SCIENCE CHINA
Life Sciences

• COVER ARTICLE •

July 2015 Vol.58 No.7: 627–638
doi: 10.1007/s11427-015-4872-x

High quality reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop

TIAN Yang[1,10,13†], ZENG Yan[4†], ZHANG Jing[8†], YANG ChengGuang[9], YAN Liang[1,5], WANG XuanJun[13], SHI ChongYing[2], XIE Jing[3], DAI TianYi[2], PENG Lei[2], ZENG HUAN Yu[1], XU AnNi[1], HUANG YeWei[13], ZHANG JiaJin[11,12], MA Xiao[13], DONG Yang[7,10], HAO ShuMei[6*] & SHENG Jun[13*]

[1]*College of Life Sciences, Jilin University, Changchun 130012, China;*
[2]*College of Food Sciences, Yunnan Agricultural University, Kunming 650201, China;*
[3]*College of Animal Sciences, Yunnan Agricultural University, Kunming 650201, China;*
[4]*College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;*
[5]*Pu'er Institute of Pu-er Tea, Pu'er 665000, China;*
[6]*School of Agriculture, Yunnan University, Kunming 650091, China;*
[7]*Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming 650093, China;*
[8]*College of Life Sciences, Huazhong University of Science and Technology, Wuhan 430074, China;*
[9]*College of Life Sciences, Wuhan University, Wuhan 430072, China;*
[10]*Yunnan Institute of Lamu, Kunming 650034, China;*
[11]*School of Science and Information Engineering, Yunnan Agricultural University, Kunming 650201, China;*
[12]*State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Kunming 650223, China;*
[13]*Key Laboratory of Pu-er Tea Science, Ministry of Education and Yunnan Agricultural University, Kunming 650201, China*

Received January 7, 2015; accepted March 10, 2015; published online May 29, 2015

The drumstick tree (*Moringa oleifera* Lam.) is a perennial crop that has gained popularity in certain developing countries for its high-nutrition content and adaptability to arid and semi-arid environments. Here we report a high-quality draft genome sequence of *M. oleifera*. This assembly represents 91.78% of the estimated genome size and contains 19,465 protein-coding genes. Comparative genomic analysis between *M. oleifera* and related woody plant genomes helps clarify the general evolution of this species, while the identification of several species-specific gene families and positively selected genes in *M. oleifera* may help identify genes related to *M. oleifera*'s high protein content, fast-growth, heat and stress tolerance. This reference genome greatly extends the basic research on *M. oleifera*, and may further promote applying genomics to enhanced breeding and improvement of *M. oleifera*.

genome, drumstick tree, *Moringa oleifera*

Citation: Tian Y, Zeng Y, Zhang J, Yang CG, Yan L, Wang XJ, Shi CY, Xie J, Dai TY, Peng L, Zeng HY, Xu AN, Huang YW, Zhang JJ, Ma X, Dong Y, Hao SM, Sheng J. High quality reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop. Sci China Life Sci, 2015, 58: 627−638, doi: 10.1007/s11427-015-4872-x

Despite doubling of per-acre yields of major grain crops since 1950, nearly one in seven people suffer from malnutrition worldwide, predominately in developing countries. While there are myriad reasons for widespread lack of foodstuffs, part of the problem is that the human food system is dominated by annual crops that are sown each year, such as corn, wheat, rice, and most of leaf vegetables. These

†Contributed equally to this work
*Corresponding author (email: shengj@ynau.edu.cn; haosm@sina.com)

crops can be quite labor and resource intensive, and are not all suited to certain ecosystems found in many developing countries. Conversely, certain perennials—plants that are sown once and live for years—that are highly water-use and nutrient cycling efficient, adaptable to a wide array of environments, have high-nutrition value, may be viable alternatives to traditional annuals. Unfortunately, to date few of these perennial plants are widely planted or consumed, with the banana, coco and pigeon pea being notable exceptions.

Recently, a small to medium-sized, evergreen or deciduous tree native to northern India, Pakistan and Nepal known as the drumstick tree (*Moringa oleifera* Lam.)—or alternatively as the horseradish tree or ben oil tree—has received increased agricultural and industrial attention. Not only can every part of the tree be used as food, medicines or for industrial purposes [1–4], but its high protein, vitamin and mineral content have made it an attractive target for wide-spread planting in some developing countries [1,3,5,6]. Moreover, *M. oleifera* grows well at altitudes from 0 to 1,800 m and in areas with rainfall between 500 and 1,500 mm per year, making it suitable for both semi-arid and arid ecosystem, which covers 37.0% of the earth's geographical area, and even larger swaths of the developing world. Despite these benefits and efforts to cultivate the tree, little basic research on *M. oleifera* has been conducted, which greatly limits its further traditional and novel applications. Here, for the first time we have sequenced the genome of *M. oleifera* and provided well-assembled and annotated genome that should prove invaluable in furthering the uses and investigations of this important perennial.

1 Materials and methods

1.1 DNA materials

DNA used for sequencing was extracted from the leaves of a one-year-old drumstick tree (*Moringa oleifera* Lam.) planted in Pu'er, Yunnan province, China. In total, over 50 μg DNA was used to construct the sequencing libraries.

1.2 Sequencing data production and processing

Whole genome shotgun sequencing was deployed to produce reads on an Illumina Hiseq2500TM. Raw sequencing data of Illumina Hiseq2500TM were obtained through three steps, image analysis, base calling and sequence analysis, yielded a total of 202 Gb raw data.

Reads with more than 10 low-quality bases (low-quality being defined as value less than 60 bases or with "N"s were filtered). Duplicated reads and reads with adaptor are also removed. Both ends of each read were trimmed by 2 bp. Totally, we obtained seven libraries of different lengths: 177, 222, 390, 503, 3,500, 11,500, and 15,000 bp. Adaptor ligation and DNA cluster preparation were performed prior to sequencing. *K*-mer frequency and correction to reduce low frequency reads (primarily caused by sequencing errors) was done in SOAPec 2.01 [7], leaving clean data suitable for *de novo* assembly.

1.3 Genome assembly

To deal with genomes with the potential high heterozygosity of the genomes, Platanus 1.2.1 [8] was used to assemble DNA fragments (reads) into contigs. The initial *K*-mer size was set 41, step size was 10, maximum difference for branch cutting was 0.3, maximum difference for bubble crush was 0.15, and *K*-mer coverage cutoff was 5. No parameter needs to be set during scaffolding using SSPACE v2.0 [9]. The final assembly was generated after gap filling with Gapcloser v1.12 in SOAPdenovo package [7]. The assembly was evaluated by mapping the reads back to the genome using SOAPaligner 2.18.

1.4 Repetitive sequence annotation

Tandem Repeats Finder (TRF) 4.04 [10] was used to identify tandem repeats in the *M. oleifera* genome, and then Repeatmasker 3.3.0 and RepeatProteinMask were used to search repeats against Repbase [11] at the DNA and protein levels, respectively. These results were combined with the *de novo* prediction via LTR_FINDER 1.05 [12] and RepeatScout [13].

1.5 Protein-coding gene annotation

A combination of homology-based and *ab initio* methods yielded 19,465 annotated protein-coding genes. Protein sequences of six plant species (*Arabidopsis thaliana* [14], *Glycine max* [15], *Oryza sativa* [16], *Populus trichocarpa* [17], *Sorghum bicolor* [18], *Selaginella moellendorffii* [19]) were used in the homology-based method. These six species were selected because they all have well assembled and annotated genomes sequences, and include angiosperm to gymnosperm species that are related to *M. oleifera,* making them excellent reference points to explore the evolutionary processes related to *M. oleifera*. In the homology-based method, we first performed tblastn setting *e*-value cutoff 10^{-5}. Blast hits with *e*-value lower than 10^{-5} in the genome were discarded, and then predicted regions were extended by 2,000 bp both upstream and downstream, and aligned against protein sequence using GeneWise [20] to identify gene structure. With the AUGUSTUS 2.5.5 [21], Genscan, and GlimmerHMM 3.0.1 [22] software packages used for gene prediction. In the *ab initio* method, the genes predicted by software were aligned to *Arabidopsis thaliana* protein sequences, with alignment rate set at 0.5. The two sets of genes were then merged using GLEAN, a software that can create consensus gene sets by integrating disparate sources of gene structure evidence.

1.6 Gene function annotation

Potential functions of the annotated genes were assigned by choosing the best alignment of genes against the TrEMBL [23], KEGG [24] and InterProscan [25] databases.

1.7 Non-coding gene annotation

tRNAscan-SE v1.23 [26] was used for *M. oleifera* tRNA annotation. We used homology method to identify rRNA. rRNA sequence data was downloaded from the Rfam [27] database to serve as a reference. INFERNAL v0.81 [28] was used to identify snRNA and miRNA.

1.8 miRNA target analysis

Mature miRNA sequences were downloaded from miRbase [29] and aligned to annotated miRNA genes via blastn. Hits longer than 16 bp were selected as potential mature miRNA sequences. Then we predicted target genes of these mature miRNA sequences using online tool psRNATarget [30].

1.9 Gene families

Four other species including *Vitis vinifera*, *Cajanus cajan*, *Carica papaya*, *Malus pumila* and software OrthoMCL 1.4 [31] were used to identify gene clusters. First, we conducted pairwise alignment using blastp with *e*-value cutoff of 10^{-5}. Then OrthoMCL was used with all parameters default.

1.10 Phylogenetic relationship and divergent time

Single copy gene family genes of the five woody plants (*Moringa oleifera*, *Vitis vinifera*, *Cajanus cajan*, *Carica papaya*, *Malus pumila)* obtained from gene family analysis were used for phylogenetic analysis. Multiple sequence alignments were performed using MUSCLE 3.8.31 [32]. Four-fold degenerate sites were extracted from each gene and concatenated into one linear sequence for each species, in order to construct a neighbor joining tree using PhyML 3.0. To estimate the divergence time of each species, we used known divergence time information between plant species from the public resource, TIMETREE (http://www.timetree.org/). Using data generated from the phylogenic tree, we estimated divergence times with MCMCTREE in paml 4.4 [33].

1.11 Gene family contraction and expanding

CAFE 2.1 [34] was used to screening gene family expansion and contraction history.

1.12 Positive selection analysis

Blast was performed to align the coding sequence data of *M. oleifera* and *Carica papaya* in order to find the gene pairs with the best alignments. The resulting 5,601 orthologous gene pairs were aligned again using lastz as a preparation for KaKs_ Calculator 1.2 [35], which finally yielded a dataset of each gene pair's *Ka*/*Ks* ratio. Alignment in Figure 3 and Figure S10 was produced by multiple alignment tool MUSCLE [32] and picture was generated by ClustalX [36].

2 Results

2.1 Genome assembly of *M. oleifera*

We obtained 457× coverage DNA sequencing data for the *M. oleifera* sample (summary of sequencing data used for the assembly is presented in Table S1, and 17-mer frequency distribution is shown in Figure S1). Based on the 17-mer frequency distribution, the estimated genome size was estimated at 315 Mb (Table S2), and further flow cytometry indicated that the nuclear genome size (*c*-value) of *M. oleifera* was comparable and/or smaller than that of *Oryza sativa*. The final contig and scaffold N50 were 123 kb and 1.14 Mb, respectively (Table 1), with over 80% (231 Mb) of the total sequence represented in 262 scaffolds. The final quality of genome assembly was comparable to recently published high-quality reference plant genomes [19]. In total, 95.67% reads could be re-mapped to the assembly, further confirmed the quality of our genome assembly (Table S3).

The genome size of woody plant ranges from 280 Mb, such as *Prunus mume* [37], to 221.8 Gb for *Pinus taeda* [38].

Table 1 Summary statistics of *M. oleifera* genome assembly

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	4,165	4,362	5,792	1,382
N80	30,989	1,914	150,929	262
N70	60,562	1,261	396,940	147
N60	91,660	880	736,902	93
N50	123,008	611	1,140,476	61
Longest	1,070,888		6,788,971	
Average size	6,911		8,677	
Total number (>1,000bp)		13512		10,494
Total	287,419,725	41,586		33,332

Here, the genomes of *M. oleifera* proved to be among the smallest, being even smaller than rice. Paired with *M. oleifera's* fast-growth, high seed production, and adaptation to arid and semi-arid environments, the small size of the *M. oleifera* genome makes it not just an attractive perennial, but a viable model for functional genomic studies aimed at better characterizing the woody plant biology.

2.2 Annotation of *M. oleifera* Genome

A combination of homology and *ab initio* methods allowed us to annotate 19,465 high-confidence protein-coding loci in the *M. oleifera* genome with a mean coding length of 3,354.22 bp and an average of 5.42 exons per gene (Table S4). Further gene structure-based evaluation to confirm the annotation of protein-coding genes (distribution of mRNA structure statistics are in Figures S2 and S3) showed that 93.74% of *M. oleifera* genes have homologs in the TrEMBL protein database, and 72.67% could be classified by Swiss-Prot [23]. In total, 94.01% of the genes have either known homologs or can be functionally classified with InterPro, GO, KEGG, Swiss-Prot or TrEMBL databases [39] (Table S5).

Structure- and homology-based analyses identified 148,820,058 bp repetitive elements, covering most types of plant transposable elements. Most of the repeats were *de novo* predicted. Curiously, only 10.1% of the repeats detected by homologous method, perhaps reflecting phylogenetic distance of *M. oleifera* from other plants with published genomes. Together with numerous truncated repetitive elements, these elements make up 51.45% of the *M. oleifera* genome (Table S6), while 136 Mb of the repeats were transposable elements (TE) that make up 47.10% of the *M. oleifera* genome (Table S7; distribution of TE divergence rate is shown in Figures S4 and S5). An overview of annotated non-coding RNA (ncRNA) genes is shown in Table S8. In total, we predicted 87 mature miRNAs and 369 potential target genes of these miRNAs (Table S9). GO (gene ontology) [40] enrichment analysis of these genes using Ontologizer [41] (Figure S6) showed that 25 of 26 enriched terms were concentrated in cellular biological process regulation.

Previous studies found that intracellular tRNA level may be correlated with tRNA gene copy number [42]. Here, 1,777 tRNA genes reside in *M. oleifera* genome, but only 388 in *Carica papaya* and 600 in *Vitis vinifera*, which may related to *M. oleifera's* markedly high protein synthesis ability.

2.3 Phylogenetic and whole genome duplication analysis

M. oleifera was originally classified into *Rhoeadales* in *Flora of China* based on its morphology, but mounting molecular evidences suggest it belongs to *Brassicales* [43]. Here, four woody plants with published genomes from the *Dicotyledons* clade—*Vitis vinifera* [44], *Cajanus cajan* [45], *Carica papaya* [46], and *Malus domestica* [47]—were used to construct a phylogenetic tree of *M. oleifera* (Figure 1A shows the divergence time of each branch). The tree showed that *Carica papaya* is the most closely related species to *M. oleifera*, suggesting placement of this species in *Brassicales*. And we analyzed the phylogeny of four *Brassicales* (*Arabidopsis thaliana*, *Brassica rapa*, *carica papaya*, and *Moringa oleifera*; shown in Figure S7). Further whole genome duplication analysis of these four *Brassicales* indicated that whole genome duplication (WGD) events took place several times in *Brassicales* (Figure 1C). Such WGD events help clarify some of the history of *Brassicales;* for example, *Carica papaya* was previously known to have not experienced the At-β WGD [46], and our data suggests that neither *M. oleifera* nor *Carica papaya* have experienced any recent WGD events (Figure 1C). Instead, our analyses indicate that the last WGD events of these two species took place before they diverged from *A. thaliana*, that is the At-γ WGD; a finding further supported by calculating the *Ks* between the paralogous genes of *M. oleifera* (Figure 1D) which showed only one obvious peak where $Ks \approx 1.8$.

2.4 *M. oleifera*-specific gene families and genes

Gene family is often an assemblage of genes with approximately the same function. Species-specific gene families add raw materials to the generation of discrepancy against other species [48,49]. Here, we carried out gene family clustering analysis on all protein-coding genes of *M. oleifera*. Comparative analysis of *M. oleifera* with *Vitis vinifera*, *Cajanus cajan*, *Carica papaya* and *Malus domestica* showed that these five different plant species possess similar numbers of gene families, with a core set of 10,215 shared genes (Figure 1b). Compared to other species, however, *M. oleifera* has markedly fewer single-copy families and unclustered genes (distribution of gene clusters in *M. oleifera* genome was shown in Table S10 and Figure S8). Of 12,298 gene families in *M. oleifera*, 198 gene families were *M. oleifera*-specific, including a total of 812 genes (GO enrichment analysis of these genes is in Figure S9). Calculating gene family contraction and expansion on each branch in the phylogenetic tree showed that that *M. oleifera* has 560 expanded gene families—the smallest of the five wood plant species with published genomes—and 2,611 contracted gene families (Figure 1A), making it comparatively smaller with a compact genome.

Curiously, four *SKP1* genes and 18 F-box domain containing genes were identified as members of *M. oleifera*-specific families. Gene *SKP1* is a protein crucial to the cell cycle controlling [50] that helps coordinated the ubiquitination and degradation of phase specific proteins to maintain the cell cycle, with the F-box motif maintaining the association between these proteins [51]. Rather confusingly,

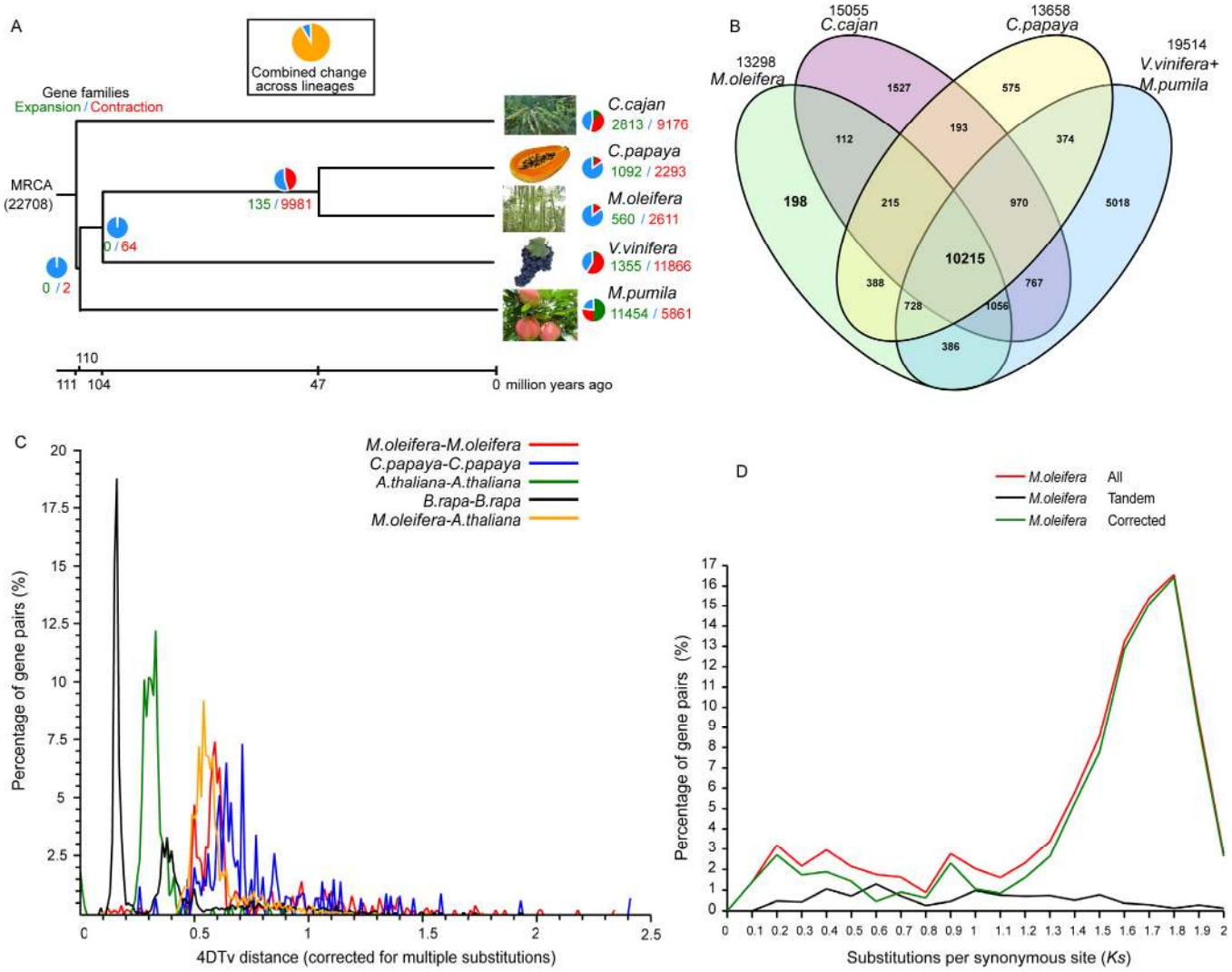

Figure 1 Comparative genomic analysis. A, Phylogenetic tree of five woody plants, *M. oleifera*, *Vitis vinifera*, *Cajanus cajan*, *Carica papaya* and *Malus domestica.* Estimated divergence time ranges from 47 million years ago (mya) to 110 mya. Numbers of gene family contractions and expansions are shown in the pie chart. B, Gene family analysis of *M. oleifera* against *Vitis vinifera*, *Cajanus cajan*, *Carica papaya* and *Malus domestica*. *Vitis vinifera* and *Malus domestica* were merged into one dataset. C, 4DTv (fourfold degenerate third-codon transversion) analysis of M. oleifera using *Arabidopsis thaliana*, *Brassica rapa* and *Carica papaya*. D，Distribution of *Ks* between paralogou gene pairs.

four *SKP1* and 18 F-box containing genes *M. oleifera* were in the *M. oleifera*-specific gene families while another seven *SKP1* genes and 104 F-box containing genes were not. In theory, there may be two potential explanations: First, these represented genes may be newly derived and may play a role in *M. oleifera*'s fast-growth and heat tolerance, while second these genes may simply be redundant and accumulated many mutations. We also found three *BET V 1* genes in the *M. oleifera*-specific gene families. *BET V 1* was first found in birch tree pollen as an allergen [52], but more functions of this gene have been discovered later, including its role as a steroid carrier [53]. *BET V 1* genes are potential factors for the *M. oleifera*'s fast growth as this gene family is related to the binding of many ligands, including ABA, lipids and steroids. These *M. oleifera*-specific genes may be functionally important to *M. oleifera* and warrant further investigation.

2.5 Positively selected genes in *M. oleifera* genome

Positively selected genes often have functions that favor the organism's adaptation and establishment in an area. To investigate which genes may be associated with certain traits that have made *M. oleifera*'s successful, we conducted positive selection analysis. Blast and KaKs_Calculator [54] compared orthologs between *M. oleifera* and each one of *Carica papaya*, *Vitis vinifera* and *Malus domstica,* and respectively found 566, 399, 112 genes of *M. oleifera* with *Ka*/*Ks* ratio>1 (significance, $P<0.05$; see Table S11, Table S12, Table S13). We further found four genes that overlap among the three gene sets (Figure 2). We also found two genes (annotated gene: *lamu_GLEAN_10016878*, *lamu_*

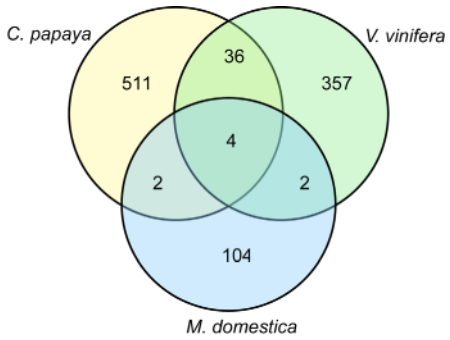

Figure 2 Positively selected genes identified from pairwise comparisons between *M. oleifera* and *Carica papaya*, *Vitis vinifera* and *Malus domestica*.

GLEAN_10011614) with alignment length longer than half the gene size, indicating they may be specifically and strongly selected in *M. oleifera*. Gene *lamu_GLEAN_10016878* is functionally annotated as a Myb/SANT-like DNA binding domains. The SANT domain is ubiquitous in chromatin regulatory proteins and it is often involved in histone acetylation, deacetylation and ATP-dependent chromtin remodeling process [55]. More importantly, many proteins containing the Myb/SANT domain have DNA binding activity and are related to gene regulation; these regulators of different functions often do not resemble each other. However, both ends of this gene and some segments of the central region is highly conserved. However some parts of central regions vary (see screening of the alignment of orthologous genes in Figure S10) [56,57].

Gene *lamu_GLEAN_10011614* is supposed to be a ribosomal protein S6e, which is highly conserved in vertebrates, invertebrates, and fungi [58]. In eukaryotes, ribosomal proteins synthesized in cytoplasm are imported into nucleoplasm before associating with newly transcribed pre-rRNA to form a 90S complex, which is processed into a 60S and a 40S ribosomal subunit and subsequently exported into cytoplasm [59]. Ribosomal proteins assist the maturation and functioning of pre-18S RNA and ribosome [60]. The S6e amino acid sequence has twonucleolar binding sequence (Nobis) and severalnuclear localization signal (NLS) sequence according to Kundu-Michalik's study [58]. We identified Nobis1's N-terminus by recognizing (G)RVRL pattern and inferred the C-terminus according to the length of Nobis1 revealed in previous study [58]. Based on the Kundu-Michalik's study, we also proposed a rough border of the Nobis2 frame. And some other elements such as NLS and phosphorylation sites on this sequence were tentatively presented in Figure 3. Phosphorylation states of S6 often act as a switch and regulate cell processes through phosphorylation cascade [61]. Positive selection of the *lamu_GLEAN_10011614* gene elegantly serves as further molecular evidence that the protein synthesis machinery of *M. oleifera* has likely experienced strong evolutionary rewiring to produce more proteins.

2.6 Analysis of transcription factor families

Transcription factors regulate gene expression, making them crucial and diverse in organisms ranging from microbes to high plants and animals [62,63]. Analysis of previously identified transcription factors has yielded a large amount of information on gene expression patterns. Taking the *Arabidopsis thaliana* transcription factor families in the TAIR database (http://arabidopsis.org/browse/genefamily/index.jsp) [64] as reference, we identified a total of 939 transcription factors (Table S14) in the *M. oleifera* genome using the blastp with *P*-value<10^{-20}. Interestingly, our positive selection analysis of *M. oleifera* against each of *Vitis vinifera*, *Carica papaya*, and *Malus domestica* uncovered 43 transcription factors were under positive selection in various classes including ABI3VP1, AP2-EREBP, Alfin-like, C2C2-Dof, C2C2-Gata, C2H2, C3H, CPP, E2F-DP,

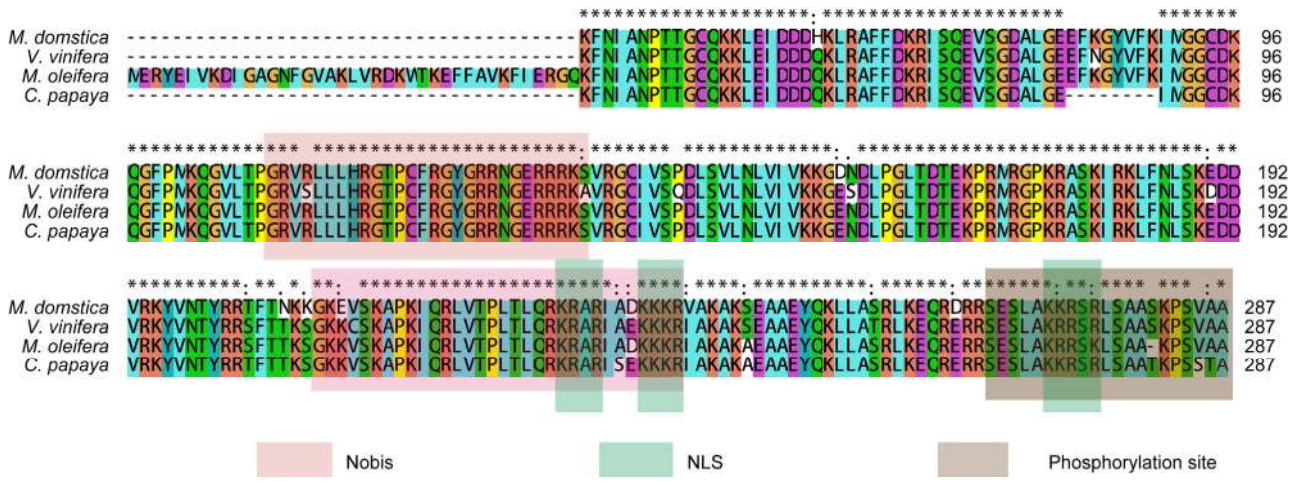

Figure 3 Alignment of *M. oleifera* gene *lamu_GLEAN_10011614* and its orthologous in *Vitis vinifera*, *Cajanus cajan*, *Carica papaya*, *Malus domestica*. Postulated elements like Nobis, NLS, and phosphorylation sites have been marked in colored rectangles.

G2-like, GRAS, Homeobox, MADS, MYB, NAC, PHD, Trihelix, WRKY, bHLH transcription factors. Among these, the WRKY transcription factors are particularly interesting, since they were previously suggested to play important roles in response to various abiotic stress, including cold, heat, water deficiency, excessive salt, nutrient starvation, and variable light condition. Here, we found five copies under positive selection. Similarly, the C2H2 transcription factors—a superfamily that plays important roles in defense responses and various other physiological processes in plants—have four copies under positive selection. Meanwhile, the AP2-EREBP transcription factors that were previously implicated in hormone, sugar and redox signaling in context of abiotic stresses such as cold and drought, had two copies under positive selection, while the C3H transcription factors with some copies being reported to response to drought stress have two copies under positive selection (Table S15). The fact that all of these transcription factors involved in stress response were found to be under positive selection may account for *M. oleifera*'s adaptation to both heat and drought stress present in arid environments.

2.7 HSP genes

Heat stress is a serious threat to crop production, which may be exacerbated by changes in global climates. Accordingly, the high temperature tolerance of *M. oleifera* [1] may prove quite useful. In many species of plants, Heat shock proteins (HSPs) or stress-induced proteins participate in many primary stress response such as drought, salinity, cold and hot temperatures and chemicals [65–67]. Using the heat shock proteins sequences of *Arabidopsis thaliana* we downloaded from HSPIR (http://pdslab.biochem.iisc.ernet.in/hspir/chaperone.php) [68] as reference, we identified a total of 133 heat shock proteins. Based on their nature of functions and molecular mass, HSPs are classified broadly into six major families, namely Hsp70 (25 copies in *M. oleifera* genome), Hsp40 (J-proteins, 52 copies in *M. oleifera* genome), Hsp60 (chaperonins, 17 copies in *M. oleifera* genome), Hsp90 (three copies in *M. oleifera* genome), Hsp100 (Clp proteins, nine copies in *M. oleifera* genome) and small heat shock proteins (27 sHsps copies in *M. oleifera* genome) (Table S16).

We further checked the HSP genes' *Ka*/*Ks* ratio between *M. oleifera* and *Carica papaya*, and found that the average *Ka*/*Ks* ratio of HSP genes was higher than that of the background (Table S17). HSP genes that have positive selection features against any of *Carica papaya*, *Vitis vinifera* and *Malus domstica* were collected and shown in Table S18. These genes may potentially be related to the heat tolerance that is one characteristic of *M. oleifera*.

2.8 Brassinosteroid signal transduction pathway

Brassinosteroid is a kind of plant hormone with a regulatory function in cell elongation and cell division, which can significantly promote plant growth. Previous reports suggest that that brassinosteroids help plants get through environmental stresses such as cold, drought and heat. Here, we analyzed the brassinosteroid signal transduction pathway in *M. oleifera* and found that the *BAK1* (BRI1 associated receptor kinase 1) gene expanded in *M. oleifera* with 29 copies, as compared to five copies in *A. thaliana* genome (Figure S11). Furthermore, we also noticed one copy of the *BAK1* gene was also under positive selection when compared against *Vitis vinifera*. *BAK1* plays a major role in transducing the BR signal, and loss-of-function mutation of *BAK1* caused a weak dwarf phenotype [69].

2.9 γ-aminobutyrate (GABA) bio-synthesis and sitosterol bio-synthesis pathways in *M. oleifera*

We analyzed GABA bio-synthesis and sitosterol biosynthesis pathways in *M. oleifera*—both of which are important hormone pathways in plants—and annotated all genes in the pathways. 4-Aminobutyrate or GABA is a ubiquitous, four carbon, non-protein amino acid found in higher plants, animals, fungi and bacteria. In plants, the concentration of GABA is markedly stimulated by a variety of stress conditions, *e.g.* hypoxia, temperature shock, mechanical manipulation and damage, water stress and phytohormones [70,71]. Here, we found that GABA is synthesized almost exclusively by the irreversible α-decarboxylation of L-glutamate by glutamate decarboxylase (GAD; annotated gene: *lamu_GLEAN_10006873*, *lamu_GLEAN_10006874, lamu_GLEAN_10004957, lamu_GLEAN_10007711, lamu_GLEAN_10007712, lamu_GLEAN_10007713*) [72,73]. Subsequently, GABA is catabolized by GABA transaminase (GABA-T; annotated gene: *lamu_GLEAN_10002543*) and succinate semialdehyde dehydrogenase (SSADH; annotated gene: *lamu_GLEAN_10008793*, *lamu_GLEAN_10008794*) to succinate, an important Kerbs cycle metabolite [73]. The only other enzyme of glutamate metabolism known to be stimulated by Ca^{2+} in plants is glutamate dehydrogenase (GDH; annotated gene: *lamu_GLEAN_10005665*), a mitochondrial enzyme (Figure S12).

To understand the sterol biosynthesis genes in *M. oleifera,* we tried to draw the major sterol biosynthetic pathway operating in most higher plants [74]. Sitosterol is a typical plant membrane reinforcement, at the expanse of campesterol [75]. Campesterol can be used to produce brassinosteroids, which were reported to have observable growth-promoting effects in many plants. The gene *STM2*, of which two copies were found in *M. oleifera*, plays a critical role in balancing the ratio of campesterol to sitosterol to satisfy both growth requirements and membrane integrity (Figure S13).

3 Discussion

At the time of this study, no genomes of species in the family *Moringaceae* were available, making the present *M. oleifera* genome data a valuable reference for further studies on both *M. oleifera* and other species in this important plant family. Due to a dearth of related research, this present study is far from conclusive on many fronts, and is instead suggestive of many further lines of inquiry into the unique characteristics of *M. oleifera* that remain to be explored. In particular, the gene cluster analysis reveals that *M. oleifera* possesses a remarkably small amount of single copy genes, and small amount of *M. oleifera* specific gene families. Taken alongside the fact that the annotated *M. oleifera* genes were fewer than any other resolved higher plants indicate that *M. oleifera* has a compact genome, which may, in part, be responsible or underlie its comparatively fast growth and rapid cell proliferation.

In the present study, we concentrated on the indistinct relationship between the genome content characters and the phenotypic traits, identified a number of genes or gene families that might account for the high protein content, heat tolerance, drought resistance, and fast growth of *M. oleifera*. The gene list provided by our analyisis is important not only for the future functional studies of *M. oleifera*, but also for future efforts in breeding and improvement of *M. oleifera*, both of which may help promote *M. oleifera* as a viable perennial crop in regions of the world where food shortages are endemic or the local environment cannot support more traditional annual crops.

1 Olson ME, Fahey JW. *Moringa oleifera*: a multipurpose tree for the dry tropics. Revista Mexicana De Biodiversidad, 2011, 82: 1071–1082

2 Horwath M, Benin V. Theoretical investigation of a reported antibiotic from the "Miracle Tree" *Moringa oleifera*. Computational and Theoretical Chemistry, 2011, 965: 196–201

3 Makkar HPS, Becker K. Nutrients and antiquality factors in different morphological parts of the *Moringa oleifera* tree. J Agr Sci, 1997, 128: 311–322

4 Palada MC. Moringa (*Moringa oleifera* Lam.): A versatile tree crop with horticultural potential in the subtropical United States. Hortscience, 1996, 31: 794–797

5 Oliveira JTA, Silveira SB, Vasconcelos IM, Cavada BS, Moreira RA. Compositional and nutritional attributes of seeds from the multiple purpose tree *Moringa oleifera* Lamarck. J Sci Food Agr, 1999, 79: 815–820

6 Amaglo NK, Bennett RN, Lo Curto RB, Rosa EAS, Lo Turco V, Giuffrida A, Lo Curto A, Crea F, Timpo GM. Profiling selected phytochemicals and nutrients in different tissues of the multipurpose tree *Moringa oleifera* L., grown in Ghana. Food Chem, 2010, 122: 1047–1054

7 Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 2012, 1: 18

8 Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res, 2014, 24: 1384–1395

9 Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 2011, 27: 578–579

10 Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 1999, 27: 573–580

11 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res, 2005, 110: 462–467

12 Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res, 2007, 35: W265–268

13 Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics, 2005, 21 Suppl 1: i351–358

14 Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 2000, 408: 796–815

15 Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. Genome sequence of the palaeopolyploid soybean. Nature, 2010, 463: 178–183

16 Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science, 2002, 296: 92–100

17 Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, 2006, 313: 1596–1604

18 Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS. The

Sorghum bicolor genome and the diversification of grasses. Nature, 2009, 457: 551–556
19 Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, Ashton NW, Axtell MJ, Barker E, Barker MS, Bennetzen JL, Bonawitz ND, Chapple C, Cheng C, Correa LG, Dacre M, DeBarry J, Dreyer I, Elias M, Engstrom EM, Estelle M, Feng L, Finet C, Floyd SK, Frommer WB, Fujita T, Gramzow L, Gutensohn M, Harholt J, Hattori M, Heyl A, Hirai T, Hiwatashi Y, Ishikawa M, Iwata M, Karol KG, Koehler B, Kolukisaoglu U, Kubo M, Kurata T, Lalonde S, Li K, Li Y, Litt A, Lyons E, Manning G, Maruyama T, Michael TP, Mikami K, Miyazaki S, Morinaga S, Murata T, Mueller-Roeber B, Nelson DR, Obara M, Oguri Y, Olmstead RG, Onodera N, Petersen BL, Pils B, Prigge M, Rensing SA, Riano-Pachon DM, Roberts AW, Sato Y, Scheller HV, Schulz B, Schulz C, Shakirov EV, Shibagaki N, Shinohara N, Shippen DE, Sorensen I, Sotooka R, Sugimoto N, Sugita M, Sumikawa N, Tanurdzic M, Theissen G, Ulvskov P, Wakazuki S, Weng JK, Willats WW, Wipf D, Wolf PG, Yang L, Zimmer AD, Zhu Q, Mitros T, Hellsten U, Loque D, Otillar R, Salamov A, Schmutz J, Shapiro H, Lindquist E, Lucas S, Rokhsar D, Grigoriev IV. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science, 2011, 332: 960–963
20 Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res, 2004, 14: 988–995
21 Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res, 2004, 32: W309–312
22 Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics, 2004, 20: 2878–2879
23 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003, 31: 365–370
24 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30
25 Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. Nucleic Acids Res, 2005, 33: W116–120
26 Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997, 25: 955–964
27 Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res, 2013, 41: D226–232
28 Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013, 29: 2933–2935
29 Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res, 2008, 36: D154–158
30 Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res, 2011, 39: W155–159
31 Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res, 2003, 13: 2178–2189
32 Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 2004, 32: 1792–1797
33 Yang ZH. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol, 2007, 24: 1586–1591
34 De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics, 2006, 22: 1269–1271
35 Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics, 2006, 4: 259–263
36 Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 1997, 25: 4876–4882
37 Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z, Fan G, Xing Z, Han C, Pan H, Zhong X, Shi W, Liang X, Du D, Sun F, Xu Z, Hao R, Lv T, Lv Y, Zheng Z, Sun M, Luo L, Cai M, Gao Y, Yin Y, Xu X, Cheng T. The genome of *Prunus mume*. Nat Commun, 2012, 3: 1318
38 Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, Neale DB. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics, 2010, 11: 420
39 Camon E, Barrell D, Brooksbank C, Magrane M, Apweiler R. The Gene Ontology Annotation (GOA) Project--Application of GO in SWISS-PROT, TrEMBL and InterPro. Comp Funct Genomics, 2003, 4: 71–74
40 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25: 25–29
41 Bauer S, Grossmann S, Vingron M, Robinson PN. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics, 2008, 24: 1650–1651
42 Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol, 1997, 268: 322–330
43 Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci U S A, 2010, 107: 18724–18728
44 Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 2007, 449: 463–467
45 Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang SP, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, Jackson SA. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol, 2012, 30: 83–89
46 Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Perez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature, 2008, 452: 991–996
47 Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A,

Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagne D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouze P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R. The genome of the domesticated apple (*Malus x domestica Borkh.*). Nat Genet, 2010, 42: 833–839

48 Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, Hetru C, Hoa NT, Hoffmann JA, Kanzok SM, Letunic I, Levashina EA, Loukeris TG, Lycett G, Meister S, Michel K, Moita LF, Muller HM, Osta MA, Paskewitz SM, Reichhart JM, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu J, Zheng L, Bork P, Kafatos FC. Immunity-related genes and gene families in *Anopheles gambiae*. Science, 2002, 298: 159–165

49 Shuai B, Reynaga-Pena CG, Springer PS. The lateral organ boundaries gene defines a novel, plant-specific gene family. Plant Physiol, 2002, 129: 747–761

50 Connelly C, Hieter P. Budding yeast SKP1 encodes an evolutionarily conserved kinetochore protein required for cell cycle progression. Cell, 1996, 86: 275–285

51 Bai C, Sen P, Hofmann K, Ma L, Goebl M, Harper JW, Elledge SJ. SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. Cell, 1996, 86: 263–274

52 Breiteneder H, Pettenburger K, Bito A, Valenta R, Kraft D, Rumpold H, Scheiner O, Breitenbach M. The Gene Coding for the Major Birch Pollen Allergen Betvl, Is Highly Homologous to a Pea Disease Resistance Response Gene. Embo J, 1989, 8: 1935–1938

53 Markovic-Housley Z, Degano M, Lamba D, von Roepenack-Lahaye E, Clemens S, Susani M, Ferreira F, Scheiner O, Breiteneder H. Crystal structure of a hypoallergenic isoform of the major birch pollen allergen Bet v 1 and its likely biological function as a plant steroid carrier. J Mol Biol, 2003, 325: 123–133

54 Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics, 2010, 8: 77–80

55 Boyer LA, Latek RR, Peterson CL. The SANT domain: a unique histone-tail-binding module? Nat Rev Mol Cell Biol, 2004, 5: 158–163

56 Barg R, Sobolev I, Eilon T, Gur A, Chmelnitsky I, Shabtai S, Grotewold E, Salts Y. The tomato early fruit specific gene Lefsm1 defines a novel class of plant-specific SANT/MYB domain proteins. Planta, 2005, 221: 197–211

57 Mohrmann L, Kal AJ, Verrijzer CP. Characterization of the extended Myb-like DNA-binding domain of trithorax group protein Zeste. J Biol Chem, 2002, 277: 47385–47392

58 Kundu-Michalik S, Bisotti MA, Lipsius E, Bauche A, Kruppa A, Klokow T, Kammler G, Kruppa J. Nucleolar binding sequences of the ribosomal protein S6e family reside in evolutionary highly conserved peptide clusters. Mol Biol Evol, 2008, 25: 580–590

59 Fromont-Racine M, Senger B, Saveanu C, Fasiolo F. Ribosome assembly in eukaryotes. Gene, 2003, 313: 17–42

60 Ferreira-Cerca S, Poll G, Gleizes PE, Tschochner H, Milkereit P. Roles of eukaryotic ribosomal proteins in maturation and transport of pre-18S rRNA and ribosome function. Mol Cell, 2005, 20: 263–275

61 Ruvinsky I, Meyuhas O. Ribosomal protein S6 phosphorylation: from protein synthesis to cell size. Trends Biochem Sci, 2006, 31: 342–348

62 Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res, 2003, 31: 374–378

63 Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science, 2000, 290: 2105–2110

64 Poole RL. The TAIR database. Methods Mol Biol, 2007, 406: 179–212

65 Morimoto RI. Cells in stress: transcriptional activation of heat shock genes. Science, 1993, 259: 1409–1410

66 Lindquist S, Craig EA. The heat-shock proteins. Annu Rev Genet, 1988, 22: 631–677

67 Lindquist S. The heat-shock response. Annu Rev Biochem, 1986, 55: 1151–1191

68 R RK, N SN, S PA, Sinha D, Veedin Rajan VB, Esthaki VK, D'Silva P. HSPIR: a manually annotated heat shock protein information resource. Bioinformatics, 2012, 28: 2853–2855

69 Nam KH, Li J. BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. Cell, 2002, 110: 203–212

70 Bown AW, Shelp BJ. The Metabolism and Functions of [gamma]-Aminobutyric Acid. Plant Physiol, 1997, 115: 1–5

71 Narayan VS, Nair PM. Metabolism, Enzymology and Possible Roles of 4-Aminobutyrate in Higher-Plants. Phytochemistry, 1990, 29: 367–375

72 Chung I, Bown AW, Shelp BJ. The production and efflux of 4-aminobutyrate in isolated mesophyll cells. Plant Physiol, 1992, 99: 659–664

73 Tuin LG, Shelp BJ. In-Situ [C-14] Glutamate Metabolism by Developing Soybean Cotyledons .1. Metabolic Routes. J Plant Physiol, 1994, 143: 1–7

74 Benveniste P. Biosynthesis and accumulation of sterols. Annu Rev Plant Biol, 2004, 55: 429–457

75 Schaeffer A, Bronner R, Benveniste P, Schaller H. The ratio of campesterol to sitosterol that modulates growth in Arabidopsis is controlled by STEROL METHYLTRANSFERASE 2;1. Plant J, 2001, 25: 605–615

Supporting Information

Figure S1 17-mer frequency distribution of reads data. A minor peak is at about half the depth of the main peak, indicating the heterozygosity of *M. oleifera*.

Figure S2 Distribution of exon numbers in annotated mRNA sequences. Annotation data of *Vitis vinifera*, *Populus trichocarpa*, *Selaginella moellendorffii* were used in parallel with *Moringa oleifera*'s annotation data.

Figure S3 Distribution of mRNA length, CDS length, exon length and intron length. Alongside the genome annotation of *Moringa oleifera*, we used *Vitis vinifera*, *Populus trichocarpa*, *Selaginella moellendorffii* genome annotation as references.

Figure S4 Distribution of Divergence Rate of each Type of *Moringa oleifera*'s TE. Divergence rate was calculated between the identified TE elements in the genome by homology-based method and the consensus sequence in the Repbase.

Figure S5 Distribution of Divergence Rate of each Type of *Moringa oleifera*'s TE. Divergence rate was calculated between the identified TE elements in the genome by de novo method and the consensus sequence in the predicted TE library.

Figure S6 GO analysis of the miRNA targeted gene. Note: *P*-value cutoff was set at 0.0005 to optimize the image.

Figure S7 Phylogenetic tree of four Brassicales (*Arabidopsis thaliana, Brassica rapa, Carica papaya,* and *Moringa oleifera*).

Figure S8 Orthologous gene distribution among five woody plants. *Vitis vinifera, Cajanus cajan, Carica papaya, Malus pumila* serve as the out group species to search the orthologous genes.

Figure S9 GO enrichment analysis of the *M. oleifera* specific gene family. These genes belong to the family that only exists in *M. oleifera* but not in *Vitis vinifera*, *Cajanus cajan*, *Carica papaya*, *Malus pumila*. *P*-value cutoff is set 0.01.

Figure S10 Alignment of *M. oleifera* gene *lamu_GLEAN_10011614* and its orthologous in *Vitis vinifera*, *Cajanus cajan*, *Carica papaya*, *Malus domestica*. The central region has more variations and N terminal is highly conserved.

Figure S11 Brassinosteroid signal transduction pathway in *M. oleifera* and *Arabidopsis thaliana*. The number near the left square brackets indicates the copies of this gene in the *Arabidopsis thaliana* genome, while the other indicates the copy number in the *M. oleifera* genome.

Figure S12 Simplified metabolic diagram the GABA shunt in relation to the Krebs cycle. GAD, glutamate decarboxylase; GABA-T, GABA transaminase; GDH, glutamate dehydrogenase; SSADH, succinic semialdehyde dehydrogenase.

Figure S13 Biosynthesis of (24ξ)-24-methyl cholesterol (campesterol) and (24R)-24-ethyl cholesterol (sitosterol) in Arabidopsis thaliana and *M. oleifera* genome. CPI, cyclopropyl sterol isomerase; SMT, sterol methyltransferase; OBT14DM, obtusifoliol-14-demethylase; SMO, sterol 4-methyl oxidase; DWF1, gene encoding the Δ5-sterol-Δ24-reductase. (isomerase); FACKEL, gene encoding the Δ8,14-sterol-Δ14-reductase; HYDRA1, gene encoding the Δ8-Δ7-sterol isomerase; DWF5, gene encoding the Δ5,7-sterol-Δ7-reductase; DWF7, Δ7-sterol-C5(6)-desaturase.

Table S1 Statistics of raw data

Table S2 Statistics of 17-mer analysis

Table S3 Statistics of mapping reads to the genome assembly

Table S4 Statistics of gene annotation

Table S5 Overview of gene function annotation.

Table S6 Statistics of Repeats in *M. oleifera* Genome

Table S7 TEs Content in the Assembled *Moringa oleifera* Genome

Table S8 Annotated ncRNA in the genome

Table S9 Predicted miRNA target genes

Table S10 Overview of gene family clustering among *M. oleifera* and *V. vinifera*, *C. cajan*, *C. papaya*, *M. pumila*

Table S11 Positively selected genes of *M. oleifera* against *Carica papaya*

Table S12 Positively selected genes of *M. oleifera* against *Vitis vinifera*

Table S13 Positively selected genes of *M. oleifera* against *Malus pumila*

Table S14 Identified transcription factors

Table S15 Positively selected transcription factor genes in *M. oleifera*

Table S16 Heat Shock Proteins (HSPs) in the *M. oleifera* genome

Table S17 HSP genes' *Ka/Ks* ratio and a comparison with the background

Table S18 Positively selected HSP genes in *M. oleifera*

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.