Article

Not peer-reviewed version

TrAnnoScope: A Modular Snakemake Pipeline for Comprehensive Full-Length Transcriptome Analysis and Functional Annotation

Aysevil Pektas , Frank Panitz , Bo Thomsen *

Posted Date: 7 November 2024

doi: 10.20944/preprints202411.0489.v1

Keywords: RNA-Seq; reproducible pipeline; high-performance computing (HPC); transcriptome analysis; functional annotation; ISO-seq; snakemake; long-read sequencing

 doi:10.20944/preprints202411.0489.v1

Article

TrAnnoScope: A Modular Snakemake Pipeline for Comprehensive Full-Length Transcriptome Analysis and Functional Annotation

Aysevil Pektas [1], Frank Panitz [1,2] and Bo Thomsen [1,*]

[1] Department of Molecular Biology and Genetics, Aarhus University, 8000 Aarhus, Denmark; aysevilpektas@mbg.au.dk (A.P.); frank.panitz@luke.fi (F.P.)
[2] Applied Statistical Methods, Natural Resources Institute Finland (Luke), 20520 Turku, Finland
* Correspondence: bo.thomsen@mbg.au.dk

Abstract: Transcriptome assembly and functional annotation are essential for understanding gene expression and biological function. Nevertheless, many existing tools lack the flexibility to integrate both short- and long-read sequencing data or fail to provide a complete, customizable workflow for transcriptome analysis. Here, we present TrAnnoScope, a comprehensive transcriptome analysis pipeline that provides a complete, customizable workflow capable of efficiently processing and integrating short- and long-read sequencing data to generate high-quality, full-length transcripts with detailed functional annotation. The pipeline encompasses steps from quality control to functional annotation, employing a range of tools and established databases, such as SwissProt, Pfam, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes. As a case study, TrAnnoScope was applied to RNA-Seq data from zebra finch brain, ovary, and testis tissues. TrAnnoScope produced comprehensive transcriptome, demonstrating strong alignment with the reference genome (99.63%) and capturing a significant proportion of nearly complete protein sequences (92.7%). The functional annotation process yielded extensive matches to known protein databases and successfully assigned relevant functional terms to majority of the transcripts. As such, TrAnnoScope successfully integrates multiple sequencing technologies to generate comprehensive transcriptomes with minimal user input. Its modular design, flexibility, and ease of use make it a valuable tool for researchers working with complex datasets, particularly for non-model organisms.

Keywords: RNA-Seq; reproducible pipeline; high-performance computing (HPC); transcriptome analysis; functional annotation; ISO-seq; snakemake; long-read sequencing

1. Introduction

RNA sequencing (RNA-Seq) has become a powerful tool for detecting novel transcripts, understanding gene expression, cataloging protein-coding genes, and revealing the biological functions of genes [1–3]. Additionally, RNA-Seq has unlocked the study of non-model organisms without the need for a reference genome, through de novo transcriptome analysis [4]. However, there are several challenges associated with RNA-Seq analysis, such as sequencing errors and fragmentation resulting from technological limitations, in addition to issues, such as repetitive regions and overlapping genes due to transcriptome complexity [2,4].

Short-read technologies have lower error rates and provide a higher coverage than long-read sequencing technologies. Nevertheless, transcriptomes generated exclusively from short reads often suffer from fragmentation and incomplete transcript reconstructions due to the erroneous computational predictions of isoforms. In contrast, long-read sequencing technologies can capture full-length (FL) transcripts and resolve isoform complexity; however, they retain a higher error rate and lower throughput. Accurate transcriptome assembly is crucial for downstream analyses, including functional genomics, gene discovery, and the elucidation of complex biological processes [4]. Hybrid approaches that combine short- and long-read technologies can overcome the weaknesses

 doi:10.20944/preprints202411.0489.v1

of each technology and improve transcriptome coverage and accuracy to obtain known and novel transcripts [2,5].

Transcriptome generation and annotation are challenging because of the complexity of the procedures, the need to select appropriate tools, and the significant computational resources required [3]. Several RNA-Seq pipelines offer an interconnected collection of tools designed to automate the process, such as RNAflow [6] and RASflow [7], which primarily focus on differential expression analysis, while others, such as TransXpress [8], TransPi [9], and Pincho [10], focus on de novo transcriptome assembly and functional annotation. However, these tools depend on short-read sequencing for analysis.

Several toolkits, such as Functional IsoTranscriptomics Analysis (FIT) [11], IsoTools [12], TAGET [13], and nf-core/isoseq [14] utilize the properties of long-read sequencing technologies for transcriptome analysis. However, they are primarily designed to function with reference annotations, and currently, Trans2Express [15] is the only reproducible protocol for non-model organisms. It enables de novo hybrid transcriptome assembly using both the Illumina and Oxford Nanopore Technologies (ONT) platforms, aiming to recover a single transcript per gene for transcriptome characterization and gene expression analysis. However, this approach may lead to the loss of important information relating to alternative splicing and isoform diversity and limit the detection of novel transcripts or isoforms, especially in the less-studied regions of a transcriptome; a more comprehensive approach that captures multiple isoforms is essential for fully elucidating the functional potential of genes [16]. Furthermore, Trans2Express offers limited flexibility, restricting users from selecting and combining tools within the pipeline to meet their specific research objectives.

Here, we present TrAnnoScope, a comprehensive FL transcriptome and annotation pipeline that integrates Illumina short-and long-read data through a number of key steps, including data preprocessing, long-read error correction, contamination removal, quality assessment, and comprehensive functional annotation. The pipeline was designed to improve transcriptome accuracy and completeness by leveraging the strengths of long reads using short reads and thorough preprocessing without the need for reference annotation. This integration can resolve complex isoform structures, identify novel genes, and provide a complete transcriptome representation. TrAnnoScope is highly modular, allowing users to customize workflows and integrate different components to suit their research goals. In addition, it supports parallel execution and cluster computing, enabling the faster processing of larger datasets. It is sufficiently versatile to be used in many research areas, from gene discovery and transcriptome profiling to comparative genomics and the study of complex biological systems. As a result, TrAnnoScope provides a powerful, reproducible approach for conducting efficient bioinformatics analyses for large-scale transcriptome analyses.

2. Materials and Methods

2.1. Components of the TrAnnoScope Pipeline

We implemented our pipeline using Snakemake owing to its simplicity and ability to automate complex workflows while managing dependencies [17]. The TrAnnoScope pipeline consists of several modules, starting with a Python script (Figure 1A) that simplifies the setup by installing the necessary dependencies and databases for TrAnnoScope, allowing users to install only the components essential for their needs. Following this, the pipeline includes quality control of Illumina reads (Figure 1B), preprocessing of Illumina reads (Figure 1E), preprocessing of PacBio reads (Figure 1C), contamination removal, error correction using Illumina reads, isoform clustering and classification (Figure 1F), quality assessment (Figure 1D), and annotation (Figure 1G). A configuration file is provided for users to customize the tool parameters according to their needs. To address the time-intensive steps, such as contamination removal and annotation, TrAnnoScope supports parallel execution by dividing input files for faster processing. Detailed instructions are available on the TrAnnoScope GitHub page, [https://github.com/aysevllpkts/TrAnnoScope accessed on 6 Nov 2024], and users can run the pipeline fully or selectively based on their needs.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 November 2024 doi:10.20944/preprints202411.0489.v1

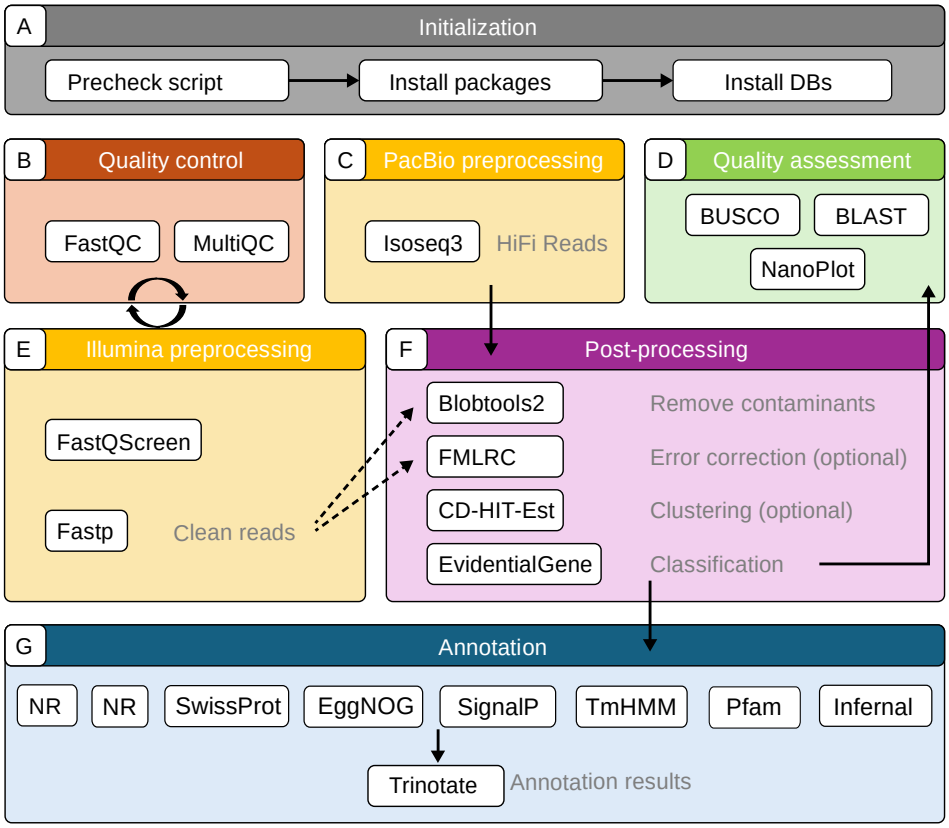

Figure 1. Overview of the TrAnnoScope pipeline consisting of the following steps: (**A**) initialization, (**B**) quality control of Illumina reads, (**C**,**E**) preprocessing of Illumina and PacBio reads, respectively; (**D**) quality assessment; (**F**) post-processing including the removal of contaminants, error correction, and cluster and classification steps, and (**G**) annotation.

2.2. Pipeline Implementation

2.2.1. Quality and Preprocessing for Illumina Reads

The first module of the TrAnnoScope pipeline focuses on performing the quality control of Illumina short reads, which is an essential step for ensuring data reliability. `FastQC` [18] is used to generate quality reports for individual samples, and `MultiQC` [19] combines these reports into a comprehensive summary of all Illumina datasets.

To detect potential contamination, `FastQScreen` [20] maps the reads against a set of reference databases. TrAnnoScope allows users to use the default FastQScreen Genome database or create a custom database to filter out undesired sequences, such as species-specific mitochondrial RNA (mtRNA) or ribosomal RNA (rRNA) sequences. For adapter trimming and quality filtering, `fastp` [21] is employed due to its speed, low memory usage, and detailed quality reports. Additionally, it detects and removes poly-G tails from Illumina NextSeq/NovaSeq data. Ensuring high-quality reads, such as via contamination removal and error correction, is crucial for downstream analysis, as poor-quality reads can lead to inaccurate results. A final quality control step is performed for preprocessed reads using `FastQC` and `MultiQC` to confirm that they are suitable for downstream analysis.

2.2.2. Preprocessing of PacBio Reads

The `Isoseq3` package [22] is used to preprocess the PacBio reads, which is essential for converting raw sequencing data into high-quality, FL transcripts. The module begins by converting subreads into circular consensus sequences (CCSs), which are self-corrected sequences, using the `ccs` tool to ensure a higher accuracy in the downstream analysis. The `lima` tool is then used to

 doi:10.20944/preprints202411.0489.v1

demultiplex and trim primer sequences from the CCS reads to obtain FL reads. Further, `isoseq3 refine` removes chimeric sequences, which can originate from multiple transcripts, and retains only reads with poly-A tails, producing FL non-chimeric (FLNC) reads that improve read quality. `Isoseq3 cluster/cluster2` groups reads based on their sequence similarity to identify unique transcripts and their isoforms, offering insights into alternative splicing and transcript diversity, both of which are crucial for understanding gene expression dynamics. This preprocessing step currently applies only to the PacBio reads. However, if the user has Nanopore reads in the FASTA format, they can still be used for the subsequent steps of the pipeline.

2.2.3. Contamination Removal

Contaminants in RNA-Seq data can significantly impact the quality and accuracy of results, leading to biased gene expression and the incorrect identification of splice variants. These contaminants can be introduced at various stages of the RNA-Seq process [23,24].

To address this, the TrAnnoScope pipeline uses `Blobtools2` [25] to remove microbial and cross-contamination from long reads. `Blobtools2` analyzes, visualizes, and filters assemblies based on the GC content, coverage, and taxonomic information. This tool is especially useful for de novo data, helping to identify and remove contaminants, thereby improving the quality of assemblies. The input files for `Blobtools2` are generated using `Bowtie2` [26] for coverage data, Benchmarking Universal Single-Copy Orthologs (`BUSCO`) for taxonomic classification [27], and `BLAST` [28] against the National Center for Biotechnology Information (NCBI) nucleotide (NT) database for taxonomic information.

2.2.4. Error Correction

Error correction can be beneficial for improving the accuracy of long reads, which typically have higher error rates than short reads [29]. In TrAnnoScope, the `FMLRC` tool is available as an optional step for correcting errors in long reads by leveraging complementary Illumina reads. `FMLRC` utilizes a multi-string Burrows–Wheeler transform and FM index to retrieve k-mer frequencies and construct de Bruijn graphs from short reads. It performs two passes with short and long k-mer values to correct unsupported regions in long reads, resulting in a more comprehensive correction process [30]. Owing to its efficiency and accuracy, `FMLRC` is a robust choice for error correction [29].

2.2.5. Clustering and Classification

To eliminate redundancy and reduce the complexity of the transcript data, TrAnnoScope employs two tools for clustering and classification: `CD-HIT-Est` [31], which is provided as an optional tool, and EvidentialGene [32]. `CD-HIT-Est` clusters transcripts based on their sequence similarity to remove redundancy within each sample, whereas `EvidentialGene` classifies the transcripts as primary and alternate forms based on their quality and potential function across the combined dataset.

2.2.6. Quality Assessment

TrAnnoScope includes a quality assessment step that utilizes several tools to evaluate the transcriptome comprehensively. `NanoPlot` [33] generates descriptive statistics, such as the mean, median, and N50 values, providing a clear overview of transcript continuity. `BUSCO` [27] further assesses transcriptome completeness against a user-defined lineage, ensuring the presence of elements for the organism of interest. Additionally, transcriptome quality is evaluated by comparing the number of FL or nearly FL transcripts against known protein databases, using an approach similar to the `Trinity` method of counting FL transcripts [34]. However, we implemented an in-house Bash script that calculates the percentage of high-coverage proteins present in the transcriptome compared to SwissProt by default or a user-defined custom database for closely related organisms. Together, these methods provide a detailed and accurate measure of transcriptome quality.

 doi:10.20944/preprints202411.0489.v1

2.2.7. Annotation

Functional annotation of the transcriptome is performed using `Trinotate` [35], an advanced annotation suite created for the automated functional annotation of transcriptomes. `Trinotate` integrates multiple sequence databases, including Pfam [36], SwissProt [37], SignalP [38], TMHMM [39], EggNOG [40], and Infernal [41], to provide a comprehensive annotation. Additionally, TrAnnoScope provides homology searches against NCBI (non-redundant protein) NR and NT databases. Users can select the databases they wish to use for annotation in the configuration file. However, they must manually download and prepare the necessary files for the NT and NR databases prior to the annotation process. To facilitate this, Bash scripts to automate the downloading and indexing of the databases were provided.

To accelerate the annotation process, we implemented a strategy that splits the input files into user-defined chunks, enabling the parallel execution of homology searches. This approach significantly reduces the time required to annotate large datasets. After completing the homology searches, the results are parsed using modified `Trinotate` helper scripts to generate detailed annotation files enriched with Gene Ontology (GO) [42] from SwissProt, Pfam and EggNOG. For further functional insights, we implemented in-house R scripts to create plots, including GO, Kyoto Encyclopedia of Genes and Genomes (KEGG) [43], Eukaryotic Orthologous Groups (KOG) [44], and species distributions (based on the NCBI NR database).

2.2.8. Data Selection

To demonstrate the functionality and versatility of our pipeline, we processed publicly available RNA-Seq reads from the zebra finch (*Taeniopygia guttata*), an avian model used to study the neural mechanisms of local learning and social behavior [45]. This species is notable for its complex vocalizations, ease of breeding in captivity, and pronounced sexual dimorphism, making it a valuable model for understanding vocal learning and its implications for human speech and language development [45,46].

RNA-Seq data were obtained from the NCBI Sequence Read Archive [47] (SRR8551559, SRR8551563, SRR8551565, SRR8551567, SRR8551558, SRR8551562, SRR8551564, and SRR8551566). This dataset includes Illumina NextSeq 500 paired-end reads (2 × 76 bp) and long-read sequences generated via PacBio SMRT Sequel from various tissues, including the brain, ovary, and testis (Table S1). These complementary data types provided an ideal scenario for evaluating the ability of the pipeline to integrate and interpret both short and long reads for a comprehensive transcriptome assembly. TrAnnoScope was executed on the GenomeDK cluster using SLURM for all analysis steps. Detailed information about the outputs and runtime is provided in Supplementary File S3.

To fully utilize TrAnnoScope, users must provide both long and short reads. However, the pipeline is flexible: users can choose to run the quality control and preprocessing modules with only Illumina reads or execute the preprocessing, clustering and classification, quality assessment, and annotation modules with only long reads.

2.2.9. Mapping to the Zebra Finch Reference Genome

The final transcriptome generated by the TrAnnoScope pipeline was mapped to the current zebra finch reference genome (RefSeq: GCF_003957565.2) using `minimap2 -ax splice --secondary=no -C5` parameters. Alignment statistics were obtained using the `samtools flagstats` option.

3. Results and Discussion

In this study, we applied TrAnnoScope to RNA-Seq data from the zebra finch (*Taeniopygia guttata*) to evaluate its effectiveness in transcriptome assembly and functional annotation. The dataset included Illumina reads and PacBio long-read sequences from brain, ovary, and testis tissues. Our primary objective was to assess the ability of the pipeline to integrate and interpret these complementary data types to achieve a comprehensive transcriptome assembly. This section details

 doi:10.20944/preprints202411.0489.v1

the results obtained, focusing on the key findings related to preprocessing, contamination removal, error correction, clustering and classification, quality assessment, and functional annotation.

In the preprocessing step for Illumina reads, we evaluated the read quality before and after data processing using `FastQC`, with the quality metrics compiled into a single report for concise visualization via `MultiQC`. Contaminants, including rRNA, mtRNA, and other potential contaminants, were removed using `FastQScreen`. This process involved hits against the LSU_Ref and SSU_Ref Silva databases v.138 [48], the zebra finch mitochondrial genome (NCBI Reference Sequence: NC_007897.1), and the `FastQScreen` database of vectors, adapters, and GRCm38 rRNA. For each sample, only minor hits were detected in the rRNA databases, primarily mitochondrial reads from the zebra finch (~5%); approximately 95% of the reads had no-hits (Figure S1). Reads that did not map to these databases (no-hits) were retained for downstream analyses. Adapter sequences and low-quality bases were trimmed using `fastp`. Table 1 presents the preprocessing statistics for each Illumina sample. Following preprocessing, the number of retained high-quality reads ranged from 27,446,223 to 33,201,568 per sample, ensuring robust data for the downstream assembly.

Table 1. Preprocessing step results for Illumina reads.

Steps/Samples	Brain_2	Brain_5	Ovary_2	Testis_5
Raw	32,217,548	29,323,820	33,201,568	28,474,620
FastQScreen	30,396,892	28,010,526	31,965,452	27,981,280
fastp	29,777,552	27,446,223	31,089,662	27,519,646

The initial preprocessing of the PacBio raw data (subreads) followed the `Isoseq3` package for each sample. CCSs were generated using the default minimum number of subreads (default: 3). FL transcripts were identified, and primers were removed using `lima` with `–peek-guess`, while `isoseq3 refine` was used to remove poly-A tails and artificial concatemers to obtain FLNC reads. High-quality FL consensus sequences were obtained using the `isoseq3 cluster2` with the `–singletons` parameter. By default, `isoseq3 cluster2` retains isoforms that are represented by at least two FLNC reads. To capture rare but potentially significant isoforms, the `--singletons` option was employed for this analysis to include these single-read isoforms in the consensus sequences.

To eliminate potential contamination from PacBio reads, `BlobTools2` was employed to retain only vertebrate sequences and other hits for further analysis. Except for the ovary_2 sample, which contained hits from the *Annelia phylum*, all other samples contained only vertebrate sequences and other hits (Figure S2). Additionally, rRNA and mitochondrial fragments were identified and removed by aligning the reads against the NCBI nucleotide database (NT, retrieved on 9 February 2024) using `Blastn`. High-quality, clean FL reads from each sample were corrected using `FLMRC` with default parameters. The error-correction step is provided as an option for users. With the advancements in sequencing technologies, the accuracy of long reads has been gradually increasing. The error rate of Nanopore sequencing has improved from approximately 64% for R7 to approximately 84–95% for R9.4 [49]. In contrast, the PacBio platform, utilizing the CCS approach, achieves a greater than 99% consensus accuracy [50,51]; however, systematic errors can still persist, especially in homopolymeric regions [52,53]. Depending on the accuracy of the sequencing platform utilized, users can proceed with downstream analysis without an additional error-correction step. In our study, we chose to perform the error correction because our analysis revealed that the `BUSCO` scores of the error-corrected clean reads were superior to those of the non-error-corrected clean reads (Figure S3). This finding aligns with the literature suggesting that error correction can still enhance the overall quality of transcriptomic data [51]. Table 2 shows the number of reads obtained at each preprocessing step. To remove redundancy, `CD-HIT-Est` was first employed for each sample (Table 2), and subsequently, `EvidentialGene` was used for the combined sample for the further classification of mRNA reads and predicted protein sequences (Table 3).

 doi:10.20944/preprints202411.0489.v1

Table 2. Preprocessing steps for PacBio reads.

Steps/Samples	Brain_2	Brain_5	Ovary_2	Testis_5
Raw	444,968	717,758	483,419	729,821
CCS	124,615	56,773	198,608	42,832
FL	93,967	15,245	172,050	25,332
FLNC	89,017	15,103	168,336	25,095
Clustered	47,129	10,229	80,405	18,101
Contamination	46,769	9990	80,166	17,993
Error-correction	46,769	9990	80,166	17,993
CD-Hit-Est	23,636	6175	37,838	11,284

Table 3. Descriptive statistics of the zebra finch transcriptome obtained using TrAnnoScope.

Steps/Samples	Transcriptome
Total isoforms	39,984
mean, median, N50 transcripts	3097.7|2794|4108
Total proteins	39,984
mean, median, N50 proteins	398.2|271|597
Full-length proteins (EvidentialGene)	86.7%
Transcriptome completeness	C: 79.1% [S: 38.2%, D: 40.9%], F: 4.7%, M: 16.2%, n: 3354

After `EvidentialGene`, 39,984 transcripts were obtained with an average transcript length of 3097.7 bp, a median of 2794 bp, and an N50 of 4108 bp, indicating the robustness of the assembly and the inclusion of long, high-quality transcripts (Table 3). Among the 39,984 predicted proteins, the average protein length was 398.2 amino acids with a median of 271 amino acids and an N50 of 597 amino acids. Notably, 86.7% of the transcripts were classified as complete proteins, further demonstrating the effectiveness of the pipeline for generating FL sequences. For FL representation analysis, a total of 26,141 transcripts matched the zebra finch protein sequences (GCF_003957565.2) with an e-value threshold of 1×10^{-20}. Of these, 24,560 transcripts presented as nearly FL (>70% coverage) relative to the zebra finch reference protein sequences. Among these, 10,575 transcripts were classified as FL transcripts with 100% coverage (Figure 2A, Table S2).

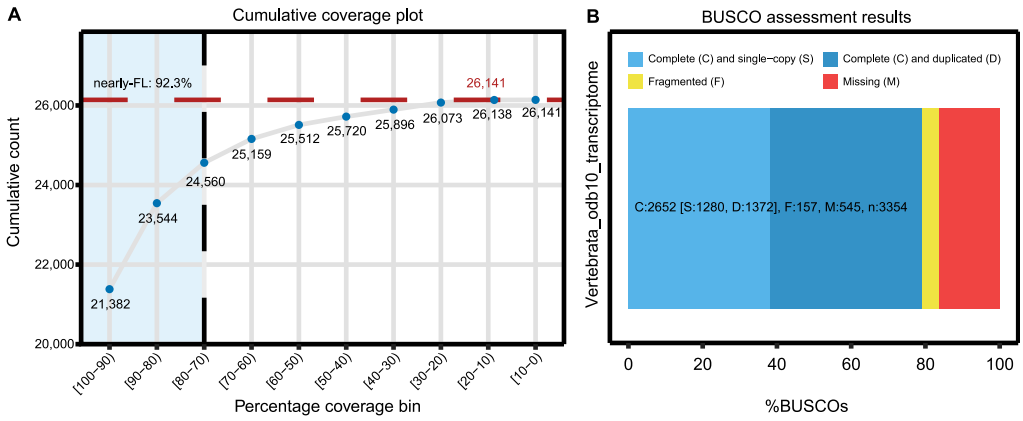

Figure 2. (**A**) Cumulative dot plot for significant hits from BLASTP (e-value= 1×10^{-20}) against the zebra finch proteome. (**B**) BUSCO assessment of the transcriptome against vertebrata_odb10 showing the distribution of complete (C), single-copy (S), duplicated (D), fragmented (F), and missing (M) BUSCOs to evaluate transcriptome completeness and quality.

The `BUSCO` assessment results demonstrated the completeness of the transcriptome based on the presence of BUSCOs from the vertebrate lineage (Figure 2B). The analysis revealed that FL transcripts matched 83.8% of the single-copy orthologs in BUSCOs, comprising 79.1% complete orthologs and

 doi:10.20944/preprints202411.0489.v1

4.7% fragmented orthologs out of a total of 3354 orthologs. Additionally, 16.2% of the orthologs were classified as missing (Table 3). While these BUSCO result indicates a reasonably high level of completeness, it also highlights potential limitations in the transcriptome assembly. The observed BUSCO completeness score can be attributed to the limited sampling of only three specific tissues. BUSCO scores assess the presence of conversed, single-copy orthologs expected in a comprehensive transcriptome. By focusing on a smaller number of tissues only a subset of actively expressed transcripts was captured, which reduces the overall BUSCO score. This result does not indicate poor data quality but reflects the targeted nature of the sampling strategy [54]. To improve the completeness of the transcriptome, future study could incorporate additional tissue samples to capture broader range of gene expression. Despite the limitations, the `BUSCO` results, combined with the descriptive statistics, underscore the ability of the pipeline to accurately obtain transcriptomes from Illumina and PacBio data, providing a reliable foundation for subsequent functional analyses.

The alignment of our transcriptome to current zebra finch genome yielded a high mapping rate of 99.63%, underscoring the accuracy and reliability of the TrAnnoScope pipeline for generating high-quality transcriptomic data. Despite this high mapping rate, a total of 154 transcripts did not map to the genome.

The identification of 154 transcripts that did not map to genome presents an intriguing opportunity for further exploration into the genomic landscape if the zebra finch. Among these unmapped transcripts, a significant portion (136 transcripts) exhibited notable hits in homology searches (see Supplementary File S2). It indicates that these transcripts likely represent real biological data rather than artifacts. Interestingly, 7 of these 136 transcripts aligned with sequences from previous zebra finch genome assemblies that are not present in the current genome version. This suggests that the unmapped transcripts may represent genomic regions that have been lost or altered in the latest assembly. This observation highlights the ongoing refinement of genomic resources and emphasizes the importance of considering multiple assembly versions in genomic data analysis. The rest of the unmapped 136 transcripts aligned with closely related species, particularly those within the passerine bird family. This indicates that these transcripts may possess functional relevance, potentially aligning with genes conserved across closely related passerine species and suggesting a shared heritage that may be critical for understanding evolutionary relationships in this group.

Among the remaining unmapped transcripts lacking homology search results, eight showed hits only for SignalP and TmHMM, indicating that they might encode peptides with specific targeting signals or transmembrane domains. This finding hints at their potential functional roles within cellular processes. However, the remaining ten transcripts, which yielded no significant information, raise questions about their biological significance. These transcripts could represent novel genes, warranting further exploration into their functions. Alternatively, they may be artifacts, emphasizing the need for additional validation.

The functional annotation of the assembled transcriptome was conducted using a comprehensive set of databases via `Trinotate` (e-value 1×10^{-5}), providing valuable insights into the roles and characteristics of the predicted proteins (Table 4). The annotation revealed broad coverage across multiple databases, enhancing confidence in the functional assignments. Out of the 39,984 transcripts, 70.7% (28,274) had significant hits against the UniProt/SwissProt database using `Blastx`, while 62.7% (25,051) were confirmed through `Blastp` searches. Domain-based searches using Pfam identified conserved protein domains in 59.2% (23,689) of transcripts, highlighting their protein-coding potential.

Table 4. Overview of the annotation results.

Database	Hits (%)
UniProt/SwissProt Blastx	28,274 (70.7%)
UniProt/SwissProt Blastp	25,051 (62.7%)
Pfam Domains	23,689 (59.2%)
GO	28,399 (71.0%)
KOG	26,409 (66.0%)

 doi:10.20944/preprints202411.0489.v1

KEGG	26,097 (65.3%)
Transmembrane Domains (TmHMM)	7155 (17.9%)
Signal Peptides (SignalP)	2305 (5.8%)
Non-coding RNAs (Infernal)	169 (0.4%)
Non-redundant protein DB (NR Blastx)	33,049 (82.7%)
Non-redundant protein DB NR Blastp	28,303 (70.8%)
Non-redundant nucleotide DB (NT Blastn)	39,827 (99.6%)

GO terms from SwissProt, Pfam and EggNOG were assigned to 71.0% (28,399) of transcripts, categorizing them into biological processes, molecular functions, and cellular components (Figure 3). Additionally, KOG classifications were found for 66.0% (26,409) of transcripts, offering insights into their evolutionary relationships and potential functional roles (Figure 4). The KEGG pathway analysis annotated 65.3% (26,097) of transcripts, linking them to various metabolic and signaling pathways (Figure 5).

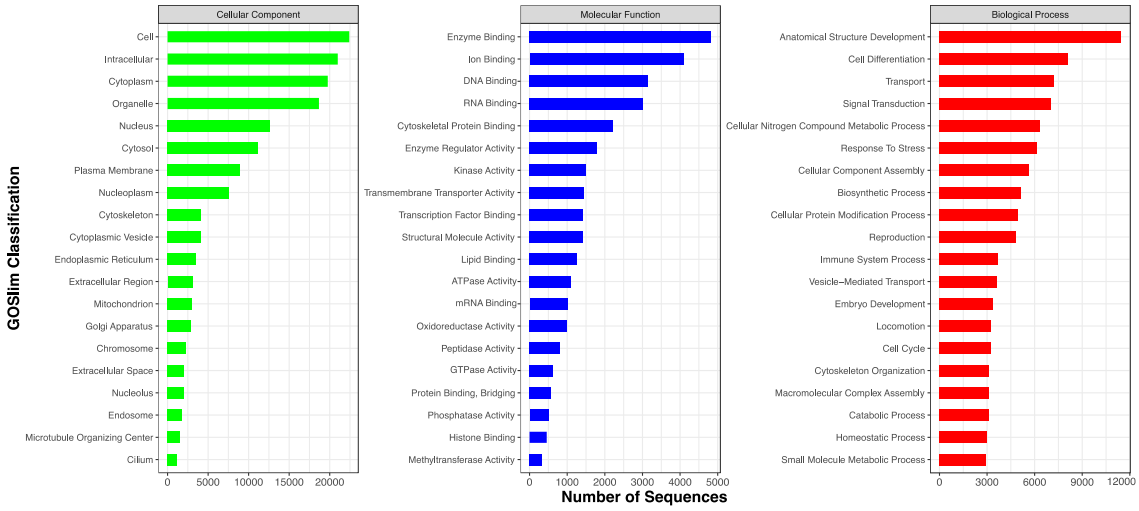

Figure 3. Distribution of top 20 Gene Ontology terms for the cellular components, molecular functions, and biological processes of zebra finch.

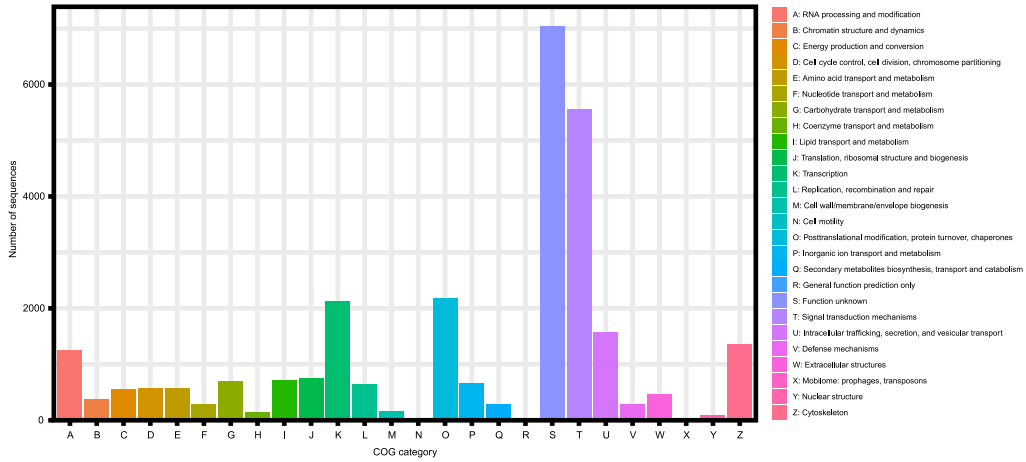

Figure 4. KOG classification of zebra finch.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 November 2024 doi:10.20944/preprints202411.0489.v1

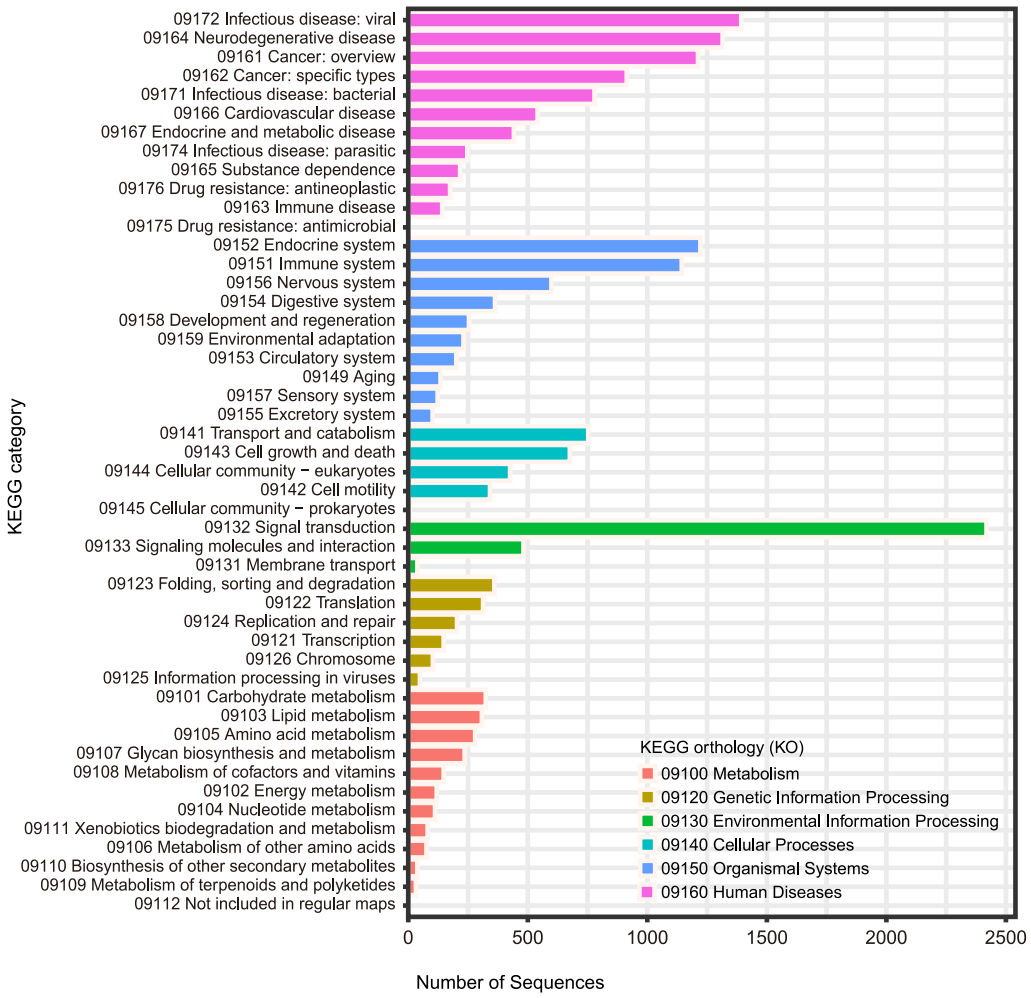

Figure 5. KEGG classification of zebra finch.

Structural and localization predictions identified transmembrane domains in 17.9% (7155) of transcripts, as predicted using TMHMM, and signal peptides in 5.8% (2305), as predicted using SignalP. A small portion of the transcripts (0.4%, 169) were annotated as non-coding RNAs using Infernal.

Further annotation against the NR database using `Blastx` produced hits for 82.7% (33,049) of transcripts, while `Blastp` hits were obtained for 70.8% (28,303). Nearly all transcripts (99.6%, 39,827) matched in the NT database through `Blastn`, indicating the strong conservation of these sequences against the known sequences. These extensive annotation results underscore the robustness of the transcriptome, with most transcripts being functionally annotated across various databases, providing a rich resource for downstream biological analyses.

The species distribution of the transcripts obtained from the TrAnnoScope pipeline underscores the effectiveness of our approach in functional annotation of zebra finch data (Figure 6). The `Blastx` homology search against the NR database revealed that a significant number of hits (17,121) were assigned to the zebra finch, indicating a strong alignment between our assembled transcripts and existing annotations. This result highlights the robustness of the TrAnnoScope pipeline in processing RNA-Seq data and achieving meaningful annotations that are crucial for downstream analyses. Furthermore, the presence of additional hits to closely related species, such as the society finch (*Lonchura striata domestica*), Gouldian finch (*Chloebia gouldiae*), and canary (*Serinus canaria*) can be attributed to high degree of genetic similarity among these species. This observation reinforces the notion that functional conservation is common among closely related species, facilitating the identification of homologous genes and conserved biological functions. Moreover, the identification

 doi:10.20944/preprints202411.0489.v1

of transcripts that align with other birds in the Passeriformes order, such as starlings (*Lamprotornis superbus*), swallows (*Hirundo rustica rustica*), sparrows (*Melospiza melodia maxima, Passer montanus*), and Réunion grey white-eyes (*Zosterops borbonicus*), reflects the shared ancestry within this diverse avian group.

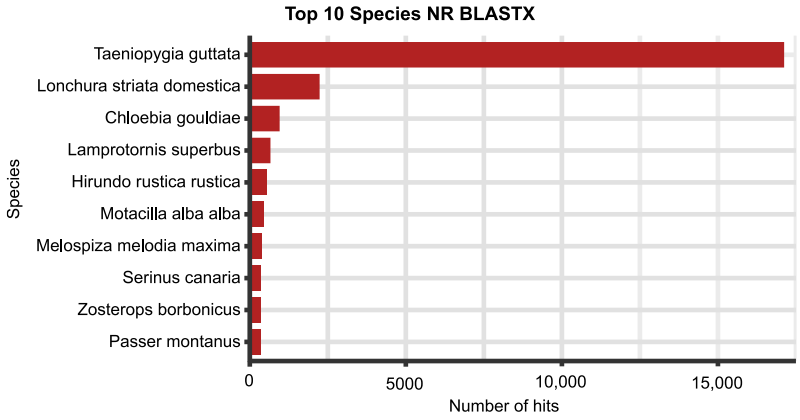

Figure 6. Top 10 species distribution of the transcript sequences of the zebra finch against the NR database.

The successful functional annotation of these transcripts not only validates the efficiency of the TrAnnoScope pipeline but also provides valuable insights into the evolutionary relationships among these avian species. The alignment of our data with that of the zebra finch and its relatives may contribute to a deeper understanding of the genetic basis of traits relevant to adaptation and survival in varying environments. Additionally, these findings can serve as a foundation for future studies aimed at exploring gene function, expression patterns, and evolutionary dynamics within the Passeriformes order and beyond.

While the TrAnnoScope pipeline has demonstrated its effectiveness in processing and annotating transcriptomic data, there are several areas for future improvement. One significant limitation is the current support for long-read preprocessing, which is restricted to PacBio reads. Incorporating preprocessing capabilities for ONT reads would greatly enhance the versatility of the pipeline, allowing users to leverage the strengths of both sequencing platforms for comprehensive transcriptome assembly. Additionally, providing support for differential expression analysis within the pipeline would facilitate more in-depth investigations into gene expression patterns across various conditions and tissues. This enhancement could empower researchers to derive meaningful biological insights from their data, further expanding the utility of TrAnnoScope in the field of transcriptomics. Addressing these gaps will not only improve the overall functionality of the pipeline but also increase its appeal to a broader range of users conducting diverse transcriptomic studies.

4. Conclusions

In this study, we introduced TrAnnoScope, a comprehensive pipeline for transcriptome analysis and annotation that utilizes both short- and long-read data. Applying TrAnnoScope to zebra finch RNA-Seq data demonstrated its capability to generate high-quality transcripts and functional annotations across multiple databases. The pipeline efficiently processes large datasets, from quality control to final annotation, resulting in a transcriptome with significant functional insights.

TrAnnoScope is built with Snakemake, and its modular design allows for easy customization while requiring minimal programming skills, making it accessible to users with varying levels of expertise. Its parallelized steps and user-defined parameters enhance the speed and reliability of transcriptome analysis. Although some manual database preparation is necessary, the pipeline remains a valuable tool for researchers, particularly those working with non-model organisms. Future updates will focus on automating database management and expanding the preprocessing options for other platforms.

 doi:10.20944/preprints202411.0489.v1

Overall, TrAnnoScope is a versatile and efficient tool for transcriptomics, providing a robust platform for transcriptome analysis.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Supplementary File S1: Table S1: Overview of RNA-Seq samples and platforms, Figure S1: FastQScreen mapping results across genomes for Illumina reads, Figure S2: BlobTools contamination assessments for PacBio samples, Figure S3: Comparison of error corrected clean reads vs. non-error corrected clean reads, Table S2: Full-length representation table against zebra finch protein sequences, Supplementary File S2: Annotation and taxonomy information for unmapped hits, Supplementary File S3: Folder structure of the TrAnnoScope pipeline, along with the details of SLURM jobs and runtime for data.

Author Contributions: A.P. developed the pipeline and wrote the manuscript. B.T. and F.P. interpreted the results and edited the manuscript. B.T. supervised and acquired funding for the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by research grants from the Novo Nordisk Foundation (NNF20OC0064467), the Lundbeck Foundation (R324-2019-1625), and the Aarhus University Research Foundation (AUFF-2020-9-17).

Data Availability Statement: All sequencing reads were obtained from the Sequence Read Archive (SRA) database of the NCBI. The accession numbers can be found in the Materials and Methods section, under 'Data Selection'. TrAnnoScope pipeline can be accessed via https://github.com/aysevllpkts/TrAnnoScope.

Acknowledgments: AI-tools were used to refine the grammar and language of the draft. All subsequent edits and the content of the manuscript remain the responsibility of the author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, J.W.; Shrestha, L.; Green, G.; Leier, A.; Marquez-Lago, T.T. The hitchhikers' guide to RNA sequencing and functional analysis. *Brief. Bioinform.* **2023**, *24*, bbac529. https://doi.org/10.1093/bib/bbac529.
2. Deshpande, D.; Chhugani, K.; Chang, Y.; Karlsberg, A.; Loeffler, C.; Zhang, J.; Muszynska, A.; Munteanu, V.; Yang, H.; Rotman, J.; et al. RNA-seq data science: From raw data to effective interpretation. *Front. Genet.* **2023**, *14*, 997383. https://doi.org/10.3389/fgene.2023.997383.
3. Raghavan, V.; Kraft, L.; Mesny, F.; Rigerte, L. A simple guide to de novo transcriptome assembly and annotation. *Brief. Bioinform.* **2022**, *23*, bbab563. https://doi.org/10.1093/bib/bbab563.
4. Esteve-Codina, A. RNA-Seq Data Analysis, Applications and Challenges. In *Comprehensive Analytical Chemistry*; Jaumot, J., Bedia, C., Tauler, R., Eds.; Elsevier: Amsterdam, The Netherlands, **2018**; Volume 82, pp. 71–106.
5. Garg, R.; Jain, M. RNA-Seq for transcriptome analysis in non-model plants. *Methods Mol. Biol.* **2013**, *1069*, 43–58. https://doi.org/10.1007/978-1-62703-613-9_4.
6. Lataretu, M.; Holzer, M. RNAflow: An Effective and Simple RNA-Seq Differential Gene Expression Pipeline Using Nextflow. *Genes* **2020**, *11*, 1487. https://doi.org/10.3390/genes11121487.
7. Zhang, X.; Jonassen, I. RASflow: An RNA-Seq analysis workflow with Snakemake. *BMC Bioinform.* **2020**, *21*, 110. https://doi.org/10.1186/s12859-020-3433-x.
8. Fallon, T.R.; Calounova, T.; Mokrejs, M.; Weng, J.K.; Pluskal, T. transXpress: A Snakemake pipeline for streamlined de novo transcriptome assembly and annotation. *BMC Bioinform.* **2023**, *24*, 133. https://doi.org/10.1186/s12859-023-05254-8.
9. Rivera-Vicens, R.E.; Garcia-Escudero, C.A.; Conci, N.; Eitel, M.; Worheide, G. TransPi-a comprehensive TRanscriptome ANalysiS PIpeline for de novo transcriptome assembly. *Mol. Ecol. Resour.* **2022**, *22*, 2070–2086. https://doi.org/10.1111/1755-0998.13593.
10. Ortiz, R.; Gera, P.; Rivera, C.; Santos, J.C. Pincho: A Modular Approach to High Quality De Novo Transcriptomics. *Genes* **2021**, *12*, 953. https://doi.org/10.3390/genes12070953.
11. FIT: Functional IsoTranscriptomics Analyses. Available online: https://tappas.org/ (accessed on 11 September 2024).
12. Lienhard, M.; van den Beucken, T.; Timmermann, B.; Hochradel, M.; Borno, S.; Caiment, F.; Vingron, M.; Herwig, R. IsoTools: A flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics* **2023**, *39*, btad364. https://doi.org/10.1093/bioinformatics/btad364.

13. Xia, Y.; Jin, Z.; Zhang, C.; Ouyang, L.; Dong, Y.; Li, J.; Guo, L.; Jing, B.; Shi, Y.; Miao, S.; et al. TAGET: A toolkit for analyzing full-length transcripts from long-read sequencing. *Nat. Commun.* **2023**, *14*, 5935. https://doi.org/10.1038/s41467-023-41649-0.
14. Guizard, S.; Miedzinska, K.; Smith, J.; Smith, J.; Kuo, R.I.; Davey, M.; Archibald, A.; Watson, M. nf-core/isoseq: Simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. *Bioinformatics* **2023**, *39*, btad150. https://doi.org/10.1093/bioinformatics/btad150.
15. Kasianova, A.M.; Penin, A.A.; Schelkunov, M.I.; Kasianov, A.S.; Logacheva, M.D.; Klepikova, A.V. Trans2express—De novo transcriptome assembly pipeline optimized for gene expression analysis. *Plant Methods* **2024**, *20*, 128. https://doi.org/10.1186/s13007-024-01255-7.
16. Zhang, W.; Petegrosso, R.; Chang, J.W.; Sun, J.; Yong, J.; Chien, J.; Kuang, R. A large-scale comparative study of isoform expressions measured on four platforms. *BMC Genom.* **2020**, *21*, 272. https://doi.org/10.1186/s12864-020-6643-8.
17. Molder, F.; Jablonski, K.P.; Letcher, B.; Hall, M.B.; Tomkins-Tinch, C.H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S.O.; Kanitz, A.; et al. Sustainable data analysis with Snakemake. *F1000Research* **2021**, *10*, 33. https://doi.org/10.12688/f1000research.29032.2.
18. Babraham Bioinformatics. FastQC. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed on 8 August 2024).
19. Ewels, P.; Magnusson, M.; Lundin, S.; Kaller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.
20. Wingett, S.W.; Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **2018**, *7*, 1338. https://doi.org/10.12688/f1000research.15931.2.
21. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
22. Pacific Biosciences. Iso-Seq—Scalable De Novo Isoform Discovery from Pacbio HiFi Reads. Available online: https://isoseq.how/ (accessed on 8 August 2024).
23. Mortezaei, Z. Computational methods for analyzing RNA-sequencing contaminated samples and its impact on cancer genome studies. *Inform. Med. Unlocked* **2022**, *32*, 101054.
24. Gondane, A.; Itkonen, H.M. Revealing the History and Mystery of RNA-Seq. *Curr. Issues Mol. Biol.* **2023**, *45*, 1860–1874. https://doi.org/10.3390/cimb45030120.
25. Laetsch, D.R.; Blaxter, M.L. BlobTools: Interrogation of genome assemblies. *F1000Research* **2017**, *6*, 1287.
26. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
27. Manni, M.; Berkeley, M.R.; Seppey, M.; Zdobnov, E.M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **2021**, *1*, e323. https://doi.org/10.1002/cpz1.323.
28. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. https://doi.org/10.1186/1471-2105-10-421.
29. Fu, S.; Wang, A.; Au, K.F. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* **2019**, *20*, 26. https://doi.org/10.1186/s13059-018-1605-z.
30. Wang, J.R.; Holt, J.; McMillan, L.; Jones, C.D. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinform.* **2018**, *19*, 50. https://doi.org/10.1186/s12859-018-2051-3.
31. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
32. Gilbert, D. EvidentialGene: tr2aacds, mRNA Transcript Assembly Software. Available online: http://arthropods.eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html (accessed on 29 Oct 2024).
33. De Coster, W.; D'Hert, S.; Schultz, D.T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **2018**, *34*, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
34. Trinity. Counting Full Length Trinity Transcripts. Available online: https://github.com/trinityrnaseq/trinityrnaseq/wiki (accessed on 8 August 2024).

 doi:10.20944/preprints202411.0489.v1

35. Trinotate: Transcriptome Functional Annotation and Analysis. Available online: https://github.com/Trinotate/Trinotate/wiki (accessed on 21 October 2024).
36. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.
37. UniProt, C. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.
38. Teufel, F.; Almagro Armenteros, J.J.; Johansen, A.R.; Gislason, M.H.; Pihl, S.I.; Tsirigos, K.D.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **2022**, *40*, 1023–1025. https://doi.org/10.1038/s41587-021-01156-3.
39. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **2001**, *305*, 567–580. https://doi.org/10.1006/jmbi.2000.4315.
40. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernandez-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.
41. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.
42. Gene Ontology, C.; Aleksander, S.A.; Balhoff, J.; Carbon, S.; Cherry, J.M.; Drabkin, H.J.; Ebert, D.; Feuermann, M.; Gaudet, P.; Harris, N.L.; et al. The Gene Ontology knowledgebase in 2023. *Genetics* **2023**, *224*, iyad031. https://doi.org/10.1093/genetics/iyad031.
43. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.
44. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A.; Koonin, E.V. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. https://doi.org/10.1093/nar/28.1.33.
45. Mello, C.V. The zebra finch, Taeniopygia guttata: An avian model for investigating the neurobiological basis of vocal learning. *Cold Spring Harb. Protoc.* **2014**, *2014*, 1237–1242. https://doi.org/10.1101/pdb.emo084574.
46. Hauber, M.E.; Louder, M.I.; Griffith, S.C. The Natural History of Model Organisms: Neurogenomic insights into the behavioral and vocal development of the zebra finch. *eLife* **2021**, *10*, e61849. https://doi.org/10.7554/eLife.61849.
47. Leinonen, R.; Sugawara, H.; Shumway, M. International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **2011**, *39*, D19–D21. https://doi.org/10.1093/nar/gkq1019.
48. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glockner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. https://doi.org/10.1093/nar/gks1219.
49. Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K.F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **2021**, *39*, 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.
50. Ferrarini, M.; Moretto, M.; Ward, J.A.; Surbanovski, N.; Stevanovic, V.; Giongo, L.; Viola, R.; Cavalieri, D.; Velasco, R.; Cestaro, A.; et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genom.* **2013**, *14*, 670. https://doi.org/10.1186/1471-2164-14-670.
51. Tvedte, E.S.; Gasser, M.; Sparklin, B.C.; Michalski, J.; Hjelmen, C.E.; Johnston, J.S.; Zhao, X.; Bromley, R.; Tallon, L.J.; Sadzewicz, L.; et al. Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. *G3* **2021**, *11*, jkab083. https://doi.org/10.1093/g3journal/jkab083.
52. Sacristan-Horcajada, E.; Gonzalez-de la Fuente, S.; Peiro-Pastor, R.; Carrasco-Ramiro, F.; Amils, R.; Requena, J.M.; Berenguer, J.; Aguado, B. ARAMIS: From systematic errors of NGS long reads to accurate assemblies. *Brief. Bioinform.* **2021**, *22*, bbab170. https://doi.org/10.1093/bib/bbab170.

 doi:10.20944/preprints202411.0489.v1

53. Pourmohammadi, R.; Abouei, J.; Anpalagan, A. Error analysis of the PacBio sequencing CCS reads. *Int. J. Biostat.* **2023**, *19*, 439–453. https://doi.org/10.1515/ijb-2021-0091.
54. Waterhouse, R.M.; Seppey, M.; Simao, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* **2018**, *35*, 543-548, doi:10.1093/molbev/msx319.