OXFORD

BIOLOGY Methods & Protocols

Biology Methods and Protocols, 2023, bpad013

https://doi.org/10.1093/biomethods/bpad013

Advance Access Publication Date: 7 July 2023

Methods Article

Rapid genome functional annotation pipeline anchored to the house sparrow (*Passer domesticus*, Linnaeus 1758) genome reannotation

Melisa Eliana Magallanes-Alba [1,2], Agustín Baricalla[3], Natalia Rego[4,5], Antonio Brun[1,6,7], William H. Karasov[2], Enrique Caviedes-Vidal [1,2,7,*]

[1]Instituto Multidisciplinario de Investigaciones Biológicas (IMIBIO-SL), Consejo Nacional de Investigaciones Científicas y Técnicas, San Luis, San Luis 5700, Argentina

[2]Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, Madison, WI 53706, USA

[3]Centro de Investigaciones y Transferencia del Noroeste de la Provincia de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas, Pergamino, Buenos Aires 2700, Argentina

[4]Bioinformatics Unit, Institut Pasteur de Montevideo, Montevideo, Montevideo 11200, Uruguay

[5]Laboratorio de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Montevideo 11400, Uruguay

[6]Departamento Kinesiología y Fisiatría. Facultad de Ciencias de la Salud, Universidad Nacional de San Luis, San Luis , San Luis 5700, Argentina

[7]Departamento de Biología. Facultad de Química, Bioquímica y Farmacia, Universidad Nacional de San Luis, San Luis, San Luis 5700, Argentina

*Correspondence address: Instituto Multidisciplinario de Investigaciones Biológicas de San Luis. Almte. Brown 907, D5700 ANW, San Luis. Argentina. Email: enrique.caviedes@gmail.com

Abstract

The house sparrow (*Passer domesticus*) is a valuable avian model for studying evolutionary genetics, development, neurobiology, physiology, behavior, and ecology, both in laboratory and field-based settings. The current annotation of the *P. domesticus* genome available at the Ensembl Rapid Release site is primarily focused on gene set building and lacks functional information. In this study, we present the first comprehensive functional reannotation of the *P. domesticus* genome using intestinal Illumina RNA sequencing (RNA-Seq) libraries. Our revised annotation provides an expanded view of the genome, encompassing 38592 transcripts compared to the current 23574 transcripts in Ensembl. We also predicted 14717 protein-coding genes, achieving 96.4% completeness for Passeriformes lineage BUSCOs. A substantial improvement in this reannotation is the accurate delineation of untranslated region (UTR) sequences. We identified 82.7% and 93.8% of the transcripts containing 5′- and 3′-UTRs, respectively. These UTR annotations are crucial for understanding post-transcriptional regulatory processes. Our findings underscore the advantages of incorporating additional specific RNA-Seq data into genome annotation, particularly when leveraging fast and efficient data processing capabilities. This functional reannotation enhances our understanding of the *P. domesticus* genome, providing valuable resources for future investigations in various research fields.

Keywords: gene annotation; RNA-Seq; annotation pipeline; intestine; birds; *Passer domesticus*

Introduction

Studies in the past decade herald the potential power of "omics" analyses to advance the understanding of mechanistic bases underlying adaptations and differences in phenotypes within and across avian species [1–4]. Good-quality genome sequences and functional genome annotations represent essential resources for these studies. RNA sequencing (RNA-Seq) has been previously used to improve them, including (i) correcting predicted gene structures [5], (ii) detecting new alternative splicing isoforms [6], and (iii) discovering new genes and new transcripts [7, 8]. The available resources for RNA-Seq data in genome annotation vary from case to case based on factors such as ontogenetic state, organ and tissue specificity, presence of pathogens, and other additional factors. Thus, the potential for gene annotation improvement may always exist for those genes and transcripts expressed less widely, mainly in studies of specific tissues, ontogenetic states, or different experimental conditions.

We are using the omnivorous avian model *Passer domesticus* (house sparrows) to study physiological pathways and underlying mechanisms associated with nutritional flexibility [9, 10]. The house sparrow is the most widely distributed wild bird in the world and is somewhat unique among wild avian species in its close association with humans, not only in the agricultural environment, where presumably this association first evolved, but also in urban areas. Given our focus on intestinal functional studies, accurate gene identification is crucial due to the diverse and essential functions attributed to the single layer of epithelial cells (i.e. the enterocytes) that line the small intestine of vertebrates [10]. As of now, the genome of the house sparrow (GCA_001700915.1) has been sequenced using Illumina HiSeq technology, and its current gene set annotation was determined using Ensembl Rapid release (https://rapid.ensembl.org/Passer_domesticus_GCA_001700915.1/Info/Index), which is based on the Ensembl Genebuild Method [11]. The protein-coding annotation

Received: April 27, 2023. **Revised:** June 26, 2023. **Editorial Decision:** June 27, 2023. **Accepted:** July 05, 2023

method utilizes an RNA-Seq pipeline as the primary data source for model generation. However, the current annotation is solely based on Illumina RNA-Seq data from a single individual's muscle tissue, which limits the representation of other tissues. Therefore, it is likely that a considerable proportion of genes are misannotated in the reference gene set. Indeed, this problem became apparent during our analysis of RNA-Seq samples from house sparrows of different ages feeding on different diets. In the current Ensembl annotation, it was found that only 45.3% of its transcripts contain 5′-untranslated regions (UTRs), while 48.6% of the transcripts contain 3′-UTRs. Numerous studies emphasize the significance of short open reading frames within 5′-UTRs for fine-tuning gene functions [12]. On the other hand, 3′-UTR sequences are commonly required for experiment design [13]. Our objective was to generate an enhanced genome annotation version specifically tailored to intestinal tissue. To accomplish this, we utilized intestinal tissue expression data from a previous experiment involving wild-hatched house sparrow nestlings, which were later raised in environmental chambers on a semi-synthetic starch-based diet [14]. Initially, we employed Illumina RNA-Seq technology to assess RNA expression in the experimental samples. Subsequently, we utilized the GeMoMa pipeline [15] to construct the structural and functional annotation. Additionally, we aimed to optimize the annotation process by utilizing modest or low computational resources, specifically 20 CPUs and 100 GB RAM. The implemented GeMoMa-based pipeline proved to be highly effective in providing a valuable resource for re-annotating the genome.

Material and methods

Study site and sample collection

All samples were obtained from a previous experiment with house sparrow nestlings aimed at studying the phenotypic flexibility of the intestinal enzymes, maltase, sucrase, and aminopeptidase-N [14]. Briefly, 3 days post-hatch nestlings were collected from their wild nests located in dairy barns and a parking ramp near the Department of Forest and Wildlife Ecology on the University of Wisconsin–Madison campus. Animals were housed individually in environmentally controlled chambers in our laboratory. At their arrival and continuously for 6 days, the birds were fed a diet with a high-starch and low-casein content. Five nestlings were euthanized with CO_2 and dissected to remove the small intestine, and the medial regions were preserved in RNAlater (Invitrogen; Carlsbad, California, USA). Additional procedure details follow those in Rott, Caviedes-Vidal, and Karasov [14]. All experimental procedures were approved by the University of Wisconsin–Madison ethics committee (permit no. RARC A-0570-0-03-14).

RNA extraction, quantification, and integrity control

Total RNA was isolated from frozen tissue using the PureLink™ RNA Mini Kit (Invitrogen; Carlsbad, California, USA) according to the manufacturer's instructions. Before all the procedures we decontaminated equipment, benchtops, glassware, and plasticware using RNase AWAY™ Surface Decontaminant (Thermo Scientific; Waltham, MA, USA). All samples were assessed for purity and integrity using a NanoDropOne Spectrophotometer and an Agilent 2100 BioAnalyzer. Afterward, they were submitted to the University of Wisconsin–Madison Biotechnology Center for quantification.

The Illumina® TruSeq® Stranded mRNA Sample Preparation kit (Illumina Inc., San Diego, California, USA) was used to construct the libraries. For each library preparation, mRNA was purified from 1000 ng total RNA using poly-T oligo-attached magnetic beads. Subsequently, each poly-A-enriched sample was fragmented using divalent cations under elevated temperature. Fragmented RNA was synthesized into double-stranded cDNA using SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, California, USA) and random primers for first-strand cDNA synthesis followed by second-strand synthesis using DNA Polymerase I and RNase H for removal of mRNA. Double-stranded cDNA was purified by paramagnetic beads (Agencourt AMPure XP beads, Beckman Coulter). The cDNA products were incubated with Klenow DNA Polymerase to add an 'A' base (Adenine) to the 3′-end of the blunt DNA fragments. DNA fragments were ligated to Illumina's unique dual adapters, which have a single "T" base (Thymine) overhang at their 3′-end. The adapter-ligated DNA products were purified by paramagnetic beads. Adapter-ligated DNA was amplified in a Linker-mediated PCR reaction for 10 cycles using Phusion™ DNA Polymerase and Illumina's PE genomic DNA primer set and then purified by paramagnetic beads. Quality and quantity of the finished libraries were assessed on the ATTI Fragment Analyzer (Agilent Technologies, Inc., Santa Clara, California, USA) and Qubit® dsDNA HS Assay Kit (Invitrogen, Carlsbad, California, USA), respectively. Libraries were standardized to 2 nM. Paired-end 2 × 150 bp sequencing was performed on an Illumina NovaSeq6000 sequencer obtaining an average of 28.4 million per sample (Supplementary Table S1). Raw reads were submitted to the National Center of Biotechnology Information (NCBI) to the Sequence Read Archive under the Bioproject: PRJNA785148.

Passer domesticus genome annotation

The publicly available genome assembly for *P. domesticus* (GCA_001700915.1) was "cleaned" to remove some repetitive contigs that are contained in other scaffolds, sorted by length, and masked using Funannotate v1.8.1 [16].

The *P. domesticus* genome assembly was annotated using GeMoMa v.1.9 [16]. The GeMoMaPipeline function was executed to complete the full pipeline, with a maximum intron size of 200 kb. For reference, three closely related species genomes were utilized: *Passer montanus* (GCA_014805655.1), *Pyrgilauda ruficollis* (GCA_017590135.1), and *Onychostruthus taczanowskii* (GCA_017590135.1). These genomes were obtained from NCBI [17] and were selected based on their close phylogenetic relationship, belonging to the same genus or family. Moreover, their annotated transcriptomes displayed high completeness levels within the Passeriformes lineage in BUSCO analysis (99.2% in all cases). The intestinal RNA-Seq data generated (Supplementary Table S1), and also muscle RNA-Seq data available on Ensembl ([18]; https://ftp.ensembl.org/pub/rapid-release/species/Passer_domesticus/GCA_001700915.1/rnaseq/) were aligned to the *P. domesticus* genome using HISAT2 v.2.2.1 [19], and the aligned bam file incorporated to the pipeline. The RNA-Seq data were previously analyzed by Rcorrector [20] to correct sequencing errors, and TrimGalore [21] to remove low-quality reads and trim adapters with the parameters as default to avoid potential troubleshooting during the alignment step. GeMoMa utilizes mapped RNA-Seq data to predict UTRs, identify introns, and determine splice sites. However, it is constrained by the gene models established for the selected reference organisms, potentially missing any expressed genes absent in the chosen reference genomes and annotations. To address this limitation, the aligned short reads were assembled into transcripts using StringTie v2.2.1 [22]. The resulting GFF file was then integrated into the GeMoMaPipeline using its external annotation option.

The structural annotation obtained from GeMoMa was analyzed using BUSCO [23] to estimate the completeness of the gene protein dataset specific to the Passeriformes lineage. Furthermore, we utilized miniprot [24] to directly align the Passeriforme protein set to the genome in order to recover any missing genes from the BUSCO reference. This approach produced a complementary second structural annotation. Lastly, we merged the structural annotations from both GeMoMa and miniprot using the 'agat_sp_merge_annotations.pl script' [25]. To assess the potential of our RNA-Seq data in enhancing Ensembl gene annotations, we conducted an analysis that involved determining the number of expressed genes detected, as well as the mean and median read counts per sample. This evaluation considered the Ensembl gene models (Supplementary Table S1). To accomplish this task, we used featureCounts [26], running it on the obtained HISAT2 bam files with the following parameters: -s 2 -t "gene" -g 'ID' –minOverlap 1 –fracOverlap 0 –fracOverlapFeature 0 -p -C.

To obtain functional annotations for the final set of protein-coding genes, we performed a search of each predicted protein sequence against the InterPro protein database using InterProScan v5.57–9.0 [27]. Subsequently, we derived functional annotations from the Carbohydrate-Active Enzymes, Clusters of Orthologous Groups of proteins, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes databases using EggNOG-mapper v2.1.9 [28]. The obtained results were then incorporated into the annotation using AGAT's script: 'agat_sp_manage_functional_annotation' [25] (Fig. 1).

Comparison of annotations

The annotation summaries were generated using the 'agat_sp_functional_statistics.pl script' [25]. To minimize duplicates, we utilized the 'agat_sp_keep_longest_isoform.pl' script [25] to select the longest isoform for each protein-coding gene. We assessed the genome's completeness and both annotations (Ensembl and ours – PasserD) using BUSCO v5.3.2 [24], employing the single-copy orthologs reference sets from both the Passeriformes and Aves lineages (from OrthoDB passeriformes_odb10 and aves_odb10 databases, [29]). For manual inspection of gene annotations across the genome, we employed the Integrative Genomics Viewer [30].

Computing resources were provided by the University of Wisconsin–Madison Center for High Throughput Computing of the Department of Computer Sciences (University of Wisconsin–Madison, 53706 WI, USA).

Results and discussion

Reannotation of the *P. domesticus* genome

In this study, we generated an updated gene annotation for the *P. domesticus* genome (GCA_001700915.1), which is now referred to as PasserD (PasserD.gff; Supplementary File 2). We then conducted a comprehensive comparison between PasserD and the previous annotation, referred to as Ensembl, to assess differences and improvements. The reannotation process (Fig. 1) yielded a final set of 15614 protein-coding genes and 38592 transcripts, resulting in an average of 2.6 transcript isoforms per gene at the whole-genome scale. In the PasserD annotation, 37179 transcripts were found to have 3′-UTRs or 5′-UTRs, representing approximately 96.3% of all the annotated transcripts. The mean number of exons per gene increased from 12.5 in the prior annotation to 15.3 in PasserD. The PasserD annotation also contains alternatively spliced or alternatively initiated transcripts. We used eggNOG-mapper and InterProScan to perform comprehensive annotation of the transcripts. The analysis revealed that a significant portion of the total transcripts, specifically 87.2% (33660 transcripts), were successfully assigned GO terms. Moreover, 94.8% (36 567 transcripts) of the transcripts received complementary functional annotations. These attributes were not present in the Ensembl annotation, indicating that our

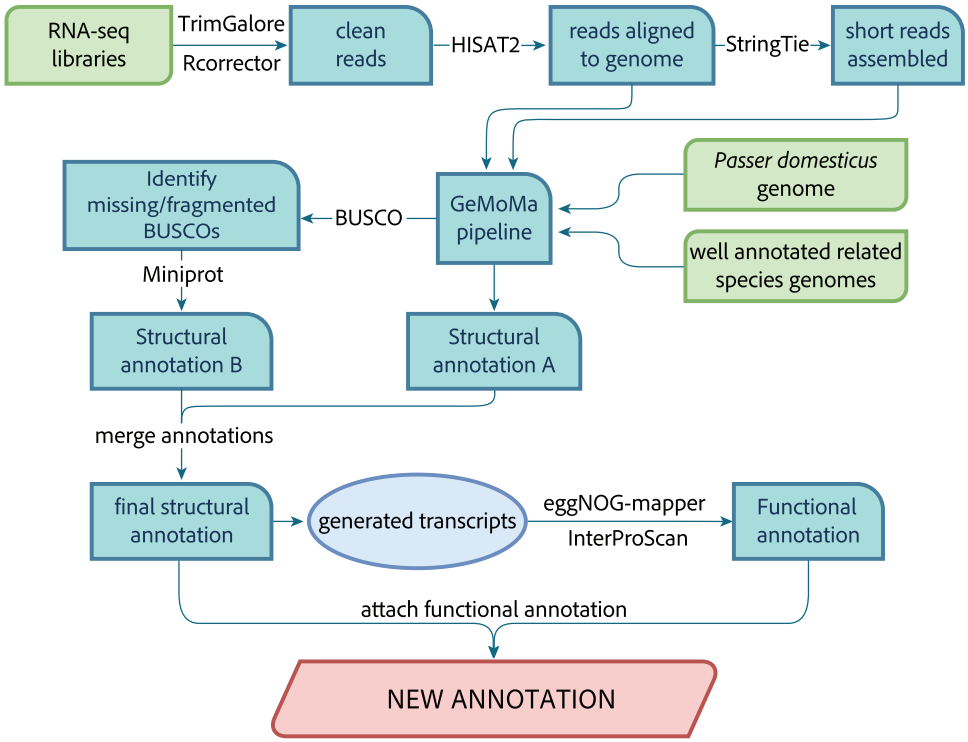

Figure 1. Annotation workflow for *P. domesticus* protein-coding genes.

annotation approach provided additional valuable information (Table 1).

Evaluation and assessment of assembly completeness

To evaluate the improvement in the reannotation, we compared the performance of both annotations, PasserD and Ensembl, using BUSCO in transcriptome mode on the near-universal single-copy orthologs gene set in the Passeriformes lineage. The PasserD reannotation covered 96.4% of the 10 844 Passeriformes gene set, whereas the previous Ensembl annotation achieved a completeness of 88.4% using the same gene set. This result demonstrates that PasserD most successfully captures a significant proportion of the known conserved genes within the Passeriformes lineage. Additionally, we assessed the performance of both annotations in the Aves lineage-conserved orthologous gene set (8338 genes). Completeness of PasserD reannotation was 96.4%, while Ensembl annotation accounted for 90.7% of the same gene set. This evaluation allowed us to gage the effectiveness of the PasserD reannotation in improving the annotation coverage of conserved genes in the Passeriformes lineage, as well as the broader Avian gene set (Fig. 2).

Table 1. Summary of Ensembl and PasserD annotations

Type	Ensembl	PasserD
Protein-coding genes		
Number of genes	14188	15614
Number of transcripts	22926	38592
Number of exons	291615	641666
Transcripts with a 5′-UTR	10379	31917
Transcripts with a 3′-UTR	11147	36234
Transcripts with 5′- and 3′-UTR	NA[1]	28128
Mean gene length (bp)	26244	36368
Mean transcript length (bp)	33393	53729
Mean number of transcripts per gene	1.6	2.6
Mean CDS length (bp)	1987	2467
Mean number of exons per CDS	12.5	15.3
Mean exon length (bp)	192	355
Mean 5′-UTR length (bp)	270	784
Mean 3′-UTR length (bp)	698	2970
Transcripts with associated GO terms	0	33660
Transcripts with associated gene name	0	36525
Transcripts with functional annotation	0	36567
Complete Aves lineage BUSCOs, %	89.20	96.40
Missing Aves lineage BUSCOs, %	7.80	2.30

[1] not available value

Prediction of gene functions

In the PasserD reannotation (Supplementary File 3), eggNOG assigned specific GO terms to 33469 transcripts, which were not included in the previous annotation. The new annotation demonstrates an improvement over the previous annotation, which can be illustrated through the following two examples. First, let us consider the *PASSERD08661* gene, which encodes the sucrase–isomaltase (SI) enzyme. This enzyme is expressed on the luminal side of the intestinal brush border membrane and plays a crucial role in the digestion of dietary carbohydrates. It facilitates the breakdown of starch derivatives, such as maltose, isomaltose, malto-oligosaccharides, and sucrose [31, 32]. In PasserD, SI exhibits extended evidence of additional exons and splicing sites, resulting from the integration of two previously split genes. This integration has led to the formation of a new

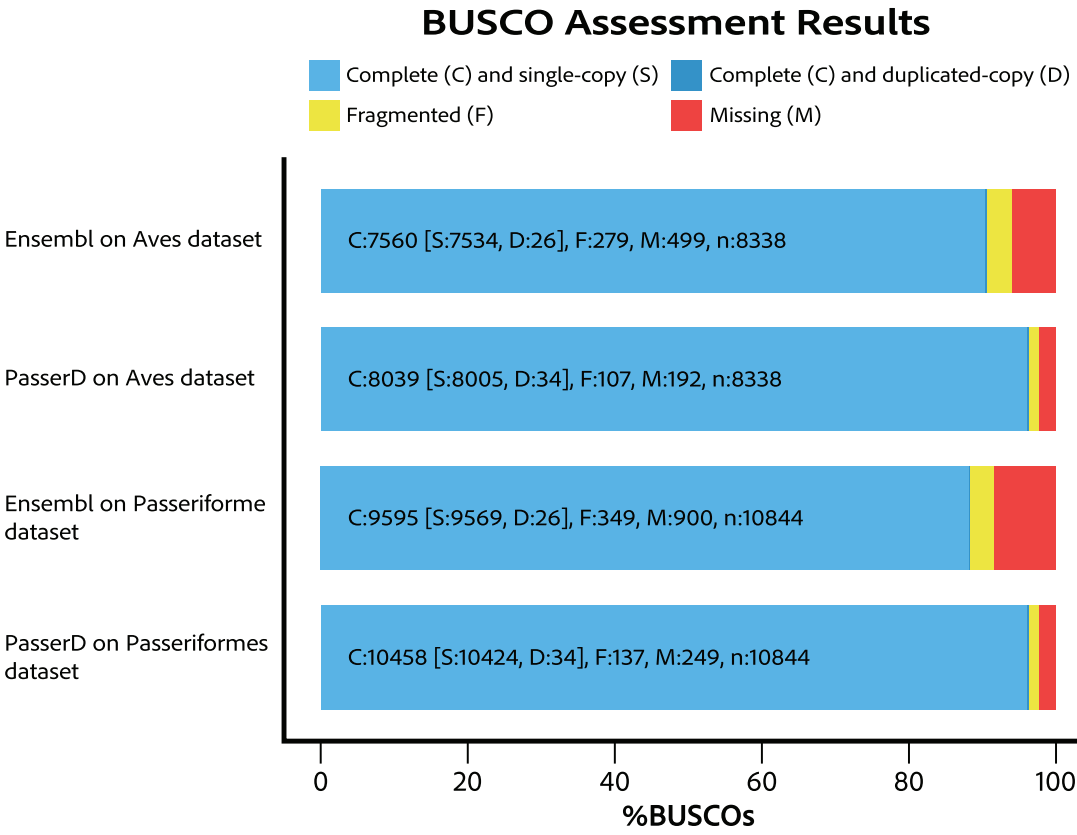

Figure 2. To assess the gene set completeness of the *Passer domesticus* genome annotations, PasserD and Ensembl, we utilized BUSCO 5.2.2 [23] in transcriptome mode. This analysis involved two distinct datasets: the Passeriformes and Aves lineage datasets.

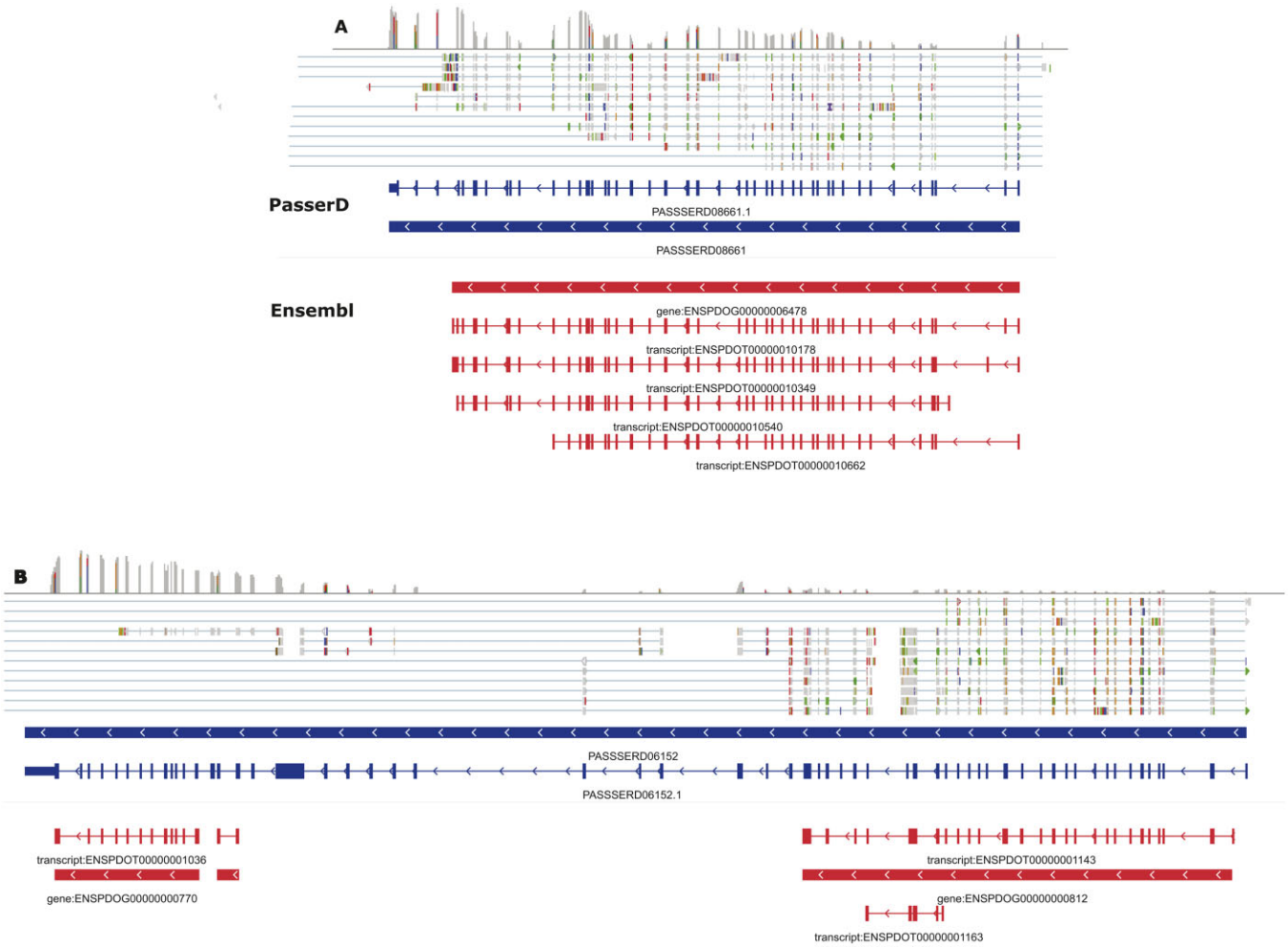

Figure 3. Examples of known genes with improved annotation. IGV views of the transcript models encoding (**A**) SI and (**B**) MUC2 according to PasserD (blue) and Ensembl (red) annotations with RNA-Seq mapped reads. The blue and red lines with gene IDs below represent the whole gene. The blue and red lines with mRNA IDs below show the exon and intron gene structure in detail. The thinner bars at the transcript ends represent UTRs, the thick bars represent exons forming the coding sequence (CDS), and the thin lines represent the introns. The arrowheads within the lines indicate transcriptional orientation.

fully annotated gene with a longer transcript length, as depicted in Fig. 3A. The second example involves the gene *PASSERD06152*. In Ensembl, it was annotated as two distinct genes, but in PasserD, it is considered as a single gene. This gene is responsible for encoding Mucin 2 (MUC2), a member of the mucin protein family. Mucins are high molecular weight glycoproteins produced by various epithelial tissues. The protein encoded by this gene is secreted and forms a protective, insoluble mucous barrier in the gut lumen [33] (Fig. 3B). Both examples demonstrate an expansion of UTR regions.

Furthermore, we discovered complete or fragmented genes in PasserD that were not present in the Ensembl annotation. When we examined their overrepresentation in the Biological Process domain using Panther 17.0 [34], we observed enrichment in GO terms related to cellular processes associated with the assembly, arrangement, or disassembly of a cilium. Cilia are slender projections found on the surface of eukaryotic cells. This result is consistent with the fact that the included RNA-Seq data are of small intestinal origin, and it is expected to improve the gene annotation of those genes expressed exclusively and/or predominantly in this tissue. As in mammals, the intestinal epithelium in birds is a single layer of cells, mainly composed of columnar absorptive cells (enterocytes) [35]. In such cells, the apical surface area is greatly increased by microvilli, tightly packed finger-like projections of the membrane into the lumen. These microvilli may be considered as the primary cellular interface between the lumen and the inside of the organism in vertebrates, housing numerous proteins related to hydrolytic and absorptive processes. The exceptionally uniform size, orientation, and density of microvilli are generated by a complex network of proteins and signaling molecules, such as Villin (VIL1) and Plastin 1 (PLS1), which bundle actin filaments in coordination with Ezrin (EZR) and Myosin 1a (MYO1A) that cross-link the core actin bundles to their surrounding membrane [36]. The genes for all these proteins are found in the new annotation PasserD (Table 2), and the improved annotation of these genes is depicted in Supplementary Fig. S1.

In terms of computing resources, annotation pipelines typically require a high-performance server-grade computer with a significant amount of RAM and multiple CPU cores to handle the computational demands efficiently. An additional benefit obtained using the proposed pipeline is that all the required steps to prepare the data (e.g. Funannotate, Trimgalore, HISAT2, or StringTie) can be run in a personal workstation (8–16 cores and 16–32 GB of RAM) and then a computational system with more resources can be used to complete the GeMoMa step. The new

Table 2. Gene enrichment analysis of GO terms of the Biological Process domain of the new genes annotated in PasserD using the overrepresentation test (Fisher's test statistics) in PANTHER [34]

GO biological process complete	Ref.	New	exp	f.e.	o	*P*-value	FDR
cilium flagellum assembly (GO: 0120316)	36	9	1.56	5.76	+	8.26E-05	3.01E-02
cilium movement (GO: 0003341)	170	21	7.38	2.84	+	4.95E-05	2.16E-02
microtubule-based process (GO: 0007017)	809	62	35.13	1.76	+	3.99E-05	2.02E-02
motile cilium assembly (GO: 0044458)	62	12	2.69	4.46	+	5.10E-05	2.10E-02
cilium assembly (GO: 0060271)	322	35	13.98	2.5	+	2.83E-06	4.44E-03
organelle assembly (GO: 0070925)	803	61	34.87	1.75	+	5.56E-05	2.18E-02
organelle organization (GO: 0006996)	3026	176	131.39	1.34	+	8.04E-05	3.00E-02
plasma membrane bounded cell projection assembly (GO: 0120031)	418	40	18.15	2.2	+	1.35E-05	9.60E-03
cell projection assembly (GO: 0030031)	430	41	18.67	2.2	+	1.03E-05	8.06E-03
cilium organization (GO: 0044782)	355	38	15.41	2.47	+	2.25E-06	5.03E-03

From left to right columns contain: the name of the annotation data category, Ref.: number of genes in the reference list used (*Homo sapiens*), New: number of genes of the new annotation PasserD used, EXP: number of expected genes based on the reference list, f.e.: fold enrichment factor of the genes in the uploaded list over the expected, o: indicates if the category is overrepresented (+), *P*: raw value *P*-value determined by Fisher's exact test ($P < 0.05$ is considered statistically significant), FDR: False discover rate calculated by the Benjamini–Hochberg procedure (a critical value of 0.05 was used to filter results).

annotation of the house sparrow genome was completed in 4 h using 20 CPUs, and the peak RAM usage was lower than 100 GB (high-end workstation or entry-level server) and a load of 6 BAM files of 5–6 GB each. Even faster computational times can be achieved using additional CPUs in each step, and run times are both a function of the evidence dataset presented for alignment as well as the gene density of a genome, but the observed throughput of $>200\,\mathrm{Mb\,h^{-1}}$ demonstrates that even the largest of eukaryotic genomes could be annotated in a reasonable period of time.

Conclusions

The genome annotation of protein-coding genes is a necessary step for all downstream analyses, and the choice of used resources and methods significantly impacts annotation quality and completeness. In this study, we utilized efficient and resource-friendly computational tools and an optimized annotation pipeline to enhance the annotation of the *P. domesticus* genome. To achieve this improvement, we made use of Illumina short reads obtained from intestinal tissue. They not only updated the gene models but also increased alternatively spliced isoforms and achieved a total of 80.9% and 93.11% of transcripts containing 5′- and 3′-UTRs, respectively. Given the material used, we extend the ongoing development of genomic resources for house sparrows focusing on functional genomics [37].

The house sparrow transcriptome data provided here can serve as a reference for comparative genomics, phylogenomics, and the analysis of differential gene expression. They offer an exciting opportunity to explore the mechanistic base of processes within and between species. Comparisons like these provide a valuable resource that advances various fields, including ecological and evolutionary physiology, conservation biology, and ecotoxicology. As a result, the impact of such research is expected to provide important insights for both animal and human biomedical scientists.

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Author contributions

Melisa Magallanes (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review and editing [lead]), Agustin Baricalla (Conceptualization [equal], Data curation [supporting], Formal analysis [supporting], Investigation [supporting], Methodology [equal], Software [equal], Validation [supporting], Visualization [supporting], Writing – original draft [supporting], Writing – review and editing [supporting]), Natalia Rego (Conceptualization [equal], Data curation [supporting], Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Software [supporting], Supervision [equal], Validation [equal], Visualization [supporting], Writing – original draft [supporting], Writing – review and editing [supporting]), Antonio Brun (Investigation [equal], Resources [equal], Writing – original draft [supporting], Writing – review and editing [supporting]), William H. Karasov (Conceptualization [supporting], Funding acquisition [equal], Investigation [supporting], Project administration [supporting], Resources [supporting], Supervision [supporting], Writing – original draft [supporting], Writing – review and editing [supporting]), and Enrique Caviedes-Vidal (Conceptualization [supporting], Funding acquisition [lead], Project administration [supporting], Resources [lead], Supervision [lead], Validation [lead], Writing – original draft [supporting], Writing – review and editing [supporting].

Funding

The UW–Madison Center for High-Throughput Computing (CHTC) is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the OSG Consortium, which is supported by the National Science Foundation and the US Department of Energy's Office of Science. This work was supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) (PIP 834), the Universidad Nacional de San Luis (Ciencia y Técnica 2-0814), the National Science Foundation (IOS-1354893), and the Department of Forest and Wildlife Ecology, University of Wisconsin–Madison.

Conflict of interest statement. None declared.

Data availability

RNA-Seq data for this project have been deposited in NCBI website under BioProject ID PRJNA785148. Passer domesticus genome version 1.0 (GCA_001700915.1) available at: https://ftp.ensembl.org/pub/rapidrelease/species/Passer_domesticus/GCA_001700915.1/ensembl/genome/. Other supporting data are included as

Supplementary files: (A) NCBI accession numbers of RNA-Seq data are in Table S1 in Supplementary file 1, the gff file in Supplementary file 2 and the functional annotation is described in Supplementary file 3.

References

1. Künstner AXEL, Wolf JBW, Backström N *et al.* Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Mol Ecol* 2010;**19**:266–76.
2. Balakrishnan CN, Mukai M, Gonser RA *et al.* Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior. *PeerJ* 2014;**2**:e396.
3. Marasco V, Fusani L, Pola G, Smith S. Data on the *de novo* transcriptome assembly for the migratory bird, the Common quail (*Coturnix coturnix*). *Data Brief* 2020;**32**:106041.
4. Frias-Soler RC, Villarín Pildaín L, Wink M *et al.* A revised and improved version of the northern wheatear (*Oenanthe oenanthe*) transcriptome. *Diversity* 2021;**13**:151.
5. Larsen PE, Trivedi G, Sreedasyam A *et al.* Using deep RNA sequencing for the structural annotation of the Laccaria bicolor mycorrhizal transcriptome. *PLoS One* 2010;**5**:e9780.
6. Filichkin SA, Priest HD, Givan SA *et al.* Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res* 2010;**20**:45–58.
7. Guttman M, Garber M, Levin JZ *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 2010;**28**:503–10.
8. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–5.
9. Karasov WH, Caviedes-Vidal E. Adaptation of intestinal epithelial hydrolysis and absorption of dietary carbohydrate and protein in mammals and birds. *Comparative Biochemistry and Physiology, Part A* 2021;**253**:119860.
10. Karasov WH, Martinez del Rio C, Caviedes-Vidal E. Ecological physiology of diet and digestive systems. *Annu Rev Physiol* 2011;**73**:69–93.
11. Aken BL, Ayling S, Barrell D *et al.* The Ensembl gene annotation system. *Database* 2016;**2016**:baw093.
12. Dvorak P, Leupen S, Soucek P. Functionally significant features in the 5′ untranslated region of the ABCA1 gene and their comparison in vertebrates. *Cells* 2019;**8**:623.
13. Mayr C. What are 3′ UTRs doing? *Cold Spring Harb Perspect Biol* 2019;**11**:a034728.
14. Rott KH, Caviedes-Vidal E, Karasov WH. Intestinal digestive enzyme modulation in house sparrow nestlings occurs within 24 hours of a change in diet composition. *J Exp Biol* 2017;**220**:2733–42.
15. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol* 2019;**1962**:161–77.
16. Palmer JM, Stajich J. *Funannotate v1.8.1: Eukaryotic Genome Annotation*. Zenodo, 2020.https://zenodo.org/record/4054262 (accessed February 23, 2022)
17. Sayers EW, Cavanaugh M, Clark K *et al.* GenBank 2023 update. *Nucleic Acids Res* 2023;**51**:D141–4.
18. Cunningham F, Achuthan P, Akanni W *et al.* Ensembl 2019. *Nucleic Acids Res* 2019;**47**:D745–51.
19. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15.
20. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaSci* 2015;**4**:s13742-015-0089-y.
21. Krueger F, James F, Ewels P *et al. TrimGalore: V 0. 6.7*. Zenodo, 2021. https://zenodo.org/record/5127899 (accessed February 23, 2022).
22. Shumate A. The Development and Application of Computational Methods for Genome Annotation. *Doctoral dissertation*. Johns Hopkins University, 2022. http://jhir.library.jhu.edu/handle/1774.2/67233.
23. Manni M, Berkeley MR, Seppey M *et al.* BUSCO: assessing genomic data quality and beyond. *Curr Protoc* 2021;**1**:e323.
24. Li H. Protein-to-genome alignment with miniprot. *Bioinformatics* 2023; 39:btad014.
25. Dainat J, Hereñú D, Davis E. *et al.* NBISweden/AGAT: AGAT-v1.1.0. Zenodo 2023. https://zenodo.org/record/7950165 (accessed May 19, 2023).
26. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
27. Jones P, Binns D, Chang H-Y *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;**30**:1236–40.
28. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 2021;**38**:5825–9.
29. Kriventseva EV, Kuznetsov D, Tegenfeldt F *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2019;**47**:D807–d811.
30. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**:178–92.
31. Caviedes-Vidal E, Afik D, Martinez del Rio C *et al.* Dietary modulation of intestinal enzymes of the house sparrow (*Passer domesticus*): testing an adaptive hypothesis. *Comp Biochem Physiol A Mol Integr Physiol* 2000;**125**:11–24.
32. Semenza G, Auricchio S Small-Intestinal Disaccharidases. In The Online Metabolic and Molecular Bases of Inherited Disease, eds: C. R. Scriver, A. L. Beaudet, W. S. Sly et al. New York, NY: McGraw-Hill Education., **1989**, 2975–97.
33. Johansson MEV, Hansson GC. Immunological aspects of intestinal mucus and mucins. In Ratcliffe MJH (ed.), *Encyclopedia of Immunobiology*. Oxford: Academic Press, 2016, 381–8.
34. Thomas PD, Ebert D, Muruganujan A *et al.* PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci* 2022;**31**:8–22.
35. Starck JM. Shaping up: how vertebrates adjust their digestive system to changing environmental conditions. *Animal Biol* 2003;**53**:245–57.
36. Crawley SW, Shifrin DA, Grega-Larson NE *et al.* Intestinal brush border assembly driven by protocadherin-based intermicrovillar adhesion. *Cell* 2014;**157**:433–46.
37. Singh PK, Singh RP, Singh RL. From gene to genomics: tools for improvement of animals. In Mondal S, Singh RL (eds), *Advances in Animal Genomics*. Academic Press, 2020, 13–32.