DE NOVO DRAFT ASSEMBLY AND ANNOTATION OF THE *MORINGA OLEIFERA* GENOME FROM PAKISTAN

M. T. Pervez [1†], S. H. Abbas[2, 1†] and M. E. Babar [3] and T. Hussain [1*]

[1] Department of Biological Sciences, Virtual University of Pakistan
[2] Center for Bioinformatics and Computational Biology, The Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai 200241, China
[3] University of Veterinary and Animal Sciences (UVAS), Lahore, Pakistan
†These authors have contributed equally to this work and share first authorship.
*Corresponding author's e-mail: tanveer.hussain@vu.edu.pk

ABSTRACT

Moringa oleifera, commonly known as the horseradish (drumstick tree), is a fast-growing, mineral-rich evergreen tree in the family Moringaceae, with wide applications in herbal medicine, agriculture, and livestock production. Despite its importance, genomic resources remain limited, with only one public draft assembly previously available. This study presents a de novo draft genome of *Moringa oleifera* from Pakistan, sequenced by the Illumina HiSeq platform using paired-end libraries. The assembly comprises ~205.2 Mb across 13,872 scaffolds (N50 = 17,279 bp) and includes 26,215 predicted protein-coding genes. Assembly quality, supported by completeness assessments, indicates the recovery of 83.9% of embryophyte BUSCOs and 84.68% of CEGMA core genes. This Pakistan-origin reference fills a vital geographic and technical gap for *M. oleifera*, enabling robust functional annotation and comparative analysis in the *Moringa* genus. It provides an immediate foundation for marker development and downstream applications in molecular breeding, biotechnology, and stress adaptation research.

Keywords*: Moringa oleifera*, genome, assembly, annotation, Pakistan

https://doi.org/10.36899/JAPS.2026.4.0088 Published first online May 01, 2026

INTRODUCTION

Moringa oleifera (*M. oleifera*; **NCBI Taxonomy ID:** txid3735), commonly known as the horseradish (drumstick) tree, is a fast-growing, mineral-rich species in the family Moringaceae. It is found naturally in regions of South Asia (Pakistan, northern India, Bangladesh, Nepal, and Afghanistan) (Gopalakrishnan *et al.*, 2016; Bhattacharya *et al.*, 2018; Oyeyinka and Oyeyinka, 2018). It has multipurpose applications in the fields of herbal medicine, agriculture, and livestock (Vergara-Jimenez *et al.*, 2017; Islam *et al.*, 2021; Pareek *et al.*, 2023). Although the leaves are the most widely consumed part of the plant, the whole plant is edible and highly nutritious, and it is increasingly being accepted as a possible alternative food source to fight malnutrition, especially among children and infants (Gopalakrishnan *et al.*, 2016; Oyeyinka and Oyeyinka, 2018). Nutritionally, *M. oleifera* is an excellent source of proteins, vitamin C, β-carotene, potassium, and calcium. It is also rich in natural antioxidant compounds, including ascorbic acid, flavonoids, carotenoids, and phenolics (Brilhante *et al.*, 2017; Table S1). It contains higher levels of several important nutrients than conventional sources, more vitamin C than oranges, more vitamin A than carrots, more calcium than milk, more potassium than bananas, and more iron than spinach (Gopalakrishnan *et al.*, 2016).

Moringa oleifera has a long history of traditional use in medicine; its leaves have been used to treat various conditions, including depression, anxiety, psoriasis, and asthma (Vergara-Jimenez *et al.*, 2017). Beyond its medicinal and economic value, *M. oleifera* also holds ecological importance. Due to its ability to flourish under low-rainfall conditions and withstand long-term drought, it is considered a pioneer species for afforestation. Given its diverse therapeutic applications, ecological adaptability, and nutritional richness, *M. oleifera* represents a promising candidate for future scientific research. In particular, exploring the genetic potential of this species as an alternative food source is increasingly relevant because of climate change and food security challenges.

Despite the high economic and medicinal value of *M. oleifera*, its usage for genetic improvement, breeding, and other purposes is largely limited due to its limited genomic resources. To date, complete genome assemblies have been

reported for accessions originating from China (Tian *et al.*, 2015) and India (Shyamli *et al.*, 2021), but no reference genome from Pakistan is available. This study presents a de novo, annotated draft genome of *M. oleifera* from Pakistan using Illumina paired-end data. The objectives were to (i) assemble and annotate a Pakistan origin reference, (ii) benchmark assembly contiguity and completeness using standard metrics (N50, BUSCO), and (iii) provide functionally annotated gene models to enable downstream applications in molecular breeding, biotechnology, and stress adaptation research.

MATERIALS AND METHODS

Plant materials & DNA extraction: Young, healthy leaves of *Moringa oleifera* were collected from the Botanical Garden of Government College University (GCU), Bagh-e-Jinnah, Lahore, Pakistan (31.556747° N, 74.328372° E) (Fig. 1). Leaf tissue was silica-gel dried and processed at the Virtual University of Pakistan (VUP) laboratory. Genomic DNA was extracted following a modified protocol of Ahmed *et al.*, (2009), in which the final ether precipitation step was replaced with silica-column purification to improve DNA purity. DNA integrity was verified on 1% agarose gel (EtBr) and visualized under UV; quantity and purity were assessed on a Multiskan GO µDrop plate reader (A260/280 and A260/230 ratios)

Whole-genome sequencing & raw-read quality control: High-molecular-weight DNA was shipped at ambient temperature to Novogene (Hong Kong) for Illumina library preparation and sequencing (paired-end 2 × 150 bp, HiSeq platform). The resulting raw sequencing data underwent quality control using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and AfterQC (Chen *et al.*, 2017). Adapter sequences and low-quality bases were trimmed, and reads were filtered using the following explicit criteria: minimum Phred quality (Q20), sliding-window trimming (4-bp window), minimum post-trim read length of 50 bp, removal of reads with > 5% ambiguous bases (N), and discarding exact PCR duplicates. After filtering, a total of 121.4 Gb of clean, high-quality data were retained. **Genome size and heterozygosity estimation**

Filtered reads were analyzed with Jellyfish (Manekar and Sathe, 2018) to generate k-mer frequency spectra at k = 21, 25, and 31. Genome size, repeat content, and heterozygosity were inferred using GenomeScope 2.0 (http://genomescope.org/genomescope2.0/) with default settings (Fig. 3). Selected k values are the standard for GenomeScope profiling and are independent of the assembly *k* chosen.

De novo assembly & completeness evaluation: Assemblies were generated with ABySS (GNU v.4.2.1), which is based on the *de Bruijn* graph algorithm and is well-suited for genomes with high heterozygosity (Simpson *et al*., 2009), and SOAPdenovo (v1.05), which is effective for short-read data generated by Illumina platforms (Luo *et al.*, 2012), using multiple k-mer sizes (k = 31, 71, 91, 121). For each assembler and k-mer, contigs were scaffolded using default parameters. Scaffolds < 1 kb were excluded from downstream analysis to reduce fragmentation and spurious short sequences. Assembly quality was evaluated using the number of contigs and scaffolds, scaffold length distributions, and overall genome completeness. The best-performing assemblies were identified by comparing these metrics across k-mer sizes. Assembly statistics, including total number of contigs and scaffolds (≥ 1 kb), N50/N75, L50, and percentage of Ns, were assessed using QUAST v5.0.2 (Gurevich *et al.*, 2013). Completeness was estimated using BUSCO v3.0 (Simão *et al.*, 2015; embryophyta_odb9, n = 1440; genome mode) and CEGMA (Parra *et al.*, 2007). The final assembly was selected by jointly maximizing completeness (BUSCO single-copy), contiguity (higher N50, lower L50), and compactness (fewer scaffolds, lower %Ns), while minimizing duplicated or fragmented BUSCOs and QUAST misassembles across k-mer settings.

Repeat identification & gene annotation: As no species-specific TE library exists for *M. oleifera*, a de novo custom repeat library was built with RepeatModeler v2.0.1 (Flynn *et al.*, 2020) on scaffolds ≥ 1 kb. This tool integrates two complementary algorithms, RECON (v1.08) and RepeatScout (v1.0.5), to predict and classify repetitive sequences. The resulting custom repeat library was then used with RepeatMasker (v4.1.0), which scanned the genome assembly to annotate and mask identified repeat elements (Akulova *et al.*, 2020; Flynn *et al.*, 2020). Protein-coding genes were predicted using BRAKER2, which combines GeneMark-ET and AUGUSTUS for evidence-guided gene prediction (Stanke *et al.*, 2006; Brůna *et al.*, 2020; Brůna *et al.*, 2021). Where available, external protein hints were supplied to improve exon-intron boundary accuracy. Gene models shorter than 150 bp or overlapping masked repeats by > 80% were filtered to avoid TE-related artifacts. Predicted proteins were functionally annotated with InterProScan to assign domains and Gene Ontology (GO) terms, and were queried against NCBI nr using BLASTP (E-value ≤ 1e-5; minimum alignment coverage ≥ 50%; maximum target sequences = 20) (Boratyn *et al.*, 2013; Blum *et al.*, 2021).

RESULTS

Raw Illumina reads were quality checked and filtered using FastQC and AfterQC, followed by k-mer profiling (Jellyfish) and genome modeling (GenomeScope 2.0). A total of 124.2 Gb of raw Illumina PE150 sequencing data were generated for *Moringa oleifera* (~395× coverage). Following quality filtering, 121.4 Gb of high-quality reads were retained for downstream analysis. Over 92% of bases had a Phred quality score ≥ 30, indicating excellent sequencing accuracy (Fig. S1; Table S2). Short-read assemblies were generated with ABySS and SOAPdenovo across multiple k-mer sizes, and assembly statistics were computed with QUAST.

Genome size estimation & assembly quality: The estimated genome size of *Moringa oleifera* ranged between ~210.76 Mb and 220.20 Mb, with heterozygosity of 1.84–2.03% and unique content of 79.4–80.0% (Fig. 3). The k-mer spectra (k =21, 25, 31) suggest that the *M. oleifera* genome contains a high proportion of repeats and substantial heterozygosity, which pose challenges for short-read assembly.

Among the assemblies evaluated, the selected build totals ~205.2 Mb, comprising 13,872 scaffolds and N50 = 17,279 bp (NCBI accessions: GenBank assembly GCA_021560355.1 [Moringa_Oleifera_1.0]; WGS master JAKELQ010000001-JAKELQ010013872). Across k-mer settings, the selected build (ABySS, k = 91) maximized contiguity and completeness among the short-read assemblies (Table 1; Table S3). A previously published *M. oleifera* assembly from the Indian cultivar 'Bhagya' is also available (ASM2139783v1). Because that resource derives from different germplasm and a distinct sequencing/assembly strategy, it is consider a complementary reference, whereas the present study provides the first Pakistan-origin genome tailored to regional germplasm (Table S3).

Assembly completeness assessments indicated robust gene recovery; BUSCO completeness was 83.9% and CEGMA completeness was 84.68%, confirming the presence of most conserved plant gene sets (Table 2). These results collectively indicate that the assembled *M. oleifera* genome is highly suitable for downstream genomic and functional studies. Together, these references enable side-by-side assessment of South Asian *M. oleifera* germplasm and provide a practical baseline for marker development, trait-mapping, and comparative studies in Pakistan, with the choice of reference guided by study material and objectives.

Repeat landscape and annotation: The assembled GC content is 34.5%, and repetitive element analysis indicated that 51.86% of the assembly (106.45 Mb of 205.25 Mb) consists of interspersed repeat regions, including retroelements (15.88%), DNA transposons (9.07%), and unclassified interspersed elements (26.92%) (Table 3; Table S4). In total, 26,215 protein-coding genes were identified in the genome. Functional annotation revealed that 81.07% of these genes matched known protein domains or biological pathways, indicating broad annotation coverage and functional relevance.

Figure 1 Sampling site and plant material for the Pakistan-origin *Moringa oleifera genome*. (A) Satellite base-map of Lahore highlighting Bagh-e-Jinnah with an inset zoom on the GCU Botanical Garden sampling area (31.556747° N, 74.328372° E) © Google. (B) Representative morphology of *M. oleifera,* including plantation view, leaves (used for extraction), fruit pod/seeds, and flowers

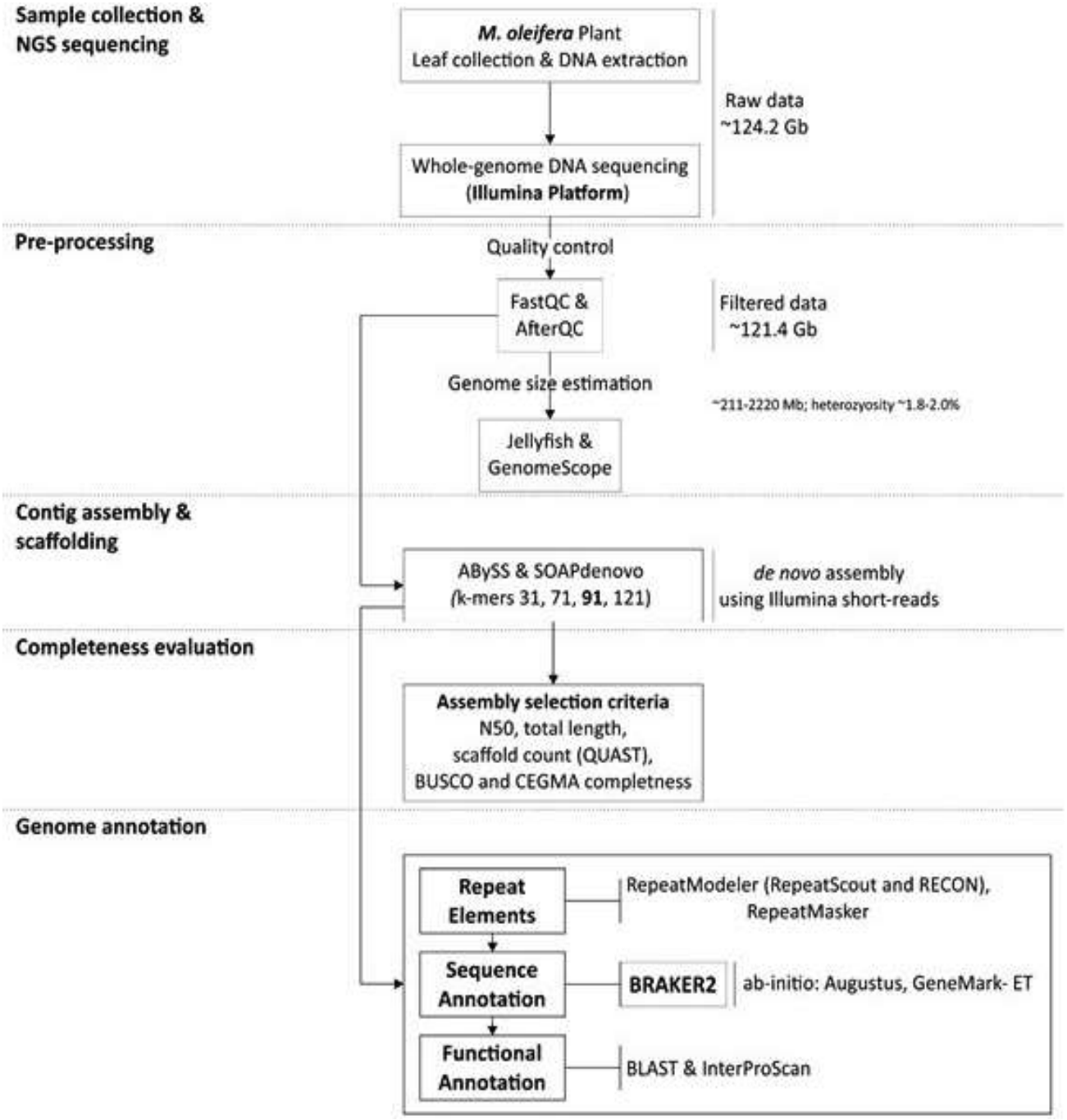

Figure 2: Workflow for *Moringa oleifera* genome assembly and annotation. Fresh leaves were collected and subjected to Illumina PE150 whole-genome sequencing. Reads were quality filtered (FastQC/AfterQC) and used for genome survey (Jellyfish/GenomeScope), followed by de novo assembly with ABySS and SOAPdenovo across multiple k-mers. The best assembly was selected using QUAST statistics and completeness (BUSCO/CEGMA), then repeats were identified (RepeatModeler/RepeatMasker) and genes predicted with BRAKER2 (GeneMark-ET + AUGUSTUS) and functionally annotated (InterProScan/BLAST)

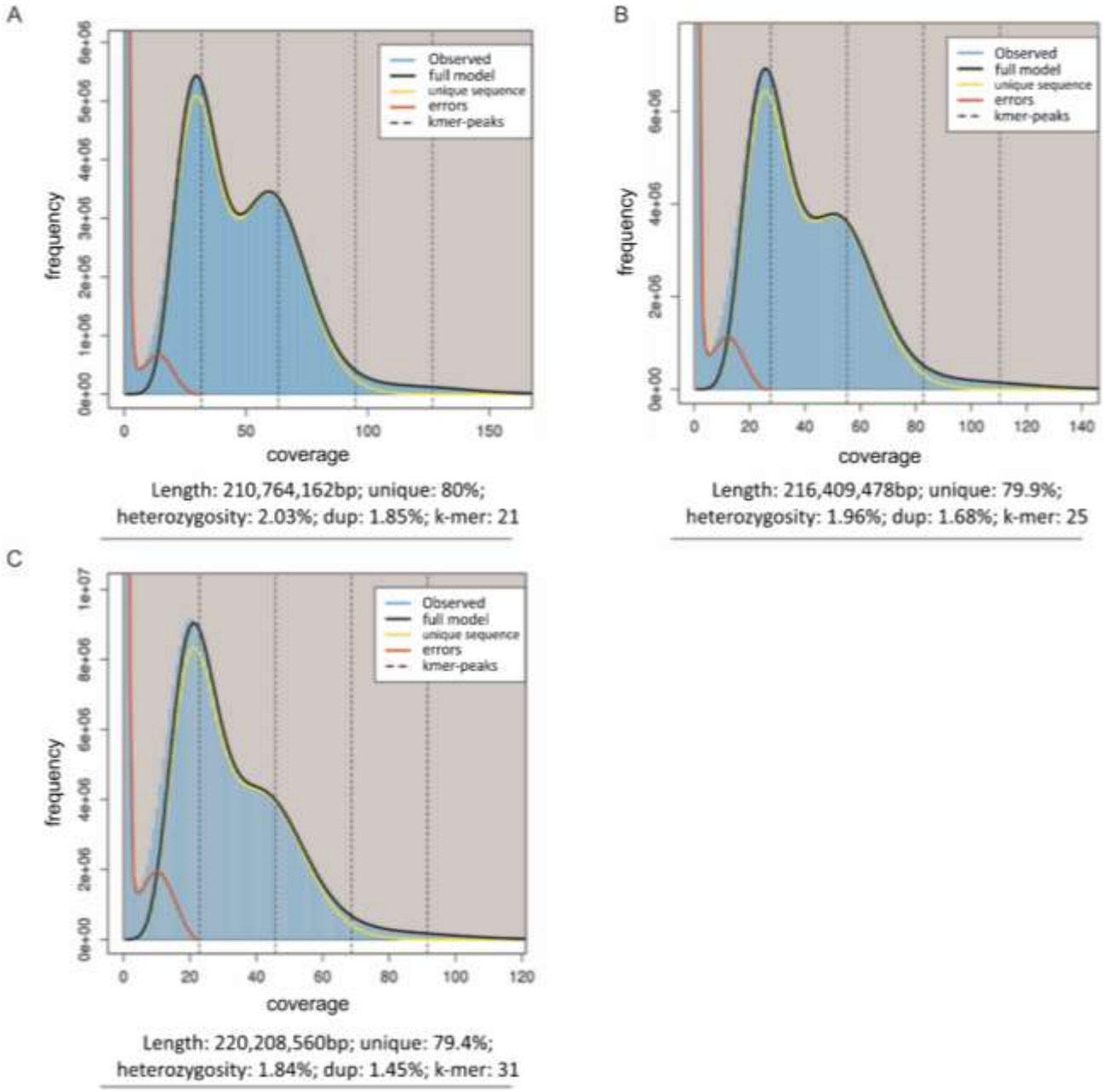

Figure 3: Genome survey of *Moringa oleifera* using k-mer frequency spectra (GenomeScope). K-mer models for (A) K = 21, (B) k -25, and (C) k = 31. The fitted distributions indicate an estimated size of ~210—220 Mb, unique sequence ~79.4—80.0%, heterozygosity ~1.84—2.03%, and short-duplication rate ~1.45—1.85% across k values. Vertical dashed lines mark the error peak and the heterozygous/homozygous coverage modes. Estimates at k = 31 (Length 220,208,560 bp; unique 79.4%; heterozygosity 1.84%; duplication 1.45%) were used to guide assembly parameterization.

Table 1 Assembly statistics for candidate de novo assemblies of the Pakistan-origin *Moringa oleifera* genome. Metrics were computed with QUAST on scaffolds ≥1 kb. The selected build (ABySS, k = 91) shows higher contiguity and a more compact assembly size relative to SOAPdenovo. Full metric panels and additional k-mers are provided in Table S3.

Moringa oleifera Genome Assembly Statistics		
	Assembler	
Parameter	**ABySS (k = 91)**	**SOAPdenovo**
Total scaffolds	13,872	492,699
Total length (Mb)	205.2483	262.013
Total length (bp)	205,248,313	262,012,530
Longest scaffold (bp)	395,443	27836
Scaffold N50 (bp)	17279	1232

Table 2 Genome assembly completeness assessment. BUSCO was run against the embryophyta_odb9 dataset (n = 1440) and CEGMA against 248 core eukaryotic genes. The selected assembly (ABySS, k = 91) shows substantially higher completeness than SOAPdenovo.

Assembler	**BUSCO (n:1440, embryophyta_odb9)**			**CEGMA (n = 248)**		
	Complete%	Fragmented %	Missing (%)	Complete (%)	Complete + Partial (%)	Missing (%)
ABySS (k = 91)	**83.9**	7.5	8.6	**84.68**	**96.77**	3.23
SOAPdenovo	26.2	17.6	56.2	26.61	60.89	39.11

Notes: BUSCO categories: "Complete" includes single-copy and duplicated BUSCOs; "Fragmented" and "Missing" follow BUSCO defaults. CEGMA values shown as "Complete" and "Complete + Partial".

Table 3 Repeat content in the assembled genome of *Moringa oleifera*. Repeat annotations were generated with RepeatMasker v4.1.0 using a species-specific library built by RepeatModeler v2.0.1 (RECON + RepeatScout). Statistics are computed on scaffolds ≥1 kb (total assembly length 205,248,313 bp; 13,872 sequences).

Class	**Number of elements**	**Length occupied (bp)**	**Percentage of genome (%)**
Retroelements	36483	32585386	15.88
SINEs	142	29487	0.01
LINEs	6971	6813393	3.32
LTR elements	29370	25742506	12.54
DNA transposons	23141	18617872	9.07
Unclassified-Interspersed	88757	55245172	26.92
Total interspersed repeats	–	**106448430**	**51.86**
Small RNA	8772	10312333	5.02
Satellites	260	79406	0.04

Supplementary Table S1 Reported properties, applications and uses of *Moringa oleifera* plant parts.

***Moringa oleifera* Leaves**	**Properties**	▪ Antioxidant ▪ Flavonoids ▪ Minerals and Vitamins ▪ Amino acids	(Gopalakrishnan *et al.*, 2016; Gopalakrishnan *et al.*, 2016; Oyeyinka and Oyeyinka, 2018; Vergara-Jimenez *et al.*, 2017)
	Applications	▪ Food supplements ▪ Source of proteins	
	Benefits	▪ Therapeutic ▪ Stabilizes cell structure ▪ Promotes metabolism	
***Moringa oleifera* seeds**	**Properties**	▪ Antioxidant ▪ Anti-microbial ▪ Anti-inflammatory ▪ Phenolic ▪ Bio-active ▪ Compounds ▪ Anti-fungal	(Brilhante et al., 2017; Gopalakrishnan *et al.*, 2016)
	Applications	▪ Oil extraction ▪ Water purification ▪ Lubricant ▪ Medicinal	
	Benefits	▪ Skin remedies ▪ Liver Health	
Ben oil (*M. oleifera*)	**Properties**	▪ Anti-aging ▪ Antioxidant ▪ Exfoliant ▪ Preservative ▪ Anti-inflammatory ▪ High oxidation stability	(Vergara-Jimenez *et al.*, 2017; Pareek *et al.*, 2023)
	Applications	▪ Soap ▪ Lubricant ▪ Food (Edible oil) ▪ Cosmetics (skin, hair, perfume base)	
	Benefits	▪ Skin remedies ▪ Rheumatic oil for arthritic joints ▪ Excellent moisturizer	

Supplementary Table S2 Sequencing summary for the *Moringa oleifera* (*M. oleifera*) genome (Illumina HiSeq, PE150). Raw and post-filtering ("clean") statistics are shown. Read counts are in millions (M); bases are in gigabases (Gb); quality is reported as the percentage of bases with Q≥20 (Q20) or Q≥30 (Q30).

Metric	**Raw reads**	**Clean reads**
Read length (bp)	150 bp	149 bp
Read pairs (million)	173.11	172.12
Total bases (Gb)	124.20	121.40
Q20 bases (%)	97.1	97.3
Q30 bases (%)	92.1	92.5
GC content of reads (%)	37.77	37.59

Supplementary Table S3 Assembly statistics across ABySS k-mer values. QUAST metrics are shown for scaffolds ≥1,000 bp. The selected build (k = 91) maximized contiguity (N50 = 17,279 bp) with the fewest scaffolds.

Assembly (k)	Longest scaffold	Scaffolds (>= 1000 bp)	N50
31	974,48	23,545	6,340
71	312,914	28,711	11,614
91	**395,443**	**13,872**	**17,279**
121	318,204	35,488	14,614

Note: Statistics computed with QUAST v5.0.2 on scaffolds ≥1,000 bp

Supplementary Table S4 Repeat content by class and subclass in the assembled genome of *Moringa oleifera*. Repeat annotations were generated with RepeatMasker v4.1.0 using a species-specific library built by RepeatModeler v2.0.1 (RECON + RepeatScout). Statistics are computed on scaffolds ≥1 kb (total assembly length 205,248,313 bp; 13,872 sequences).

Class	Number of elements	Length occupied (bp)	Percentage of genome (%)
Retroelements	36483	32585386	15.88
SINEs	142	29487	0.01
LINEs	6971	6813393	3.32
LTR elements	29370	25742506	12.54
DNA transposons	23141	18617872	9.07
Unclassified-interspersed	88757	55245172	26.92
Total interspersed repeats	–	106448430	51.86
Rolling circles	410	321098	0.16
Small RNA	8772	10312333	5.02
Satellites	260	79406	0.04
Simple repeats	74014	2945867	1.44
Low complexity	18918	942157	0.46

Note: Percentages are relative to the assembly (205,248,313 bp). "Total interspersed repeats" = Retroelements + DNA transposons + Unclassified-interspersed. Other classes (e.g., simple repeats, low complexity) are listed separately and may overlap with interspersed repeats; therefore values need not sum to 100%.

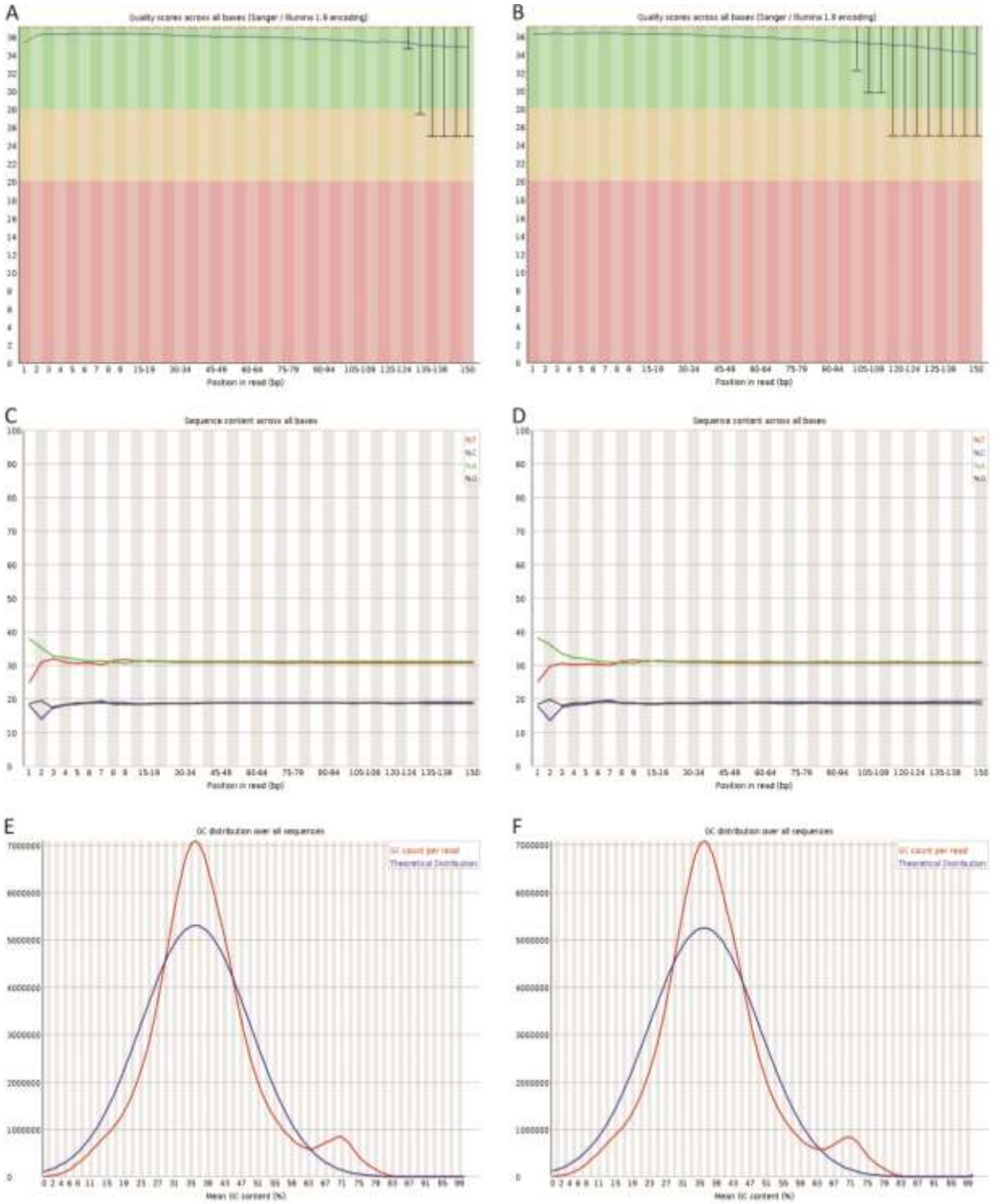

Figure S4. Illumina read quality for the Moringa oleifera genome (PE150). (A–B) Per-base Phred scores for read 1 (A) and read 2 (B) show uniformly high quality across cycles (majority ≥Q30). (C–D) Per-base nucleotide composition profiles for read 1 (C) and read 2 (D) are stable after the initial cycles, consistent with random genomic libraries. (E–F) GC-content distributions for read 1 (E) and read 2 (F) are unimodal and centered near the assembly GC (≈35–38%), indicating low contamination. Summary metrics: raw data ~124.2 Gb; clean data ~121.4 Gb; ≥92% bases Q≥30.

DISCUSSION

In the past decade, next-generation sequencing, especially whole-genome and whole-exome sequencing, has become a highly effective approach in crop genomics, providing deep insights into genomic architecture and enabling advances in agronomic traits and agricultural productivity (Edwards and Batley, 2010; Pérez-de-Castro *et al.*, 2012; Schreiber *et al.*, 2018). Today, these techniques allow researchers to resolve the complex genetic structure of crops and identify variants linked to desirable traits, thereby stimulating the breeding of improved crop varieties with increased resilience, yield, and nutritional value (Buch *et al.*, 2023; Panahi *et al.*, 2025). In addition, genome sequencing and annotation provide a useful platform for studying mechanisms governing plant growth promotion, protection, and colonization (Guo *et al.*, 2015).

This study presents a de novo, short-read draft genome for *M. oleifera* originating from Pakistan (~205.2 Mb; 13,872 scaffolds; N50 = 17,279 bp), predicted 26,215 protein coding genes, and recovers ≈ 83.9% embryophyte BUSCOs together with 84.68% CEGMA genes. The best de novo assembly was obtained at k-mer 91 by ABySS. The genome has a GC content of 34.5%. Interspersed repeats account for 51.86% of the assembly, comprising retroelements (15.88%), DNA transposons (9.07%), and unclassified interspersed repeats (26.92%). These metrics establish a usable reference for downstream analysis and provide gene models with broad functional support. This short-read assembly prioritizes gene recovery and completeness under heterozygosity and repeat constraints; the previously published long-read assembly from India offers higher contiguity but derives from different germplasm and technology, It is therefore consider a complementary reference rather than a direct comparator. (Tian *et al.*, 2015; Shyamli *et al.*, 2021). Together, the two assemblies expand the genomic toolkit for *Moringa* by providing coverage across South Asian germplasm: the present study contributes the first Pakistan-origin reference, enabling population-level analysis, marker development, and trait mapping anchored to regional material, whereas the Bhagya resource offers a high-contiguity scaffold for species-wide comparative work. In practice, the choice of reference should be guided by study material and objectives, and the availability of both resources enables side-by-side assessment of Pakistan and India germplasm where appropriate (Table S3).

The pipeline prioritized completeness and gene recovery under short-read constraints: (i) exploring multiple k-mers and assemblers with objective selection criteria (N50/L50, %Ns, BUSCO); (ii) de novo repeat modeling before gene prediction to reduce TE-derived artifacts; and (iii) BRAKER2-guided structural annotation to improve exon-intron boundary accuracy. The resulting gene set shows broad functional support (≈ 81.07% with domain and/or pathway evidence), indicating that annotation quality is suitable for downstream analysis even when contiguity is modest.

The gene catalog and repeat landscape can be leveraged instantly to pursue biological ends, for example, families involved in stress response, antioxidant metabolism, and primary nutrient pathways can be queried to generate refined hypotheses for tolerance and nutritional traits (Garcia-Caparros *et al.*, 2021; Rao and Zheng, 2025). This Pakistan-origin reference can facilitate the discovery of markers and aid allele mining in local breeding programs, as well as comparative population genetics.

Short-read assemblies often fail to fully resolve long repeats and highly heterozygous regions, which likely explains the lower contiguity and smaller assembled size compared with long-repeat-aware builds (Eisenhofer *et al.*, 2024). Future improvements will incorporate long reads, Hi-C scaffolding, and transcriptome-assisted curation to increase contiguity, enhance repeat resolution, and improve BUSCO recovery. Combining several accessions into a *Moringa* pangenome is expected to reveal structural variation that underlies regional adaptation and traits, adding additional utility to the Pakistan and Bhagya references.

Taken together, this work delivers a Pakistan-origin reference genome for *M. oleifera* that is immediately useful for functional genomics and breeding applications in the region and that complements the existing Bhagya assembly. The availability of both resources opens a practical path for side-by-side comparative analyses across South Asian germplasm and provides a stronger foundation for molecular breeding, biotechnology, and evolutionary studies in *Moringa*.

Conclusion: This study provides a de novo, annotated draft genome of *M. oleifera* originating from Pakistan (~205.2 Mb; 13,872 scaffolds; N50 = 17,279 bp; 26,215 genes; ≈83.9% BUSCO). This Pakistan-origin reference offers an immediate resource for marker development, functional genomics, and comparative analyses within *Moringa*, particularly for regional germplasm and breeding programs. By making these data publicly available, community use is enabled for trait discovery in nutrition and stress adaptation. While contiguity is constrained by short-read sequencing, the resource establishes a practical baseline that can be upgraded with long reads and Hi-C, and integrated into future *Moringa* pangenome efforts.

Funding: This work was supported by Virtual University of Pakistan.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability Statement: The assembled genome has been deposited to NCBI under GenBank assembly accession GCA_021560355.1 (Moringa_Oleifera_1.0). The corresponding WGS project is JAKELQ000000000 (this study uses version JAKELQ010000000, scaffolds JAKELQ010000001–JAKELQ010013872; BioProject PRJNA765934; BioSample SAMN21592341).

Author contributions: MT. P. conceived and supervised the study. SH. A. developed the methodology, implemented the pipeline, and drafted the manuscript. ME. B. and T. H. validated data, review results and revised the final version. All authors read and approved the final manuscript.

REFERENCES

Ahmed, I., M. Islam, W. Arshad, A. Mannan, W. Ahmad and B. Mirza (2009). High-quality plant DNA extraction for PCR: an easy approach. J. Appl. Genet. 50(2): 105–107. doi: 10.1007/BF03195661.

Akulova, V.S., V.V. Sharov, A.I. Aksyonova, Y.A. Putintseva, N.V. Oreshkova, S.I. Feranchuk, D.A. Kuzmin, I.N. Pavlov, Y.A. Litovka and K.V. Krutovsky (2020). De novo sequencing, assembly and functional annotation of *Armillaria borealis* genome. BMC Genomics 21(7): 534. doi: 10.1186/s12864-020-06964-6

Bhattacharya, A., P. Tiwari, P.K. Sahu and S. Kumar (2018). A review of the phytochemical and pharmacological characteristics of *Moringa oleifera*. J. Pharm. Bioallied Sci. 10(4): 181–191. doi: 10.4103/JPBS.JPBS_126_18.

Blum, M., H.Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, *et al.,* (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 49(D1): 344–354. doi: 10.1093/nar/gkaa977.

Boratyn, G.M., C. Camacho, P.S. Cooper, G. Coulouris, A. Fong, N. Ma, T.L. Madden, W.T. Matten, S.D. McGinnis, Y. Merezhuk, *et al.,* (2013). BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 41: W29–W33. doi: 10.1093/nar/gkt282.

Brilhante, R.S.N., J.A. Sales, V.S. Pereira, D. d.S.C.M. Castelo-Branco, R.d.A. Cordeiro, C.M.d.S. Sampaio, M.d.A.N. Paiva, J.B.F.d. Santos, J.J.C. Sidrim and M.F.G. Rocha (2017). Research advances on the multiple uses of *Moringa oleifera*: a sustainable alternative for socially neglected population. Asian Pac. J. Trop. Med. 10(7): 621–630. doi: 10.1016/j.apjtm.2017.07.002.

Brůna, T., K.J. Hoff, A. Lomsadze, M. Stanke and M. Borodovsky (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 3(1). doi: 10.1093/nargab/lqaa108.

Brůna, T., A. Lomsadze and M. Borodovsky (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinforma. 2(2). doi: 10.1093/nargab/lqaa026.

Buch, K., A. Kaushik, U. Mishra, S. Beese, S. Samanta and R. Singh (2023). Unravelling the complexity of plant breeding through modern genetic techniques and tools: a review. Int. J. Plant Soil Sci.35: 97-105. doi: 10.9734/ijpss/2023/v35i213950.

Chen, S., T. Huang, Y. Zhou, Y. Han, M. Xu and J. Gu (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18(3): 80. doi: 10.1186/s12859-017-1469-3

Edwards, D. and J. Batley (2010). Plant genome sequencing: applications for crop improvement. Plant Biotechnol. J. 8(1): 2–9. doi: 10.1111/j.1467-7652.2009.00459.x.

Eisenhofer, R., J. Nesme, L. Santos-Bay, A. Koziol, S.J. Sørensen, A. Alberdi and O. Aizpurua (2024). A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. Microbiol. Spectr. 12(4): e03590-23. doi: 10.1128/spectrum.03590-23.

Flynn, J.M., R. Hubley, C. Goubert, J. Rosen, A.G. Clark, C. Feschotte and A.F. Smit (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117(17): 9451–9457. doi: 10.1073/pnas.1921046117.

Garcia-Caparros, P., L. De Filippis, A. Gul, M. Hasanuzzaman, M. Ozturk, V. Altay and M.T. Lao (2021). Oxidative stress and antioxidant metabolism under adverse environmental conditions: a review. Bot. Rev. 87: 421-466. doi: 10.1007/s12229-020-09231-1.

Gopalakrishnan, L., K. Doriya and D.S. Kumar (2016). *Moringa oleifera*: A review on nutritive importance and its medicinal application. Food Sci. Hum. Wellness 5(2): 49–56. doi: 10.1016/j.fshw.2016.04.001.

Guo, S., X. Li, P. He, H. Ho, Y. Wu and Y. He (2015). Whole-genome sequencing of *Bacillus subtilis XF-1* reveals mechanisms for biological control and multiple beneficial properties in plants. J. Ind. Microbiol. Biotechnol. 42(6): 925–937. doi: 10.1007/s10295-015-1612-y.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29(8): 1072–1075. doi: 10.1093/bioinformatics/btt086.

Islam, Z., S.R. Islam, F. Hossen, K. Mahtab-ul-Islam, M.R. Hasan, and R. Karim (2021). *Moringa oleifera* is a prominent source of nutrients with potential health benefits. Int. J. Food Sci. 6627265. doi: 10.1155/2021/6627265.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, *et al.*, (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 1(1): 18. doi: 10.1186/2047-217X-1-18.

Manekar, S.C. and S.R. Sathe (2018). A benchmark study of k-mer counting methods for high-throughput sequencing. Gigascience. 7(12): giy125. doi: 10.1093/gigascience/giy125.

Oyeyinka, A.T. and S.A. Oyeyinka (2018). *Moringa oleifera* as a food fortificant: recent trends and prospects. J. Saudi Soc. Agric. Sci. 17(2): 127–136. doi: 10.1016/j.jssas.2016.02.002

Panahi, B., N.H. Gharajeh and H.M. Jalaly (2025). Advances in barley germplasm diversity characterization through next-generation sequencing approach. Genet. Resour. Crop Evo. 72: 3829-3843. doi: 10.1007/s10722-024-02196-9.

Pareek, A., M. Pant, M.M. Gupta, P. Kashania, Y. Ratan, V. Jain, A. Pareek and A.A. Chuturgoon (2023). *Moringa oleifera*: an updated comprehensive review of its pharmacological activities, ethnomedicinal, phytopharmaceutical formulation, clinical, phytochemical, and toxicological aspects. Int. J. Mol. Sci. 24(3): 2098. doi: 10.3390/ijms24032098.

Parra, G., K. Bradnam and I. Korf (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23(9): 1061–1067. doi: 10.1093/bioinformatics/btm071.

Pérez-de-Castro, A.M., S. Vilanova, J. Cañizares, L. Pascual, J.M. Blanca, M.J. Díez, J. Prohens and B. Picó (2012). Application of genomic tools in plant breeding. Curr. Genomics. 13(3): 179–195. doi: 10.2174/138920212800543084.

Rao, M.J. and B. Zheng (2025). The role of polyphenols in abiotic stress tolerance and their antioxidant properties to scavenge reactive oxygen species and free radicals. Antioxidants 14(1): 74. doi: 10.3390/antiox14010074.

Schreiber, M., N. Stein and M. Mascher (2018). Genomic approaches for studying crop evolution. Genome Biol. 19(1): 140. doi: 10.1186/s13059-018-1528-8.

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva and E.M. Zdobnov (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19): 3210–3212. doi: 10.1093/bioinformatics/btv351.

Shyamli, P.S., S. Pradhan, M. Panda and A. Parida (2021). De novo whole-genome assembly of *Moringa oleifera* helps identify genes regulating drought stress tolerance. Front. Plant Sci. 12: 766999. doi: 10.3389/fpls.2021.766999.

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J.M. Jones and I. Birol (2009). ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6): 1117–1123. doi: 10.1101/gr.089532.108.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34: 435–439. doi: 10.1093/nar/gkl200.

Tian, Y., Y. Zeng, J. Zhang, C.G. Yang, L. Yan, X.J. Wang, C.Y. Shi, J. Xie, T.Y. Dai, L. Peng, *et al.*, (2015). High quality reference genome of drumstick tree (*Moringa oleifera Lam.*), a potential perennial crop. Sci. China Life Sci. 58(7): 627–638. doi: 10.1007/s11427-015-4872-x.

Vergara-Jimenez, M., M.M. Almatrafi and M.L. Fernandez (2017). Bioactive components in *Moringa Oleifera* leaves protect against chronic disease. Antioxidants 6(4): 91. doi: 10.3390/antiox6040091.