Schlinger, B.A., 2018. 11β-HSD Types 1 and 2 in the Songbird Brain. Front. Endocrinol. 9, 86. https://doi.org/10.3389/fendo.2018.00086.
Roberts, R.J., Carneiro, M.O., Schatz, M.C., 2013. The advantages of SMRT sequencing. Genome Biol. 14 (7), 405. https://doi.org/10.1186/gb-2013-14-7-405.
Rundstrom, P., Creanza, N., 2021. Song learning and plasticity in songbirds. Curr. Opin. Neurobiol. 67, 228–239. https://doi.org/10.1016/j.conb.2021.02.003.
Sahraeian, S.M.E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P.T., Au, K.F., Bani Asadi, N., Gerstein, M.B., Wong, W.H., Snyder, M.P., Schadt, E., Lam, H.Y.K., 2017. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat. Commun. 8 (1), 59. https://doi.org/10.1038/s41467-017-00050-4.
Salzberg, S.L., 2019. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 20 (1), 92. https://doi.org/10.1186/s13059-019-1715-2.
Sohrabji, F., Nordeen, E.J., Nordeen, K.W., 1990. Selective impairment of song learning following lesions of a forebrain nucleus in the juvenile zebra finch. Behavioral and Neural Biology 53 (1), 51–63. https://doi.org/10.1016/0163-1047(90)90797-a.
Spool, J.A., Macedo-Lima, M., Scarpa, G., Morohashi, Y., Yazaki-Sugiyama, Y., Remage-Healey, L., 2021. Genetically identified neurons in avian auditory pallium mirror core principles of their mammalian counterparts. Curr. Biol. 31 (13), 2831–2843.e6. https://doi.org/10.1016/j.cub.2021.04.039.
Srivastava, A., George, J., Karuturi, R.K.M., 2019. Transcriptome Analysis. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, pp. 792–805. https://doi.org/10.1016/B978-0-12-809633-8.20161-1.
Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., Conesa, A., 2018. SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28 (3), 396–411. https://doi.org/10.1101/gr.222976.117.
The UniProt Consortium, 2019. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 47 (D1), D506–D515. https://doi.org/10.1093/nar/gky1049.
Vierra, M.N., Kingan, S.B., Tseng, E., Clark, T., Hon, T., Rowell, W.J., Mountcastle, J., Fedrigo, O., Jarvis, E.D., Korlach, J., 2017. From RNA to Full-Length Transcripts: The PacBio Iso-Seq Method for Transcriptome Analysis and Genome Annotation. Genome10K and Genome Science Conference Abstracts.
Wang, L., Jiang, X., Wang, L., Wang, W., Fu, C., Yan, X., Geng, X., 2019. A survey of transcriptome complexity using PacBio single-molecule real-time analysis combined with Illumina RNA sequencing for a better understanding of ricinoleic acid biosynthesis in Ricinus communis. BMC Genomics 20 (1), 456. https://doi.org/10.1186/s12864-019-5832-9.
Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., Ware, D., 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat. Commun. 7 (1), 11708. https://doi.org/10.1038/ncomms11708.

Warren, W.C., Clayton, D.F., Ellegren, H., Arnold, A.P., Hillier, L.W., Künstner, A., Searle, S., White, S., Vilella, A.J., Fairley, S., Heger, A., Kong, L., Ponting, C.P., Jarvis, E.D., Mello, C.V., Minx, P., Lovell, P., Velho, T.A.F., Ferris, M., Wilson, R.K., 2010. The genome of a songbird. Nature 464 (7289), 757–762. https://doi.org/10.1038/nature08819.

Wu, P.-Y., Phan, J.H., Wang, M.D., 2012. The Effect of Human Genome Annotation Complexity on RNA-Seq Gene Expression Quantification. IEEE International Conference on Bioinformatics and Biomedicine Workshops IEEE International Conference on Bioinformatics and Biomedicine 2012, 712–717. https://doi.org/10.1109/BIBMW.2012.6470224.

Wyman, D., Mortazavi, A., Berger, B., 2019. TranscriptClean: Variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. Bioinformatics 35 (2), 340–342. https://doi.org/10.1093/bioinformatics/bty483.

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., England, W., Chu, S.-H., Spitale, R.C., Tenner, A.J., Wold, B.J., Mortazavi, A., 2020. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. BioRxiv 672931. https://doi.org/10.1101/672931.

Zhang, J., Liu, C., He, M., Xiang, Z., Yin, Y., Liu, S., Zhuang, Z., 2019. A full-length transcriptome of Sepia esculenta using a combination of single-molecule long-read (SMRT) and Illumina sequencing. Mar. Geonomics 43, 54–57. https://doi.org/10.1016/j.margen.2018.08.008.

Zhao, S., Zhang, B., 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16 (1). https://doi.org/10.1186/s12864-015-1308-8.