Journal of Applied Biology & Biotechnology Vol. 14(5), pp. 132-141, September-October, 2026
Available online at http://www.jabonline.in
DOI: 10.7324/JABB.2026.315063

De novo transcriptome assembly and RNA quality optimization in *Gigantochloa levis* using Illumina sequencing

Mohammad Rahmat Derise[1], Anis Adilah Mustafa[1], Kenneth Francis Rodrigues[1], Julius Kulip[2,3], Intan Mariana Maliki[1], Chee Wei Yew[1], Wilson Thau Lym Yong[1*]

[1]Biotechnology Research Institute, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia.
[2]Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia.
[3]Botanicals Sabah Sdn. Bhd., Kota Kinabalu, Malaysia.

ARTICLE INFO

Article history:
Received on: 25/03/2026
Accepted on: 04/06/2026
Available Online: 25/07/2026

Key words:
Bamboo transcriptomics, *de novo* transcriptome assembly, *Gigantochloa levis*, RNA extraction, RNA sequencing

ABSTRACT

Gigantochloa levis is an economically important tropical bamboo; however, molecular studies on this species remain limited due to difficulties in obtaining high-quality RNA and the scarcity of genomic resources. This study aimed to develop an optimized RNA extraction protocol and generate a *de novo* transcriptome assembly for *G. levis*. Leaf, shoot, and stem tissues were processed using a modified TransZol Up Plus RNA Kit, and RNA quality was evaluated by spectrophotometry, agarose gel electrophoresis, and Agilent bioanalyzer analysis. High-quality RNA was subsequently sequenced using the Illumina NovaSeq 6000 platform. Sequencing generated 57.6 GB of data with more than 98% clean reads, low error rates, and high base-calling accuracy. *De novo* assembly using Trinity produced 1,056,042 contigs, providing a comprehensive transcriptomic dataset for the species. The optimized extraction protocol consistently yielded RNA integrity numbers above 8.0 across all tissues. Benchmarking universal single-copy orthologs (BUSCO) analysis demonstrated high transcriptome completeness, with 90.7% and 90.6% complete BUSCOs identified in the Trinity and CD-HIT-EST transcript sets, respectively. The CD-HIT-EST assembly also showed reduced duplicated BUSCOs and increased single-copy BUSCOs, indicating effective removal of redundant transcripts. All sequencing data have been deposited in the National Center for Biotechnology Information Sequence Read Archive. Collectively, this study establishes a reliable RNA extraction workflow and provides a valuable transcriptomic resource that will facilitate future functional genomics, comparative transcriptomics, and plant biotechnology research in bamboo and other lignocellulosic species.

1. INTRODUCTION

Bamboo species belong to the subfamily Bambusoideae of the grass family Poaceae and are ecologically and economically important due to their rapid growth, high biomass productivity, and broad environmental adaptability [1]. As fast-growing perennial grasses with exceptional carbon assimilation capacity, bamboos contribute significantly to climate change mitigation through carbon sequestration, soil stabilization, and ecosystem restoration [2]. Beyond their ecological value, bamboo species serve as important renewable bioresources for rural communities and global industries, providing raw construction materials, engineered wood products, furniture, textiles, pulp and paper, handicrafts, and food production [3]. The growing global demand for sustainable and low-carbon materials has further strengthened the importance of bamboo as an environmentally responsible alternative to conventional timber resources. Consequently, the global bamboo industry continues to expand, driven by increasing industrial applications and growing recognition of bamboo as a sustainable bioresource [4]. In the context of Industry 5.0, bamboo is increasingly recognized as a highly engineered, traceable, and sustainable bioresource integrated into a low-emission and circular bioeconomy [1]. Recent advances in bamboo genomics and biotechnology have further highlighted the importance of molecular resources for improving biomass production, stress resilience, climate adaptation, and value-added industrial applications in bamboo species [5,6].

Southeast Asia is one of the major centers of bamboo diversity and hosts numerous native and cultivated species with considerable ecological and commercial value. Among these, *Gigantochloa levis* (Buluh Beting) is an important tropical bamboo widely used for construction, furniture production, and traditional handicrafts [7]. Species within the genus *Gigantochloa* belong to the family Poaceae and are characterized by pseudospikelet-type inflorescences clustered at flowering branch nodes, typically containing two to four florets and one terminal imperfect floret. The palea is distinctly two-keeled and ciliate along the margins, representing a diagnostic feature that differentiates *Gigantochloa* from related genera within the *Bambusa–Dendrocalamus–Gigantochloa* complex [8]. Similar to other members of Bambusoideae, *Gigantochloa* species produce caryopsis-type fruits, although flowering and fruiting events are infrequent because

*Corresponding Author
Wilson Thau Lym Yong, Biotechnology Research Institute, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia. E-mail: wilsonyg @ ums.edu.my

of their long and irregular flowering cycles [9,10]. The culms of *G. levis* are valued for their straightness, strength, and durability, making the species particularly suitable for load-bearing structures and high-quality artisanal products.

Although *G. levis* is recognized as commercially important in Southeast Asia, species-specific annual production volume and global revenue data remain limited because bamboo production statistics are generally reported collectively at the genus or industry level rather than for individual species [11]. Current global bamboo production is estimated to exceed 30 million tonnes annually, with major production and utilization concentrated in China, India, Indonesia, Vietnam, Thailand, Malaysia, and the Philippines [12,13]. The global bamboo and rattan market has been estimated to generate more than United States dollars (USD) 60 billion annually through applications in construction, furniture, textiles, engineered biomaterials, pulp and paper, food products, and bioenergy [14]. In Malaysia and other Southeast Asian countries, *G. levis* is among the bamboo species commonly utilized for structural construction, woven products, scaffolding, and cottage industries because of its favorable culm properties and rapid growth characteristics [15,16]. However, publicly curated production and revenue datasets specifically dedicated to *G. levis* remain unavailable, highlighting the broader lack of species-level economic and molecular documentation for tropical bamboo resources. Previous studies on *G. levis* have mainly focused on its morphology, anatomy, mechanical properties, and chemical composition [17,18]. Although these studies have improved understanding of its industrial potential, the molecular mechanisms underlying culm development, fiber differentiation, secondary cell wall biosynthesis, and responses to biotic and abiotic stresses remain poorly understood.

This limited understanding of bamboo molecular biology is largely attributable to the scarcity of genomic and transcriptomic resources for *G. levis* and bamboo species in general. Compared with well-characterized grasses such as rice, maize, and *Brachypodium*, bamboo species remain underrepresented in genomic databases because of their large and highly repetitive genomes, irregular flowering cycles, and complex polyploid evolutionary histories that complicate genome assembly and annotation [6]. Consequently, molecular investigations in bamboo have often been restricted to targeted gene analyses or limited transcriptomic studies, leaving many ecologically and economically important taxa insufficiently explored [19]. Recent advances in bamboo genomics have highlighted the importance of integrating transcriptomics, stress biology, and genetic engineering approaches to better understand bamboo growth, development, and environmental adaptation [6]. In addition, population-level genomic studies have demonstrated that bamboo species possess unique genetic adaptations associated with climatic and environmental responses, further underscoring the importance of molecular resources for bamboo improvement and conservation [20]. The development of transcriptomic resources is, therefore, increasingly important in plant science, as these datasets provide fundamental insights into gene expression, regulatory pathways, biomass formation, lignification, and stress-response mechanisms [21,22]. Such knowledge is essential for accelerating molecular breeding, developing climate-resilient crops, improving sustainable biomass utilization, and supporting future bioeconomy-driven industries.

RNA sequencing has emerged as an effective approach for generating molecular resources in non-model plant species. *De novo* transcriptome assembly enables reconstruction of expressed gene sets without requiring a reference genome and facilitates the identification of functional pathways, regulatory networks, and stress-responsive genes associated with growth, development, and biomass formation [23]. Such datasets are essential for comparative genomics, molecular breeding, conservation strategies, and biotechnology applications. In bamboo, transcriptomic information is particularly valuable for understanding genes involved in cellulose biosynthesis, lignin deposition, vascular differentiation, and rapid culm elongation, all of which directly influence biomass quality and industrial performance [24]. Moreover, transcriptome resources can support the identification of molecular markers and candidate genes relevant to stress tolerance, carbon sequestration efficiency, and adaptation to changing environmental conditions [25,26]. However, transcriptome sequencing depends critically on the availability of high-quality RNA. Bamboo tissues contain high levels of polysaccharides, polyphenols, lignin precursors, and other secondary metabolites that interfere with enzymatic reactions and compromise RNA purity and integrity [27]. Commercial RNA extraction kits frequently produce degraded or contaminated RNA when applied to bamboo tissues, highlighting the need for optimized extraction strategies tailored to species-specific biochemical characteristics. Establishing reproducible and reliable RNA extraction workflows is, therefore, essential for obtaining high-quality transcriptomic datasets from bamboo and other lignocellulosic plant species.

In this study, we addressed a major methodological bottleneck by developing an optimized RNA extraction protocol specifically adapted for leaf, shoot, and stem tissues of *G. levis*. This improved workflow enabled the successful isolation of high-quality RNA suitable for transcriptomic analysis and facilitated the generation of a *de novo* transcriptome assembly for *G. levis* using Illumina NovaSeq sequencing. The resulting dataset, deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), provides a publicly accessible transcriptomic resource for this species. This resource establishes an important molecular foundation for future studies on gene expression, developmental regulation, secondary cell wall biosynthesis, and stress adaptation in *G. levis*. Furthermore, the methodological framework developed here provides practical value for transcriptomic investigations in other recalcitrant bamboo and lignocellulosic plant species. Collectively, this study contributes a robust platform for functional genomics, comparative transcriptomics, and plant biotechnology research aimed at improving bamboo utilization, advancing sustainable biomaterial innovation, and supporting low-emission circular bioeconomy initiatives.

2. MATERIALS AND METHODS

2.1. Plant Material Collection and Identification

Healthy and mature plants of *G. levis* cultivated at the Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Malaysia, were used for leaf, shoot, and stem tissue collection. Species identification was initially performed by a certified plant taxonomist based on morphological characteristics, and a voucher specimen was deposited at the Forest Research Centre, Sabah Forestry Department (voucher specimen number: SAN88611). Species identity was further confirmed through molecular barcoding using the *matK* and *rbcL* markers according to previously described methods [28]. The obtained sequences showed 100% homology with *G. levis*, corresponding to GenBank accession numbers OR483837 (*matK*) and OR483854 (*rbcL*). Freshly harvested tissues were washed with sterile distilled water to remove surface debris and cut into approximately 1–2-cm segments using sterile scissors. Samples were immediately snap-frozen in liquid nitrogen, transferred into sterile cryovials, and stored at –80°C in an ultra-low-temperature freezer (Thermo Scientific, USA) until RNA extraction.

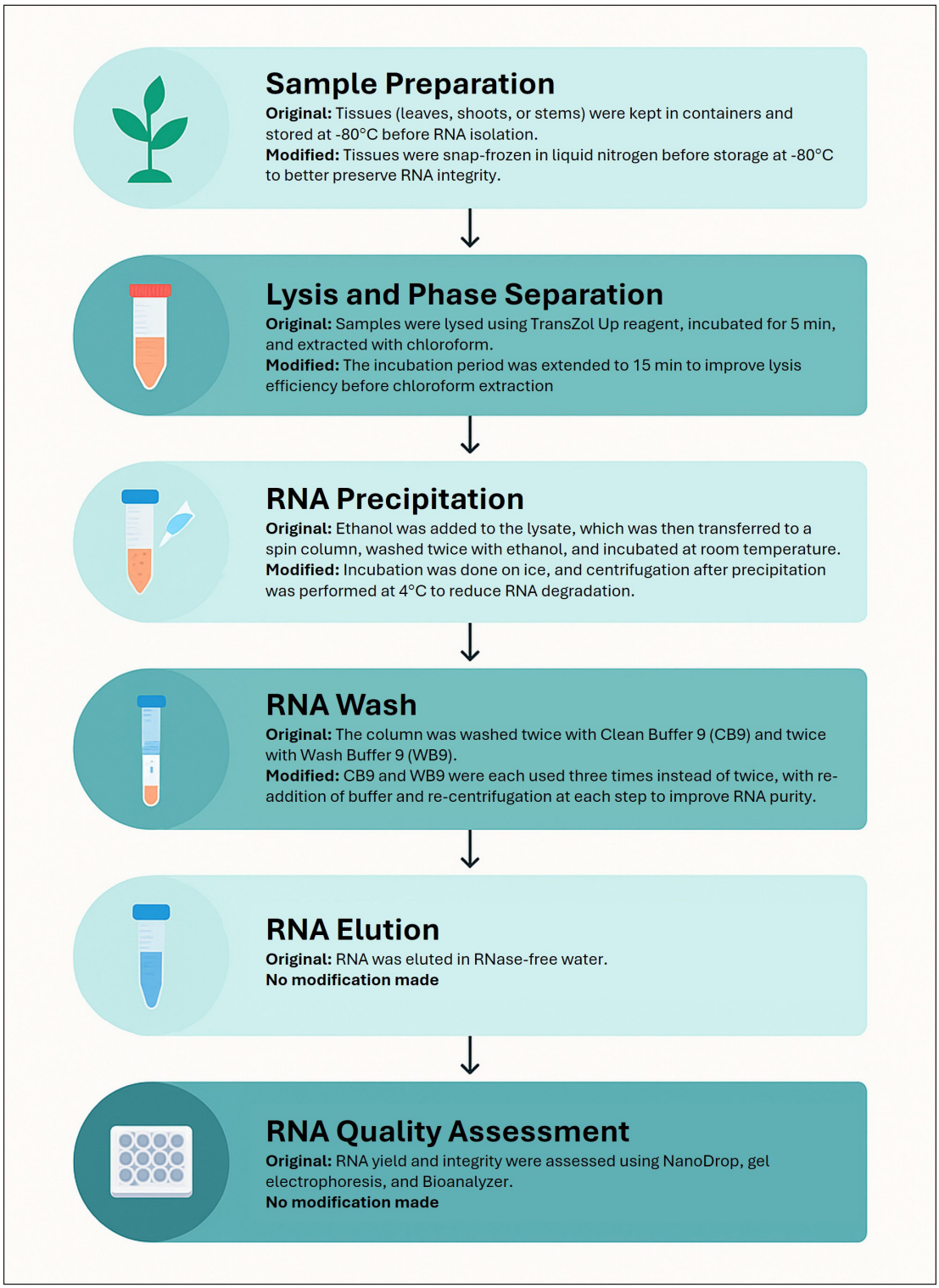

Figure 1. Overview of the optimized RNA extraction protocol using the TransZol Up Plus RNA Kit (Transgen Biotech, Beijing, China), with modifications to the manufacturer's protocol highlighted.

2.2. RNA Extraction and Quality Assessment

Total RNA was extracted using the TransZol Up Plus RNA Kit (TransGen Biotech, Beijing, China) following the manufacturer's protocol with targeted modifications to improve RNA recovery and purity from polysaccharide- and polyphenol-rich bamboo tissues. Frozen tissues were ground into a fine powder in pre-chilled mortars using liquid nitrogen before immediate transfer into tubes containing TransZol Up reagent. Samples were thoroughly homogenized and incubated to ensure complete cell lysis, followed by chloroform-mediated phase separation. RNA precipitation and centrifugation steps were performed at low temperature using a refrigerated microcentrifuge (Eppendorf, Germany) to minimize RNA degradation. Additional washing steps

were incorporated to improve the removal of secondary metabolites and residual contaminants. The resulting RNA pellets were briefly air-dried and dissolved in RNase-free water (Invitrogen, USA).

RNA concentration and purity were measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific, USA), and A260/280 and A260/230 absorbance ratios were recorded. RNA integrity was initially assessed by electrophoresis on 1% agarose gels prepared using RNase-free reagents and subsequently evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) with the RNA 6000 Nano Kit to obtain RNA integrity number (RIN) values. Three biological replicates were prepared for each tissue type during RNA extraction and quality assessment. Before sequencing, equal quantities of RNA from the biological replicates of each tissue type were pooled to generate a representative RNA sample for library preparation. Only RNA samples with RIN values ≥7.0 and purity ratios >2.0 were used for downstream analyses.

2.3. Library Preparation and RNA Sequencing

Total RNA with RIN values ≥7.0 was used for cDNA library preparation using the Illumina TruSeq Stranded mRNA Library Prep Kit (Illumina Inc., San Diego, CA). Library preparation involved mRNA enrichment using oligo (dT) magnetic beads, RNA fragmentation, first- and second-strand cDNA synthesis, adapter ligation, and polymerase chain reaction (PCR) amplification. Library quality and fragment size distribution were evaluated using the Agilent 2100 Bioanalyzer before sequencing. Paired-end sequencing was subsequently performed on the Illumina NovaSeq 6000 platform (Illumina Inc., USA) to generate transcriptomic data for each tissue type. Raw sequencing reads were generated in a file format storing sequencing reads and their quality scores (FASTQ) format [29].

2.4. Read Preprocessing and Quality Control

Raw sequencing reads were processed using fastp (version 0.23.2) to remove adapter sequences, trim low-quality bases, and discard reads containing >10% ambiguous nucleotides. Quality metrics, including Q20 and Q30 values, Guanine–cytosine (GC) content, and per-base sequence quality, were generated during preprocessing. Clean reads obtained after filtering were retained for all downstream analysis.

2.5. *De novo* Transcriptome Assembly and Evaluation

Clean reads from the pooled leaf, shoot, and stem samples were combined and assembled *de novo* using Trinity (Galaxy Version 2.15.1 + galaxy1) implemented on the Galaxy platform [30]. A combined transcriptome assembly was generated to establish a comprehensive reference transcriptome for *G. levis*. Cluster Database at High Identity with Tolerance–Expressed Sequence Tag (CD-HIT-EST) clustering analysis (Galaxy Version 4.8.1 + galaxy0) was subsequently performed to reduce transcript redundancy and improve assembly compactness, resulting in a non-redundant transcript dataset. Assembly statistics, including total contig number, N50 and N90 values, mean contig length, and contig length distribution, were calculated using the FASTA Statistics tool (Galaxy Version 2.0).

Transcriptome completeness and biological representation were evaluated using benchmarking universal single-copy orthologs (BUSCO) analysis (Galaxy Version 5.8.0 + galaxy2) with the embryophyta_odb12 dataset (N = 2,026). BUSCO metrics, including percentages of complete single-copy, complete duplicated, fragmented, and missing orthologs, were used to assess transcriptome completeness and assembly quality. Both the Trinity-assembled transcripts and the CD-HIT-EST non-redundant transcript datasets were evaluated independently to determine the impact of redundancy reduction on transcriptome completeness.

3. RESULTS

3.1. Evaluation of the Optimized RNA Extraction Protocol

The optimized RNA extraction protocol (Fig. 1) consistently produced high-quality RNA from all *G. levis* tissues examined and yielded higher RNA quantity with improved overall extraction performance compared with the original manufacturer's protocol. Using the original TransZol Up Plus RNA Kit protocol, RNA concentrations ranged from approximately 420 to 640 ng/µl (Table 1), whereas the modified protocol increased RNA yields to approximately 500 to 920 ng/?l across leaf, shoot, and stem tissues (Table 2). Spectrophotometric analysis showed that both protocols produced A260/280 and A260/230 ratios above 2.0, indicating minimal contamination by proteins, polysaccharides, and phenolic compounds.

Agarose gel electrophoresis revealed intact RNA with clearly defined 25S and 18S ribosomal RNA bands and no visible smearing or degradation in samples obtained using the optimized protocol (Fig. 2). On the contrary, RNA extracted using the unmodified kit protocol showed partial degradation and reduced band clarity. Agilent Bioanalyzer analysis further confirmed high RNA integrity (Fig. 3), with RIN values consistently above 8.0 across all tissue types analyzed using the optimized protocol (Table 2). In comparison, RNA obtained

Table 1. Quality assessment of RNA isolated from *G. levis* tissues using the original protocol of the TransZol Up Plus RNA Kit (Transgen Biotech, Beijing, China).

Tissue	RNA concentration (ng/µl)	A260/280	A260/230	RIN	rRNA ratio
Leaf	419.8 ± 68.22	2.18 ± 0.01	2.22 ± 0.08	2.40	0.4
Shoot	643.33 ± 36.95	2.16 ± 0.01	2.20 ± 0.04	2.60	0.1
Stem	430.37 ± 32.19	2.15 ± 0.01	2.31 ± 0.07	6.10	1.2

Table 2. Quality assessment of RNA isolated from *G. levis* tissues using the modified protocol of the TransZol Up Plus RNA Kit (Transgen Biotech, Beijing, China).

Tissue	RNA concentration (ng/µl)	A260/280	A260/230	RIN	rRNA ratio
Leaf	649.80 ± 17.76	2.16 ± 0.02	2.25 ± 0.03	8.30	1.9
Shoot	918.37 ± 68.42	2.17 ± 0.02	2.28 ± 0.03	9.10	2.0
Stem	501.1 ± 46.79	2.15 ± 0.02	2.21 ± 0.01	9.20	2.0

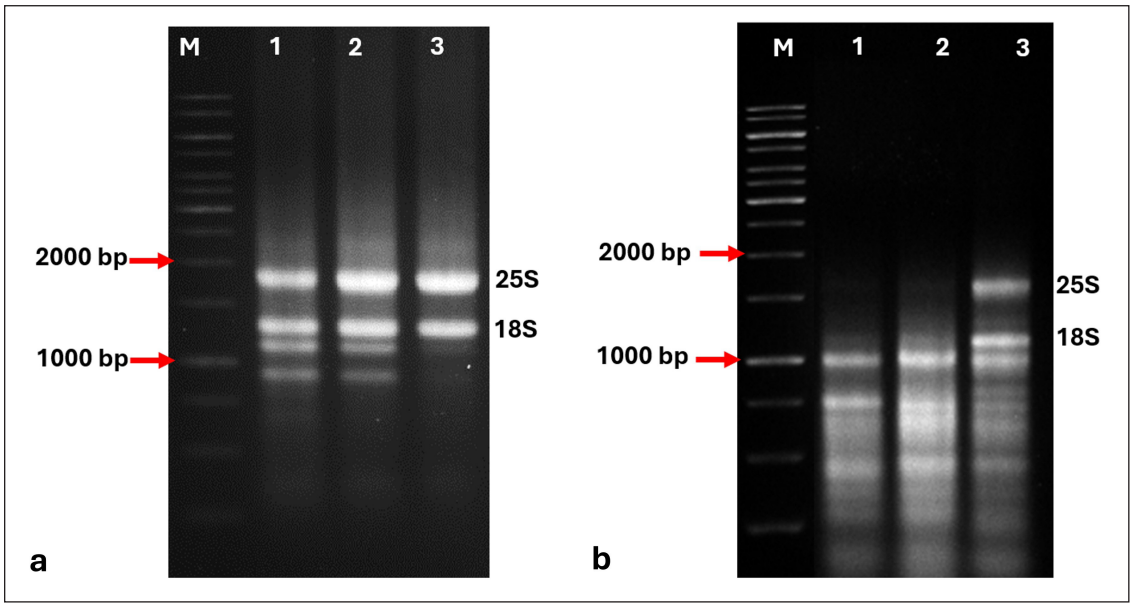

Figure 2. Electrophoresis of total RNA extracted from different tissues of *G. levis* using the modified (a) and original (b) RNA isolation protocols of the TransZol Up Plus RNA Kit. Total RNA was extracted from leaf, shoot, and stem tissues, with three biological replicates per tissue. Samples were separated on a 1% agarose gel and visualized under UV light. Lane M: Thermo Scientific O'GeneRuler 1 kb DNA ladder; Lane 1: leaf; Lane 2: shoot; Lane 3: stem.

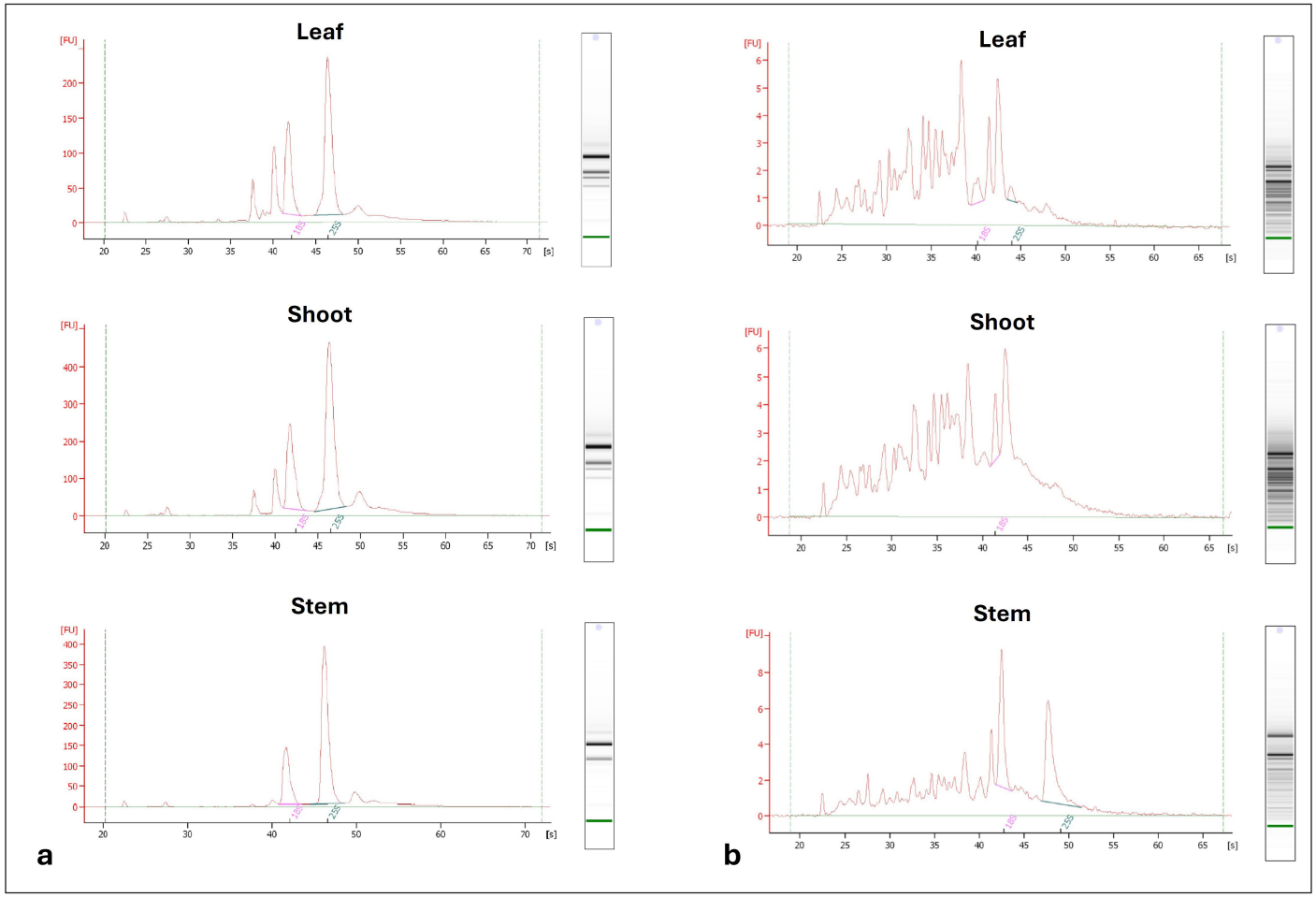

Figure 3. Electropherogram and gel-like image showing total RNA quality from *G. levis* using the modified (a) and original (b) protocols of the TransZol Up Plus RNA Kit. Data were generated using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano assay following the manufacturer's protocol. Electropherograms represent one of three biological replicates per tissue. RIN values were recorded to assess the effect of protocol modification on RNA quality.

Table 3. Sequencing read statistics for *G. levis* tissues.

Tissue	Raw reads	Clean reads	Effective (%)	Error (%)	Q20 (%)	Q30 (%)	GC (%)
Leaf	234,823,160	233,287,514	99.35	0.03	96.14	90.24	44.41
Shoot	275,274,528	270,064,610	98.11	0.03	96.34	90.63	44.74
Stem	267,626,012	262,322,764	98.02	0.03	96.05	90.12	47.01

Table 4. Contig statistics of the Trinity and CD-HIT-EST transcriptome assemblies of *G. levis*.

Parameter	Statistics	
	Trinity transcripts	CD-HIT-EST transcripts
Contig L50	192,229	154,282
Contig N50	983	885
Contig L90	756,812	617,492
Contig N90	286	272
Contig length (max)	19,016	19,016
Contig length (min)	171	179
Contig length (mean)	694	642
Contig length (median)	424	390
Contig length (std)	759	714
Contig number (bp)	733,197,014	543,207,784
Contig number (seq)	1,056,042	844,933
GC (%)	45.14	44.85

using the unmodified protocol showed substantially lower RIN values ranging from 2.4 to 6.1 (Table 1), indicating lower RNA integrity and greater degradation.

3.2. Sequencing Output and Read Quality

High-throughput Illumina sequencing generated a total of 57.6 GB of raw data, corresponding to the total uncompressed FASTQ file size obtained from the sequencing run. Sequencing statistics for each tissue type are summarized in Table 3. Leaf tissue generated approximately 234.8 million raw reads, while shoot and stem tissues produced approximately 275.3 and 267.6 million raw reads, respectively.

Following quality filtering, more than 98% of reads were retained as clean reads for all tissue types. Sequencing error rates remained low at approximately 0.03%, while Q20 and Q30 values consistently exceeded 96% and 90%, respectively. GC content ranged from 44.4% to 47.0%, which is consistent with values commonly reported for plant transcriptomes. The high percentage of clean reads and sequencing quality metrics demonstrated the suitability of the dataset for downstream *de novo* transcriptome assembly and analysis.

3.3. *De novo* Transcriptome Assembly

The *de novo* transcriptome assembly generated a large and comprehensive set of transcripts from the pooled *G. levis* tissues. Trinity assembly produced 1,056,042 contigs with a total assembled length of 733,197,014 bp, a mean contig length of 694 bp, an N50 value of 983 bp, and a GC content of 45.14% (Table 4). The assembled transcriptome also contained long transcript sequences, with maximum contig lengths reaching 19,016 bp, indicating broad transcript representation and overall assembly stability.

CD-HIT-EST clustering analysis was subsequently performed to reduce transcript redundancy and improve assembly compactness. The resulting non-redundant transcript dataset contained 844,933 contigs with a total assembled length of 543,207,784 bp, a mean contig length of 642 bp, an N50 value of 885 bp, and a GC content of 44.85%. The maximum contig length remained unchanged at 19,016 bp, indicating retention of long transcript sequences after redundancy reduction. Both the Trinity and Cluster Database at High Identity with Tolerance–Expressed Sequence Tag (CD-HIT-EST) transcript sets, datasets were subsequently used for transcriptome completeness evaluation and downstream analyses.

3.4. Assembly Completeness and Quality Assessment

BUSCO analysis using the embryophyta_odb12 dataset demonstrated high transcriptome completeness for both the Trinity and CD-HIT-EST transcript datasets (Fig. 4). Complete BUSCOs accounted for 90.7% and 90.6% of the Trinity and CD-HIT-EST assemblies, respectively. The Trinity assembly contained 84.3% duplicated BUSCOs, whereas the CD-HIT-EST assembly showed reduced duplicated BUSCOs at 74.8% and increased single-copy BUSCOs from 6.5% to 15.8%, indicating effective transcript redundancy reduction after clustering. Fragmented BUSCOs accounted for 7.4% in both transcript datasets, while missing BUSCOs remained low at 1.9% and 2.0% for the Trinity and CD-HIT-EST assemblies, respectively (Table 5). These results indicate that both transcriptome assemblies retained high biological completeness following redundancy reduction.

3.5. Public Deposition of the Transcriptome Dataset

All raw sequencing reads generated in this study have been deposited in the NCBI SRA under BioProject accession number PRJNA1034257. In addition, the Trinity transcript FASTA files,

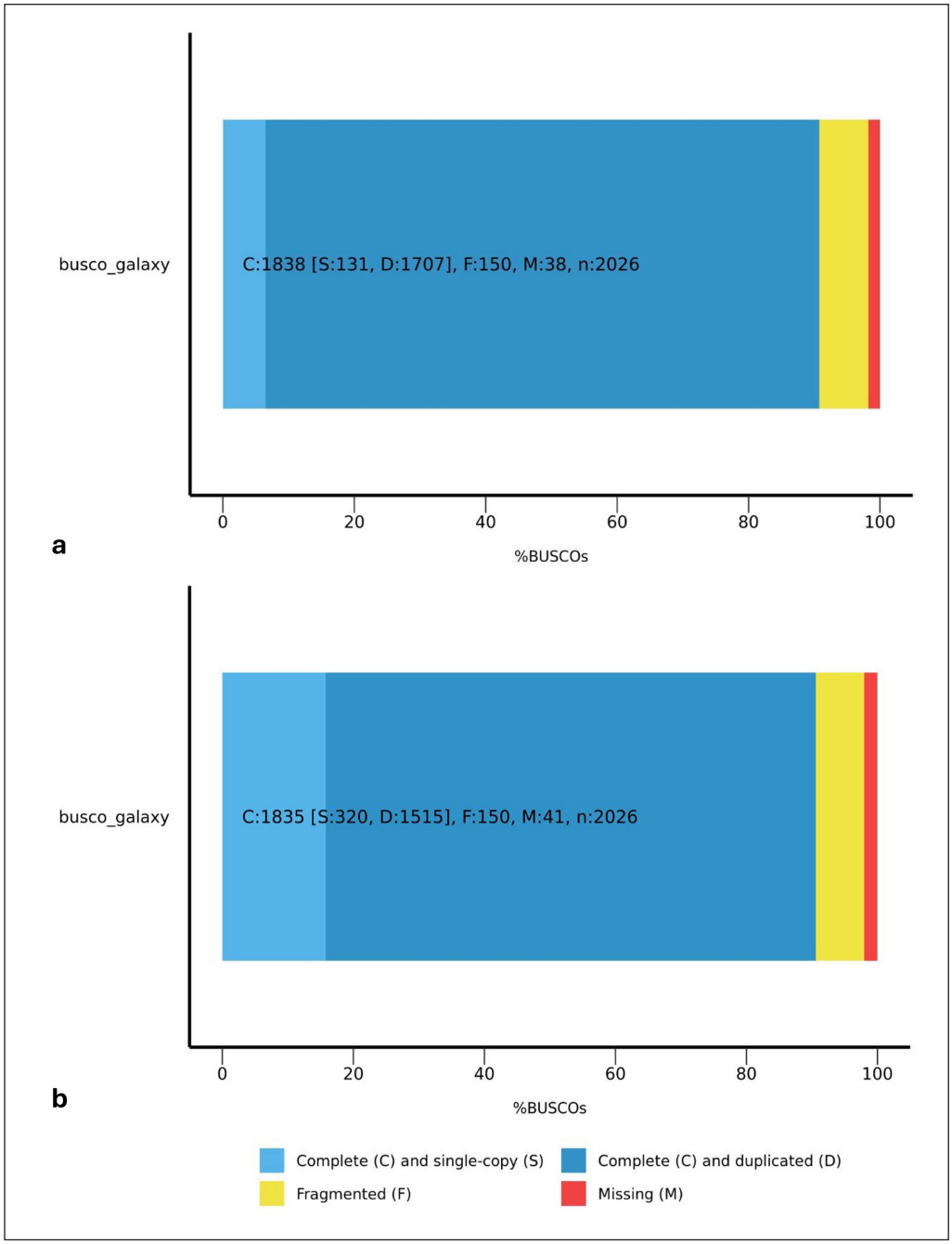

Figure 4. BUSCO completeness assessment of the Trinity (a) and CD-HIT-EST (b) *de novo* transcriptome assemblies of *G. levis* using the embryophyta_odb12 dataset in Galaxy v5.8.0 + galaxy2, showing percentages of complete single-copy, complete duplicated, fragmented, and missing BUSCOs.

Table 5. BUSCO analysis statistics for the Trinity and CD-HIT-EST transcriptome assemblies of *G. levis*.

Parameter	Statistics (Percentage)	
	Trinity transcripts	CD-HIT-EST transcripts
Complete BUSCOs	1,838 (90.7%)	1,835 (90.6%)
Complete and single-copy BUSCOs	131 (6.5%)	320 (15.8%)
Complete and duplicated BUSCOs	1,707 (84.3%)	1,515 (74.8%)
Fragmented BUSCOs	150 (7.4%)	150 (7.4%)
Missing BUSCOs	38 (1.9%)	41 (2.0%)
Total BUSCO groups searched	2026	2026

CD-HIT-EST non-redundant transcript FASTA files, CD-HIT-EST clustering outputs, and BUSCO assessment outputs for both transcript datasets have been deposited in the Figshare Repository. These publicly available datasets provide a transcriptomic resource for *G. levis* and support future functional genomics, comparative transcriptomics, and molecular investigations in bamboo and other lignocellulosic plant species.

4. DISCUSSION

The extraction of high-quality RNA from bamboo tissues remains technically challenging because bamboo contains high levels of polysaccharides, polyphenols, and lignin-associated secondary metabolites that interfere with nucleic acid isolation and downstream enzymatic reactions. These biochemical characteristics frequently compromise RNA purity and integrity in lignocellulosic plant species [31]. The optimized extraction strategy developed in this study successfully addressed these challenges and enabled the consistent recovery of high-integrity RNA from multiple *G. levis* tissues. Compared with the original protocol, the modified workflow produced higher RNA yields together with improved RNA integrity, as reflected by consistently high RIN values across all tissues examined. The ability to reproducibly isolate high-quality RNA is essential for transcriptome sequencing, particularly in recalcitrant plant species with complex biochemical compositions [32,33]. Therefore, the optimized protocol presented here provides a practical and reproducible framework for transcriptomic studies in bamboo and other lignified plant species.

The sequencing dataset generated in this study provided extensive transcriptome coverage for *G. levis*. High Q20 and Q30 scores, low sequencing error rates, and GC contents within the expected range for grass transcriptomes indicated that the sequencing data were of high quality and suitable for *de novo* transcriptome assembly [34,35]. In non-model plant species lacking reference genomes, sequencing quality and depth are critical for accurate transcript reconstruction and transcriptome completeness. The deep sequencing coverage achieved in this study likely improved the recovery of long transcript sequences and enhanced the representation of low-abundance transcripts and transcript variants. Such transcriptomic resources are important for advancing the understanding of regulatory pathways involved in plant growth, development, biomass accumulation, and responses to environmental stress.

The *de novo* transcriptome assembly generated a large and comprehensive transcript dataset for *G. levis*. Following CD-HIT-EST redundancy reduction, the non-redundant transcript dataset retained high assembly quality while reducing transcript duplication, thereby improving assembly compactness and downstream usability. The retention of long transcript sequences after clustering further indicated the robustness of the assembly process. BUSCO analysis additionally demonstrated high transcriptome completeness, with more than 90% complete BUSCOs identified in both the Trinity and CD-HIT-EST transcript datasets. These completeness values are comparable with those reported for transcriptome assemblies in other bamboo species and non-model plants [23,36]. Although duplicated BUSCOs remained relatively high, transcript duplication is commonly observed in *de novo* transcriptome assemblies and may reflect transcript isoforms, gene family expansions, and alternative splicing events frequently reported in grasses [37,38]. The reduction in duplicated BUSCOs and corresponding increase in single-copy BUSCOs following CD-HIT-EST clustering indicate that redundancy reduction effectively improved assembly representation while preserving biological completeness.

Despite increasing interest in bamboo biotechnology and sustainable biomaterials, molecular resources for many tropical bamboo species remain limited. Existing genomic and transcriptomic studies have focused primarily on a relatively small number of bamboo taxa, restricting broader molecular investigations and comparative analyses [19,25]. The transcriptomic resource generated in this study, therefore, contributes additional molecular data for *G. levis* and expands the available genomic resources for tropical bamboo research. Although functional annotation and pathway-level analyses were beyond the scope of the present study, the transcriptome dataset provides a valuable foundation for future investigations involving gene expression profiling, comparative transcriptomics, and candidate gene discovery [39,40]. In particular, the dataset may facilitate future exploration of genes potentially associated with lignin biosynthesis, cellulose deposition, fiber formation, stress adaptation, and other traits relevant to bamboo growth and industrial utilization.

The publicly accessible datasets generated in this study, together with the optimized RNA extraction workflow, provide a useful platform for future plant biotechnology and functional genomics research in bamboo and other recalcitrant lignocellulosic species. In the context of sustainable bioresource development and Industry 5.0 initiatives, bamboo is increasingly recognized as a renewable and engineered biomaterial suitable for low-emission and circular bioeconomy applications. Consequently, the availability of reliable transcriptomic resources will support future efforts in molecular breeding, sustainable biomass utilization, biomaterial innovation, and climate-resilient crop development.

5. CONCLUSION

This study presents a *de novo* transcriptome resource for *G. levis*, contributing valuable molecular data to the currently limited genomic resources available for tropical bamboo species. The generation of a high-quality and publicly accessible transcriptomic dataset establishes an important molecular foundation for future studies on gene function, developmental regulation, biomass formation, and environmental adaptation in this economically and ecologically important bamboo species. In addition, the optimized RNA extraction workflow and transcriptome assembly pipeline developed in this study provide a practical and reproducible framework for transcriptomic investigations in other recalcitrant bamboo and lignocellulosic plant species. Collectively, this work supports future advances in bamboo functional genomics, comparative transcriptomics, and plant biotechnology, while contributing to sustainable biomass utilization, biomaterial innovation, and the long-term development of bamboo-based bioresources and industries.

6. LIST OF ABBREVIATIONS

BUSCO, Benchmarking universal single-copy orthologs; CD-HIT-EST, Cluster Database at High Identity with Tolerance–Expressed Sequence Tag; FASTQ, A file format storing sequencing reads and their quality scores; GB, Gigabyte; GC, Guanine–cytosine; N50, A measure of contig length representing the length above which 50% of the total assembly is contained; N90, A measure of contig length representing the length above which 90% of the total assembly is contained; NCBI, National Center for Biotechnology Information; Q20, Quality score indicating a 1% error rate; Q30, Quality score indicating a 0.1% error rate; RIN, RNA integrity number; SRA, Sequence read archive.

7. ACKNOWLEDGMENT

The authors acknowledge the Malaysian Ministry of Higher Education for funding this project through the Fundamental Research Grant

Scheme (Grant No. FRGS/1/2020/STG01/UMS/02/4). The authors also thank the Sabah Biodiversity Centre (SaBC) for granting permission to collect samples under Access License No. JKM/MBS.1000-2/2 JLD.17(134).

8. AUTHOR CONTRIBUTIONS

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work. All the authors are eligible to be author as per the International Committee of Medical Journal Editors (ICMJE) requirements/guidelines.

9. CONFLICTS OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Although one of the authors, Julius Kulip, is affiliated with Botanicals Sabah Sdn. Bhd., Kota Kinabalu, Malaysia, this affiliation had no direct or indirect influence on the study design, experimental work, results or manuscript preparation.

10. ETHICAL APPROVALS

This study does not involve experiments on animals or human subjects.

11. PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of the publisher, the editors and the reviewers. This journal remains neutral with regard to jurisdictional claims in published institutional affiliation.

12. DATA AVAILABILITY

All sequencing data have been deposited in the BioProject database of the SRA at the NCBI under accession number PRJNA1034257 and are accessible at: https://www.ncbi.nlm.nih.gov/sra/PRJNA1034257. In addition, the Trinity assembled transcript FASTA files, CD-HIT-EST transcript FASTA files, CD-HIT-EST clustering output files, and BUSCO assessment output files for both the Trinity and CD-HIT-EST transcript datasets have been deposited in Figshare and are publicly accessible at: https://doi.org/10.6084/m9.figshare.32231229.

13. USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors confirm that the manuscript was written by the authors without the use of AI tools for scientific content generation. However, AI-based language tools, including QuillBot and Grammarly, were used solely to improve English clarity, grammar, and readability to ensure effective communication of the manuscript to the scientific community. All content was carefully reviewed, verified, and approved by the authors, who take full responsibility for the accuracy, integrity, and originality of the work.

REFERENCES

1. Ahmad Z, Kumari R, Mir B, Saeed T, Firdaus F, Vijayakanth V, *et al.* Bamboo for the future: from traditional use to Industry 5.0 applications. Plants 2025;14(19):3019; doi: https://doi.org/10.3390/plants14193019
2. Pan C, Zhou G, Shrestha AK, Chen J, Kozak R, Li N, *et al.* Bamboo as a nature-based solution (NbS) for climate change mitigation: biomass, products, and carbon credits. Climate 2023;11(9):175; doi: https://doi.org/10.3390/cli11090175
3. Patel HR, Mathakia R, Mangroliya UC, Mandaliya VB. Sustainable bamboo: technological innovations and patent insights for a greener future. Adv Bamboo Sci 2025;10:100127; doi: https://doi.org/10.1016/j.bamboo.2025.100127
4. Mazzoni F, Binz C, Losacker S. What can the bamboo industry learn from timber? Resource mobilization across global innovation systems in the construction sector. Geoforum 2025;165:104378; doi: https://doi.org/10.1016/j.geoforum.2025.104378
5. Wang N, Wang W, Zhu Q. Unlocking the genetic blueprint of bamboo for climate adaption. Trends Plant Sci 2025;30(7):693–5; doi: https://doi.org/10.1016/j.tplants.2025.02.007
6. Wang W, Wu Q, Wang N, Ye S, Wang Y, Zhang J, *et al.* Advances in bamboo genomics: growth and development, stress tolerance, and genetic engineering. J Integr Plant Biol 2025;67(7):1725–55; doi: https://doi.org/10.1111/jipb.13909
7. Osman S, Ahmad M, Zakaria MN, Zakaria AM, Ibrahim Z, Abu F, *et al.* Bamboo as future bio-industrial material: physical behaviour and bending strength of Malaysia's Beting bamboo (*Gigantochloa levis*). In IOP Conf Ser Earth Environ Sci, 2022;951:012001; doi: https://doi.org/10.1088/1755-1315/951/1/012001
8. Xu ZC, Liu JX, Li DZ. *Gigantochloa glabrata* (Poaceae, Bambusoideae), a new bamboo species from Yunnan, China. PhytoKeys 2021;171:37–45; doi: https://doi.org/10.3897/phytokeys.171.59562
9. Long L, Minghui Y, Wenjing Y, Yulong D, Shuyan L. Research advance in growth and development of bamboo organs. Ind Crops Prod 2023;205:117428; doi: https://doi.org/10.1016/j.indcrop.2023.117428
10. Liu JX, Guo C, Ma PF, Zhou MY, Luo YH, Zhu GF, *et al.* The origin and morphological character evolution of the paleotropical woody bamboos. J Integr Plant Biol 2024;66(10):2242–61; doi: https://doi.org/10.1111/jipb.13751
11. Baharuddin N, Abdul Karim SR, Kassim AS, Osman Al-Edrus SS, Lee SH. Global bamboo industries: an overview. In: Md Tahir P, Lee SH, Osman Al-edrus SS, Uyup MKA (eds). Multifaceted bamboo: engineered products and other applications, Singapore, Springer, p 15–41, 2023; doi: https://doi.org/10.1007/978-981-19-9327-5_2
12. Chaudhary U, Malik S, Rana V, Joshi G. Bamboo in the pulp, paper and allied industries. Adv Bamboo Sci 2024;7:100069; doi: https://doi.org/10.1016/j.bamboo.2024.100069
13. Li Y, Wang N, Abdul Latiff AR. Development of the bamboo forest economy: reviewing China's 'bamboo as a substitute for plastic initiative' and its development. Adv Bamboo Sci 2025;11:100130; doi: https://doi.org/10.1016/j.bamboo.2025.100130
14. Bhol N, Parida S, Mallick P, Mahapatra SR, Sharma U, Das J, *et al.* Global bamboo sector: status, critical challenges and pathways for future sustainable development. Int J Multidiscip Res 2026;8(2):260272726.
15. Perumal M, Mohd Hassan NH, Abdullah N, Omar L, Zainudin J, Wasli ME. Survivorship and field growth characteristics of four selected bamboo species for the development of bamboo industry in Sarawak, Malaysian Borneo. Malaysian Appl Biol 2023;52(5):155–62; doi: https://doi.org/10.55230/mabjournal.v52i5.icfic08
16. Mustapha A, Yusoh AS, Chee Beng O, Awalludin MF, Sukri NANM, Uyup MKA. Exploring the potential of *Gigantochloa levis* and *Gigantochloa scortechinii* bamboo species for plybamboo production. Bioresources 2024;19(4):7434–49; doi: https://doi.org/10.15376/biores.19.4.7434-7449
17. Osman S, Ahmad M, Zakaria MN, Bakar BFA, Abu F, Kamarudin SH, *et al.* Variation of chemical properties, crystalline structure and calorific values of native Malaysian bamboo species. Wood Fiber Sci 2022;54(3):173–86; doi: https://doi.org/10.22382/wfs-2022-17

18. Zakikhani P, Zahari R, Bin Haji Hameed Sultan MT, Abang Abdul Majid DL. Morphological, mechanical, and physical properties of four bamboo species. BioResources 2017;12(2):2479–95; doi: https://doi.org/10.15376/biores.12.2.2479-2495
19. Basak M, Dutta S, Biswas S, Chakraborty S, Sarkar A, Rahaman T, *et al.* Genomic insights into growth and development of bamboos: what have we learnt and what more to discover?. Trees 2021;35:1771–91; doi: https://doi.org/10.1007/s00468-021-02197-6
20. Zhang RL, Li DZ, Zhang YX. Population genetic structure and climate adaptation analysis of *Brachystachyum densiflorum*. Chin Bull Bot 2025;60(3):407–24; doi: https://doi.org/10.11983/CBB24094
21. Li L, Zhou B, Liu D, Wu H, Shi Q, Lin S, *et al.* Transcriptomic complexity of culm growth and development in different types of moso bamboo. Int J Mol Sci 2023;24(8):7425; doi: https://doi.org/10.3390/ijms24087425
22. Qian Y, Jia J, Chen Z, Wang K, Li P, Gao P, *et al.* Environmental drivers and transcriptomic variations shaping Lei bamboo shoots across cultivation regions. Front Plant Sci 2025;16:1565665; doi: https://doi.org/10.3389/fpls.2025.1565665
23. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform 2022;23(2):563; doi: https://doi.org/10.1093/bib/bbab563
24. Biswas S, Rahaman T, Gupta P, Mitra R, Dutta S, Kharlyngdoh E, *et al.* Cellulose and lignin profiling in seven, economically important bamboo species of India by anatomical, biochemical, FTIR spectroscopy and thermogravimetric analysis. Biomass Bioenergy 2022;158:106362; doi: https://doi.org/10.1016/j.biombioe.2022.106362
25. Di X, Zou X, Wang Q, Sun H. Development and future prospects of bamboo gene science. Int J Mol Sci 2025;26(15):7259; doi: https://doi.org/10.3390/ijms26157259
26. Liu Y, Wu C, Hu X, Gao H, Wang Y, Luo H, *et al.* Transcriptome profiling reveals the crucial biological pathways involved in cold response in Moso bamboo (*Phyllostachys edulis*). Tree Physiol 2020;40(4):538–56; doi: https://doi.org/10.1093/treephys/tpz133
27. Rai V, Ghosh JS, Dey N. Isolation of total RNA from hard bamboo tissue rich in polyphenols and polysaccharides for gene expression studies. Electron J Biotechnol 2010;13(5):22–3; doi: https://doi.org/10.2225/vol13-issue5-fulltext-17
28. Yong WTL, Mustafa AA, Derise MR, Rodrigues KF. DNA barcoding using chloroplast matK and rbcL regions for the identification of bamboo species in Sabah. Adv Bamboo Sci 2024;7:100073; doi: https://doi.org/10.1016/j.bamboo.2024.100073
29. Pola-Sánchez E, Hernández-Martínez KM, Pérez-Estrada R, Sélem-Mójica N, Simpson J, Abraham-Juárez MJ, *et al.* RNA-seq data analysis: a practical guide for model and non-model organisms. Curr Protoc 2024;4(5):e1054; doi: https://doi.org/10.1002/cpz1.1054
30. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;8(8):1494–512; doi: https://doi.org/10.1038/nprot.2013.084
31. Carpinetti PDA, Fioresi VS, Ignez Da Cruz T, De Almeida FAN, Canal D, Ferreira A, *et al.* Efficient method for isolation of high-quality RNA from *Psidium guajava* L. tissues. PLoS One 2021;16(7):255245; doi: https://doi.org/10.1371/journal.pone.0255245
32. Sasi S, Krishnan S, Kodackattumannil P, Shamisi AA, Aldarmaki M, Lekshmi G, *et al.* DNA-free high-quality RNA extraction from 39 difficult-to-extract plant species (representing seasonal tissues and tissue types) of 32 families, and its validation for downstream molecular applications. Plant Methods 2023;19:84; doi: https://doi.org/10.1186/s13007-023-01063-5
33. Tyagi P, Singh D, Mathur S, Singh A, Ranjan R. Upcoming progress of transcriptomics studies on plants: an overview. Front Plant Sci 2022;13:1030890; doi: https://doi.org/10.3389/fpls.2022.1030890
34. Singh R, Ming R, Yu Q. Comparative analysis of GC content variations in plant genomes. Trop Plant Biol 2016;9:136–49; doi: https://doi.org/10.1007/s12042-016-9165-4
35. Zhao D, Hamilton JP, Hardigan M, Yin D, He T, Vaillancourt B, *et al.* Analysis of ribosome-associated mRNAs in rice reveals the importance of transcript size and GC content in translation. G3 Genes Genomes Genet 2017;7(1):203–19; doi: https://doi.org/10.1534/g3.116.036020
36. Zhao H, Gao Z, Wang L, Wang J, Wang S, Fei B, *et al.* Chromosome-level reference genome and alternative splicing atlas of moso bamboo (*Phyllostachys edulis*). GigaScience 2018;7(10):115; doi: https://doi.org/10.1093/gigascience/giy115
37. Liu G, Wang H, Gao H, Yu S, Liu C, Wang Y, *et al.* Alternative splicing of functional genes in plant growth, development, and stress responses. Int J Mol Sci 2025;26(12):5864; doi: https://doi.org/10.3390/ijms26125864
38. Shen Y, Qin Z, Ren G, Deng P, Ji W, Jiao C, *et al.* Complexity and regulation of age-dependent alternative splicing in *Brachypodium distachyon*. Plant Physiol 2023;192(4):2703–22; doi: https://doi.org/10.1093/plphys/kiad223
39. Huang Z, Zhu P, Zhong X, Qiu J, Xu W, Song L. Transcriptome analysis of moso bamboo (*Phyllostachys edulis*) reveals candidate genes involved in response to dehydration and cold stresses. Front Plant Sci 2022;13:960302; doi: https://doi.org/10.3389/fpls.2022.960302
40. Liu Y, Zhu C, Lin Z, Li H, Di X, Yue X, *et al.* Systematic identification and validation of the reference genes from 447 transcriptome datasets of moso bamboo (*Phyllostachys edulis*). Hortic Plant J 2025;11(3):1353–63; doi: https://doi.org/10.1016/j.hpj.2023.11.007

How to cite this article:
Derise MR, Mustafa AA, Rodrigues KF, Kulip J, Maliki IM, Yew CW, Yong WTL. *De novo* transcriptome assembly and RNA quality optimization in *Gigantochloa levis* using Illumina sequencing. J Appl Biol Biotech 2026;14(5):132-141. DOI: 10.7324/JABB.2026.315063