Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid assembly of the African savanna elephant whole blood transcriptome using short- and long-read data

Stacey Caitlin Engel
Stellenbosch University
Tanya Jane Kerr
Stellenbosch University
Gian van der Spuy
Stellenbosch University
Peter Erik Buss
South African National Parks
Léanie Kleynhans
Mater Research
Michele Ann Miller
Stellenbosch University
Tracey Jooste
t.jooste@sun.ac.za

Stellenbosch University

Research Article

Keywords: African savanna elephant, functional annotation, hybrid assembly, MGI Tech, Oxford Nanopore, RNA-Seq, transcriptome, whole blood

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11534641/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid assembly of the African savanna elephant whole blood transcriptome using short- and long-read data.

Authors: *Stacey Caitlin Engel[a,b], Tanya Jane Kerr[a,b], Gian van der Spuy[a,b], Peter Erik Buss[c], Léanie Kleynhans[a,b,d,e], Michele Ann Miller[a,b,+] and Tracey Jooste[a,b, f,+*].*

Author affiliations:

[a]South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[b]Biomedical Research Institute, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

[c]South African National Parks, Veterinary Wildlife Services, Kruger National Park, Skukuza, South Africa.

[d]Mater Research Institute - The University of Queensland, Translational Research Institute, Brisbane, Australia.

[e]Australian Infectious Diseases Research Centre, The University of Queensland, Brisbane, Australia.

[f]Crick Africa Network, The Francis Crick Institute, London, United Kingdom

[+] Co-senior authors

*Corresponding author: Tracey Jooste tjooste@sun.ac.za

Abstract

Background: Understanding of immune responses and biomarker discovery in African savanna elephants (*Loxodonta africana*) has been limited due to the paucity of transcriptomic resources. Therefore, this study aimed to generate a high-quality whole blood transcriptome for this species using both short- and long-read sequencing data and develop a workflow for transcriptome assembly in non-model species. To determine the highest quality assembly, multiple transcriptome assemblies were generated and evaluated using combinations of short- and long-read sequencing data, together with reference-guided (*StringTie, StringTie2* and *StringTie2* annotation-guided) and *de novo* (*rnaSPAdes*, *Trinity* and *RNA-Bloom2*) assembly approaches.

Results: Paired-end short-read next-generation sequencing data were generated using the MGI platform, while long-read sequencing data were produced using the third-generation Oxford Nanopore Technologies platform. From the assembly tools compared in this study, *StringTie2* with annotation-guided assembly produced the highest-quality transcriptome. This assembly exhibited high transcript continuity, low levels of unannotated transcripts, extensive functional annotation, and a BUSCO completeness score of 98.4%. Functional characterisation represented diverse biological processes, molecular functions, cellular components, and pathways relevant to immune and cellular regulation.

Conclusion: These findings support the use of this whole blood transcriptome as a reference resource for future studies related to disease in African savanna elephants. The transcriptome and associated annotations provide a foundation for future gene expression, functional genomics, and genome annotation research in this endangered species.

Keywords

African savanna elephant, functional annotation, hybrid assembly, MGI Tech, Oxford Nanopore, RNA-Seq, transcriptome, whole blood

Background

Infectious diseases, including tuberculosis (TB) and elephant endotheliotropic herpesvirus (EEHV) haemorrhagic disease (HD), represent significant threats to endangered African savannah elephant (*Loxodonta africana*) populations [1–3]. Despite the conservation importance of these diseases, little is known about the immune responses of African savanna elephants to infection.

Transcriptomic studies have substantially advanced our understanding of host immune responses in both human and veterinary medicine and facilitated the discovery of novel diagnostic biomarkers [4, 5]. Transcriptomic profiling identifies patterns of gene expression and provides insight into the molecular mechanisms underlying biological responses. Advances in both short-read and, more recently, long-read sequencing technologies have expanded opportunities for transcriptome assembly and characterisation [6–8]. Numerous guidelines are available to support transcriptome assembly and annotation, including recommendations for selecting appropriate analytical tools and workflows [9–12]. Transcriptome assembly strategies generally fall into two categories: reference-guided and *de novo* approaches. Reference-guided assemblers require a reference genome from the target species or a closely related organism, whereas *de novo* assemblers reconstruct transcripts without prior genomic information [11].

Reference guided assemblers, such as *StringTie,* reconstruct transcripts by clustering aligned sequencing reads, constructing splice graphs to identify transcript structures, and estimating transcript abundance using maximum flow algorithms [13]. Building upon the original *StringTie* framework, *StringTie2* provides improved handling of error-prone long-reads while reducing computational runtime and memory usage [14]. Reference guided approaches have been successfully applied in a range of species, including mouse-eared bats (*Myotis myotis*) [15] and the avian malaria parasite *Plasmodium relictum* [16].

In contrast, *de novo* assembly methods reconstruct transcriptomes directly from sequencing reads. The *rnaSPAdes* assembler uses de Bruijn graph-based algorithms for transcript reconstruction [17]. When both short- and long-read data are available, *rnaSPAdes* can employ the *hybridSPAdes* module, which uses short-reads to construct the assembly graph and long-reads to improve transcript resolution through alignment [18]. Similarly, *RNA-Bloom2* also uses a de Bruijn graph framework and initially applies short-reads to correct errors in long-read sequences before assembly [19]. Another widely used *de novo* assembler, *Trinity,* integrates three software modules - *Inchworm*, *Chrysalis* and *Butterfly* to reconstruct transcript sequences [20]. *De novo* assembly approaches have been successfully used to characterise transcriptomes in diverse species, including the fat-tailed dunnart (*Sminthopsis crassicaudata*) [21], fire salamander (*Salamandra salamandra*) [22], semi-wild gayal (*Bos frontalis*) [23], and spiny mouse (*Acomys cahirinus*) [24].

The quality of a transcriptome assembly is strongly influenced by the sequencing data used. Short-read next-generation sequencing (NGS) is generally more cost-effective and produces highly accurate reads [10]; however, its limited read length often prevents sequencing across multiple splice junctions, making the reconstruction of full-

length transcripts challenging [9, 10]. Conversely, long-read third-generation sequencing (TGS) can span complete transcript isoforms, reducing assembly complexity, improving transcript contiguity, and decreasing chimeric assemblies [9]. However, long-read platforms typically exhibit higher sequencing error rates than short-read sequencing technologies [10]. To leverage the strengths of both approaches, hybrid transcriptome assembly strategies that combine short- and long-read data have become increasingly popular [10]. Hybrid assemblies have been successfully generated for several species, including poinsettia (*Euphorbia pulcherrima*) [25], cultivated cardoon *(Cynara cardunculus var. altilis* (DC)) [6] and semi-wild gayal (*Bos frontalis*) [23].

A publicly available, high-quality African savanna elephant transcriptome would provide a valuable resource for investigating immune responses and disease pathogenesis in this species. However, the only published transcriptomic dataset currently available for African savanna elephants was generated from placental tissue [26], limiting its utility for studies of systemic immune responses and infectious diseases like TB and EEHV HD. In contrast, whole blood transcriptomes capture the dynamic activity of circulating immune cells and provide a comprehensive snapshot of immune and regulatory processes occurring throughout the body. Therefore, the aim of this study was to generate a high-quality whole blood transcriptome for the African savanna elephant using both short- and long-read sequencing data. Given the recent improvement of the African savanna elephant reference genome [27], both reference-guided and *de novo* assembly strategies were evaluated to produce a functionally annotated reference transcriptome. Furthermore, given that assessing transcriptome assembly quality remains challenging, particularly for *de novo* assemblies [28], a

comprehensive suite of biological and statistical metrics was used to evaluate and compare assembly performance.

Methods

RNA samples

Whole blood RNA, used for transcriptome assembly, was obtained from four free-ranging adult (> 25 years old), bull African savanna elephants in the Kruger National Park (KNP), South Africa, as part of a registered disease surveillance project. Blood collection, total RNA extraction, RNA enrichment using depletion-based methods, and RNA quality assessment prior to library preparation were performed as previously described [29]. An overview of the experimental workflow is summarized in Figure 1. The enriched RNA of each elephant was used to prepare two sequencing libraries: one for short-read NGS and one for long-read TGS.

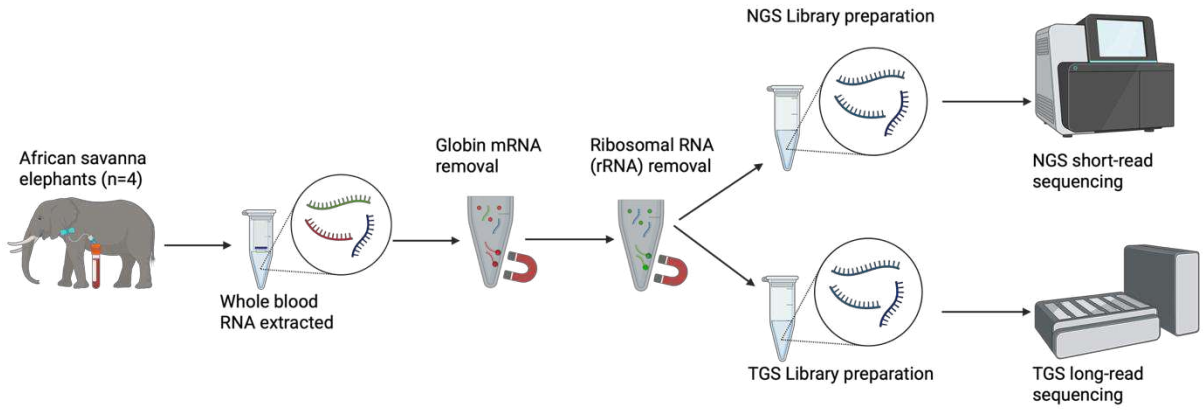


Figure 1. *Experimental workflow for RNA extraction and processing in preparation for NGS and TGS sequencing. Whole blood from African savanna elephants were collected in PAXgene blood tubes. The extracted total RNA was subjected to globin messenger RNA (mRNA) and ribosomal RNA (rRNA) depletion prior to library preparation and sequencing. Sequencing was performed using both short-read, next-generation sequencing (NGS, i.e. MGI*

Tech) and long-read, third-generation sequencing (TGS, i.e. Oxford Nanopore Technology). Created in https://BioRender.com.

Library preparation and sequencing

Short-read, next-generation sequencing (NGS)

For short-read NGS, aliquots of approximately 230 ng of enriched RNA were used for library preparation using the MGIEasy RNA Library Prep Set (MGI Tech, Shenzhen, 518000, China) according to Manual Version A4. Briefly, RNA was fragmented at 87°C for 6 minutes, reverse-transcribed into cDNA, and subjected to a series of enzymatic reactions, including end repair and A-tailing, adapter ligation, and PCR amplification. Following fragment size assessment and library quantification, PCR products from the four elephant whole blood samples were pooled in equimolar concentrations and converted into a single-stranded circular DNA library. The resulting library (40 fmol) was sequenced across four lanes on the DNBSEQ-G400RS platform (MGI Tech) at the South African Medical Research Council (SAMRC) Genomics Platform (Tygerberg, South Africa), generating 100-bp paired-end (PE100) reads.

Long-read, third generation sequencing (TGS)

For long-read sequencing, approximately 50 ng of enriched RNA was used for library preparation with the cDNA-PCR Sequencing V14 Barcoding Kit (SQK-PCB114-24, Oxford Nanopore Technology (ONT), Oxford, OX4 4DQ, UK) and recommended third-party reagents. Libraries were prepared using an in-house modified version of the manufacturer's protocol (Version PCB_9201_v114_revE_11Dec2024). Two modifications were introduced. First, during full-length transcript amplification (step 4), four separate PCR reactions were performed per sample to generate sufficient PCR

product yield. Second, PCR cycling conditions were modified to include an extension step at 65°C for 45 seconds, followed by 16 cycles of denaturation, annealing and extension.

Individual elephant samples were barcoded, and 10-20 fmol of amplified library from each sample was pooled to generate a single sequencing library. The pooled library, containing approximately 60 fmol DNA (mean fragment size ~320 bp), was sequenced on a PromethION 2 Solo instrument at the SAMRC Genomics Platform using a PromethION R10.4.1 flow cell (FLO-PRO114M, ONT). Reads of all lengths were retained, and FASTQ files were generated at 10-minute intervals throughout the sequencing run.

Raw sequence data processing

All sequence data processing, transcriptome assembly, and quality-control analyses were performed using the computational infrastructure of the Centre for High Performance Computing (CHPC), South Africa. All software packages and Bash scripts used in the analyses are listed in Additional File 1: Table S1. An overview of the bioinformatics workflow is provided in Additional File 2: Figure S1.

Short-read, next generation sequencing (NGS)

The quality of the raw sequencing data was assessed using *FastQC* (v 0.12.1) [30], and results were summarised using *MultiQC* (v 1.23) [31]. Reads were subsequently trimmed using *Trimmomatic* (v 0.36) [32] to remove adapter sequences, bases with a Phred quality score below Q20, and reads shorter than 20 bp. Quality of trimmed reads were re-evaluated to confirm successful trimming.

Processed reads were mapped to the African savanna elephant reference genome (mLoxAfr1.hap2; GCF_030014295.1) using *STAR* (v 2.7.7a) [33]. The resulting BAM

files were merged and sorted using *SAMtools* (v 1.20) [34]. Read and alignment statistics, including alignment error rates, were generated using *QualiMap* (v 2.2.1) [35].

Long-read, third generation sequencing (TGS)

Base calling and demultiplexing were initially performed during sequencing using the High Accuracy (HAC, v 4.3.0, 400bps) model in *MinKNOW* software (v 24.11.8). Subsequently, base calling was repeated using the Super Accuracy (SUP, v 4.3.0) model.

FASTQ files generated for each barcode were concatenated into a single FASTQ file for individual elephants using *SAMtools*. Read quality was assessed using *FastQC* (v 0.12.1) and *MultiQC* (v 1.23). The ONT adapter sequences were removed using *Porechop* (v0.2.4) [36], and reads shorter than 20 bp were discarded using *NanoFilt* (v2.8.0) [37]. The quality of the filtered reads was subsequently assessed using *NanoPlot* (v 1.43) [38], *FastQC* (v 0.12.1), and *MultiQC* (v 1.23).

Trimmed reads were aligned to the African savanna elephant genome (mLoxAfr1.hap2; GCF_030014295.1) using *Minimap2* (v 2.24) [39]. The resulting BAM files were merged and sorted using *SAMtools*. Read and alignment statistics, including error rates were calculated using *QualiMap* (v 2.2.1).

Transcriptome assembly

Reference-guided assemblies

StringTie (v2.2.1) [13] and *StringTie2* (v 2.2.0) [14] were used to generate reference-guided transcriptome assemblies (Additional File 3: Figure S2). Merged and sorted BAM files from both the short- and long-read datasets were provided as input to *StringTie* using the '-mix' parameter. In addition, *StringTie2* was used to generate

separate transcriptome assemblies from the short-read and long-read alignments using the '-L' parameter. These assemblies were subsequently merged using the '--merge' parameter. An annotation-guided assembly was generated using the same workflow, with the exception of transcript merging being guided by the reference annotation file using the '-G' parameter in *StringTie2*.

The resulting files, generated using the *StringTie*, *StringTie2* and *StringTie2* annotation-guided approaches, were converted to a FASTA format using *GffRead* [40]. Transcript sequences were extracted using the '-w' parameter, while the African elephant reference genome was provided using the '-g' parameter. The resulting FASTA files were used downstream for quality assessment.

De novo assemblies

To generate *de novo* transcriptome assemblies, *rnaSPAdes* (v 3.15.3) [17], *RNA-Bloom2* (v 2.0.1) [19], and *Trinity* (v 2.15.0) [41] were used (Additional File 3: Figure S2). These assemblers use trimmed sequencing reads directly as input, without reference-genome alignment.

For *rnaSPAdes*, assemblies generated using the default k-mer sizes of 29 and 43, automatically selected according to the input data, were compared, and the assembly generated with k-mer 29 was selected for subsequent analyses. Paired-end short-read (NGS) data generated on the MGI platform were first subjected to error correction using the *rnaSPAdes* '--only-error-correction' parameter. Corrected short-read (NGS) and long-read (TGS) datasets were specified in a YAML input file and assembled using the '-rna' mode (*rnaSPAdes* hybrid). In this hybrid assembly, long reads were used to improve assembly contiguity through gap closure and repeat resolution. To evaluate the contribution of long-read TGS data to assembly performance, an additional

assembly was generated using only the error-corrected short-reads as input (*rnaSPAdes* short-read). The resulting transcripts.fasta files were used for downstream quality assessment.

RNA-Bloom2 (v 2.0.1) [19] was also used to generate a *de novo* transcriptome assembly from the short-read and error-corrected long-read datasets. *RNA-Bloom2* was run with its required dependencies, including *Java* (v 17.0.2), *ntCard* [42], *Minimap2* (v 2.30) and *Racon* (v 1.5.0) [43]. Trimmed and filtered TGS (ONT) long-reads were provided using the '-long' parameter, whereas paired-end short-reads were provided using the '-left' and 'right' parameters. Default long-read assembly settings were used, including a k-mer size of 25, a minimum k-mer coverage threshold of 2, and a minimum transcript length of 200 bp. The resulting transcripts.fasta file was used for assembly quality assessment.

Finally, *Trinity* (v 2.15.0) was used to generate a *de novo* transcriptome assembly using the short-read dataset only. Forward and reverse FASTQ files were provided as input using the '--seqType fq' parameter. The resulting transcriptome FASTA file was used for downstream quality assessment.

Assembly quality control

The assembled transcriptomes were evaluated using multiple quality control metrics, as summarized in Additional File 4: Figure S3. General assembly statistics were generated using *rnaQUAST* (v 2.3.1) [44]. For each assembly, the transcriptome FASTA file was provided as input using the '--transcripts' parameter, together with the African elephant reference genome (--reference) and corresponding GTF annotation file (--gtf). The resulting output was used to generate assembly quality metrics.

Transcriptome completeness was assessed using *BUSCO* (Benchmarking Universal Single-Copy Orthologs) (v 5.8.0) [45–47], with *hmmsearch* (v 3.4) and *metaeuk* (v 7.bba0d80) as dependencies. Completeness was evaluated against the mammalia_odb12 lineage datasets using the '-m transcriptome' mode. All data visualisations were performed using the ggplot2 package (v4.0.3) [48] in R [49] (v 4.5.2) within in the RStudio environment (v 2026.1.0.392) [50].

To assess read representation, the processed short- and long-read datasets were aligned back to each assembled transcriptome using *Minimap2* (v 2.30). Long reads were aligned using the '-ax splice' parameter, whereas the paired-end short reads were aligned using the '-ax splice:sr' parameter. Alignment statistics were subsequently extracted from the resulting mapping files.

Predicted protein sequences and functional annotations were generated for each assembled transcriptome. Transcripts with read lengths <200 bp were removed from downstream analyses. Thereafter, long open reading frames (ORFs) were first identified using *TransDecoder* (v 5.2.0) (Haas, BJ. https://github.com/TransDecoder/TransDecoder) with the 'TransDecoder.LongOrfs' module. Putative coding sequences were then predicted using 'TransDecoder.Predict'. Predicted protein sequences and their corresponding lengths were extracted from the resulting .pep files using SeqKit [51] (v 0.15.0). Functional annotations of the predicted proteins were subsequently performed using eggNOG-mapper (v 2.1.13) [52], with the '--itype protein' parameter. Annotations were assigned based on eggNOG orthology relationships [53] and DIAMOND sequence similarity searches [54]. Functional redundancy within each assembled transcriptome was assessed using CD-HIT (v 4.7) at a sequence identity threshold of 95% ('-c 0.95') and a word size of 5 ('-n 5') [55–59].

Ontology annotation of assembled transcriptome

The eggnog.emapper.annotations file was used to evaluate the Gene Ontology (GO) terms and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway annotations for each assembled transcriptome. Annotations were filtered to retain only those with an E-value $\geq 1 \times 10^{-10}$ and a score ≥ 100. The GO term analyses were performed in R (v 4.5.2) [49] using the GO.db (v 3.22.0) package, whereas KEGG pathways analyses were conducted using the KEGGREST package (v 1.50.0). The GO annotations that could not be assigned to a specific ontology category were excluded from further analysis.

To identify potentially novel genes and transcripts, each assembled transcriptome was compared with the African elephant reference genome annotation (mLoxAfr1.hap2; GCF_030014295.1) using gffcompare (v 0.12.9) [40]. The "-R" parameter was used to exclude reference transcripts that did not overlap with transcripts in the assembled transcriptome.

The novel transcripts were further investigated using *BLASTn* [60] (v 2.16.0+) and *BLASTx* [61] (v 2.16.0+) to assess its similarity to other known nucleotide sequences and proteins respectively. *BLASTn* was performed using the nucleotide collection (nt) database and *BLASTx* using the non-redundant UniprotKB/SwissProt sequence database. Both tools were run using the default parameters and an E-value cut off of 1×10^{-5}.

Results

Sequence data quality control and pre-processing

The short-read sequencing generated a total of 1.31 billion paired-end reads (100 bp) across the four African savanna elephant samples. Raw read quality was exceptionally high, with Phred quality scores ranging from Q34 to Q38 [29]. Long-read sequencing generated 41.71 million reads following high-accuracy base calling. Read lengths ranged from 59 to 633,337 bp, and *FastQC* analysis indicated a peak Phred score of approximately Q27. Re-base calling using the Super Accuracy model improved read quality, increasing peak Phred scores to Q29 - Q31 (Additional File 5: Figure S4 A and B). *NanoPlot* analysis further showed that quality scores ranged from Q6 to Q20, with more than 88.9% of reads exceeding Q15 (Additional File 6: Table S2).

Following adapter removal and quality filtering, 1.22 billion paired-end short reads and 37.1 million long reads were retained for downstream analyses. The processed short-read dataset had read lengths ranging from 20 to 90 bp (mean = 89.4 bp), Phred quality scores of Q34 - Q38, and GC contents of 50 - 53% (Additional File 6: Table S2). Processed long-read data on the other hand ranged from of 20 to 633,226 bp in length, Phred quality scores of Q34 – Q36 and GC contents of 44 - 45% (Additional File 5: Figure S4 C and D; Additional File 6: Table S2). Consistent with the characteristics of TGS (ONT) sequencing, per-base quality scores showed greater variability in the long-read dataset (Q2.59 - Q35) compared to the short-read NGS (MGI) dataset.

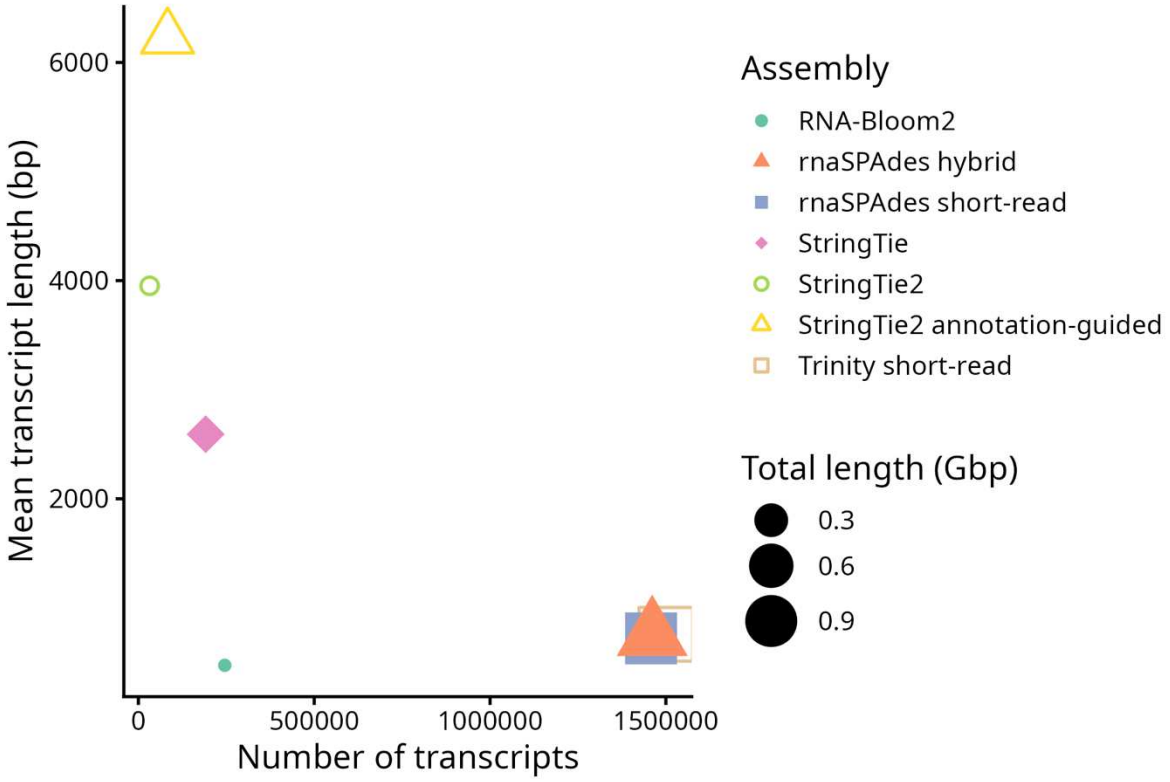
Alignment of the processed datasets to the African elephant reference genome using *STAR* and *Minimap2* resulted in unique mapping rates of 86.7 to 90.4% for short reads and from 86.4 to 88.6% for long reads (Additional File 6: Table S2). *QualiMap* analysis

indicated post-processing alignment error rates of less than 0.3% across all datasets (Additional File 6: Table S2), confirming the high quality of the sequencing data for downstream transcriptome assembly.

Transcriptome assembly basic quality

To identify the optimal transcriptome assembly, assembly statistics generated using *rnaQUAST* were compared across assemblies, including transcriptome length, transcript number, duplication ratio, transcript continuity metrics, isoform coverage, and the proportion of unannotated transcripts (Figure 2, Table 1). The African savanna elephant reference genome and annotation (mLoxAfr1.hap2; GCF_030014295.1) comprising 32,081 genes and 755,401 exons, were used as the reference for all *rnaQUAST* analyses.

Trinity generated the largest transcriptome assembly (1.14 Gbp), with the greatest number of transcripts (Figure 2, Table 1). Overall, the *de novo* assemblers, *Trinity* and *rnaSPAdes,* produced substantially larger assemblies with more transcripts than the reference-guided assemblers, although their mean transcript lengths were shorter (Figure 2). In contrast, *RNA-Bloom2* generated a comparatively compact assembly, with fewer transcripts and shorter mean transcript lengths than the other *de novo* assemblers (Figure 2).



344

345 ***Figure 2.*** *Mean transcript length (base pairs; bp) and number of transcripts for African*
346 *savanna elephant whole blood transcriptome assemblies as determined by rnaQUAST (v*
347 *2.3.1). Size of each point corresponds to the total assembly length (Gbp) of each assembly.*

348 Assemblies with larger transcriptome sizes and higher transcript numbers generally
349 exhibited higher duplication ratios (Table 1). *Trinity* and *rnaSPAdes* showed higher
350 duplication ratios than the reference-guided assemblies, whereas *StringTie,*
351 *StringTie2,* and the *StringTie2* annotation-guided assembly exhibited the lowest
352 duplication ratio. Although *RNA-Bloom2* is a *de novo* assembler, it produced a smaller
353 transcriptome (with fewer and shorter transcripts) and consequently displayed a lower
354 duplication ratio compared to other *de novo* approaches.

355 Transcript continuity was assessed using the N50 statistic and the proportion of
356 transcripts exceeding 1 kb. *StringTie2* produced the highest N50 value (22,680 bp,
357 Table 1), while the longest transcript (738,538 bp, Table 1) was identified in the

StringTie2 annotation-guided assembly. *StringTie2* also generated one of the smallest assemblies, containing only 32,034 transcripts (Table 1). Additionally, the *StringTie2* annotation-guided assembly contained the highest proportion of transcripts greater than 1 kb in length (80.3%, Table 1). The remaining reference-guided assemblies also contained a higher proportion of transcripts exceeding 1 kb (47 - 47.6%) than the *de novo* assemblies (5.6 - 19.7%; Table 1), consistent with greater transcript continuity in the reference-guided assemblies.

Table 1. Basic quality statistics of African savanna elephant whole blood transcriptomes assembled using different tools.

Assembly tool	Total length (Mbp)	Longest transcript (bp)	Duplication ratio	N50 (bp)	Transcripts over 1 kb (%)	Primary mapping rate (%)		Unannotated (%)
						NGS	TGS	
StringTie	495.8	122 308	1.4	7 592	47	68.7- 72.6	78.0-80.6	61.9
StringTie2	126.6	368 759	1.2	**22 680**	47.6	51.8- 57.6	80.5-83.0	23.2
StringTie2 annotation-guided	516.2	**738 538**	**1.1**	13 206	**80.3**	65.5 - 74.2	81.4- 83.9	**11.7**
Trinity short-read	1 136.6	49 817	2.3	1 174	19.7	**95.4- 95.6**	85.3- 87.7	69.8
rnaSPAdes short-read	1 045.0	51 060	2.2	1 583	18.1	93.8- 94.5	85.4- 87.8	73.4
rnaSPAdes hybrid	1 090.7	50 969	2.4	1 704	18.6	92.8- 94.5	85.4- 87.8	73.2
RNA-Bloom2	116.3	4 797	1.6	535	5.6	48.2- 52.7	**88.3- 90.7**	54.6

Mbp: million base pair, bp: base pair, TGS: Third Generation Sequencing (Oxford Nanopore Technology), NGS: Next Generation Sequencing (MGI), N50: represents the length of the shortest contig in the set of longest sequences that collectively cover at least 50% of the total assembled genome length. Values in bold indicate the best-performing assembly for each metric. Statistics generated by rnaQUAST (v 2.3.1).

Read representation was assessed by mapping the processed sequencing datasets back to each assembled transcriptome (relative to the African savanna elephant reference genome mLoxAfr1.hap2; GCF_030014295.1) using *Minimap2* (v2.30). Prior to transcriptome assembly, 86.7-90.4% of short-reads and 86.4-88.6% of long reads mapped to the African savanna elephant reference genome. Short-reads generally exhibited higher mapping efficiencies to the *de novo* transcriptome assemblies than to the reference genome, with the exception of *RNA-Bloom2* (Table 1). In contrast, long-read mapping rates were lower for the reference-guided assemblies, whereas mapping efficiencies for the *de novo* assemblies were comparable to those observed against the reference genome. Among the *de novo* assemblers, *RNA-Bloom2* achieved the highest long-read mapping rate (Table 1).

The proportion of unannotated transcripts varied substantially among assemblies. The *StringTie2* annotation-guided assembly contained the lowest proportion of unannotated transcripts (11.7%; Table 1), whereas *rnaSPAdes* exhibited the highest proportion (Table 1). The *rnaSPAdes* assemblies were also characterised by high transcript numbers, elevated duplication ratios, and shorter mean transcript lengths relative to other assemblies (Table 1).

Assembly completeness was evaluated using the mean isoform, mean exon and database (reference genome) coverage metrics. The *StringTie2* annotation-guided assembly achieved the highest mean isoform coverage (1.0), exon (0.1) and database coverage (1.0; Table 2), indicating near-complete representation of the annotated reference file. Most other assemblies exhibited similar isoform coverage values (0.5), except *RNA-Bloom2*, which had low database coverage (Table 2).

The ability of each assembly to reconstruct annotated isoforms was further assessed using the isoform assembly metric. The *StringTie2* annotation-guided assembly achieved the highest score (1.0; Table 2), indicating near-complete reconstruction of annotated isoforms. The remaining reference-guided assemblies, *StringTie* (0.5) and *StringTie2* (0.5), also outperformed the *de novo* assemblers which achieved scores ranging from 0.3 to 0.4 (Table 2). These results suggest that approximately 50% of the annotated isoforms were reconstructed by the reference-guided assemblies, while the *de novo* assemblers reconstructed a smaller proportion of the annotated isoforms.

Table 2. Transcriptome completeness and reconstruction metrics for the assembled African savanna elephant whole blood transcriptomes.

Tool	**Mean isoform coverage**	**Mean exon coverage**	**Database coverage**	**Isoform assembly**
StringTie	0.5	0.9	0.2	0.5
StringTie2	0.5	0.9	0.1	0.5
StringTie2 annotation-guided	**1.0**	**1.0**	**1.0**	**1.0**
Trinity short-read	0.5	0.9	0.3	0.4
rnaSPADES short-read	0.5	0.9	0.3	0.4
rnaSPADES hybrid	0.5	0.9	0.3	0.4

RNA-Bloom2	0.3	0.8	0.1	0.3

Metrics generated by rnaQUAST (v 2.3.1) using the publicly available African savanna elephant reference genome and annotation file. Values in bold indicate the best-performing assembly for each metric.

To evaluate the contribution of long-read sequencing to assembly performance, the *rnaSPAdes* hybrid assembly was compared with an assembly generated using short-read data alone. Inclusion of the long-read data had little effect on overall assembly size or transcript counts (Figure 2, Table 2). However, the proportion of transcripts exceeding 1 kb increased from 264,522 (18.1%) in the short-read assembly to 270,849 (18.5%) in the hybrid assembly (Table 1). The *rnaSPAdes* hybrid assembly also exhibited a 6.6% increase in 95%-assembled genes and in the proportion of 95%-assembled isoforms relative to the *rnaSPAdes* short-read only assembly. Together, these results indicate that incorporation of long-read sequencing data improved assembly completeness, despite having only a modest effect on the overall assembly statistics.

Transcriptome completeness

To further evaluate assembly quality, transcriptome completeness was assessed using *BUSCO,* which identifies conserved universal single-copy orthologs expected to be present in mammalian transcriptomes. The *StringTie2* annotation-guided assembly achieved the highest *BUSCO* completeness score (98.4%; Figure 3), whereas *RNA-Bloom2* showed the lowest completeness score (13.1%; Figure 3). A substantial proportion of the complete *BUSCOs* identified in the *StringTie2* annotation-guided assembly were classified as duplicated (57.3%). The *de novo* assemblies, generated by *Trinity* and *rnaSPAdes*, also contained relatively high proportions of complete

BUSCOs but exhibited higher levels of *BUSCO* duplications, compared to the reference-guided assemblies (Figure 2).

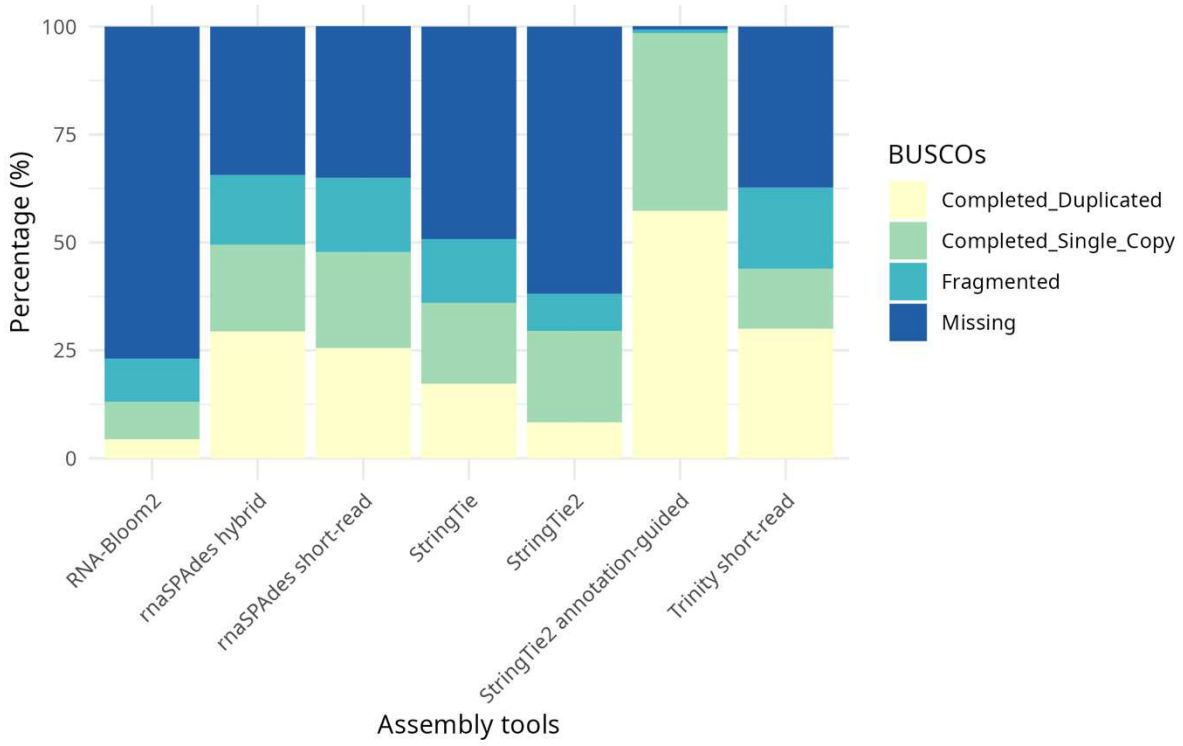


Figure 3: *BUSCO completeness scores (%) of African savanna elephant whole blood transcriptome assemblies generated using different assembly tools as measured by BUSCO (v5.8.0). The mammalia_odb12 lineage dataset was used to assess each of the assembled transcriptomes and determine the transcriptome completeness based on the proportions of completed and duplicated, completed and single copy, fragmented, and missing BUSCO genes present in each assembly.*

The *Trinity* assembly contained the highest proportion of fragmented *BUSCOs* (18.8%, Figure 3), whereas the *StringTie2* annotation-guided assembly contained the lowest proportion (0.8%, Figure 3). The *RNA-Bloom2* assembly exhibited the highest proportion of missing BUSCOs (76.9%, Figure 3). Collectively, these results indicate

that the *StringTie2* annotation-guided assembly provided the most complete representation of conserved mammalian orthologs among the evaluated assemblies.

Functional assessment of transcriptomes

To further assess transcriptome completeness and coding potential, ORFs were predicted using *TransDecoder* (Table 3). Reference-guided assemblies contained a greater proportion of transcripts with predicted ORFs compared to *de novo* assemblies (Table 3). The *StringTie2* annotation-guided assembly had the highest proportion of transcripts containing at least one predicted ORF (83.3%), whereas the *rnaSPAdes* assemblies had the lowest proportion. These differences suggest variation among assemblies in the representation of complete protein-coding transcripts.

Table 3. Summary of coding-region and protein annotation statistics for assembled African elephant whole blood transcriptomes.

Assembly Tool	**Coding regions with ORFs (%)**	**Average protein length (aa) (min-max)**	**N50 (aa)**	**Functionally annotated proteins with ORF (%)**	**Redundancy (%)**
StringTie	37.5	207.1 (85 - 8 570)	229	63.5	18.0
StringTie2	54.6	259.3 (86 - 6 552)	303	81.7	17.0
StringTie2 annotation-guided	**83.3**	**603.7 (85 - 34 606)**	**835**	**92.7**	47.0

Trinity short-read	12.1	213.1 (85 - 16 486)	241	50.1	32.0
rnaSPAdes short-read	10.7	247.8 (85 - 14 834)	320	58.3	33.0
rnaSPAdes hybrid	11.2	256 (85 - 14 834)	337	55.0	35.0
RNA-Bloom2	15.0	162.9 (85 - 1 275)	165	70.4	**13.0**

The African savanna elephant whole blood transcriptome was generated using multiple tools. The proportions of transcripts containing open reading frames (ORFs; %) were predicted using TransDecoder (v 5.2.0), average protein length (aa) and N50 (aa) was obtained using SeqKit (v 0.15.0), functional annotation of predicted proteins (%) was assessed using eggNOG-mapper (v 2.1.13) and protein redundancy (%) was determined by CD-HIT (v 4.7). Bold text indicates the best-performing value for each metric. aa: amino acid, ORF: open reading frame.

Predicted protein lengths were evaluated as an additional indicator of transcript functionality. Proteins predicted from the *StringTie2* annotation-guided assembly had the greatest mean length (603.7 amino acids (aa)), the broadest length distribution, and the longest predicted protein (34,606 aa; Table 3). The reference-guided assemblies also exhibited higher proportions of transcripts containing predicted coding regions compared to the *de novo* assemblies. Furthermore, the *StringTie2* annotation-guided assembly achieved the highest protein-length N50 value (Table 3), indicating a larger proportion of long protein-coding sequences.

Functional annotation of predicted proteins was performed using *eggNOG-mapper*. Proteins were considered annotated when assigned at least one functional descriptor,

including a gene description, Gene Ontology (GO) term, KEGG orthology (KO), Cluster of Orthologous Genes (COG) category, or a Protein Families (Pfam) domain. Among all assemblies, the *StringTie2* annotation-guided assembly achieved the highest annotation rate, with 92.7% of predicted proteins receiving at least one functional annotation (Table 3).

Protein redundancy was assessed using *CD-HIT*. The *StringTie2* annotation-guided assembly showed the highest level of protein redundancy (47.0%), whereas *RNA-Bloom2* exhibited the lowest redundancy (13.0%; Table 3). The remaining reference-guided assemblies showed lower levels of redundancy than the *rnaSPAdes* and *Trinity* assemblies (Table 3). Overall, the *StringTie2* annotation-guided assembly combined high coding potential, long predicted protein sequences, and extensive functional annotation.

Functional annotation of top-performing transcriptomes

Functional annotations generated with *eggNOG-mapper* were further evaluated for the top-performing *de novo (rnaSPAdes hybrid)*, reference-guided (*StringTie2*), and annotation-guided (*StringTie2* annotation-guided) assemblies. Following filtering to remove low-confidence annotations, 92.3% of annotations were retained in the *rnaSPAdes* hybrid assembly, 96.0% in the *StringTie2* assembly, and 97.5% in the *StringTie2* annotation-guided assembly. The high proportion of retained annotations suggests strong support for the majority of functional assignments.

Gene Ontology (GO) terms were classified into Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) ontologies and compared among the three assemblies. Most GO terms were shared across all three assemblies, although a larger number of annotations were shared between the *StringTie2* annotation-guided

and *rnaSPAdes* hybrid assemblies (Figure 4). *StringTie2* contained the fewest annotated GO terms, consistent with its smaller transcriptome size relative to the other two assemblies.

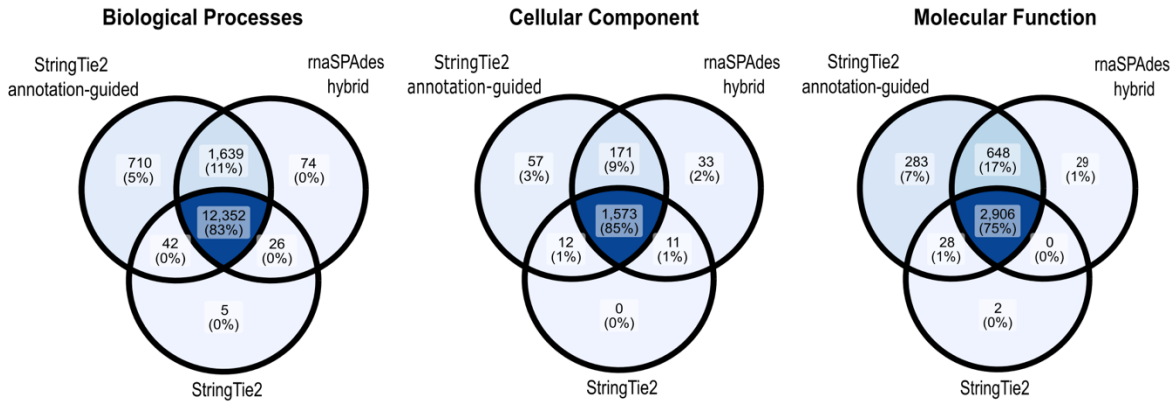
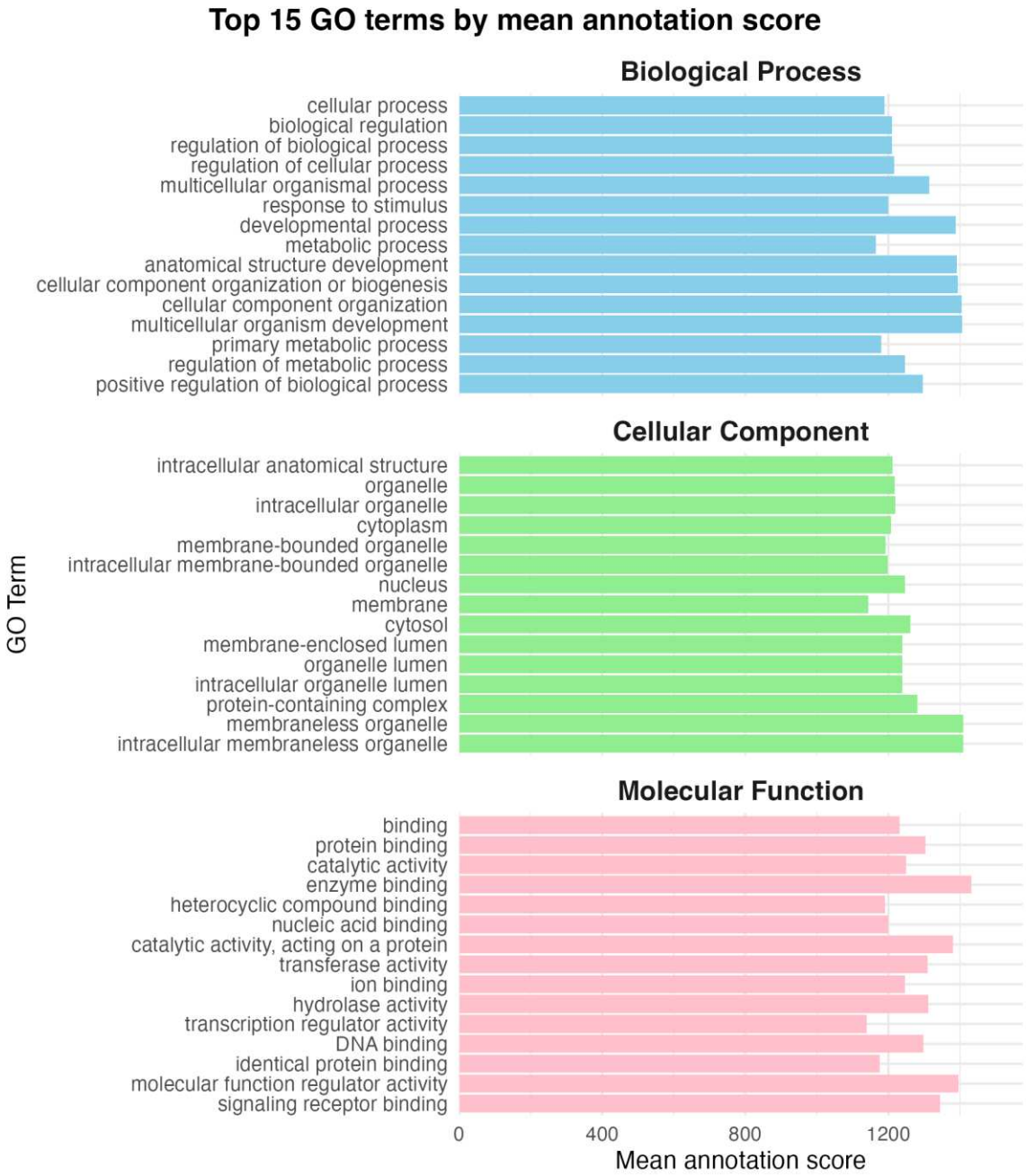


Figure 4. *Venn diagrams depicting the overlap of Gene Ontology (GO) terms among the StringTie2, StringTie2 annotation-guided and rnaSPAdes hybrid African savanna elephant whole blood transcriptome assemblies. GO terms were assigned using eggNOG-mapper and classified using the GO.db package in R Studio. Terms were grouped into Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) ontologies.*

As the *StringTie2* annotation-guided assembly consistently outperformed the other assemblies across multiple quality metrics, its functional annotations were investigated for specific biological processes and immune-related functions. GO terms were ranked according to annotation frequency, and the top 15 most highly represented terms from each ontology category were examined (Figure 5).



506

507 ***Figure 5.*** *Top 15 gene ontology (GO) terms identified in the African savanna elephant whole*
508 *blood transcriptome for the Biological Process (BP), Cellular Component (CC) and Molecular*
509 *Function (MF) ontologies. Go terms were assigned using eggNOG-mapper and annotated by*
510 *the GO.db package in R studio. Terms were ranked according to their total scores, and the 15*
511 *highest-ranking terms within each category are shown.*

The most abundant BP terms were primarily associated with general cellular and metabolic processes (Figure 5). The dominant CC terms included intracellular structures such as the cytosol, cytoplasm, membrane, nucleus and lumen. Within the MF category, the most frequently represented terms were associated with binding and catalytic activities. Together, these results indicate that the transcriptome encompasses a broad range of cellular functions and biological processes.

KEGG pathway annotations were also examined for the *StringTie2* annotation-guided assembly (Table 4). Most annotated transcripts were associated with metabolic pathways, which also received the highest annotation scores. Additional highly represented pathways were associated with cancer-associated pathways, viral infections, endocytosis and signalling pathways (Table 4). These are consistent with the presence of conserved cellular, regulatory, and signalling mechanisms within the assembled transcriptome.

Table 4. Top KEGG Pathways identified in the *StringTie2* annotation-guided assembly of African elephant whole blood transcriptome.

KEGG Pathway	**Transcripts**	**Total score**	**Mean score**	**Pathway name**
ko01100	4303	4137783	961.6	Metabolic pathways
ko05200	2045	2321922	1135.4	Pathways in cancer
ko05165	1580	2075378	1313.5	Human papillomavirus infection
ko04740	3953	2033190	514.3	Olfactory transduction

ko04144	1393	1622599	1164.8	Endocytosis
ko04010	1303	1552717	1191.7	MAPK signaling pathway
ko04151	1277	1488532	1165.7	PI3K-Akt signaling pathway
ko04015	982	1469808	1496.8	Rap1 signaling pathway
ko05202	1164	1396612	1199.8	Transcriptional mis-regulation in cancer
ko04120	758	1322793	1745.1	Ubiquitin mediated proteolysis
ko04810	889	1310576	1474.2	Regulation of actin cytoskeleton
ko05205	904	1257268	1390.8	Proteoglycans in cancer
ko05166	1236	1253724	1014.3	Human T-cell leukemia virus 1 infection
ko05169	1145	1190606	1039.8	Epstein-Barr virus infection
ko05206	862	1148125	1331.9	MicroRNAs in cancer

Pathways were assigned using eggNOG-mapper, ranked according to their total scores and annotated using KEGGREST package in R studio. Only the top 15 pathways are presented.

The *StringTie2* annotation-guided assembly contained 11.7% unannotated transcripts according to the *rnaQUAST* analysis (Table 2). To identify potential novel transcripts and genes, the assembly was compared with the current African elephant reference genome annotation using *gffcompare.* Transcripts were classified as novel (class code

'u') if they displayed no overlap with annotated reference transcripts. This analysis identified 2,231 putative novel genes, representing 6.1% of assembled genes of which 99 were annotated by *eggNOG-mapper*, encoding 2,342 putative novel transcripts, representing 2.8% of assembled transcripts of which 119 were annotated. The top BLASTx results showed no similarity to known proteins in 2081 (88.9%) of these transcripts. However, examination of sequence similarity against the nucleotide database, revealed that only 338 (14.4%) had no known sequence matches.

Discussion

The lack of a whole blood reference transcriptome has limited investigation of immune gene expression in African savanna elephants (*Loxodonta africana*). In this study, multiple transcriptome assemblies were generated and evaluated using combinations of short-read NGS and long-read TGS data together with reference-guided (*StringTie, StringTie2* and *StringTie2* annotation-guided) and *de novo* (*rnaSPAdes*, *Trinity* and *RNA-Bloom2*) assembly approaches. Assembly performance was assessed using complementary metrics, including assembly statistics, read representation, transcript completeness, and functional annotation. Collectively, these analyses identified the *StringTie2* annotation-guided assembly as the highest quality transcriptome generated in this study. The assembly exhibited high transcript continuity, low levels of unannotated transcripts, extensive functional annotation, and a BUSCO completeness score of 98.4%. Functional characterisation further demonstrated representation of diverse biological processes, molecular functions, cellular components, and pathways relevant to immune and cellular regulation. These findings support the use of this

557 whole blood transcriptome as a reference resource for future studies of gene
558 expression and host responses related to disease in African savanna elephants.

559 The short-read NGS and long-read TGS datasets used for transcriptome assembly
560 were of high quality and achieved high mapping efficiencies against the African
561 elephant reference genome. As expected, the Oxford Nanopore long-reads displayed
562 lower per-base accuracy than short-read sequencing data, reflecting the higher error
563 rates typically associated with TGS technologies [62]. However, the complementary
564 strengths of the two sequencing platforms provided extensive coverage, while
565 facilitating the reconstruction of longer transcript sequences. The low alignment error
566 rates observed following processing and filtering indicated that both datasets were
567 suitable for transcriptome assembly.

568 Distinct performance patterns were observed between reference-guided and *de novo*
569 assembly approaches. The *de novo* assemblers generally produced larger
570 transcriptomes with greater numbers of shorter transcripts but also exhibited higher
571 duplication rates and lower transcript contiguity. In contrast, the reference-guided
572 assemblies generated fewer transcripts, greater transcript contiguity, lower duplication
573 ratios, and improved reconstruction of annotated isoforms and exons. A similar trend
574 was noted in a previous study where reference guided assemblies had longer
575 transcripts, as noted by higher N50 values, compared to *de novo* transcriptome
576 assemblies of the blood-sucking insect, *Rhodnius prolixus* [63].

577 Among the *de novo* assemblers, *RNA-Bloom2* differed substantially from *Trinity* and
578 *rnaSPAdes*, producing a smaller transcriptome with fewer transcripts and a lower
579 duplication rate. Although some assembly characteristics were more like those
580 observed for the reference-guided assemblies, *RNA-Bloom2* exhibited relatively low

transcriptome completeness and the highest proportion of missing BUSCO genes. Overall, the reference-guided approaches consistently outperformed *de novo* assemblies across most quality metrics. When a reference genome was available this was true for other non-model species as well [63,64].

Comparison of the *rnaSPAdes* assemblies generated from short-read data alone and those from combined short-read and long-read datasets demonstrated the value of incorporating long-read sequencing data into transcriptome assembly. Although inclusion of long reads had only modest effects on overall assembly size and transcript counts, it increased the proportion of longer transcripts and improved the recovery of assembled genes and isoforms. These findings are consistent with previous reports demonstrating improved transcript reconstruction and completeness when long-read data are incorporated into *rnaSPAdes* hybrid assemblies [18]. Hybrid assemblies also out performed assemblies using either short- or long-read data alone using other assembly tools [63]. Similar to the results in this study, hybrid assemblies were more complete and included longer transcripts [65]. Together, these results support the use of hybrid sequencing strategies for transcriptome assembly when both short-read and long-read data are available.

A hybrid Trinity assembly was also attempted however the assembly could not be completed with the available computational resources and was terminated before completion. Similar computational constraints associated with hybrid Trinity assemblies have been reported previously [18] and *de novo* assmeblies [64]. Consequently, the Trinity hybrid assembly was excluded from further comparisons. Overall, these results support the use of hybrid assembly approaches that integrate long-read and short-read sequencing data to improve the quality and completeness of the transcriptome.

Transcriptome completeness was further evaluated using BUSCO. The *StringTie2* annotation-guided assembly achieved the highest completeness score (98.4%), indicating that nearly all expected conserved mammalian orthologs were represented. Although a completeness score of 100% would not be expected from whole blood samples collected from healthy adult males, as many genes are tissue-, developmental-stage-, or condition-specific, the completeness observed in the present study is comparable to that reported for high-quality transcriptomes generated in other non-model species [21, 66]. The low proportion of fragmented BUSCOs further supports the integrity of the assembled transcriptome.

Considering the assembly statistics, transcript completeness metrics, and functional annotation results, the *StringTie2* annotation-guided assembly was identified as the optimal transcriptome. This outcome was expected given the recent improvements to the African savanna elephant reference genome and annotation resources [27]. *StringTie2* was specifically developed to improve transcript reconstruction from both short- and long-read sequencing data, and has previously demonstrated improved transcript recovery when reference annotations are incorporated during assembly [14]. Similar studies have also reported enhanced transcript reconstruction and BUSCO recovery using *StringTie2* compared with alternative reference-guided and *de novo* assembly methods [14, 68, 68].

Despite its overall superior performance, the *StringTie2* annotation-guided assembly contained a relatively high proportion of duplicated BUSCOs and exhibited the highest level of protein redundancy. Several factors may contribute to these observations, including the retention of alternative transcript isoforms and closely related coding sequences. Since the assembly was guided by an existing annotation and demonstrated strong performance across multiple quality metrics, these duplicated

sequences are likely a reflection of genuine biological variation rather than assembly artefacts. Retaining these transcripts therefore preserves potentially important isoform diversity and may improve the utility of the resource for future transcriptomic analyses. Furthermore, *StringTie2* has consistently been reported to recover a greater number of non-redundant transcripts than other assemblers [67, 68], suggesting that our decision to retain these transcripts in the assembly may improve transcriptome completeness and biological representation.

Functional annotation further supported the completeness of the assembled transcriptome. GO annotations encompassed a broad range of biological processes, molecular functions, and cellular components, while KEGG analyses identified pathways associated with cellular metabolism, signalling, endocytosis, immune function, and host–pathogen interactions. A small number of GO terms could not be assigned to a specific ontology due to database limitations or mis-assemblies and were therefore excluded from further analyses. Importantly, no major functional categories appeared under-represented, suggesting that the transcriptome captured a diverse set of biological functions relevant to whole blood gene expression. Since single-tissue reference are rarely complete [63], novel genes and transcripts were expected due to the different tissue source from the current available reference. The identification of more than 2,200 putative novel genes and over 2,300 putative novel transcripts in this transcriptome further highlights the potential of this resource to refine existing African elephant genome annotations.

A limitation of this study was the relatively small number of individuals included in the transcriptome assembly. Additional samples representing different age classes, sexes, and health statuses may increase transcript diversity and improve the representation of low-abundance transcripts. Furthermore, all blood samples were collected from

free-ranging elephants that were chemically immobilised. Although the influence of immobilization on the elephant blood transcriptome is currently unknown, physiological stress responses during capture and drug administration may affect the expression of some genes [69]. Nevertheless, the large sequencing depth and the use of complementary short- and long-read sequencing technologies provided extensive transcript coverage and enabled the generation of a highly complete reference transcriptome. As additional transcriptomic datasets become available, the resource presented here can be further expanded and refined.

Conclusions

A high-quality whole blood reference transcriptome for the African savanna elephant was generated using a StringTie2 annotation-guided assembly approach that combined short- and long-read sequencing data and merged with the guidance of the current reference annotation file. The resulting transcriptome exhibited high completeness, extensive functional annotation, and representation of a broad range of biological and immune-related pathways. In addition to providing a valuable resource for future investigations of immune responses and infectious diseases in African savanna elephants, this study presents a validated workflow for transcriptome assembly in non-model species when both short- and long-read sequencing data are available. The transcriptome and associated annotations provide a foundation for future studies of gene expression, functional genomics, and genome annotation in this endangered species.

List of abbreviations

%: percentage

aa: amino acid

AU: Absorbance Unit

BAM: Binary Alignment/Map

BLAST: Basic Local Alignment Search Tool

bp: base pair

BP: biological process

BUSCO: Benchmarking Universal Single-Copy Orthologs

CC: cellular component

CHPC: Centre for High Performance Computing

DNA: deoxyribonucleic acid

EEHV: Elephant Endotheliotropic Herpesvirus

EEHVHD: Elephant Endotheliotropic Herpesvirus Haemorrhagic Disease

fmol: femtomole

GC: guanine (G) and cytosine (C)

GO: gene ontology

kb: kilobases

KEGG: Kyoto Encyclopaedia of Genes and Genomes

KNP: Kruger National Park

MF: molecular function

mRNA: messenger RNA

NCBI: National Centre for Biotechnology Information

ng: nanogram

NGS: next generation sequencing

ONT: Oxford Nanopore Technology

ORF: open reading frame

PBMC: peripheral blood mononuclear cells

PCR: polymerase chain reaction

RINe: RNA integrity number

RNA: ribonucleic acid

rRNA: ribosomal RNA

SAMRC: South African Medical Research Council

TB: tuberculosis

TGS: third-generation sequencing

Additional Material

Additional File 1: Table S1. Software and parameters used for sequencing read quality control, transcriptome assembly, and transcriptome quality assessment.

Additional File 2: Figure S1: Overview of the data-processing workflow for short-read next generation sequencing (NGS; blue) and long-read third-generation sequencing (TGS; green).

Additional File 3: Figure S2. Overview of the reference guided and *de novo* transcriptome assembly workflows and tools used in this study.

Additional File 4: Figure S3. Overview of the tools used to assess the quality of the whole blood transcriptomes assembled for African savanna elephants, and their respective roles in the analysis workflow.

Additional File 5: Figure S4. MultiQC summary of long-read third generation sequencing (TGS) RNA-seq data before and after adapter removal and read length filtering (reads < 20 bp), assessed using FastQC (v0.12.1).

Additional File 6: Table S2. Summary of sequencing statistics RNA-seq reads derived from African savanna elephant whole blood samples.

Declarations

Ethics approval and consent to participate

South African Veterinary Council (SAVC)-registered wildlife veterinarians were responsible for all procedures, including immobilisation of animals, blood collection, and BAL fluid sampling. Ethical approval for this study was granted by the Stellenbosch University Animal Care and Use Research Ethics Committee (SU-ACU-2021–21572; SU-ACU-2024–30181) and the Stellenbosch University Biological and Environmental Safety Research Ethics Committee (SU-BEE-2021–22561; SU-BES-2025-32836). Section 20 approval was granted by the South African Department of

741 Agriculture, formerly the Department of Agriculture, Land Reform and Rural
742 Development (12/11/1/7/6; 12/11/1/7/6E-(JD); 12/11/1/7/6E (6192SdlR)). A
743 Threatened or Protected Species (TOPS) permit was obtained through the
744 Department of Forestry, Fisheries and Environment, formerly the Department of
745 Environmental Affairs (Standing Permit S02556; S65805; S67119 and Registration
746 Certificate 29416; 02256; 29007). An approved Research Agreement (SANParks
747 Research Agreement BUSP1511; SS717; SS1400) was obtained from South African
748 National Parks (SANParks) which included evaluation by SANParks Animal Care and
749 Use Committee. SANParks is the owner and legal custodian of the elephants used in
750 this study and provided formal approval for their inclusion in the research. All study
751 activities were conducted under SANParks authorisation and in compliance with
752 relevant permits and ethical requirements.

753 *Consent for publication*

754 Not applicable.

755 *Availability of data and materials*

756 The datasets generated and analysed in this study are publicly available on the NCBI
757 Sequencing Read Archive (SRA) under the BioProject accession number
758 PRJNA1499697. The assembled transcriptome and annotation are further available
759 on the NCBI GenBank-Transcriptome Shotgun Assembly (TSA) database under the
760 TSA submission number GLQV00000000.

761 No specific software or code was developed for this study. The experiments and
762 bioinformatics were performed in accordance with the protocols detailed in the
763 Methods section. The versions and parameters of all software used are listed in
764 Additional File 1: Table S1.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the South African Medical Research Council (SAMRC) Centre for Tuberculosis Research, National Research Foundation (NRF) South African Research Chair Initiative (SARChI Grant 86949), NRF Competitive Program for Rated Researcher Grant (CPRR Grant 129268), NRF Black Academics Advancement Programme (BAAP Grant NFSG23040589700), The Crick Africa Network: African Career Acceleration Award (CAN ACA Grant), International Elephant Foundation (IEF), American Association of Zoo Veterinarians (AAZV) Wild Animal Health Fund (WAHF) (AAZV WAHF Grant 2024 #7) and The Tennessee Elephant Sanctuary. Funding to Stacey Engel was provided through a Department of Science and Innovation (DSI) – National Research Foundation (NRF) Doctoral Scholarship.

Authors' contributions

SCE performed laboratory process, assembled the transcriptomes, analysed data and drafted the manuscript. TJK assisted with sample processing, provided supervision and edited the manuscript. GDS provided supervision and editing the manuscript. PEB assisted with sample collection requiring veterinarian skills and manuscript review. LK provided supervision and edited the manuscript. MAM assisted with sample collection from elephants, provided supervision, funding acquisition and edited the manuscript. TJ assisted in training SCE with laboratory processes and data analysis, provided supervision, funding acquisition and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors wish to acknowledge Simomotfo Intimate Shongwe, Guy Hausler, Tebogo Manamela, Leana Freese, and Dr. Lufuno Netshitavhadulu from the South African National Parks (SANParks) Veterinary Wildlife Services, Kruger National Park (KNP), Skukuza, South Africa. The authors would also like to thank Dr. Ansia van Coller and the South African Medical Research Council (SAMRC) Genomics Platform, South Africa for their contributions to sequencing and for providing access to the MGI and ONT sequencing platforms. The authors further acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project. The authors used Microsoft Copilot (Microsoft Corporation, Redmond, Wahington 98052, USA) to assist with grammar checking, language editing, and improvement of manuscript readability. All scientific content, interpretations, and final manuscript revisions were reviewed and approved by the authors.

References

1. Maslow JN, Mikota SK. Tuberculosis in elephants - A reemergent disease: diagnostic dilemmas, the natural history of infection, and new immunological tools. Vet Pathol. 2015;52:437–40. doi.org/10.1177/0300985814568357.

2. Long SY, Latimer EM, Hayward GS. Review of elephant endotheliotropic herpesviruses and acute hemorrhagic disease. ILAR Journal. 2016;56:283–96. doi.org/10.1093/ilar/ilv041.

3. Gobush KS, Edwards CTT, Balfour D, Wittemyer G, Maisels F, Taylor RD. *Loxodonta africana* (amended version of 2021 assessment). The IUCN Red List of Threatened Species 2022: eT181008073A223031019. 2022. doi.org/10.2305/IUCN.UK.2021-1.RLTS.T181008073A181022663.en

4. Chaussabel D, Pascual V, Banchereau J. Assessing the human immune system through blood transcriptomics. BMC Biol. 2010;8:84. doi.org/10.1186/1741-7007-8-84.

5. Yan Y, Zhu S, Jia M, Chen X, Qi W, Gu F, et al. Advances in single-cell transcriptomics in animal research. J Anim Sci Biotechnol. 2024;15:102. doi.org/10.1186/s40104-024-01063-y.

6. Puglia GD, Prjibelski AD, Vitale D, Bushmanova E, Schmid KJ, Raccuia SA. Hybrid transcriptome sequencing approach improved assembly and gene annotation in *Cynara cardunculus (L.).* BMC Genom. 2020;21:317. doi.org/10.1186/s12864-020-6670-5.

7. Ramberg S, Høyheim B, Østbye T-KK, Andreassen R. A *de novo* full-length mRNA transcriptome generated from hybrid-corrected PacBio long-reads improves the transcript annotation and identifies thousands of novel splice variants in Atlantic salmon. Front Genet. 2021;12-2021. doi.org/10.3389/fgene.2021.656334

8. Han SW, Jewell S, Thomas-Tikhonenko A, Barash Y. Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads. Genome Res. 2024;34:1624–35. doi.org/10.1101/gr.278659.123.

9. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671–82. doi.org/10.1038/nrg3068.

832 10. Luo Y, Liao X, Wu F-X, Wang J. Computational approaches for transcriptome
833 assembly based on sequencing technologies. Curr Bioinform. 2020;15:2–16.
834 doi.orgdoi.org/10.2174/1574893614666190410155603.

835 11. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to *de novo* transcriptome
836 assembly and annotation. Brief Bioinform. 2022;23:bbab563.
837 doi.org/10.1093/bib/bbab563.

838 12. Jackson DJ, Cerveau N, Posnien N. *De novo* assembly of transcriptomes and
839 differential gene expression analysis using short-read data from emerging model
840 organisms – a brief guide. Front Zool. 2024;21:17. doi.org/10.1186/s12983-024-
841 00538-y.

842 13. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL.
843 *StringTie* enables improved reconstruction of a transcriptome from RNA-seq reads.
844 Nat Biotechnol. 2015;33:290–5. doi.org/10.1038/nbt.3122.

845 14. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome
846 assembly from long-read RNA-seq alignments with *StringTie2*. Genome Biol.
847 2019;20:278. doi.org/10.1186/s13059-019-1910-1.

848 15. Huang Z, Gallot A, Lao NT, Puechmaille SJ, Foley NM, Jebb D, et al. A nonlethal
849 sampling method to obtain, generate and assemble whole blood transcriptomes from
850 small, wild mammals. Mol Ecol Resour. 2016;16:150–62. doi.org/10.1111/1755-
851 0998.12447.

852 16. Videvall E, Paxton KL, Campana MG, Cassin-Sackett L, Atkinson CT, Fleischer
853 RC. Transcriptome assembly and differential gene expression of the invasive avian

malaria parasite *Plasmodium relictum* in Hawai'i. Ecol Evol. 2021;11:4935–44. doi.org/10.1002/ece3.7401.

17. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. *rnaSPAdes*: a *de novo* transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8:giz100. doi.org/10.1093/gigascience/giz100.

18. Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, et al. Extending *rnaSPAdes* functionality for hybrid transcriptome assembly. BMC Bioinform. 2020;21:302. doi.org/10.1186/s12859-020-03614-2.

19. Nip KM, Hafezqorani S, Gagalova KK, Chiu R, Yang C, Warren RL, et al. Reference-free assembly of long-read transcriptome sequencing data with *RNA-Bloom2*. Nat Commun. 2023;14:2940. doi.org/10.1038/s41467-023-38553-y.

20. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. *Trinity*: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011;29:644–52. doi.org/10.1038/nbt.1883.

21. Ibeh N, Feigin CY, Frankenberg SR, McCarthy DJ, Pask AJ, Romero IG, et al. *De novo* transcriptome assembly and genome annotation of the fat-tailed dunnart (*Sminthopsis crassicaudata*). Gigabyte. 2024;2024:1–16. doi.org/10.46471/gigabyte.118.

22. Libro P, Chiocchio A, De Rysky E, Di Martino J, Bisconti R, Castrignanò T, et al. *De novo* transcriptome assembly and annotation for gene discovery in *Salamandra salamandra* at the larval stage. Sci Data. 2023;10:330. doi.org/10.1038/s41597-023-02217-9.

876 23. Gao Z, Wang B, Lu Y, Chong Y, Li M, Hong J, et al. *De novo* transcriptome
877 assembly and annotation of the semi-wild Gayal (*Bos frontalis*). Sci Data.
878 2025;12:589. doi.org/10.1038/s41597-025-04927-8.

879 24. Mamrot J, Legaie R, Ellery SJ, Wilson T, Seemann T, Powell DR, et al. *De novo*
880 transcriptome assembly for the spiny mouse (*Acomys cahirinus*). Sci Rep.
881 2017;7:8996. doi.org/10.1038/s41598-017-09334-7.

882 25. Vilperte V, Lucaciu CR, Halbwirth H, Boehm R, Rattei T, Debener T. Hybrid *de novo*
883 transcriptome assembly of poinsettia (*Euphorbia pulcherrima Willd. Ex Klots*ch) bracts.
884 BMC Genom. 2019;20:900. doi.org/10.1186/s12864-019-6247-3.

885 26. Hou Z-C, Sterner KN, Romero R, Than NG, Gonzalez JM, Weckle A, et al.
886 Elephant transcriptome provides insights into the evolution of eutherian placentation.
887 Genome Biol Evol. 2012;4:713–25. doi.org/10.1093/gbe/evs045.

888 27. Shi M, Chen F, Sahu SK, Wang Q, Yang S, Wang Z, et al. Haplotype-resolved
889 chromosome-scale genomes of the Asian and African Savannah Elephants. Sci Data.
890 2024;11:63. doi.org/10.1038/s41597-023-02729-4.

891 28. Hölzer M, Marz M. *De novo* transcriptome assembly: A comprehensive cross-
892 species comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8:giz039.
893 doi.org/10.1093/gigascience/giz039.

894 29. Engel SC, Jooste T, Kerr TJ, Spuy GD van der, Buss PE, Kleynhans L, et al.
895 Evaluation of human-specific RNA enrichment kits for transcriptome sequencing in a
896 non-model organism: the African savanna elephant (*Loxodonta africana*). 2026.
897 Preprint at doi.org/10.21203/rs.3.rs-10341759/v1.

30. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

31. Ewels P, Magnusson M, Lundin S, Käller M. *MultiQC*: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8. doi.org/10.1093/bioinformatics/btw354.

32. Bolger AM, Lohse M, Usadel B. *Trimmomatic*: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. doi.org/10.1093/bioinformatics/btu170.

33. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. *STAR*: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. doi.org/10.1093/bioinformatics/bts635.

34. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008. doi.org/10.1093/gigascience/giab008.

35. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4. doi.org/10.1093/bioinformatics/btv566.

36. Wick R. rrwick/Porechop. 2017. https://github.com/rrwick/porechop

37. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. *NanoPack*: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9. doi.org/10.1093/bioinformatics/bty149.

920 38. De Coster W, Rademakers R. *NanoPack2*: population-scale evaluation of long-
921 read sequencing data. Bioinformatics. 2023;39:btad311.
922 doi.org/10.1093/bioinformatics/btad311.

923 39. Li H. New strategies to improve *minimap2* alignment accuracy. Bioinformatics.
924 2021;37:4572–4. doi.org/10.1093/bioinformatics/btab705.

925 40. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020; 9:
926 ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.1.

927 41. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De*
928 *novo* transcript sequence reconstruction from RNA-seq using the *Trinity* platform for
929 reference generation and analysis. Nat Protoc. 2013;8:1494–512.
930 doi.org/10.1038/nprot.2013.084.

931 42. Mohamadi H, Khan H, Birol I. ntCard: a streaming algorithm for cardinality
932 estimation in genomics data. Bioinformatics. 2017;33:1324–30.
933 doi.org/10.1093/bioinformatics/btw832.

934 43. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome
935 assembly from long uncorrected reads. Genome Res. 2017;27:737–46.
936 doi.org/10.1101/gr.214270.116.

937 44. Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. *rnaQUAST*: a
938 quality assessment tool for *de novo* transcriptome assemblies. Bioinformatics.
939 2016;32:2210–2. doi.org/10.1093/bioinformatics/btw218.

940 45. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. *BUSCO*:
941 assessing genome assembly and annotation completeness with single-copy
942 orthologs. Bioinformatics. 2015;31:3210–2. doi.org/10.1093/bioinformatics/btv351.

943 46. Seppey M, Manni M, Zdobnov EM. *BUSCO*: Assessing Genome Assembly and
944 Annotation Completeness. In: Kollmar M, editor. Gene Prediction: Methods and
945 Protocols. New York, NY: Springer; 2019. p. 227–45. doi.org/10.1007/978-1-4939-
946 9173-0_14.

947 47. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel
948 and streamlined workflows along with broader and deeper phylogenetic coverage for
949 scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–
950 54. doi.org/10.1093/molbev/msab199.

951 48. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag. 2016.

952 49. R Core Team. R: A language and environment for statistical computing. R
953 Foundation for Statistical Computing, Vienna, Austria. 2021.

954 50. Posit team. Studio: Integrated Development Environment for R. Posit Software.
955 2026.

956 51. Shen W, Le S, Li Y, Hu F. *SeqKit*: A cross-platform and ultrafast toolkit for FASTA/Q
957 file manipulation. PLoS One. 2016;11:e0163962.
958 doi.org/10.1371/journal.pone.0163962.

959 52. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J.
960 eggNOG-mapper v2: Functional annotation, orthology assignments, and domain

961 prediction at the metagenomic scale. Mol Biol Evol. 2021;38:5825–9.
962 doi.org/10.1093/molbev/msab293.

963 53. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook
964 H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated
965 orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res.
966 2019;47:D309–14. doi.org/10.1093/nar/gky1085.

967 54. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale
968 using DIAMOND. Nat Methods. 2021;18:366–8. doi.org/10.1038/s41592-021-01101-
969 x.

970 55. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to
971 reduce the size of large protein databases. Bioinformatics. 2001;17:282–3.
972 doi.org/10.1093/bioinformatics/17.3.282.

973 56. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds
974 up clustering of large protein databases. Bioinformatics. 2002;18:77–82.
975 doi.org/10.1093/bioinformatics/18.1.77.

976 57. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of
977 protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
978 doi.org/10.1093/bioinformatics/btl158.

979 58. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and
980 comparing biological sequences. Bioinformatics. 2010;26:680–2.
981 doi.org/10.1093/bioinformatics/btq003.

982 59. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-
983 generation sequencing data. Bioinformatics. 2012;28:3150–2.
984 doi.org/10.1093/bioinformatics/bts565.

985 60. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA
986 sequences. J Comput Biol. 2000;7:203–14. doi.org/10.1089/10665270050081478.

987 61. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped
988 BLAST and PSI-BLAST: a new generation of protein database search programs.
989 Nucleic Acids Res. 1997;25:3389–402. doi.org/10.1093/nar/25.17.3389.

990 62. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and
991 its applications. Nat Rev Genet. 2020;21:597–614. doi.org/10.1038/s41576-020-
992 0236-x.

993 63. Marchant A, Mougel F, Mendonça V, Quartier M, Jacquin-Joly E, da Rosa JA, et
994 al. Comparing *de novo* and reference-based transcriptome assembly strategies by
995 applying them to the blood-sucking bug *Rhodnius prolixus*. Insect Biochem Mol Biol.
996 2016;69:25–33. doi.org/10.1016/j.ibmb.2015.05.009.

997 64. Vijay N, Poelstra JW, Künstner A, Wolf JBW. Challenges and strategies in
998 transcriptome assembly and differential gene expression quantification. A
999 comprehensive *in silico* assessment of RNA-seq experiments. Mol Ecol. 2013;22:620–
1000 34. doi.org/10.1111/mec.12014.

1001 65. Marchant A, Mougel F, Almeida C, Jacquin-Joly E, Costa J, Harry M. *De novo*
1002 transcriptome assembly for a non-model species, the blood-sucking bug *Triatoma*
1003 *brasiliensis,* a vector of Chagas disease. Genetica. 2015;143:225–39.
1004 doi.org/10.1007/s10709-014-9790-5.

66. Caputo DR, Robson SC, Werner I, Ford AT. Complete transcriptome assembly and annotation of a critically important amphipod species in freshwater ecotoxicological risk assessment: *Gammarus fossarum*. Environ Int. 2020;137:105319. doi.org/10.1016/j.envint.2019.105319.

67. Voshall A, Behera S, Li X, Yu X-H, Kapil K, Deogun JS, et al. A consensus-based ensemble approach to improve transcriptome assembly. BMC Bioinform. 2021;22:513. https://doi.org/10.1186/s12859-021-04434-8.

68. Srikakulam N, Sridevi G, Pandi G. High-quality reference transcriptome construction improves RNA-seq quantification in Oryza sativa indica. Front Genet. 2022;13-2022. doi.org/10.3389/fgene.2022.995072

69. Mancia A, Warr GW, Chapman RW. A transcriptomic analysis of the stress induced by capture–release health assessment studies in wild dolphins (*Tursiops truncatus*). Mol Ecol. 2008;17:2581–9. doi.org/10.1111/j.1365-294X.2008.03784.x.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTARYMATERIALElephantTranscriptomeAssembly2026.10.20.docx