Title
Improved Zebra Finch Brain Transcriptome Using Both Short and Long Read RNA-seq Methods

Author
He, Jingyan

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Improved Zebra Finch Brain Transcriptome Using Both Short and Long Read RNA-seq Methods

A thesis submitted in satisfaction of the requirements for the degree Master of Science in Physiological Science

by

Jingyan He

2020

ABSTRACT OF THE THESIS

Improved Zebra Finch Brain Transcriptome Using Both Short and Long Read RNA-seq Methods

by

Jingyan He

Master of Science in Physiological Science

University of California, Los Angeles, 2020

Professor Barney Schlinger, Chair

Zebra finch (*Taeniopygia guttata*) is a representative songbird species that has been widely studied to investigate the neurobiological basis of vocal learning, a rare trait shared in only a few species including human. In December 2019, an updated zebra finch genome annotation (bTaeGut1_v1.p) was released from 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-read method and PacBio single-molecule real-time (SMRT) long-read method 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 to investigate the new bird annotation for the first time. 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. Our results provide additional resources for future studies of zebra finches and other songbirds utilizing this improved annotation.

The thesis of Jingyan He is approved.

Xinshu Xiao

Stephanie A. White

Barney Schlinger, Committee Chair

University of California, Los Angeles

2020

DEDICATION

This thesis is dedicated to my parents and friends for their support, partnership, and encouragement throughout my academic endeavor.

TABLE OF CONTENTS

ABSTRACT..... ii

COMMITTEE PAGE iii

DEDICATION iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

ACKNOWLEDGMENTS viii

INTRODUCTION 1

MATERIALS AND METHODS 4

OVERVIEW OF THE RNA-SEQ ANALYSIS PIPELINE 4

HIGH-THROUGHPUT RNA-SEQ DATASETS 5

SHORT-READ RNA-SEQ DATA ANALYSIS 5

CONSTRUCTION OF AN IMPROVED ANNOTATION BASED ON SHORT-READ ANALYSIS 6

LONG-READ RNA-SEQ DATA ANALYSIS PIPELINE 7

GENERATION OF THE FINAL ANNOTATION 8

QUANTIFICATION OF TRANSCRIPTS IN THE FINAL ANNOTATION 8

ORF/PEPTIDE PREDICTION 9

RESULTS 9

CONSTRUCTION OF IMPROVED ANNOTATION BASED ON NOVEL DISCOVERIES FROM SHORT-READ DATA 9

TALON ANNOTATION OF PACBIO FLNC READS 10

CONSTRUCTION OF FINAL ANNOTATION AND TRANSCRIPT QUANTIFICATION 11

PEPTIDE PREDICTION OF TRANSCRIPTS IN THE FINAL ANNOTATION 12

DISCUSSION 12

BIBLIOGRAPHY 26

LIST OF FIGURES

FIGURE 1. OVERVIEW OF THE ANALYSIS PIPELINE 18

FIGURE 2. COMPOSITION OF STRINGTIE ASSEMBLED TRANSCRIPTOME BASED ON TRANSCRIPT NOVELTY CLASSIFICATION AFTER TRANSCRIPT-GENE ASSIGNMENT 19

TABLE 1. SUMMARY OF IMPROVED ANNOTATION IN COMPARISON TO REFERENCE ENSEMBL ANNOTATION 20

FIGURE 3. LENGTH DISTRIBUTION OF TRANSCRIPT MODELS FROM ENSEMBL ANNOTATION AND STRINGTIE PREDICTION 20

FIGURE 4. TALON ANNOTATION RESULTS USING IMPROVED ANNOTATION AS REFERENCE 21

TABLE 2. CONTENT SUMMARY AND COMPARISON BETWEEN FINAL ANNOTATION AND ENSEMBL ANNOTATION 22

FIGURE 5. SOURCE COMPOSITION OF NOVEL GENES AND TRANSCRIPTS IN THE FINAL ANNOTATION 23

FIGURE 6. TRANSCRIPT QUANTIFICATION BASED ON FINAL ANNOTATION 24

FIGURE 7. PEPTIDE LENGTH DISTRIBUTION IN TRANSDECODER PREDICTED PROTEOME AND UNIPROT ZEBRA FINCH PROTEOME 25

ACKNOWLEDGEMENTS

I would like to express my appreciation to my committee members for all the support and guidance throughout my graduate study. Special thanks to Dr. Xinshu Xiao for providing me the opportunity to join her lab and leading me to the field of Bioinformatics. Additionally, I want to thank the Xiao Lab members, who have generously helped me develop my research and computational skillsets, especially Dr. Ling Zhang. I would also like to thank to Marisela Diaz-Vasquez for all her help throughout my master's.

INTRODUCTION

Songbirds (Order Passeriformes; Suborder Oscine) are well-established organisms for studies aimed at understanding the neural basis of vocal learning (Clayton et al., 2009; Doupe & Kuhl, 1999; Jarvis, 2019). This rare ability of acquiring vocalization through imitation of a model, is observed only in a few mammals and bird species. While from two distinct orders, songbird species and human shows numerous parallels between bird song acquisition and human speech development (Jarvis, 2004). For example, both species share corticostriatal circuits for vocalization production and demonstrate a direct connection from motor cortex to brainstem vocal motor neurons, which is a unique connection observed in vocal learners (Bolhuis et al., 2010; Jürgens, 2002; Petkov & Jarvis, 2012, Jarvis, 2004). In addition, brain regions involved in vocal learning pathways of songbirds and human show functional specialization and exhibit a convergent transcriptional profile, suggesting overlapped molecular mechanisms underlying complex vocal learning traits across the two evolutionary 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 human, songbird species that are amenable to laboratory environment such as zebra finch (*Taeniopygia guttata*) offers a favorable vocal learning model for experimental manipulation (Heston & White, 2017).

As a representative of the songbirds, the zebra finch (*Taeniopygia guttata*) genome was sequenced (Warren et al., 2010), only the second avian species subject to whole genome sequencing (Hillier et al., 2004;). This zebra finch genome assembly, the nucleotide sequence of the genome, as well as the genome annotation has been widely used. However, many studies have pointed out that the songbird annotation is incomplete (Balakrishnan et al, 2012; Fuxjager et al, 2016). In 2019, the Vertebrate Genome Project under The Genome 10K 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 zebra finch 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 & Zhang, 2015; Abril & 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). Yet, few studies have utilized the new bird genome, and 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 of a transcriptome due to its capability of capturing expressed genes in the tissue samples (Ji & Sadreyev, 2018; Salzberg, 2019). The pervasive next generation sequencing (NGS) RNA-seq methods such as Illumina-based short-read RNA-seq has been used in numerous biomedical research applications including unbiased survey of the entire transcriptome and gene expression quantification (Denoeud et al., 2008; Wang 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 & Wang, 2011). Recently, the emerging third-generation sequencing (TGS) technologies such as PacBio single-molecule real-time (SMRT) sequencing have presented an alternative powerful method for transcriptome profiling. With the advantage of producing long reads that are typically >10 Kbp long, 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 Iso-Seq 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 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 possibilities 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 peptides 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. Overall, the novel findings from our analysis provided additional sources and information for the future studies of zebra finch brain and behavior.

MATERIALS AND METHODS

Overview of the RNA-seq analysis pipeline

We utilized publicly available bioinformatic tools and pipelines to perform RNA-seq analysis for both short-read and long-read data to identify high-confident novel transcripts and genes. The Illumina-based short-read data were first processed with a previously described method for transcriptome assembly (Pertea et al., 2016). We used HISAT2 (Kim et al., 2015) and StringTie (Pertea et al., 2015) to align reads and obtain a unified transcriptome discovered from all the zebra finch tissue samples. The novel transcript and gene models from the short-read data were integrated into the reference annotation to construct a more comprehensive annotation. The improved annotation then served as the reference annotation for the long-read analysis pipeline. Next, we examined whether the short read based novel transcript prediction has supportive evidence from PacBio full-length long reads. For this purpose, we used TALON (Wyman et al., 2020), a technology-agnostic long read analysis pipeline, to annotate full-length long reads. The final annotation consists of two types of transcripts: Ensembl annotated transcripts and long reads-supported novel transcripts (Figure 1).

To further characterize the predicted novel transcripts, we performed transcript quantification using the short-read RNA-seq data that were previously used for StringTie transcriptome

assembly. The transcript expression levels in short-read data were estimated using Kallisto (Bray et al., 2016), an ultra-fast alignment-free transcript quantification program.

In addition to the transcript quantification, 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 (Figure 1).

High-throughput RNA-seq datasets

The Illumina-based short-read RNA-seq data were obtained from a previous publication from the White Lab at UCLA, Department of Integrative Biology and Physiology (Burkett et al., 2018). The study focused on the gene expression in Area X, a key vocal nucleus in zebra finch forebrain (Sohrabji et al., 1990). The RNA-seq data were generated from the Area X tissues that overexpress either FoxP2 or GFP genes from 7 juvenile male zebra finches under a critical period of vocal learning and song development. cDNA libraries for each bird sample were sequenced twice by Illumina HiSeq 2500 platform, and 50 bp long paired-end short reads for each bird 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 Iso-Seq 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).

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 files for each biological replicate were passed to StringTie v2.1.3b (Pertea et al., 2015) for transcript assembly. We provided a reference annotation 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).

Construction of an improved annotation based on short-read analysis

To accurately classify the StringTie predicted novel transcripts, a python script was written to reassign each novel transcript to a source gene based on the matched exon numbers and overlapping regions with known transcript models. In our algorithm, each novel transcript was compared with all the known transcript variants to get 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. Novel transcripts that overlapped with none of the known transcript models were considered as novel genes because these transcripts may reflect previously unidentified genes. After the assignment process, the novel transcript variants and novel genes were merged with the original Ensembl

annotation to generate an improved reference annotation for the following long-read analysis pipeline. Lastly, to further polish the improved annotation, all novel transcripts and genes without strand specificity were removed.

Long-read RNA-seq data analysis pipeline

The FLnc reads from PacBio Iso-seq methods were first aligned to the Ensembl genome assembly (bTaeGut1_v1.p, INSDC Assembly GCA_003957565.2) using Minimap2 v2.17 (Li, 2018). Minimap2 were run with default parameters for mapping PacBio Iso-seq/traditional cDNA. In addition, a specific output option `-MD` that was suggested by the TALON pipeline was turned on. The read alignment file was then passed to TranscriptClean to correct mismatches, microindels, and noncanonical splice junctions in mapped long reads, and was processed with TranscriptClean default parameters (Wyman & Mortazavi, 2019). The cleaned read alignment SAM file served as an input to the major TALON pipeline, along with the short reads-based improved annotation from previous steps. The fractions of As in the SAM file were recorded using the `talon_label_reads` utility with default options to determine if internal priming existed. The first main step of the TALON pipeline, which is the initialization of a TALON database using user-defined annotation GTF file, was carried out using `talon_initialize_database` with default parameters. Then, the annotation of FLnc CCS reads was carried out using the `talon` annotator. In the annotation step, 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 `talon_filter_transcripts` module with the specified parameter `--minCount` being `2` based on the average read coverage per unique transcript model. After the

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 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-read data was generated using the `talon_create_GTF` utility based on the filtered transcriptome and the reference annotation (Wyman et al., 2020).

Generation of the final annotation

A python script was written to construct a final annotation by integrating the novel transcripts into Ensembl annotation (bTaeGut1_v1.p). Unique novel transcript isoforms or novel genes supported by long reads created from the `talon_create_GTF` utility were merged with the Ensembl annotation to create the final annotation of zebra finch transcriptome in this project.

In summary, there are two types of novel transcripts/genes in the final annotation. One was predicted from StringTie and was supported by TALON long read annotation, and the other one was from TALON annotation alone.

Quantification of transcripts in the final annotation

Transcript quantification of Illumina-based RNA-seq data were performed using Kallisto v0.45.0 (Bray et al., 2016). To provide the FASTA formatted file for `kallisto index`, 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 & Pertea, 2020). A kallisto index was then built using the FASTA file with default parameters. The two technical replicates for each bird sample were merged for Kallisto transcript quantification. For each

biological replicate, the FASTQ files, which contain forward reads from Illumina paired-end run of two technical replicates, were concatenated tail-to-head; the same procedure was also applied to the FASTQ files containing reverse reads. The merged FASTQ files were passed to `kallisto quant` for transcript quantification with default running mode and parameters.

ORF/peptide prediction

The ORF and peptide predictions were done using TransDecoder v5.5.0 (Haas et al., 2013). First, a FASTA file was created based on the transcriptome in the final annotation GTF formatted file using the TransDecoder program utility `gtf_genome_to_cdna_fasta.pl`. The GTF file was also converted into GFF3 format using the `gtf_to_alignment_gff3.pl` utility. Then, the FASTA file was passed to `TransDecoder.LongOrfs` for open reading frame identification with default parameters. 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 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 turned on, 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.

RESULTS

Construction of improved annotation based on novel discoveries from short-read data

Transcriptome assembly was first performed to obtain all the expressed known transcripts as well as the novel transcripts using short-read data. After merging all the transcripts into a unified set of transcripts, we reassigned each novel transcript to a best matched gene. Among 20,614

StringTie-predicted novel transcripts, we successfully assigned 16,109 transcripts to a previously annotated gene, and these transcripts were considered as new transcript variants containing novel exon junctions. The remaining 4,505 novel transcripts were considered as novel isoforms expressed from unknown genes. The improved annotation was constructed by adding novel transcripts from both known and novel genes to the Ensembl annotation (Figure 2). To finalize the improved annotation, 406 novel genes without strand information were removed.

In summary, the improved annotation has a total of 26,249 genes and 59,077 transcripts, which is an expanded set compared to the Ensembl reference annotation (22,150 genes and 38,869 transcripts) (Table 1). The length distributions of transcript models from Ensembl and StringTie predictions are shown in Figure 3. 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, suggesting that additional scrutiny is needed to further examine the novel transcripts, as presented below.

Talon annotation of PacBio FLnc reads

In order to provide additional evidence to support the novel transcripts discovered from short-read RNA-seq analysis, we analyzed PacBio FLnc reads using the TALON pipeline, with the final annotation generated above as an input annotation file (including Ensembl known and StringTie-predicted novel transcripts). With a total input of 405,837 aligned FLnc reads, talon annotator identified 144,812 aligned reads that have full coverage and are 100% identical to the nucleotide sequences of the transcript models in the given annotation file. Each of the annotated read was assigned to one of the seven TALON transcript novelty categories, which include “Known”, “Incomplete splice match (ISM)”, “Novel in catalog (NIC)”, “Novel not in catalog

(NNC)", "Genomic", "Antisense", and "Intergenic" (Wyman et al., 2020). Note that the "known" reads here matched the annotations we provided to TALON, which means they may have originated from StringTie-predicted novel transcripts.

The median length of the reads in the seven categories varies from ~2,500 bp to ~5,000 bp. In particular, NIC and NNC have higher medians than other categories (Figure 4a). Consistent with the read length observation, the numbers of exons in NIC and NNC categories were also the highest among all categories (Figure 4b). The number of reads that was annotated as "known" transcripts was the highest among all the categories, following by ISM, which is the category describing partial match between the FLnc reads and the known transcript isoforms (Figure 4c). After applying a minimum read count filter ($n \geq 2$) to the novel transcript models from TALON annotation results, 12,090 distinct transcript isoforms expressed from 6,520 genes were supported by FLnc reads and were considered as high confident ones, in which the novel transcript models were required to have at least 2 read count to be retained. Among all the long reads-supported transcript models from TALON annotation, over 60% fall in the "known" transcript category, in which nearly half of the transcript models were defined by StringTie prediction based on the short-read data (Figure 4d-e).

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 gene count of 22,370 and total transcript count of 47,003 in the final annotation file. Compared to the original Ensembl annotation, the number of transcripts per gene increased from 1.75 to 2.1 (Table 2). Among all the novel genes and transcript variants, 80% genes and 41.6% transcripts were predicted based

on StringTie short-read transcriptome assembly (and supported by long reads), whereas 20% genes and 58.4% transcripts were identified by TALON annotation of long reads (Figure 5).

To further characterize the predicted novel transcripts, the transcript expression levels in the short-read data were quantified and measured via the metric transcript per million (TPM). TPM values were highly correlated among the 7 biological replicates (Figure 6a). Figure 6b shows the expression levels of known transcript models from Ensembl annotation, and those of novel transcript models from StringTie prediction or TALON annotation. We observed that 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. As shown in Figure 6c, Ensembl genes that were supported by long reads have a similar TPM distribution as the novel transcript models.

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 with the UniProt zebra finch proteome (Proteome ID: UP000007754), the total number of protein peptides increase 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 (Figure 7).

DISCUSSION

The first complete zebra finch genome was published about a decade prior to this work. Since then, the release of the bird genome has provided a powerful tool for complex genomic studies

via high-throughput approaches. Compared to the first songbird genome, the zebra finch genome annotation released in 2019 from Ensembl has largely improved in terms of its completeness. However, only few studies to date have employed the updated genome. Our study utilized publicly available high-throughput RNA-seq data generated from both NGS and TGS sequencing technologies to investigate the comprehensiveness of the new bird genome. From our analysis, we uncovered the existence of novel transcripts and genes that are laying outside the current genome annotation, which implies the imperativeness of future work on the songbird genome annotation. Notably, the new bird genome was constructed based on the genetic material from the muscle tissue of a single individual male zebra finch identified as Black17, which was the same biological sample used in the initial zebra finch genome build. Therefore, the bird genome could potentially be biased by the alleles unique to the chosen individual and the used tissue.

In recent years, the PacBio SMRT sequencing technology has gradually become a part of the standard approaches for novel transcript discovery and transcriptome profiling of different species or cell lines. Meanwhile, the computational tools and algorithms to assemble short reads generated from widely used NGS sequencing technologies have also been largely improved in their accuracy and robustness. With the availability of different sequencing methods and analysis tools, previous studies indicated the advantages of employing hybrid RNA-seq analysis approaches to gain insights of transcriptome (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 and then validations of novel findings 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 publicly available RNA-seq data.

In our study, we first adopted the widely used alignment-based transcriptome discovery tool, StringTie, to perform transcript detection and prediction using short-read RNA-seq data generated from bird brain tissue. Form the StringTie assembly, 20,614 novel transcripts were discovered. The number of the novel transcripts predicted was extremely striking as it equals to a 53% increase to the reference annotation. Considering the potential assembly false positives reported from previous studies (Pertea et al., 2015; Sahraeian et al., 2017), we visualized the StringTie predicted transcriptome in comparison to reference annotation in the Integrative Genomics Viewer (IGV) (Robinson et al., 2011), and have observed novel transcripts highly overlapped with annotated gene models. Given our primary goal of discovering high-confident novel transcripts, we classified the StringTie novel transcripts that share large overlapped regions with known genes as novel spliced transcript variants instead of a completely novel transcript originated from an unknown gene that has not been previously identified. In addition, we noticed an unsolved technical issue of StringTie assembly in the merge mode, which is the difficulty of assigning unique StringTie gene identifier to an annotated reference gene in a complex gene-rich region of the genome. To precisely classify the novel transcripts, as well as overcome the StringTie technical hindrance that would potentially bias the following long-read annotation, we reassigned the StringTie novel transcripts to a best source gene. Based on the reassignment results, ~7.6% of total StringTie predicted transcriptome, are considered as novel transcripts from unknown genes. Also, ~27.1% of total predicted transcriptome, are labeled as new transcript variants which overlapped with at least one known gene.

Next, using PacBio long reads, we confirmed 176 transcripts derived from novel genes and 3,380 novel transcript variants of known genes predicted by StringTie. According to the TALON annotation algorithm and stringent parameters defined here, full-length long reads are required to have precise matches at the exon boundaries to be assigned to a known transcript model. Therefore, the StringTie predicted novel transcript models supported by long reads are highly trustable. Among a total of 12,090 long read supported transcripts models, it is striking to see that 60.7% of them belong to the "Known" TALON-defined transcript category, among which close to half were novel transcripts predicted by StringTie. This consistency between short-read discovery and long-read data might be the result of tissue specificity, as both the short-read and long-read data were generated from brain tissues. In the meantime, this observation suggests the presence of a considerable amount 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 categories are 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 and precise definition of exon boundaries and transcript structures. Interestingly, less than 0.1% of the distinct transcript models fall into the intergenic category, an indication of improvement of our annotation.

After the long-read annotation steps, the high-confident novel transcripts from short and long read pipeline were merged with the original Ensembl annotation to construct a final annotation in our pipeline. The transcript quantification results demonstrate high transcript expression

correlation among all 7 bird samples. To look at the expression levels of newly discovered novel transcripts, we plotted the ECDF plot for annotation transcripts from different sources in the final annotation. Strikingly, the novel transcripts discovered in this project had relatively high expression levels, comparable to those of expressed Ensembl genes, strongly supporting the validity of the predictions.

Lastly, to enable future analysis at the proteome level, we predicted the possible open reading frames based on the transcript sequences. In order to gain functional insights and identify ORFs that have homology with known proteins, a universal BlastP search was performed against the entire UniProtKB/Swiss-Prot protein database to maximize the predicted ORF retention. We then retained the longest predicted ORFs that were supported by BlastP search as the best ORF for protein-coding transcript candidates. In total, we successfully predicted 42,818 peptides from the transcriptome in the final annotation, whereas the current UniProt zebra finch proteome contains 31,341 protein peptides.

In summary, our analysis has added 8,134 novel transcript variants and 220 novel genes to the Ensembl zebra finch annotation (bTaeGut1_v1.p). Moreover, we have predicted a new proteome based on the transcriptome from the final annotation. The new proteome has increased the total number of peptide sequence by ~37% compared to the current zebra finch proteome. These findings expand the latest Ensembl zebra finch annotation, representing 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 the latest TGS SMRT-seq methods.

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

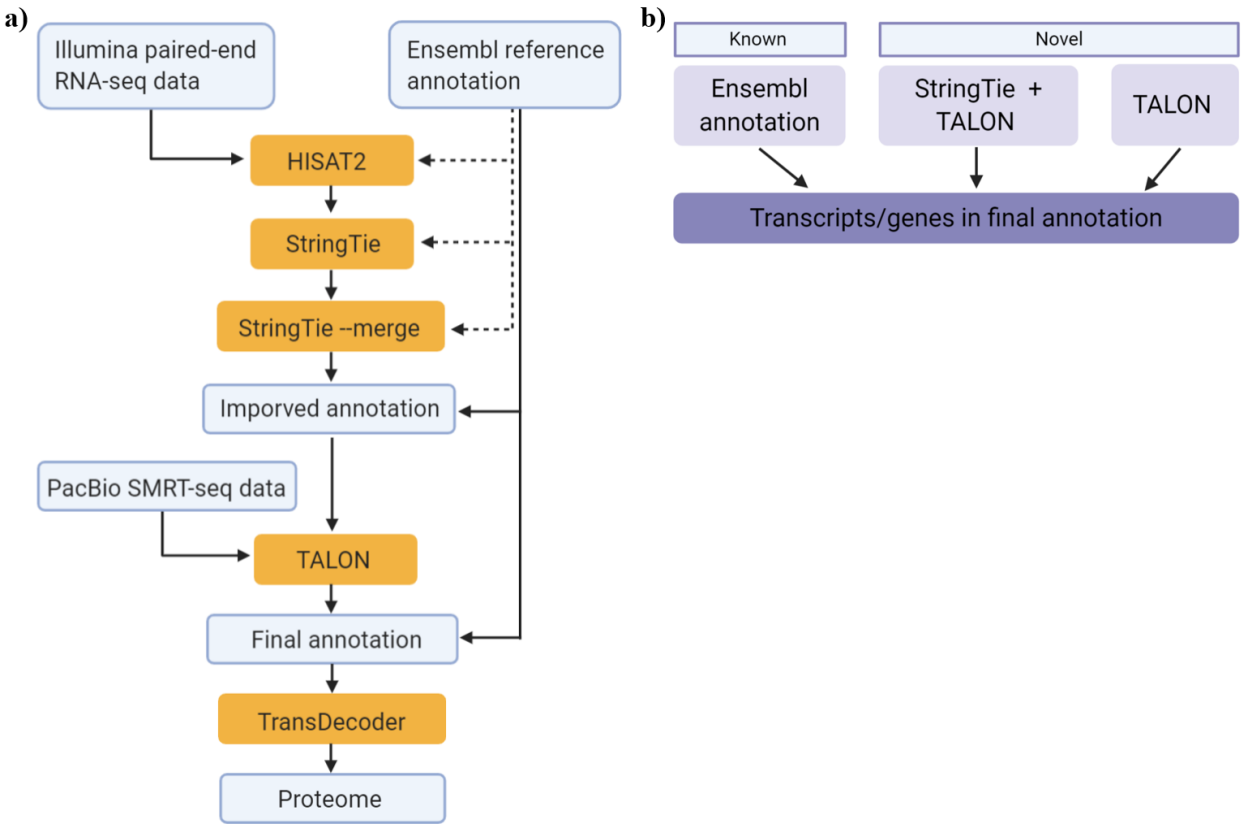

Figure 1. Overview of the analysis pipeline. a) 1. Illumina paired-end RNA-seq data were aligned to Ensembl reference genome using HISAT2. The short-read transcriptome assembly was done by StringTie, and a unified transcriptome set containing novel transcripts were generated by StringTie merge mode. The StringTie novel transcripts were processed by a customized python script and merged with Ensembl annotation transcripts to get an improved annotation. 2. The PacBio SMRT-seq generated FLnc reads were annotated against the improved 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 base 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.

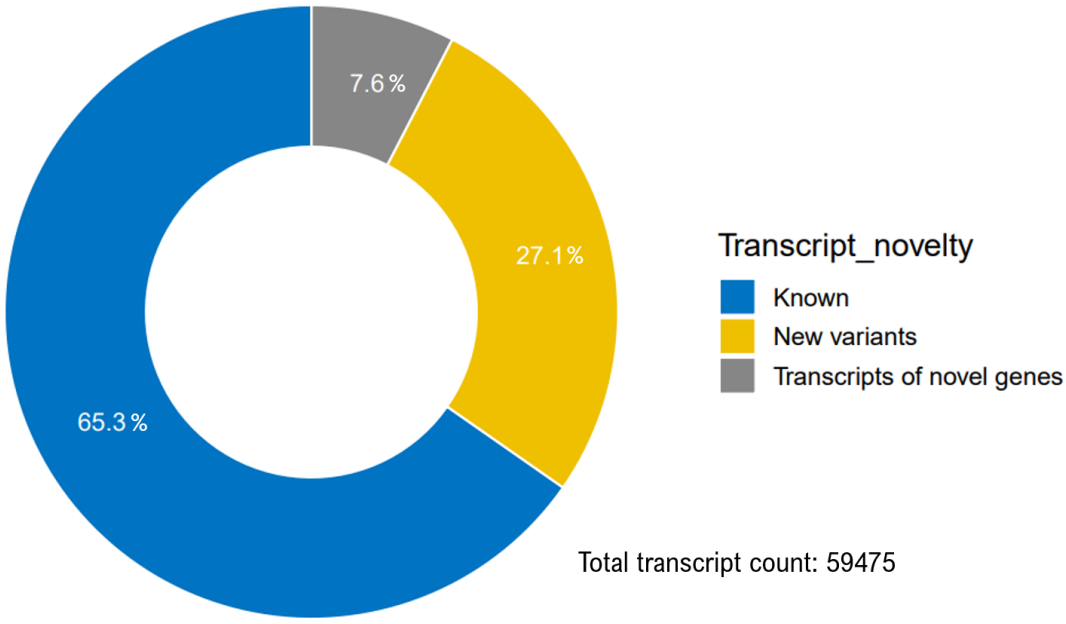

Figure 2. Composition of StringTie assembled transcriptome based on transcript novelty classification after transcript-gene assignment. Among 59,475 StringTie assembled transcript models, ~65.3% are known transcripts from Ensembl annotation. 27.1% of the transcripts are new transcript variants of known genes, and 7.6% are novel transcripts that do not have an assignable source gene.

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

	Improved 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

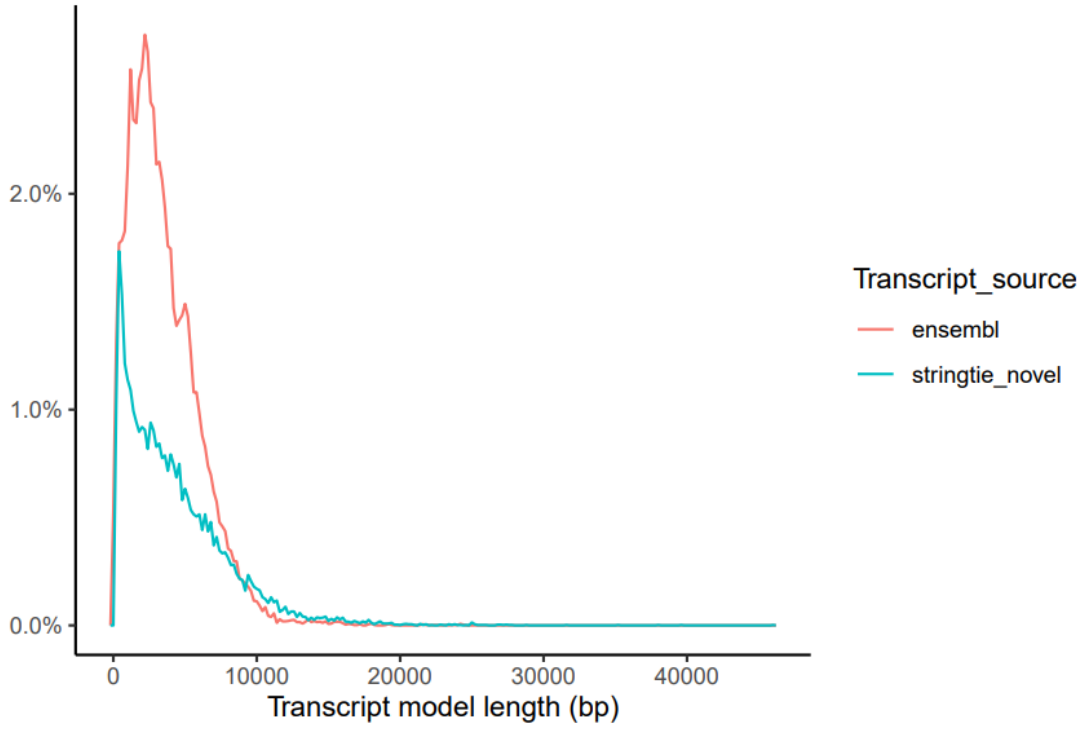

Figure 3. Length distribution of transcript models from Ensembl annotation and StringTie prediction. The vast majority of the transcript models are less than 10 Kbp long.

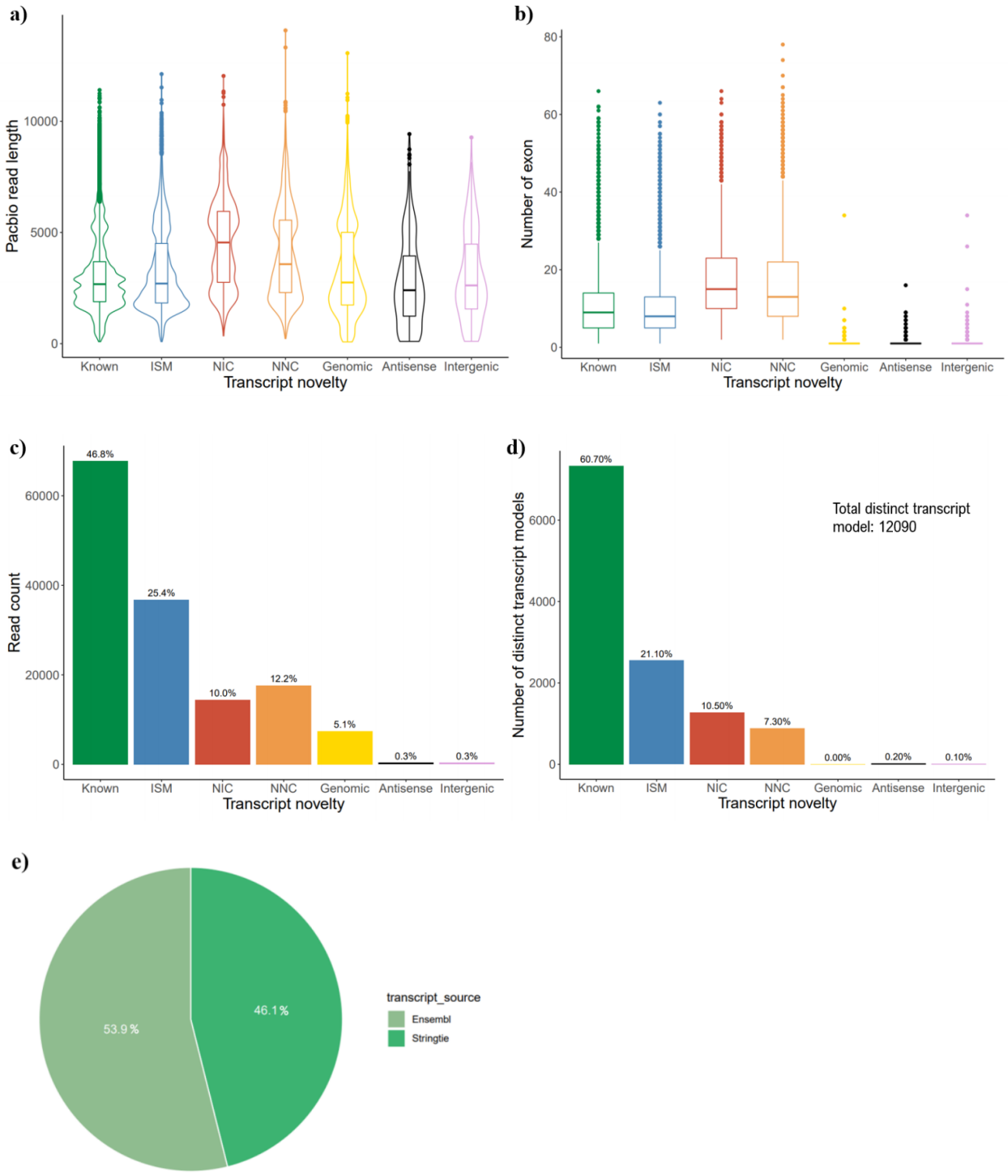

Figure 4. Talon annotation results using improved annotation as reference. a) Distribution of annotated PacBio FLnc read length in each TALON transcript novelty category. b) Exon

numbers 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

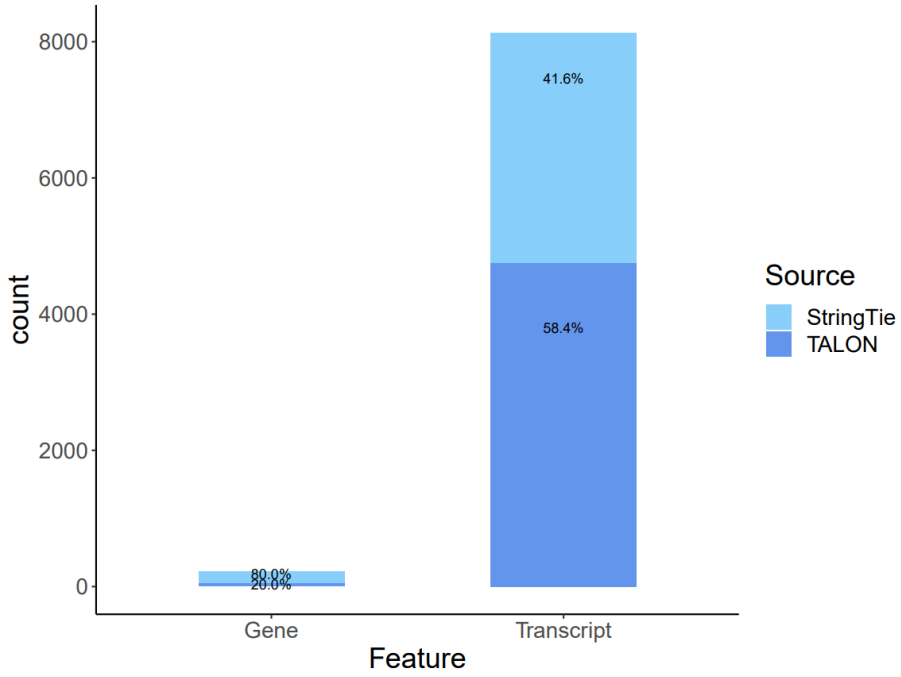

Figure 5. Source composition of novel genes and transcripts in the final annotation. Among 220 novel genes, 176 (80%) were identified from StringTie assembly, and 44 (20%) were based on TALON annotation. Among 8,134 novel transcripts, 41.6% were from StringTie assembly, and 58.4% from TALON annotation.

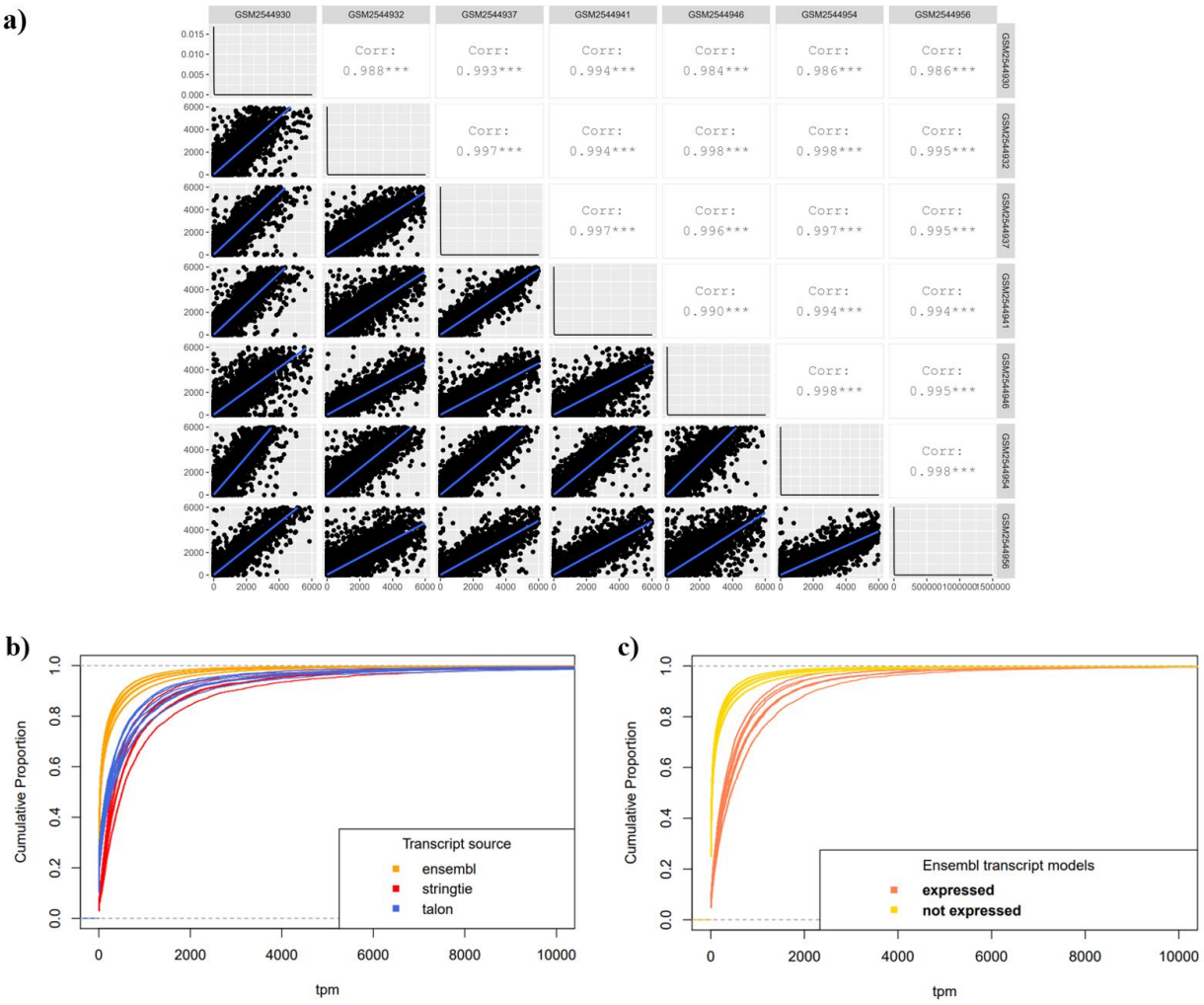

Figure 6. Transcript quantification based on final annotation. a) Transcript expression correlation among 7 zebra finch shot-read samples. b) ECDF plot of transcript TPM of all transcript models in the final annotation. c) ECDF plot of Ensembl transcript TPM that are expressed or not expressed in the long-read data.

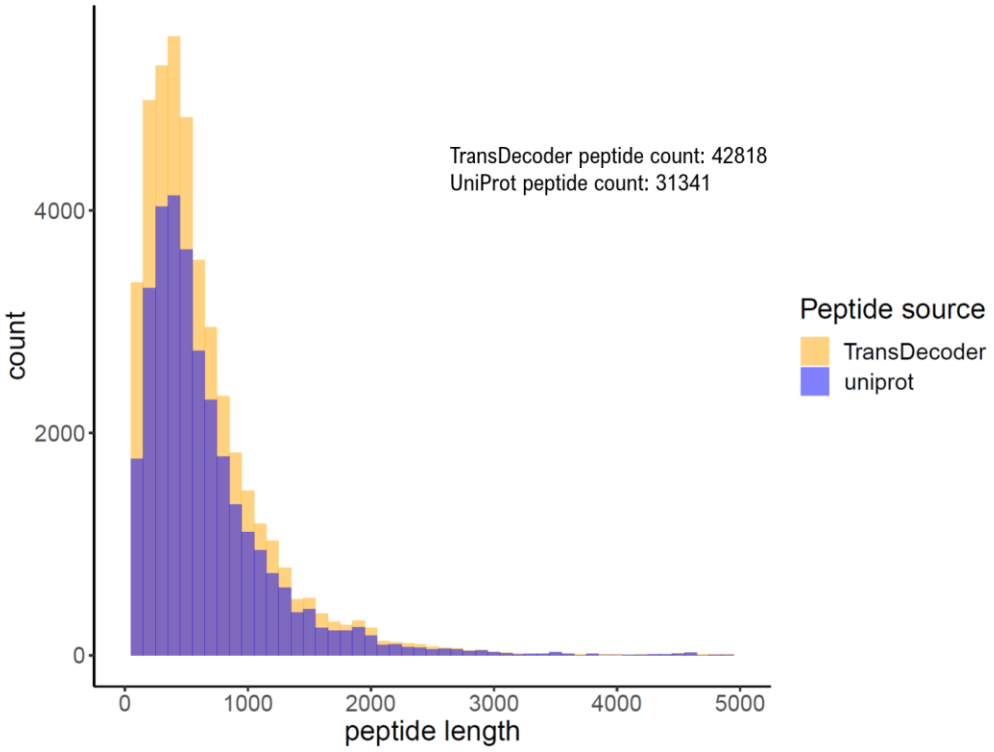

Figure 7. Peptide length distribution in TransDecoder predicted proteome and UniProt zebra finch proteome. Histogram of the distribution of peptide length (bin width = 100).

BIBILIOGRAPHY

Abril, J. F., & Castellano, S. (2019). Genome Annotation. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 195–209). Academic Press. 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., Okanoya, K., & Scharff, C. (2010). Twitter evolution: Converging mechanisms in birdsong and human speech. *Nature Reviews Neuroscience*, *11*(11), 747–759. https://doi.org/10.1038/nrn2931

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *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 Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

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. *Scientific Reports*, *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), R865-873. 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 Biology*, *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. *Annual Review of Neuroscience*, *22*, 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., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature Protocols*, *8*(8). 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. *Bioinformatics and Biology 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. *Journal of Comparative Physiology 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., … Project management: (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), 695–716. https://doi.org/10.1038/nature03154

Jarvis, E. D. (2004). Learned Birdsong and the Neurobiology of Human Language. *Annals of the New York Academy of Sciences*, *1016*, 749–777. https://doi.org/10.1196/annals.1298.038

Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science*, *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.68

Jürgens, U. (2002). Neural pathways underlying vocal control. *Neuroscience & Biobehavioral Reviews*, *26*(2), 235–258. https://doi.org/10.1016/S0149-7634(01)00068-9

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature 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. *Nature Methods*, *10*(12), 1165–1166. https://doi.org/10.1038/nmeth.2735

Li, H. (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., & Subgroup, 1000 Genome Project Data Processing. (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. https://doi.org/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, Behavior and Evolution*, *44*(4–5), 247–255. https://doi.org/10.1159/000113580

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*(10), 671–682. https://doi.org/10.1038/nrg3068

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*, 304. https://doi.org/10.12688/f1000research.23297.2

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. *Nature Protocols*, *11*(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

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. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Petkov, C. I., & Jarvis, E. (2012). Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience*, *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. *Human Molecular Genetics*, *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. *Nature Communications*, *11*(1), 2653. https://doi.org/10.1038/s41467-020-16444-w

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, *14*(7), 405. https://doi.org/10.1186/gb-2013-14-7-405

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

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. *Nature Communications*, *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 Biology*, *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

Srivastava, A., George, J., & Karuturi, R. K. M. (2019). Transcriptome Analysis. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 792–805). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20161-1

Tardaguila, M., Fuente, L. de la, Marti, C., Pereira, C., Pardo-Palacios, F. J., Risco, H. del, 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 Research*, *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 Research*, *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, 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. *Nature Communications*, *7*(1), 11708. https://doi.org/10.1038/ncomms11708

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, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

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*, *2012*, 712–717. https://doi.org/10.1109/BIBMW.2012.6470224

Wyman, D., & Mortazavi, A. (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. *Marine Genomics*, *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