



# De novo Whole-Genome Assembly of *Moringa oleifera* Helps Identify Genes Regulating Drought Stress Tolerance

P Sushree Shyamli<sup>1,2†</sup>, Seema Pradhan<sup>1†</sup>, Mitrabinda Panda<sup>1,2</sup> and Ajay Parida<sup>1\*</sup>

<sup>1</sup> Institute of Life Sciences, An Autonomous Institute Under Department of Biotechnology Government of India, NALCO Square, Bhubaneswar, India, <sup>2</sup> Regional Centre for Biotechnology, NCR Biotech Science Cluster, Faridabad, India

## OPEN ACCESS

### Edited by:

Agnieszka Zmienko,  
Institute of Bioorganic Chemistry  
(PAS), Poland

### Reviewed by:

Haidong Yan,  
University of Georgia, United States  
Jonathan Elias Maldonado,  
Pontificia Universidad Católica  
de Chile, Chile

### \*Correspondence:

Ajay Parida  
drajayparida@gmail.com

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Plant Systems and Synthetic Biology,  
a section of the journal  
Frontiers in Plant Science

**Received:** 02 September 2021

**Accepted:** 12 November 2021

**Published:** 14 December 2021

### Citation:

Shyamli PS, Pradhan S, Panda M  
and Parida A (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

Abiotic stresses, especially drought stress, are responsible for heavy losses in productivity, which in turn poses an imminent threat for future food security. Understanding plants' response to abiotic stress at the molecular level is crucially important for mitigating the impacts of climate change. *Moringa oleifera* is an important multipurpose plant with medicinal and nutritional properties and with an ability to grow in low water conditions, which makes the species an ideal candidate to study the regulatory mechanisms that modulate drought tolerance and its possible use in agroforestry system. In the present communication, we report whole-genome sequencing (WGS) of this species and assemble about 90% of the genome of *M. oleifera* var. Bhagya into 915 contigs with a N50 value of 4.7 Mb and predicted 32,062 putative protein-coding genes. After annotating the genome, we have chosen to study the heat shock transcription factor (HSF) family of genes to analyze their role in drought tolerance in *M. oleifera*. We predicted a total of 21 HSFs in the *M. oleifera* genome and carried out phylogenetic analyses, motif identification, analysis of gene duplication events, and differential expression of the HSF-coding genes in *M. oleifera*. Our analysis reveals that members of the HSF family have an important role in the plant's response to abiotic stress and are viable candidates for further characterization.

**Keywords:** *Moringa oleifera*, whole-genome sequencing (WGS), abiotic stress, HSFs, differential expression

## INTRODUCTION

*Moringa oleifera*, which belongs to the family Moringaceae, is native to the Indian subcontinent and naturalized in tropical and subtropical regions around the world. The genus *Moringa* has 13 species (Gandji et al., 2018), of which two species, viz., *M. oleifera* Lam and *M. concanensis* Nimmo, are found in India (Pandey et al., 2011). *M. oleifera* is a versatile plant and a repository of essential phytochemicals such as tannins, sterols, terpenoids, flavonoids, saponins, anthraquinones, alkaloids, and reducing sugar, present in leaves, pods, and seeds (Gopalakrishnan et al., 2016). Apart from its nutritional properties, *M. oleifera* is also known for its ability to grow in semi-arid environments. The two most common abiotic stresses, namely, hypersalinity and drought, are responsible for the majority of crop loss worldwide. Plants engage in various cellular mechanisms to deal with such extreme conditions. One of the most common mechanisms is the production of osmolytes such as proline (Liang et al., 2013). The ultimate consequence is the enhanced production of reactive oxygen species (ROS) under prolonged or severe water deficit (Hayat et al., 2012).

ROS production causes damage through membrane peroxidation. Polyphenols, which are antioxidants and produced by plants, help to retain ROS at less damaging levels. Part of the *M. oleifera* has medicinal attributes because of the possession of these antioxidants (Sofidiya et al., 2006).

The accumulation of secondary metabolites is believed to be strongly dependent on the growing conditions (Lommen et al., 2008). Polyphenols and proline often demonstrate various bioactivities such as guarding the cell against stress, involvement in metal chelation (Sharma and Dietz, 2009), and functioning as a ROS scavenger (Liang et al., 2013). An extensive range of experiments have elucidated that plants exposed to drought stress accumulate higher concentrations of secondary metabolites, antioxidants, and proline. Increased concentrations of phenols as well as nitrogen-containing substances are reported in almost all classes of natural products such as alkaloids, cyanogenic glucosides, or glucosinolates (Selmar and Kleinwächter, 2013). High concentration of cellular proline of up to 80% of amino acid pool has been reported under stress and as low as 5% under normal conditions (Delauney and Verma, 1993). It is evident that water stress plays a role in the accumulation pattern and concentrations of these compounds. Secondary metabolites have become relevant due to their crucial applications in medicinal, nutritive, and cosmetic purposes as well as their importance in plant stress physiology (Edreva et al., 2008).

A recent study investigated the presence of polyphenolic compounds, antioxidant activities, proline accumulation, and their distribution in the different parts of *M. oleifera* plant under varying water regimes. Results from the research indicated that drought stress influences the synthesis and concentration of the osmolyte proline and polyphenolic compounds in *M. oleifera* plant. Their findings showed that under water scarcity, *M. oleifera* responds by an upsurge biosynthesis and accumulation of phenolic, condensed tannin, and proline contents in all plant parts but mostly in leaves (Chitiyo et al., 2021). This indicates that accumulating osmoprotectants and antioxidant compounds such as phenolic compounds are some of the mechanisms employed by *M. oleifera* to cope up with drought-induced oxidative stress and dehydration.

Therefore, *M. oleifera* represents a promising species capable of minimizing the adverse effects of drought stress and can enhance the soil of arid regions (Boumenjel et al., 2020). In this study, we have undertaken *de novo* assembly and annotation of *M. oleifera* var Bhagya. A comparison of its genome with genomes of other plants revealed a number of orthologous groups that are important for growth and survival of plants. We have also identified the members of heat shock transcription factors (HSFs) in the genome of *M. oleifera* and analyzed their expression in response to drought stress.

## MATERIALS AND METHODS

### Plant Material, Growth Conditions, and Stress Treatment

Seeds for four varieties of *M. oleifera*, namely, Bhagya, ODC3, PKM1, and PKM2, were collected from Krishi Vigyan Kendra,

Dhenkanal, Odisha. The seeds were soaked in distilled water overnight and planted in a mix of soil rite and vermicompost in a ratio of 3:1. Plants were grown under controlled condition: 14 h/10 h: light/dark; 28°C/25°C ± 2; 65% RH, for 4 weeks. Young leaves of *M. oleifera* var. Bhagya were collected for high-quality DNA isolation for whole-genome sequencing (WGS). We subjected the 30-day-old young plants of *M. oleifera* to drought stress by withholding water for different time periods. We observed noticeable physical changes in the plants after 10 days. However, prolonged stress affected the quality of tissue and the subsequent RNA isolation yielded poor quality. Therefore, we decided to use the tissues after 10 days of drought stress (**Supplementary Figure 1**). Plants in control conditions were watered regularly as per requirement. Leaves and root tissue samples were collected from control and stress-treated plants, frozen in liquid nitrogen, and stored at −80°C. Three biological replicates were collected for each sample group.

### High-Quality Genomic DNA Extraction and Whole-Genome Sequencing

High molecular weight genomic DNA was isolated from the young leaves of *M. oleifera* var. Bhagya using the CTAB-based method (Doyle, 1991) with some modifications. DNA libraries were prepared for sequencing on PacBio Sequel and Illumina HiSeq 2500 platforms to generate long and short reads, respectively (AgriGenome Labs Pvt. Ltd, India). Raw reads were filtered to remove low-quality reads with FastQCv 0.11.9 with default parameters.<sup>1</sup>

### Genome Assembly and Annotation

The cleaned reads were subjected to Kmergenie (Chikhi and Medvedev, 2014) to predict the optimal k-value and assembly size that were found to be 96 and 392,308,674 bp, respectively. *De novo* assembly was performed at AgriGenome Pvt. Ltd<sup>2</sup> using MaSuRCA (Zimin et al., 2013), and the assembly with the default k-mer sizes using both long and short reads was selected to be optimal. The repeat sequences from the assembled genome were masked using RepeatMasker V2.2 (Tempel, 2012) with *Arabidopsis* model. CDSs were predicted from the masked assembly using Augustus V3.3.1 (Stanke et al., 2004) with *Arabidopsis* trained model. We found 32,062 predicted genes in the assembly.

The predicted genes were annotated using our in-house pipeline. The predicted genes were compared with Uniprot database using BLASTX program with E-value cutoff of 10<sup>−5</sup>. The best BLASTX hit based on query coverage, identity, similarity score, and description of each gene was filtered out. The top BLASTX hit of each gene and the organism name was extracted. The gene ontology (GO) terms Molecular Function (MF), Cellular Component (CC), and Biological Process (BP) for genes were mapped using the current annotation sets<sup>3</sup> and GO Slim terms.<sup>4</sup>

<sup>1</sup><https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

<sup>2</sup><https://www.aggenome.com/>

<sup>3</sup><http://ftp.ebi.ac.uk/pub/databases/GO/goa/>

<sup>4</sup><http://ftp.ebi.ac.uk/pub/databases/GO/goa/goslim/goaslim.map>

## Repeat Elements, Non-coding Regions, and Quality Assessment for Genome Assembly

We identified repetitive elements through both RepeatModeler v.1.0.10,<sup>5</sup> which employed RECON and RepeatScout to predict interspersed repeats, and then obtained the consensus repeat library. The Tandem Repeat Finder software<sup>6</sup> was used to identify the tandem repeats with the parameters: 2 7 7 80 10 50 2000 -1 3 -m. LTR-RTs were identified with LTR\_FINDER\_parallel (Ou and Jiang, 2019) with default parameters.

Non-coding RNAs were predicted as follows: tRNAscan-SE v 2.0 software (Lowe and Eddy, 1997) was used to predict the tRNAs with eukaryotic parameters. miRNAs and snRNAs were detected using Infernal cmscan v 1.1.4 (Nawrocki and Eddy, 2013) to search the Rfam database (Griffiths-Jones et al., 2005). The ncRNA database of *Arabidopsis thaliana* was downloaded from Phytozome v 12.1<sup>7</sup> and used for annotating the rRNA in *M. oleifera* genome sequence.

We used two strategies to assess the quality of genome assembly: BUSCO alignment and LTR Assembly index (LAI). The BUSCO pipeline<sup>8</sup> used datasets from the plant lineage (EmbryophytaOrthoDB release 9; Kriventseva et al., 2015) to assess the number of complete BUSCOs represented in the *M. oleifera* genome. LAI uses LTR-RTs to evaluate assembly continuity and has been widely used to evaluate the assembly quality of plant genomes with high contents of repetitive sequences. The LTR-RTs identified in this study were used as input for LTR\_retriever (Ou et al., 2018) to calculate LAI.

## Ortho Group Inference

Amino acid sequences for primary transcripts of 11 plant species, namely, *A. thaliana*, *Brassica oleracea*, *Carica papaya*, *Cicer arietinum*, *Citrus sinensis*, *Medicago truncatula*, *Oryza sativa*, *Populus trichocarpa*, *Ricinus communis*, *Theobroma cacao*, and *Vitis vinifera*, were downloaded from Phytozome v 12.1 (see text footnote 7). Primary transcripts for *M. oleifera* were determined with TransDecoder<sup>9</sup> and produced 23,744 genes. Orthologous groups were determined using OrthoFinder v. 2.2.0 (Emms and Kelly, 2019). The data available in the results of this analysis were used to build the species tree using FigTree.<sup>10</sup>

## The RNA-Seq Assembly of *Moringa oleifera* Transcriptome and Differential Gene Expression

The RNA-Seq reads for various tissues of *M. oleifera* var. Bhagya were downloaded from NCBI SRA database (see “Data availability”) and filtered with Trimmomatic-0.39

with default parameters (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36-threads 8) (Bolger et al., 2014). The high-quality reads were assembled into transcripts using the “-genome\_guided\_bam” feature of Trinity software (Grabherr et al., 2011), which utilized the RNA-Seq reads aligned to the *M. oleifera* genome reported in this study as input. We removed redundant transcripts and retained the longest coding representative of each transcript using cdhit (cdhit-est-c 0.98; Li and Godzik, 2006) and TransDecoder v 5.5.0 (see text footnote 9). We also used control and drought-stressed tissue samples (leaves and root) of *M. oleifera* var. Bhagya for RNA-seq. Total RNA was extracted from frozen plant tissues using RNA extraction kit (STRN50, Sigma Aldrich, St. Louis, MO, United States) following the manufacturer’s protocol. RNA concentration and purity of each sample were confirmed using Nanodrop 2000 Spectrophotometer (Thermo Scientific, Wilmington, DE, United States). The integrity of extracted RNA was also ensured by resolving on 1.2% agarose gel containing 18% formaldehyde and checking on automated electrophoresis system (Agilent 4200 TapeStation, Agilent, United States). Samples with a RIN > 7 were taken for library preparation. The libraries were prepared taking three biological replicates of each sample (Control\_Leaf, Control\_Root, Drought\_Leaf and Drought\_Root) using the TruSeq Stranded mRNA Library Prep kit (Illumina, United States). After quality assessment, the libraries were pooled and sequenced on NextSeq550 platform (Illumina, United States). The reads obtained were demultiplexed with bcl2fastq software (Illumina, United States) and filtered for quality and adapters with Trimmomatic-0.39 (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36-threads 8) (Bolger et al., 2014). The filtered reads were mapped onto the assembled transcriptome of *M. oleifera* using Bowtie2 (Langmead and Salzberg, 2012), and the *in silico* expression profile was generated with RSEM (RNA-Seq by Expectation–Maximization<sup>11</sup>) and edgeR.<sup>12</sup> Gene enrichment for the differentially expressed genes was carried out with KOBAS<sup>13</sup> and AgriGO with *A. thaliana* as reference using the Singular Enrichment Analysis (SEA) tool.<sup>14</sup>

## Identification of Heat Shock Transcription Factors in *Moringa oleifera* Genome

In order to be inclusive, we followed two strategies to identify all HSFs in the *M. oleifera* genome. First, we downloaded amino acid sequences of HSFs for all species from PlantTFDB<sup>15</sup> and used BlastP<sup>16</sup> to search protein sequences of *M. oleifera* against them to identify HSFs based on homology (with an e-value of 10<sup>-5</sup>, choosing the best alignment). Then, we

<sup>5</sup><http://www.repeatmasker.org/RepeatModeler/>

<sup>6</sup><https://tandem.bu.edu/trf/trf.html>

<sup>7</sup><https://phytozome.jgi.doe.gov/>

<sup>8</sup>[busco.ezlab.org](http://busco.ezlab.org)

<sup>9</sup><https://github.com/TransDecoder/TransDecoder>

<sup>10</sup><https://github.com/rambaut/figtree/releases>

<sup>11</sup><https://github.com/deweylab/RSEM>

<sup>12</sup><https://bioconductor.org/packages/release/bioc/html/edgeR.html>

<sup>13</sup><http://kobas.cbi.pku.edu.cn/>

<sup>14</sup><http://bioinfo.cau.edu.cn/agriGO/>

<sup>15</sup><http://planttfdb.gao-lab.org/>

<sup>16</sup><https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>



downloaded the HMM profile for DNA-binding domain (DBD) of HSFs from Pfam (PF00447) and used the HMM profile to identify DBD in those HSFs identified in the previous step. All sequences that did not contain a DBD were removed. Finally, we removed redundancy by aligning the protein sequences and removed duplicate sequences to get a final list of 21 HSFs in *M. oleifera*.

## Nomenclature and Classification of Heat Shock Transcription Factors

The HSFs were mapped onto the scaffolds in whole-genome assembly of *Moringa* and named serially as MolHSF1–MolHSF21 according to their mapped position on the genome contigs. The HSFs were classified into three classes according to their similarities to the AtHSFs. The coiled coil (heptad or HR-A/B) domains were predicted using Marcoil software.<sup>17</sup> Although MolHSFs 9, 11, and 13 did not show the presence of coiled coil structure (which is one of the requirements for a protein to qualify as a HSF), they have been included in this study due to their unusually long HSF-DBD. ProtParam tool in ExPASy (expasy.org) was used to determine pI of the peptides, and NLS were predicted using NLStradamus<sup>18</sup> and NLSMapper<sup>19</sup> and nuclear export signals (NESs) with,<sup>20</sup> respectively.

## Phylogenetic Analysis and Conserved Domain Identification

The amino acid sequences of the 21 MolHSFs were aligned by ClustalW and phylogenetic tree were constructed using MEGAX software (Kumar et al., 2018) with neighbor-joining method, Poisson correction, and 1,000 bootstraps. Domains were identified using Pfam and SMART databases and visualized using DOG 2.0.<sup>21</sup> We used MEME<sup>22</sup> to identify conserved motifs in the MolHSF protein sequences, and gene structure was analyzed using GSDS.<sup>23</sup> The protein sequences of HSFs in *Moringa* were compared to those from *Arabidopsis*, *T. cacao*, *C. papaya*, and *O. sativa*. The evolutionary history was inferred by using the maximum-likelihood method and JTT matrix-based model. The multiple sequence alignments were visualized with JalView<sup>24</sup> (Waterhouse et al., 2009).

## Analysis of Duplication Events and Promoter Sequence

The duplication events for MolHSFs were identified using default parameters of MCScanX software (Wang et al., 2012). The ratios of non-synonymous (Ka) and synonymous (Ks) substitutions were calculated with TBtools (Chen et al., 2020). Promoter sequences were isolated by extracting 2,000 bp

sequence upstream of TSS of the MolHSFs using BEDTools (Quinlan and Hall, 2010). Conserved motifs were identified with PlantCARE.<sup>25</sup>

## Isolation of RNA and Quantitative Real-Time PCR

The relative expression of the selected HSFs was quantified using qRT-PCR. The FASTA sequences of the selected transcripts were retrieved, and primers were designed using Integrated DNA Technologies (IDT) Primer design online tool by selecting the generic option and the following criterion: amplicon size of 100–150 bp; primer length of 18–23 bases; melting temperature of 57–63°C; and GC content of 40%–60% (Supplementary Table 7).

For each RNA sample, 1 µg of total RNA was treated with RNase-free DNase 1 (Sigma-AMPD1, St. Louis, United States) and reverse-transcribed to synthesize cDNA using the first-strand cDNA synthesis kit (K1612, Thermo Scientific, MA, United States). The qRT-PCR was performed on QuantStudio-5 real-time PCR system (Thermo Fisher Scientific, United States) with SYBR green chemistry (Applied Biosystems, United States) in three technical and two biological replicates. The expression was normalized by housekeeping gene actin since we observed the stable expression of *M. oleifera* actin in our analysis (Ct values were in the range of 21–22). Each reaction (5 µl SYBR Green, 1 µl template cDNA, 1 µl each of the primers (10 µM), and 2 µl RNase-free water) was performed three times with the following program: 50°C (2 min), 95°C (10 min) followed by 40 cycles of 95°C (15 s), 53°C (1 min), and melt curve stage of 95°C (15 s) and 65°C (15 s). The expression values were calculated using the comparative  $2^{-\Delta\Delta C_t}$  method. The qRT-PCR analysis is presented graphically by taking the RQ values. Correlation between the log<sub>2</sub> fold change values of *in silico* and qRT-PCR gene expression data was charted on MS Office Excel.

## RESULTS

### Assembly and Annotation of *Moringa oleifera* Genome

Sequencing the genome of *M. oleifera* var. Bhagya generated 23.02 Gb of total reads on the SMRT platform and 90.9 Gb of short reads on the Illumina sequencing platform (Supplementary Table 1). This produced a cumulative coverage of more than 300× since the reported genome size of *Moringa* is 315 Mb (Tian et al., 2015). The clean reads were assembled into 915 contigs with MaSuRCA (Zimin et al., 2013) and yielded a genome assembly with an N50 value of 4.7 Mb, the longest contig length of 13.8 Mb, and representing about 281 Mb (~90% of total genome) of *Moringa* genome (Table 1).

The BUSCO analysis revealed that 95.9% of core genes from the Embryophyta were present in the *M. oleifera*

<sup>17</sup><https://bcf.isb-sib.ch/Delorenzi/Marcoil/index.html>

<sup>18</sup><http://www.moseslab.csb.utoronto.ca/NLStradamus/>

<sup>19</sup>[http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS\\_Mapper\\_form.cgi](http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi)

<sup>20</sup><http://www.cbs.dtu.dk/services/NetNES/>

<sup>21</sup><http://dog.biocuckoo.org/>

<sup>22</sup><http://meme-suite.org/>

<sup>23</sup><http://gsds.gao-lab.org/>

<sup>24</sup><http://www.jalview.org/>

<sup>25</sup><http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>

**TABLE 1** | Attributes of the whole-genome assembly of *M. oleifera*.

<b>De novo genome assembly</b>	
Total number of contigs	915
Length of the longest contig	13,807,473 bp
Total length	281,946,330 bp
N50	4,719,167 bp
Number of N's per 100 kbp	0.25
GC%	37.82
<b>Annotation</b>	
Number of predicted protein-coding genes	31,056
Average gene length (bp)	1,842
Average CDS length (bp)	1,798

<b>Repeat elements</b>			
<b>Type</b>	<b>Number of elements</b>	<b>Length occupied (bp)</b>	<b>Percentage of sequence</b>
Retroelements	4,524	1,949,564	0.69
SINEs	1	100	0.00
Penelope	0	0	0.00
LINES	357	54,439	0.02
CRE/SLACS	0	0	0.00
L2/CR1/Rex	0	0	0.00
R1/LOA/Jockey	0	0	0.00
R2/R4/NeSL	0	0	0.00
RTE/Bov-B	26	3,871	0.00
L1/CIN4	331	50,568	0.02
LTR elements	4,166	1,895,025	0.67
BEL/Pao	0	0	0.00
Ty1/Copia	2,521	1,269,453	0.45
Gypsy/DIRS1	1,636	624,847	0.22
Retroviral	0	0	0.00
DNA transposons	2,783	443,823	0.16
hobo-Activator	170	41,476	0.01
Tc1-IS630-Pogo	37	1,820	0.00
En-Spm	0	0	0.00
MuDR-IS905	0	0	0.00
PiggyBac	0	0	0.00
Tourist/Harbinger	81	18,723	0.01
Other (Mirage,P-element, Transib)	0	0	0.00
Rolling circles	329	31,500	0.01
Unclassified	171,757	142,408,786	50.51
Total interspersed repeats		144,802,173	51.36
Small RNA	473	58,633	0.02
Satellites	69	7,634	0.00
Simple repeats	63,799	2,641,831	0.94
Low complexity	15,520	779,798	0.28

genome, of which 94.4% were complete and single-copy BUSCOs (**Supplementary Table 2**). In addition to that, LTR assembly index (LAI) score of 10.27 is an indication of a good-quality genome assembly (**Supplementary Table 2**).

A total of 32,062 protein-coding genes were predicted from the assembled genome, which came down to 31,056 after removing redundancy. Gene ontology terms were assigned using the annotations from UniProt/Swiss-Prot database and

processing them through in-house pipelines. Processes related to “binding” and “catalytic activity” were over-represented in the molecular functions category, while biological processes such as “cellular” and “metabolic processes” were most abundantly represented (**Figure 1A**). These sequences were searched against the UniProt/Swiss-Prot database,<sup>26</sup> Pfam database,<sup>27</sup> KEGG,<sup>28</sup> and COG databases,<sup>29</sup> which annotated 21,634 of the predicted genes (**Figure 1B**).

There are two previous reports of WGS and assembly of *M. oleifera*. Tian et al. (2015) reported 33,332 contigs with N50 value of 1.14 Mb and a total of 19,465 protein-coding genes, while Chang et al. (2018) reduced the number of assembled contigs to 22,329 albeit with a N50 of 0.9 Mb and predicted 18,451 protein-coding genes. In comparison with the assembly reported in this study, we assembled the genome to 915 contigs with the N50 value of 4.7 Mb and 31,056 non-redundant protein-coding genes. We also used a more recently available whole-genome assembly for *M. oleifera* in the database of NCBI (PRJNA268707), along with the one reported by Chang et al. (2018) to perform a more exhaustive comparison between the assemblies (**Supplementary Table 3**). All data presented in this report point to the better quality and accuracy of the genome assembly reported for *M. oleifera*.

## Orthologous Groups and Important Gene Families

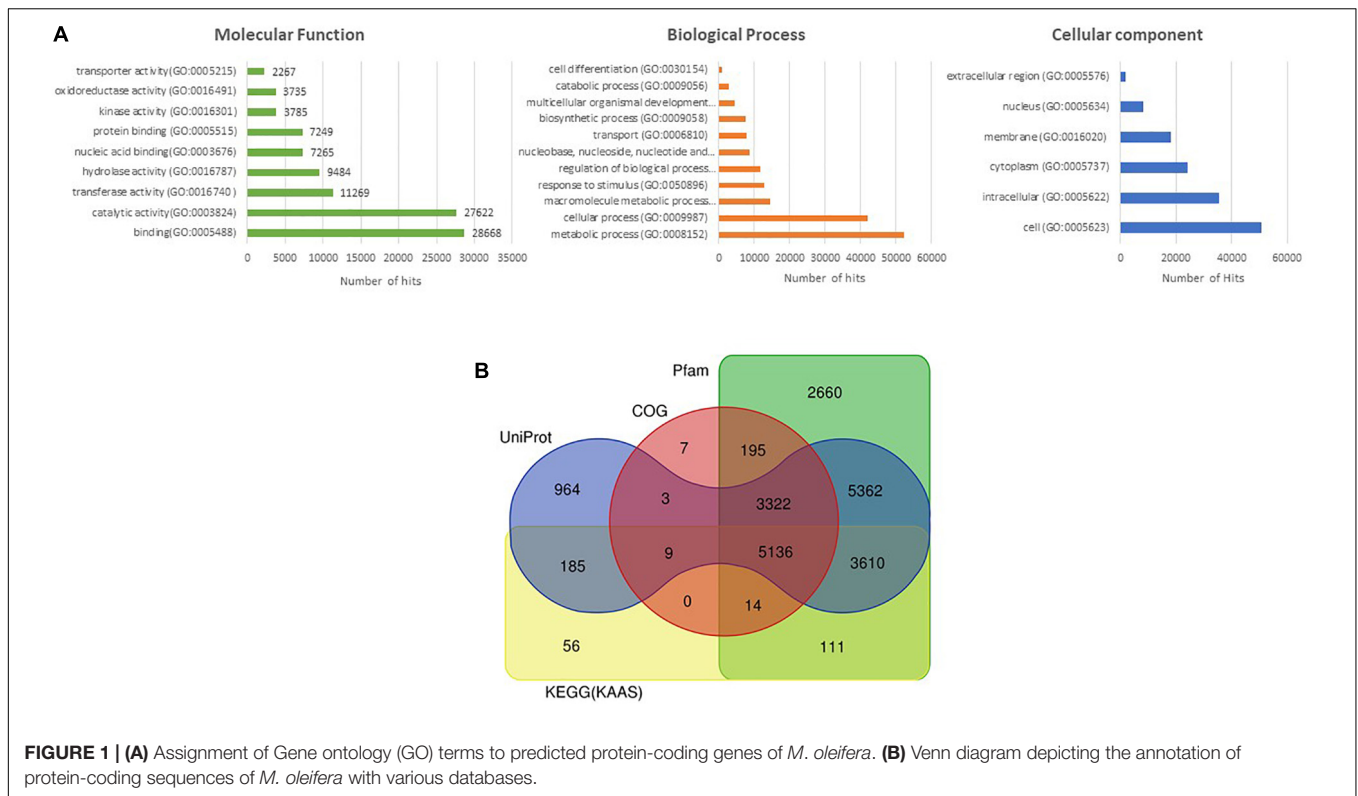
The course of evolution has ensured the conservation of gene families important for survival of organisms. One of the methods to determine these gene families is to identify the orthologous groups after a phylogenetic comparison of the protein sequences in a group of varied organisms. We identified orthogroups after comparing a total of 378,983 protein sequences from 12 different plants, namely, *M. oleifera*, *A. thaliana*, *B. oleracea*, *C. papaya*, *C. arietinum*, *C. sinensis*, *M. truncatula*, *O. sativa*, *P. trichocarpa*, *R. communis*, *T. cacao*, and *V. vinifera*. Of the 6,574 orthogroups that contained representative protein sequences from all 12 species, a total of 536 were single-copy orthogroups (**Supplementary Table 4**). The species tree constructed with the data from this analysis revealed that *M. oleifera* had more similarity with *C. papaya* and *T. cacao*, while forming a separate clade from *A. thaliana* and *B. oleracea* (**Figure 2A**). These results are also reflected in the number of orthologs shared by *M. oleifera* with the various plants. It shares the most orthologs with *T. cacao*, while sharing the least with *O. sativa* (**Figure 2B**). The protein sequences for all plants in the single-copy orthologous groups were annotated against the protein sequences of *A. thaliana* in the UniProt/Swiss-Prot database using BlastP to identify the important gene families. The analysis revealed that a number of ion channel transporters, pentatricopeptide repeat-containing proteins, chaperones, and transcription factors such as WRKY, HSFs, and CCCH zinc finger transcription factors are conserved,

<sup>26</sup><https://www.uniprot.org/uniprot/?query=reviewed:yes>

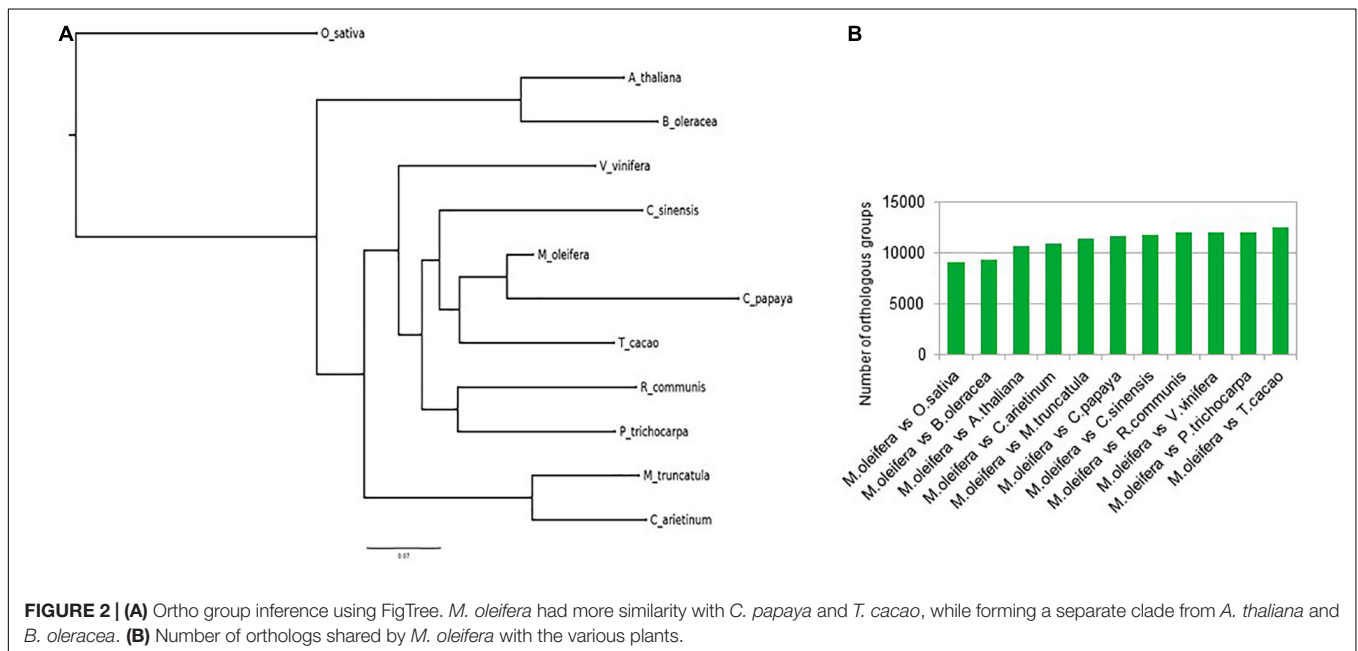
<sup>27</sup>[http://ftp.ebi.ac.uk/pub/databases/Pfam/current\\_release/Pfam-A.fasta.gz](http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.fasta.gz)

<sup>28</sup><https://www.genome.jp/kegg/kaas/>

<sup>29</sup><https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/>



**FIGURE 1 | (A)** Assignment of Gene ontology (GO) terms to predicted protein-coding genes of *M. oleifera*. **(B)** Venn diagram depicting the annotation of protein-coding sequences of *M. oleifera* with various databases.



**FIGURE 2 | (A)** Ortho group inference using FigTree. *M. oleifera* had more similarity with *C. papaya* and *T. cacao*, while forming a separate clade from *A. thaliana* and *B. oleracea*. **(B)** Number of orthologs shared by *M. oleifera* with the various plants.

indicating their importance in plant growth and survival (Supplementary Table 5).

## Simple Sequence Repeats in *Moringa oleifera* Genome

Simple sequence repeats (SSRs) are one of the most widespread molecular markers in plant genomes and have

proven to be the popular choice for estimating genetic diversity and molecular breeding (Pan et al., 2020). A total of 92,163 SSRs were identified in 594 out of the 915 contigs of the assembled genome of *M. oleifera* (Table 2). Tetranucleotide repeats were most abundant followed by di- and trinucleotide repeats. Genic SSRs are especially helpful in assaying functional diversity and marker-assisted selection

(MAS) in plants (Li et al., 2021). Therefore, we also mined SSRs in the protein-coding sequences of *M. oleifera* and identified 3,362 SSRs in 31,056 protein-coding genes, with the trinucleotide repeats being most abundant (Table 2). We have designed primers targeting genic SSRs, which can be used as a valuable resource for crop improvement (Supplementary Table 6).

## Transcriptome of *Moringa oleifera* and Differential Expression of Genes in Response to Drought Stress

The filtered, high-quality reads from leaf, root, stem, flower, and pod tissues of *M. oleifera* var. Bhagya were mapped onto the genome to generate a genome-guided transcriptome. We obtained 1,38,362 assembled transcripts, which we reduced to 1,37,132 transcripts after removing redundancy and retaining the longest CDS for each transcript. The assembly statistics and quality assessment suggest that this transcriptome assembly is of good quality (Supplementary Table 8). RNA-Seq of the control and drought-treated leaf and root tissues of *M. oleifera* was done. The RNA-Seq generated about 753.5

million reads which were filtered for quality. About 711 million reads (94.37%) were retained after removing low quality reads and adapter sequences. We mapped the high-quality reads from each sample onto the transcriptome of *M. oleifera* assembled in this study to develop a representative heat map of differentially expressed genes (Figure 3). Overall, there was a more pronounced effect of drought on leaves as compared to roots. A homology-based search against the UniProt/Swiss-Prot database revealed a number of differentially expressed genes to be transcription factors/regulators. These included Zn finger-containing transcription factors, ethylene-responsive transcription factors, WRKY, bHLH, and a number of transporters and genes associated with secondary metabolite production. In addition to these, we also found a number of heat shock proteins (HSPs) and HSFs to be differentially expressed in response to drought stress (Supplementary Table 9). Enrichment analysis revealed the over-representation of genes related to “Biosynthesis of secondary metabolites” and “response to stimuli.” This provided us with an idea of the regulatory mechanisms regulating drought stress response in *M. oleifera* (Supplementary Figure 2).

## Heat Shock Transcription Factors in the *Moringa oleifera* Genome

Heat shock transcription factors have emerged as important regulators of response to abiotic stress in plants, and although *Moringa* is mainly valued for its medicinal properties, the fact that it can withstand drought conditions makes this plant a good source of genomic resources for plant improvement. We identified 21 HSFs in the genome of *M. oleifera* (MolHSF1-MolHSF21) ranging in length from 110 to 1,530 amino acids. All of the MolHSFs were predicted to contain either NLS or NES or both (Table 3).

## Phylogenetic Analysis and Classification

The protein sequences of MolHSFs were grouped according to the homology between them. The phylogenetic tree thus generated reflects the classification of the MolHSFs into three classes, namely, A, B, and C (Figure 4A). These similarities were also reflected in the gene structures of the MolHSFs where members of a clade displayed similar intron–exon composition (Figure 4B). We aligned the members of Class A and Class B, respectively, and found the distinct features typical to the classes. The members of Class A are comprised of HR-A and HR-B motifs separated by an insert (Figure 5A), while the members of Class B contained a HR-A/B motif with no discernible insert between the two (Figure 5B). We then compared the protein sequences for HSFs from four other plant species, namely, *A. thaliana*, *O. sativa*, *C. papaya*, and *T. cacao*, with the MolHSFs. The consequent phylogenetic tree revealed that MolHSFs displayed closer homology to the Hsfs from *C. papaya* and *T. cacao*, while *O. sativa* was a clear outlier (Figure 6). These results are similar to those obtained from orthologous groups formed by comparing the protein sequences of various plants in this study.

**TABLE 2 |** Results of microsatellite (SSRs) identification in *M. oleifera*.

SSRs in whole genome	
Total number of sequences examined	915
Total size of examined sequences (bp)	281,946,330
Total number of identified SSRs	92,163
Number of SSR-containing sequences	594
Number of sequences containing more than 1 SSR	465
Number of SSRs present in compound formation	10,902
Distribution to different repeat-type classes	
Unit size	Number of SSRs
Dinucleotide repeats	27,200
Trinucleotide repeats	17,276
Tetranucleotide repeats	31,789
Pentanucleotide repeats	10,734
Hexanucleotide repeats	5,164
SSRs in protein-coding genes	
Total number of sequences examined	31,056
Total size of examined sequences (bp)	57,197,102
Total number of identified SSRs	3,632
Number of SSR-containing sequences	3,109
Number of sequences containing more than 1 SSR	366
Number of SSRs present in compound formation	88
Distribution to different repeat-type classes	
Unit size	Number of SSRs
Dinucleotide repeats	131
Trinucleotide repeats	1,815
Tetranucleotide repeats	1,414
Pentanucleotide repeats	109
Hexanucleotide repeats	163

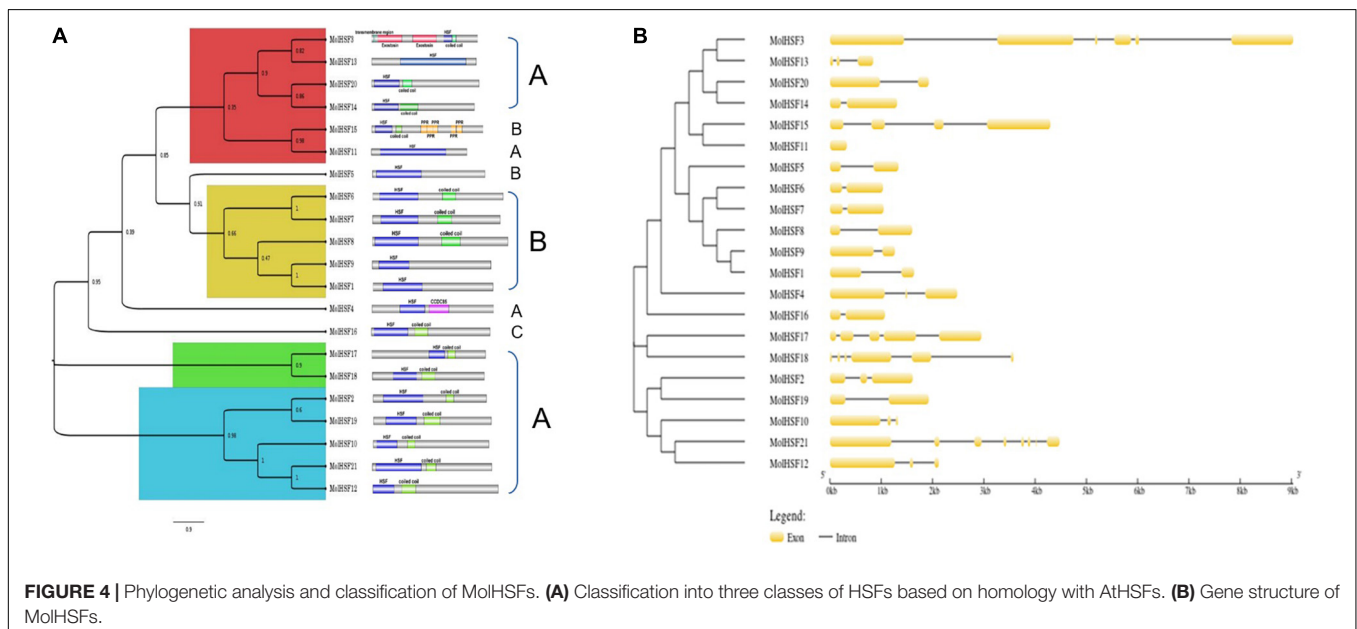






**TABLE 3** | Properties of HSFs identified in *M. oleifera*.

Gene ID	Length	Stability	pI	NLS	NES
MolHSF1	286	Unstable	6.9	P	P
MolHSF2	411	Unstable	5.67	P	P
MolHSF3	1,530	Unstable	Undetermined	P	–
MolHSF4	575	Unstable	4.95	P	P
MolHSF5	233	Unstable	8.31	P	P
MolHSF6	277	Unstable	4.95	–	P
MolHSF7	320	Unstable	5.14	–	P
MolHSF8	291	Unstable	5.96	P	–
MolHSF9	367	Unstable	Undetermined	–	P
MolHSF10	363	Unstable	4.63	–	P
MolHSF11	110	Stable	5.12	–	P
MolHSF12	472	Unstable	5.08	P (low score) Most probably cytoplasmic and nucleus	P
MolHSF13	144	Unstable	5.72	P	P
MolHSF14	398	Unstable	4.89	P	P
MolHSF15	650	Unstable	9.33	P	–
MolHSF16	325	Unstable	5.66	P	–
MolHSF17	679	Unstable	5.69	P	–
MolHSF18	445	Unstable	5.42	P	P
MolHSF19	362	Unstable	4.95	P	P
MolHSF20	397	Unstable	5.84	P	P
MolHSF21	623	Unstable	5.25	P	P



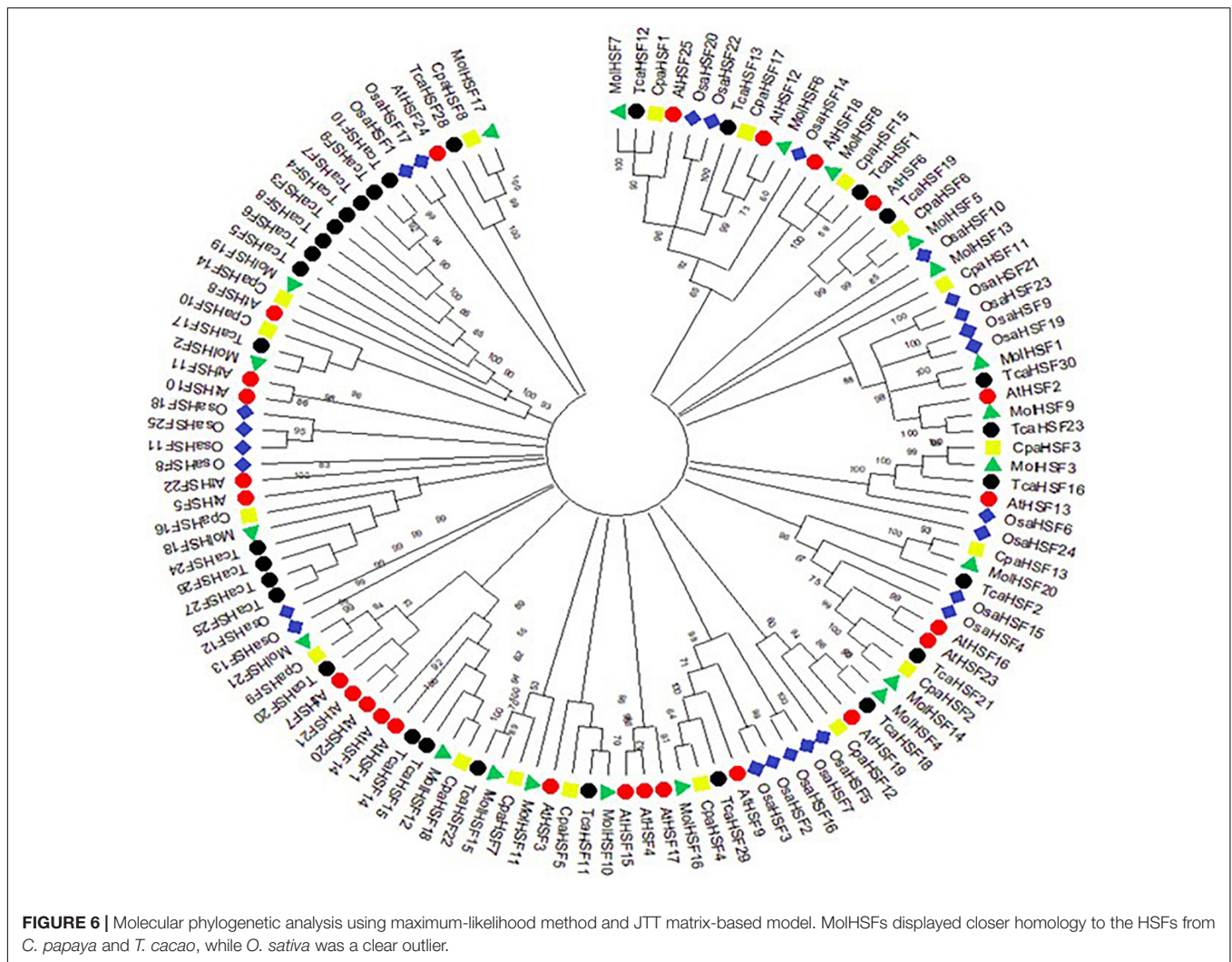
an upregulation of >5. MolHSF-8 was upregulated in root tissues of all four varieties of *M. oleifera* (Figure 8). Other than PKM1 leaf tissues, a downregulated trend was observed in six HSF genes of Bhagya, ODC3, and PKM2. Out of the 21 identified HSFs, 19 HSFs showed gene expression in leaf and root tissues of all four varieties, except MolHSF13 and MolHSF18. Moreover, the qPCR results showed that there was good consistency between the expression levels of the genes analyzed by qRT-PCR and their levels detected using RNA-seq. Consequently, the qPCR analysis results

confirmed that the data we obtained from RNA-seq are reliable (Supplementary Figure 4).

## DISCUSSION

Good-quality genome assemblies of important plant species have become a necessity for mining valuable genomic resources for genetic enhancement and crop improvement. Contemporary methods have made it possible to generate more complete





**TABLE 4 |** Identification of conserved motifs in promoter region of MolHSFs.

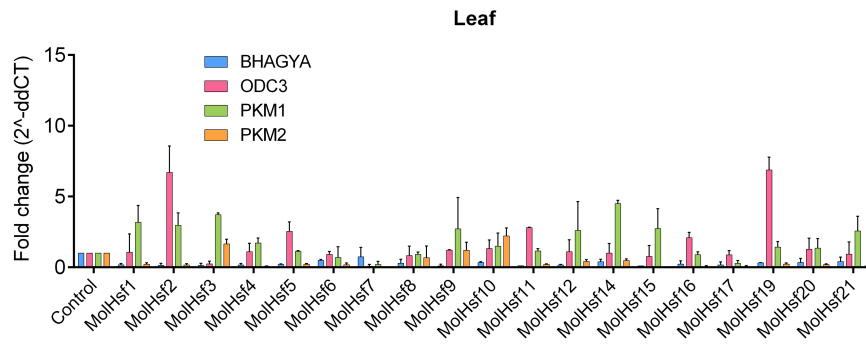
Type of Motif	Description	Motif sequence
ABRE	cis-acting element involved in the abscisic acid responsiveness	TACGTG; ACGTG; CACGTA; CACGTG; CGCAGCTGTC; CGTACGTGCA; AACCCGG
STRE	cis-regulatory element able to mediate transcriptional induction by different forms of stress	AGGGG
LTR	cis-acting element involved in low-temperature responsiveness	CCGAAA
TC-rich repeats	cis-acting element involved in defense and stress responsiveness	ATTCTCTAAC;GTTTTCTTAC
MBS	MYB binding site involved in drought inducibility	CAACTG
DRE core	cis-acting regulatory element involved in cold and dehydration response	GCCGAC
MYB	cis-acting element involved in drought responsiveness	TAACCA; CAACCA; CAACAG; CAACTG; TAACTG; TAACCA

over the course of evolution. This makes HSFs an essential part of plants' development and survival.

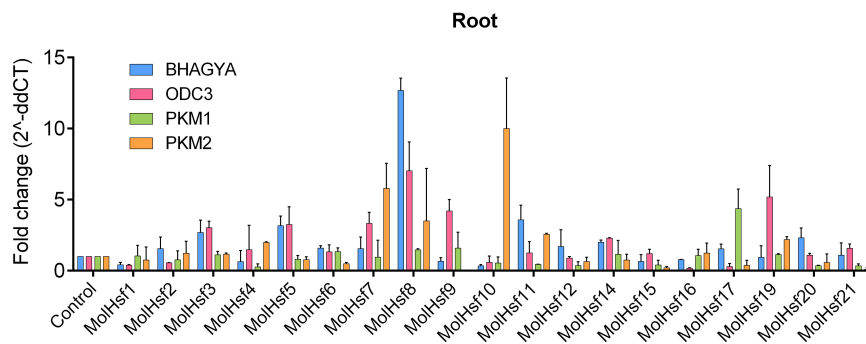
Many transcription factor families have been documented in their ability to regulate plants' response to abiotic stress. Most of these studies have also reported a number of motifs associated with stress response that are present in the promoter regions of these transcription factors. We analyzed the promoter sequences of the *MolHSF* genes and found a number of such motifs like

"ABRE," "STRE," "LTR," "MYB" etc. to be present. To further study the involvement of *MolHSFs* in drought stress response, we analyzed the expression of *MolHSFs* in the young plants of *Moringa* subjected to drought stress.

Basal-level expression of 19 *MolHSFs* was seen in Bhagya and PKM2 var. in leaf tissues. While in PKM1, most *MolHSFs* were upregulated, in ODC3 var. *MolHSF*-2,-19 were significantly upregulated. Most *MolHSFs* showed a higher expression in root



**FIGURE 7** | RT-PCR expression analysis of 19 MolHSF genes in leaf tissues of Bhagya, ODC3, PKM1, and PKM2 varieties of *M. oleifera*.



**FIGURE 8** | RT-PCR expression analysis of 19 MolHSF genes in root tissues of Bhagya, ODC3, PKM1, and PKM2 varieties of *M. oleifera*.

tissue as compared to the leaf, confirming the expectations that root is the first tissue that senses and is affected by drought stress and roots respond faster to stress than leaves, undergoing more complex gene regulation during water deprivation. MolHSF-8 presented high expression in all the four varieties of *Moringa oleifera*, suggesting an important regulatory role during drought stress. It was interesting to note that a contrasting trend was observed in the expression of MolHSFs of Bhagya and PKM1. While basal-level expression of most MolHSFs was seen in the leaf tissues of Bhagya, a high expression trend was observed in PKM1 leaf. Conversely, the trend was reversed in the root of Bhagya and PKM1 variety.

## CONCLUSION

*Moringa oleifera* is an important multipurpose plant with medicinal and nutritional properties and with the ability to tolerate drought, which makes it an ideal candidate to study the regulatory mechanisms that modulate drought tolerance and its possible use in agroforestry system. We carried out WGS of this species and assembled about 90% of the genome of *M. oleifera* var. Bhagya into 915 scaffolds with a N50 value of 4.7Mb and predicted 32,062 putative protein-coding genes. The 21 HSFs identified in *M. oleifera* were phylogenetically more similar to HSFs from *C. papaya* and *T. cacao* as compared to *A. thaliana* and *O. sativa*. Also, the Ka/Ks ratio of less than 1 for the

two pairs of tandemly duplicated MolHSFs suggests purifying selection, pointing to the conserved nature of these TFs over the course of evolution. This makes HSFs an essential part of plants' development and survival. qRT-PCR showed that *MolHSF8* could be a promising candidate for functional characterization in drought tolerance in plants. Further characterization of identified HSFs in *M. oleifera* and their functional validation in a panel of genotypes under varying abiotic stress conditions will help in divulging new sources of stress-resistant genes for improvement of this miracle plant.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: National Center for Biotechnology Information (NCBI) BioProject database under accession numbers PRJNA394193, PRJNA765946, PRJNA747889, and PRJNA756620.

## AUTHOR CONTRIBUTIONS

PSS contributed to the sample collection, experimental data collection, and visualization. SP was responsible for bioinformatics data analysis and equally contributed with PSS in experimental designing. AP and SP supervised the research work.



AP, SP, PSS, and MP wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was funded by the Institutional grant from Department of Biotechnology, Ministry of Science and Technology, Government of India.

## ACKNOWLEDGMENTS

PSS is thankful to Krishi Vigyan Kendra, Dhenkanal, Odisha, for providing seeds of four varieties used in this study. The authors are grateful to Agrigenome Pvt. Ltd. for the WGS work. PSS acknowledges DST-INSPIRE for the fellowship. MP is thankful to CSIR for the funding.

## REFERENCES

- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/BIOINFORMATICS/BTU170
- Boumenjel, A., Papadopoulos, A., and Ammari, Y. (2020). Growth response of *Moringa oleifera* (Lam) to water stress and to arid bioclimatic conditions. *Agroforest Syst.* 95, 823–833. doi: 10.1007/s10457-020-00509-2
- Chang, Y., Liu, H., Liu, M., Liao, X., Sahu, S. K., Fu, Y., et al. (2018). The draft genomes of five agriculturally important african orphan crops. *GigaScience* 8, 1–16. doi: 10.1093/gigascience/giy152
- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009
- Chikhi, R., and Medvedev, P. (2014). Informed and automated K-Mer size selection for genome assembly. *Bioinformatics* 30, 31–37. doi: 10.1093/bioinformatics/btt310
- Chitiyo, S. T., Ncube, B., Ndhkala, A. R., and Tsvuura, Z. (2021). Biochemical responses of *Moringa Oleifera* Lam. Plants to Graded Moisture Deficit. *South Afr. J. Bot.* 138, 41–49. doi: 10.1016/J.SAJB.2020.12.003
- Delauney, A. J., and Verma, D. S. (1993). Proline biosynthesis and osmoregulation in plants. *Plant J.* 4, 215–223. doi: 10.1046/J.1365-313X.1993.04020215.X
- Doyle, J. (1991). “DNA protocols for plants,” in *Molecular Techniques in Taxonomy*, eds G. M. Hewitt, A. W. B. Johnston, and J. P. W. Young (Berlin: Springer), 283–293. doi: 10.1007/978-3-642-83962-7\_18
- Edreva, A., Velikova, V., Tsonev, T., Dagnon, S., Gürel, A., Aktaş, L., et al. (2008). Stress-protective role of secondary metabolites: diversity of functions and mechanisms. *Gen. Appl. Plant Physiol.* 34, 67–78.
- Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 1–14. doi: 10.1186/s13059-019-1832-y
- Gandji, K., Chadare, F. J., Idohou, R., Salako, V. K., and Assogbadjo, A. E. (2018). Status and utilisation of *Moringa oleifera* lam?: a review. *Afr. Crop Sci. J.* 26, 137–156. doi: 10.4314/acsj.v26i1.10
- Gopalakrishnan, L., Doriya, K., and Kumar, D. S. (2016). *Moringa Oleifera*: a review on nutritive importance and its medicinal application. *Food Sci. Hum. Wellness* 5, 49–56. doi: 10.1016/j.fshw.2016.04.001
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33, D121–D124. doi: 10.1093/nar/gki081

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2021.766999/full#supplementary-material>

**Supplementary Figure 1** | *Moringa* varieties, viz., Bhagya, ODC3, PKM1, and PKM2, during drought stress treatment.

**Supplementary Figure 2** | Gene enrichment analysis for DEGs in response to drought stress in *M. oleifera*. (A) Enrichment of genes in KEGG pathways and (B) enrichment of GO terms in the DEGs.

**Supplementary Figure 3** | Conserved motifs of MoHSFs. The most conserved motifs were identified after alignment and visualized with MEME suite.

**Supplementary Figure 4** | (A) Graph depicting the trend of pattern of expression observed *in silico* (RNA-Seq) and with qRT PCR. (B) Scatter plot for the correlation between log2 values of fold change observed in digital expression analysis for RNA-Seq and validation with qRT PCR.

- Guo, J., Wu, J., Ji, Q., Wang, C., Luo, L., Yuan, Y., et al. (2008). Genome-wide analysis of heat shock transcription factor families in rice and *Arabidopsis*. *J. Genet. Genomics* 35, 105–118. doi: 10.1016/S1673-8527(08)60016-8
- Hayat, S., Hayat, Q., Alyemeni, M. N., Wani, A. S., Pichtel, J., and Ahmad, A. (2012). Role of proline under changing environments: a review. *Plant Signal. Behav.* 7, 1456–1466. doi: 10.4161/PSB.21949
- Khan, S. A., Li, M. Z., Wang, S. M., and Yin, H. J. (2018). Revisiting the role of plant transcription factors in the battle against abiotic stress. *Int. J. Mol. Sci.* 19:1634. doi: 10.3390/ijms19061634
- Kim, T., Samraj, S., Jiménez, J., Gómez, C., Liu, T., and Begcy, K. (2021). Genome-wide identification of heat shock factors and heat shock proteins in response to UV and high intensity light stress in lettuce. *BMC Plant Biol.* 21:185. doi: 10.1186/s12870-021-02959-x
- Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simão, F. A., Pozdnyakov, I. A., et al. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.* 43, D250–D256. doi: 10.1093/nar/gku1220
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Li, M., Xie, F., Li, Y., Gong, L., Luo, Y., Zhang, Y., et al. (2020). Genome-Wide Analysis of the heat shock transcription factor gene family in *Brassica juncea*: structure, evolution, and expression profiles. *DNA Cell Biol.* 39, 1990–2004. doi: 10.1089/dna.2020.5922
- Li, P. S., Yu, T. F., He, G. H., Chen, M., Zhou, Y. B., Chai, S. C., et al. (2014). Genome-wide analysis of the Hsf family in soybean and functional identification of GmHsf-34 involvement in drought and heat stresses. *BMC Genom.* 15:1009. doi: 10.1186/1471-2164-15-1009
- Li, Q., Su, X., Ma, H., Du, K., Yang, M., Chen, B., et al. (2021). Development of genic SSR marker resources from RNA-Seq Data in *Camellia japonica* and their application in the genus *camellia*. *Sci. Rep.* 11:9919. doi: 10.1038/s41598-021-89350-w
- Li, W., and Godzik, A. (2006). Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/BIOINFORMATICS/BTL158
- Li, Y. Z., Liu, Z., Zhang, K. Y., Chen, S. Y., Liu, M. J., and De Zhang, Q. (2020). Genome-Wide analysis and comparison of the DNA-binding one zinc finger gene family in diploid and tetraploid cotton (*Gossypium*). *PLoS One* 15:e0235317. doi: 10.1371/journal.pone.0235317
- Liang, X., Zhang, L., Natarajan, S. K., and Becker, D. F. (2013). Proline mechanisms of stress survival. *Antioxid. Redox Signal.* 19, 998–1011. doi: 10.1089/ARS.2012.5074

- Lommen, W. J. M., Bouwmeester, H. J., Schenk, E., Verstappen, F. W. A., Elzinga, S., and Struik, P. C. (2008). Modelling processes determining and limiting the production of secondary metabolites during crop growth: the example of the antimalarial artemisinin produced in *Artemisia annua*. *Acta Hort.* 765, 87–94. doi: 10.17660/ACTAHORTIC.2008.765.10
- Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955
- Lyu, T., Liu, W., Hu, Z., Xiang, X., Liu, T., Xiong, X., et al. (2020). Molecular characterization and expression analysis reveal the roles of Cys2/His2 zinc-finger transcription factors during flower development of *Brassica rapa* Subsp. *Chinensis*. *Plant Mol. Biol.* 102, 123–141. doi: 10.1007/s11103-019-00935-6
- Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509
- Ou, S., and Jiang, N. (2019). LTR\_FINDER\_parallel: parallelization of LTR\_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* 10, 1–3. doi: 10.1186/s13100-019-0193-0
- Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* 46:e126. doi: 10.1093/nar/gky730
- Pan, G., Chen, A., Li, J., Huang, S., Tang, H., Chang, L., et al. (2020). Genome-wide development of simple sequence repeats database for flax (*Linum usitatissimum* L.) and its use for genetic diversity assessment. *Genetic Resources Crop Evol.* 67, 865–874. doi: 10.1007/s10722-020-00882-y
- Pandey, A., Pradheep, K., Gupta, R., Nayar, E. R., and Bhandari, D. C. (2011). “Drumstick tree” (*Moringa oleifera* Lam.): a multipurpose potential species in India. *Genet. Resources Crop Evol.* 58, 453–460. doi: 10.1007/s10722-010-9629-6
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/BIOINFORMATICS/BTQ033
- Scharf, K. D., Berberich, T., Ebersberger, I., and Nover, L. (2012). The plant heat stress transcription factor (Hsf) family: structure, function and evolution. *Biochim. Biophys. Acta Gene Regulatory Mech.* 1819, 104–119. doi: 10.1016/j.bbagr.2011.10.002
- Selmar, D., and Kleinwächter, M. (2013). Stress enhances the synthesis of secondary plant products: the impact of stress-related over-reduction on the accumulation of natural products. *Plant Cell Physiol.* 54, 817–826. doi: 10.1093/PCP/PCT054
- Sharma, S. S., and Dietz, K. J. (2009). The relationship between metal toxicity and cellular redox imbalance. *Trends Plant Sci.* 14, 43–50. doi: 10.1016/j.TPLANTS.2008.10.007
- 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, 3210–3212. doi: 10.1093/bioinformatics/btv351
- Sofidiya, M. O., Odukoya, O. A., Familoni, O. B., and Inya-Agha, S. I. (2006). Free radical scavenging activity of some nigerian medicinal plant extracts. *Pakistan J. Biol. Sci.* 9, 1438–1441. doi: 10.3923/PJBS.2006.1438.1441
- 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. doi: 10.1093/nar/gkh379
- Tang, M., Xu, L., Wang, Y., Cheng, W., Luo, X., Xie, Y., et al. (2019). Genome-wide characterization and evolutionary analysis of heat shock transcription factors (HSFs) to reveal their potential role under abiotic stresses in radish (*Raphanus sativus* L.). *BMC Genomics* 20:772. doi: 10.1186/s12864-019-6121-3
- Tempel, S. (2012). Using and understanding repeatMasker. *Methods Mol. Biol.* 859, 29–51. doi: 10.1007/978-1-61779-603-6\_2
- Tian, Y., Zeng, Y., Zhang, J., Yang, C. G., Yan, L., Wang, X. J., et al. (2015). High quality reference genome of drumstick tree (*Moringa Oleifera* Lam.), a potential perennial crop. *Sci. China Life Sci.* 58, 627–638. doi: 10.1007/s11427-015-4872-x
- Wang, P., Song, H., Li, C., Li, P., Li, A., Guan, H., et al. (2017). Genome-wide dissection of the heat shock transcription factor family genes in arachis. *Front. Plant Sci.* 8:106. doi: 10.3389/fpls.2017.00106
- Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MScanX: a toolkit for detection and evolutionary analysis of gene synteny and Collinearity. *Nucleic Acids Res.* 40, 1–14. doi: 10.1093/nar/gkr1293
- Wani, S. H., Anand, S., Singh, B., Bohra, A., and Joshi, R. (2021). WRKY transcription factors and plant defense responses: latest discoveries and future prospects. *Plant Cell Rep.* 40, 1071–1085. doi: 10.1007/s00299-021-02691-8
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Sequence analysis jalview version 2-a multiple sequence alignment editor and analysis workbench. *Bioinform. Applic. Note* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033
- Yoon, Y., Seo, D. H., Shin, H., Kim, H. J., Kim, C. M., and Jang, G. (2020). The role of stress-responsive transcription factors in modulating abiotic stress tolerance in plants. *Agronomy* 10:788. doi: 10.3390/agronomy10060788
- Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Shyamli, Pradhan, Panda and Parida. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.