2018

Analysis of the transcription factors expressed in the mature seed embryos of Moringa oleifera Lam. using RNA-sequencing and de novo transcriptome assembly

Vivian A. Panes
Ateneo de Manila University

R.D Baoas
Ateneo de Manila University

Custom Citation

Panes, V.A. and Baoas, R.D. (2018). Analysis of the transcription factors expressed in the mature seed embryos of Moringa oleifera Lam. using RNA-sequencing and de novo transcriptome assembly. Acta Hortic. 1205, 705-716 DOI: 10.17660/ActaHortic.2018.1205.87 https://doi.org/10.17660/ActaHortic.2018.1205.87

Analysis of the transcription factors expressed in the mature seed embryos of *Moringa oleifera* Lam. using RNA-sequencing and de novo transcriptome assembly

V.A. Panes and R.D. Baoas

Department of Biology, School of Science and Engineering, Ateneo de Manila University, Loyola Heights, Quezon City, the Philippines.

Abstract

***Moringa oleifera* Lam. is well known for its numerous documented properties, particularly its significant applications in nutrition, therapeutics, biocontrol, energy, and bioremediation. These properties are the consequences of the vibrant physiological processes of the plant in the context of the ever-changing biotic and abiotic factors, in which transcription factors play substantial roles. Transcription factors (TFs) are the regulators of gene expression. Transcription factors enable the activation or repression of transcription. Along with the advent of ultrahigh-throughput sequencing technologies such as RNA sequencing (RNA-Seq), in combination with bioinformatics techniques, the investigation of the TFs of *M. oleifera* was made possible. This research aimed to identify transcripts encoding for transcription factors in the mature embryos of *Moringa oleifera* Lam. through RNA-sequencing and de novo transcriptome assembly (SOAP and Trinity assemblies); and determine their gene expression levels. In this study, the cataloguing and functional annotation of highly expressed TFs in *M. oleifera* were performed. Annotations were made based on BLAST, plant TF databases, TAIR, NCBI, gene2go, KEGG and ATTED-II. Highly expressed transcripts were homologs of *A. thaliana*. Other putative TFs were homologous to *Theobroma cacao*. Highly expressed putative TFs from SOAP as well as highly expressed TFs from TriAnn showed involvement in various seed processes. Some of the TFs were associated with non-seed related functions. It is recommended that validation of the functions of these putative *M. oleifera* transcription factors be performed through quantitative real-time PCR which can quantify the abundance and expression of TF genes in the mature seed embryos of *M. oleifera* in real time. Validation of genes encoding for TFs using quantitative realtime PCR which is an efficient method for the detection and quantitation of gene expression and can shed light on the functions of the transcription factors encoded by the TF transcripts particularly in their involvement in the many attributes of the seed embryos of *M. oleifera* such as in the developmental process, production of antioxidants, oil biosynthesis and stress response.**

Keywords: RNA-Seq, annotations, *Moringa oleifera*, gene expression

INTRODUCTION

Moringa oleifera Lam. thrives in tropical and sub-tropical regions. It originated in the Himalayan areas of India, Nepal and Pakistan. It is also native in Africa and western Asia including the Anatolian and Arabian peninsulas. Its distribution extends to the Philippines, Cambodia and the Americas (Anwar et al., 2007; Rajalakshmi et al., 2017). *M. oleifera* is dubbed as the "miracle tree" referring to its numerous practical uses (Abdull Razis et al., 2014; Leone et al., 2016). One of the reasons for the diversity of applications humanity is able to utilize *M. oleifera* is due to the varied cellular and molecular properties imparted to the plant by protein transcription factors (TFs). The dynamism of biological processes of *M. oleifera* is in part due to the TFs expressed during seed maturation. TFs are proteins that recognize specific sequences of DNA that are often called *cis*-regulatory sequences, because they must be on the same chromosome to the genes they control (Todeschini et al., 2014).

TFs bind to these *cis*-regulatory sequences, which are dispersed throughout the genomes, and this binding puts in motion a series of reactions that ultimately specify which genes are to be transcribed and at what rate (Todeschini et al., 2014). TFs enable the activation or repression of transcription by either allowing or inhibiting access of the RNA polymerase to the *cis*-regulatory element of a gene promoter (Liu et al., 2013). Approximately 10% of the protein-coding genes of most organisms are devoted to TFs, making them one of the largest classes of proteins in the cell. In most cases, a given TF recognizes its own *cis*-regulatory sequence, which is different from those recognized by all other regulators in the cell (Todeschini et al., 2014). Thus, TFs are important particularly in the context of responding to an ever-changing environment. This study focused on the analysis of the TFs in the mature embryos of *M. oleifera* Lam. The transcriptome consists of all species of RNAs in a cell or tissue of an organism at a specific developmental stage or physiological condition. The transcriptome includes both coding and non-coding RNAs (Liu et al., 2013). This study primarily aimed to identify the transcripts encoding for TFs, determine the expression levels of these transcripts and to categorize their corresponding putative functions based on annotations. This study is significant in that the identification of transcripts for the TFs and their expression levels in the mature embryos of *M. oleifera* will open the potential in manipulating these genes through induction or suppression of their expression, modifying pathways wherein these TFs are involved particularly in the important and interesting attributes of *M. oleifera* such as in lipid biosynthesis, antioxidant synthesis and stress response.

MATERIALS AND METHODS

Plant material for total RNA extraction

Mature seed embryos of *M. oleifera* were obtained from Muñoz, Nueva Ecija, Central Luzon, the Philippines. The samples were immediately frozen in liquid nitrogen and stored in -80°C ultralow freezer. Total RNAs was extracted from the mature seed embryos of *M. oleifera* Lam. using the Ambion mirVana total RNA isolation Kit (Life Technologies Inc., Carlsbad, CA., USA) following the manufaturer's protocol. Extracted RNA was qualified and quantified using a Nanodrop 1000 Spectrophotometer (Thermo Fisher, Waltham, MA, USA).

cDNA construction, RNA sequencing and de novo sequence assembly

A total of 20 µg of total RNA was used for cDNA library construction. cDNA library construction and normalisation were performed through Ambry Genetics, Alisso Viejo, CA, USA. The resulting library was sequenced using Illumina High Seq. and assembled de novo using the SOAP (Short Oligonucleotide Analysis Package) and Trinity assemblers (Ambry Genetics, Aliso Viejo, CA, USA).

Selection of unigenes encoding for transcription factors

Unigenes selected from the mature *M. oleifera* transcriptome encoding for transcription factors were identified by cross-referencing the identifiers of these unigenes with the identifiers of TFs catalogued in the Plant Transcription Factor Databases (TFDBs). The consultation of datasets in the current study involved the following steps: 1) identifiers of the unigenes (GenBank Protein Accessions for SOAPAnn unigenes and UniProt Entry Names for TriAnn unigenes) were converted to the type of identifier of entries in the dataset of interest, then 2) converted identifiers were matched with identifiers in the said dataset. Queries on the online sites of the databases were performed using identifiers applicable to each database. The second selection method involved assignment of transcripts into expression level groups based on the methodology by Toung et al. (2011). The selected unigenes were cross-referenced, through the use of identifiers against TAIR (The Arabidopsis Information Resource), NCBI gene2go for GO (Gene Ontology); AtTFDB (*Arabidopsis thaliana* Transcription Factor Data Bases); DATF (Database of Arabidopsis Transcription Factors); PlnTFDB (Plant Transcription Factor Database) for Transcription Factor families.

RESULTS AND DISCUSSION

Ten TF families with the most unigenes in the SOAP assembly

The Trihelix family contained the most unigenes in two TFDBs (Figure 1). ARABIDOPSIS 6BINTERACTING PROTEIN1-LIKE1 (ASIL1) and ASIL2, are 2 trihelix proteins, expressed prominently in the seed. These proteins are known to inhibit the operation of maturation in early embryogenesis, repressing maturation processes until the heart stage (Barr et al., 2012). There were 11 unigenes categorized as *C2H2* in AtTFDB. Titan like (*TTL*) controls nuclear division in the endosperm (Lu et al., 2012). TRANSPARENT TESTA 1 (*TT1*) engages in the development of the seed coat endothelium. THE INDETERMINATE DOMAIN1 (*IDD1*) on ENHYDROUS (*ENY*) maintains seed maturation such as suppression of photosynthesis (Lu et al., 2012).

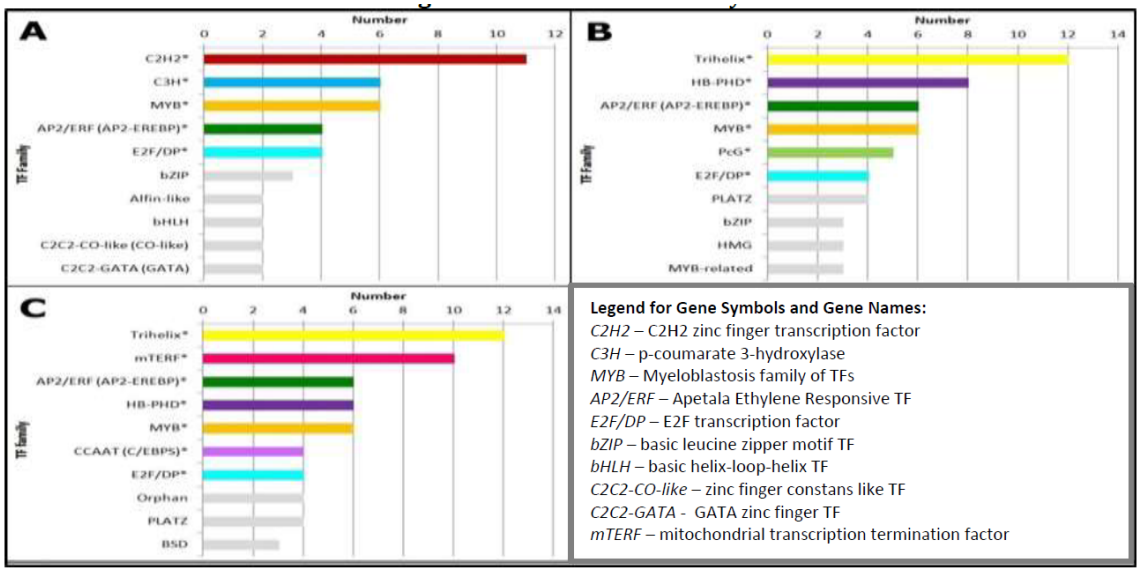

Figure 1. Numbers of SOAP Assembly DB-selected unigenes (putative TFs) classified into families in A) AtTFDB (*Arabidopsis thaliana* Transcription Factor Data Base), B) DATF (Database of Arabidopsis Transcription Factors), and C) PlnTFDB (Plant Transcription Factor Data Base) according to SOAP de novo Trans-Annotation. Other gene symbols and gene names not written in the box: *HB-PHD* – Hobeobox Cys4-His-Cys3 motif; *CCAAT (C/EBPS)* – CCAAT Enhancer Binding Proteins; *PLATZ* – plant AT-rich sequence and zinc-binding proteins; *BSD* – Blasticidin S Deaminase; *PcG* – Polycomb Group.

The TF family *mTERF* had 10 transcripts categorized in PlnTFDB. mTERF proteins facilitate transcription initiation as well as modulation of DNA replication in the mitochondria. There were 8 and 6 unigenes categorized as *HB-PHD* TFs on DATF and PlnTFDB respectively. The *HB-PHD* category is one of the six subdivisions within the *HB/HD* family (Chew et al., 2013), distinguished from the other subgroups Cys4-His-Cys3 motif. Members of this group participate in transcriptional regulation involving chromatin (Plesch et al., 1997; Rocha et al., 2005). PlnTFDB had 4 unigenes in the *CCAAT* (*C/EBPS*) group. The family of NF-Y TFs consists of members defined by their subunits: NF-YA (CBFB), NF-YB (CBF-A), and NF-YC (CBF-C) which may act as activators or repressors (Romier et al., 2003). A member of the NF-YB Subgroup is LEC1, also known as NF-YB9, which is a regulator of seed maturation (Siefers et al., 2009; Graeber et al., 2012). Among the genes if affects are *FUS3* and *ABI3*, having known control of storage proteins for the seed (Yamamoto et al., 2009). All 3 TFDBs categorized four unigenes into the *E2F/DP* family involved in the regulation of cell cycle and cell proliferation.

There were 4, 6 and 6 unigenes belonging to the *AP2-EREBP* superfamily in AtTFDB, DATF and PlnTFDB, respectively. The *AB-EREBP* superfamily are involved in seed development. APETALA (*AP2*) participates in the development of the seed coat (Jofuku et al., 1994). *AP2* determines seed size and influences both the maternal tissues and the embryo (Ohto). There were six unigenes belonging to the *C3H* family, according to AtTFDB. The *C3H* family maybe involved early in the filling of the kernel of *Z. mays* seeds (Liu et al., 2005). There were six unigenes classified into the MYB family in all TFDBs. The *R2R3MYB*-class is involved in the development of the endosperm (Ambawat et al., 2013). *MYB118* modulates the synthesis of storage nutrients and the partitioning of the embryo and endosperm (Ambawat et al., 2013). Five unigenes were categorized according to DATF into the *PcG* family which participate in seed development. Some of these were *FERTILIZATION INDEPENDENT SEED2* (*FIS2*) and *MEDEA* which regulate the proliferation of both the embryo and the endosperm (Grossniklaus et al., 1998; Köhler et al., 2003).

Top ten families of TF unigenes identified from the Trinity transcriptome assembly

Among the TF families having high number of detected TFs across TFDBs for the Trinity Assembly were *HB*, *MYB*-related and *ARR-B* (*GARP-ARR-B*) (Figure 2). Families with consistently high numbers of transcripts categorized into them, based on two or more plant TFDBs were: *HB* (*Homeobox*) [AtTFDB:1474, DATF:1461 and PlnTFDB:1474], *MYB*-related (AtTFDB:526, DATF:341 and PlnTFDB:454) and *HB-PHD* (DATF:594 and PlnTFDB:578). Some of the TFs identified from Trinity had also been identified in the SOAP Assembly (*HB-PHD*, *MYB/MYB* related, *CH3H*, *C2H2* and *AP2-EREBP*). Hence, this section outlined only *HB* (Homeobox), *JUMONJI*, *bHLH* and *ARR-B* (*GARP-ARR-B*).

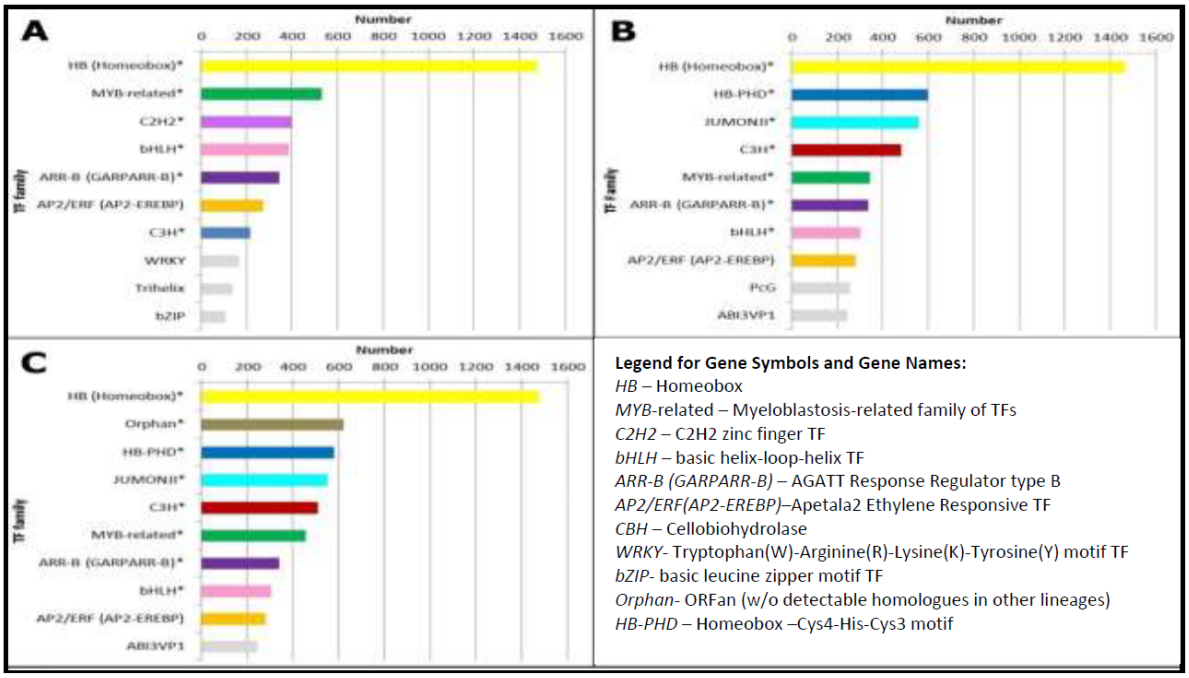

Figure 2. Numbers of Trinity Assembly DB-selected unigenes (putative TFs) classified into families in A) AtTFDB (*Arabidopsis thaliana* Transcription Factor Data Base), B) DATF (Database of Arabidopsis Transcription Factors), and C) PlnTFDB (Plant Transcription Factor Data Base) according to Trinity Annotation. Other gene symbols and gene names that are not written in the box were: *Jumonji*; *C3H – p-coumarate 3-hydroxylase*; *ABI3VP1* – Abscissic Acid Insensitive VP1 TF; *PcG* – Polycomb Group.

The *Jumonji* family was among the top five groups with the most transcripts in DATF

and PlnTFDB, with 555 and 549, respectively. The JUMONJI (*JMJ*) TF family takes its name from "jumonji" meaning cross-like (Takeuchi et al., 2006). These TFs generally participate in chromosome remodeling and floral regulation (Takeuchi et al., 2006). Two *A. thaliana JMJ* TFs, *JMJ*20 and 22 promote seed germination. The two are histone arginine demethylases which repress GIBBERELLIN3 β-HYDROXYLASE1 and 2 (GA3 ox1 and 2) (Footitt et al., 2013). *JMJ*20 and 22 are dependent on active phytochrome B (*PHYB*), signaling them to increase levels of GA for seed germination (Cho et al., 2012). There were 383 unigenes in the AtTFDB: 296 in the DATF and 303 in the PlnTFD classified as *bHLH* TFs. One bHLH TF is reported to be a homolog to RETARDED GROWTH OF EMBRYO1 (*RGE1*) of *A. thaliana* (Xue et al., 2012) which is expressed in the endosperm but regulates the growth of embryo following the heart stage. *bHLH* is a member of subfamily *1b* (Kondou et al., 2008). Two associated bHLH TFs in the seed of *A. thaliana* were ZHOUPI (*ZOU*) and INDUCER OF CBP EXPRESSION 1 (*ICE1*) expressed in the embryo, endosperm and testa (Chinnusamy et al., 2003). Their function involves the breakdown of the endosperm (Chinnusamy et al., 2003). The former also affects development of the embryo cuticle coordinated by GSO1 and GSO2 receptor kinases, via activation of ABNORMAL LEAF SHAPE1 (*ALE1*), a subtilisin serine protease, the expression of which is restricted in the endosperm (Ulf et al., 2007).

TF Families of unigenes that are upregulated in the SOAP assembly

The classification of transcripts across TFDBs varied (Table 1). Some unigenes had two out of three or three out of three consensus on classification across TFDBs.

Table 1. TF families of ten selected unigenes with FPKM values at the top 25th percentile according to SOAP annotation.

TAIR ID	Subject description[1]	TF family		
		AtTFDB	DATF	PlnTFDB
AT3G14180	Sequence specific DNA binding TF	Trihelix	Trihelix	Trihelix
AT4G22140	PHD finger protein	-	HB-PHD	HB-PHD
AT5G51230	Polycomb group protein EMB flower	C2H2	-	-
AT5G51980	Transducin/WD40 repeat-like superfamily protein	-	C3H	C3H
AT2G02740	Single-stranded DNA binding protein WHY3	WHIRLY	WHIRLY	WHIRLY
AT4G38960	B-box type zinc finger acting protein	C2C2-CO-like	-	ORPHAN
AT2G30410	Tubulin folding co-factor A protein KIESEL	TCP	-	-
AT4G36860	LIM domain acting protein	-	LIM	-
AT1G56200	Embryo defective protein 1303	C2H2	-	-
AT3G18870	Mitochondrial transcription termination factor family protein	-	-	mTERF

[1]The "subject description" lists the UniProt Protein Name of the unigenes.

The sequence-specific DNA binding transcription factor was the most upregulated unigene based on the FPKM values at the top 25th percentile according to SOAP annotation. Sequence specific DNA binding transcription factor is the other name for ASIL2 (Ref.). Seed maturation in *A. thaliana* requires the coordination of ASIL2 in conjunction with ASIL1 and HDA6/SIL1 (Barr et al., 2012). It is known to repress seed maturation during early embryogenesis as well as after germination (Willmann et al., 2011). The PHD finger and bromo adjacent homology domain containing protein also known as EARLY BOLTING IN SHORT DAYS (EBS), plays a role in seed dormancy (Sung and Amasino, 2004). It is related to chromatin remodeling factors (Luo and Dean, 1999). It has other reported functions, including roles in flower induction and floral homeotic gene expression (Piñeiro et al., 2003; López-González et al., 2014). The expression of Embryonic Flower 2 (EMF2) in developing embryos is up until the globular stage and in the endosperm is up until cellularization (Yoshida et al., 2001). EMF2's function is indicated to be for repression of seed development. This function is consistent with the role of Polycomb group protein (Pcg) in gene silencing through modification of histones (Schatlowski et al., 2008). Tansducin/WD40 repeat-like

superfamily protein share a number of functions with its WD40 repeat protein family relatives (Gachomo et al., 2014). Transducin is expressed in the embryo and endosperm and suppresses germination (Gachomo et al., 2014). Transducin induces ethylene production increasing the activity of hydrolases in the cell walls of the endosperm cap, thereby leading to seed germination (Gachomo et al., 2014). *WHIRLY* is involved in response and resistance to disease via salicylic acid (SA) signaling (Desveaux et al., 2005). The B-box type zinc-finger containing proteins does not seem to have functions involving the seed (Khanna et al., 2009). The roles of its members include responses to light and circadian rhythm (Khanna et al., 2009). An example is BBX19, which regulates flowering time negatively (Khanna et al., 2009). One TF the KIESEL (tubulin folding co-factor) is involved in cell division (Lu et al., 2010). KIESEL acts as a molecular post-chaperonin and participates in the β- tubulin-folding- pathway expressed in the early development of the endosperm (Lu et al., 2010). Similarly, the mTERF family protein is involved in embryo development associated with the synthesis of amino acids, nucleotides and vitamins, for the development of the embryo (Zhao et al., 2014). The LIM domain-containing protein is involved in determining seed size (Maturana et al., 2011). The DA1- RELATED PROTEIN1 (DAR1) functions to maintain the size of organs and seeds by setting the length of time those plant parts can develop (Maturana et al., 2011).

The identification of the upregulated unigenes in the SOAP annotation encoding for TFs in the mature seed embryos of *M. oleifera* and the determination of their putative functions are crucial because it shed light in understanding the processes taking place in the *M. oleifera* mature seed embryos. The putative functions of the aforementioned TFs showed that these TFs are the major regulators of important processes such as embryogenesis, seed maturation, determination of the seed size, seed germination, early development of the endorsperm and seed dormancy. These TFs are situated at, or near the top of the transcriptional cascades thus controlling the expression genes significant for the processes in the *M. oleifera* embryos. *M. oleifera* is known for its many nutritional and medicinal properties, hence understanding of the functions of these TFs can shed light on how the M. oleifera seeds acquire those highly beneficial medicinal and nutritional properties. As the embryo increases in size and mass, storage products such as oils, proteins and starch were deposited. These accumulated storage reserves contribute to the beneficial medicinal and nutritional properties of *M. oleifera* seeds.

Top ten highly expressed TF unigenes for the Trinity assembly

The unigenes identified in *Moringa* with complete consensus classification were bHLH96 (*bHLH*); TCP19(*TCP*); *WRI1*, *AIL5* and At2g41710 (*AP2-EREBP*); and NF-YA3 (*CCAAT*) (Table 2).

Of the ten representative unigenes (Table 2) seven have putative functions in relation to seed processing. Ethylene-responsive transcription factor *WRI1* is important in the production and storage of nutrients for embryo development (Kilaru et al., 2015). Its regulatory mechanism links carbohydrate glycolytic activity to FA metabolism. The activity of *WRI1* is also connected to stress response (Focks and Benning, 1998). A homolog in *R. communis*, 30069.m00440, is similarly involved in oil metabolism (Xu et al., 2013). The WRI1 TF is not restricted to the seed, its oil biosynthetic roles in other tissue have been documented. An ortholog, EgWRI1, was found in oil palm mesocarp tissue (Ma et al., 2013). Another ortholog has been found in *Persea americana* (avocado) mesocarp (Kilaru et al., 2015). The AP2-like ethylene responsive transcription factor is known for names such as PLETHORA (PLT5), EMBRYOMAKER (EMK), and AIL5. PLT5 functions in the maintenance of embryonic identity in developing and mature embryo (Jofuku et al., 1994). Functions of PLT5 involving meristematic regions particularly in leaf development and phyllotaxis have been demonstrated (Prasad et al., 2011; Pinon et al., 2013). Nuclear transcription factor Y subunit A-3 participates in the regulation of early embryogenesis, particularly from the globular to the torpedo stages (Siefers et al., 2009; Graeber et al., 2012). It may affect the transport of auxin, which is responsible for cell division (Siefers et al., 2009; Graeber et al., 2012). Transcription factor bHLH is expressed in the seed and functions in the breakdown of

the endosperm and development of the embryo (Kondou et al., 2008). The B3 domain-containing protein REM11 functions for seed maturation and embryo development. The HMG function by negatively regulating seed germination, especially in the context of high salinity and low water levels (Kwak et al., 2007; Pedersen et al., 2010). The second, zinc finger CCCH domain containing protein 20 with a number of epithets (i.e., AtC3H20, AtTZF2, and *A. thaliana* oxidation related Zinc Finger 1 [AtOZF1]) is not involved in repression of seed germination but is restricted in the plasma membrane and plays a role in defense and tolerance response against oxidative stress involving both ABA and jasmonic acid. TCP19 does not directly influence seed development but rather functions in regulating the senescence of leaves (Danisman et al., 2013).

Table 2. TF families of the ten selected representative unigenes with FPKM values at the top 25th percentile according to annotation using Trinity.

TAIR ID	Subject description	TF family		
		AtTFDB	DATF	PlnTFBD
AT1G20693	High mobility group B protein 2	-	HMG	HMG
AT1G72210	Transcription factor bHLH96	bHLH	bHLH	bHLH
AT5G51910	Transcription factor TCP19	TCP	TCP	TCP
AT3G54320	Ethylene-responsive transcription factor WRI1	AP2/ERF (APS-EREBP)	AP2/ERF (APS-EREBP)	AP2/ERF (APS-EREBP)
AT2G41710	AP2-like ethylene-responsive transcription factor At2g41710	AP2/ERF (APS-EREBP)	AP2/ERF (APS-EREBP)	AP2/ERF (APS-EREBP)
AT2G24681	B3 domain-containing protein REM11	-	-	-
AT5G57390	AP2-like ethylene-responsive transcription factor AIL5	AP2/ERF (APS-EREBP)	AP2/ERF (APS-EREBP)	AP2/ERF (APS-EREBP)
AT2G19810	Zinc finger CCCH domain-containing protein 20	-	C3H	C3H
AT1G72830	Nuclear transcription factor Y subunit A-3	CCCAATHAP2 (NF-YA/CBF-B)	CCAATHAP2 (NF-YA/CBF-B)	CCAAT (C/EBPS)
AT5G13820	Telomere repeat binding protein 4	MYB-related	-	-

The classification of the upregulated unigenes in the Trinity annotation encoding for TFs in the mature seed embryos of *M. oleifera* and the their putative functions are significant since it enabled greater insight into the processes occurring in the *M. oleifera* mature seed embryos such as the production and storage of nutrients for seed development, seed oil biosynthesis and metabolism. *M. oleifera* is known for its oil (the Ben) oil which is a polyunsaturated fatty acid and is a healthy oil. These TFs control the expression of genes involved in oil biosynthesis. The maturation of seed embryos marked the beginning of reserve deposition. Embryonic development in plants particularly for *M. oleifera* is the most crucial and consequently also the most sensitive to environmental stresses, hence, aside from the TFs involved in embryogenesis and seed maturation some TFs identified are also involved in defense and tolerance against oxidative stress.

Gene ontology (GO) functions indicate involvement of many unigenes in transcription

The GO processes that are mostly associated with selected unigenes for both annotations show that many of these unigenes (3,290) are sequence specific DNA-binding transcription factors (Figure 3), giving merit to their selection as putative TFs. Many of the DB-selected unigenes (5,241) have DNA-templated, transcriptional regulatory roles, which clearly indicate their involvement in transcription. The unigenes that are involved in the regulation of transcription could contain activator or repressor domains (Guo et al., 2008a; Du et al., 2012; Gupta et al., 2015). Meanwhile, many unigenes are involved in processes specific to a component of the cell, tissue, or pathway; this observation is based on the top ten biological processes (Figure 3).

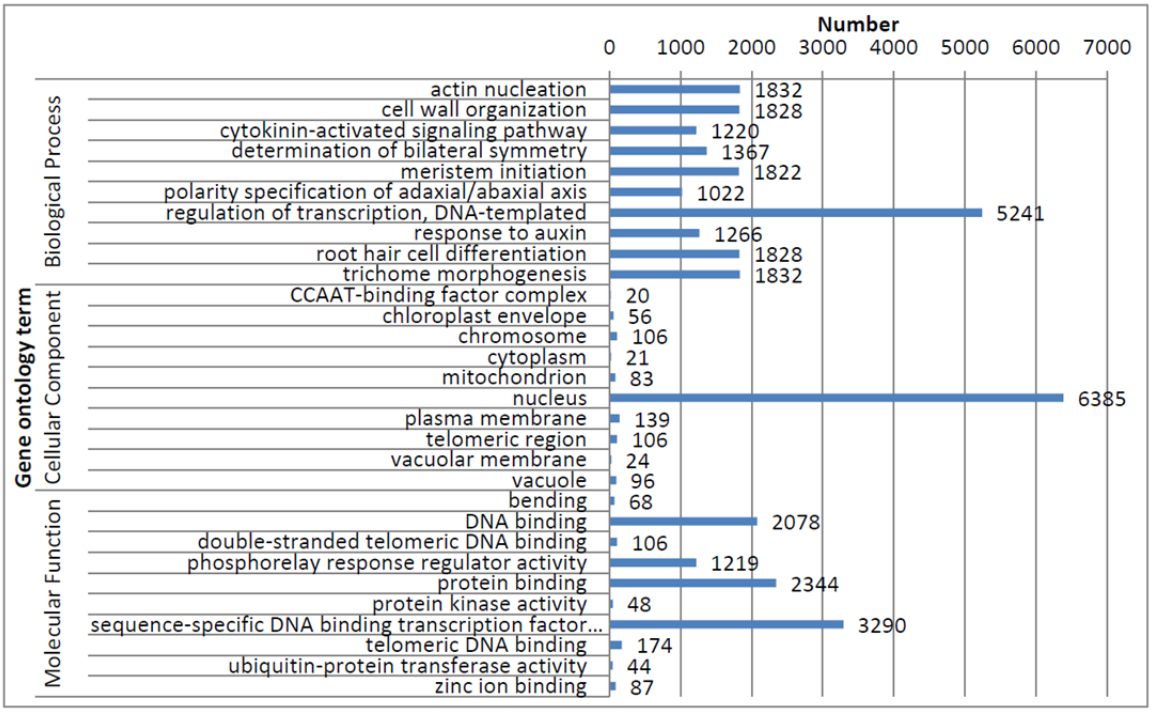

Figure 3. Top 10 unigenes per GO classes most-associated with the top-25th-percentile-FPKM, DB-selected unigenes of involving molecular function, cellular component and biological process.

Furthermore, many of the transcripts (6,385) have GO functions that are localized in the nucleus (Figure 3). It has been reported that some TFs have nuclear localization signal (NLS) domain that restricts them to the nucleus (Guo et al., 2008a; Du et al., 2012; Gupta et al., 2015). Also, transcription happens in the nucleus, chloroplasts, and mitochondria, hence, TF functions were expressed in these organelles. Furthermore, DNA-binding and protein-binding activities are associated with TFs, due to the presence in many of DNA-binding domains (DBDs) and protein-protein interaction domains (Guo et al., 2008a; Du et al., 2012; Gupta et al., 2015).

The putative functions identified through GO annotations confirmed the identity of the unigenes as TFs in the *M. oleifera* mature seed embryos because of their DNA-binding and protein-binding activities. TFs are typically classified as activators or repressors of gene expression. Activators recruit coactivators, resulting in gene activation, while repressors recruit corepressors, leading to transcriptional repression (Boyle and Després, 2010). The TFs are largely accounted as the major controllers of embryogenesis and seed maturation because the repressing and activating functions of this factor are carried out at the same locus through the same *cis*-element in a promoter. Mutation of the genes encoding these TFs can have severe detrimental effects on the seed maturation of *M. oleifera* seed embryos, including reduction including seed oil accumulation. Conversely, increasing levels of their mRNA can lead to increased seed oil accumulation.

CONCLUSIONS AND RECOMMENDATION

Unigenes encoding for transcription factors in *M. oleifera* mature seed embryos were characterized through species identity, gene expression levels, TF family classification and gene ontology. Putative functions of the *M. oleifera* TFs in the Trinity and SOAP annotations were determined. The Trihelix family contained the most unigenes identified to inhibit the operation of maturation in early embryogenesis. The *C2H2* controls nuclear division in the endosperm, involved in the development of the seed coat endothelium and maintains seed maturation. The NF-Y TFregulates seed maturation. There were 4 unigenes in the *E2F/DP*

family that are involved in the regulation of cell cycle and cell proliferation. There were 16 unigenes in the *AP2-EREBP* superfamily involved in seed development. *AP2* participates in the development of the seed coat and determines seed size. Six unigenes were classified into the *MYB* family involved in the development of the endosperm and synthesis of storage nutrients as well as embryo and endosperm partitioning. Five unigenes involved in seed development were identified in the *PcG* family. The *Jumonji* family was among the top five groups with the most transcripts in the Trinity annotation. Jumonji participates in floral regulation and promote seed germination. The *bHLH* TFs involved in the breakdown of the endosperm were also identified in the Trinity annotation. Most transcripts identified in the Trinity annotation were also identified in the SOAP annotation.

The sequence-specific DNA binding TF known to represses seed maturation during early embryogenesis and after germination was the most upregulated unigene in the SOAP annotation based on FPKM values. It was followed by the *PHD* finger which plays a role in seed dormancy, flower induction and floral homeotic gene expression. EMF2's function is indicated to be for repression of seed development which is consistent with the role of Polycomb group protein (Pcg) in gene silencing by histone modification. Tansducin/WD40 induces ethylene production leading to seed germination. KIESEL acts as a molecular post-chaperonin expressed in the early development of the endosperm. mTERF is involved in the synthesis of amino acids, nucleotides and vitamins for embryonic development. The LIM domain-containing protein is involved in determining seed size. The *HMG* is an upregulated unigene based on the 25th percentile FPKM values identified to negatively regulate seed germination. The ethylene-responsive transcription factor *WRI1* is another upregulated unigene important in the production and storage of nutrients for embryo development. Nuclear transcription factor Y participates in the regulation of early embryogenesis affecting the transport of auxin responsible for cell division. The B3 domain-containing protein REM11 functions for seed maturation and embryo development. The second, zinc finger CCCH domain-containing protein plays a role in defense and tolerance response against oxidative stress.

The identification of the upregulated unigenes in the SOAP and Trinity annotations encoding for TFs in the mature seed embryos of *M. oleifera* and its putative functions are crucial because it enabled greater insight in the processes occurring in the *M. oleifera* mature seed embryos. The putative functions of these TFs show that they are the major regulators of important processes such as embryogenesis, seed maturation, production and storage of nutrients for seed development, oil metabolism, determination of the seed size, seed germination, early development of the endorsperm and seed dormancy. These TFs are situated at the top of the transcriptional cascades thus controlling the expression of genes significant for the processes in the *M. oleifera* embryos. *M. oleifera* is known for its many nutritional and medicinal properties, hence understanding the functions of these TFs shed light on how the *M. oleifera* seeds obtained its beneficial properties. As the embryo increases in size, and mass, storage products such as oils, proteins and starch were deposited. These accumulated storage reserves contribute to the beneficial and nutritional properties of *M. oleifera* seeds. *M. oleifera* is known for its oil (the Ben) oil which is a polyunsaturated fatty acid and is a healthy oil. These TFs control the expression of genes involved in oil biosynthesis. Embryonic development in plants particularly for *M. oleifera* is the most crucial and consequently also the most sensitive to environmental stresses, hence, other than the TFs involved in embryogenesis and seed maturation some TFs identified are involved in defense and tolerance against oxidative stress.

Unigenes categorized based on Gene Ontology (GO) from the SOAP and Trinity annotations show that many of these unigenes (3,290) are sequence specific DNA-binding transcription factors. A good number of unigenes (5,241) have DNA-templated, transcriptional regulatory roles indicating their involvement in transcription. Majority of the transcripts (6,385) have GO functions localized in the nucleus because transcription happens in the nucleus. The TFs are largely accounted as the major controllers of embryogenesis and seed maturation because the repressing and activating functions of these TFs are carried out at the same locus through the same *cis*-element in a promoter.

Mutation of the genes encoding these TFs can have severe detrimental effects on the seed maturation of *M. oleifera* seed embryos, including reduction and seed oil accumulation. Conversely, increasing levels of their mRNA can lead to increased seed oil accumulation.

The validation of the TF unigenes identified from the mature embryo of *M. oleifera* should be experimentally validated through real-time qPCR which is an efficient method for the detection of gene expression occurring at a specific stage of development as in the mature embryo of *M. oleifera*. It is an efficient method that is less laborious and time consuming and can detect resolution twenty times more than the traditional PCR.

Literature cited

Abdull Razis, A.F., Ibrahim, M.D., and Kntayya, S.B. (2014). Health benefits of *Moringa oleifera.* Asian Pac. J. Cancer Prev. *15* (*20*), 8571–8576 https://doi.org/10.7314/APJCP.2014.15.20.8571. PubMed

Ambawat, S., Sharma, P., Yadav, N.R., and Yadav, R.C. (2013). *MYB* transcription factor genes as regulators for plant responses: an overview. Physiol Mol Biol Plants *19* (*3*), 307–321 https://doi.org/10.1007/s12298-013-0179-1. PubMed

Anwar, F., Latif, S., Ashraf, M., and Gilani, A.H. (2007). *Moringa oleifera*: a food plant with multiple medicinal uses. Phytother Res *21* (*1*), 17–25 https://doi.org/10.1002/ptr.2023. PubMed

Barr, M.S., Willmann, M.R., and Jenik, P.D. (2012). Is there a role for trihelix transcription factors in embryo maturation? Plant Signal Behav *7* (*2*), 205–209 https://doi.org/10.4161/psb.18893. PubMed

Boyle, P., and Després, C. (2010). Dual-function transcription factors and their entourage. Plant Signal Behav *5* (*6*), 629–634 https://doi.org/10.4161/psb.5.6.11570. PubMed

Chew, W., Hrmova, M., and Lopato, S. (2013). Role of Homeodomain leucine zipper (HD-Zip) IV transcription factors in plant development and plant protection from deleterious environmental factors. Int J Mol Sci *14* (*4*), 8122–8147 https://doi.org/10.3390/ijms14048122. PubMed

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B.H., Hong, X., Agarwal, M., and Zhu, J.K. (2003). ICE1: a regulator of cold-induced transcriptome and freezing tolerance in *Arabidopsis.* Genes Dev. *17* (*8*), 1043–1054 https://doi.org/10.1101/gad.1077503. PubMed

Cho, J.N., Ryu, J.Y., Jeong, Y.M., Park, J., Song, J.J., Amasino, R.M., Noh, B., and Noh, Y.S. (2012). Control of seed germination by light-induced histone arginine demethylation activity. Dev. Cell *22* (*4*), 736–748 https://doi.org/10.1016/j.devcel.2012.01.024. PubMed

Danisman, S., van Dijk, A.D.J., Bimbo, A., van der Wal, F., Hennig, L., de Folter, S., Angenent, G.C., and Immink, R.G.H. (2013). Analysis of functional redundancies within the Arabidopsis TCP transcription factor family. J. Exp. Bot. *64* (*18*), 5673–5685 https://doi.org/10.1093/jxb/ert337. PubMed

Desveaux, D., Maréchal, A., and Brisson, N. (2005). Whirly transcription factors: defense gene regulation and beyond. Trends Plant Sci. *10* (*2*), 95–102 https://doi.org/10.1016/j.tplants.2004.12.008. PubMed

Du, H., Yang, S.S., Liang, Z., Feng, B.R., Liu, L., Huang, Y.B., and Tang, Y.X. (2012). Genome-wide analysis of the MYB transcription factor superfamily in soybean. BMC Plant Biol. *12* (*1*), 106 https://doi.org/10.1186/1471-2229-12-106. PubMed

Focks, N., and Benning, C. (1998). wrinkled1: A novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. Plant Physiol. *118* (*1*), 91–101 https://doi.org/10.1104/pp.118.1.91. PubMed

Footitt, S., Huang, Z., Clay, H.A., Mead, A., and Finch-Savage, W.E. (2013). Temperature, light and nitrate sensing coordinate Arabidopsis seed dormancy cycling, resulting in winter and summer annual phenotypes. Plant J. *74* (*6*), 1003–1015 https://doi.org/10.1111/tpj.12186. PubMed

Gachomo, E.W., Jimenez-Lopez, J.C., Baptiste, L.J., and Kotchoni, S.O. (2014). GIGANTUS1 (GTS1), a member of Transducin/WD40 protein superfamily, controls seed germination, growth and biomass accumulation through ribosome-biogenesis protein interactions in *Arabidopsis thaliana.* BMC Plant Biol. *14* (*1*), 37 https://doi.org/10.1186/1471-2229-14-37. PubMed

Graeber, K., Nakabayashi, K., Miatton, E., Leubner-Metzger, G., and Soppe, W.J. (2012). Molecular mechanisms of seed dormancy. Plant Cell Environ. *35* (*10*), 1769–1786 https://doi.org/10.1111/j.1365-3040.2012.02542.x. PubMed

Grossniklaus, U., Vielle-Calzada, J.P., Hoeppner, M.A., and Gagliano, W.B. (1998). Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. Science *280* (*5362*), 446–450 https://doi.org/10.1126/science.280.5362.446. PubMed

Guo, A.Y., Zhu, Q.H., Gu, X., Ge, S., Yang, J., and Luo, J. (2008a). Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family. Gene *418* (*1-2*), 1–8 https://doi.org/10.1016/j.gene.2008.03.016. PubMed

Gupta, S., Malviya, N., Kushwaha, H., Nasim, J., Bisht, N.C., Singh, V.K., and Yadav, D. (2015). Insights into structural and functional diversity of Dof (DNA binding with one finger) transcription factor. Planta *241* (*3*), 549–562 https://doi.org/10.1007/s00425-014-2239-3. PubMed

Jofuku, K.D., den Boer, B.G., Van Montagu, M., and Okamuro, J.K. (1994). Control of *Arabidopsis* flower and seed development by the homeotic gene APETALA2. Plant Cell *6* (*9*), 1211–1225 https://doi.org/10.1105/tpc.6.9.1211. PubMed

Khanna, R., Kronmiller, B., Maszle, D.R., Coupland, G., Holm, M., Mizuno, T., and Wu, S.H. (2009). The *Arabidopsis* B-box zinc finger family. Plant Cell *21* (*11*), 3416–3420 https://doi.org/10.1105/tpc.109.069088. PubMed

Kilaru, A., Cao, X., Dabbs, P.B., Sung, H.J., Rahman, M.M., Thrower, N., Zynda, G., Podicheti, R., Ibarra-Laclette, E., Herrera-Estrella, L., et al. (2015). Oil biosynthesis in a basal angiosperm: transcriptome analysis of *Persea Americana* mesocarp. BMC Plant Biol. *15* (*1*), 203 https://doi.org/10.1186/s12870-015-0586-2. PubMed

Köhler, C., Hennig, L., Spillane, C., Pien, S., Gruissem, W., and Grossniklaus, U. (2003). The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. Genes Dev. *17* (*12*), 1540–1553 https://doi.org/10.1101/gad.257403. PubMed

Kondou, Y., Nakazawa, M., Kawashima, M., Ichikawa, T., Yoshizumi, T., Suzuki, K., Ishikawa, A., Koshi, T., Matsui, R., Muto, S., and Matsui, M. (2008). RETARDED GROWTH OF EMBRYO1, a new basic helix-loop-helix protein, expresses in endosperm to control embryo growth. Plant Physiol. *147* (*4*), 1924–1935 https://doi.org/10.1104/pp.108.118364. PubMed

Kwak, K.J., Kim, J.Y., Kim, Y.O., and Kang, H. (2007). Characterization of transgenic Arabidopsis plants overexpressing high mobility group B proteins under high salinity, drought or cold stress. Plant Cell Physiol. *48* (*2*), 221–231 https://doi.org/10.1093/pcp/pcl057. PubMed

Leone, A., Spada, A., Battezzati, A., Schiraldi, A., Aristil, J., and Bertoli, S. (2016). *Moringa oleifera* sees and oil: characteristics and uses for human health. Int J Mol Sci *17* (*12*), 2141 https://doi.org/10.3390/ijms17122141. PubMed

Liu, P.P., Koizuka, N., Martin, R.C., and Nonogaki, H. (2005). The BME3 (Blue Micropylar End 3) GATA zinc finger transcription factor is a positive regulator of Arabidopsis seed germination. Plant J. *44* (*6*), 960–971 https://doi.org/10.1111/j.1365-313X.2005.02588.x. PubMed

Liu, Y., Huang, Z., Ao, Y., Li, W., and Zhang, Z. (2013). Transcriptome analysis of yellow horn (*Xanthoceras sorbifolia* Bunge): a potential oil-rich seed tree for biodiesel in China. PLoS ONE *8* (*9*), e74441 https://doi.org/10.1371/journal.pone.0074441. PubMed

López-González, L., Mouriz, A., Narro-Diego, L., Bustos, R., Martínez-Zapater, J.M., Jarillo, J.A., and Piñeiro, M. (2014). Chromatin-dependent repression of the Arabidopsis floral integrator genes involves plant specific PHD-containing proteins. Plant Cell *26* (*10*), 3922–3938 https://doi.org/10.1105/tpc.114.130781. PubMed

Lu, L., Nan, J., Mi, W., Wei, C.H., Li, L.F., and Li, Y. (2010). Crystallization and preliminary X-ray analysis of tubulin-folding cofactor A from *Arabidopsis thaliana.* Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *66* (*8*), 954–956 https://doi.org/10.1107/S1744309110023900. PubMed

Lu, X., Li, Y., Su, Y., Liang, Q., Meng, H., Li, S., Shen, S., Fan, Y., and Zhang, C. (2012). An *Arabidopsis* gene encoding a C2H2-domain protein with alternatively spliced transcripts is essential for endosperm development. J. Exp. Bot. *63* (*16*), 5935–5944 https://doi.org/10.1093/jxb/ers243. PubMed

Luo, R.X., and Dean, D.C. (1999). Chromatin remodeling and transcriptional regulation. J. Natl. Cancer Inst. *91* (*15*), 1288–1294 https://doi.org/10.1093/jnci/91.15.1288. PubMed

Ma, W., Kong, Q., Arondel, V., Kilaru, A., Bates, P.D., Thrower, N.A., Benning, C., and Ohlrogge, J.B. (2013). Wrinkled1, a ubiquitous regulator in oil accumulating tissues from Arabidopsis embryos to oil palm mesocarp. PLoS ONE *8* (*7*), e68887 https://doi.org/10.1371/journal.pone.0068887. PubMed

Maturana, A.D., Nakagawa, N., Yoshimoto, N., Tatematsu, K., Hoshijima, M., Tanizawa, K., and Kuroda, S. (2011). LIM domains regulate protein kinase C activity: a novel molecular function. Cell. Signal. *23* (*5*), 928–934 https://doi.org/10.1016/j.cellsig.2011.01.021. PubMed

Pedersen, D.S., Merkle, T., Marktl, B., Lildballe, D.L., Antosch, M., Bergmann, T., Tönsing, K., Anselmetti, D., and Grasser, K.D. (2010). Nucleocytoplasmic distribution of the Arabidopsis chromatin-associated HMGB2/3 and HMGB4 proteins. Plant Physiol. *154* (*4*), 1831–1841 https://doi.org/10.1104/pp.110.163055. PubMed

Piñeiro, M., Gómez-Mena, C., Schaffer, R., Martínez-Zapater, J.M., and Coupland, G. (2003). EARLY BOLTING IN SHORT DAYS is related to chromatin remodeling factors and regulates flowering in *Arabidopsis* by repressing FT.

Plant Cell *15* (*7*), 1552–1562 https://doi.org/10.1105/tpc.012153. PubMed

Pinon, V., Prasad, K., Grigg, S.P., Sanchez-Perez, G.F., and Scheres, B. (2013). Local auxin biosynthesis regulation by PLETHORA transcription factors controls phyllotaxis in *Arabidopsis.* Proc. Natl. Acad. Sci. U.S.A. *110* (*3*), 1107–1112 https://doi.org/10.1073/pnas.1213497110. PubMed

Plesch, G., Störmann, K., Torres, J.T., Walden, R., and Somssich, I.E. (1997). Developmental and auxin-induced expression of the Arabidopsis prha homeobox gene. Plant J. *12* (*3*), 635–647 https://doi.org/10.1046/j.1365-313X.1997.d01-15.x. PubMed

Prasad, K., Grigg, S.P., Barkoulas, M., Yadav, R.K., Sanchez-Perez, G.F., Pinon, V., Blilou, I., Hofhuis, H., Dhonukshe, P., Galinha, C., et al. (2011). *Arabidopsis* PLETHORA transcription factors control phyllotaxis. Curr. Biol. *21* (*13*), 1123–1128 https://doi.org/10.1016/j.cub.2011.05.009. PubMed

Rajalakshmi, R., Rajalakshmi, S., and Parida, A. (2017). Evaluation of the genetic structure in drumstick (*Moringa oleifera* Lam.) Using SSR markers. Curr. Sci. *112* (*6*), 1250–1256 https://doi.org/10.18520/cs/v112/i06/1250-1256.

Rocha, G.C.G., Corrêa, R.L., Borges, A.C.N., Pereira de Sá, C.B., and Alves-Ferreira, M. (2005). Identification and characterization of homeobox genes in Eucalyptus. Genet. Mol. Biol. *28* (*3 suppl.*), 511–519 https://doi.org/10.1590/S1415-47572005000400005.

Romier, C., Cocchiarella, F., Mantovani, R., and Moras, D. (2003). The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. J. Biol. Chem. *278* (*2*), 1336–1345 https://doi.org/10.1074/jbc.M209635200. PubMed

Schatlowski, N., Creasey, K., Goodrich, J., and Schubert, D. (2008). Keeping plants in shape: polycomb-group genes and histone methylation. Semin. Cell Dev. Biol. *19* (*6*), 547–553 https://doi.org/10.1016/j.semcdb.2008.07.019. PubMed

Siefers, N., Dang, K.K., Kumimoto, R.W., Bynum, W.E., 4th, Tayrose, G., and Holt, B.F., 3rd. (2009). Tissue-specific expression patterns of Arabidopsis NF-Y transcription factors suggest potential for extensive combinatorial complexity. Plant Physiol. *149* (*2*), 625–641 https://doi.org/10.1104/pp.108.130591. PubMed

Sung, S., and Amasino, R.M. (2004). Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. Nature *427* (*6970*), 159–164 https://doi.org/10.1038/nature02195. PubMed

Takeuchi, T., Watanabe, Y., Takano-Shimizu, T., and Kondo, S. (2006). Roles of jumonji and jumonji family genes in chromatin regulation and development. Dev. Dyn. *235* (*9*), 2449–2459 https://doi.org/10.1002/dvdy.20851. PubMed

Todeschini, A.L., Georges, A., and Veitia, R.A. (2014). Transcription factors: specific DNA binding and specific gene regulation. Trends Genet. *30* (*6*), 211–219 https://doi.org/10.1016/j.tig.2014.04.002. PubMed

Toung, J.M., Morley, M., Li, M., and Cheung, V.G. (2011). RNA-sequence analysis of human B-cells. Genome Res. *21* (*6*), 991–998 https://doi.org/10.1101/gr.116335.110. PubMed

Ulf, S., Stalberg, K., Stymne, S., and Ronne, H. (2007). A family of eukaryotic lysophospholipid acetyl transferases with broad specificity. FEBS Lett. *2* (*23*), 305–309 https://doi.org/10.1016/j.febslet.2007.12.020.

Willmann, M.R., Mehalick, A.J., Packer, R.L., and Jenik, P.D. (2011). MicroRNAs regulate the timing of embryo maturation in *Arabidopsis.* Plant Physiol. *155* (*4*), 1871–1884 https://doi.org/10.1104/pp.110.171355. PubMed

Xu, W., Li, F., Ling, L., and Liu, A. (2013). Genome-wide survey and expression profiles of the AP2/ERF family in castor bean (*Ricinus communis* L.). BMC Genomics *14* (*1*), 785 https://doi.org/10.1186/1471-2164-14-785. PubMed

Xue, L.J., Zhang, J.J., and Xue, H.W. (2012). Genome-wide analysis of the complex transcriptional networks of rice developing seeds. PLoS ONE *7* (*2*), e31081 https://doi.org/10.1371/journal.pone.0031081. PubMed

Yamamoto, A., Kagaya, Y., Toyoshima, R., Kagaya, M., Takeda, S., and Hattori, T. (2009). Arabidopsis NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. Plant J. *58* (*5*), 843–856 https://doi.org/10.1111/j.1365-313X.2009.03817.x. PubMed

Yoshida, N., Yanai, Y., Chen, L., Kato, Y., Hiratsuka, J., Miwa, T., Sung, Z.R., and Takahashi, S. (2001). EMBRYONIC FLOWER2, a novel polycomb group protein homolog, mediates shoot development and flowering in *Arabidopsis.* Plant Cell *13* (*11*), 2471–2481 https://doi.org/10.1105/tpc.13.11.2471. PubMed

Zhao, Y., Cai, M., Zhang, X., Li, Y., Zhang, J., Zhao, H., Kong, F., Zheng, Y., and Qiu, F. (2014). Genome-wide identification, evolution and expression analysis of *mTERF* gene family in maize. PLoS ONE *9* (*4*), e94126 https://doi.org/10.1371/journal.pone.0094126. PubMed