Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

Genome-wide identification and *in silico* characterization of PEBP gene family in Avocado (*Persea americana*)

A thesis presented by

Sandra Elena Rivas Morales

Submitted to the

School of Engineering and Sciences

in partial fulfillment of the requirements for the degree of

Master of Science in Biotechnology

Monterrey, Nuevo León, June 2021

Dedication

This thesis is dedicated to my family, whose love accompanies me, teaches me, and takes care of me every day.

I also dedicate this work to my teachers, as their patience and willingness to share their knowledge nourished my passion for science.

Finally, but not less important, this work is dedicated to the memory of the people we lost in the fight against COVID 19.

Acknowledgements

In 2020 we faced the most challenging health crises in this century. Adapting to the new lifestyle was a stressful process for all people, including those employed in science, but at the same time, the need for scientific research was never more visible.

With this said, I would like to express my deep gratitude to my advisors, who guided me to develop a project that adapted to the new normality.

This research would not have been possible if Dr. Urrea had not offered me the opportunity to join a project of international extent. I also appreciate the sharp analysis of Dr. Rocío Díaz because her point of view always made me see a different perspective. To Dr. Alejandro Pereira, I must thank him for accompanying me on the path of bioinformatics, as it would have taken me so much time to learn on my own.

I cannot stop mentioning the help I received from Dr. Jorge Salazar in my training in the laboratory and while I was writing my thesis. Nor would I want to exclude the contribution of my colleagues in the research group, Lili, Sara, Dr. Perla Ramos, Alvaro, and Jessi, as their extensive experience and enthusiasm were enlightening.

Likewise, I thank the Instituto Tecnológico y de Estudios Superiores de Monterrey, for make me feel welcome and granting me an academic scholarship. In addition, it would not have been possible to carry out this academic degree without the financial support of CONACYT # 968967.

In addition to the academic support, I am indebted to my family and friends. Whether they advised me on technical issues or gave me emotional support, their attentions and amazing presences fed my spirit and provided me the strength to continue.

I could not imagine myself on this path without the influence of my parents. All those afternoons spent in the laboratory beside my mom and my dad's love for nature engraved in me the enthusiasm to answer questions and explore new horizons.

A significant part of the experience as a master's student was the friendships, I was able to build. The people I met two years ago, quickly went from classmates to a network of love and acceptance.

Finally, I am grateful to have people like my cousins Oziel y Edgar, who showed me infinite patience every time I needed guidance on technology matters. I also want to express my admiration for the commitment they put into everything they do, and I aspire to be as kind and dedicated as they are.

I consider myself fortunate to have brave, intelligent, and empathetic people as my companions, like my childhood friends Nelly and Vale, with whom I have shared dreams and chaotic energy. I can always expect the sincerest opinions from them. It is likewise a blessing to be friends with Rossana, an exceptional teacher, and an even better friend, and Eliud, the one that does not need so many words to understands so much.

Without all this people, the brightness of my days would be less.

Genome-wide identification and *in silico* characterization of PEBP gene family in Avocado (*Persea americana*)

By Sandra Elena Rivas Morales

Abstract

As key regulators of plant architecture, phosphatidylethanolamine-binding proteins (PEBPs) integrate internal and environmental stimulus. That makes them interesting targets for biotechnological applications, especially in plant breeding.

The six PEBP genes in *Arabidopsis thaliana* and a recently identified new member, gene AT5G01300, are considered as the model members, but there is a variable range of genes of the PEBP family in each species. Now at days, complete genome publications of non-model species like *Persea americana* have made it possible to explore the extension and function of this family.

In this study, we aim to perform a manual curation of the PEBP genes in a variety and a cultivar of *P. americana*, place those genes in a phylogenetic context by comparison with different clades of the vegetal kingdom, and infer functionality. To do that, we performed a BLAST+ search against three *P. Americana* genomes, we characterized them *in silico* and reconstruct the phylogenetic relationships with other 13 species PEBP sequences in RAxML.

We found 27 putative PEBP genes in *P. americana* classified in four sub-clades, FT-like, TFL-like, MFT-like, and a 4th sub-clade. The 4th sub-clade contains the most divergent sequences, including AT5G01300, with partial DPDxP and GHIR motifs, and was rarely found by other authors. Antecedents about the 4th clade show that the proteins from this clade are principally expressed in seed, bud, and flower, but functional characterization in Arabidopsis did not show any effect over flowering phenotype.

On the other hand, characterization *in silico* of predicted PEBP proteins from *P. americana,* revealed a specific pattern in 4th clade proteins, meanwhile, the cis-acting elements on the 2k bp up region, provided evidence of the functional

diversification between the 4th clade and the FT/TFL sub-clade and supported the closer relationship with the MFT-sub-clade found in the phylogeny.

Based on our results we propose an HMM-based methodology to identify proteins from the PEBP family in plants. We also recognized a different group of PEBP genes with an unknown function, present in almost all the species we selected. Finally, this information could bring insights into the evolution of the flowering process in perennial species, contribute to the understanding of the role of the PEBP family in *P. Americana* and add information to the description of a new clade of PEBP proteins.

identificación a nivel genómico y caracterización *in silico* de la familia de genes PEBP en aguacate (*Persea americana*)

Por Sandra Elena Rivas Morales

Resumen

Como reguladores clave de la arquitectura de la planta, las proteínas de unión a fosfatidiletanolamina (PEBP) integran estímulos internos y ambientales. Eso los convierte en objetivos interesantes para aplicaciones biotecnológicas, especialmente en el fitomejoramiento.

Los seis genes de PEBP en *Arabidopsis thaliana* y un nuevo miembro identificado recientemente, el gen AT5G01300, se consideran miembros del modelo, pero existe un rango variable de genes de la familia PEBP en cada especie. Hoy en día, las publicaciones completas del genoma de especies no modelo como Persea americana han hecho posible explorar la extensión y función de esta familia.

En este estudio, nuestro objetivo es realizar una curación manual de los genes PEBP en una variedad y un cultivar de *P. americana*, colocar esos genes en un contexto filogenético por comparación con diferentes clados del reino vegetal e inferir la funcionalidad. Para ello, realizamos una búsqueda BLAST+ contra tres genomas de *P. americana*, los caracterizamos *in silico* y reconstruimos las relaciones filogenéticas con secuencias de PEBP de otras 13 especies en RAxML.

Encontramos 27 genes de PEBP putativos en *P. americana* clasificados en cuatro sub-clados, FT-like, TFL-like, MFT-like y un cuarto sub-clado. El 4º sub-clado contiene las secuencias más divergentes, incluido AT5G01300, con motivos DPDxP y GHIR parciales, y rara vez se encontró reportado por otros autores. Los antecedentes sobre el 4º sub-clado muestran que las proteínas de este grupo se expresan principalmente en semillas, brotes y flores, pero la caracterización funcional en Arabidopsis no mostró ningún efecto sobre el fenotipo de floración.

Por otro lado, la caracterización *in silico* de las proteínas putativas PEBP de *P. americana*, reveló un patrón específico en las proteínas del 4o sub-clado, mientras que los elementos reguladores cis en 2k pb rio arriba, proporcionaron evidencia de la diversificación funcional entre el 4o sub-clado. y los sub-clados FT/TFL, y apoyó la relación más estrecha con el sub-clado MFT que se reportó en la filogenia.

Con base en nuestros resultados, proponemos una metodología basada en HMM para identificar proteínas de la familia PEBP en plantas. También reconocimos un sub-clado de genes PEBP altamente divergente, de función desconocida, presente en casi todas las especies que seleccionamos. Finalmente, esta información podría aportar nuevo conocimiento sobre la evolución del proceso de floración en especies perennes, contribuir a la comprensión del papel de la familia PEBP en *P. americana* y agregar información a la descripción de un nuevo sub-clado de proteínas PEBP.

List of figures

List of tables

Contents

Chapter 1. General Introduction

1.1 Introduction

The phosphatidylethanolamine-binding proteins (PEBPs) are a conserved group of proteins present in all taxa of bacteria, animals, and plants, involved in diverse signaling pathways (Chautard *et al.*, 2004).

In pants, the PEBP family is known for controlling plant architecture and transition from vegetative to reproductive phase (Karlgren *et al*., 2011). Until today, most of the information about this family has been described based on the six identified proteins of the model species *A. thaliana*: FLOWERING LOCUST (FT), TWIN SISTER OF FT (TSF), MOTHER OF FT (MFT), TERMINAL FLOWER 1(TFL1), BROTHER OF FT (BFT) and ARABIDOPSIS THALIANA CENTRORADIALIS HOMOLOGUES (ATC) (Ziv *et al*., 2014).

These six proteins are classified into three sub-clades according to their phylogenetic, relationships and functions (Carmona *et al*., 2007), the ancestral clade Mother of FT and TFL-like (MFT-like), represented by AtMFT; the flowering inhibitors TFL-like, which includes AtTFL1, AtBFT, and ATC (Gao *et al*., 2017); and the integrators of flowering stimuli FT-like that are AtFT and AtTSF (Li *et al*., 2016). Also, few recent reports have included a 4th clade with proteins with unknown function (Dong *et al*., 2020).

Findings of an early flowering phenotype by overexpression of FT-like genes or mutation in TFL-like genes (Pasriga *et al*., 2019) have raised interest in these clades, making them the best characterized among the PEBP family. However, evidence suggests that the number of PEBP genes may vary from species to species, and the duplication events could be related to adaptation (Kikuchi *et al*., 2009) transforming them into a biotechnological target and making it necessary to study the complete PEBP family of genes in each species (Cai *et al*., 2019; G. Wang *et al*., 2018).

In perennial species, long juvenile periods prevent the breeding of new varieties, even more, when crops face problems due to rapid climatic change and pathogens (Lavi *et al*., 1992). Avocado (*Persea americana*) is a perennial tropical fruit that has stirred interest in the international culinary market due to its high nutritional value, consistency, and unique flavor (Ortega Tovar, 2003). In 2019 SAGARPA estimated that gains from Mexican exportations of this product reached 3,201 million dollars, avocado crops struggle with environmental changes that affect their phenology and plagues (Álvarez Bravo *et al*., 2017).

It takes from three to fifteen years for an avocado tree to get to the reproductive phase, depending on the cultivar (Lahav & Lavi, 2009); that, is why gene segregation by selective crosses can take too long. In addition to this, there is limited information about the genetic background of flowering and maturation in this species. However, the recent publication of the complete genome of one variety and one cultivar of *P. americana* may have open the door for a deeper understanding of the genetic background of these processes, including poorly studied genes in avocado like PEBP family, setting up the ground for future biotechnological strategies.

There is only one reported PEBP protein for *P. americana,* which is an FT-like protein and has been functionally characterized (Ziv *et al*., 2014). There is also one report of a TFL-like member in the genre Persea (Gao *et al*., 2017). The identification of the rest of the members of the family may increase the understanding of the flowering process in this economically important crop tree and could open new biotechnological possibilities to control plant architecture.

1.2 Hypothesis

The genomes from *P. americana* var. *drymifolia* and cultivar Hass possess the same number of members from the phosphatidylethanolamine-binding proteins family, with conserved characteristics that allow us to infer their role in plant architecture.

1.3 Objectives

General objective:

Identify and characterize *in silico* possible members of the phosphatidylethanolamine-binding proteins (PEBP) gene family in drymifolia variety and a cultivar Hass of *P. americana*.

Specific objectives:

1. Identify and retrieve sequences of the PEBP family from two reported varieties of *P. americana* (cultivar Hass and var. d*rymifolia*), by homology search and confirmation with the PEBP domain.

2. Reconstruct a phylogenetic tree of the PEBP gene family of avocado by comparing it with 13 more species, to gain insights into the evolutionary history of this family.

3. Infer ortholog relationships of PEBP genes from *P. americana* and the 13 species to determine the gene duplications.

4. Describe the gene structure of the predicted members of the PEBP family in avocado variety *drymifolia* and cultivar Hass, including intron-exon organization and conserved motifs.

5. Identify regulatory cis-elements in the 2k bp upstream region of the predicted PEBP gene sequences and infer sub-clades petters of functionality.

6. *In silico* determination of the subcellular localization of predicted PEBP proteins and their characteristics.

1.4 Thesis structure

The present thesis is composed of 5 chapters: the first one present the general context of the topic addressed in this work, the hypothesis, and the objectives; chapter 2 reviews the available literature about the subject of flowering, PEBP family, avocado phenology, and biotechnological potential. The 3rd chapter describes the methodology selected for the experiments, and chapter 4 details the

results with their discussion. Finally, chapter 5 summarizes the general conclusions and suggests the following steps for PEBP proteins investigation in avocado.

Chapter 2. Literature review

2.1 Flowering molecular pathways

Flowering is a trait typical of the angiosperm; it is the process in which the meristem changes its vegetative growth to reproductive tissue, resulting in the development of sexual organs. The transition from vegetative to reproductive phases is a critical step for plant adaptation. Each species has evolved to synchronize their flowering time with the weather, available resources (Franks *et al*., 2007), or the pollinators seasoning (Fantinato *et al*., 2018).

Flowering can result from the stimulation of environmental factors as light quality, photoperiod, temperature (thermosensory flowering pathway), and endogenous signals like aging and hormones. Besides common flowering pathways like the ones mentioned before, sugar budget and ROS signaling also contribute to the determination of flowering time (Dally *et al*., 2015) (**Fig 1**).

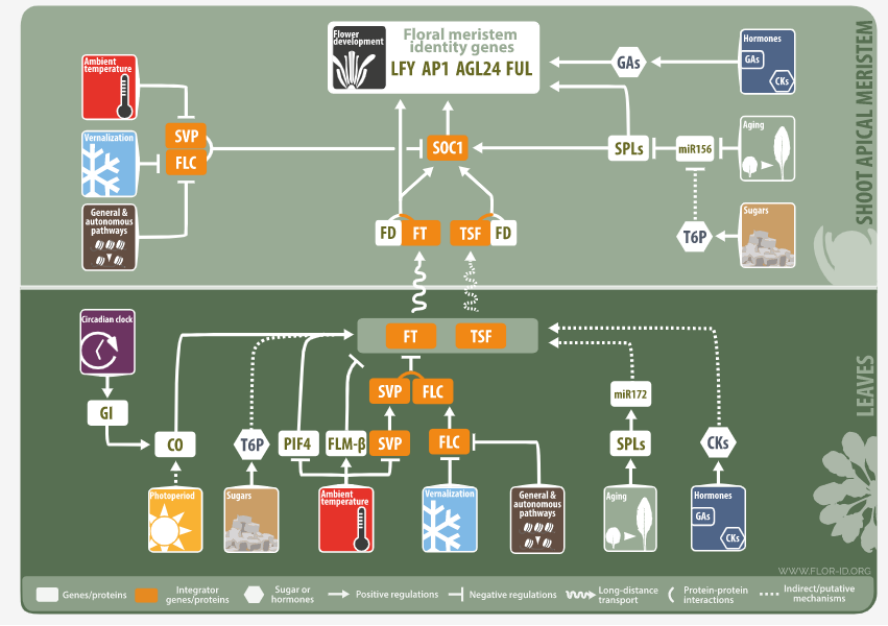


Figure 1. General scheme of molecular flowering pathways. Taken from the Flowering Interactive Database (Bouché et al., 2016).

Studies of the multiple molecular pathways for flowering have identified molecules that serve as integrators of several stimuli; those molecules are called florigens. Florigen integrators are conjunction of long-distance signaling molecules produced in leaves and transported to the shoot apex that regulate the transition from the vegetative state to the reproduction state of the meristem by activating cascade expression with SOC1 or directly activating the expression of the floral identity genes (Boss *et al*., 2004).

2.1.1 Light influencing floral initiation

Light is related to the flowering time either by the time of the exposure or its quality. Concerning the time of exposure, plants are classified as requiring Long- or short-day. It has been found that in long-day plants, blue light-dependent interaction between FMN (flavin mononucleotide), containing the photoreceptor FKF1, and the gigantean (GI) protein results in the transcription of CONSTANS (CO), a B-box zinc-finger transcription factor, reaching robust peaks in the afternoon (Baudry *et al.*, 2010; Sawa *et al*., 2007).

The progressive accumulation of CO activates the expression of FLOWERING LOCUS T (FT) (Wenkel *et al*. 2006; Kumimoto *et al*. 2008, 2010). That is possible because the NF-Y transcription factor and CO form a complex via its conserved CCT domain and bind to the CCAAT box sequence, around 5.3 kb upstream of the transcription start site of FT (Cao *et al*., 2014).

It is worth mentioning that the organs responsible for the perception of light are the leaves and due to that, the activation and accumulation of FT occurs in the companion cells of the phloem (Y. Yang et al., 2007). Then FT interacts with the endoplasmic reticulum protein called FT-INTERACTING PROTEIN 1 (FTIP1), which will mediate the movement of FT from companion cells. FT is small enough to move freely through plasmodesmata from one cell to another but they need the help of a chaperon called NaKR1 (described and characterized in *Arabidopsis*) to

continue its journey until it reaches the apical meristem (Zhu *et al*., 2016). Once it is there, it will bind with FD (also B-Zip) another transcription factor, in a complex with the protein 14-3-3, and together they will activate the genes that determinate floral identity, such as AEPTALA 1 (AP1), FRUITFUL, and SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1(SOC1) in meristem; FUL, SEPALLATA3 (SEP3) and SWEET10 in leaves (Teper-Bamnolker & Samach, 2005). Transcriptome studies have shown that the first transcript to be detected in meristems is the SOC1 mRNA, at the time it triggers the expression of numerous genes involved in the transition to floral meristem (Immink *et al*., 2012).

2.1.2 Temperature activation of flowering

Plants constantly sense environmental changes in light or temperature, the molecular mechanism by which temperature affects flowering time is called the thermosensory flowering pathway, and it still has a long journey of discovery (Blázquez *et al*., 2003). Research efforts have described two mechanisms of temperature response, the vernalization, resulted after prolonged treatment of cold temperature, and the growth ambient temperatures that can be defined as the physiological non-stressful temperature range of a given species, and its manipulation can stimulate flowering or even substitute the long-day conditions (Li *et al*., 2016).

Plant species with winter growth habit develop vernalization for protection of the floral meristems from frost damage (Hepworth *et al*., 2002) The main participant of the vernalization mechanism is the protein Flowering Locus C (FLC), this protein belongs to the MADS-box clade and it prevents the expression of FT and SOC1 by direct binding to the chromatin of those genes (Kim *et al*., 2009). Orthologous of FLC in perennial plants like citrus are differentially regulated, revealing their implication in the evolution of this group of plants (Zhang *et al*., 2009)

In Arabidopsis, the vernalization process consists of three stages: activation of vernalization, dynamic reprogramming of FLC during cold treatment, and epigenetic silencing of FLC by the trimethylation of Lys27 of histone H3 (Song *et al*., 2013). Previous to vernalization, FRIGIDA (FRI) keep FLC expressed by

forming the FRI complex and binding to the FLC locus as a transcriptional activator (Choi *et al*., 2011).

The ambient temperature pathway for flowering induction may reflect the adaptation of each plant to the native environment. In *A. thaliana*, for example, increase in the temperature from 23° to 27°C is enough to induce flowering under short-day conditions (Balasubramanian *et al*., 2006) meanwhile in avocado and other tropical and sub-tropical trees, flowering is induced by low ambient temperatures (Wilkie *et al*., 2008) and the temperature sterility threshold is usually upper to 30°C (Slot & Winter, 2016).

One of the principal components of the ambient temperature response is SHORT VEGETATIVE PHASE (SVP), an important MADS-domain transcription factor that represses FT and interacts with FLC (Jeong *et al*., 2007). It also promotes the early flowering myb protein (EFM) that is a relevant transcription factor with a role in directly repressing FT on the leaf vasculature (Yan *et al*., 2014).

Besides SVP, splicing of the Flowering locus M (FLM) plays an important role in the response to ambient temperature. At lower temperatures, expression levels of the form FLMβ increase, and then form a complex with SVP that binds to the upstream region of FT, TWIN SISTER OF FT (TSF), and SOC1 to repress flowering. When the temperatures increase SVP is degraded and the complex is dissolved, allowing the plant to flower (Hwan *et al*., 2014).

2.1.3 miRNA in regulation of flowering pathways

Recently, researchers have put more attention into the small non-coding sequences of RNA (miRNA), since they participate in cross talking between many flowering pathways (Kim & Sung, 2014). In age-related flowering, the major regulators and best-characterized microRNAs are MiR156 and miR172 (Spanudakis & Jackson, 2014). Their sequential action marks the transition between the juvenile and adult phase; miR156 is highly expressed during the juvenile period and decreases along with the plant age, while miR172 is more expressed in the adult phase of the plant (Wu *et al*., 2009).

Ahsan *et al*. (2019) found that MiR156 binds to SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) for degradation signaling, and that regulation is conserved in annual and perennial plants. From the three horticultural tree crops studied in the research of Ahsan and collaborators, it resulted interesting that only avocado showed abundance of miR172, repressor of APETALA2 (AP2)-like, related to age (Ahsan *et al*., 2019).

2.1.4 Gibberellins and hormone-related flowering

Among hormonal controls of flowering, the gibberellin (GA) molecular pathway has been studied extensively. GA promotes plant growth, seed germination, and accelerates flowering (Langridge, 1957) by the repression of RGL1 (Park *et al*., 2013), or the activation of LFY, an integrator of floral stimulus in short days (Blázquez *et al*., 1998). Besides, GA also regulates the expression of SOC1 and FT in long days to promote flowering (Hisamatsu & King, 2008).

2.2 Phosphatidylethanolamine-binding proteins (PEBP)

PEBP proteins are a highly conserved group of proteins between all taxa. They were identified for the first time in bovine brain cytosol and were linked to lipid metabolism (Bernier *et al*., 1986). Since then, Phosphatidylethanolamine-binding proteins have been described in many taxa, with a wide range of biological functions. In insects, for example, proteins from this family have been found to contribute to the immune system (Tang *et al*., 2019); in humans, PEBP is highly expressed in tumors and is related to cancer-activate signaling pathways (Luo *et al*., 2019).

Proteins from the PEBP family share a very similar domain structure but their main characteristic is the big central B-sheet and its anion binding pocket, which give them their name and can also bind to phosphate groups and phospholipids (Al-Mulla *et al*., 2013).

2.2.1 PEBP in plants, phylogeny, function, and conserved amino acids

As well as in the other taxa, PEBP functions in the plant have diverse roles, but in general terms, they are involved in the architecture, development, and reproductive initiation (Hiraoka *et al*., 2013). Researchers have put significant effort into the identification and characterization of PEBP members in each species because there is no fixed number of genes for all plants (**Table 1**). Additional to the variable number of genes, it has been found that members of the PEBP family with similar sequences may have a completely different effect on the plant (Mimida *et al*., 2001).

Until now, the PEBP gene family has been divided into three sub-clades according to their biochemical properties, functions, and sequence similitude. The sub-clades are FLOWERING LOCUS T-like (FT-like), TERMINAL FLOWER1-like (TFL1-like), and MOTHER OF FT AND TFL1- like (MFT-like) (Kobayashi & Weigel, 2007). Phylogenetic analysis has shown that before the emergence of the seed-producing plants, there were only MFT-like genes in mosses and liverworts (Hedman *et al*., 2009), then a duplication event gave origin to the other two subfamilies. That event may be related to the development of flowers in plants (Karlgren *et al*., 2011) (**Fig. 2**).

Table 1. Number of PEBP genes identified in plants.

Species	Common name	Number of genes identified	References
Pyrus bretschneideri	Pear	10	Zhao et al., 2020.
Pyrus communis		5	Zhao et al., 2020.
Pyrus betuleafolia		9	Zhao et al., 2020.
Triticum dicoccoides	Wheat	38	Dong et al., 2020.
Triticum urartu		16	Dong et al., 2020.
Aegilops tauschii		22	Dong et al., 2020.
Triticum aestivum L.		76	Dong et al., 2020.
Sorghum bicolor	Sorghum	19	Wolabu et al., 2016.

Gossypium hirsutum L.	Upland cotton	9	Zhang et al., 2016.
Phyllostachys heterocycla	Moso bamboo	5	Yang et al., 2019.
Zea mays	Maize	25	Danilevskaya et al., 2008.
Actinidia chinensis	Kiwifruit	13	Voogd et al., 2017.
Glycine max	Soybean	23	Wang et al., 2015.
Populus nigra var. italica Koehne	Lombardy poplar	9	Igasaki et al., 2008.

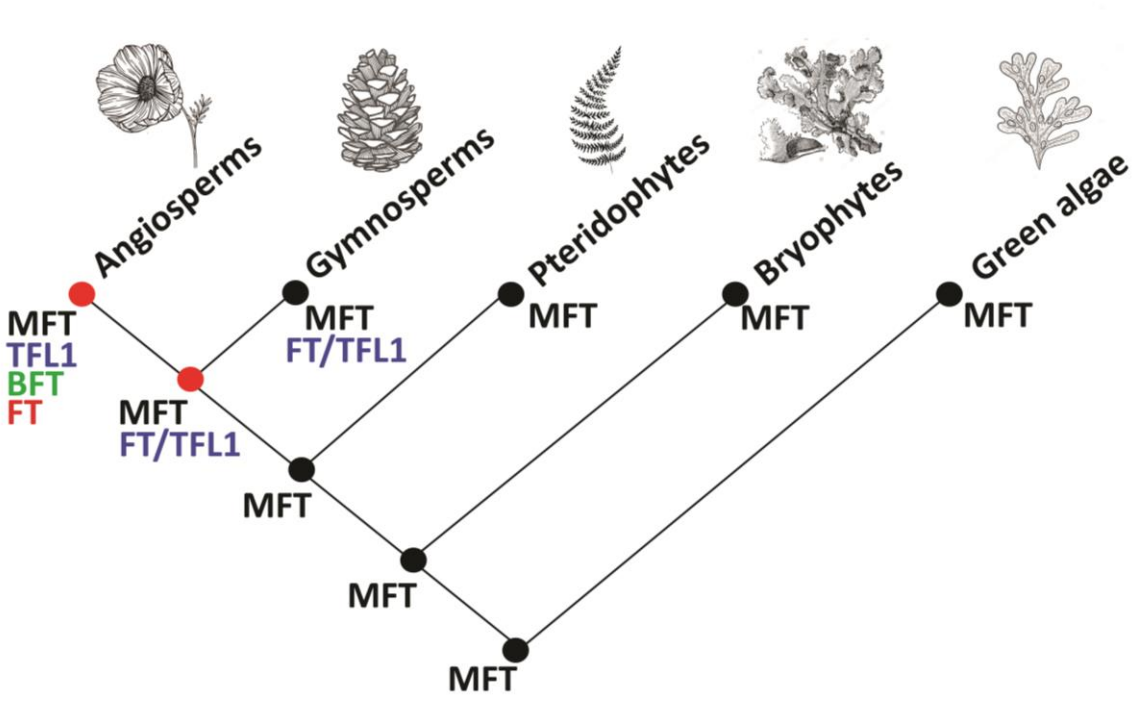


Figure 2. Representation of the evolutionary history of PEBP family and its sub-clades in Plants. Red circles represent the point of divergence of PEBP sub-clades. Modified from Pin & Nilsson, 2012.

Differences between the three sub-clades remain minimal, but there are some determinant changes in the amino acid sequences that can modify their functions (**Table 2**). For example, the Tyr85/His88 and Gln140/ Asp144, which are supposed to bind and form the anion binding pocked and are used to distinguish FT-like proteins from TFL-like (Zhen Wang *et al*., 2017).

Table 2. Critical amino acids for identification of plant PEBP.

Description	sequence	Reference
Residues which form a hydrogen bond in TFL1, but not FT, and which are likely the most critical residues for distinguishing FT and TFL1 activity	Tyr85/His88 and Gln140/ Asp144	(Mackenzie et al., 2019)
anion-binding sites	DPDXP (Asp-Pro-Asp-X-Pro) & GIHR (Gly-Ile-His-Arg)	(Mackenzie et al., 2019)
conserved segment B of FT proteins	positions 128–141, exon 4	(Ahn et al., 2006)
Amino acid sequences of PEBP proteins at the 14-3-3 interaction interfac	DLRxF	(Mackenzie et al., 2019)
Pro residue found in most MFT-genes	P237	(Hedman et al., 2009)
Diagnostic of true FT genes	LYN triad	(Ahn et al., 2006)

2.2.2 **FT-like sub-clade**

Members of FT-like sub-clade are widely recognized as florigen components (Shalit et al., 2009), which means that they regulate the transition from vegetative to reproductive phase, and also play an important role in the control of sugar storage (Andrés *et al*., 2020), fruit set, vegetative growth, stomatal control (Kinoshita *et al*. 2011) and tuberization (Navarro *et al*., 2011). They are also related to the efficient use of water (Robledo *et al*., 2019) and in some cases, the overexpression of FT has had adverse effects like dwarfism or infertile flowers (Lin *et al*., 2019).

The exact mechanism by which FT orthologs respond to different stimuli in each species may vary, but all the FT-like proteins are mobile signals that travel from the leaves to the shoot apical meristem (SAM), where it conforms to the flowering activation complex (FAC) (Wigge 2011).

In Arabidopsis, FT triggers reproductive transition under a long-day photoperiod due to the accumulation of CONSTANS (CO) protein (Baudry *et al*., 2010). CO forms a complex with the transcription factor NF-Y and then joins to the CCAAT box situated approximately 5.3 kb upstream of the transcription start site of FT (Cao *et al*., 2014). Once translated, the FT protein interacts with the FTIP1 (FT-

interacting protein 1) of the endoplasmic reticulum, which mediates its movement through phloem companion cells (Y. Yang *et al*., 2007).

FT protein is large enough to move freely through the plasmodesmata (20-25 kDa) and it has been reported to join the MaKR1 chaperon to continue its journey until the apical meristem (Zhu *et al*., 2016). In the SAM, 14-3-3 interacts with FT in the cytoplasm and then the complex is translocated to the nucleus, where it binds to the bZIP transcription factor FD, resulting in the formation of the FAC, which triggers the expression of the MADS-box genes and marks the irreversible transition to the reproductive meristem (Dally *et al*., 2015).

FAC structure is predicted to form a W-shaped protein, with two 14-3-3 proteins joined by their N-terminal regions and one FT-monomer attached to their c-terminal region. The inner corners of the “W” are positive pockets that interact with the phosphorylated C-terminal region of the FD proteins (Taoka *et al*., 2011).

FT and its homologous have similar tertiary structures to the PEBP found in animals, dominated by a large B sheet at the center and the anion binding pocket formed with the C-terminal peptide (Tang *et al*., 2019). Evidence until now has shown that increasing in phosphatidylcholine (PC): phosphatidylethanolamine (PE) ratio accelerates flowering, but *in vitro* studies have only shown FT binding to a (PC) (Nakamura *et al*., 2014).

2.2.3 TFL1-like sub-clade

TFL-like proteins share a highly similar structure with FT-like, there are only 39 non-conservative residues between TFL1 and FT in Arabidopsis (Ho & Weigel, 2014), but they carry out the opposite function, keeping an undifferentiated apical meristem by competing for the FD transcription factor (Hanzawa *et al*., 2005).

Ahn and collaborators found that an external loop structure (residues 128–145), together with the adjacent peptide segment, contributed to the opposite FT and TFL1 activities (Ahn *et al*., 2006), and Ho and Weigel (2014) found that specific mutations at the four Glu-109, Trp-138, Gln-140, and Asn-152 residues could transform FT into a TFL1-like floral repressor.

Their role in plant architecture is very important since it promotes vegetative growth (Asadi Khanouki *et al*., 2020) and in some cases, homologs from this sub-clade are involved in another reproduction process like tuberization (Shi *et al*., 2018).

2.2.4 MFT-like sub-clade

Proteins from the MFT-like sub-clade are also the only kind of PEBP's found in gymnosperms and non-vascular plants. They are related to gamete and sporophyte development in bryophytes, and in the angiosperms, their expression pattern is associated with seed germination and fruit maturation (Xi *et al*., 2010). Research on the early evolution of the PEBP family has shown that MFT-like protein has an ancestral character, and later duplication in this sub-clade raises the other two sub-clades (Hedman *et al*., 2009).

In *A. thaliana*, as well as some other plants, MFT-like genes are up-regulated by abscisic acid (ABA) and brassinosteroids (BR) (M. Wang *et al*., 2019) and repressed by Auxins, salicylic acid (SA), and methyl jasmonate (MeJa) (Chen *et al*., 2018).

2.3 Duplication, expression patterns and adaptation

Duplicated genes are the primary source for evolution, because of relaxation of selective pressure over the new copy allows functional diversification (Zhang, 2003). The PEBP genes have a notorious role in plant evolution, duplications in this family coincide with relevant evolutionary changes, like the emergence of FT/TFL sub-clade, and its major contribution to seed plant evolution (Karlgren *et al*., 2011).

Relevance of the PEBP family is even evident in short periods, an interesting example is *Ambrosia artemisiifolia*, an annual invasive eudicot. Experiments showed that the invasive population of this plant displayed an early flowering phenotype compared with the native population under non-inductive long-day photoperiods, associated with the expression of FT and TFL homologs. Also, analysis of the flowering time of a hybrid of the two populations demonstrated a dominant inherited early flowering phenotype (Kralemann *et al*., 2018).

Another example of the importance of the PEBP family in adaptation is in the domesticated crops like it was described in soybean (Thakare *et al*., 2011) and sunflower (Blackman *et al*., 2010).

Several duplications in PEBP genes in recent lineages have been identified with different spatial-temporal expression patterns that are taken as evidence of different functions (Komiya *et al*., 2008; Hagiwara *et al*., 2009). In sugar beet (*Beta vulgaris*) a pair of paralogs PEBP proteins, phylogenetically identified as FT-like that even contain Tyr-85 and Gln-140 conserved residues, were found with antagonistic functions, whereas BvFT2 is essential for flowering, BvFT1 acts as a flowering repressor (Pin *et al*., 2010).

2.4 Avocado, taxonomy, and phenology

The fruit of *P. americana* Mill is a berry with fleshy mesocarp and endocarp, of a single dicotyledonous seed. It has an ovoid and globular shape, and its size is very variable, with a rough and firm or soft and smooth shell. Depending on the variety, its color turns dark green or black when it reaches maturity.

The tree that produces it can reach 20 meters, but its average size is 12 m and it takes approximately 3 to 15 years (Lahav & Lavi, 2009). It has abundant branches with high sensitivity to sunburn and low temperatures, strong wind currents, or excess production (Godínez *et al*., 2000).

Genus Persea belongs to the Lauraceae family and is made up of two subgenera: Eriodaphne and Persea. This genus is distinguished by having pubescences on both sides of the sepals (KOPP, 1966).

The distribution of this genus is concentrated in the central region of Mexico to the Caribbean, passing through Peru, Bolivia, Ecuador, and Brazil. However, it is in Mexico and Brazil where the greatest number of species of this genus are found, greatly compromising them to the conservation of the avocado's genetic diversity.

In Mexico, it is possible to identify several horticultural varieties, grouped within *P. americana* Mill species. The great variety of ecological conditions and natural selection has produced races adapted to multiple habitats (de la Luz Sánchez-

Pérez, 1999). Although artificial selection has contributed greatly since its domestication in pre-Columbian times (Turner and Miksiek, 1984). Thanks to this, we find an abundant genetic wealth in the country, which in turn represents a wealth of resources available for its characterization and use.

The patented avocado variety “Méndez No.1” (patent ID USPP11173P), is cultivated in a large area of the state of Jalisco and Michoacán, and its relevance lies in the harvest of its fruit in summer, when avocado prices reach their maximum peak, unlike the “Hass” cultivar which ripens its fruit in winter (Cossio-Vargas *et al*., 2008). In addition, the Méndez variety, in contrast to the “Hass” cultivar, presents a faster floral development (187 days in summer, 222 days in winter), and its winter vegetative flow can produce flowers in summer (Salazar-García *et al*., 2018), this and other characteristics are listed in the following table.

Table 3. Phenology comparison between Méndez and Hass varieties of *P. americana*

Variety	**Méndez**	**Hass**
Harvest	Summer	Winter
Vegetative growth flow	Summer: August –February. winter: February	Higher vegetative growth in winter.
Flowering	September y February	Produced by a single vegetative growth in winter.
Floral development	Summer: 6,23 m (187 d) Winter:7,4 m (222 d)	Sumer: 7,5 Winter:11,5
Fruit maturation	September/October & July	8 months (March-November)
Root growth flow	2 growth flows, the biggest in summer (June – August)	Normally two growth Flow: spring and summer.
Fruit fall	March and July	June

2.5 Importance of PEBP identification in avocado.

Plant reproduction has been a key feature to the fitness of the species. The genes involved in the phenology have always represented an interesting point for conservation and overcoming reproduction problems provoked by the fast climate change (Tzfira *et al*., 1998).

In avocado, phenologic behavior is dependent on the climatic conditions, although its management and genetic background are also important factors, the flowering time and fruit set may be delayed, overlapped, or shortened primarily due to the temperature and water availability (Bartoli, 2013).

Technological advances have made the study of the genetic basis of plant reproduction attractive to improve the commercial characteristics of crops. It has directed the researchers to explore biotechnological solutions that include manipulation of molecular flowering pathways (Park *et al*., 2014).

This approach is currently being evaluated by several investigators and each one has developed its methodology, but a central focus of those strategies is the overexpression, characterization, and monitoring of genes called “integrators” of the floral pathways. Some examples of these integrators are FLOWERING LOCUS T (FT), TWIN SISTER OF FT (TSF), SUPPRESSOR OF OVEREXRESSION OF CO 1 (SOC1), and LEAFY (LFY) (Wigge *et al*., 2005). They have the characteristic of being activated by more than one stimulus, but it does not mean that they are not tissue or stimuli specific. Some experiments have successfully attempted to induce flowering by biotechnological intervention like (Pasriga *et al*., 2019) that obtained early flowering phenotype in rice by overexpression of RICE FLOWERING LOCUS T1 (RFT1).

Since it was described as a floral response integrator, FT has been a popular target for rapid cycle breeding in tropical and perennial trees. During the characterization of FT in *A. thaliana* and other species, researchers developed a transgenic and stable overexpression of the gene by Agrobacterium-mediated transformation.

There are several considerations to define a methodology to obtain a plant overexpressing FT, starting with the selection of the gene. Most of the research started with the known sequence of *A. thaliana* FT, characterizing the flowering gene of the plant of interest in *Arabidopsis*, or trying to induce flowering in the plants of interest overexpressing the genes of *Arabidopsis*. In any case, the homology of the sequence could be insufficient to produce the expected results, so a deep bioinformatics analysis is necessary. In this work, we will make special emphasis on the PEBP family genes of avocado as relevant targets for biotechnological improvement of plant architecture.

Chapter 3. Materials and Methods

3.1 Sequences identification and phylogenetic reconstruction

3.1.1 Genome databases and identification of PEBP homologs

The genome sequences of one variety and one cultivar of *P. americana* were obtained from the NCBI database, two of them were from the cultivar Hass (GCA_008087245.1 and GCA_002908915.1) and one corresponding to variety d*rymifolia* (GCA_008033785.1) (Rendón-Anaya *et al*., 2019). It´s worth mentioning that these three genomes lack annotation files, and the genome GCA_002908915.1 was sequenced in Hainan University, from leaf of a mature tree (project accession PRJNA412302). Additional to Persea genomes, we selected and downloaded 13 plant proteomes from the Phytozome v12 database (Goodstein *et al*., 2012) to complement the phylogenetic analysis, based on their reports of functionally characterized proteins from the PEBP family, its role as model plants, and its economic importance (**Table 4**).

Table 4. Group, order, and species number of the selected organisms for the phylogenetic analysis.

Group	Order	Species	Source
Moss	Funariales	*Physcomitrum patens*	NCBI
Basal angiosperms	Amborellales	*Amborella trichopoda*	Phytozome
	Laurales	*P. americana* cultivar Hass (Hainan)	NCBI
		P. americana cultivar Hass (CINVESTAV)	NCBI
		P. americana var. *drymifolia*	NCBI
	Brassicales	*Arabidopsis thaliana*	Phytozome
Eudicots	Malvales	*Theobroma cacao*	Phytozome
	Sapindales	*Citrus sinensis*	Phytozome
	Malpighiales	*Populus trichocarpa*	Phytozome
	Fabales	*Medicago truncatula*	Phytozome
	Rosales	*Malus domestica*	Phytozome
	solanales	*Solanum lycopersicum*	Phytozome
	Vitales	*Vitis vinifera*	Phytozome
	Caryophylalles	*Beta vulgaris*	Phytozome
Monocots	Poales	*Oryza sativa*	Phytozome
		Sorghum bicolor	Phytozome

To identify sequences from the PEBP family on the *P. americana* genomes, we used the six identified protein sequences from *A. thaliana* as query on a BLAST search against the tree genomes (**Table 5**).

Table 5. Accession number of the genes used as queries for the BLAST searches.

Gene	Accesion
AtFT	**BAA77838.1**
AtTFL1	**NM_120465.3**
AtTSF	**NM_118156.2**
AtBFT	**NM_125597.2**
AtMFT	**NM_101672.4**
ATC	**NM_128315.4**
PaFT	**KM023154.1**
PaTFL	**KY933634.1**

All the analyses were carried out in the UBUNTU GNU/Linux operating system (https://ubuntu.com/). For the identification of PEBP proteins in the avocado genomes, we use the tblastn program with the following parameters: -max_target_seqs 50 -outfmt '6 qseqid sseqid qlen slen qstart qend sstart send length pident positive mismatch gaps evalue bitscore qcovs stitle' -num_threads 1. The source code of the BLAST+ v2.11.0 program was obtained from the NCBI repository (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/).

Contigs matching against PEBP sequences were retrieved using the EMBOSS tools (Rice & Bleasby, 2000), and analyzed with the FGENESH+ gene prediction program (Solovyev, 2004) to obtain the complete CDS and amino acid sequences; for these analyses, the *A. thaliana* PEBP proteins and specific gene-finding parameters were used. We used the HMM model domain for the PEBP family (PF01161), downloaded from Pfam (v.2.6) to identify using HMMER v3.3.2 the presence of the PEBP conserved domain across the predicted sequences. Finally, the sequences were aligned to manually curate them.

3.1.2 Phylogenetic Analysis

To reconstruct the phylogenetic relationships among PEBP proteins from plants, we performed HMM searches in the proteomes from the 13 selected species, using the PEBP Pfam model and the HMMER program. Sequences were retrieved by species in individual fasta files for a latter multiple sequence alignment in MEGA 10.1.8 using MUSCLE, and then manually curated. We also checked the chromosomal, scaffold, or contig localization, and sequences with short alignments with the PEBP Pfam model, isoforms of trunk proteins were excluded from further analysis.

After the filtration, we used MEGA 10.1.8 to calculate the best empirical evolutionary model using the Maximum Likelihood statistical method and considering all sites as useful data. The model with the lowest Bayesian Information Criterion (BIC) was considered as the best model to describe the substitution pattern. The next step was aligning the sequences with HMMER

(v3.3.2), for that we worked with two approaches, as to align command “hmmalign” only use the section of the proteins that encompass the Pfam model, so in one strategy we used complete sequences and then unaligned regions were aligned with clustal W; and the second one, using the option “trimm”, which will cut the unaligned sections.

Phylogenetic trees were constructed in MEGA and RAxML v 8.2.12 (Stamatakis, 2014) using the Maximum Likelihood algorithm, the Jones-Taylor-Thornton (JTT) model, and non-uniformity of evolution rates described by Gamma distribution. Automatic bootstrap computation was performed in RAxML by the command –autoMRE, also parsimony random seed and the following parameters -f a -x 12345 -p 12345. For MEGA analysis, 500 iterations were used for bootstrap. The summarized methodology is showed in **Fig. 3**.

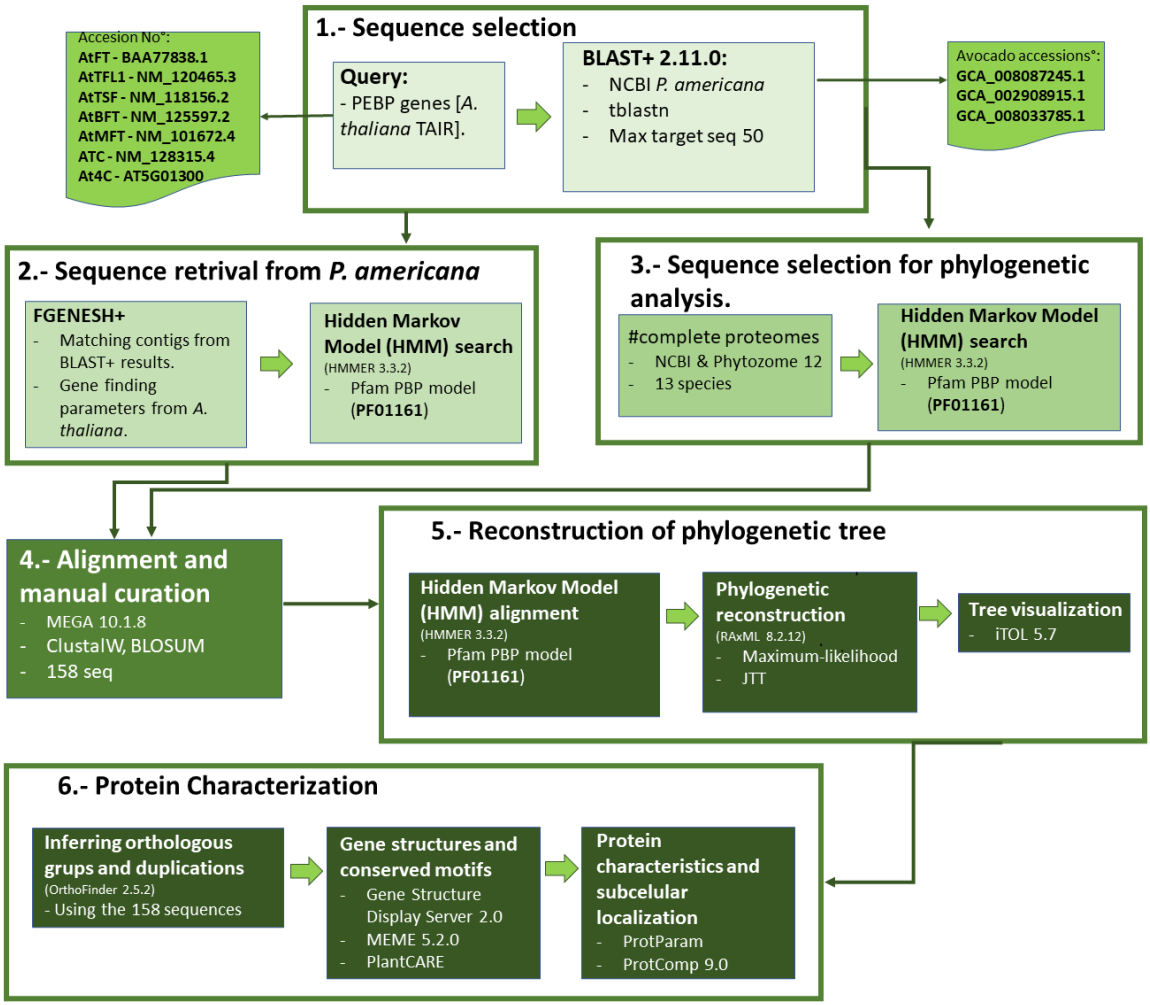


Figure 3. Graphical summary of methodology. 1) *A. thaliana* PEBP sequences were used as query for BLAST+ search against *P. americana* genomes. 2) Predicted PEBP sequences were retrieved from *P. americana* genomes and presence of PEBP domain was corroborated. 3) Proteome search for PEBP sequences in 13 species using the PEBP domain model. 4) Filter of isoforms and trunk proteins. 5) Analysis of all the sequences for the phylogenetic reconstruction and tree visualization. 6) Characterization of avocado PEBP predicted proteins and gene structure analysis.

3.2 Inferring orthologous groups and duplications

Sequences from all the taxa included in the phylogenetic analysis were put in one fasta file for each species. The 13 files were submitted to OrthoFinder (Emms & Kelly, 2019) without alignment to perform a general automatic examination, which includes inference of the orthogroups of the species (OG's) and gene duplication events that are referenced in a calculated gene tree. We used DIAMOND as a search algorithm (Buchfink *et al*., 2015), instead of BLAST (the default option)

because of its increased speed in short sequences and similar degree of sensitivity.

3.3 Gene structures and conserved motifs

3.3.1 Intron-exon organization

CDS sequences of the putative PEBP genes were obtained from the FGENESH+ analysis, and then they were used in a BLAST search for the complete nucleotide sequence in the NCBI platform. Two files with the CDS and the gene sequences of the three *P. americana* varieties were depicted with the program Gene Structure Display Server 2.0 (http://gsds.cbi.pku.edu.cn).

3.3.2 Conserved motifs

Conserved motifs in proteins from each sub-clade on the phylogenetic tree were identified by means of Multiple Expectation Maximization for Motif Elicitation software (MEME 5.2.0) (Bailey & Elkan, 1994) using the optimized parameters of minimum motif length of 6, maximum of 30, maximum motif output set at 10, and one occurrence per sequence. Only motifs with an E-value lower than 1e-4 will be retained (M. Wang *et al*., 2019).

3.3.3 Cis-acting elements

Genomic sequences from 2000 bp upstream at the transcription start site of the *P. americana* sequences were obtained and submitted to PlantCARE database (Zhao *et al*., 2020). The platform gives the type and frequency of the cis-elements for each sequence, as well as the enrichment. Results were inspected manually and organized in an excel sheet.

3.4 Protein characterization and subcellular localization

Reliable sequences from *P. americana* were submitted to ExPASy ProtParam (Gasteiger *et al*., 2005) to predict the isoelectric point (pI), molecular weight (MW),

and instability index. For the subcellular localization estimation, we will use the package ProtComp 9.0.

Chapter 4. Results and discussion

4.1 **Identification of PEBP homologs and phylogenetic analysis**

*High quality alignments and tree figures are available in the following link: https://drive.google.com/drive/folders/1GrIV7bnRduusZ8htOOpYXH2As-qWEhe5?usp=sharing *

Using the six PEBP's of *A. thaliana* as queries, Blastn searches found 20 and 21 matching sequences against the contigs of Hass genomes from Hainan and CINVESTAV respectively, but only in 18 contigs for the *drymifolia* variety. All contigs with matching sequences were extracted in individual Fasta files and then processed with FGENESH+ (Solovyev, 2007). Results from FGENESH+ predicted a single PEBP sequence in each contig, although in some of the contigs no sequences were found. A total of 20 sequences from the CINVESTAV Hass genome, 16 from the Hainan genome, and 14 from *drymifolia* variety were used in the next analysis (**Table 6**).

Table 6. *P. americana* accession numbers and numbers of identified PEBP sequences identified.

Genome	Accession	# of retrieved seq (blast+)	# of predicted seq (FGENESH+)	# of filtered seq (HMMER)	TFL-like	FT-like	MFT-like	4th clade *
Hass CINVESTAV	GCA_008087245.1	21	20	11	5	3	3	1
Hass Hainan	GCA_002908915.1	20	16	6	4	1	1	1
Drymifolia CINVESTAV	GCA_008033785.1	18	14	7	2	2	3	1

*** Sequences retrieved after the preliminary phylogenetic analysis with MEGA 10.1.8, see at section 4.1.2.**

Amino acid sequences predicted from FGENESH+ were analyzed with HMMER 3.3.2 using the Pfam model PF01161 to confirm the presence of the conserved domains. Only 11, 6, and 7 sequences were confirmed to follow the PEBP model; their provisional names and size are shown in **Table 7**. All the amino acid sequences found in the Hass Hainan genome were found 100% identical to the sequences from drymifolia variety, and only two proteins were found exclusively in the Hass CINVESTAV genome.

Table 7. Identities and characteristics of putative genes of PEBP family in *P. americana*.

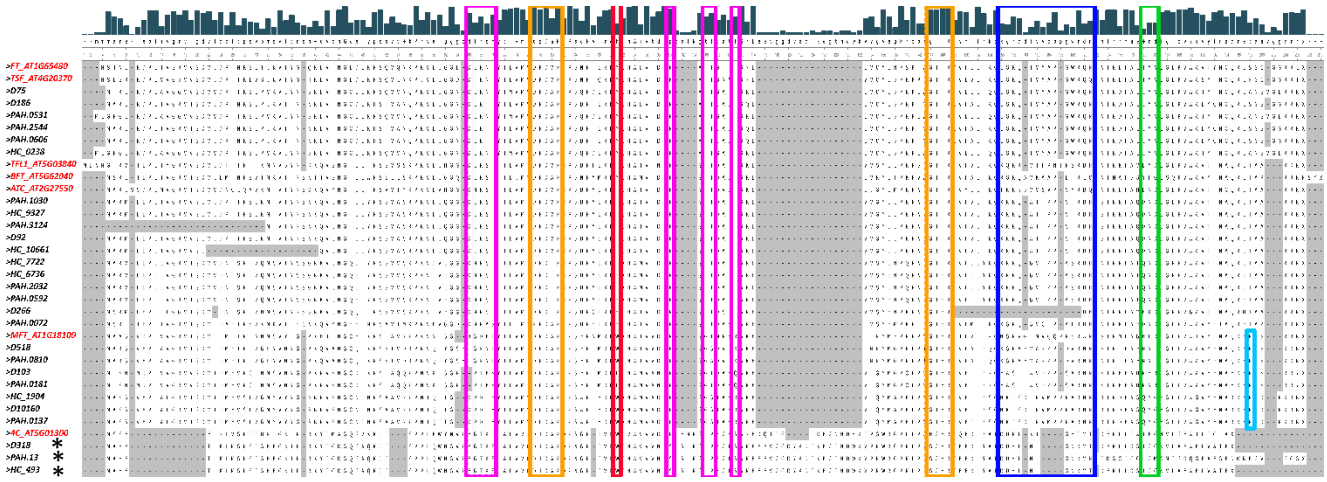
Genome	Sequence ID for this study	Contig/ scaffold id	Strand	Exons	CDS	Amino acid	Sub-clade
Hass HAINAN	HC_10661	tig00010661	+	5	465	154	TFL-like
	HC_9327	tig0009327	-	4	522	173	TFL-like
	HC_7722	tig0007722	-	4	522	173	TFL-like
	HC_6736	tig0006736	+	4	522	173	TFL-like
	HC_0238	tig0000238	+	5	534	177	FT-like
	HC_1904	tig0001904	-	4	525	174	MFT-like
	HC_493*	tig0000493	+	2	501	166	4th
Hass CINVESTAV	PAH.0072	Ctg0072	-	4	522	173	TFL-like
	PAH.0592	Ctg0592	+	4	522	173	TFL-like
	PAH.1030	Ctg1030	-	4	522	173	TFL-like
	PAH.2032	Ctg2032	+	4	522	173	TFL-like
	PAH.3124	Ctg3124	+	4	444	147	TFL-like
	PAH.0606	Ctg0606	-	4	525	174	FT-like
	PAH.2544	Ctg2544	-	4	525	174	FT-like
	PAH.0531	Ctg0531	-	5	534	177	FT-like
	PAH.0810	Ctg0810	+	4	525	174	MFT-like
	PAH.0137	Ctg0137	+	4	525	174	MFT-like
	PAH.0181	Ctg0181	+	4	519	172	MFT-like
	PAH.13*	Ctg0013	+	2	531	176	4th
***DRYMIFOLIA* CINVESTAV**	D266	Scf00266	-	5	459	152	TFL-like
	D186	Scf00186	-	4	525	174	FT-like
	D75	Scf00075	-	4	528	175	FT-like
	D103	Scf00103	+	4	519	172	MFT-like
	D518	Scf00518	+	4	525	174	MFT-

							like
	D10160	Scf10160	-	4	525	174	MFT-like
	D92	Scf00092	+	4	522	173	TFL-like
	D318*	Scf00318	-	2	501	166	4th

***Sequences retrieved after the preeliminar phylogenetic analysis with MEGA 10.1.8, see at section 4.1.2.**

Amino acid sequences obtained from the HMMER filtration were aligned and manually inspected to find the key amino acids previously reported in the literature for plant PEBP. Most of the sequences found in the three genomes had the highly conserved anion binding motif DPDxP (Asp-Pro-Asp-X-Pro) & GIHR (Gly-Ile-His-Arg), except for the sequence PAH.0072, which present a substitution in the Asp77 for an Asn77 (**Fig. 4**). Conservation of Tyr85/His88 and Gln140/ Asp144 was found in 17 sequences. Those residues are predicted to form a hydrogen bond network (Nakamura *et al*., 2019) and have an important role in determining flowering time.

LYN triad is a characteristic motif on the c segment of the fourth exon that distinguishes FT-like proteins, and it was found in six of the sequences which also have Tyr85 residues (**Fig. 4**). Also, Pro residues at the C-terminus identified in 7 sequences of the total number of sequences predicted from the three genomes, which is a particular feature of the MFT-like proteins (Hedman *et al*., 2009).



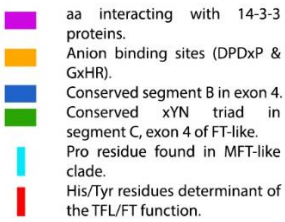


Figure 4. Alignment of the predicted PEBP proteins identified from the genome wide search analyses. Key amino acids are highlighted with colored boxes under the alignment, and *A. thaliana* ID sequences are tagged in red. Color code of the boxes are detailed below the alignment. *Sequences retrieved after the preliminary phylogenetic analysis with MEGA 10.1.8, see at section 4.1.2

4.1.2 Phylogenetic reconstruction

Complete proteomes from 13 plant species were downloaded from Phytozome and analyzed with HMMER software to identify PEBP homologs proteins. A total of 185 annotated proteins suited the Pfam model (PF01161). The sequences were retrieved, aligned, and manually filter. Sequences corresponding to isoforms, with incomplete domains, or truncated proteins, were excluded from further analysis. The Bit-score was also a filtration parameter. Finally, 131 proteins plus the avocado sequences, were selected to perform the phylogenetic reconstruction (**Table 8**).

Table 8. Total sequences found in the proteome-wide search.

Group	Order	Specie	# of total seq	# of final seq	TFL-like	FT-like	MFT-like	4th sub-clade
Moss	Funariales	*Physcomitrella patens*	25	9	0	0	7	2
Basal angiosperm	Amborellales	*Amborella trichopoda*	8	4	1	1	2	
eudicots	Brassicales	*Arabidopsis thaliana*	8	7	3	2	1	1
	Malvales	*Theobroma cacao*	8	7	3	1	2	1
	Sapindales	*Citrus sinensis*	9	8	3	3	1	1
	Malpighiales	*Populus trichocarpa*	10	7	3	2	1	1
	Fabales	*Medicago truncatula*	20	13	4	6	1	2
	Rosales	*Malus domestica*	16	9	6	1	1	1

	Solanales	*Solanum lycopersicum*	17	12	5	4	2	1
	Vitales	*Vitis vinifera*	7	6	3	0	2	1
	Caryophylalles	*Beta vulgaris*	9	9	3	3	2	1
monocots	Poales	*Oryza sativa*	22	20	4	13	2	1
		Sorghum bicolor	26	20	4	13	2	1
		Total	185	131	42	49	26	14

We performed two alignments with HMMER, the first option using the command "hmmalign" without option "trim", and the N and C terminus aligned with clustal W, showed large gaps generated by a few sequences; the second alignment, performed with the command "hmmalign" but including the option –trim to cut the unaligned fragments was more compact and resulted in smaller gap extension (**Fig. 5**). Using MEGA 10.1.8 we searched for the best fitting substitution model for the sequences, including the *P. americana* putative proteins. The Jones-Taylor-Thornton model (JTT) (Jones *et al*., 1992) with non-uniformity of evolutionary rates modeled by Gamma distribution (+G) resulted as the model that better describes the substitution pattern, according to their Bayesian Information Criterion (BIC; the lowest the best).

With model information, we calculated a preliminary tree on MEGA using Maximum likelihood (**Fig. 6**), which showed the presence of a 4th clade, different from the three clades, however, there were no avocado or *Amborella* sequences in that clade. Regarding Avocado sequences, we did not look for these sequences at the beginning of the analysis. Using the sequences from the 4th clade of all species, we performed BLAST+ searches against *P. americana* genomes and following the same filter process as in the previous search. Three more sequences were found, one for each genome (marked with an asterisk in **Table 7** and **Fig. 4).**

Figure 5. Multiple sequence alignment using HMMER program and the Pfam model PF01161. A total of 158 filter PEBP sequences are grouped by sub-clade and their initial and terminal regions were cut by the option –trim. Colored bars next to sequence names represent the different sub-clades established in Figure 6.

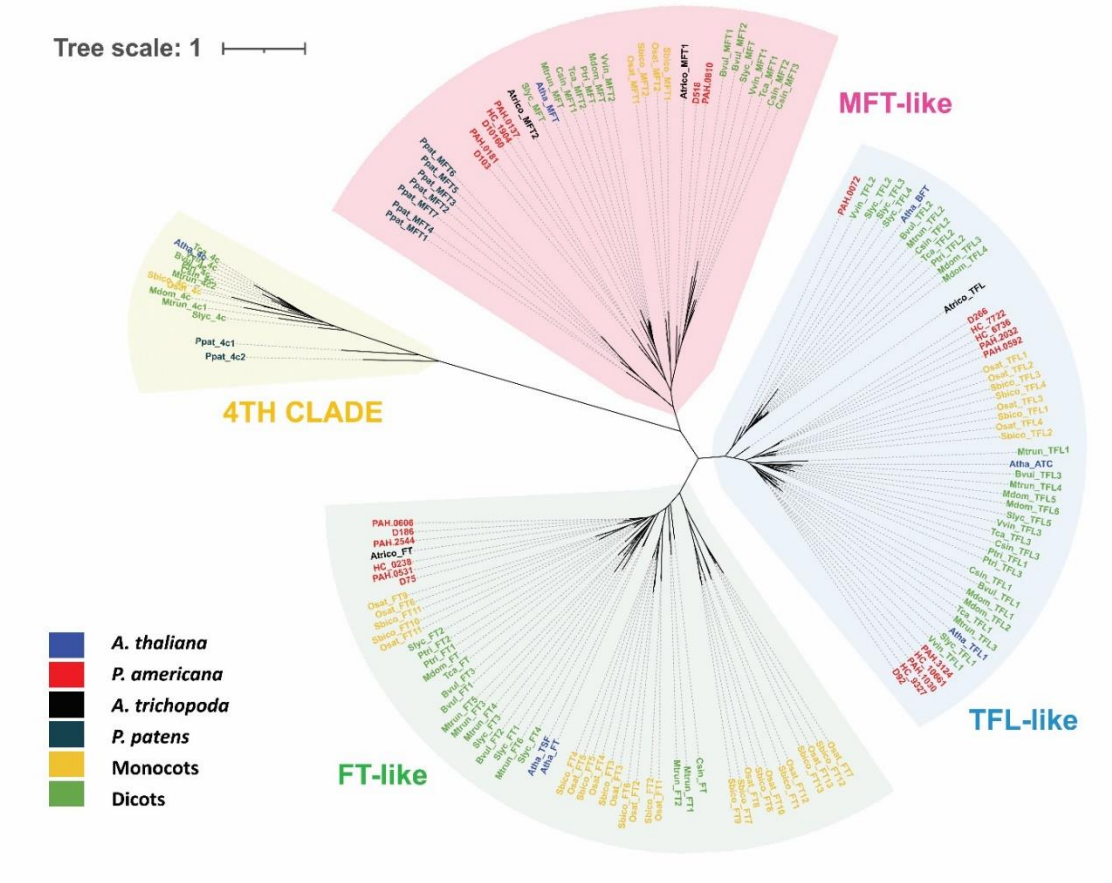


Figure 6. Preliminary phylogenetic reconstruction of PEBP family in plants. The tree was generated in MEGA using Maximum Likelihood.

Phylogenetic reconstruction was performed with the two alignments in RAxML 8.2.12 with automatic bootstrap calculation, parsimony random seed, and the model PROTGAMMAJTT: raxmlHPC -f a -x 12345 -p 12345 -# autoMRE -m PROTGAMMAJTT. The trees showed in this document are the result of the trimmed alignment from HMMER (**Fig. 7** and **8**).

Backing up the results of the MEGA tree, both RAxML trees confirmed the existence of four major sub-clades: FT-like, TFL-like, MFT-like, and a 4th clade that has been identified as PEBP-like by (Dong *et al*., 2020). The 4th clade or PEBP-like is immersed in the MFT-like clade, this close relation may imply an ancestral character of the sequences, besides that, at least one sequence of this clade was found in each species, with exception of *A. trichopoda*, where we only found similar truncated proteins (data not shown).

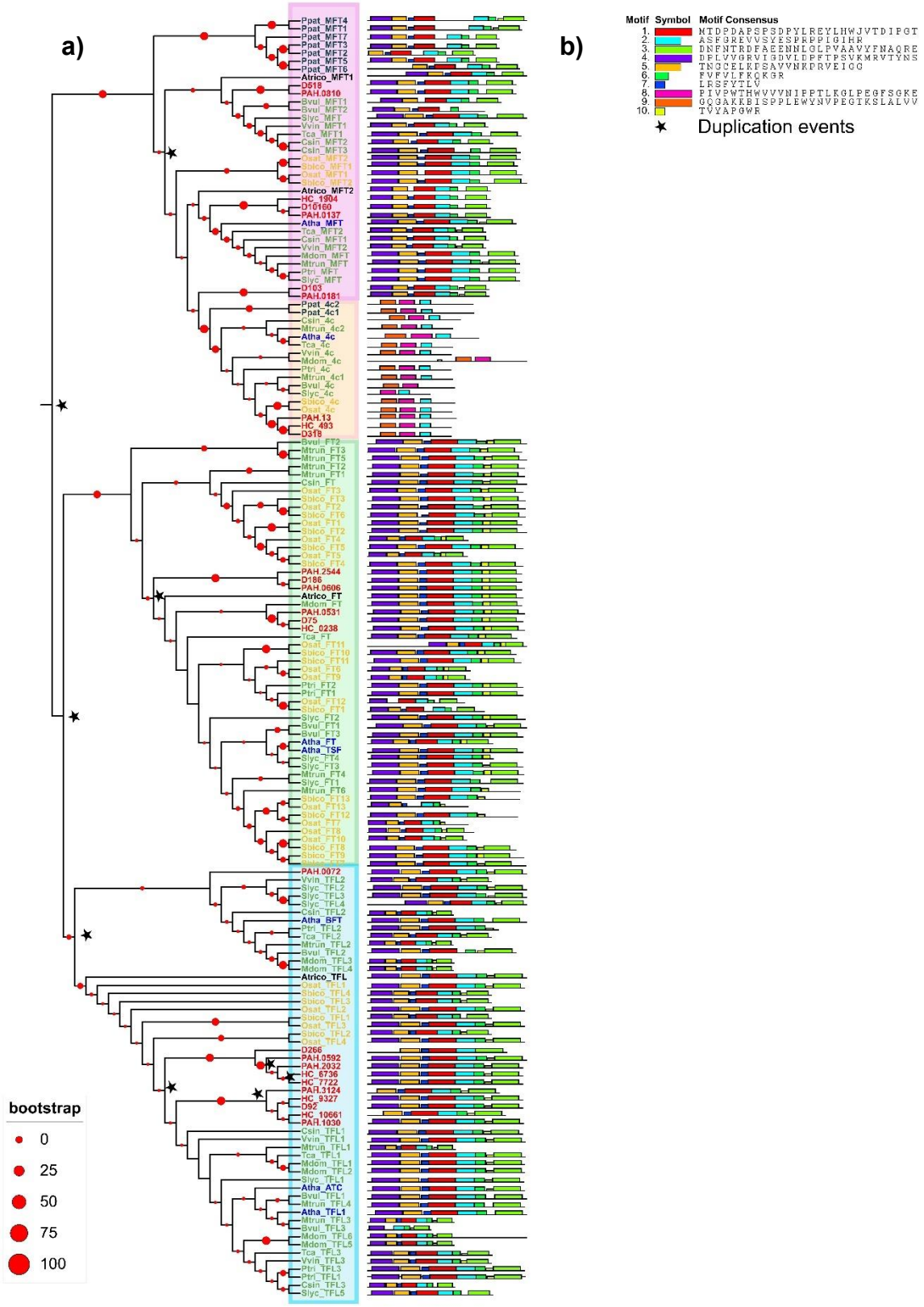

a)
b)
Motif Symbol Motif Consensus
1. MTDPDAPSPSDPYLREYLHWJVTDIPGT
2. ASFGREVVSVESPRPPIGIHR
3. DNFNTRDFAEENNLGLPVAAVVFNAQRE
4. DFLVVGRVIGDVLDPFTPSVKMRVTYNS
5. TNGCELKPSAVVNKDRVEIGG
6. FVFVLFKQKGR
7. LRSFYTLV
8. PIVPWTHWVVVNIPPTLKGLPEGFSGKE
9. GQGAKKBISPPLEWYNVPEGTKSLALVV
10. TVYAPGWR
Duplication events
bootstrap
0
25
50
75
100

Figure 7. a) Phylogenetic reconstruction of PEBP family in plants. Bootstrap is represented by red circles at the middle of the branches. *P. americana* sequences are highlighted in red, and *A. thaliana* in blue. b) Distribution of conserved motifs from each sub-clade, obtained with MEME. Colors representing each motif are described in the table.

Large divergence of the sequences from the 4th clade is reflected on the bootstrap value of its branch, but it produces a bias in the tree that makes it difficult to appreciate the topology of the rest of the sub-clades. Because of that, we included in Fig. 8 an unrooted tree with actual branch size, which is generated from the same trimmed alignment and same parameters in RAxML

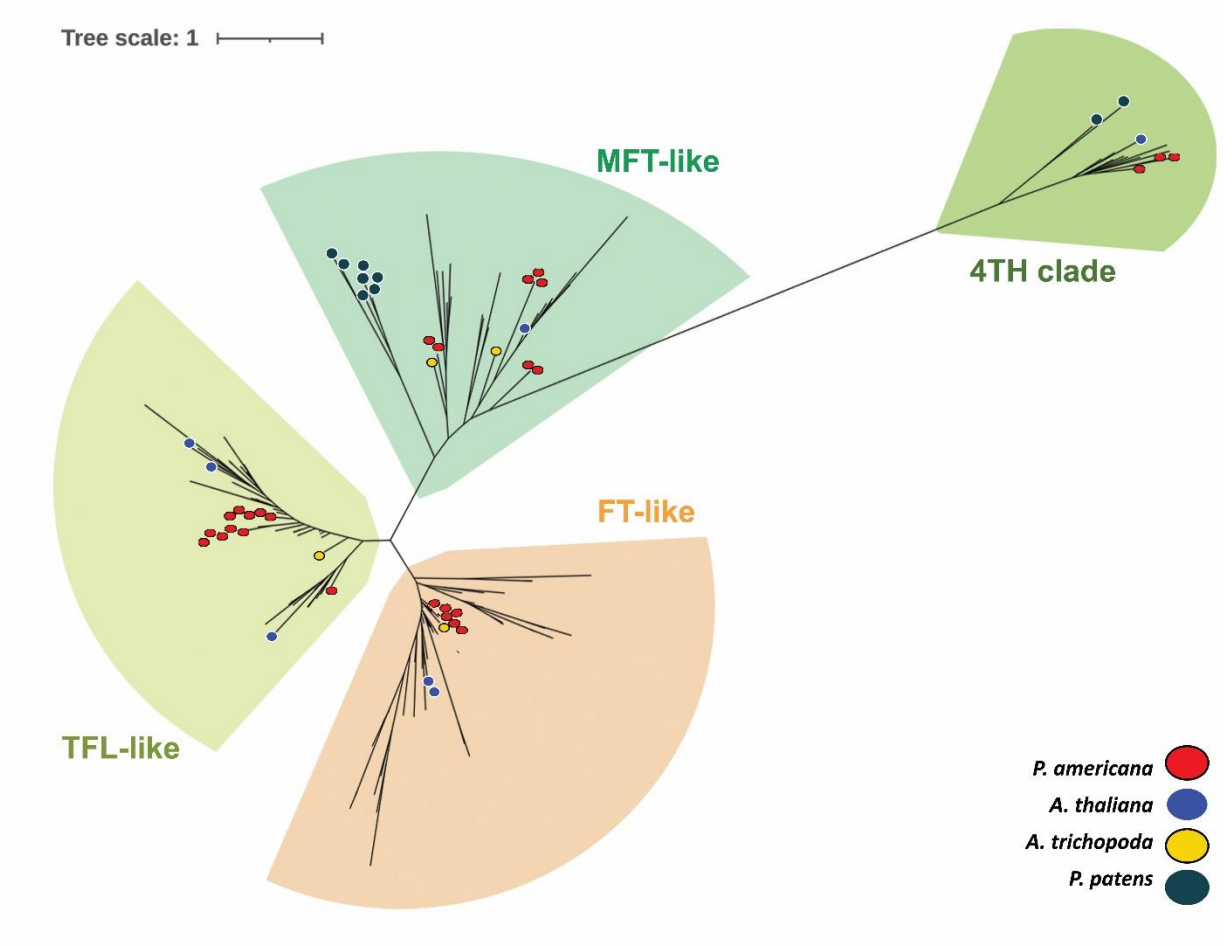


Figure 8. Unrooted phylogenetic tree of the PEBP family from 14 plant species showing the four well defined sub-clades for this gene family.

Alignments of this clade revealed important differences in the highly conserved DPDxP and GIHR motifs among all taxa (**Fig. 9**) and none of the Tyr85/His88 and Gln140/ Asp144 match with the FT/TFL characteristics.

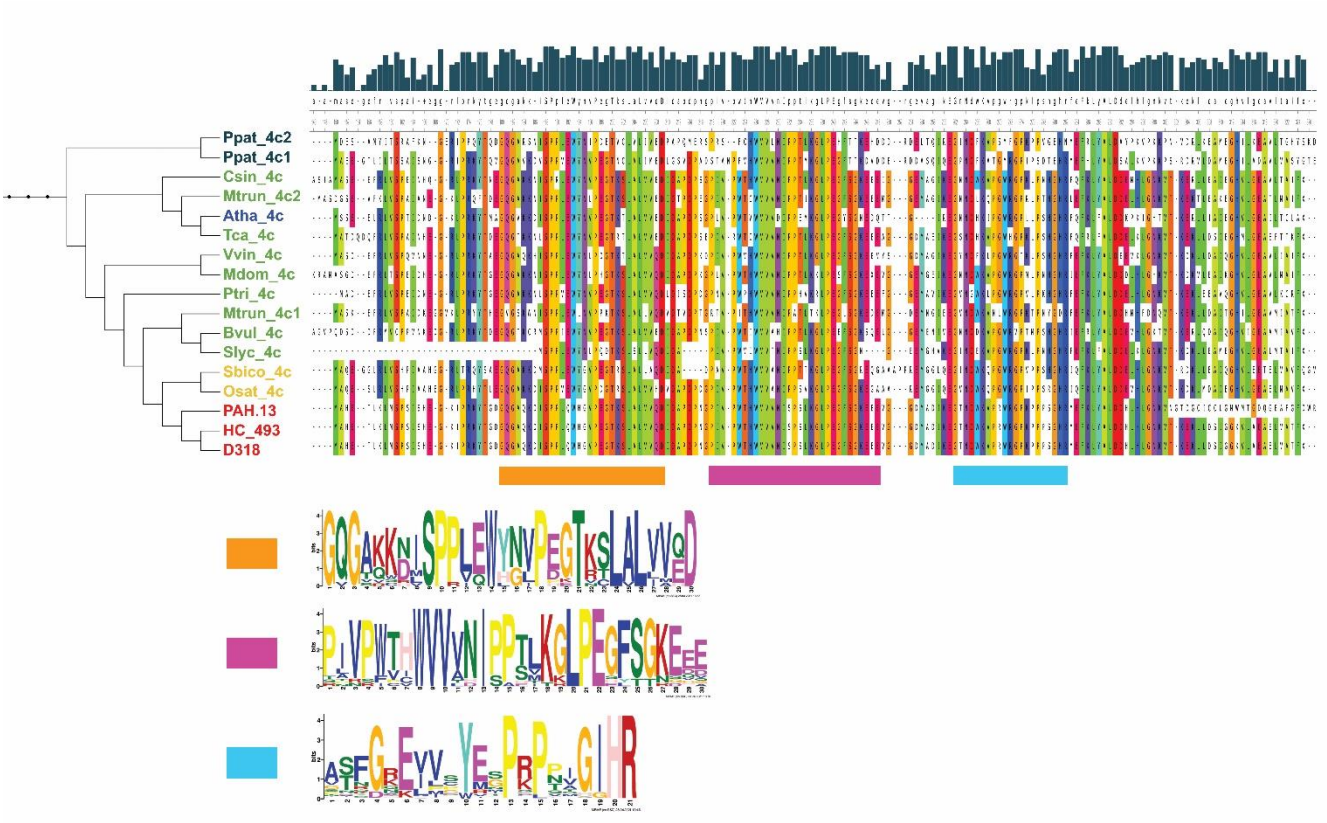


Figure 9. Alignment of the members of the 4th sub-clade, showing sites with conserved motifs in color bars and the logo obtained from MEME analysis. *P. americana* sequences ID are in red and *A. thaliana* in blue.

4.2 Infer orthologous relationships

Raw sequences from the phylogeny were separated by species in different FASTA files and then used to detect orthologous groups (OGs) by means of the OrthoFinder program. The automated general analysis was performed with the OrthoFinder -f command, using DIAMOND instead of BLAST, because of its increased speed in short sequences and similar degree of sensitivity (Buchfink *et al*., 2015).

Among key OrthoFinder's results stands out that 100% of the 158 genes were assigned to a single orthogroup, meaning that the orthogroup was strongly supported.

The program also made a duplication analysis using the gene tree as a reference, indicating that some sequences from *P. americana* might be duplicated. Notorious duplication event occurred in node 77 from the gene tree that separates FT-like and MFT-like clades, according to results sequences PAH.531, HC_0238, and D75 from the FT-like subfamily are duplicates of the sequences PAH.0810, D518, PAH.0137, D10661, HC_1904, PAH.0181 and D103 from the MFT-like. Also, sequence PAH.0072 from the TFL-like clade was found as a duplicate of the sequences from the 4th clade, PAH.13, D318, and HC_493 (**Table 9**).

Table 9. Duplication events from the gene tree of *P. americana* PEBP sequences.

Node	Gene 1	Sub-family	Gene 2	Sub-family
n47	HC_6736	TFL-like	HC_7722	TFL-like
n48	PAH.0592	TFL-like	PAH.2032	TFL-like
n49	PAH.3124, HC_10661	TFL-like	PAH.1030, D92, HC9327	TFL-like
n44	PAH.3124, HC_10661, PAH.1030, D92, HC9327, D266	TFL-like	HC_6736, HC_7722, PAH.0592, PAH.2032	TFL-like
n3	PAH.0072	TFL-like	PAH.3124, HC_10661, PAH.1030, D92, HC9327, D266, HC_6736, HC_7722, PAH.0592, PAH.2032	TFL-like
n3	PAH.0072	TFL-like	PAH.13, D318, HC_493	4th clade
n102	PAH.0531, HC_0238, D75	FT-like	D186, PAH.2544, PAH.0606	FT-like
n77	PAH.531, HC_0238, D75	FT-like	PAH.0810, D518, PAH.0137, D10661, HC_1904, PAH.0181, D103	MFT-like
n134	PAH.0810, D518	MFT-like	PAH.0137, D10160, HC_1904, PAH.0181, D103	MFT-like
n15	D266, HC_6736, HC_7722, PAH.592, PAH.2032, PAH.3124, HC_10661, PAH.1030, D92, HC_9327	TFL-like	PAH.13, D318, HC_493	4th clade

It is worth to mention, that some of the amino acid sequences from the three avocado genomes were found identical for the program RAxML, whether they were trimmed or not. It is expected that sequences from the genomes of the same variety (PAH and HC) may be identical, but further analysis of the up region and intron-exon organization may provide more information.

4.3 Gene structure and conserved motifs

Conserved motifs were predicted with Multiple Expectation Maximization for Motif Elicitation (*MEME*) in sequences of all selected taxa divided by sub-clades, motif width ranged between 6 and 30, with a maximum of 10 motifs and one occurrence per sequence. At least 35 motifs with an E-value of < 1e-4 (Ma *et al*., 2014) were found in PEBP sequences, 8 in each sequence of TFL and MFT-like proteins, 9 in the 4th clade, and 10 for the FT sub-clade (**Fig. 7b**). Three larger motifs with an e-value lower than 3e-270 were found in the 4th clade that was not found in the rest of the sequences, and motifs disposition were more dissimilar in monocots species in all the sub-clades (**Fig. 7b**), however, a clear pattern can be appreciated for the PEBP sequences from each sub-clade.

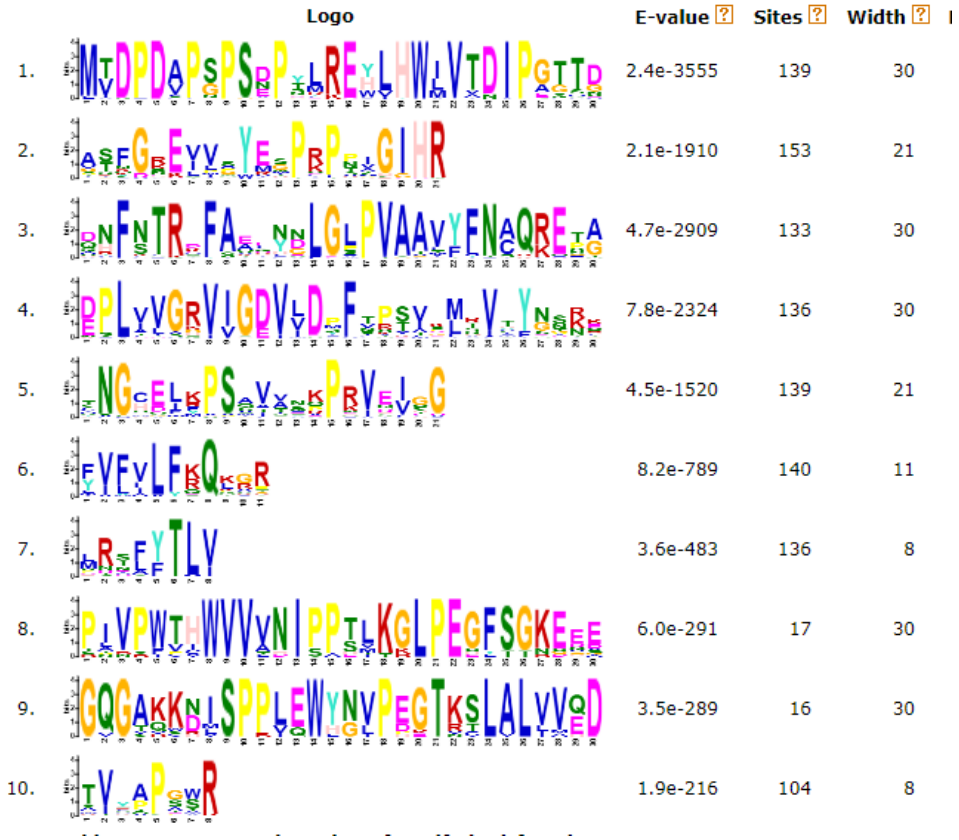


Figure 10. Logos of the conserved motifs found with MEME in the four sub-clades. The statistical significance is represented with the E-value. "Sites" are the number of sites contributing to the construction of the motif. Each motif describes a pattern of fixed "width", as no gaps are allowed in MEME.

To gain further insight into the structural diversity of the intron-exon organization, complete genes for the avocado sequence were download from NCBI and analyzed with the CDS in the Gene Structure Display Server v2 (Hu *et al*., 2014). The average of the coding region for all the genes was 515 bp, with ranges between 444 and 534 bp. The genomic organization seemed highly conserved among lineages and sub-families, most sequences of FT, MFT, and TFL-like sub-clades present four exons and three introns, except for the sequences D266 and HC_10661 which have 5 exons (**Fig. 11**).

Clear differences among gene size can be seen between the 4th clade, with less than 1000 bp, and the other three, that reach the 4000 bp, also genes in the 4th clade have only two exons and one intron.

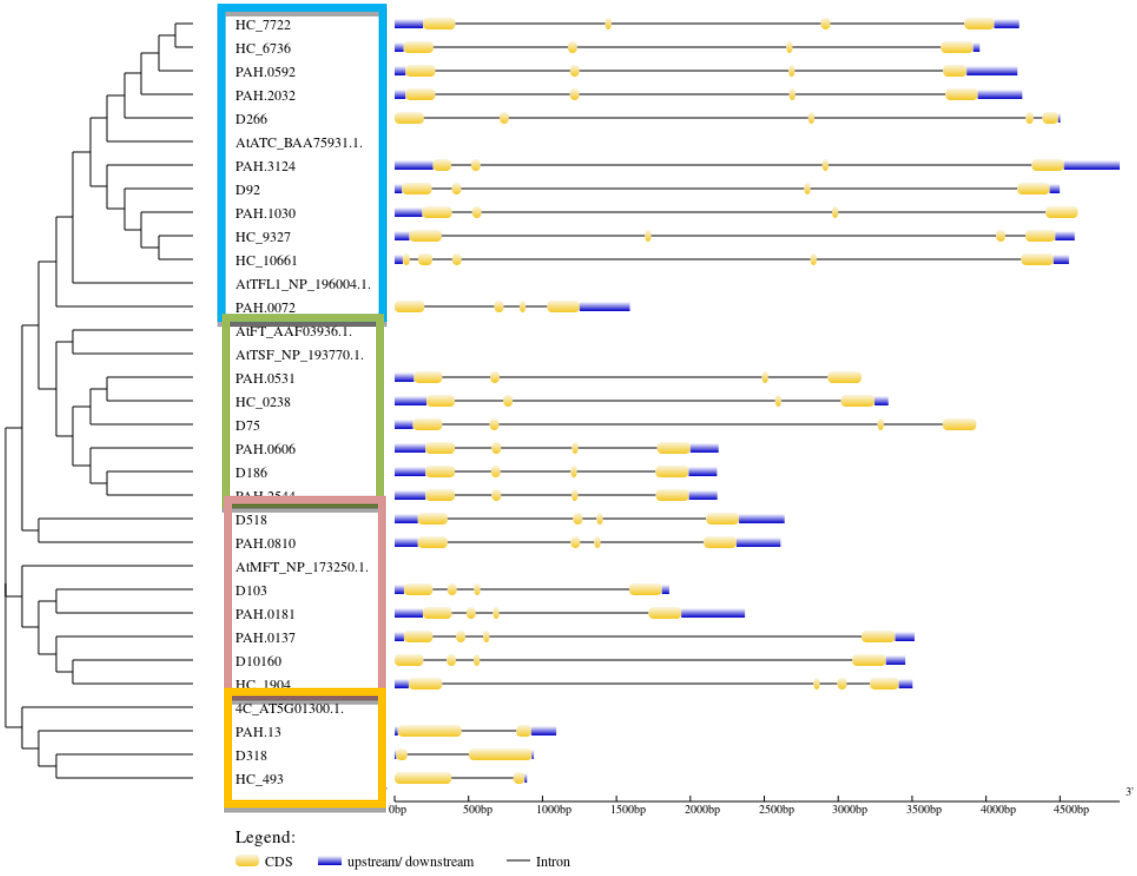


Figure 11. Organization of exons and introns from the PEBP putative genes from avocado. They are grouped according to phylogeny.

4.4 Regulatory elements and functional divergence

In order to infer the functionality of the genes, we downloaded 2 kb upstream the translation initiation codons and submit the sequences to PlantCARE database.

The program found a total of 70 different types of cis-regulatory elements in the 27 sequences, 17 of those elements are related to light responsiveness, 12 are cis-elements involved in the hormonal response, 5 stress-related motifs, and only 4 binding sites from different proteins were identified. All sequences presented TATA-box and CAAT-box motifs, which are common promotor elements, also each sequence from the 4 clades had at least 4 light-responsive and a minimum of 2 hormonal responsive elements (**Table 10**).

ABRE cis-acting element, involved in abscisic acid responsiveness was found in all the sequences from the four clades, but ABRE3a and ABRE4 were found exclusively in the 4th and TFL-like clade. Some cis-elements were found in all clades, but not in all sequences, like the MeJa responsive cis-elements CGTCA-motif and TGACG-motif, ARE motif, which is essential for the regulation of anaerobic induction, and TCA element for the salicylic acid response (**Table 11** and **Fig. 12**).

On the other way, some elements can only be found in certain sub-clades or elements that are lost in all the sequences of one sub-clade. That is the case of the AuxRR-core, an element for auxins response and the binding site Box III, found solely in proteins from FT-like clade. Another interesting element found in one clade is NON-box, related to the meristem-specific activation, that belongs to the MFT-like clade.

Table 10. Number of regulatory elements per sequence and total number of cis-elements identified per function.

Gene ID	Sub-clade	type of elements per sequence	type of LIGHT elements per seq	type of HORMONAL elements per seq	Promotor/ enhancers	Stress-related	Binding sites
HC_493	4th	33	7	5	2	1	1
PAH.13	4th	33	7	5	2	1	1
D318	4th	37	7	5	3	1	1
HC_0238	FT-like	35	7	7	3	3	1
PAH.0606	FT-like	28	4	5	3	3	1
PAH.2544	FT-like	30	5	5	3	4	1
PAH.0531	FT-like	36	7	8	3	3	1
D186	FT-like	29	4	5	3	3	1
D75	FT-like	31	6	6	3	4	1
HC_1904	MFT-like	38	8	5	3	3	1
PAH.0810	MFT-like	32	8	6	2	1	1
PAH.0137	MFT-like	37	8	5	3	3	1
PAH.0181	MFT-like	29	9	4	2	2	1
D103	MFT-like	29	8	4	2	2	1
D518	MFT-like	19	4	2	2	1	0
D10160	MFT-like	28	8	3	3	3	0
HC_10661	TFL-like	40	7	7	2	2	1
HC_9327	TFL-like	40	7	7	2	2	1
HC_7722	TFL-like	31	7	4	3	2	0
HC_6736	TFL-like	34	8	6	2	1	0
PAH.0072	TFL-like	25	7	2	2	3	1
PAH.0592	TFL-like	31	8	6	2	1	0
PAH.1030	TFL-like	37	6	6	2	2	1
PAH.2032	TFL-like	30	7	4	3	2	0
PAH.3124	TFL-like	39	6	8	2	2	1
D266	TFL-like	33	6	5	2	4	1
D92	TFL-like	40	7	7	2	2	1
total # of cis-element		70	17	12	3	5	4

Table 11. Identification symbol, name of the cis-elements, type and function of the cis-acting elements showed in figure 12.

Cis-element type	No° of cis-element	ID	Function
Hormonal responsiveness elements	A	**ABRE**	cis-acting element involved in the abscisic acid responsiveness
	B	**ABRE3a**	cis-acting element involved in the abscisic acid responsiveness
	C	**ABRE4**	cis-acting element involved in the abscisic acid responsiveness
	D	**AuxRR-core**	cis-acting regulatory element involved in auxin responsiveness
	F	**CGTCA-motif**	cis-acting regulatory element involved in the MeJA-responsiveness
	G	**TGACG-motif**	cis-acting regulatory element involved in the MeJA-responsiveness
	H	**GARE-motif**	gibberellin-responsive element
	I	**P-box**	gibberellin-responsive element
	J	**TATC-box**	cis-acting element involved in gibberellin-responsiveness
	K	**TCA-element**	cis-acting element involved in salicylic acid responsiveness
	L	**TGA- element**	auxin-responsive element
Stress-related elements	M	**ARE**	cis-acting regulatory element essential for the anaerobic induction
	N	**DRE core**	cis-acting element involved in dehydration, low-temp, salt stresses
	Ñ	**MBS**	MYB binding site involved in drought-inducibility
	O	**LTR**	cis-acting element involved in low-temperature responsiveness
	P	**TC-rich repeats**	cis-acting element involved in defense and stress responsiveness
localized expression	Q	**CAT-box**	cis-acting regulatory element related to meristem expression
	R	**GCN4_motif**	cis-regulatory element involved in endosperm expression
	S	**NON-box**	cis-acting regulatory element related to meristem specific activation
binding sites	T	**AT-rich element**	binding site of AT-rich DNA binding protein (ATBP-1)
	U	**Box III**	protein binding site
	V	**CCAAT-box**	MYBHv1 binding site
	W	**MBSI**	MYB binding site involved in flavonoid biosynthetic genes regulation
	X	**MSA-like**	cis-acting element involved in cell cycle regulation
	Y	**circadian**	cis-acting regulatory element involved in circadian control

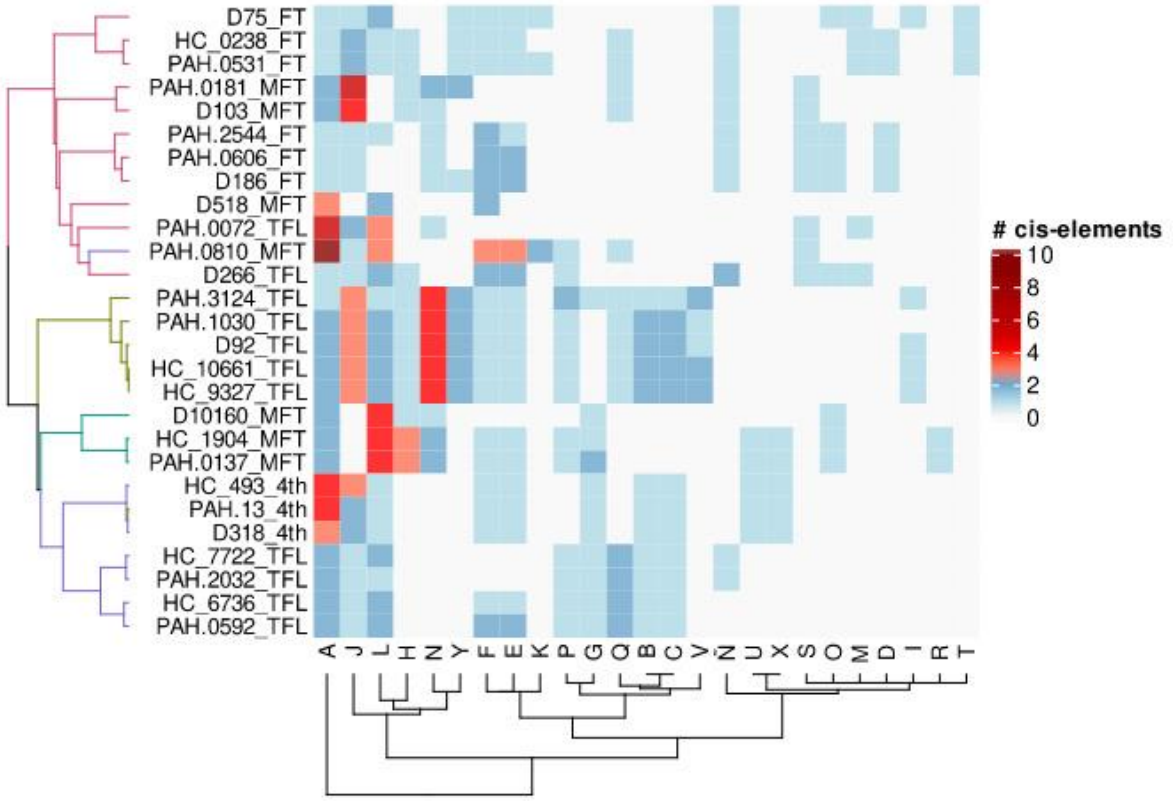


Figure 12. Heatmap of the clustering of Cis-elements found in 2000 bp upstream of the identified PEBP sequences of avocado tree. Categories in X axis are coded according to Table 11. Colored boxes represent the abundance of cis-elements per category in X axis.

Proteins from the 4th clade lacked most of the elements for stress response, besides elements for auxins response. This clade has only one kind of binding site element, the CCAAT-box, and one cis-element related to the cell-cycle regulation, MSA-like, both of which are shared with MFT-like sub-clade.

The TFL-like sequences presented cis-elements related to localized expression, the meristem expression CAT-box and GCN4_motif for the endosperm expression. FTL-like sequences also have the binding site involved in the flavonoid biosynthetic gene regulation, MBSI.

FT. TFL and MFT sub-clades have at least one sequence that presents elements for circadian control. AT-rich element for protein binding and three stress-related elements MBS, LTR, TC-rich repeats for drought-inducibility, low-temperature responsiveness, and stress-defense response, respectively.

4.5 **Prediction of Subcellular organization and protein parameters**

To complement the characterization of the putative PEBP sequences from *P. americana* previously identified, predictions of subcellular localization, molecular weight, instability index, and theoretical isoelectric point were made with ProtParam from Expasy and ProtComp v9 (**Table 12**).

Table 12. Computed parameters for the putative PEBP proteins from *P. americana.*

Sequence ID	subclade	Mw	pI	instability index	subcell-loc
HC_493	4to	18188.63	5.96	37.07	extracellular
PAH.13	4to	19213.59	5.43	35.92	extracellular
D318	4to	18188.63	5.96	37.07	extracellular
HC_0238	FT-like	19907.61	6.12	44.85	cytoplasmic
PAH.0606	FT-like	19730.33	7.73	47.73	cytoplasmic
PAH.2544	FT-like	19730.33	7.73	47.73	cytoplasmic
PAH.0531	FT-like	19907.61	6.12	44.85	cytoplasmic
D186	FT-like	19730.33	7.73	47.73	cytoplasmic
D75	FT-like	19840.52	6.73	45.52	cytoplasmic
HC_1904	MFT-like	19238.03	9.1	59.32	cytoplasmic
PAH.0810	MFT-like	19217.06	9.2	37.60	cytoplasmic
PAH.0137	MFT-like	19238.03	9.1	59.32	cytoplasmic
PAH.0181	MFT-like	18900.66	7.75	51.81	cytoplasmic
D103	MFT-like	18897.66	7.79	50.93	cytoplasmic
D518	MFT-like	19206.03	9.2	39.06	cytoplasmic
D10160	MFT-like	19238.03	9.1	59.32	cytoplasmic
HC_10661	TFL-like	17490.96	9.5	50.79	cytoplasmic
HC_9327	TFL-like	19656.45	9.63	50.62	cytoplasmic
HC_7722	TFL-like	19507.09	9.16	48.59	cytoplasmic
HC_6736	TFL-like	19507.09	9.16	48.59	cytoplasmic
PAH.0072	TFL-like	19563.61	8.97	43.80	cytoplasmic
PAH.0592	TFL-like	19507.09	9.16	48.59	cytoplasmic
PAH.1030	TFL-like	19656.45	9.63	50.62	cytoplasmic
PAH.2032	TFL-like	19507.09	9.16	48.59	cytoplasmic
PAH.3124	TFL-like	16777.96	9.64	53.97	cytoplasmic
D266	TFL-like	16963.19	6.5	44.21	cytoplasmic
D92	TFL-like	19656.45	9.63	50.62	cytoplasmic

There was very little variation between the molecular weight of the proteins, with exception of HC_10661, PAH.3124, and D266, most of the sequences ranged between 18 and 19 kD. Subcellular localization of FT, TFL, and MFT-like proteins were predicted in the cytoplasm, which was expected according to the *in vivo* characterization reported in previous investigations (Jin *et al*., 2019). On the other hand, 4th clade was indicated as extracellular proteins and their instability index was much lower than the average of the rest of PEBP sequences, meaning that their stability in a test tube is higher.

4.6 Discussion

4.6.1 Identification and characterization of *P. americana* PEBP genes

We have identified 12, 7, and 8 putative PEBP genes in the genome of Hass cultivar from CINVESTAV, Hass cultivar from Hainan University, and drymifolia variety from CINVESTAV, respectively. We were expecting to find the same number of genes in the three genomes since they were from the same species, however, the assembly level of the two Hass genomes were only up to contigs, and especially in the Hass Hainan genome, we found large regions of indeterminate amino acids. The low coverage of the sequencing and the genome assembly compromises the certainty of the number of PEBP genes found in the Hass Hainan genome, but the 100% identity with the amino acid sequences found the other two genomes may support the presence of these genes in *P. americana*.

Before this work was conducted, only two sequences of PEBP genes were reported in the avocado tree. The first one corresponds to a member of the TFL1 sub-clade, and it was used in phylogenetic reconstruction of the TFL-like sub-clade (Gao *et al*., 2017). The authors concluded that Basal angiosperms, including Persea, did not possess any TFL1 duplications, however, in our results, a larger number of genes were assigned to this clade. Among the 11 genes from the TFL sub-clade, PAH.1030, HC_9327 and D92 showed identical amino acid sequences to the TFL1 homolog reported in NCBI (KY933634.1) for *P. americana,* one for

each genome analyzed (**Fig.13**), which could be interpreted as a positive control for our search methodology.

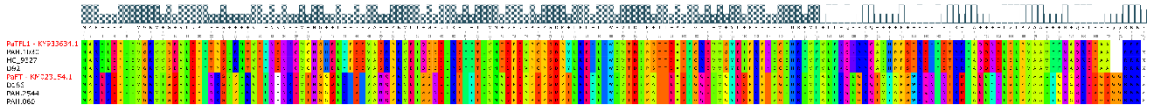


Figure 13. Alignment of the two reported PEBP proteins (PaFT and PaTFL1 with an identity of 100% regarding the retrieved proteins in this study.

The second report available was for an orthologous of *A. thaliana* FT; this protein was monitored during the off and on periods of Hass avocado cultivars, increasing before the start of the "on" period, and functionally characterize by ectopic expression in *A. thaliana*. Both experiments of this paper proved the relationship between the expression of PaFT and the start of the flowering time (Ziv *et al*., 2014). From our results, three out of six identified proteins from the FT-like genes, one for each genome analyzed, resulted identical to the PaFT sequence reported before and published in the NCBI (KM023154.1), those proteins are D186, PAH.2544, and PAH.0606 (**Fig. 11**).

Most of the sequences from FT and TFL-like proteins showed high conservation on key amino acids involved in substrate binding activity and the interaction with 14-3-3 proteins, except for the sequence PAH.0072 from the TFL-like sub-clade. This sequence presents a substitution in the short motif DPDxP, that forms a binding pocket (Nakamura *et al*., 2019), and is conserved even among mammalian PEBP proteins (Banfield & Brady, 2000). The substitution consists of Asp77 replaced by an Asn. Both amino acids have different biochemical properties, asparagine is a polar amino acid with a neutral R group, while aspartic acid has a negatively charged R group (Rodwell *et al*., 2018).

Another interesting change in the PAH.0072 sequence is in position 70, which is part of the amino acids conforming to the interaction loop with the 14-3-3 proteins (Taoka *et al*., 2011). Minimum changes in amino acid sequences may cause different conformation of the protein interaction sites and therefore a different effect over the phenotype (Hanzawa *et al*., 2005).

Even so, there have been some cases in which proteins classified in certain sub-clade and sharing the conserved residues Tyr-85/His-88, produced an unexpected phenotype. That occurs with a homologous of FT in *Fragaria ananassa* (with only 13 non-conserved substitutions compared to other members of FT sub-clade of *Rosa hybrida* and only 1 amino acid difference in segment B with *A. thaliana* FT) that showed an effect on tobacco phenotype more likely to the TFL proteins (Zhen Wang *et al*., 2017).

Another example of a protein from one sub-clade, performing the function of another, was described in sugar beet by Pin *et al*. (2010), where a pair of FT homologous were found to have antagonist activity on inducing flowering. In this sense, relaxation of selection pressure on the second copy of PEBP genes in any sub-clade may be an important source of genetic material for evolution, allowing functional diversification (Blackman *et al*., 2010).

The difference in the expression patterns is the first evidence of functional diversification of PEBP genes from the same sub-clade, as is proposed by Cai *et al*. (2019). They found a second FT-like gene in the London plane (*Platanus acerifolia* Willd), displaying a divergent pattern of expression during dormant stages of the plant and show a different response to environmental stimuli than the first copy.

Taking that information into account, phylogenic classification of PEBP proteins maybe not enough to predict the effect of a specific sequence over the plant metabolism and phenotype, making it more relevant to complement with *in silico* and functional characterization. Differences in the sequence of PAH.0072, as well as proteins from the 4th sub-clade, make them interesting targets to the next stage of the investigation.

4.6.2 Phylogenetic relationships of PEBP family

The evolutionary history of the PEBP family in plants showed that all the analyzed genes found in the 13 species belong to the same orthogroup, but phylogeny reconstruction divided these sequences into 4 major sub-clades. Three of the sub-clades have been widely characterized in previous investigations (Karlgren et al., 2011) but the 4th clade, which is immersed in the MFT clade, has only been identified in recent years (Dong *et al*., 2020).

However, evidence of the presence of this 4th clade has been mentioned from the first phylogenetic study of the MFT clade by Hedman *et al*., (2009). The authors described a group of proteins with partial PEBP-like characteristics, with divergent N-terminal but more conserved in C-terminal and almost conserved DPDxP and GHIR motifs, consistent with our sequences from the 4th clade. An important detail is that the PEBP-like sequences in their work were found exclusively in the species of Marchantia, Selaginella, Ginkgo, and Pinus, while we found it in most of the species used for the phylogenic reconstruction, also, Hedman *et al*. did not use those sequence in further analysis because of their large divergence.

Over the next years, other authors identified proteins with atypical characteristics for the MFT or FT sub-clades. Zhang *et al*. (2016) identified 9 PEBP genes in three cotton species *G. raimondii*, *G. arboretum*, and *G. hirsutum*; two of the identified PEBP genes (GhPEBP1 and GhPEBP2) showed more similarity with PEBP genes of bacteria and archaea. Those proteins belonged to *Gossypium hirsutum* and they have two exons each, like proteins from the 4th clade found in our study, but their size was very variant.

Also, phylogenetic analysis done by Zhang and collaborators (2016) with PEBP proteins from other species, confirmed the presence of a 4th clade that was named PEBP-like. They measured the tissue-specific expression of the PEBP-like/4th clade proteins, finding that these proteins had higher expression in buds and flowers for GhPEBP1 and leaves, apical shoots, buds, and flowers for GhPEBP2. More importantly, they made a functional characterization of one of the FT-like proteins and one of the PEBP-like proteins, finding a consistent early flowering on

the transgenic FT overexpressing lines, but they did not find any pattern in the phenotype from the lines overexpressing PEBP-like (Zhang *et al*., 2016).

In 2019, Wang and collaborators conducted a study on PEBP proteins in four different species of cotton (*G. raimondii*, *G. arboretum*, and *G. hirsutum,* plus *G. barbadense)* revealed more sequences than the previous studies (M. Wang *et al*., 2019). They also found sequences from a divergent clade but unlike the paper of Zhang *et al*., their 4th clade (named FT-like), only showed sequences from cotton species. However, both articles present the intron/exon organization, and the sequences from their PEBP-like/FT-like clade and both results coincide with the organization that we found in our 4th clade.

Wang *et al*. (2019), performed an expression profile of the 20 PEBP genes that they found in *G. hirsutum* and it was noticeable that the fourth FTL genes were expressed with specificity in different tissues; GhFTL2A was highly expressed in stamen, GhFTL1A and GhFTL1D were preferentially expressed in 25dpa fiber, and GhFTL2D had higher expression in more tissues, root, leaf, petal, and ovule.

Another genome-wide identification published in 2019 as well, this time in *Arachis duranensis* and *Arachis ipaensis,* presented evidence of a different clade from FT, TFL, and MFT sub-clades (Jin *et al*., 2019). In contrast with the previously described investigations, in this article phylogenetic analysis did not recognize the existence of a 4th clade, but it shows a large divergence of three sequences from the MFT sub-clade.

The three divergent sequences from the MFT clade share some characteristics with the sequences from our identified 4th clade, like the two-intron organization, their closer phylogenetic relationship with the MFT sub-clade, and a motif organization different from every other sub-clade. Subcellular prediction of these proteins as well as our results differed from the rest of the other three sub-clades, but in this paper from Jin *et al*. (2019), it was inconclusive.

Cis-acting elements like MSA-like, related to the cell cycle regulation were found exclusively in proteins of the MFT/4th clade, which agree with our results for this

group of proteins. At the same time, expression of the three genes from the MFT/4th clade (Aradu60NUI, Aradu23179, and AraipV0B0S) was higher in seed samples, with two of the genes also very expressed in main stem leaves, flowers, and pistils.

In this work, it was not possible for us, to compare expression levels of PEBP genes in avocado, and the only antecedent about PEBP members in this species is addressed by (Ziv *et al*., 2014), who linked the expression of avocado FT to its flower induction. Even so, it is possible to correlate cis-acting elements with expression patterns from the literature; In *A. thaliana,* for example, 4th clade protein was found to be strongly expressed in seed and under drought-stress treatment (by revisiting the Arabidopsis RNA-seq Database, http://ipf.sustech.edu.cn/pub/athrdb/), that information supports the results from the expression of the Arachis sequences. With that said, and the results we got for the 4th clade sequences from avocado, we infer that HC_493, PAH.13, and D318 may be expressed in tissues with a high rate of cellular division as meristems or seed and activated under hydric stress.

Finally, (Dong *et al*., 2020) searched for PEBP genes in common wheat, *Triticum dicoccoides*, *Triticum urartu*, and *Aegilops tauschii*. In this work, their phylogenetic analysis supported the existence of a 4th clade called PEBP-like again, which also included the *A. thaliana* gene AT5G01300 that we found in our results. An interesting fact about their PEBP-like clade is that the subcellular localization was found cytoplasmic while AtFT and AtTSF were found extracellular, that is the opposite result than the one we got for the retrieved PEBP's from *P. americana*.

Despite these 5 articles, there is a notorious lack of information about the proteins from the 4th clade, starting from the recognition of the clade itself. Most of the identification effort for the PEBP resulted in the differentiation of only three sub-clades FT, TFL, and MFT-like (Z. Yang *et al*., 2019; Zhao *et al*., 2020) and none 4th clade proteins. The reason behind that may be in three principal aspects: the search method for the genes in their species of interest, the sampling method for

taxa and proteins for the phylogenetic reconstruction, and the methodology for phylogenetic analysis.

The majority of papers reporting only three sub-clades employ a sequence-based homology search, which means that they use proteins from a model organism as query against the database from their species of interest, some examples of that strategy are the work of (Klintenäs *et al*., 2012), where they use similarity BLAST search with amino acid sequences from the PEBP family members from Arabidopsis and Populus as query against expressed sequence tag (EST) nucleotide collections from NCBI and Ancestral Angiosperm Genome Project; Another example is found by (Zheng Wang *et al*., 2015), they used the homology approach with BLAST and six members of PEBP in *A. thaliana*, obtaining 23 sequences from PEBP family on soybean, apparently none of them from the 4th clade.

This kind of methodology may be biased if the query sequences come from a very distant species, losing resolution to find large divergent sequences (Selzer *et al*., 2018). In the case of the PEBP family, the high grade of conservation of the domain among all taxa may validate the similarity search (Banfield & Brady, 2000), but an important bias may result from the fact that there is not an identified protein from the 4th clade in *A. thaliana,* the most common choice as a model (Mimida *et al*., 2001). That means that proteins from these sub-clades are not included in most of the phylogeny works and may also be the reason why our results did not show any 4th clade proteins at the first attempt of identification against *P. americana* genomes.

In some cases, the objective of the research is to identify members of just one sub-group of the proteins, like (Wu *et al*., 2019), that used BLAST and tBLASTn to find members of the FT and TFL sub-clade on the genome database of Petunia, and (Nakano *et al*., 2015) who looked for PEBP sequences in the genome of Fragaria. In those cases, the homology-based approach could be more justified, however, it is not accurate to talk about whole PEBP family analysis if we do not include all the sub-clades.

An interesting example of how the change of strategy may impact the final identification is the two papers about cotton mentioned above, in which, second research about the PEBP family with a model-based search resulted in 20 sequences whereas the first attempts only obtained 9 genes identified.

Selecting taxa for a phylogenetic reconstruction could be an important factor to consider, a poor sampled analysis could not reflect some phylogenetic relationships or evolution patterns. An important difference between previous investigations is the intended approach, those papers interested in phylogeny and evolution did include a wider selection of species (Klintenäs *et al*., 2012). On the contrary, papers focused on identification and characterization used principally model species with characterized PEBP proteins, as it happens in the Arachis paper (Jin *et al*., 2019), where proteins with 4th clade characteristics were successfully found using amino acid sequence from the PEBP conserved domain in a BLAST search, but the phylogenetic analysis did not support the 4th clade.

At the same time, protein search for phylogenetic reconstruction, in most cases, is biased by the selection of functionally characterized or already annotated proteins. As the 4th clade is still not characterized, divergent proteins found in the PEBP family are considered part of any other sub-clade like in Arachis, ignored like Hedman *et al*. and Lu *et al*., 2019, or simply they were not found.

Except for Hedman *et al*. (2009) that used MFT proteins of *Physcomitrella patens* for the BLAST search, the rest of the authors used a model-based search. Among them, the only other paper that used the model-based search but did not found 4th clade proteins was (Zhao *et al*., 2020). In our experiment the HMM-based search allowed us to find a larger number of 4th clade proteins, at least one for each species. The reason behind our results may be because searches based on Hidden Markov Models use probabilistic modeling (Choo *et al*., 2004), enabling a more flexible and robust search.

Finally, the accuracy of the phylogenetic tree may depend on the parameters used to compute the results, such as the alignment algorithm, the model of evolution, or the tree calculation program. A common selection for protein alignment is clustal W

(Thompson *et al*., 1997), it functions under the fact that similar sequences are usually homologous, but we found that due to the divergence of some sequences, it was better to align proteins by using the HMM model for PEBP proteins instead. Our analysis found that the best model to describe the evolution pattern in the proteins from the PEBP family, was JTT model with a gamma distribution, (Lu *et al*., 2019; Tribhuvan *et al*., 2020), used the same model, with some variants in the distribution parameter.

Ultimately, a deep phylogeny reconstruction requires a robust program and generates multiple trees to compare. (Liu *et al*., 2016) used two methodologies to determinate the best tree, they compare Maximum Likelihood performed on PHYML2.4.4 (Guindon & Gascuel, 2003) and Bayesian inference performed in MrBayes 3.1 (Ronquist & Huelsenbeck, 2003), while (Zheng *et al*., 2016) compare trees obtained with neighbor-joining in MEGA 5.0 against Maximum Likelihood in RAxML 7.0.4.

Surprisingly, from the papers reporting the 4th clade, 4 of them used MEGA to calculate the phylogenetic tree and 3 of those even used the neighbor-joining method, which has been proved to generate less reliable results than Maximum likelihood (Whelan and Morrison, 2017). That may mean that the most critical step in the identification of a 4th clade for PEBP family is the search methodology.

Chapter 5. Conclusions and future work

The commercialization of avocado fruit has been raising benefits for Mexico´s economy. It is estimated that by the year 2030, avocado demand at a global level will double, however, fast-changing climatic conditions may affect the production of this fruit. Also, avocado crops require a large amount of water, compromising the hydrologic resources of the countries that cultivate avocado on a big scale.

Because of that, it is imperative to explore the genetic resources of the avocado to find alternatives for the improvement of production. The *in silico* identification of the PEBP genes is the first step to understand the functions of this group of proteins in *P. americana*. Until today, the information available on PEBP proteins in other species shows that there is a conserved function as integrators for flowering, germination, and stress responsiveness, which invites us to think that in avocado, the function of the sequences found may also be related to these processes.

Thanks to this, there is a possibility of manipulating the proteins of the FT sub-clade to develop early flowering phenotypes, since they are known to act as integrators of flowering stimuli. In addition, the mobility of the FT sub-clade proteins makes it theoretically possible to implement transient transformation strategies, avoiding the generation of transgenic organisms, in a way that it is possible to segregate genes of interest in avocado by means of artificial crosses and in less time.

The low quality of the genome assembly compromises the certainty of the number of sequences found in each genome, for example, we did not find the same number of genes in two genomes of the same cultivar. Analysis of an available annotated proteome from leaves, roots, or seed of avocado, could add more resolution to the search of PEBP genes in this species.

From the total 27 PEBP sequences found in *P. americana*, 11 were classified as TFL-like in the phylogenic analysis, and their regulatory elements suggest a function on the light and stress responsiveness, as well as a localized expression in meristem.

According to the phylogenetic tree, from the 6 sequences of the FT sub-clade, three of them were identical to the FT amino acid sequence previously reported and characterized in *A. thaliana*, implying their role as florigen signal.

Finally, MFT and 4th clade proteins shared some interesting regulatory elements, like MSA-like element, which is involved in the cell-cycle regulation and hormonal response to gibberellin, abscisic acid, and salicylic acid. Even so, MFT-like sequences presented an element of tissue-specific expression for meristem, unique of this sub-clade, and genes from the 4th lade only have one type of cis-element related to stress response, ARE, involved in anaerobic induction. Those differences add evidence of the functional divergence between the two sub-clades, MFT and 4th, found in the phylogenetic tree.

Compared to monocots, *P. americana* has a fewer number of members from the FT sub-clade, but the same or more sequence from the TFL sub-clade. As TFL genes usually function to inhibit flowering and vegetative growth, a larger number of copies may be involved in their growth habits. On the other hand, perennial tropical fruits like *C. sinensis* and *M. domestica* had a closer number of total PEBP sequences, also the number of TFL sequences in *M. domestica* is larger than in the avocado.

All the information from the PEBP sequences from avocado will be useful for future investigations, but we found other research opportunities too, the existence of a fourth clade of PEBP family, with members in each species analyzed and very little information reported in the literature. The 4th clade showed a different motif arrangement, a characteristic intron-exon organization that can be distinguished from the other sub-clades, regulatory elements related to cell cycle regulation, and the predicted subcellular localization was found as extracellular. The antecedents show that this 4th sub-clade is not implied in flowering time, but papers neither determine the effect of this protein group on the phenotype.

We propose the recognition of the 4th sub-clade as a member of the PEBP family to perform more accurate phylogenetic analysis in this family in plants and awake the interest over the function of the proteins from the 4th clade.

5.1 Future work

HMM-based search has been effective for the detection of even highly divergent sequences, and recent publication of the seed proteome from avocado by (Juarez-Escobar et al., 2021) may allow complementary analysis for the identification of PEBP sequences.

Besides the proteome search, there are sets of gene expression of *P. americana* reported in the GEO database from NCBI. Three datasets report samples and conditions that may of interest to the project: The transcriptional profiling of Persea americana flowers (GEO accession GSE13737), Transcriptome response of avocado roots subjected to flooding and infection by the oomycete Phytophthora cinnamomi (GEO accession GSE81297), and Transcriptome analysis of an incompatible *Persea americana-Phytophthora cinnamomi* interaction reveals the involvement of SA- and JA-pathways in a successful defense response (accession GSE119635). These expression profiles are sequenced from buds, flowers, initiating fruits, leaves tissue and roots, under different treatments of stress, which may be good options for the analysis *in silico* of the expression patterns of PEBP genes.

The next steps on the investigation of PEBP genes in *P. americana* would be the direct amplification of each sequence found in this work from genomic DNA, to confirm the presence of the genes. Once the presence and identity of the genes are confirmed, we may sequence the amplicon and follow their expression in different tissue under different growth phases so we may distinct functional divergence among the members of the same sub-clade and determinate the best candidates to clone.

Up to this point, with all the information gathered during the phase of bioinformatic analysis and the confirmation of the physical presence of the PEBP genes, the logical next step is the functional characterization. To achieve that, it will be necessary to standardize the transformation and corroboration techniques, which

later may be convenient if we intend to use any of the sequences found for a biotechnological application like inducing early flowering in avocado.

We still lack much information about the specific role of PEBP proteins in *P. americana*, but this work has settled the groundwork for a larger investigation in the whole family of phosphatidylethanolamine-binding proteins, either in bioinformatics or in lab experimentation.

References

Ahn, J. H., Miller, D., Winter, V. J., Banfield, M. J., Jeong, H. L., So, Y. Y., Henz, S. R., Brady, R. L., & Weigel, D. (2006). A divergent external loop confers antagonistic activity on floral regulators FT and TFL1. *EMBO Journal*, *25*(3), 605–614. https://doi.org/10.1038/sj.emboj.7600950

Ahsan, M. U., Hayward, A., Irihimovitch, V., Fletcher, S., Tanurdzic, M., Pocock, A., Beveridge, C. A., & Mitter, N. (2019). Juvenility and vegetative phase transition in tropical/subtropical tree crops. *Frontiers in Plant Science*, *10*(June). https://doi.org/10.3389/fpls.2019.00729

Al-Mulla, F., Bitar, M. S., Taqi, Z., & Yeung, K. C. (2013). RKIP: Much more than Raf Kinase inhibitory protein. In *Journal of Cellular Physiology* (Vol. 228, Issue 8, pp. 1688–1702). J Cell Physiol. https://doi.org/10.1002/jcp.24335

Álvarez Bravo, A., Salazar García, S., Ruiz Corral, J. A., & Medina García, G. (2017). Escenarios de cómo el cambio climático modificará las zonas productoras de aguacate 'hass' en Michoacán. *Revista Mexicana de Ciencias Agrícolas*, *19*, 4035. https://doi.org/10.29312/remexca.v0i19.671

Andrés, F., Kinoshita, A., Kalluri, N., Fernández, V., Falavigna, V. S., Cruz, T. M. D., Jang, S., Chiba, Y., Seo, M., Mettler-Altmann, T., Huettel, B., & Coupland, G. (2020). The sugar transporter SWEET10 acts downstream of FLOWERING LOCUS T during floral transition of Arabidopsis thaliana. *BMC Plant Biology*, *20*(1), 1–14. https://doi.org/10.1186/s12870-020-2266-0

Asadi Khanouki, M., Rezanejad, F., & Millar, A. A. (2020). Sequence and functional analysis of a TERMINAL FLOWER 1 homolog from Brassica juncea: a putative biotechnological tool for flowering time adjustment. *GM Crops and Food*, *11*(2), 79–92. https://doi.org/10.1080/21645698.2019.1707340

Balasubramanian S, Sureshkumar S, Lempe J, Weigel D, 2006. Potent induction of Arabidopsis thaliana flowering by elevated growth temperature. PLoS Genetics 2, e106.

Banfield, M. J., & Brady, R. L. (2000). The structure of Antirrhinum centroradialis protein (CEN) suggests a role as a kinase regulator. *Journal of Molecular Biology*, *297*(5), 1159–1170. https://doi.org/10.1006/jmbi.2000.3619

Bartoli, J. (2013). Manual técnico del cultivo de aguacate Hass (Persea americana L.). *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Baudry, A., Ito, S., Song, Y.H., Strait, A.A., Kiba, T., Lu, S., Henriques, R., Pruneda-Paz, J.L., Chua, N.H., Tobin, E.M., Kay, S.A., and Imaizumi, T. (2010). F-box proteins FKF1 and LKP2 act in concert with ZEITLUPE to control Arabidopsis clock progression. Plant Cell 22, 606–622.

Bernier, I., Tresca, J. P., & Jollès, P. (1986). Ligand-binding studies with a 23 kDa protein purified from bovine brain cytosol. *Biochimica et Biophysica Acta (BBA)/Protein Structure and Molecular*, *871*(1), 19–23. https://doi.org/10.1016/0167-4838(86)90128-7

Blackman, B. K., Strasburg, J. L., Raduski, A. R., Michaels, S. D., & Rieseberg, L. H. (2010). The Role of Recently Derived FT Paralogs in Sunflower Domestication. *Current Biology*, *20*(7), 629–635. https://doi.org/10.1016/j.cub.2010.01.059

Blázquez MA, Green R, Nilsson O, Sussman MR, Weigel D. (1998). Gibberellins promote flowering of Arabidopsis by activating the LEAFY promoter. Plant Cell, 10:791–800.

Blázquez, M. A., Ahn, J. H., & Weigel, D. (2003). A thermosensory pathway controlling flowering time in Arabidopsis thaliana. *Nature Genetics*, *33*(2), 168–171. https://doi.org/10.1038/ng1085

Boss, P. K., Bastow, R. M., Mylne, J. S., & Dean, C. (2004). Multiple pathways in the decision to flower: Enabling, promoting, and resetting. *Plant Cell*, *16*(SUPPL.), 18–32. https://doi.org/10.1105/tpc.015958

Bouché, F., Lobet, G., Tocquin, P., & Périlleux, C. (2016). FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Research*, *44*(D1), D1167-D1171.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59-60.

Cai, F., Shao, C., Zhang, Y., Bao, Z., Li, Z., Shi, G., Bao, M., & Zhang, J. (2019). Identification and characterisation of a novel FT orthologous gene in London plane with a distinct expression response to environmental stimuli compared to PaFT. *Plant Biology*, *21*(6), 1039–1051. https://doi.org/10.1111/plb.13019

Cao, S., Kumimoto, R.W., Gnesutta, N., Calogero, A.M., Mantovani, R., and Holt, B.F., 3rd (2014). A distal CCAAT/NUCLEAR FACTOR Y complex promotes chromatin looping at the FLOWERING LOCUS T promoter and regulates the timing of flowering in Arabidopsis. Plant Cell 26, 1009–1017

Carmona, M. J., Calonje, M., & Martínez-Zapater, J. M. (2007). The FT/TFL1 gene family in grapevine. *Plant molecular biology*, *63*(5), 637-650.

Chautard H, Jacquet M, Schoentgen F, Bureaud N, Benedetti H. 2004. Tfs1p, a member of the PEBP family, inhibits the Ira2p but not the Ira1p Ras GTPase-activating protein in Saccharomyces cerevisiae. Eukaryotic Cell 3: 459–470.

Chen, Y., Xu, X., Chen, X., Chen, Y., Zhang, Z., Xuhan, X., Lin, Y., & Lai, Z. (2018). Seed-specific gene MOTHER of FT and TFL1 (MFT) involved in embryogenesis, hormones and stress responses in dimocarpus longan lour. *International Journal of Molecular Sciences*, *19*(8), 3–6. https://doi.org/10.3390/ijms19082403

Choi, K., Kim, J., Hwang, H. J., Kim, S., Park, C., Kim, S. Y., & Lee, I. (2011). The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in Arabidopsis, by recruiting chromatin modification factors. *The Plant Cell*, *23*(1), 289-303.

Choo, K. H., Tong, J. C., & Zhang, L. (2004). Recent applications of hidden Markov models in computational biology. *Genomics, proteomics & bioinformatics*, *2*(2), 84-96.

Cossio-Vargas, L. E., Salazar-García, S., González-Durán, I. J. L., & Medina-Torres, R. (2008). *Fenología del aguacate 'hass' en el clima semicálido de nayarit, méxico*. *14*(3), 319–324.

Dally, N., Jung, C., & Blumel, M. (2015). *Flowering time regulation in crops — what did we learn from Arabidopsis ? 32*, 121–129. https://doi.org/10.1016/j.copbio.2014.11.023

de la Luz Sánchez-Pérez, J. (1999). *RECURSOS GENÉTICOS DE AGUACATE (Persea americana Mill.) Y ESPECIES AFINES EN MÉXICO*. *1101*, 7–18.

Dong, L., Lu, Y., & Liu, S. (2020). Genome-wide member identification, phylogeny and expression analysis of PEBP gene family in wheat and its progenitors. *PeerJ*, *8*, 1–22. https://doi.org/10.7717/peerj.10483
Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*(1), 1-14.
Fantinato, E., Del Vecchio, S., Giovanetti, M., Acosta, A. T. R., & Buffa, G. (2018). New insights into plants co-existence in species-rich communities: The pollination interaction perspective. *Journal of Vegetation Science*, *29*(1), 6–14. https://doi.org/10.1111/jvs.12592
Franks, S. J., Sim, S., & Weis, A. E. (2007). Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(4), 1278–1282. https://doi.org/10.1073/pnas.0608379104
Gao, J., Huang, B. H., Wan, Y. T., Chang, J., Li, J. Q., & Liao, P. C. (2017). Functional divergence and intron variability during evolution of angiosperm TERMINAL FLOWER1 (TFL1) genes. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-13645-0
Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, 571-607.
Godínez, M., Martínez, M., Melgar, N., Méndez, W. (2000). El Cultivo de Aguacate. Guía Técnico PROFRUTA – MAGA. Guatemala.
Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., ... & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, *40*(D1), D1178-D1186.
Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.
Hagiwara W, Uwatoko N, Sasaki A, Matsubara K, Nagano H, Onishi K, Sano Y. 2009. Diversification in flowering time due to tandem FT-like gene duplication, generating novel Mendelian factors in wild and cultivated rice. Molecular Ecology 18: 1537–1549.
Hanzawa, Y., Money, T., & Bradley, D. (2005). A single amino acid converts a repressor to an activator of flowering. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(21), 7748–7753. https://doi.org/10.1073/pnas.0500932102
Hedman, H., Källman, T., & Lagercrantz, U. (2009). Early evolution of the MFT-like gene family in plants. *Plant Molecular Biology*, *70*(4), 359–369. https://doi.org/10.1007/s11103-009-9478-x
Hepworth SR, Valverde F, Ravenscroft D, Mouradov A, Coupland G (2002) Antagonistic regulation of flowering-time gene SOC1 by CON- STANS and FLC via separate promoter motifs. EMBO J 21: 4327–4337
Hiraoka, K., Yamaguchi, A., Abe, M., and Araki, T. (2013). The florigen genes FT and TSF modulate lateral shoot outgrowth in Arabidopsis thaliana. Plant Cell Physiol. 54, 352–368. doi: 10.1093/pcp/pcs168
Hisamatsu T, King RW. The nature of floral signals in Arabidopsis. II. Roles for FLOWERING LOCUS T (FT) and gibberellin. JExp Bot 2008, 59:3821–3829.
Ho, W. W. H., & Weigel, D. (2014). Structural features determining flower-

promoting activity of Arabidopsis FLOWERING LOCUS T. *Plant Cell*, *26*(2), 552–564. https://doi.org/10.1105/tpc.113.115220
Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., Gao, G., (2014). GSDS 2.0: An upgraded gene feature visualization server. Bioinformatics, 31, 1296–1297.
Hwan Lee, J., Sook Chung, K., Kim, S.K., and Ahn, J.H. (2014). Post-translational regulation of SHORT VEGETATIVE PHASE as a major mechanism for thermoregulation of flowering. Plant Signal Behav 9, e28193.
Immink, R. G., Posé, D., Ferrario, S., Ott, F., Kaufmann, K., Valentim, F. L., ... & Angenent, G. C. (2012). Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. *Plant physiology*, *160*(1), 433-449
Jeong, H. L., Seong, J. Y., Soo, H. P., Hwang, I., Jong, S. L., & Ji, H. A. (2007). Role of SVP in the control of flowering time by ambient temperature in Arabidopsis. *Genes and Development*, *21*(4), 397–402. https://doi.org/10.1101/gad.1518407
Jin, H., Tang, X., Xing, M., Zhu, H., Sui, J., Cai, C., & Li, S. (2019). Molecular and transcriptional characterization of phosphatidyl ethanolamine-binding proteins in wild peanuts Arachis duranensis and Arachis ipaensis. *BMC Plant Biology*, *19*(1), 1–16. https://doi.org/10.1186/s12870-019-2113-3
Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275-282.
Juarez-Escobar, J., Guerrero-Analco, J. A., Zamora-Briseño, J. A., Elizalde-Contreras, J. M., Bautista-Valle, M. V., Bojórquez-Velázquez, E., Loyola-Vargas, V. M., Mata-Rosas, M., & Ruíz-May, E. (2021). Tissue-specific proteome characterization of avocado seed during postharvest shelf life. *Journal of Proteomics*, *235*(October 2020), 104112. https://doi.org/10.1016/j.jprot.2021.104112
Karlgren, A., Gyllenstrand, N., Källman, T., Sundström, J. F., Moore, D., Lascoux, M., & Lagercrantz, U. (2011). Evolution of the PEBP gene family in plants: Functional diversification in seed plant evolution. *Plant Physiology*, *156*(4), 1967–1977. https://doi.org/10.1104/pp.111.176206
Kikuchi, R., Kawahigashi, H., Ando, T., Tonooka, T., & Handa, H. (2009). Molecular and functional characterization of pebp genes in barley reveal the diversification of their roles in flowering. *Plant Physiology*, *149*(3), 1341–1353. https://doi.org/10.1104/pp.108.132134
Kim, D. H., Doyle, M. R., Sung, S., & Amasino, R. M. (2009). Vernalization: winter and the timing of flowering in plants. *Annual Review of Cell and Developmental*, *25*, 277-299.
Kim DH, Sung S. 2014. Genetic and epigenetic mechanisms underlying vernalization. Arabidopsis Book 12, e0171.
Kinoshita T., Ono N., Hayashi Y., Morimoto S., Naka- mura S., Soda M., Kato Y., Ohnishi M., Nakano T., Inoue S.I., Shimazaki K.I. (2011) FLOWERING LOCUS T regulates stomatal opening. Current Biol- ogy, 21, 1232–1238.
Klintenäs, M., Pin, P. A., Benlloch, R., Ingvarsson, P. K., & Nilsson, O. (2012). Analysis of conifer FLOWERING LOCUS T/TERMINAL FLOWER1-like genes provides evidence for dramatic biochemical evolution in the angiosperm FT lineage. *New Phytologist*, *196*(4), 1260–1273. https://doi.org/10.1111/j.1469-

8137.2012.04332.x
Kobayashi, Y., & Weigel, D. (2007). Move on up, it's time for change—mobile signals controlling photoperiod-dependent flowering. *Genes & development*, *21*(19), 2371-2384
Kralemann, L. E. M., Scalone, R., Andersson, L., & Hennig, L. (2018). North European invasion by common ragweed is associated with early flowering and dominant changes in FT/TFL1 expression. *Journal of Experimental Botany*, *69*(10), 2647–2658. https://doi.org/10.1093/jxb/ery100
Kumimoto R., Adam L., Hymus G., Repetti P., Reuber T., Marion C., Hempel F. & Ratcliffe O. (2008) The Nuclear Factor Y sub- units NF-YB2 and NF-YB3 play additive roles in the promotion of flowering by inductive long-day photoperiods in Arabidopsis. Planta 228, 709–723.
Kumimoto R.W., Zhang Y., Siefers N. & Holt B.F. (2010) NF-YC3, NF-YC4 and NF-YC9 are required for CONSTANS-mediated, photoperiod-dependent flowering in Arabidopsis thaliana.
Komiya R, Ikegami A, Tamaki S, Yokoi S, Shimamoto K. (2008) Hd3a and RFT1 are essential for flowering in rice. Development 135: 767–774.
Lahav, E., & Lavi, U. (2009). Avocado genetics and breeding. In *Breeding plantation tree crops: tropical species* (pp. 247-285). Springer, New York, NY.
Langridge, J. (1957). Effect of day-length and gibberellic acid on flowering of Arabidopsis. Nature 180, 36–37.
Lavi, U., Lahav, E., Degani, C., & Gazit, S. (1992). The Genetics of the Juvenile Phase in Avocado and its Application for Breeding. *Journal of the American Society for Horticultural Science*, *117*(6), 981–984. https://doi.org/10.21273/jashs.117.6.981
Li, L., Li, X., Liu, Y., & Liu, H. (2016). Flowering responses to light and temperature. *Science China Life Sciences*, *59*(4), 403–408. https://doi.org/10.1007/s11427-015-4910-8
Lin, T., Chen, Q. X., Wichenheiser, R. Z., & Song, G. qing. (2019). Constitutive expression of a blueberry FLOWERING LOCUS T gene hastens petunia plant flowering. *Scientia Horticulturae*, *253*(May), 376–381. https://doi.org/10.1016/j.scienta.2019.04.051
Liu, Y. Y., Yang, K. Z., Wei, X. X., & Wang, X. Q. (2016). Revisiting the phosphatidylethanolamine-binding protein (PEBP) gene family reveals cryptic FLOWERING LOCUS T gene homologs in gymnosperms and sheds new light on functional evolution. *New Phytologist*, *212*(3), 730–744. https://doi.org/10.1111/nph.14066
Lu, Y., Chen, W., Zhao, L., Yao, J., Li, Y., Yang, W., Liu, Z., Zhang, Y., & Sun, J. (2019). Different divergence events for three pairs of PEBPs in Gossypium as implied by evolutionary analysis. *Genes and Genomics*, *41*(4), 445–458. https://doi.org/10.1007/s13258-018-0775-0
Luo, Z. K., Chen, Q. F., Qu, X., & Zhou, X. Y. (2019). The roles and signaling pathways of phosphatidylethanolamine-binding protein 4 in tumors. In *OncoTargets and Therapy* (Vol. 12, pp. 7685–7690). Dove Medical Press Ltd. https://doi.org/10.2147/OTT.S216161
Ma, W., Noble, W.S., Bailey, T.L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP, Nat. Protoc. 9 1428–1450.

Mackenzie, K. K., Coelho, L. L., Lütken, H., & Müller, R. (2019). Phylogenomic analysis of the PEBP gene family from Kalanchoë. *Agronomy*, *9*(4), 1–16. https://doi.org/10.3390/agronomy9040171

Mimida, N., Goto, K., Kobayashi, Y., Araki, T., Ahn, J. H., Weigel, D., Murata, M., Motoyoshi, F., & Sakamoto, W. (2001). Functional divergence of the TFL1-like gene family in Arabidopsis revealed by characterization of a novel homologue. *Genes to Cells*, *6*(4), 327–336. https://doi.org/10.1046/j.1365-2443.2001.00425.x

Nakamura, Y., Andrés, F., Kanehara, K., Liu, Y.-c., Dörmann, P. & Coupland, G. (2014). Arabidopsis florigen FT binds to diurnally oscillating phospholipids that accelerate flowering. Nature communications, 5, p. 3553.

Nakano, Y., Higuchi, Y., Yoshida, Y., & Hisamatsu, T. (2015). Environmental responses of the FT/TFL1 gene family and their involvement in flower induction in Fragaria×ananassa. *Journal of Plant Physiology*, *177*, 60–66. https://doi.org/10.1016/j.jplph.2015.01.007

Nakamura, Y., Lin, Y. C., Watanabe, S., Liu, Y. chi, Katsuyama, K., Kanehara, K., & Inaba, K. (2019). High-Resolution Crystal Structure of Arabidopsis FLOWERING LOCUS T Illuminates Its Phospholipid-Binding Site in Flowering. *IScience*, *21*, 577–586. https://doi.org/10.1016/j.isci.2019.10.045

Ortega Tovar, M. Á. (2003). *Valor nutrimental de la pulpa fresca de aguacate Hass*. 741–748.

Park J, Nguyen KT, Park E, Jeon J-S, Choi G. (2013). DELLA proteins and their interacting RING finger proteins repress gibberellin responses by binding to the promoters of a subset of gibberellin-responsive genes in Arabidopsis. Plant Cell, 25:927–943.

Park, S. J., Jiang, K., Tal, L., Yichie, Y., Gar, O., Zamir, D., Eshed, Y., & Lippman, Z. B. (2014). Optimization of crop productivity in tomato using induced mutations in the florigen pathway. *Nature Genetics*, *46*(12), 1337–1342. https://doi.org/10.1038/ng.3131

Pasriga, R., Yoon, J., Cho, L. H., & An, G. (2019). Overexpression of RICE FLOWERING LOCUS T 1 (RFT1) Induces Extremely Early Flowering in Rice. *Molecules and Cells*, *42*(5), 406–417. https://doi.org/10.14348/molcells.2019.0009

Pin, P. A., Benlloch, R., Bonnet, D., Wremerth-Weich, E., Kraft, T., Gielen, J. J. L., et al. (2010). An antagonistic pair of FT homologs mediates the control of flowering time in sugar beet. Science 330, 1397–1400. doi: 10.1126/science.1197004

Pin, P. A., & Nilsson, O. (2012). The multifaceted roles of FLOWERING LOCUS T in plant development. *Plant, Cell and Environment*, *35*(10), 1742–1755. https://doi.org/10.1111/j.1365-3040.2012.02558.x

Rendón-Anaya, M., Ibarra-Laclette, E., Méndez-Bravo, A., Lan, T., Zheng, C., Carretero-Paulet, L., Perez-Torres, C. A., Chacón-López, A., Hernandez-Guzmán, G., Chang, T. H., Farr, K. M., Brad Barbazuk, W., Chamala, S., Mutwil, M., Shivhare, D., Alvarez-Ponce, D., Mitter, N., Hayward, A., Fletcher, S., … Herrera-Estrella, L. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National*

Academy of Sciences of the United States of America, *116*(34), 17081–17089. https://doi.org/10.1073/pnas.1822129116
Rice, P. Longden, I. and Bleasby, A. (2000). *EMBOSS: The European Molecular Biology Open Software Suite*. Trends in Genetics 16, (6) pp276—277.
Robledo, J. M., Thompson, A. J., Zsögön, A., Medeiros, D., Peres, L. E. P., & Araújo, W. L. (2019). *Control of water - use efficiency by florigen*. *October*, 1–11. https://doi.org/10.1111/pce.13664.
Rodwell, V. W., Bender, D. A., Botham, K. M., Kennelly, P. J., & Weil, P. A. (2018). *Harper's illustrated biochemistry*. New York (NY): McGraw-Hill Education.
Salazar-García, S., Ibarra-Estrada, M. E., & González-Valdivia, J. (2018). Phenology of "Méndez" avocado in Southern Jalisco, México. *Agrociencia*, *52*(7), 991–1003.
Sawa, M., Nusinow, D.A., Kay, S.A., and Imaizumi, T. (2007). FKF1 and GIGANTEA complex formation is required for day-length meas- urement in Arabidopsis. Science 318, 261–265.
Selzer, P. M., Marhöfer, R. J., & Koch, O. (2018). Sequence Comparisons and Sequence-Based Database Searches. In *Applied Bioinformatics* (pp. 35-50). Springer, Cham.
Shalit A, Rozman A, Goldshmidt A, Alvarez JP, Bowman JL, Eshed Y, Lifschitz E. 2009. The flowering hormone florigen functions as a general systemic regulator of growth and termination. Proceed- ings of the National Academy of Sciences USA 106: 8392–839
Shi, L., Group, P. S., & Group, P. (2018). *Research on the allelic variance of FT-like PEBP proteins in Solanum tuberosum*.
Slot, M., & Winter, K. (2016). The effects of rising temperature on the ecophysiology of tropical forest trees. In *Tropical Tree Physiology* (pp. 385-412). Springer, Cham.
Solovyev, V., & Director of Bioinformatics. (2007). Statistical approaches in eukaryotic gene prediction. *Handbook of statistical genetics*.
Song, J., Irwin, J., & Dean, C. (2013). Remembering the prolonged cold of winter. In *Current Biology* (Vol. 23, Issue 17, pp. R807–R811). Cell Press. https://doi.org/10.1016/j.cub.2013.07.027
Spanudakis, E., & Jackson, S. (2014). *The role of microRNAs in the control of flowering time*. *65*(2), 365–380. https://doi.org/10.1093/jxb/ert453
Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
Tang, M., Dong, Z., Guo, P., Zhang, Y., Zhang, X., Guo, K., An, L., Liu, X., & Zhao, P. (2019). Functional analysis and characterization of antimicrobial phosphatidylethanolamine-binding protein BmPEBP in the silkworm Bombyx mori. *Insect Biochemistry and Molecular Biology*, *110*, 1–9. https://doi.org/10.1016/j.ibmb.2019.03.011
Taoka, K. I., Ohki, I., Tsuji, H., Furuita, K., Hayashi, K., Yanase, T., Yamaguchi, M., Nakashima, C., Purwestri, Y. A., Tamaki, S., Ogaki, Y., Shimada, C., Nakagawa, A., Kojima, C., & Shimamoto, K. (2011). 14-3-3 proteins act as intracellular receptors for rice Hd3a florigen. *Nature*, *476*(7360), 332–335.

https://doi.org/10.1038/nature10272
Teper-Bamnolker, P., & Samach, A. (2005). The flowering integrator FT regulates SEPALLATA3 and FRUITFULL accumulation in Arabidopsis leaves. *The Plant Cell*, *17*(10), 2661-2675
Thakare, D., Kumudini, S., & Dinkins, R. D. (2011). The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in Arabidopsis. *Planta*, *234*(5), 933–943. https://doi.org/10.1007/s00425-011-1450-8.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence align- ment aided by quality analysis tools. Nucleic Acids Res 25, 4876–4882.
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
Tribhuvan, K. U., Das, A., Srivastava, H., Kumar, K., Durgesh, K., Sandhya, Mithra, S. V. A., Jain, P. K., & Gaikwad, K. (2020). Identification and characterization of PEBP family genes reveal CcFT8 a probable candidate for photoperiod insensitivity in C. cajan. *3 Biotech*, *10*(5), 1–12. https://doi.org/10.1007/s13205-020-02180-x
Turner, B. L. II, and C.H. Miksiek. 1984. Economic plant species associated with prehistoric agriculture in the Maya lowlands. Economic Botany 38(2): 179-173.
Tzfira, T., Zuker, A., & Altman, A. (1998). genetic transformation and its application to future forests. *Tibtech*, *16*(3), 439–446.
Wang, G., Wang, P., Gao, Y., Li, Y., Wu, L., Gao, J., Zhao, M., & Xia, Q. (2018). Isolation and functional characterization of a novel FLOWERING LOCUS T homolog (NtFT5) in Nicotiana tabacum. *Journal of Plant Physiology*, *231*, 393–401. https://doi.org/10.1016/j.jplph.2018.10.021
Wang, M., Tan, Y., Cai, C., & Zhang, B. (2019). Identification and expression analysis of phosphatidy ethanolamine-binding protein (PEBP) gene family in cotton. *Genomics*, *111*(6), 1373–1380. https://doi.org/10.1016/j.ygeno.2018.09.009
Wang, Zhen, Yang, R., Devisetty, U. K., Maloof, J. N., Zuo, Y., Li, J., Shen, Y., Zhao, J., Bao, M., & Ning, G. (2017). The divergence of flowering time modulated by FT/TFL1 is independent to their interaction and binding activities. *Frontiers in Plant Science*, *8*(May), 1–16. https://doi.org/10.3389/fpls.2017.00697
Wang, Zheng, Zhou, Z., Liu, Y., Liua, T., Li, Q., Ji, Y., Li, C., Fang, C., Wang, M., Wu, M., Shen, Y., Tang, T., Jianxin, M., & Tian, Z. (2015). Functional evolution of phosphatidylethanolamine binding proteins in soybean and arabidopsis. *Plant Cell*, *27*(2), 323–336. https://doi.org/10.1105/tpc.114.135103
Wenkel S., Turck F., Singer K., Gissot L., Le Gourrierec J., Samach A. & Coupland G. (2006) CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of Arabidopsis. The Plant Cell 18, 2971–2984
Wigge, P. A., Kim, M. C., Jaeger, K. E., Busch, W., Schmid, M., Lohmann, J. U., &

Weigel, D. (2005). Integration of spatial and temporal information during floral induction in Arabidopsis. *Science*, *309*(5737), 1056–1059. https://doi.org/10.1126/science.1114358
Wigge P.A. (2011) FT, A mobile developmental signal in plants. Current Biology 21, 374–378.
Wilkie JD, Sedgley M, Olesen T. (2008). Regulation of floral initiation in horticultural trees. Journal of Experimental Botany 59, 3215–3228.
Whelan S, Morrison DA. (2017). Inferring trees. Methods Mol Biol. 1525:349–377
Wu, G., Park, M. Y., Conway, S. R., Wang, J. W., Weigel, D., and Poethig, R. S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. Cell 138, 750–759. doi: 10.1016/j.cell.2009.06.031
Wu, L., Li, F., Deng, Q., Zhang, S., Zhou, Q., Chen, F., Liu, B., Bao, M., & Liu, G. (2019). Identification and Characterization of the FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family in Petunia. *DNA and Cell Biology*, *38*(9), 982–995. https://doi.org/10.1089/dna.2019.4720
Xi, W., Liu, C., Hou, X., & Yu, H. (2010). MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in Arabidopsis. *The Plant Cell*, *22*(6), 1733-1748.
Yan, Y., Shen, L., Chen, Y., Bao, S., Thong, Z., and Yu, H. (2014). A MYB-domain protein EFM mediates flowering responses to envi- ronmental cues in Arabidopsis. Dev Cell 30: 437–448.
Yang, Y., Klejnot, J., Yu, X., Liu, X., & Lin, C. (2007). *Florigen (II): It is a Mobile Protein*. *49*(12), 1665–1669. https://doi.org/10.1111/j.1744-7909.2007.00614.x
Yang, Z., Chen, L., Kohnen, M. V., Xiong, B., Zhen, X., Liao, J., Oka, Y., Zhu, Q., Gu, L., Lin, C., & Liu, B. (2019). Identification and Characterization of the PEBP Family Genes in Moso Bamboo (Phyllostachys heterocycla). *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-51278-7
Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, *18*(6), 292-298.
Zhang J-Z, Li Z-M, Mei L, Yao J-L, Hu C-G. (2009). PtFLC homolog from trifoliate orange (Poncirus trifoliata) is regulated by alternative splicing and experiences seasonal fluctuation in expression level. Planta, 229:847–859.
Zhang, X., Wang, C., Pang, C., Wei, H., Wang, H., Song, M., Fan, S., & Yu, S. (2016). Characterization and functional analysis of PEBP Family genes in upland cotton (Gossypium hirsutum L.). *PLoS ONE*, *11*(8), 1–20. https://doi.org/10.1371/journal.pone.0161080
Zhao, S., Wei, Y., Pang, H., Xu, J., Li, Y., Zhang, H., Zhang, J., & Zhang, Y. (2020). Genome-wide identification of the PEBP genes in pears and the putative role of PbFT in flower bud differentiation. *PeerJ*, *2020*(4), 1–19. https://doi.org/10.7717/peerj.8928
Zheng, X. M., Wu, F. Q., Zhang, X., Lin, Q. B., Wang, J., Guo, X. P., Lei, C. L., Cheng, Z. J., Zou, C., & Wan, J. M. (2016). Evolution of the PEBP gene family and selective signature on FT-like clade. *Journal of Systematics and Evolution*, *54*(5), 502–510. https://doi.org/10.1111/jse.12199
Zhu, Y., Liu, L., Shen, L., & Yu, H. (2016). NaKR1 regulates long-distance movement of FLOWERING LOCUS T in Arabidopsis. *Nature Plants*, *2*(June). https://doi.org/10.1038/nplants.2016.75

Ziv, D., Zviran, T., Zezak, O., Samach, A., & Irihimovitch, V. (2014). Expression profiling of FLOWERING LOCUS T-like gene in alternate bearing “Hass” avocado trees suggests a role for PaFT in avocado flower induction. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0110613