Bangladesh J. Plant Taxon. **32**(1): 1-16, 2025 (June) DOI: https://doi.org/10.3329/bjpt.v32i1.82388

DE NOVO NUCLEAR GENOME ASSEMBLY AND ANNOTATION OF *AIZOON CANARIENSE* L. (AIZOACEAE): UNCOVERING GENOMIC ADAPTATIONS OF A MEDICINAL HERB TO ARID ENVIRONMENTS

REEM LAFI SALEEM ALOFI[1], MOHAMMAD AJMAL ALI[1*], MONA SOLAIMAN ALWAHIBI[1], SHEIKH SUNZID AHMED[2], M. OLIUR RAHMAN[2*], RAJESH MAHATO[3], SOO-YONG KIM[4] AND JOONGKU LEE[5]

[1] *Department of Botany and Microbiology, College of Science, King Saud University, Riyadh-11451, Saudi Arabia*
[2] *Department of Botany, Faculty of Biological Sciences, University of Dhaka, Dhaka 1000, Bangladesh*
[3] *ArrayGen Technologies Private Limited, Undri, Pune-411060, Maharashtra, India*
[4] *International Biological Material Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea*
[5] *Department of Environment and Forest Resources, College of Agricultural Life Science, Chungnam National University, Daejeon, South Korea*

Keywords: Nuclear genome; *Aizoon canariense*; GAME v.1; Annotation; GO analysis; KEGG pathway.

Abstract

In this investigation, whole genome sequencing and assembly of medicinally important species *Aizoon canariense* L. were performed to unveil its nuclear genome. The assembled nuclear genome length was 661.14 Mb, with an N50 value of 25,334 bp. The genome was largely homozygous, with a low level of heterozygosity ranging from 0.077 to 0.078%. BUSCO assessment exhibited 91.8% completeness based on the Viridiplantae database. Orthologous gene-based analysis revealed that the highest number of genes were associated with replication, recombination, and repair category. Gene Ontology (GO)-based annotation identified 5,814 genes involved in biological processes (BP), 25,137 genes linked to cellular components (CC), and 16,644 genes associated with molecular functions (MF). Pathway enrichment analysis identified the protein modification pathway representing the highest number of genes (807), whereas the pigment biosynthesis pathway exhibiting the lowest number of genes (75). A total of 20,448 repeat elements across 57 distinct types were identified in the assembled genome, with LTR/Gypsy being the most abundant, comprising 7,532 copies and spanning approximately 27,750 bp. These findings lay a strong foundation for future research on the molecular mechanisms of stress resilience in arid ecosystems targeting *A. canariense*.

Introduction

Aizoon canariense L. is a native desert herb found in Saudi Arabia, belonging to the family Aizoaceae Martinov that holds significant pharmacological properties. The Aizoaceae, commonly known as the ice plant family, consists of 127 genera and approximately 1,860 species, primarily distributed in the tropical and subtropical regions of South Africa, with some species found in Australia. These plants are often referred to as "stone plants" or "carpet weeds" (Bittrich and Hartmann, 1988; Leistner, 2000). *A. canariense* is an annual or perennial herb, characterized by its prostrate growth and thick-stemmed morphology, stems reaching up to 40 cm in length and often

*Corresponding authors. Email: alimohammad@ksu.edu.sa ; oliur.bot@du.ac.bd

exhibiting a papillose texture. The leaves are subcircular to obovate, entire, decurrent at the base, and covered with fine hairs. The flowers are solitary and sessile, with perianth segments that are yellowish inside and greenish or reddish and pilose outside. The fruit is star-shaped, red or pink, and depressed at the center (El-Amier and Al-Hadithy, 2020).

Nature-derived medicines have long been a cornerstone of traditional healing systems, offering a rich source of bioactive compounds with therapeutic potential. Unlike synthetic drugs, which are often associated with side effects and environmental concerns due to complex chemical synthesis, plant-based medicines provide a more biocompatible and sustainable alternative (Bhardwaj *et al.*, 2024; Rather *et al.*, 2025). *A. canariense*, a medicinally significant desert herb, exemplifies the potential of nature-derived treatments. Traditionally used to treat ailments such as hypertension and digestive disorders, recent studies have highlighted its cytotoxic, antioxidant, and anti-inflammatory properties (Yonbawi *et al.*, 2021). The diverse phytochemicals in *A. canariense*, including flavonoids, alkaloids, and saponins, contribute to its pharmacological profile, many of which remain unexplored in the context of modern drug development (Bakr *et al.*, 2021). Moreover, its ability to thrive in extreme desert environments suggest the presence of unique stress-related metabolites that may inspire novel therapeutics. In light of growing concerns over antibiotic resistance and the adverse effects of synthetic drugs, exploring plant-derived medicines from *A. canariense* may offer a promising avenue for safer, more effective, and environmentally sustainable drug discovery, further underscoring the need for decoding its genomic blueprint.

Adapted to the harsh environmental conditions of Saudi Arabia, *A. canariense* exhibits unique physiological and biochemical traits that enable its survival in arid ecosystems (Baeshen *et al.*, 2021). Understanding its nuclear genome assembly is crucial for unravelling the genetic basis of its resilience to drought, high salinity, and intense heat. The nuclear genome in plants serves as the central repository of genetic information, regulating vital biological processes such as growth, development, stress responses, and the biosynthesis of secondary metabolites (Hu *et al.*, 2021). With the rapid advancements of next-generation sequencing (NGS) technologies, particularly Illumina sequencing, and the emergence of sophisticated bioinformatics tools, plant genome decoding has become more efficient and accessible than ever before (Raza and Ahmad, 2019). In the past, large genome sizes, high repeat content, and structural complexities posed significant challenges to genome assembly. However, modern sequencing technologies and computational pipelines have now facilitated high-quality nuclear genome reconstruction (Miller, 2001; Taber *et al.*, 2014).

With reference to genomic adaptations of desert plants, Baeshen *et al.* (2021) performed *de novo* transcriptome assembly of several species native to Saudi Arabia to investigate their responses to abiotic stress, identifying thousands of stress-responsive genes and transcription factors associated with drought and salinity tolerance across multiple taxa. A full-length transcriptome analysis of *Stipagrostis pennata* revealed rapid transcriptomic evolution, indicating that desert adaptation in *S. pennata* is driven by dynamic gene regulation and functional diversification of stress-related genes (Ding *et al.*, 2021). More recently, Alharbi *et al.* (2024) conducted a comprehensive genomic survey of 51 plant species from Saudi Arabia, with a primary focus on chloroplast genomes, growing interest in nuclear genomes, and limited exploration of mitochondrial data, emphasizing the need for more extensive nuclear genome sequencing and further research into adaptive trait genomics. In this context, assembling and annotating the nuclear genome of *A. canariense* may offer valuable insights into genes responsible for abiotic stress tolerance, metabolic pathways linked to medicinally significant compounds, and evolutionary adaptations that enhance its survival in harsh desert conditions. Furthermore, a well-

annotated genome provides a foundation for comparative genomic studies with other stress-tolerant species, enabling the identification of both conserved and novel adaptive traits (Vu *et al.*, 2015). These insights not only broaden our understanding of desert plant biology but also hold significant promise for biotechnological applications, including the development of climate-resilient crops and the identification of bioactive compounds for pharmaceutical and agricultural advancements.

Despite its ecological and medicinal significance, the nuclear genome of *A. canariense* remains unexplored. Therefore, in the present study, we aim to generate the first whole-genome assembly of *A. canariense* using next-generation sequencing approach, focusing on assembling and annotating the genome to identify genes linked to stress tolerance and medicinal properties. The findings will enhance our current understanding of desert plant resilience and open new avenues for biotechnological and pharmaceutical applications.

Materials and methods

Specimen collection

A. canariense was collected from Medina, Saudi Arabia (24º42′42.5″N, 39º28′18.2″E; altitude: 828 m). Species identification was carried out based on the morphological characteristics of leaves and fruits (Fig. 1). The voucher specimen has been deposited in the King Saud University Herbarium (KSUH) in Riyadh, Saudi Arabia, under the collection code Alofi, R.L.S & Ali, M.A. 2023-1.

Genome sequencing

Total genomic DNA was extracted from silica gel-dried leaves using the Qiagen DNA Extraction Kit. Paired-end sequencing was performed on a NovaSeq 6000 platform, generating 151 bp reads. The quality of the NGS reads was assessed using FASTQC v.0.12.1 and fastp v.0.20.1 tools (Chen *et al.*, 2018; Ahmed and Rahman, 2025). The raw sequencing data are publicly available on NCBI under the SRA accession ID SRR31760782.

Nuclear genome analysis

The paired-end sequencing reads were processed using the Genomic Analysis Made Easy (GAME) v.1 pipeline to estimate the pre-assembly genome size, construct the nuclear genome assembly, annotate the assembled genome, and characterize repetitive DNA elements (Ali *et al.*, 2024). The GAME v.1 software was developed using Python to provide a user-friendly, fast, free, and automated GUI-based solution for plant genome assembly and annotation.

Genome size estimation and assembly

GenomeScope v.1.0 integrated within GAME v.1 platform was employed to estimate genome size, heterozygosity, and repeat content based on k-mer frequency distribution derived from high-throughput sequencing reads (Ranallo-Benavidez *et al.*, 2020). The nuclear genome was subsequently assembled using the GATB Minia pipeline, an efficient de Bruijn graph-based assembler optimized for large-scale genomic datasets (Drezen *et al.*, 2014).

Assembly quality assessment

The completeness and accuracy of the assembled genome were evaluated using QUAST v.5.3.0 module, which provided key assembly metrics such as contig length distribution, N50 values, and genome coverage (Gurevich *et al.*, 2013). The completeness of the genome assembly was evaluated using BUSCO v.5.8.0 based on benchmarking universal single-copy orthologs (Seppey *et al.*, 2019).

Functional annotation

Gene prediction and functional annotation were conducted using Augustus v.3.4.0 (Stanke *et al.*, 2004), an *ab initio* gene prediction tool. Repeat elements within the assembled genome were identified and masked using RepeatMasker to characterize repetitive DNA sequences (Chen, 2004).

Data visualization

The results of the analysis, including genome size estimation, assembly statistics, gene annotation, and repeat content, were visualized using ggplot in R, providing graphical representations of the key genomic features (Valero-Mora, 2010).

Results and Discussion

SRA reads and quality control

Whole-genome sequencing was performed on the Illumina NovaSeq 6000 platform, generating a total of 178,737,199 sequencing reads, corresponding to 54 Gbp of raw data with a total file size of 16.6 Gb (SRA accession: SRR31760782). FastQC quality control analysis revealed satisfactory results for both forward and reverse reads, with high base-call accuracy (Q32–Q36) across most positions, showing only a slight decline toward the end of the reads. The Q scores, also known as Phred scores, represent the accuracy of each base call in sequencing data. The observed high Q scores suggest that the sequencing data is of very high quality, with a low error rate. The slight decline toward the end of the reads is a common occurrence and typically reflects a natural decrease in sequencing accuracy as the sequencing process progresses. This finding is consistent with previous studies (Ahmed and Rahman, 2024, 2025).

Fig. 1. Morphology of *Aizoon canariense* collected from Medina of Kingdom of Saudi Arabia. A. Habit (×1), B. Flowering twig (×3).

The fastp-based quality control analysis revealed that the raw data set contained 357.47 million reads (53.98 Gbp), with Q20 and Q30 percentages of 97.10% and 92.38%, respectively (Table 1). After filtering, 344.94 million high-quality reads (96.49%) were retained, totaling 52.02 Gbp, with improved Q20 (97.89%) and Q30 (93.48%) values. The mean read length remained almost consistent (151 bp before filtering and 150 bp after), and the GC content showed a slight reduction from 39.87% to 39.75%. A low duplication rate (4.82%) indicated minimal PCR bias, while stringent filtering effectively removed low-quality reads (3.45%), those with excessive ambiguous bases (0.0027%), and short reads (0.0487%). These results underscore the high accuracy and reliability of the sequencing data, ensuring an optimal dataset for genome assembly of *A. canariense* and its downstream analyses by minimizing sequencing artifacts and preserving authentic genomic information.

Genome characterization using K-mer-based analysis

GenomeScope analysis of the *A. canariense* nuclear genome yielded essential genomic estimates through k-mer profiling, enabling precise genome characterization without a reference genome (Fig. 2). The haploid genome size was estimated to be between 849.09 Mb and 849.40 Mb, with a substantial proportion of unique sequences (81.6%) and repetitive content ranging from 156.35 Mb to 156.40 Mb, indicating a moderately complex genome (Table 2). The heterozygosity rate was low (0.0766–0.0777%), suggesting a largely homozygous genome and limited genetic variation within the sampled population. The model fit was high (98.39–99.46%), confirming the accuracy of the k-mer-based predictions, while the sequencing read error rate remained minimal (0.1419%), ensuring the high quality of the dataset.

Table 1. Summary of quality control metrices for paired-end Illumina reads processed using the fastp tool.

Summary	
fastp version	0.20.1
Sequencing	Paired end (151 cycles + 151 cycles)
Mean length before filtering	151 bp, 151 bp
Mean length after filtering	150 bp, 150 bp
Duplication rate	4.817741 %
Insert size peak	0
Before filtering	
Total reads	357.474398 M
Total bases	53.978634 G
Q20 bases	52.414873 G (97.102999%)
Q30 bases	49.863750 G (92.376829%)
GC content	39.872787%
After filtering	
Total reads	344.941404 M
Total bases	52.023063 G
Q20 bases	50.922957 G (97.885349%)
Q30 bases	48.631723 G (93.481084%)
GC content	39.745233%
Filtering result	
Reads passed filters	344.941404 M (96.494016%)
Reads with low quality	12.349180 M (3.454563%)
Reads with too many N	9.748000 K (0.002727%)
Reads too short	174.066000 K (0.048693%)

Table 2. Genome characteristics of *A. canariense* estimated using the GenomeScope module of GAME v1.

Property	Minimum	Maximum
Heterozygosity	0.0766417%	0.0777496%
Genome haploid length	849,091,440 bp	849,396,167 bp
Genome repeat length	156,345,157 bp	156,401,267 bp
Genome unique length	692,746,283 bp	692,994,900 bp
Model fit	98.3917%	99.4597%
Read error rate	0.141877%	0.141877%

To validate the effectiveness of GenomeScope in characterizing the nuclear genome *of A. canariense*, its results were compared with those obtained for *Spiraea crenata* (Laczkó *et al.*, 2024). Although *A. canariense* and *S. crenata* belong to different families (Aizoaceae and Rosaceae, respectively), the comparison highlights the utility of GenomeScope in estimating key genomic parameters across diverse plant taxa. The haploid genome size of *S. crenata* (232.33 Mb) was considerably smaller than that of *A. canariense*, illustrating variations in genome expansion between the species. The proportion of unique sequences in *S. crenata* (60.2%) was lower than in *A. canariense* (81.6%), suggesting that *A. canariense* harbored a higher fraction of non-repetitive sequences. Additionally, *S. crenata* exhibited a higher heterozygosity rate (0.834%) compared to *A. canariense* (0.0766–0.0777%), implying greater genetic diversity within the sampled population. The k-mer sequencing coverage (Kcov) was higher in *S. crenata* (16.5×) compared to *A. canariense* (6.57×), suggesting a higher depth of sequencing in *S. crenata*, which could potentially lead to more reliable genome assembly. Both species exhibited low sequencing error rates (*S. crenata*: 0.0621%, *A. canariense*: 0.1419%), further bolstering the reliability of the NGS data.

Gaultheria prostrata (Ericaceae) nuclear genome assembly revealed a model error rate of about 0.493% (Lin *et al.*, 2024). In contrast, the assembly of *A. canariense* in the present study achieved a significantly lower model error rate of 0.1419%, nearly five times lower than that reported for *G. prostrata* (Fig. 2). For large genome assemblies, error rates of less than 0.5% are usually considered as acceptable since they lessen the possibility of distortion in repeated regions and structural variants (Ali *et al.*, 2024). The significantly lower error rate in *A. canariense* underscores the robustness of our genome assembly, ensuring a more accurate representation of the nuclear genome and reinforcing its suitability for downstream analyses and comparative genomic studies.

Nuclear genome assembly and quality assessments

The GATB-Minia pipeline successfully assembled the nuclear genome of *A. canariense*, generating a total of 181,780 contigs with a cumulative length of approximately 660.44 Mb (Table 3). The largest contig reached 474,998 bp, reflecting the assembly's capability to reconstruct long genomic fragments. The N50 value, a key metric for assembly continuity, was 7,092 bp, indicating that at least half of the total assembly length was represented by contigs of this size or longer. The assembly exhibited a GC content of 39.14%, aligning with expected values for plant genomes. The total number of contigs exceeding 1,000 bp was 128,133, with a combined length of ~622.78 Mb, while 38,857 contigs were longer than 5,000 bp. Notably, 218 contigs exceeded 50,000 bp, reflecting the presence of relatively long, high-confidence sequences. The absence of N's per 100 kbp indicated a clean assembly with no unresolved base calls. These results demonstrate the

effectiveness of the GATB-Minia pipeline in assembling the nuclear genome, providing a strong foundation for downstream gene annotation and comparative genomic analyses.

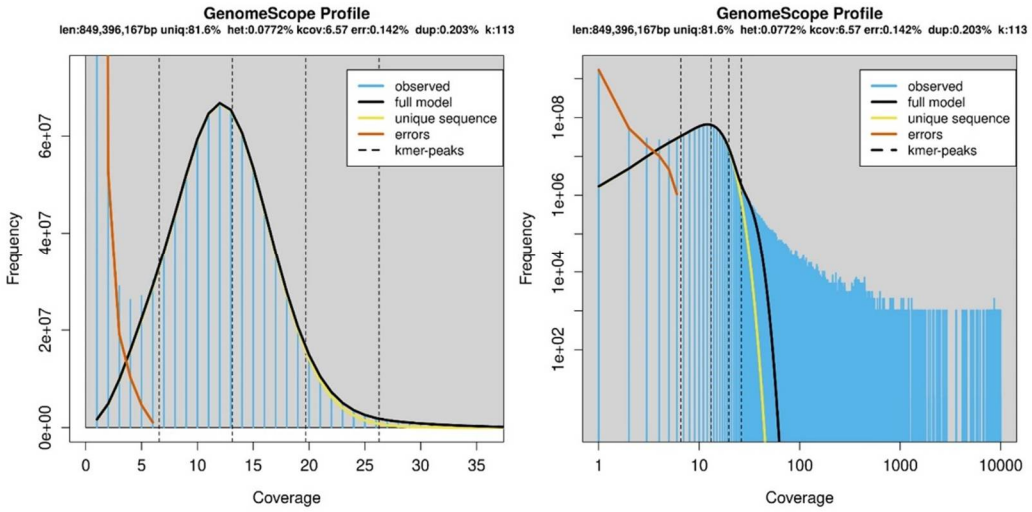
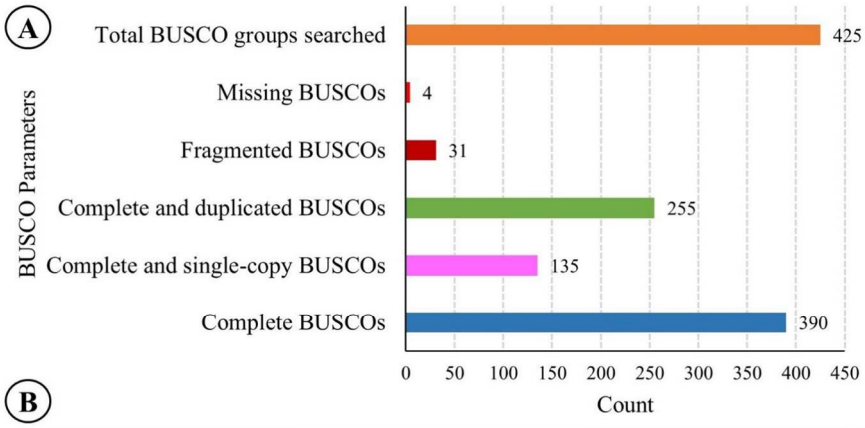


Fig. 2. Pre-assembly genomic characterization of the nuclear genome of *A. canariense* using GenomeScope module.



BUSCO Assembly Statistics	
Number of scaffolds	164058
Number of contigs	188226
Total length	686543321
Percent gaps	0.444%
Scaffold N50	24 KB
Contig N50	14 KB

Fig. 3. BUSCO evaluation of the nuclear genome of *A. canariense*. A. Genome completeness, B. Assembly statistics.

Table 3. Post-assembly assessment of scaffolds and contigs via QUAST tool.

Assembly metrices	Scaffolds	Contigs
Sequences ≥ 0 bp	164058	279492
Sequences ≥ 1,000 bp	49824	128133
Sequences ≥ 5,000 bp	27928	38857
Sequences ≥ 10,000 bp	18837	13994
Sequences ≥ 25,000 bp	7183	1135
Sequences ≥ 50,000 bp	1865	218
Cumulative size (≥ 0 bp)	686543321	692180555
Cumulative size (≥ 1,000 bp)	638256886	622779075
Cumulative size (≥ 5,000 bp)	587307098	413363501
Cumulative size (≥ 10,000 bp)	521466937	238850598
Cumulative size (≥ 25,000 bp)	333766500	55564350
Cumulative size (≥ 50,000 bp)	150608110	26771333
Total sequence count	83246	181780
Longest single sequence	1419080	474998
Overall assembled length	661142997	660442599
GC content (%)	39.13	39.14
N50	25334	7092
N90	4387	1440
auN	49500.0	15686.7
L50	7057	24896
L90	29576	104616
N's per 100 kbp	460.59	0.00

The GATB-Minia pipeline assembled 83,246 scaffolds for the *A. canariense* nuclear genome, with a total length of ~661.14 Mb (Table 3). The largest scaffold reached 1,419,080 bp, indicating successful reconstruction of extensive genomic regions. The assembly had an N50 value of 25,334 bp, suggesting enhanced scaffold continuity compared to previous scaffold-based assemblies. The GC content was 39.13%, aligning with expectations for plant genomes. Of the total scaffolds, 49,824 exceeded 1,000 bp, contributing to ~638.26 Mb, while 27,928 scaffolds were longer than 5,000 bp. Remarkably, 1,865 scaffolds surpassed 50,000 bp, highlighting the presence of long, high-confidence sequences. The L50 value of 7,057 indicated that the shortest 50% of the assembly was contained in 7,057 scaffolds, while the L90 value of 29,576 reflected a substantial proportion of smaller fragments. The auN value of 49,500 supported the overall quality of the assembly. However, the presence of 460.59 N's per 100 kbp suggested some unresolved regions. Despite this, the results demonstrate the efficacy of the GATB-Minia pipeline in producing a robust and high-quality scaffolded genome assembly, facilitating downstream annotation and comparative genomic analyses.

Compared to *Cymbopogon citratus* (DC.) Stapf, which had a total genome size of 364.44 Mb and an N50 value of 4,347 bp, our *A. canariense* assembly exhibited significantly higher contiguity and scaffold length (Chakravartty and Neelapu, 2024). The longest scaffold in *A. canariense* (1,419,080 bp) was more than 20 times longer than that of *C. citratus* (67,673 bp), indicating a superior reconstruction of genomic regions. Additionally, the L50 of *A. canariense* (7,057) was considerably lower than that of *C. citratus* (23,781), suggesting fewer but larger scaffolds covered half of the assembled genome. A higher proportion of large scaffolds was also

observed in *A. canariense*, with 22.6% exceeding 10,000 bp, compared to only 3.3% in *C. citratus*. However, while *C. citratus* reported no unresolved regions, *A. canariense* exhibited 460.59 N's per 100 kbp, indicating the presence of some assembly gaps. Overall, the *A. canariense* genome assembly demonstrated greater scaffold continuity, longer genomic fragments, and a more contiguous structure, making it a more comprehensive resource for genomic studies.

The BUSCO analysis of the nuclear genome assembly of *A. canariense*, using the Viridiplantae database, revealed a high level of completeness, with 91.8% of the searched BUSCOs identified as complete (Fig. 3). Of these, 31.8% were single-copy genes, while 60.0% were duplicated, suggesting a substantial proportion of retained gene duplications within the assembly. Additionally, 7.3% of BUSCOs were fragmented, and only 0.9% were entirely missing, indicating minimal gene loss and a well-represented gene space. The assembly consisted of 164,058 scaffolds and 188,226 contigs, with an estimated genome size of approximately 686.54 Mb. The scaffold and contig N50 values were 24 kb and 14 kb, respectively, indicating a moderately continuous assembly that captures large genomic segments while balancing contiguity and completeness. The percentage of gaps in the assembly was 0.444%, suggesting that most of the genomic regions were successfully resolved with minimal ambiguity.

When compared to *Phoenix roebelenii* and *Cymbopogon citratus*, the genome assembly of *A. canariense* shows markedly superior genome completeness. The overall BUSCO completeness score for *A. canariense* (91.8%) surpasses those of *P. roebelenii* (84.2%) and *C. citratus* (60.9%), reflecting a more comprehensive representation of conserved plant orthologs (Chakravartty and Neelapu, 2023, 2024). Moreover, the proportion of duplicated BUSCOs in *A. canariense* (60.0%) is significantly higher than in *P. roebelenii* (4.3%) and *C. citratus* (2.1%), suggesting greater retention of gene duplications in *A. canariense*, potentially reflecting evolutionary processes such as whole-genome duplications. The missing BUSCOs are considerably lower in *A. canariense* (0.9%) compared to *P. roebelenii* (10.2%) and *C. citratus* (17.7%), demonstrating the assembly's greater completeness and minimal gene loss. Additionally, the percentage of fragmented BUSCOs in *A. canariense* (7.3%) is also lower than in *C. citratus* (21.4%), highlighting the greater continuity and higher-quality assembly in *A. canariense*. The BUSCO metrices validate the robustness of the *A. canariense* genomic assembly and underscore its potential for downstream functional genomics and evolutionary studies.

Functional annotation of the nuclear genome

The COG (Clusters of Orthologous Genes)-based functional annotation of the *A. canariense* nuclear genome revealed a wide array of functional categories, highlighting key biological processes and metabolic pathways. The most abundant category was replication, recombination, and repair, comprising 2,015 annotated genes, indicating a strong genomic stability mechanism and active DNA maintenance processes (Fig. 4). Signal transduction mechanisms were the second most represented group, with 1,296 genes, suggesting complex regulatory networks for cellular communication. A significant number of genes were also assigned to general function prediction (1,227) and mobilome elements, including prophages and transposons (1,151), indicating the presence of uncharacterized proteins and mobile genetic elements, respectively. Key metabolic pathways were well-represented, including carbohydrate transport and metabolism (993 genes), amino acid transport and metabolism (739 genes), and lipid transport and metabolism (655 genes), indicating a well-developed metabolic framework. Genes involved in translation, ribosomal structure, and biogenesis (950 genes) and post-translational modification, protein turnover, and chaperones (835 genes) underscore the importance of protein synthesis and maintenance. The Transcription-related genes (628) reflect robust gene expression regulation, while energy production and conversion (610) genes indicate an active cellular energy metabolism. Secondary

metabolite biosynthesis, transport, and catabolism category was represented by 495 genes, pointing to the plant's potential in producing bioactive compounds. Genes related to defense mechanisms (347) and cell wall/membrane/envelope biogenesis (687) suggest adaptations for environmental stress responses and structural integrity. The presence of nucleotide transport and metabolism (203), intracellular trafficking and secretion (113), and cell motility (91) genes further enriches the genomic functional landscape. Less common but notable categories included chromatin structure and dynamics (18), RNA processing and modification (16), cytoskeleton components (11), and extracellular structures (2), while nuclear structure-related genes were missing (Fig. 4).

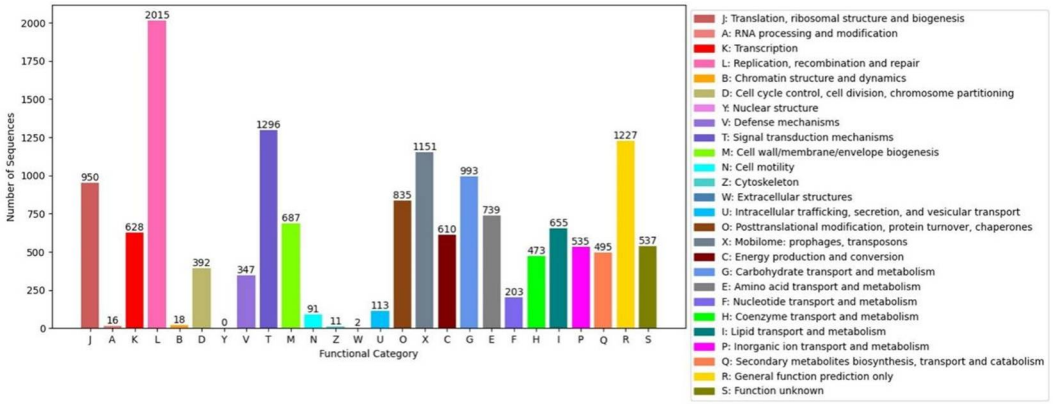


Fig. 4. Functional characterization of the *Aizoon canariense* nuclear genome elucidating classification of nuclear genes based on COG functional categories.

The Gene Ontology (GO)-based functional annotation provided a comprehensive overview of the biological processes (BP), cellular components (CC), and molecular functions (MF) of *A. canariense* nuclear genome (Fig. 5). In the biological process category, 5,814 genes were annotated, with proteolysis being the most abundant category (993 genes), emphasizing its critical role in protein turnover and cellular regulation. Conversely, the response to abscisic acid was the least represented (445 genes), indicating a more specialized function in stress adaptation mechanisms. In the cellular component category, 25,137 genes were identified, with the nucleus exhibiting the highest representation (4,801 genes), highlighting its essential role in genetic regulation. In contrast, the Golgi apparatus had the lowest gene count (1,027 genes), reflecting its specialized role in protein modification and transport. The molecular function category included 16,644 genes, with ATP binding being the most prevalent (4,557 genes), underscoring its fundamental role in energy metabolism and enzymatic activity. The mRNA binding class was the least represented (900 genes), indicating its selective role in gene expression regulation (Fig. 5).

A comparative analysis with the GO-based annotation of *Chenopodium pallidicaule* Aellen revealed striking similarities, reinforcing the robustness of the *A. canariense* genome annotation. In both species, proteolysis emerges as the most enriched biological process, emphasizing its conserved role in protein homeostasis (Ali *et al.*, 2024). Likewise, within the cellular component domain, the nucleus is the most represented structure, validating its universal significance in transcriptional control. Furthermore, ATP binding is the dominant molecular function in both genomes, underscoring its critical role in cellular energetics. These consistent patterns suggest a

shared functional framework between *A. canariense* and *C. pallidicaule*, further validating the accuracy and biological relevance of the GO classification in *A. canariense*.

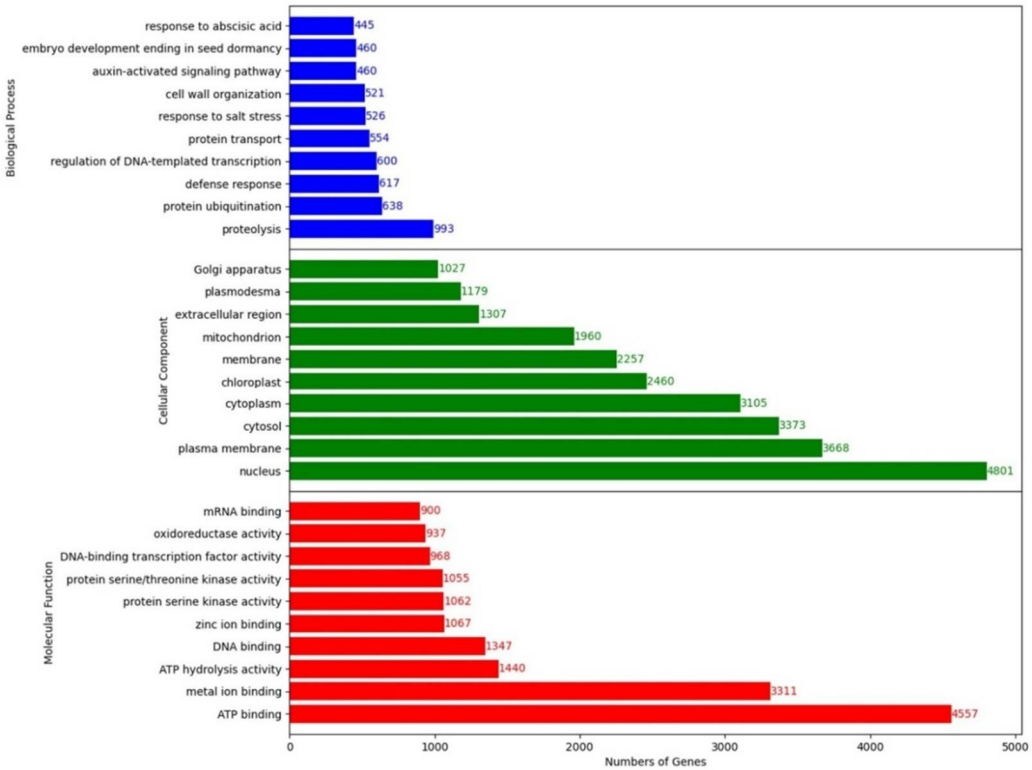


Fig. 5. Functional characterization of the *Aizoon canariense* nuclear genome illustrating gene ontology-based annotation across biological processes, cellular components, and molecular functions.

Pathway-based annotation represents a pivotal component of genome analysis, as it provides functional insights into the biochemical and metabolic pathways within an organism. Mapping genes to specific pathways may help to identify key biological processes, elucidates metabolic capacities, and infer species-specific adaptations. This approach is particularly instrumental in uncovering mechanisms underlying stress responses, biosynthesis of secondary metabolites, and complex regulatory networks. These insights hold significant implications for agriculture, biotechnology, and evolutionary biology (Wang *et al.*, 2022).

The KEGG pathway-based annotation of the *A. canariense* nuclear genome revealed the distribution of functional genes across ten major pathway categories, highlighting key metabolic and regulatory processes (Fig. 6). Among these, the protein modification pathway exhibited the highest gene representation, comprising 807 genes, emphasizing the significance of post-translational modifications in cellular function and proteome stability. Protein ubiquitination, a crucial process for protein degradation and cellular homeostasis, ranked second with 632 genes, while amino acid biosynthesis, essential for fundamental metabolic activities, was the third most abundant pathway with 402 genes. In contrast, pathway with the lowest gene representation included pigment biosynthesis, with only 75 genes, indicating a relatively specialized and limited role in pigment formation. Similarly, porphyrin-containing compound metabolism, associated with

heme and chlorophyll biosynthesis, had 81 genes, while fatty acid biosynthesis, a vital pathway for membrane lipid production, contained 83 genes.

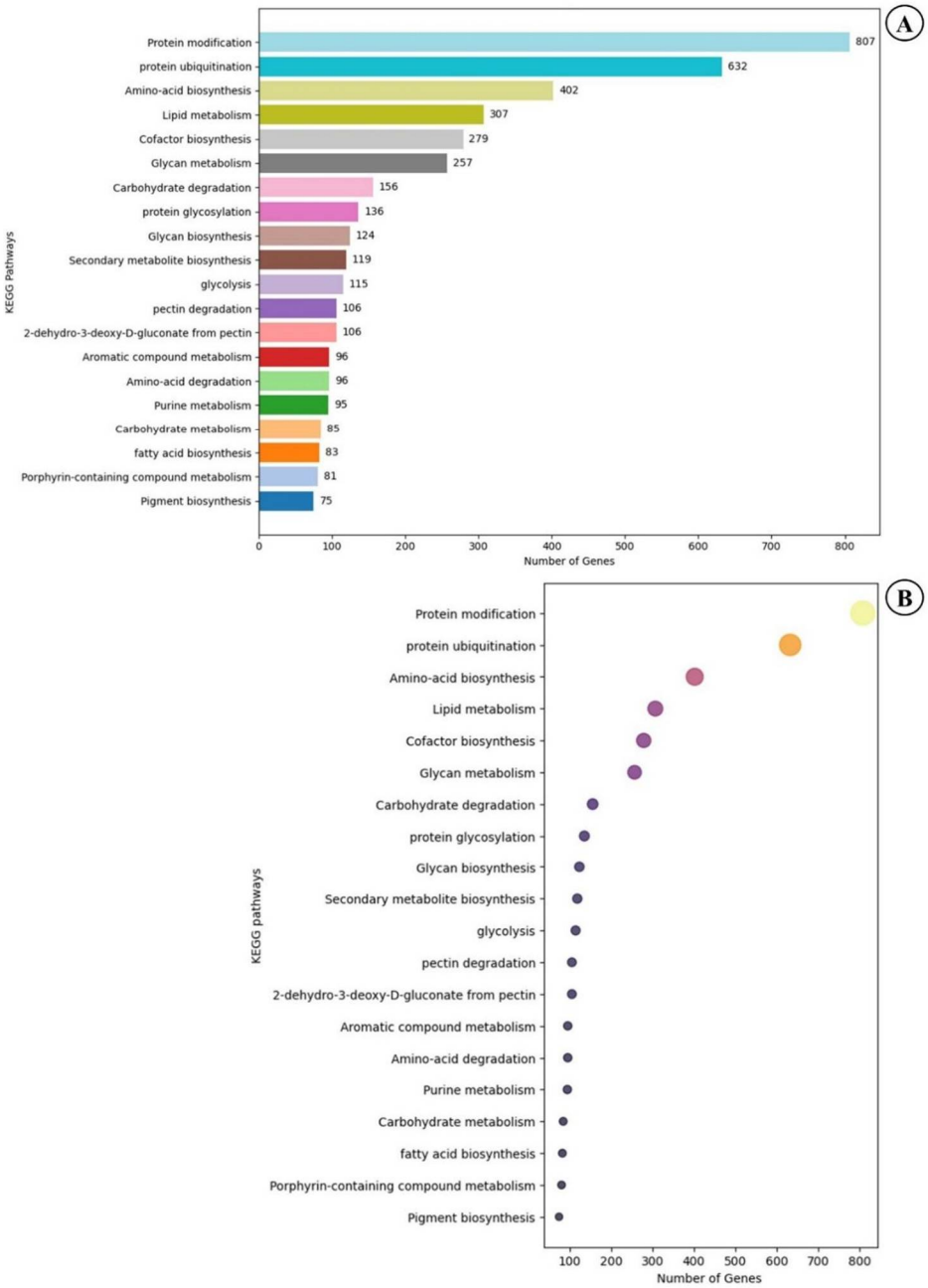


Fig. 6. KEGG pathway-based functional classification of nuclear genes in *A. canariense*. A. Histogram representation of gene distribution across different pathways, B. Dot plot visualization of pathway enrichment.

In comparison with the nuclear genome of *C. pallidicaule* (Ali *et al.*, 2024), *A. canariense* exhibited a similar pattern of pathway enrichment. In *C. pallidicaule*, the highest number of genes was associated with the protein modification pathway (460 genes), followed by protein ubiquitination (335 genes) and amino acid biosynthesis (151 genes). Notably, *A. canariense* showed significantly higher gene counts in these pathways, 807, 632 and 402 genes respectively, suggesting a more extensive involvement in protein regulation and primary metabolic processes (Fig. 6). Moreover, while *C. pallidicaule* had the lowest gene count in pyruvate from D-glyceraldehyde 3-phosphate metabolism (35 genes), *A. canariense* exhibited comparatively higher representation in low-count pathways, such as pigment biosynthesis (75 genes) and porphyrin metabolism (81 genes). These findings validate the pathway-based annotation approach and underscores the functional diversity of both species. The higher gene representation in *A. canariense* may suggest possible adaptations related to stress response, metabolic plasticity, or genome evolution, reinforcing the importance of pathway annotation in understanding plant genomic functionality (Kellogg and Bennetzen, 2004).

The RepeatMasker analysis identified 20,448 repeat elements across 57 distinct repeat types within an assembly length of 686,543,321 bp. The total length of repeat sequences accounted for 269,099 bp. The most abundant category was LTR/Gypsy, with 7,532 copies spanning 27,750 bp (Fig. 7).

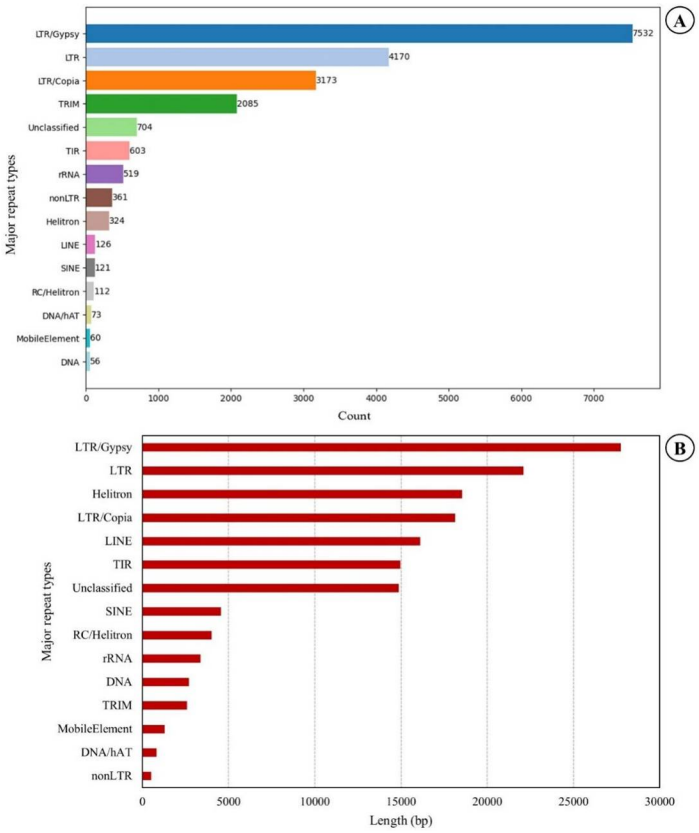


Fig. 7. Evaluation of repetitive elements in the nuclear genome of *A. canariense* using RepeatMasker module. A. Major repeat types showing number of genes, B. Major repeat types showing length distribution.

Other major repeat classes included LTR (4,170 copies) and LTR/Copia (3,173 copies), contributing 22,104 bp and 18,129 bp, respectively. DNA transposons such as TRIM (2,085 copies) and TIR (603 copies) were also prevalent, along with smaller contributions from Helitron, LINE, and SINE elements. Unclassified repeats and other mobile elements further contributed substantial sequence lengths, forming a diverse set of repetitive sequences within the genome. In contrast, the RepeatMasker analysis of *C. pallidicaule* revealed significantly lower repeat content, spanning only 223,973 bp, with a more limited set of repeat classes (Ali *et al.*, 2024). Although LTR elements such as Copia and Gypsy were also present in *C. pallidicaule*, they were less abundant, contributing to smaller proportions of the genome. The findings in *A. canariense* revealed a more complex and diverse repeat landscape with a greater variety of repeat types and higher repeat counts. This suggests that *A. canariense* possesses a more expansive set of repetitive sequences compared to *C. pallidicaule*, potentially reflecting differences in genome size, complexity, and structural organization.

In conclusion, the comprehensive whole-genome sequencing and subsequent analyses of *Aizoon canariense*, a medicinally important desert plant, have provided valuable insights into its genomic structure and functional potential. The successful nuclear genome assembly, spanning 661.14 Mb, laid the foundation for functional annotation through COG and Gene Ontology classifications, revealing a diverse array of genes involved in critical biological processes, cellular components, and molecular functions. RepeatMasker analysis identified a variety of repetitive sequences, highlighting the complexity and diversity of its genome. These findings contribute significantly to the understanding of the genomic architecture of *A. canariense*, offering a deeper insight into its potential for bioactive compound discovery and adaptation to harsh desert environments. The present investigation lays the groundwork for future research aimed at exploring the molecular mechanisms underpinning its medicinal properties and stress resilience, thus paving the way for further applications in pharmacological and biotechnological fields.

Acknowledgements

The authors extend their appreciation to Ongoing Funding Research Program (ORF-2025-306), King Saud University, Riyadh, Saudi Arabia, for funding this work. This research was also supported by the KRIBB Initiative Program [KGM1172511] of the Republic of Korea.

References

Ahmed, S.S. and Rahman, M.O. 2024. Deciphering the complete chloroplast genome sequence of *Meconopsis torquata* Prain: Insights into genome structure, comparative analysis and phylogenetic relationship. Heliyon **10**: e36204.

Ahmed, S.S. and Rahman, M.O. 2025. Comparative genomics and phylogenetic analysis of complete chloroplast genome of *Scaphium scaphigerum* (Wall. *ex* G. Don) G. Planch. Dhaka Univ. J. Biol. Sci. **34**(1): 119–143.

Alharbi, S.A., Alzahrani, A.A., Alluqmani, S.M., Albdour, A.M., Alzahrani, E.A. and Almsaoudi, S.J. 2024. Genomic insights into Saudi Arabian plant biodiversity: Progress and future directions. Electronic J. Univ. Aden Basic & Appl. Sci. **5**(4): 342–354.

Ali, M.A., Mahato, R. and Lee, J. 2024. Genomic Analysis Made Easy (GAME V1): An automated software for plant genome assembly and annotation from illumina sequencing. Bangladesh J. Plant Taxon. **31**(2): 225–238.

Baeshen, M.N., Ahmed, F., Moussa, T.A.A., Abulfaraj, A.A., Jalal, R.S., Noor, S.O., Baeshen, N.A. and Huelsenbeck, J.P. 2021. A comparative analysis of *de novo* transcriptome assembly to understand the abiotic stress adaptation of desert plants in Saudi Arabia. Appl. Ecol. & Environ. Res. **19**(3): 1753–1782.

Bakr, R.O., El-Behairy, M.F., Elissawy, A.M., Elimam, H. and Fayed, M.A. 2021. New adenosine derivatives from *Aizoon canariense* L.: *In vitro* anticholinesterase, antimicrobial, and cytotoxic evaluation of its extracts. Molecules **26**(5): 1198.

Bhardwaj, D., Tiwari, D., Upadhye, V.J., Ramniwas, S., Rautela, I., Ballal, S., Kumar, S., Bhat, M., Sharma, S., Kumar, M.R., Pandey, P. and Khan, F. 2024. Discovery of new natural phytocompounds: The modern tools to fight against traditional bacterial pathogens. Curr. Pharmaceut. Biotechnol. **26**: e13892010344474.

Bittrich, V. and Hartmann, H.E. 1988. The Aizoaceae - a new approach. Bot. J. Linn. Soc. **97**(3): 239–254.

Chakravartty, N. and Neelapu, N.R. 2023. The *de novo* genome assembly (nuclear, chloroplast and mitochondria) of ornamental plant pygmy date palm *Phoenix roebelenii*. J. Appl. Biol. Biotech. **11**: 113–122.

Chakravartty, N. and Neelapu, N.R.R. 2024. The genome assembly of lemon grass to identify the genes in *de novo*. J. Appl. Biol. Biotech. **12**: 100–149.

Chen, N. 2004. Using repeat masker to identify repetitive elements in genomic sequences. Curr. Prot. Bioinform. **5**(1): 4–10.

Chen, S., Zhou, Y., Chen, Y. and Gu, J. 2018. fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinform. **34**(17): i884–i890.

Ding, X., Zhang, T. and Ma, L. 2021. Rapidly evolving genetic features for desert adaptations in *Stipagrostis pennata*. BMC Genomics **22**: 846.

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

El-Amier, Y.A. and Al-Hadithy, O.N. 2020. Phytochemical constituents, antioxidant and allelopathic activities of *Aizoon canariense* L. on *Zea mays* (L.) and associated weeds. Plant Arch. **20**(1): 303–310.

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

Hu, Q., Ma, Y., Mandáková, T., Shi, S., Chen, C., Sun, P., Zhang, L., Feng, L., Zheng, Y., Feng, X., Yang, W., Jiang, J., Li, T., Zhou, P., Yu, Q., Wan, D., Lysak, M.A., Xi, Z., Nevo, E. and Liu, J. 2021. Genome evolution of the psammophyte *Pugionium* for desert adaptation and further speciation. Proc. Natl. Acad. Sci. **118**(42): e2025711118.

Kellogg, E.A. and Bennetzen, J.L. 2004. The evolution of nuclear genome structure in seed plants. Am. J. Bot. **91**(10): 1709–1725.

Laczkó, L., Jordán, S., Póliska, S., Rácz, H.V., Nagy, N.A., Molnár V.A. and Sramkó, G. 2024. The draft genome of *Spiraea crenata* L. (Rosaceae) – The first complete genome in tribe Spiraeeae. Sci. Data **11**(1): 219.

Leistner, O.A. 2000. Seed plants of southern Africa: Families and genera. *Strelitzia* **10**. National Botanical Institute, Pretoria.

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.

Miller, W. 2001. Comparison of genomic DNA sequences: Solved and unsolved problems. Bioinform. **17**(5): 391–397.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. **11**(1): 1432.

Rather, H.A., Maqbool, I., Bhat, K.A., Gautam, V., John, A., Nazir, A. and Ayaz, A. 2025. Different types of phytocompounds with immunomodulatory activities. *In:* Mahajan, R., Shehjar, F., Zargar, S.M., Masoodi, K.Z. and Shah, Z.A. (Eds), Role of medicinal plants in autoimmune diseases. Academic Press, New York, pp. 261–289.

Raza, K. and Ahmad, S. 2019. Recent advancement in next-generation sequencing techniques and its computational analysis. Int. J. Bioinform. Res. Appl. **15**(3): 191–220.

Seppey, M., Manni, M. and Zdobnov, E.M. 2019. BUSCO: Assessing genome assembly and annotation completeness. *In*: Kollmar, M. (Ed.), Gene Prediction. Methods in Molecular Biology, Vol. **1962**. Humana, New York, pp. 227–245.

Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. **32**: W309–W312.

Taber, K.A.J., Dickinson, B.D. and Wilson, M. 2014. The promise and challenges of next-generation genome sequencing for clinical care. JAMA Intern. Med. **174**(2): 275–280.

Valero-Mora, P.M. 2010. ggplot2: Elegant graphics for data analysis. J. Stat. Softw. **35**: 1–3.

Vu, G.T., Schmutzer, T., Bull, F., Cao, H.X., Fuchs, J., Tran, T.D., Jovtchev, G., Pistrick, K., Stein, N., Pecinka, A., Neumann, P., Novak, P., Macas, J., Dear, P.H., Blattner, F.R., Scholz, U. and Schubert, I. 2015. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. The Plant Genome **8**(3): 1–14.

Wang, J., Xie, J., Chen, H., Qiu, X., Cui, H., Liu, Y., Sahu, S.K., Fang, D., Li, T., Wang, M., Chen, Y., Liu, H., Zhang, J. and Wang, B. 2022. A draft genome of the medicinal plant *Cremastra appendiculata* (D. Don) provides insights into the colchicine biosynthetic pathway. Commun. Biol. **5**(1): 1294.

Yonbawi, A.R., Abdallah, H.M., Alkhilaiwi, F.A., Koshak, A.E. and Heard, C.M. 2021. Anti-proliferative, cytotoxic and antioxidant properties of the methanolic extracts of five Saudi Arabian flora with folkloric medicinal use: *Aizoon canariense*, *Citrullus colocynthis*, *Maerua crassifolia*, *Rhazya stricta* and *Tribulus macropterus*. Plants **10**(10): 2073.

(Manuscript received on 20 January, 2025; revised on 2 June, 2025)