Bangladesh J. Plant Taxon. **31**(2): 225-238, 2024 (December) DOI: https://doi.org/10.3329/bjpt.v31i2.78750

GENOMIC ANALYSIS MADE EASY (GAME V1): AN AUTOMATED SOFTWARE FOR PLANT GENOME ASSEMBLY AND ANNOTATION FROM ILLUMINA SEQUENCING

MOHAMMAD AJMAL ALI[1,*], RAJESH MAHATO[2] AND JOONGKU LEE[3]

[1]*Department of Botany and Microbiology, College of Science, King Saud University, Riyadh 11451, Saudi Arabia.*
[2]*ArrayGen Technologies Private Limited, Undri, Pune-411060, Maharashtra, India.*
[3]*Department of Environment and Forest Resources, College of Agricultural Life Science, Chungnam National University, Daejeon, South Korea.*

Keywords: GAME v1; Python; Short reads; Plant genome; Assembly; Annotation.

Abstract

The recent development and affordable accessibility of the next-generation high-throughput sequencing technology and artificial intelligence have propelled more researchers to get involved in genomics and to the threshold of a new beginning in understanding, utilizing, and conserving the biodiversity. However, one of the biggest challenges for the analysis of high-throughput sequencing reads is the whole genome assembly and annotation. Availability of user-friendly free software that manages all types of sequenced DNA to be used in a local environment is lacking. Hence, the Genomic Analysis Made Easy (GAME v1) software has been developed using Python to provide a user-friendly, fast, free, and automated GUI-based solution for plant genome assembly and annotation. The software performs on a Linux-based operating environment with a minimum of 16 GB RAM and 100 GB disc space, fully automated from the installation to execution, thus requiring minimal bioinformatics expertise for the execution. The GAME v1 generates detailed quality reports of the raw reads, GenomeScope heterozygosity report, QUAST contigs and scaffolds results, BUSCO summary plot, COG functional annotation chart, GO chart, NCBI and UniProt annotations, KEGG pathway distribution graph, and RepeatMasker plots. The nuclear genome of *Chenopodium pallidicaule* retrieved from NCBI was assembled and annotated successfully using GAME v1, revealing preliminary genome size estimate of around 419.54 Mb based on GenomeScope analysis prior to assembly. The final assembly, as assessed by QUAST, unveiled a total length of 285.85 Mb (0.29 Gb) containing 23,806 genes. This automated solution will facilitate plant genomics research by revealing the underlying insights of draft nuclear genomes. The software is available at https://arraygen.com/game

Introduction

The time-scale has witnessed tremendous changes in nucleic acid sequencing technology, moving from first generation of sequencing used for sequencing of a short oligonucleotides to next generation high-throughput sequencing used for the whole genome and transcriptome analysis (Satam *et al.*, 2023). Advances in high-throughput sequencing technologies, including long-read platforms like Oxford Nanopore and Pacific Biosciences, as well as cost-effective short-read platforms like Illumina sequencing, have significantly accelerated genome sequencing efforts over the past decade. These advancements have been complemented by the development of sophisticated computational tools (Tian *et al.*, 2024), the integration of automation and artificial intelligence (Caudai *et al.*, 2021), and the enhanced availability of high-quality genomic resources, such as chromosome-level de novo genome assemblies (Shirasawa *et al.*, 2021) and detailed genome annotations (Ejigu *et al.*, 2020).

*Corresponding author: alimohammad@ksu.edu.sa

Together, these innovations have ushered in a new era of genomics, enabling ambitious projects like the Open Green Genomes initiative (https://phytozome-next.jgi.doe.gov/) and the Earth BioGenome Project (https://www.earthbiogenome.org/) to sequence the genomes of all complex life on Earth.

In genomics, generating high-quality genome assemblies and annotations has become essential for understanding the biology of any species (Jung *et al.*, 2020). However, existing software for genome assembly and annotation often demands high-performance hardware (Kathiresan *et al.*, 2017), advanced bioinformatics expertise (Helmy *et al.*, 2016), and costly subscriptions (El-Metwally *et al.*, 2013). Each tool has its unique strengths and limitations depending on the specific application (Rice and Green, 2019). Despite this, selecting open-access computational tools capable of managing diverse DNA sequence types in local environments with minimal hardware requirements and reduced execution times remains a significant challenge (Kathiresan *et al.*, 2017). This issue is particularly pronounced for scientists in developing countries, where limited computational resources hinder participation in the genomic revolution (Helmy *et al.*, 2016).

To address these barriers, Genomic Analysis Made Easy (GAME v1) has been developed. GAME v1 is an innovative, user-friendly, and fully automated GUI-based software designed to simplify the complex and resource-intensive process of plant genome assembly and annotation. Built using Python, GAME v1 combines speed, efficiency, and accessibility, making it an ideal tool for researchers with limited computational resources or bioinformatics expertise. By integrating multiple steps into a single streamlined workflow, the software minimizes the technical challenges typically associated with genome analysis. GAME v1 is available for free, ensuring that scientists, particularly having resource-constrained settings, can actively participate in cutting-edge genomic research without the need for expensive hardware or paid subscriptions. With its intuitive interface and robust performance, GAME v1 empowers researchers to focus on scientific discovery while democratizing access to advanced genomic tools.

Materials and Methods

GAME v1 was developed on a Linux platform using Python to create an automated, user-friendly pipeline for plant genome assembly and annotation from Illumina sequencing data. It integrates numerous bioinformatics tools to streamline complex genomic workflows. The apt and pip3 packages ensure seamless installation and management of dependencies, while Java facilitates the execution of cross-platform tools used in various pipeline stages. For sequence quality assessment and preprocessing, FastQC evaluates raw read quality, and fastp performs adapter trimming and quality filtering to ensure high-quality inputs for assembly (Jung *et al.*, 2020). The GATB Minia Pipeline was chosen as the primary genome assembler due to its high performance with low-memory requirements, making it particularly suitable for resource-constrained environments (Drezen *et al.*, 2014). While assemblers such as MaSuRCA or SPAdes offer robust assembly capabilities, their higher computational demands and memory footprints can be prohibitive for researchers with limited hardware resources. GATB Minia excels in efficiently assembling genomes from large Illumina datasets, aligning with the objectives of GAME v1 to provide a lightweight and accessible solution (https://github.com/GATB/gatb-minia-pipeline). Tools like Kmergenie optimize k-mer selection, a critical step in improving assembly accuracy, and QUAST evaluates the quality of assembled contigs through metrics such as N50 and misassemblies (Gurevich *et al.*, 2013). For sequence alignment, BWA maps reads to reference sequences, supporting downstream analysis. Functional annotation is carried out using Prodigal for gene prediction, Augustus for advanced gene model prediction, and BUSCO, which assesses

the completeness of genome assemblies based on conserved single-copy orthologs (Simão *et al.*, 2015).

RepeatMasker identifies and masks repetitive DNA elements, improving the annotation process (Tarailo-Graovac and Chen, 2009). For homology-based functional annotation, Diamond, HMMER (hmmscan), and MetaEuk enable rapid and sensitive protein sequence comparisons. Visualization and statistical analysis are conducted using ggplot2 in R, which generates high-quality plots for Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analyses, providing functional insights (Wickham, 2011). Supporting tools like Git ensure version control and collaboration during software development, while CMake automates the compilation and build process for integrated tools (Perez-Riverol *et al.*, 2016). Utilities like gzip and curl handle data compression and retrieval, facilitating efficient processing of large genomic datasets. BLASTn performs nucleotide sequence similarity searches. Validated with *Chenopodium pallidicaule* Aellen (SRR4425239) as a test case, GAME v1 integrates these tools into a single pipeline, reducing computational barriers and enabling assembly and annotation of plant genomes efficiently, regardless of prior bioinformatics expertise. The Debian package file of the software can be download from https://arraygen.com/game, and extracted followed by command “sudo dpkg -i GAME.deb” in a new terminal to install. The installation time depends on the speed of the internet.

Results and Discussion

The minimum system requirements for running GAME v1 include Ubuntu 22.04.4 LTS operating system (64-bit), 16 GB of RAM, a processor with at least 4 cores, and a minimum of 2.0 TB hard disk space. However, these requirements may increase based on the size of the raw data being processed. Larger datasets (raw data) demand additional RAM and disk space to ensure optimal execution by GAME v1. Installation and testing were conducted on a Lenovo ThinkStation C30 equipped with an Intel® Xeon® E5-2620 CPU @ 2.00 GHz (12 cores), 128 GB of RAM, an 8.0 TB disk, and NV106 graphics, running Ubuntu 22.04.4 LTS (GNOME version 42.9) with the Wayland windowing system. Under these conditions, installation of the software and all dependencies over a 100 Mbps internet connection required approximately two hours. The successful operation of GAME v1 is contingent upon the proper installation of all dependent tools, including those necessary for genome assembly, annotation, and downstream analyses. Once installation is complete, users can launch GAME v1 by specifying the project path and entering the desired project name in the intuitive welcome interface (Fig. 1A). This streamlined setup ensures a seamless start to genome analysis workflows, and generates detailed quality reports of the raw reads, GenomeScope heterozygosity report, QUAST contigs and scaffolds results, BUSCO summary plot, COG functional annotation chart, GO chart, NCBI and UniProt annotations, KEGG pathway distribution graph, and RepeatMasker plots.

A total of 46.1 GB of paired-end short-read data (*C. pallidicaule*, SRR4425239) was retrieved from the NCBI SRA database. The FASTQ files, containing forward and reverse reads, were seamlessly uploaded into GAME v1 using the software's "Browse" option, enabling efficient integration into the genome assembly workflow (Fig. 1B), and it was executed with the tool setting (Fig. 1C). Under the tools setting, the number of threads were provided according to 10 available threads in the CPU (Fig. 1C). On execution, the software first checked the installation of the dependent tools. The successful execution competed in 17 hours, and generated all the results in various HTML, PDF and Excel files. The results of quality control analysis are shown in Table 1. Before filtering, the dataset comprised 389.56 million reads totaling 38.96 billion bases, with 96.86% of bases having a quality score of Q20 or higher and 90.72% achieving Q30 or higher.

The GC content of the raw data was 38.53%. After filtering, the dataset was reduced to 350.65 million reads with 35.03 billion bases, indicating the removal of low-quality reads and adapters. The proportion of Q20 bases increased to 98.82%, while Q30 bases rose to 94.55%, reflecting a significant improvement in overall sequence quality. The GC content post-filtering was slightly reduced to 38.21%, indicating the maintenance of sequence composition integrity.

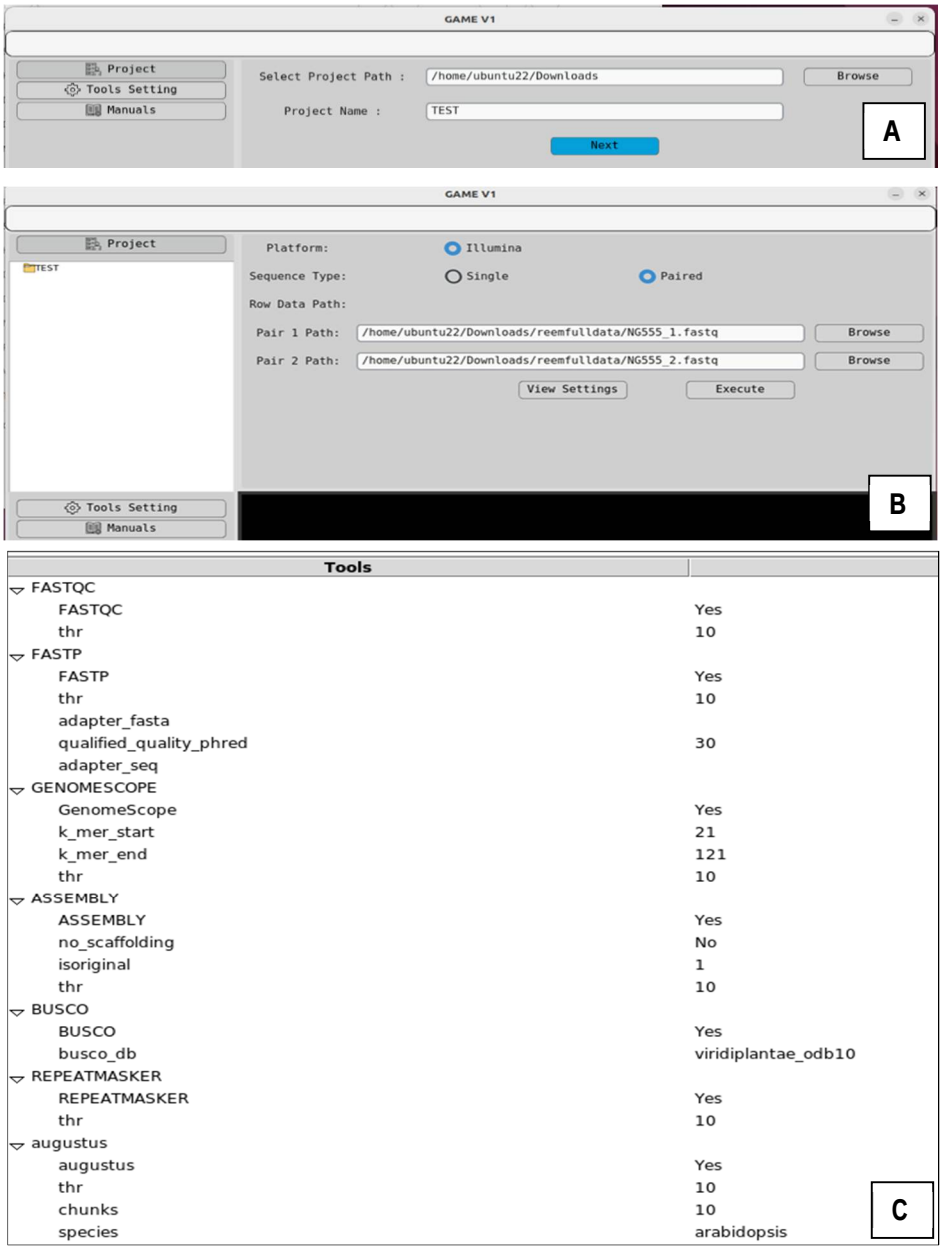

Fig. 1. GAME v1 welcome interface and tools setting. A. Interface of the GAME v1. B. Uploading of the sequence data. C. Details of the tools setting.

These results demonstrate the efficiency of fastp in enhancing data quality for downstream genome assembly and analysis. The filtering process retained 350.65 million reads, accounting for 90.01% of the total input reads, ensuring a high-quality dataset for downstream analysis. A total of 38.90 million reads (9.99%) were removed due to low quality, while 9,220 reads (0.0024%) were excluded because they contained excessive ambiguous bases (Ns). Notably, no reads were discarded for being too short, reflecting the robustness of the dataset's original length distribution. These results highlight the effectiveness of the filtering process in maintaining high-quality reads while minimizing data loss.

Table 1. Quality control analysis of the paired-end Illumina reads generated by fastp tool prebuilt in GAME v1.

QC analysis fastp report	
Summary	
fastp version	0.20.1 (https://github.com/OpenGene/fastp)
Sequencing	Paired end (100 cycles + 100 cycles)
Mean length before filtering	99bp, 99bp
Mean length after filtering	99bp, 99bp
Duplication rate	3.636400%
Insert size peak	169
Before filtering	
Total reads	389.563646 M
Total bases	38.955270 G
Q20 bases	37.732166 G (96.860233%)
Q30 bases	35.338756 G (90.716238%)
GC content	38.531857%
After filtering	
Total reads	350.650934 M
Total bases	35.031077 G
Q20 bases	34.616826 G (98.817475%)
Q30 bases	33.120725 G (94.546695%)
GC content	38.210978%
Filtering result	
Reads passed filters	350.650934 M (90.011206%)
Reads with low quality	38.903492 M (9.986428%)
Reads with too many N	9.220000 K (0.002367%)
Reads too short	0 (0.000000%)

The genomic properties of *C. pallidicaule* were assessed using the GenomeScope model with a k-mer size of 83, following trimming of forward reads to remove adapter sequences and poly-G artifacts. The analysis revealed a low heterozygosity rate, ranging from 0.0642% to 0.0660%, suggesting that the genome is predominantly homozygous (Table 2, Fig. 2). Such low heterozygosity levels often indicate a history of inbreeding, which can arise due to self-pollination, small population size, or geographic isolation. Alternatively, it may signify strong selective pressures that favor genomic stability by purging deleterious alleles and conserving adaptive traits.

This genetic uniformity could enhance the species' fitness in its specific ecological niche. It may also reduce its ability to adapt to rapidly changing environmental conditions or emerging stressors. Understanding heterozygosity is thus crucial for conservation strategies and for elucidating the evolutionary processes shaping the genetic makeup of *C. pallidicaule* (Ellestad *et al.*, 2022). The estimated haploid genome length was approximately 419.33 to 419.54 Mb, with a repeat length ranging from 102.62 to 102.67 Mb, highlighting the repetitive regions that play a key role in structural and functional aspects of the genome. The unique genome length, spanning 316.71 to 316.87 Mb, provides a measure of the non-redundant regions that are essential for understanding the functional genomics of this species. These unique sequences likely encode genes critical for adaptation, stress response, and metabolic processes specific to *C. pallidicaule*. The model fit, ranging from 97.99% to 99.35%, underscores the robustness of the k-mer-based analysis, ensuring that the genome features are accurately captured. Additionally, the extremely low read error rate of 0.1068% reflects the high quality of the input sequencing data, further validating the reliability of the estimates. In a recent study, *de novo* assembly of *Gaultheria prostrata* Kalm *ex* L. (Ericaceae) nuclear genome showed model error rate of about 0.493% (Lin *et al*, 2024). In our study, GAME v1 successfully executed GenomeScope and the model error rate was much lower than that of *G. prostrata*, further validating the GAME v1 execution protocol. Error rates below 0.5% are generally regarded as acceptable for large genome assemblies, as they minimize the risk of misrepresentation in genomic features such as repetitive regions or structural variants. Our error rate, which is nearly five times lower, reinforces the reliability of our assembly and confirms that it meets, and indeed surpasses, industry and research standards for *de novo* genome projects (Lin *et al.*, 2024). The insights into heterozygosity and repeat content may aid in designing strategies for assembling repetitive regions and identifying potential genomic markers (Dai *et al.*, 2016).

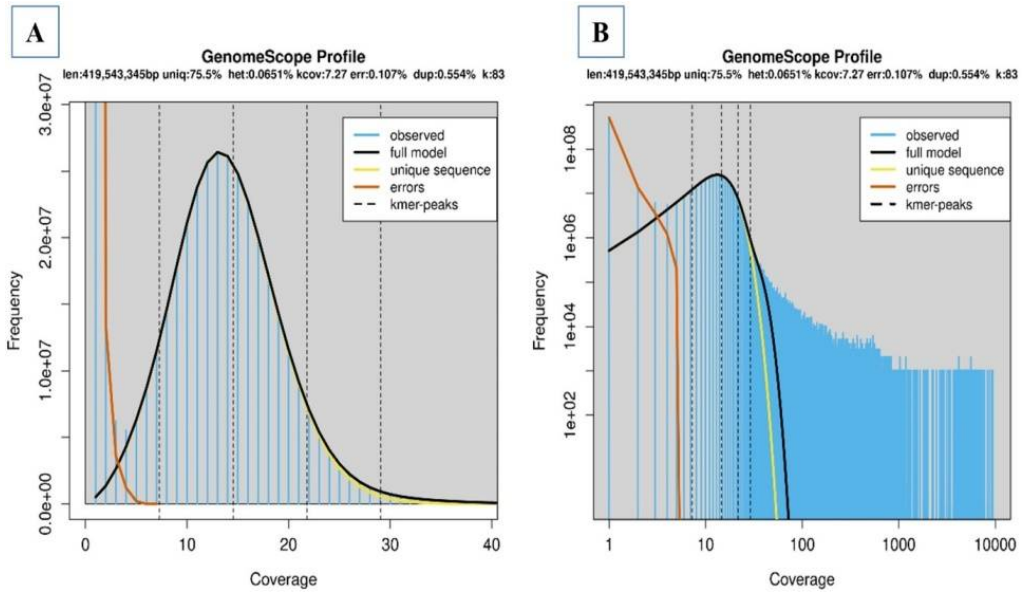

Fig. 2. GenomeScope profile of the sequenced Illumina paired-end reads of *C. pallidicaule*. A. K-mer frequency ranging upto 3.0×10^7. B. K-mer frequency ranging upto 1.0×10^9.

GATB minia pipeline was able to generate the assembled contigs and scaffolds files successfully. QUAST evaluation of the contigs revealed significant insights on the assembled contigs (Table 3). A total of 341,893 contigs were generated, of which 58,302 contigs were longer

than 1,000 bp, 17,041 contigs exceeded 5,000 bp, and 5,852 contigs were greater than 10,000 bp in length. Additionally, 315 contigs were longer than 25,000 bp, and 3 contigs exceeded 50,000 bp. The total length of the assembled sequences was 320,381,369 bp, with 261,461,981 bp in contigs ≥ 1,000 bp and 166,570,318 bp in contigs ≥ 5,000 bp. The largest contig assembled was 53,042 bp, and the assembly had a GC content of 36.27%.

Table 2. Genomic estimates of *C. pallidicaule* genome using GenomeScope module of GAME v1.

Properties	Minimum	Maximum
Heterozygosity	0.0642223%	0.0659812%
Genome Haploid Length	419,333,197 bp	419,543,345 bp
Genome Repeat Length	102,622,500 bp	102,673,929 bp
Genome Unique Length	316,710,697 bp	316,869,415 bp
Model Fit	97.9918%	99.3459%
Read Error Rate	0.106822%	0.106822%

Table 3. Evaluation of contigs and scaffolds assembled via GATB Minia pipeline inside GAME v1.

Characteristics	Contigs	Scaffolds
Contigs/Scaffolds (≥ 0 bp)	341893	270380
Contigs/Scaffolds (≥ 1000 bp)	58302	46511
Contigs/Scaffolds (≥ 5000 bp)	17041	16906
Contigs/Scaffolds (≥ 10000 bp)	5852	7963
Contigs/Scaffolds (≥ 25000 bp)	315	1130
Contigs/Scaffolds (≥ 50000 bp)	3	70
Total length (≥ 0 bp)	320381369	316150614
Total length (≥ 1000 bp)	261461981	272974537
Total length (≥ 5000 bp)	166570318	204363590
Total length (≥ 10000 bp)	88104240	140718618
Total length (≥ 25000 bp)	9669339	38029954
Total length (≥ 50000 bp)	155557	4196083
Contigs/Scaffolds	85317	64228
Largest Contig/Scaffold	53042	96620
Total length	281042957	285845886
GC (%)	36.27	36.33
N50	6394	9805
N90	1246	1717
auN	8339.5	12948.7
L50	12442	8186
L90	50659	34531
N's per 100 kbp	0.00	18.89

The N50 value, a critical metric for evaluating genome assembly quality, provides insight into the contiguity of assembled sequences. For *C. pallidicaule*, the N50 value was recorded at 6,394 bp, notably higher than the N50 of *Cymbopogon citratus*, which measured 4,347 bp (Chakravartty and Neelapu, 2024). This comparison highlights that the *C. pallidicaule* genome assembly possesses greater contiguity, indicating that its sequences are, on average, longer and potentially more complete. Higher N50 values are significant as they often reflect better assembly

performance, facilitating downstream analyses such as gene prediction and structural annotation. Conversely, the lower N50 value of *C. citratus* suggests that its assembly may have more fragmented sequences, possibly due to sequencing limitations, assembly strategies, or inherent genomic complexity. These differences emphasize the importance of optimizing sequencing and assembly methods to achieve high-quality genomic assemblies for comparative and functional studies. The auN was calculated as 8,339.5 bp. The L50 and L90 values, indicating the number of contigs required to cover 50% and 90% of the genome assembly, were 12,442 and 50,659, respectively. No ambiguous bases (N's) were detected in the assembly, as reflected by 0.00 N's per 100 kbp. Regarding the scaffold assembly, a total of 270,380 scaffolds were generated, with 46,511 scaffolds being at least 1,000 bp in length and 7,963 exceeding 10,000 bp. The largest scaffold reached a length of 96,620 bp. The assembly achieved a total length of 285,845,886 bp, with a GC content of 36.33%. The N50 was 9,805 bp, and the L50 was 8,186 bp, indicating moderate continuity. The auN was 12,948.7 bp, reflecting the overall quality of the assembly. Additionally, there were 18.89 Ns per 100 kbp, indicating some unresolved regions, which are typical in draft genome assemblies. These metrics collectively highlighted a robust scaffold assembly with potential for further refinement.

The quality of the genome assembly was evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs), which assesses genome completeness based on the presence of conserved orthologous genes. The analysis revealed that the assembly contained 269 complete BUSCOs (C), representing 63.29% of the total searched BUSCO groups (Table 4). Among these, 264 BUSCOs were identified as complete and single-copy (S), while 5 BUSCOs were complete and duplicated (D). Additionally, 143 BUSCOs were classified as fragmented (F), and 13 BUSCOs were reported as missing (M). Overall, the evaluation was conducted against 425 BUSCO groups. The assembly statistics further supported the BUSCO results, highlighting the structural attributes of the genome.

Table 4. Genome completeness analysis using Viridiplantae database in BUSCO module of GAME v1.

BUSCO results	
Complete BUSCOs (C)	269
Complete and single-copy BUSCOs (S)	264
Complete and duplicated BUSCOs (D)	5
Fragmented BUSCOs (F)	143
Missing BUSCOs (M)	13
Total BUSCO groups searched	425
Assembly statistics	
Number of scaffolds	341893
Number of contigs	341893
Total length	320381369
Percent gaps	0.000%
Scaffold N50	15 KB
Contigs N50	15 KB

The total length of the assembly was 320,381,369 bp, distributed across 341,893 contigs, with no gaps detected (0.000% gaps). The contig N50 and scaffold N50 values were both 15 kb, indicating that half of the total genome length was covered by contigs or scaffolds of this length or longer. These results demonstrate a moderate level of genome completeness and structural integrity, emphasizing the assembly's suitability for downstream analyses such as gene annotation

and functional studies. The presence of complete single-copy BUSCOs suggests a significant representation of the essential genes within the assembly, while the fragmented and missing BUSCOs highlight areas for potential improvement in assembly quality (Manni *et al.*, 2021).

In a recent study of *Cymbopogon citratus* L. nuclear genome, complete BUSCOs were recorded as 60.90% and missing BUSCOs were noted as 17.7% (Chakravartty and Neelapu, 2024). In the present investigation, GAME v1 revealed better results for *C. pallidicaule* with complete and missing BUSCOs of about 63.29% and 3.06%, respectively. This outcome reinforced the nuclear genome assembly of *C. pallidicaule* using GAME v1. Functional analysis using COG (Clusters of Orthologous Groups) database revealed annotations for 24 different categories while no records were observed for nuclear structure (Y) and extracellular structures (W) (Fig. 3A).

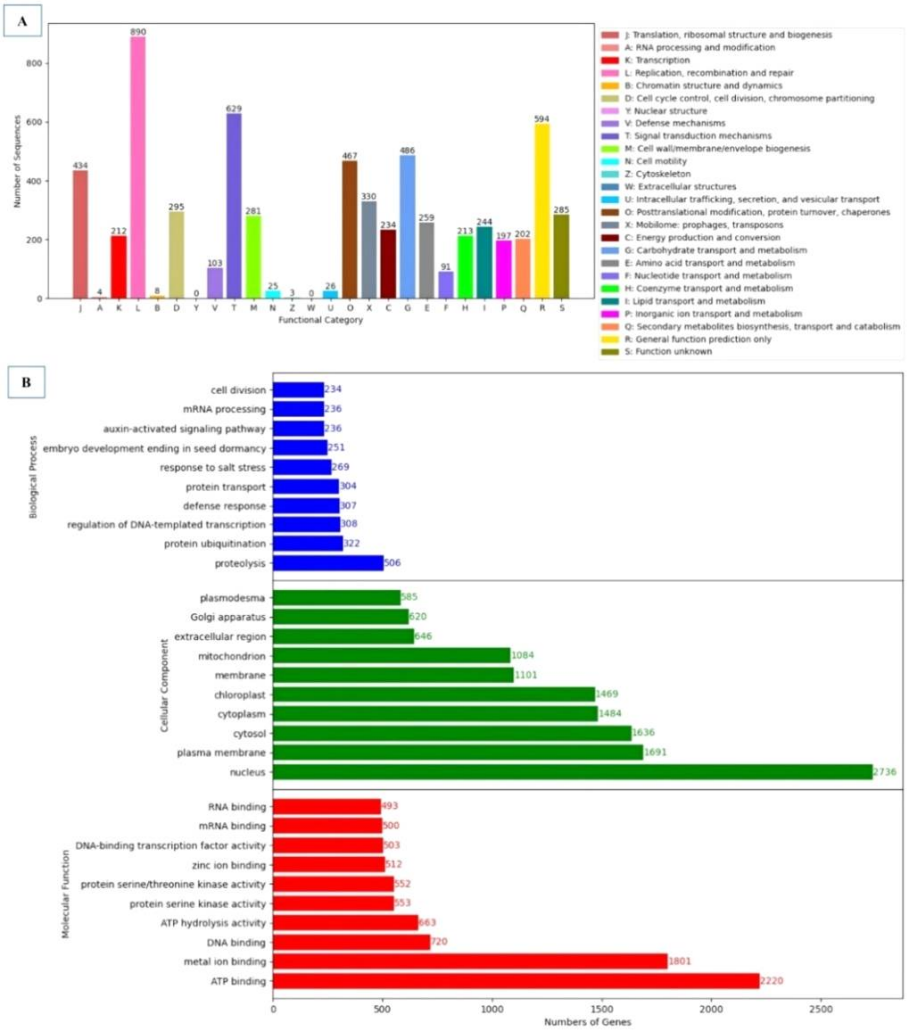

Fig. 3. Functional annotation of the nuclear genome of *C. pallidicaule*. A. Annotation based on COG identifiers, B. Annotation based on gene ontology.

The highest number of genes (890) were predicted to be involved in the replication, recombination, and repairing processes. The lowest number of genes (3) were predicted to have cytoskeleton-related functions. In *C. pallidicaule*, the cytoskeleton might play crucial role in cellular processes such as cell division, intracellular transport, and maintaining cell shape. Identifying cytoskeleton-related genes provides insights into these critical physiological functions. Gene ontology (GO)-based analysis categorized the nuclear genome into three major functional groups: biological processes, cellular components, and molecular functions, providing comprehensive insights into the functional roles of the predicted genes.

A total of 23,806 genes were identified, distributed across cellular components (13,052 genes), molecular functions (8,517 genes), and biological processes (2,737 genes) (Fig. 3B). Within the cellular component category, the nucleus exhibited the highest representation with 2,736 genes, while the plasmodesma showed the lowest with 585 genes. For molecular functions, ATP binding dominated with 2,220 genes, underscoring its critical role in energy-dependent cellular processes, whereas RNA binding was the least represented, involving 493 genes. In the biological process category, proteolysis had the highest gene count, reflecting its essential role in protein turnover and cellular regulation, while cell division showed the lowest representation. This GO analysis highlights the functional complexity of the genome and provides valuable insights into the biological and molecular mechanisms operating within *C. pallidicaule*.

Pathway distribution analysis using KEGG database unraveled a total of 2028 genes involved in various pathways (Fig. 4). The highest number of genes (460) were found to play functional roles in the protein modification pathways. The lowest gene count (35) was observed in the pyruvate from D-glyceraldehyde 3-phosphate pathway. A total of 335 genes were involved in the protein ubiquitination pathway and another 151 genes were functional in the amino acid biosynthesis pathway. The gene count for carbohydrate metabolism and purine metabolism pathways was same (40). In the same way, aromatic compound metabolism, porphyrin-containing compound metabolism, and fatty acid biosynthesis pathways showed very similar gene count (37).

RepeatMasker analysis revealed repetitive elements spanning 223,973 base pairs (bp), representing approximately 0.069% of the nuclear genome. A variety of repeat classes were identified, each contributing distinct lengths and gene counts (Figs 5 and 6). Long terminal repeats (LTRs) were among the most prominent categories. 4,132 LTR/Copia elements were identified, contributing 20,230 bp (0.006%), while 3,947 LTR/Gypsy elements spanned 23,312 bp (0.007%). Additional LTR-related elements, including general LTR sequences, contributed 20,575 bp (0.006%) across 3,184 genes. DNA transposons included 267 TIR (Terminal Inverted Repeat) elements, covering 13,231 bp (0.004%), and 202 Helitron elements, spanning 18,502 bp (0.006%). Smaller contributions were observed from TRIM (Terminal Repeat Retrotransposons in Miniature) elements, which accounted for 2,268 bp (0.001%) across 279 genes. Unclassified repeats contributed 14,773 bp (0.005%) and were represented by 285 genes. LINEs (Long Interspersed Nuclear Elements) made up 17,576 bp (0.005%) across 41 genes, while SINEs (Short Interspersed Nuclear Elements) contributed 2,171 bp (0.001%) with 7 genes. Additional elements, such as rRNA-related repeats spanning 3,132 bp (0.001%) across 202 genes, and other simple repeats at 374 bp, were also identified. The analysis reflected a minimal presence of repetitive sequences in the genome, with no single category dominating the genome's structure. This low repeat content suggests a compact and efficient genomic organization, potentially limiting the influence of non-coding repetitive regions on gene functionality.

The sequencing of life on Earth is revolutionizing basic biological research and transforming our understanding of phylogenetics, evolution, ecology, conservation, agriculture, bioindustry, and medicine (Hiller *et al.*, 2012; Henry, 2022). These advances are pivotal in building a sustainable future, ensuring biosecurity, and fostering an innovative bio-economy (Blaxter *et al.*, 2022).

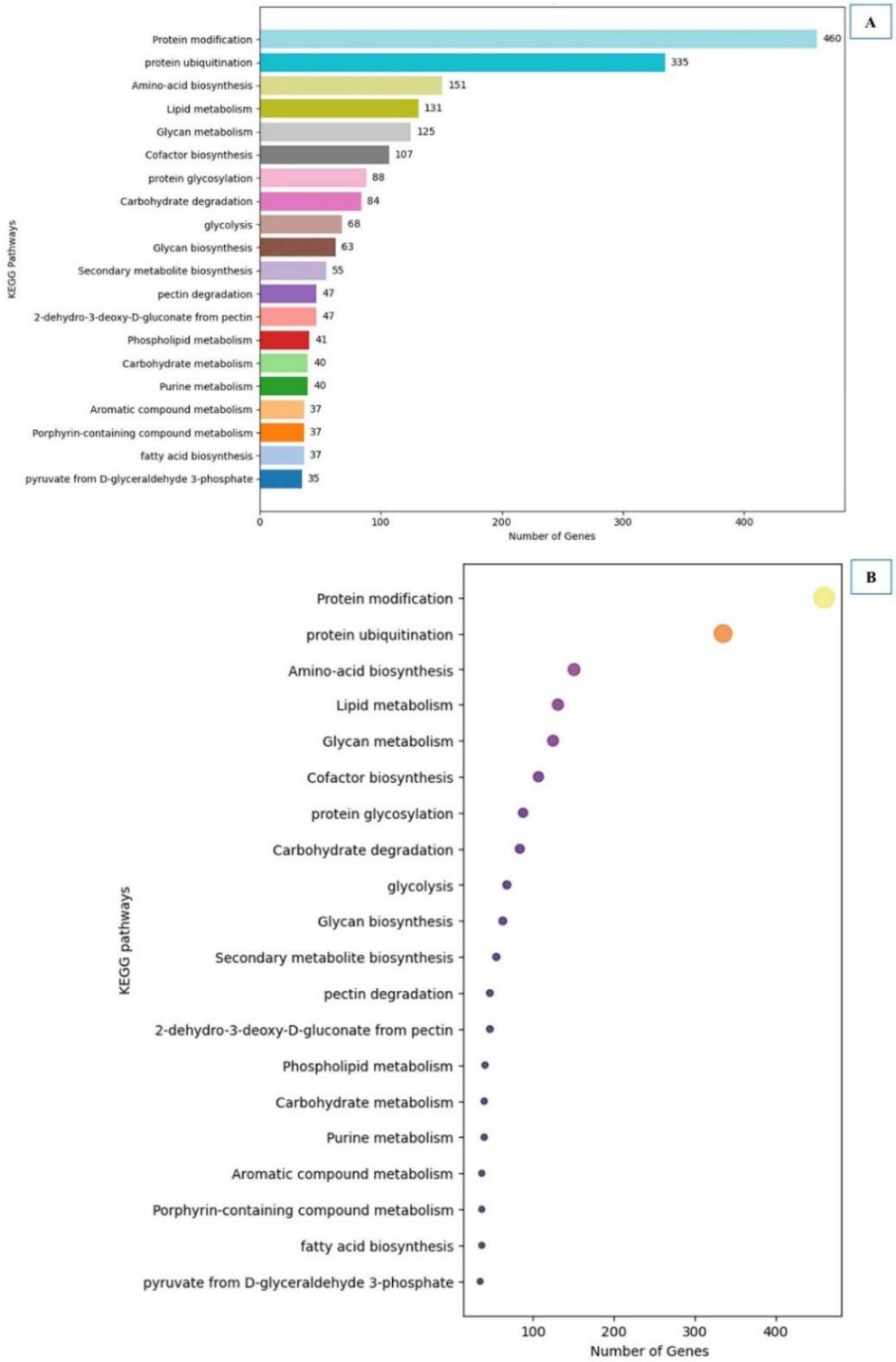

Fig. 4. KEGG pathway analysis showing the distribution of nuclear genes across various biological pathways. A. Histogram plot. B. Dot plot.

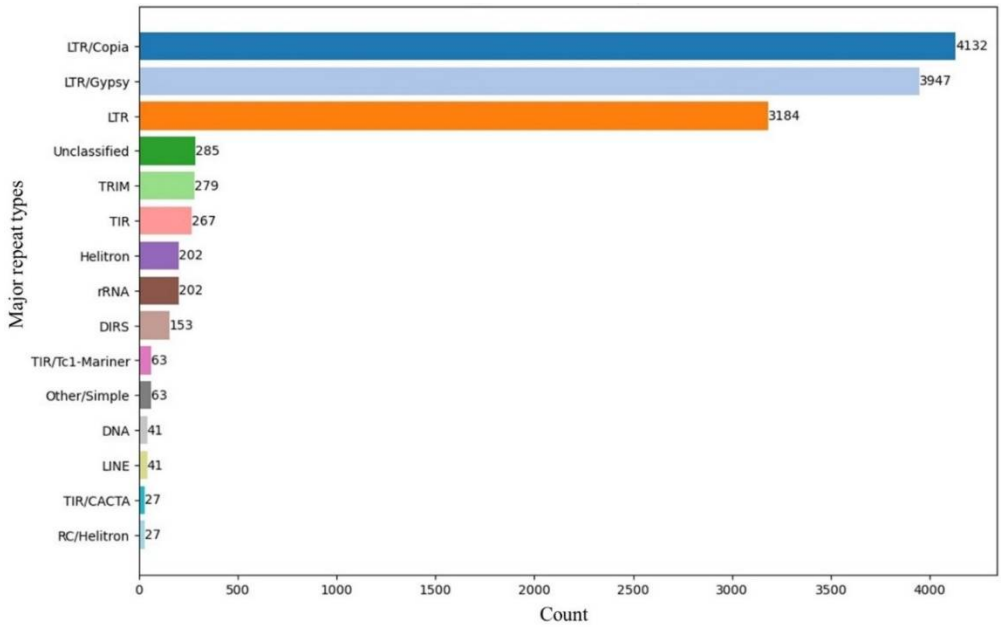

Fig. 5. Analysis of major repeat structures in the nuclear genome of *C. pallidicaule* using the RepeatMasker module of GAME v1 showing gene count.

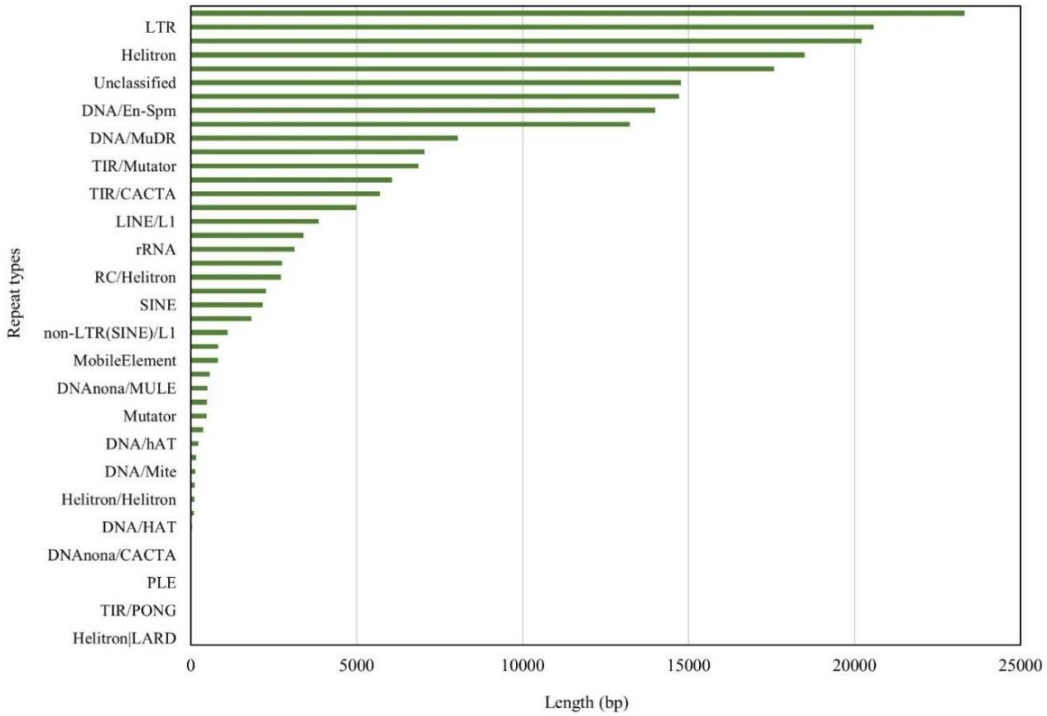

Fig. 6. Analysis of repeat structures in the nuclear genome of *C. pallidicaule* using the RepeatMasker module of GAME v1 showing the length (bp) of the repeat structures.

In plants, genomic resources have the potential to simplify and expedite experimental gene function analysis (Rodríguez Del Río *et al.*, 2024), enable metabolic pathway engineering for enhanced drought and heat tolerance (Liu *et al.*, 2023), facilitate data mining of bioactive compounds (Liu *et al.*, 2022), and improve crop breeding strategies (Henry, 2022). As sequencing costs decrease and data volumes surge, computational analysis has emerged as a significant bottleneck (Pucker *et al.*, 2022). Efficient downstream processing of short-read datasets is critical after sequencing (Zhang *et al.*, 2011), with whole-genome assembly and annotation being among the most challenging tasks. While numerous tools exist for genome assembly and annotation, they often require advanced bioinformatics skills (Helmy *et al.*, 2016) and, in some cases, costly subscriptions (El-Metwally *et al.*, 2013), such as BLAST2GO (Conesa and Götz, 2008).

While GAME v1 offers a robust, user-friendly platform for the assembly and annotation of higher plant genomes, but limited to short-read sequencing data generated by Illumina platforms. An enhanced GAME v2 is under development, which promises to overcome the current limitations by improving assembly performance and expanding analysis capabilities.

Acknowledgement

The authors extend their appreciation to the Researchers Supporting Project number (RSP2025R306), King Saud University, Riyadh, Saudi Arabia.

References

Blaxter, M., Archibald, J.M., Childers, A.K., Coddington, J.A., Crandall, K.A., Di Palma, F., Durbin, R., Edwards, S.V., Graves, J.A.M., Hackett, K.J., Hall, N., Jarvis, E.D., Johnson, R.N., Karlsson, E.K., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Lindblad-Toh, K., Lopez, J.V., Moran, N.A., Robinson, G.E., Ryder, O.A., Shapiro, B., Soltis, P.S., Warnow, T., Zhang, G. and Lewin, H.A. 2022. Why sequence all eukaryotes? Proc. Natl. Acad. Sci. USA **119**(4): e2115636118.

Caudai, C., Antonella, G., Filippo, G., Loredana, L.P., Veronica, M., Emanuele, S., Allegra, V. and Teresa, C. 2021. AI applications in functional genomics. Comp. Struct. Biotech. J. **19**: 5762–5790.

Chakravartty, N. and Neelapu, N.R.R. 2024. The *de novo* genome assembly of lemon grass to identify the genes in essential oil production. J. App. Biol. Biotechnol. **12**(2): 100–149.

Conesa, A. and Götz, S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genom. **2008**: 619832.

Dai, B., Guo, H., Huang, C., Zhang, X. and Lin, Z. 2016. Genomic heterozygosity and hybrid breakdown in cotton (*Gossypium*): Different traits, different effects. BMC Genet. **17**: 1–11.

Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P. and Lavenier, D. 2014. GATB: Genome assembly & analysis tool box. Bioinformatics **30**(20): 2959–2961.

Ejigu, G.F. and Jung, J. 2020. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biol. **9**(9): 295.

Ellestad, P., Pérez-Farrera, M.A. and Buerki, S. 2022. Genomic insights into cultivated Mexican *Vanilla planifolia* reveal high levels of heterozygosity stemming from hybridization. Plants **11**(16): 2090.

El-Metwally, S., Hamza, T., Zakaria, M. and Helmy, M. 2013. Next-generation sequence assembly: Four stages of data processing and computational challenges. PLoS Comput. Biol. **9**: e1003345.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics **29**(8): 1072–1075.

Helmy, M., Mohamed, A. and Kareem, A.M. 2016. Limited resources of genome sequencing in developing countries: Challenges and solutions. Appl. Trans. Genomics **9**: 15–19.

Henry, R.J. 2022. Progress in plant genome sequencing. Appl. Biosci. **1**: 113–128.

Hiller, M., Schaar, B.T., Indjeian, V.B., Kingsley, D.M., Hagey, L.R. and Bejerano, G. 2012. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep. **2**(4): 817–823.

Jung, H., Ventura, T., Chung, J.S., Kim, W.J., Nam, B.H., Kong, H.J., Kim, Y.O., Jeon, M.S. and Eyun, S.I. 2020. Twelve quick steps for genome assembly and annotation in the classroom. PLoS Comput. Biol. **16**(11): e1008325.

Kathiresan, N., Temanni, R., Almabrazi, H., Syed, N., Jithesh, P.V. and Al-Ali, R. 2017. Accelerating next generation sequencing data analysis with system level optimizations. Sci. Rep. **7**:, 9058.

Lin, Y.J., Ding, X.Y., Huang, Y.W. and Lu, L. 2024. First *de novo* genome assembly and characterization of *Gaultheria prostrata*. Front. Plant Sci. **15**: 1456102.

Liu, S., Zenda, T., Tian, Z. and Huang, Z. 2023. Metabolic pathways engineering for drought or/and heat tolerance in cereals. Front. Plant Sci. **14**: 1111875.

Liu, X., Gong, X., Liu, Y., Liu, J., Zhang, H., Qiao, S., Li, G. and Tang, M. 2022. Application of high-throughput sequencing on the chinese herbal medicine for the data-mining of the bioactive compounds. Front. Plant Sci. **13**: 900035.

Manni, M., Berkeley, M.R., Seppey, M. and Zdobnov, E.M. 2021. BUSCO: Assessing genomic data quality and beyond. Curr. Prot. **1**(12): e323.

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F.D.V., Fufezan, C., Ternent, T., Eglen, S.J., Katz, D.S., Pollard, T.J., Konovalov, A., Flight, R.M., Blin, K. and Vizcaíno, J.A. 2016. Ten simple rules for taking advantage of Git and GitHub. PLoS Comput. Biol. **12**(7): e1004947.

Pucker, B., Irisarri, I., de Vries, J. and Xu, B. 2022. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quant. Plant Biol. **11**(3): e5.

Rice, E.S. and Green, R.E. 2019. New Approaches for Genome Assembly and Scaffolding. Annu. Rev. Anim. Biosci. **7**: 17–40.

Rodríguez Del Río, Á., Giner-Lamia, J., Cantalapiedra, C.P., Botas, J., Deng, Z., Hernández-Plaza, A., Munar-Palmer, M., Santamaría-Hernando, S., Rodríguez-Herva, J.J., Ruscheweyh, H.J., Paoli, L., Schmidt, T.S.B., Sunagawa, S., Bork, P., López-Solanill,a E., Coelho, L.P. and Huerta-Cepas, J. 2024. Functional and evolutionary significance of unknown genes from uncultivated taxa. Nat. **626**(7998): 377–384.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R.P., Banday, S., Mishra, A.K., Das, G. and Malonia, S.K. 2023. Next-generation sequencing technology: Current trends and advancements. Biol. **12**: 997.

Shirasawa, K., Harada, D., Hirakawa, H., Isobe, S. and Kole, C. 2021. Chromosome-level *de novo* genome assemblies of over 100 plant species. Breed. Sci. **71**(2): 117–124.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19): 3210–3212.

Tarailo-Graovac, M. and Chen, N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics **25**(1): 4–10.

Tian, D., Xu, T., Kang, H., Luo, H., Wang, Y., Chen, M., Li, R., Ma, L., Wang, Z., Hao, L., Tang, B., Zou, D., Xiao, J., Zhao, W., Bao, Y., Zhang, Z. and Song, S. 2024. Plant genomic resources at National Genomics Data Center: Assisting in data-driven breeding applications. aBIOTECH **5**: 94–106.

Wickham, H. 2011. ggplot2. Wiley Interdiscip. Rev. Comput. Stat. **3**(2): 180–185.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. and Shen, B. 2011. A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. PLoS ONE **6**(3): e17915.

(Manuscript received on 22 September 2024; revised on 28 November 2024)