POLYPLOIDIZATION EVENTS SHAPED THE TRANSCRIPTION FACTOR REPERTOIRES IN LEGUMES (FABACEAE)

KANHU CHARAN MOHARANA

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO – UENF

CAMPOS DOS GOYTACAZES/RJ
FEVEREIRO-2020

POLYPLOIDIZATION EVENTS SHAPED THE TRANSCRIPTION FACTOR REPERTOIRES IN LEGUMES (FABACEAE).

KANHU CHARAN MOHARANA

Thesis submitted to the Centro de Biociências e Biotecnologia, da Universidade Estadual do Norte Fluminense, as partial fulfillment of requirements for obtaining the degree of Doctor in Biosciences and Biotechnology.

Advisor: Dr. Thiago Motta Venancio.

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE
DARCY RIBEIRO – UENF

CAMPOS DOS GOYTACAZES/RJ
FEVEREIRO-2020

FICHA CATALOGRÁFICA
UENF - Bibliotecas
Elaborada com os dados fornecidos pelo autor.

M697 Moharana, Kanhu Charan.

Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae). / Kanhu Charan Moharana. - Campos dos Goytacazes, RJ, 2020.

85 f. : il.
Inclui bibliografia.

Tese (Doutorado em Biociências e Biotecnologia) - Universidade Estadual do Norte Fluminense Darcy Ribeiro, Centro de Biociências e Biotecnologia, 2020.
Orientador: Thiago Motta Venancio.

1. whole genome duplication. 2. phylogenomics. 3. segmental duplication. 4. nodulation. 5. Soybean. I. Universidade Estadual do Norte Fluminense Darcy Ribeiro. II. Título.

CDD - 570

POLYPLOIDIZATION EVENTS SHAPED THE TRANSCRIPTION FACTOR REPERTOIRES IN LEGUMES (FABACEAE)

KANHU CHARAN MOHARANA

Thesis presented to the Centro de Biociências e Biotecnologia, da Universidade Estadual do Norte Fluminense, as partial fulfillment of requirements for obtaining the degree of Doctor in Biosciences and Biotechnology.
Advisor: Dr. Thiago Motta Venancio.

Date of approval:17/02/2020

Examination Committee:

Dr. Thiago Motta Venancio – UENF

Dr. Vanildo Silveira – UENF

Gustavo L. Rezende
Dr. Gustavo Lazzaro Rezende – UENF

Dr. Rodrigo Nunes da Fonseca - UFRJ

CAMPOS DOS GOYTACÀZES/RJ
FEVEREIRO-2020

ACKNOWLEDGEMENTS

I would like to express my thankfulness to everyone who has helped me personally and professionally during the last four years of my life in successfully completing my Ph.D.

Firstly, I would like to thank my advisor, Professor Thiago Motta Venancio for giving me opportunity to pursue my doctoral research in his lab. His experienced opinion and sound advice have helped me in shaping my research to its successful conclusion. He has also provided me with academic freedom and opportunities to gain experience from various collaborative projects.

Besides my advisor, I would like to thank Prof. Katia Valevski Sales Fernandes, Prof. Gonçalo Souza Filho, Prof. Gustavo Lazzaro Rezende and Prof. Clicia Grativol Gaspar, for their insightful comments and encouragement which incented me to widen my research from various perspectives.

I am grateful to my lab colleagues - both past and present: Francisnei Pedrosa Da Silva, Dr. Rajesh Gazara, Fabrício Brum, Fabrício Almeida, Dr. Filipe Matteoli, Hemanoel Passarelli, Késia Dias, Dayana Kelly, Dr. Lupis Ribeiro Gomes Neto and Daniella Canedo for being there in my hard times and including me in their social circle. I also owe thanks to, Kayan Ventury, for those small corridor talks.

I am very grateful to UENF for this opportunity to enroll in the PhD program. Many thanks to those working in PGBB program, especially Prof. Arnoldo Rocha Façanha, for being supportive in furnishing all required official documents. I gratefully acknowledge CAPES, FAPERJ and CNPq for providing financial support for my research.

Without the support and encouragement of my family and relatives I would never have started down this path. I would like to thank my sister Itishree Maharana and brother Subhendu Maharana for always being there for me, without which it would have been difficult to live in a foreign country. I express a huge thank you to my parents for providing me every opportunity to pursue my interests, for their sacrifices, enduring encouragement, and their unconditional faith in my abilities.

RESUMO

Leguminosas (Fabaceae) são ricas fontes de proteína e óleo, com grande impacto na economia de vários países. Recentemente, muitas leguminosas, principalmente soja, são alvos de pesquisas genômicas. Os fatores de transcrição (*transcription factors*, TF) são essenciais para o crescimento e desenvolvimento adequados das plantas. Estudos de sequenciamento e associação de genoma identificaram vários TFs envolvidos em várias características agronômicas. Aqui relatamos uma análise filogenômica das principais famílias de TF em leguminosas e sua potencial associação com características importantes, como fixação de nitrogênio e desenvolvimento de sementes. Utilizamos domínios de ligação ao DNA dos TFs para rastrear sistematicamente os genomas de 15 espécies de leguminosas e 5 de não-leguminosas. A porcentagem de TFs variou de 3-8% dos complementos gênicos. Grupos ortólogos de TFs (OG) em espécies existentes foram usados para estimar o tamanho dos OG nos nós ancestrais usando um modelo genético de nascimento e morte, o que nos permitiu identificar expansões específicas de certas linhagens. Juntas, a análise de OG e a taxa de substituições sinônimas (Ks) entre pares de genes mostram que as principais expansões de TFs estão fortemente associadas a eventos conhecidos de duplicação de genoma inteiro (WGD) nas linhagens de leguminosas (~ 58 million years ago) e *Glycine sp.* (~ 13 mya), que representam uma grande fração do repertório de TFs de *Glycine max e Phaseolus vulgaris*. Dos 3407 TFs de *Gl. max*, 1808 e 676 podem ser rastreados até um único homeólogo em *Ph. vulgaris* e *Vitis vinifera*, respectivamente. Encontramos uma tendência para que os TFs expandidos nas leguminosas sejam transcritos preferencialmente nos nódulos, sugerindo seu recrutamento no início da evolução da nodulação no clado das leguminosas. Também encontramos expansões de TFs na duplicação de genoma inteiro de *Glycine sp*., que foram seguidas por perda de genes na soja selvagem *Gl. soja*, incluindo alguns genes localizados em importantes loci de herança quantitativa (largura, comprimento e área foliar, densidade foliar, densidade de ramos, conteúdo de ácidos graxos e número de sementes por vagem). Juntas, nossas descobertas sugerem fortemente os papéis das duas duplicações de genoma inteiro na formação dos repertórios de TF nas linhagens das leguminosas e *Glycine sp.*, com implicações importantes para entender aspectos básicos da biologia das leguminosas e da soja.

ABSTRACT

Legumes (Fabaceae) are rich sources of protein and oil, with great impact in the economy of several countries. Recently many legumes, particularly soybean, are targets of genomic research. Transcription factors (TF) are essential for proper plant growth and development. Genome resequencing and association studies have pinpointed several TFs involved in several agronomic traits. Here we report a phylogenomic analysis of major TF families in legumes and their potential association with important traits such as nitrogen fixation and seed development. We used TF DNA-binding domains to systematically screen the genomes of 15 legume and 5 non-legume species. The percentage of TFs ranged from 3-8% of the gene complements. TF orthologous groups (OG) in extant species were used to estimate OG sizes in ancestor nodes using a gene birth-death model, which allowed us to identify lineage-specific expansions. Together, OG analysis and rate of synonymous substitutions (Ks) between gene pairs show that major TF expansions are strongly associated with known whole-genome duplication (WGD) events in the legume (~58 million years ago) and *Glycine* (~13 mya) lineages, which account for a large fraction of the *Phaseolus vulgaris* and *Glycine max* TF repertoires. Out of the 3,407 *Gl. max* TFs, 1,808 and 676 can be traced back to a single homeolog in *Ph. vulgaris* and *Vitis vinifera*, respectively. We found a trend for TFs expanded in legumes to be preferentially transcribed in nodules, suggesting their recruitment early in the evolution of nodulation in the legume clade. We also found TF expansions in the *Glycine* WGD that were followed by gene loss in the wild soybean *Glycine soja,* including some genes located within important quantitative trait loci (leaf width, leaf length, leaf area, branch density, fatty acid content, and number of seeds per pod). Together, our findings strongly support the roles of two WGDs in shaping the TF repertoires in the legume and *Glycine* lineages, with important implications to understand basic aspects of legume and soybean biology.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TF	: Transcription factors
DBD	: DNA binding domains
GBH	: Gene dosage balance theory
WGD	: Whole genome duplication
WGT	: Whole genome triplication
MYA	: Million years ago
HMM	: Hidden Markov model
SD	: Segmental duplicates
TD	: Tandem duplicates
PD	: Proximal duplicates
rTE	: Retro-transposons mediated duplication
dTE	: DNA-transposon mediated duplication (dTE)
DD	: Dispersed duplicates (DD)
Ka	: Non- synonymous mutation rate
Ks	: Synonymous mutation rate
BLAST	: Basic Local Alignment Tool
AP2	: APETALA 2
ERF	: Ethylene Responsive Factor
RAV	: Related to ABI3/VP1
ARF	: Auxin response factor
BBR-BPC	: Barley B Recombinant (BBR) - BASIC PENTACYSTEINE1 (BPC1)
BES1	: BRI1-EMS-SUPPRESSOR
bHLH	: Basic helix loop helix
bZIP	: Basic leucine zipper
Dof	: DNA binding with one finger
CO-like	: CONSTANS-like
LSD	: LESION SIMULATING DISEASE 1 (LSD1)
C2H2	: CCHH (Zn)

C3H	: CCCH (Zn)
CAMTA	: Calmodulin binding transcription factors
CPP	: Cystein-rich polycomb-like protein
DBB	: Double B-box zinc finger
E2F/DP	: E2 factor protein and DP protein
EIL	: Ethylene-Insentive3 (EIN3)-like protein3 (EIL3)
FAR1	: FAR-RED IMPAIRED RESPONSE1
LFY	: LEAFY
G2-like	: Golden2 (G2)-like
ARR-B	: Type-B phospho-accepting response regulator
GeBP	: GLABROUS1 enhancer-binding protein
GRAS	: GAI, RGA, and SCR
GRF	: GROWTH-REGULATING FACTOR
TALE	: Three Amino acid Loop Extension
WOX	: WUS homeobox-containing protein family
HB	: Homeobox
HB-PHD	: HB-PHD finger
HB-other	: HB-other
HRT-like	: Hairy-Related transcription-factor-like
HSF	: Heat shock factor
LBD	: ASYMMETRIC LEAVES2/LATERAL ORGAN BOUNDARIES
M-type	: MADS-type I
MIKC	: MADS-type II
MYB	: Myb proto-oncogene protein
NAC	: NAM, ATAF1, 2 and CUC2
NF-X1	: Nuclear factor, X-box binding 1
NF-YA	: Nuclear factor Y subunit A
NF-YB	: Nuclear factor Y subunit B
NF-YC	: Nuclear factor Y subunit C
Nin-like	: NODULE INCEPTION

NZZ/SPL	: SPOROCYTELESS/NOZZLE
S1Fa-like	: S1Fa-like
TCP	: TEOSINTE-LIKE1, CYCLOIDEA, and PROLIFERATING CELL FACTOR1
ZF-HD	: Zinc finger homeodomain protein
SBP	: SQUAMOSA promoter binding protein
SRS	: SHI RELATED SEQUENCE
SAP	: STERILE APETALA
STAT	: Signal Transducers and Activators of Transcription
VOZ	: Vascular plant One-Zinc finger

TABLE OF CONTENT

1. INTRODUCTION

After the Orchidaceae and Asteraceae, Legumes (Fabaceae) are the third largest Angiosperm (flowering plants) family, comprising nearly 20,000 species with tremendous morphological and ecological variations (Lewis, 2005), encompassing from arctic dwarf herbs to tropical forest trees. Members of Fabaceae are often identified by specialized fruit, referred as legume, which in turn have also given the original family name. Flower shapes in this family varies greatly and ranges from radially symmetric (actinomorphic) to bilaterally symmetric (zygomorphic) and asymmetric flowers. However, the classical butterfly-like (papilionoid) flower is consistently observed throughout the Papilionoideae sub-family. While some members of legumes grow as weeds in cereal agriculture, many others are major grain crops being source of protein and oil. In fact, legumes are second most agriculturally important family to human after grass family (Poaceae). Out of the six legume sub-families (i.e. Cercidoideae, Detarioideae, Duparquetioideae, Dialioideae, Caesalpinioideae, Faboideae (Papilionoideae), and Mimosoideae), Papilionoideae alone accounts for two-thirds of the total number of species, including many of them are economically important crops (Graham and Vance, 2003; Cardoso et al., 2012; Azani et al., 2017). These economically important crops are mostly cultivated for their grains, which are rich sources of dietary proteins; soybean (*Glycine max*) grains are also rich in lipids and are a major source of cooking oil and biofuels. Trends in international trade of pulses (edible legume seeds) are very exciting and have grown rapidly. The total exports of pulses, for instance, throughout the globe have doubled between 1990 and 2012 from 6.6 to 13.4 million tons (http:// www.fao.org/pulses-2016/en/).

Besides, being an important source of food and oil, cultivation of legumes contributes to sustainable agriculture. Many legume species are characterized with specialized tissue in roots called nodules, where certain types of bacteria (e.g. *Rhizobium*, *Bradyrhizobium*) reside in a symbiotic relationship. These bacteria fix atmospheric nitrogen (N_2), reducing or even eliminating the requirement for nitrogen fertilizers. This natural ability to recharge soil with nitrogen makes legumes an important part of the ecosystem and highlights the possible reason for their predominance in diverse habitats. For this reason, legumes are often rotated with cereal crops to boost soil fertility. According to one estimate, without such naturally fixed nitrogen by legumes, humans would require to consume 288 billion kilograms of additional fuel to chemically synthesize ammonia-based fertilizers (Young and Bharti, 2012).

1.1 Legume genomics

Genomics is the study of whole genomic content of an organism using advanced in molecular biology techniques and elements from classical genetics. Advent of genomics often considered to coincide with the conceptualization of human genome project in 1986. The primary aim of this project was to produce a finished DNA sequence of human genome which led to many technological improvements such as automated and low cost DNA sequencing and informatics algorithms to analyze the sequencing data. Thus broadly speaking heart of genomics lies in the combination of sequencing methods, and bioinformatics techniques to assemble the fragmented genome sequences, and analyze the structure and function of genomes.

Dramatic and rapid changes in climatic condition lead to extreme conditions like flood, drought and other situations that result in crop loss. Today it is a major challenge before plant scientists to combat such challenges to produce better crop varieties with greater crop production. Similar to how availability of human genome sequence and understanding of disease biology at molecular level, revolutionized traditional medicine and health care, availability of genome sequence of model plant species such as *Arabidopsis* has profound impact in crop improvement (Varshney et al., 2005; Hood and Rowen, 2013). But the question arises how sequencing genome of a plant could help in improving crop production? Many modern day crop plants are results of repeated rounds of selective domestication by early human civilizations. Traditionally crop scientists use breeding programs to improve cultivars. Later application of laws of genetics and quantitative genetics made breeding

programs closer to science. Crop breeding, generally, follows a lengthy cycle of steps:
i. evaluation of available crop varieties for desirable phenotypes (genetic diversity), ii. selection of superior phenotypes; iii. crossing selected plants to create a hybrid; and iv. back to evaluation step, then restarting the process. The resulting hybrid may possess the superior genotypes that can be tested and developed into varieties. As one can realize this process is lengthy and much more complex, since many types of phenotypes need to be evaluated (e.g. disease resistance, stress adaptation, yield, quality, etc.). Genomics based studies enables direct identification of key genes involved in important biochemical pathways and can be exploited to produce a superior plant.(Jackson et al., 2011)

Due their immense importance in ecology and agronomy many legumes are target of genomic research. To date, genome sequencing data of at least 15 legume species are publicly available, including wild and domesticated modern soybean (*Gl. soja* and *Gl. max*) (Kim et al., 2010; Schmutz et al., 2010; Xie et al., 2019), barrel clover (*Medicago truncatula*) (Young et al., 2011; Tang et al., 2014), bird’s-foot trefoil (*Lotus japonicas*)(Sato et al., 2008), Common Bean (*Ph. vulgaris*) (Schmutz et al., 2014), Chickpea (*Cicer reticulatum*, *Cicer arietinum*) (Varshney et al., 2013; Parween et al., 2015; Gupta et al., 2017), pigeonpea (*Cajanus cajan*) (Varshney et al., 2011), adzuki bean (*Vigna angularis*) (Kang et al., 2015; Yang et al., 2015), mung bean (*Vigna radiata*) (Kang et al., 2014), narrow-leaved Lupine (*Lupinus angustifolius*)(Hane et al., 2017), Chinese liquorice (*Glycyrrhiza uralensis*) (Mochida et al., 2017) and peanut (*Arachis duranensis* and *Arachis ipaensis*)(Bertioli et al., 2016; Chen et al., 2016). While all of the above mentioned genomes are from Papilionoideae subfamily, genomic data from other subfamilies is underrepresented. Recently, draft genomes from non- Papilionoideae species became available (e.g. *Chamaecrista fasciculata* and *Mimosa pudica*) (Griesmann et al., 2018). The availability of several legume genomes enables *in silico* comparative analyses and a more detailed investigation of legume specific traits.

1.1.1 Soybean genomics: challenges and opportunities

The modern soybean is the most economically important legume species, accounting for ~70% of dietary proteins and ~61% of total edible oil (http://soystats.com/, accessed on Feb 27, 2017). In addition, soybean oil is also used in the plastic and biofuel industries. A commercial grade fatty acid called lecithin, obtained from soybean oil, is also used in food and pharmaceutical industry as protective coating. Further, nitrogen fixation makes soybean an integral part of integrative agriculture. The increasing demand for food, animal feed and oil has increased soybean cultivation worldwide, sometimes at the unfortunate cost of deforestation. Nearly 6% of the total arable land is used for soybean agriculture and it is expected to increase in future (Hartman et al., 2011).

The genus *Glycine* is divided into two sub-genera, one of which includes the modern cultivar of soybean [*Gl. max* (L.) Merrill; n=20] and its wild-soybean progenitor [*Gl. soja* Sieb. & Zucc.; n=20]. It is currently accepted that the transition from wild to domesticated soybean took place in Central China between 5,000 and 9,000 years ago through a gradual process that involved an intermediary species, *Gl. gracilis* (Han et al., 2016; Sedivy et al., 2017). Other lines of evidence support independent domestication events in East Asia (Korea and Japan) (Zhou et al., 2015; Sedivy et al., 2017). Artificial selection during domestication involved several distinct traits, such as pod shattering, seed hardness, adaptation to different photoperiods, flowering time, and stress resistance (Zhou et al., 2015; Sedivy et al., 2017). Due to domestication and breeding, *Gl. max* has a very narrow genetic pool and is thus susceptible to many biotic and abiotic stresses that can affect yield (Chan et al., 2012). Traditionally, wild cultivars remain the main sources of novel genes in breeding programs. However, the advent of genomics and modern molecular biology techniques in breeding programs brought a revolution to the field. In 2010, genome sequencing in both modern and wild-soybean have been published (Kim et al., 2010; Schmutz et al., 2010), which were followed by many transcriptome studies (Libault et al., 2010a; Severin et al., 2010; Wang et al., 2014; Bellieny-Rabelo et al., 2016; Kim et al., 2016; Song et al., 2016). Later, several studies have been conducted to re-sequence modern- and wild-soybean cultivars (Lam et al., 2010; Zhou et al., 2015; Maldonado Dos Santos et al., 2016).

1.2 Complexity in angiosperm genomes and polyploidization events

The haploid set of chromosome possessed by an organism is referred as genome. Genome of *Ar. thaliana* with just five chromosomes was selected for genome sequences since it had smaller genome size than most plant genomes. In angiosperms the average genome size is nearly 6000 Mega basepairs (Mbp) per haploid genome which is approximately twice the size of the human genome.(Gregory et al., 2007). Genome sizes of many important crop plants, for example wheat with 15 Gbp, are even larger. One of the most striking observations among all angiosperms is that many of them have experienced one or more round of polyploidization at some point of their evolutionary history and these might be contributing to such expanded genome size (De Bodt et al., 2005; Van de Peer et al., 2009; Crepet, 2013). Polyploidization, also known as whole genome duplication (WGD), results in a sudden multiplication of a given genome, which generates a genomic shock and is often lethal. However, a few WGD events produce viable polyploid individuals. There are two different types of polyploidization: auto- and allo-polyploidization. Auto-polyploidy refers to multiplication of genetically identical chromosome sets within a single (sub) species, whereas allopolyploidy imply the multiplication of chromosome sets via hybridization of two genetically different (sub)species.

Comparative genomics studies revealed that most angiosperms species carry signatures of the ancient WGD events (often referred as paleopolyploidization) in their genomes. For example, detailed analysis of the model plant species *Ar. thaliana* genome sequence revealed three ancient polyploidization events (Bowers et al., 2003). The oldest round is termed as *gamma* (γ), which is shared by all eudicots (Bowers et al., 2003) and was also hypothesized to be shared with the ancestral of monocots (Simillion et al., 2002; Bowers et al., 2003). Apart from the gamma hexaploidy event, there are two lineage specific duplicate regions in the *Ar.*

thaliana genome (**Figure 1**), which are called *beta* (β) and *alpha* (α). With the publication of the grape (*Vitis vinifera*)(Jaillon et al., 2007) and papaya (*Carica papaya*)(Ming et al., 2008) genomes, the history of early angiosperm WGD events was better elucidated.

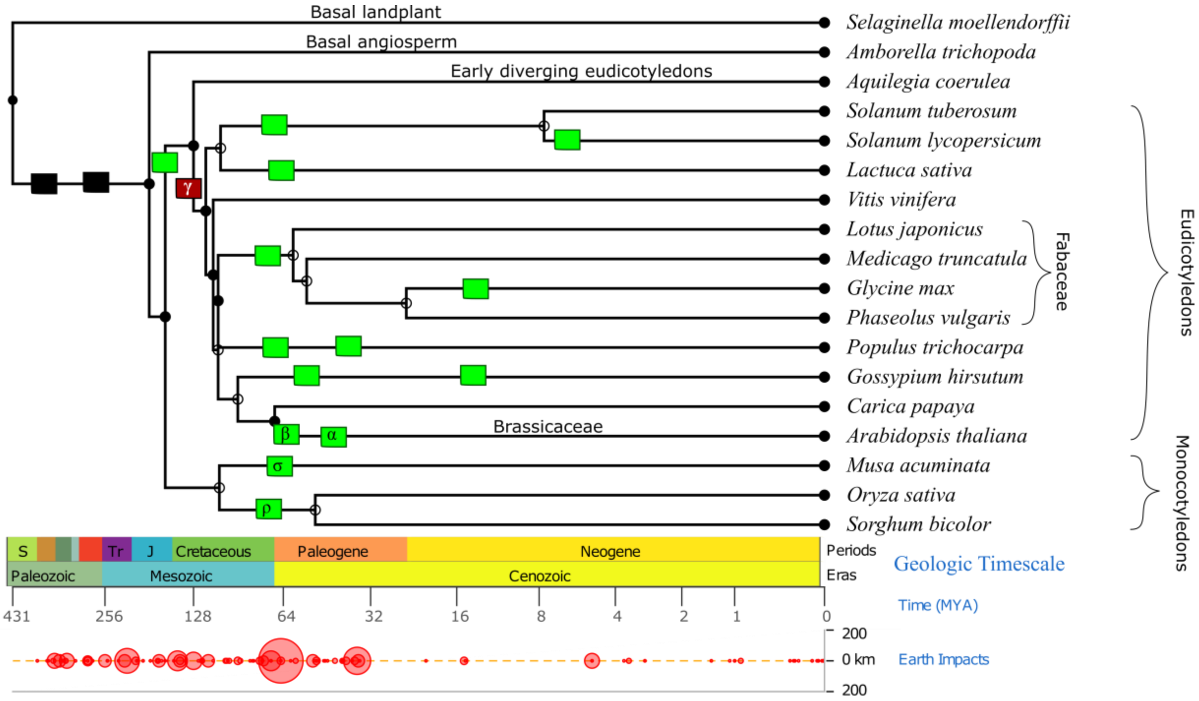

Figure 1: A simplified phylogenetic tree of several representative angiosperms showing occurrence of polyploidization events. Whole genome duplication (WGD, green box) and triplication (WGT, brown box) events along several plant lineages. The gamma WGT at the base of eudicots and the two consecutive WGDs in *Arabidopsis thaliana* have been labeled in Greek letters. Figure generated on www.timetree.org.

Monocotyledons are not devoid of ancient WGD events. By analyzing the available genome sequences of *Oryza sativa* and other grass genomes, an ancient ~70 mya WGD (i.e. *Rho* (ρ) duplication event) is shared by all monocots (Paterson et al., 2004). When the genomes of rice and sorghum were analyzed to reconstruct the gene order in their last common ancestor, an additional ancient WGD, referred as *sigma* (σ), was identified (Tang et al., 2010). Later, with the sequencing of the banana genome, it was inferred that *sigma* happened after the split between *Poales* (order of grasses) and *Zingiberales* (order of banana) (D’Hont et al., 2012). Further, by comparing such genomic regions with the grape genome, it was found that the *sigma* took place after the monocot-eudicot split (Tang et al., 2010). Similarly, more recent WGD events were also found in the banana lineage (D'Hont et al., 2012).

In addition to the WGD events shared by most angiosperms many plant lineages show traces of independent and recent genome duplication (Blanc and Wolfe, 2004b; Schlueter et al., 2004). Some of these most diverse and species rich clades, namely *Brassicaceae*, *Poaceae*, *Fabaceae*, *Solanaceae*, and *Asteraceae*, have all been suggested to have undergone WGD right before their diversification. However, the precise timing of such events remains unclear (Van de Peer et al., 2009). Interestingly many independent WGDs like those in Rice, *Medicago*, Tomato, *Lactuca sativa*, cotton, poplar, and banana happened ~60-70 mya (Tuskan et al., 2006; Fawcett et al., 2009). Recently, it has been suggested burst of WGD events is linked to the Cretaceous-Paleogene (K-Pg) (also called Cretaceous-Tertiary; KT) mass extinction event (**Figure 1**), which is described as the most recent large-scale mass extinction of plant and animal species (Vanneste et al., 2014).

The phenotypic changes due to change in of genome size should be studied in the context of increase in gene copy. A gene is a segment of DNA that carries information to code for a protein or RNA. Upon duplication of a gene it will produce 2x amount of gene product affecting the external phenotype of the organism. In addition to large scale duplication events due to polyploidization, many small scale events had also contributed to the emergence of multiple set of gene copies (Cannon et al., 2004). Such small-scale gene duplication events include tandem and proximal duplicates which typically arise through unequal crossing over of homologous chromosomes or localized transposon activities (Proulx et al., 2011). The relative contribution of such duplication modes is currently a subject of intense research (Panchy et al., 2016).

1.3 Consequences of wgd on the diversification of duplicate genes

Plant polyploidy has been proposed as a mean of speciation since 1901, as initially proposed by Hugo de Vries on *Oenothera lamarkiana* mut (Nei and Nozawa, 2011). There are two contrasting views regarding consequences of polyploidy genomes: 1) although WGD events are frequent in plants, they are unimportant and have no long term beneficial effect to the main process of evolution (Stebbins Jr, 1950); 2) WGD events provide the raw material for the creating novel phenotypes (Ohno, 1970; Levin, 1983). However, with availability of complete genomes, it is now widely accepted that WGD events have huge adaptive potential and is associated with the huge success of angiosperms and their explosive radiation in a short period of time (Crepet, 2013), although some recent reports remain questioning this hypothesis (Dodsworth et al., 2016; Kellogg, 2016).

Upon a “successful” WGD, there are massive rounds of gene loss and genome rearrangements in a process called *diploidization* (Blanc et al., 2000). It has been suggested that selective pressure for homologous chromosome pairing during mitotic cell division promotes ploidy reduction by the elimination of chromosomal segments and creation of neo-chromosomes (Albalat and Cañestro, 2016). Such substantial and immediate genomic rearrangements play an important role in speciation.

Whole genomes from several plant species sharing polyploidization events are often compared to reconstruct ancestral chromosomes and identify genomic rearrangements (Endress and Doyle, 2009). Homologous chromosome segments with similar gene orders (one-to-one relationship) are called *syntenic blocks* (Tang et al., 2008). These syntenic blocks from closely related species can be compared using bioinformatics approaches, for example, to study species divergence, as done in Brassicaceae (Murat et al., 2015) and Poaceae (Murat et al., 2010). Recently, the comparison of 10 legume genomes uncovered several large scale evolutionary events (Lee et al., 2017).

1.3.1 Fate of redundant gene copies

Gene loss through gradual diploidization is another important post-polyploidization process. At time zero, duplication events create two redundant copies of a gene, which may remain as redundant copy or may have three possible fates (**Figure 2**)(Sémon and Wolfe, 2007): (i) one of the copies loses its functionality (i.e. pseudogenization); (ii) both copies accumulate mutations and become subfunctionalized, with two genes being required to perform the functions of a single original gene, or both copies performing the same functions in different sites, times or conditions and; (iii) one of the copies may evolve a new function by neofunctionalization (Wolfe, 2001; Thomas, 2006).

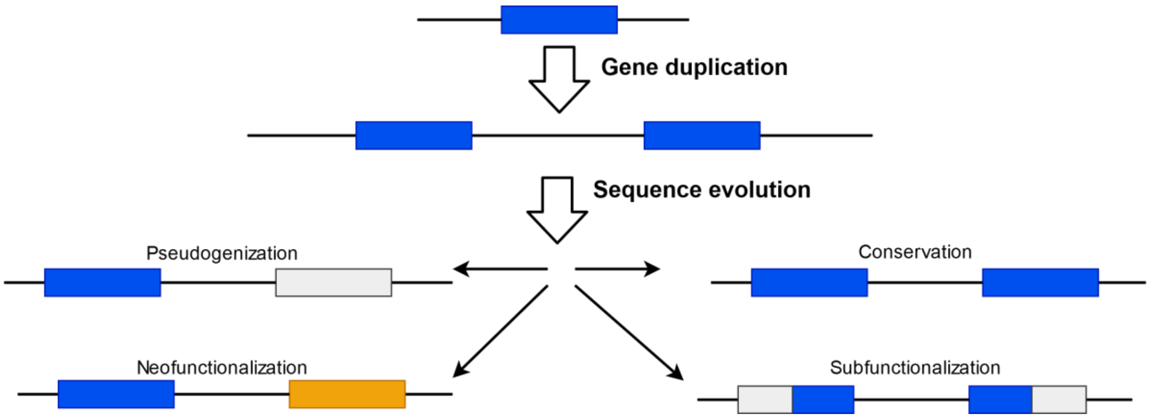

Figure 2: Schematic representation showing possible fates of duplicate genes. Gene duplication event creates redundant copy of a gene (each box represent one gene). With time if the duplicate copies do not accumulate mutations they remain conserved. In pseudogenization one the paralog becomes inactive (gray box). On neofunctionalizaation, one copy retains the ancestral function while the other paralog acquires new function (orange box). In case of subfunctionalization both paralogs accumulate mutation and both complement each other to perform the initial function (change of shade).

However, the most common fate of a duplicate gene copy is pseudogenization. To understand pseudogenization in detail, the rate of gene loss in many a paleopolyploid plant genomes has been studied (Thomas, 2006). In soybean, the rate of gene loss has been shown to slow down gradually; 1.28% of genes per million years following early legume WGD and 4.36% genes per million years following early 58 mya *Glycine* WGD (Schmutz et al., 2010). Few other duplicates tend to stay in duplicate copies. Further, most duplicates created during the *Glycine* WGD are under strong purifying pressure (Roulin et al., 2013).

1.3.2 Diploidization is not uniform across gene functions

Diploidization is not random (Thomas, 2006) and some gene families (e.g. TFs and signal transduction genes) are more prone to retain duplicated copies than others (Blanc and Wolfe, 2004a; Lehti-Shiu et al., 2017). Different mechanistic explanations have been proposed for this phenomenon, out of which the *gene balance hypothesis* (GBH) is the most accepted one. According to this hypothesis, upon a WGD, genes with many interaction partners have higher probability of being retained in duplicates, since alterations in the stoichiometry their protein products tend to be deleterious (Birchler and Veitia, 2007; Freeling, 2009; Birchler and Veitia, 2011). Retained copies then typically evolve via subfunctionalization (i.e. duplicates acquire complementary functions) or neofunctionalization (i.e. one of the copies evolves a new function) (Freeling et al., 2015).

1.4 Transcription factor families

Virtually all major biological processes are at least partially regulated at the transcriptional level by specific DNA-binding transcription factors (TFs), which specifically bind to *cis*-regulatory elements of target genes by means of DNA binding domains (DBDs). Many specific TFs bind to motifs located on the promoter region near the transcription start site and help general TFs to form a stable transcription initiation complex. Other TFs bind to distant regulatory sequences, such as enhancers or silencers, and can either stimulate or repress transcription of the associated genes (**Figure 3**)(Gonzalez, 2016). TFs bind to the regulatory region of target genes in a sequence specific manner. This sequence specificity is achieved by the DBD. Like other protein domains, DBDs have a stable structure and can be isolated from the TF protein without losing its activity. This enables in-depth study of such domains using crystallographic techniques. TFs sharing similar DBDs and are categorized into one TF family, which often show similar sequence specificity. For instance, TFs from ARF family recognize *TGTCTC* or *GAGACA* motifs. DBDs from several TF families are unable to bind with the DNA motif alone and require formation of a dimer (Gonzalez, 2016), such as those from the bZIP and bHLH families. They interact with the negatively charged DNA using their basic amino acid. Besides, DBD TFs may have other protein domains which interact with other proteins to enhance or repress transcription. Broadly speaking, the collective regulatory actions of TFs drive gene expression in different conditions. Because of their key regulatory roles, TFs have been extensively demonstrated to be critical for plant evolution and adaptation to multiple environments (Doebley and Lukens, 1998; Lehti-Shiu et al., 2017).

Proteins with shared conserved domains are often categorized into protein families. Hence a group of TFs having common DBD can be classified into a TF-family. As DBDs are highly conserved within a TF family, they can be used in the genome-wide identification of TF genes. Availability of several plant genome sequences has allowed the genome-wide identification of TFs using computational approaches. At first, nearly 1,800 TF genes were reported in *Arabidopsis thaliana* using the presence or absence of DBDs criteria. They represented more than 7% of all protein-coding genes in the species (Riechmann et al., 2000).

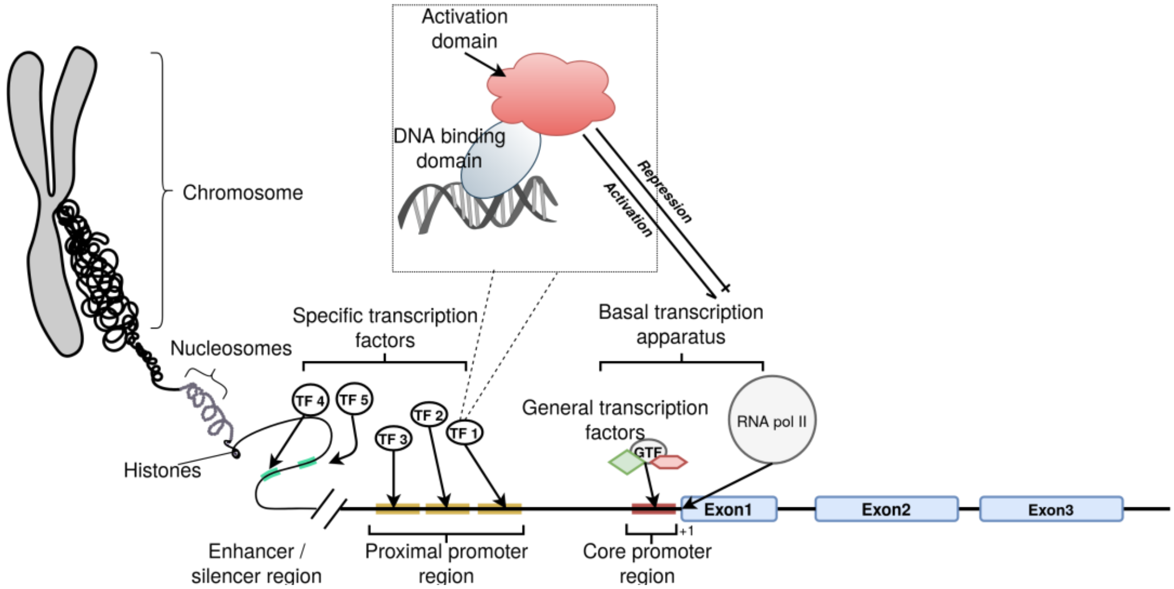

Figure 3: Transcription factor (TF) binding sites in eukaryotic genes. By means of their DNA-binding domains, TFs bind to specific motifs and drive the expression of a target gene. Binding sites can be located near or distant from the transcription start site (TSS). Adapted from (Gonzalez, 2016).

Over the years, several plant genomes had their TF families analyzed using computational approaches, resulting in the construction of several TF databases (**Table 1**) by following certain rules regarding the presence of DBDs and other domains (Riaño-Pachón et al., 2007; Mochida et al., 2010; Wang et al., 2010). Additionally, nuclear-localization signals, transcription-activation domains and oligomerization sites are also utilized to identify and classify TFs. PlantTFDB 4.0 is the most comprehensive plant TF database, containing TFs from 156 sequenced plant species. Each TF family is typically represented by sequence profiles and a hidden Markov model (HMM) in large databases such as Pfam and InterPro, which host a large collection of publicly available HMM models. The availability of these resources fueled the large scale functional and evolutionary analysis of plant TFs to explore their diversity.

Table 1: Plant transcription factor databases.

Websites	Acronym	Website address	Plant species
Arabidopsis transcription factor database	AtTFDB	https://agris-knowledgebase.org/AtTFDB/	Arabidopsis
PlantTF databases	PlantTFDB	http://cbi.planttfdb.pku.edu.cn	Multiple species
Plant TF database	PlnTFDB	http://plntfdb.bio.uni-potsdam.de/v3.0/	Multiple specie
RIKEN Arabidopsis TF database	RARTF	http://rarge.gsc.riken.jp/rartf/	Arabidopsis
Grass transcription factor database	GrassTFDB	https://grassius.org/grasstfdb.php	maize, sugarcane, sorghum and rice

Currently, over 50 TF families have been identified in plants (Jin et al., 2017). Members of several TF families have been intensively studied with regard to their important association with key biological processes, such as development, growth and defense. Comparative analysis showed that, although many TF families are present in all eukaryotic lineages, their sizes vary considerably (Lespinet et al., 2002; Nagata et al., 2016; Lehti-Shiu et al., 2017). It has been reported that plant TF families are usually larger than their animal counter parts (Shiu, 2005). The number of TFs also varies greatly between plant species. For example, Lehti-Shiu et al. found that Soybean has 27 times higher number of TFs as compared to marine green algae *Ostreococcus tauri* (Lehti-Shiu et al., 2017). They also suggested that the number of TFs in a species correlates with the presence of polyploidization events in the lineage. This view, combined with the fact that plants are more tolerant to polyploidization than animals, also explains the higher number of TFs in the former.

Genome resequencing and association studies have pinpointed involvement of TFs in controlling important agronomic traits. For instance, *SHAT1-5* (NAC family TF) promote shattering resistance by increasing lignification of fiber cap cells (Dong et al., 2014). The progress in sequencing technologies and automation over the past 12 years unleashed the power of comparative and population genomics in the identification of key genes involved in domestication and improvement of soybean and other crops, as illustrated by the discovery of many Quantitative Trait Loci (QTL) involved in commercially important traits (e.g. seed weight and oil content) by the resequencing of 302 soybean accessions (Zhou et al., 2015).

1.5 Computational approaches used to identify duplication events

1.5.1 Detecting syntenic blocks

Because of their profound impact on evolutionary trajectory, regions of chromosome originating from WGD events are of special interest. However, other disruptive factors such as high level of gene loss immediately after WGD, translocations, chromosomal rearrangements and recombination, complicate the identification of duplicated segments with conserved gene orders or syntenic blocks. Such factors are more severe in case of ancient duplication events. Therefore, bioinformatics approaches mainly focus on identifying remnants of such large duplicated segments, provided that there is reasonable protein homology (Van de Peer, 2004).

To infer homology between different chromosomal segments, each of these chromosomes is represented as a list of genes. These genes are represented by their genomic coordinates and are, then sorted according to their position on that chromosome. These lists are then used to find homologous regions. In practice, such gene homologs are determined by sequence comparison tools such as BLAST (Altschul et al., 1990; Altschul et al., 1997). The homology information is stored in a gene homology matrix (GHM) (Van de Peer, 2004). In a GHM, segments originating from large scale duplication events appear as diagonal lines; whereas tandem duplication (other local small scale duplicates) appear as vertical or horizontal lines. Among the tools available to analyze duplicated gene regions, DAGChainer (Haas et al., 2004) is one of the most popular.

1.5.2 Dating duplication events

In addition to identifying WGD, dating such events is also important. If many paralogous pairs have similar divergence times, a WGD event can be inferred. One of the popular methods to identify WGD and infer their ages is by analyzing the distribution of the rate of synonymous substitution per synonymous site (Ks) among duplicate genes (Lynch and Conery, 2000; Van de Peer, 2004). The time of divergence between two paralogs can be calculated as $T = \frac{Ks}{2\lambda}$

where λ is mean rate of synonymous substitution in the species(Van de Peer, 2004)

2. OBJECTIVES

One of the major goals of evolutionary biology has been to identify the genetic changes underlying phenotypic differences between organisms, and to distinguish the evolutionary forces responsible for these changes. Variation in gene family size among species may have important roles in speciation and adaptation. Variation in family sizes is influenced by both, large (e.g. WGD) and small scale duplications. On the other hand, gene loss results in contraction of gene families. In relatively rare conditions, creation of *de novo* genes results in the emergence of new gene families. With the availability of whole genomes from multiple species and robust computational capacity, many studies have been focusing on analyzing large changes in the size of gene families (Lehti-Shiu et al., 2017).

Although legumes have one or more shared phenotypes with other angiosperms, they have distinguishing features. For example, symbiotic nitrogen fixation is largely restricted to the legume family and few non-legumes. Bigger seed size is also an important trait found in legumes like soybean and lacking in widely studied species like *Arabidopsis*.

According to currently accepted hypothesis that wild soybean domestication took place in Central China between 5,000 and 9,000 years ago and through agriculture practices the modern soybean *Gl. max* came to existence (Han et al., 2016; Sedivy et al., 2017). The progress in sequencing technologies and automation over the past 12 years unleashed the power of comparative and population genomics in pinpointing key genes involved in domestication and improvement of soybean and other crops, as illustrated by the discovery of many Quantitative Trait Loci (QTL) involved in commercially important traits (e.g. seed weight and oil content) by the resequencing of 302 soybean accessions (Zhou et al., 2015).

Plant science has been highly benefited from *Arabidopsis* genome sequencing project. Even it is considered as model species it is insufficient to explain molecular basis of important biological processes such as symbiotic nitrogen fixation, which is largely restricted to legumes. We hypothesized that genome-wide duplication and subsequent retention of TFs might have contributed in creating such legume-specific features. Briefly, we aimed to perform large-scale comparative analysis of predicted TFs from 15 legume and five non-legume species whose genome has been sequenced and available in public domain. *Se. moellendorffii*, *Am. trichopoda* and *Aq. coerulea* were used as a representative outgroups of land plants. Similarly, *Vi. vinifera* was used as outgroup of all legumes. *Arabidopsis* was included because it is the most widely studied model plant species.

Thus the main objectives of this study are as following:

i. Identify the major TF in all sequenced legumes and perform a comparative analysis.
ii. Identify the lineage specific amplification of TF families.
iii. What are the major forces causing TF family amplification?
iv. Which families have amplified and relate such expansions with legume specific traits?
v. Which TFs have possible role in soybean domestication?

3. MATERIAL AND METHODS

3.1.1 Genome sequences and annotations

Genome sequencing and annotation data for 15 legume species were obtained from public sources (**Table 2**). In addition to this we also obtained similar information for 5 non-legume species. Most of the analyses performed here were based on the predicted protein sequence encoded by the longest splicing isoform (when more than one were available).

Table 2: Plant species used in this study and their corresponding genome assembly versions. Non-legume species are marked with asterisks.

Scientific name	Genome assembly version	Genes	Chromosome number	Reference
Cajanus cajan	GCA 000340665.1	48,331	11	(Varshney et al., 2011)
Phaseolus vulgaris	Pvulgaris 218 v1.0	27,197	11	(Schmutz et al., 2014)
Vigna radiata	Vradiata ver6	35,143	11	(Kang et al., 2014)
Vigna angularis	Vigan1.1	34,172	11	(Yang et al., 2015)
Glycine max	Gmax 275 Wm82.a2.v1	56,044	20	(Schmutz et al., 2010)
Glycine soja	W05v1.0	55,539	20	(Xie et al., 2019)
Cicer reticulatum	WCGAP v1.0	25,680	8	(Gupta et al., 2017)
Cicer arietinum	ASM33114v1	33,107	8	(Jain et al., 2013; Varshney et al., 2013; Parween et al., 2015)
Medicago truncatula	Mtruncatula 285 Mt4.0v1	50,894	8	(Young et al., 2011; Tang et al., 2014)
Glycyrrhiza uralensis	Draft-genome.20151208	34,445	8	(Mochida et al., 2017)
Lotus japonicus	Build 3.0	39,734	6	(Sato et al., 2008)
Lupinus angustifolius	v1.0	33,076	20	(Hane et al., 2017)
Arachis ipaensis	Araip1.0	46,410	10	(Bertioli et al., 2016)
Arachis duranensis	Aradu1.0	42,562	10	(Bertioli et al., 2016; Chen et al., 2016)
Chamaecrista fasciculata	version.1	21,781	8	(Griesmann et al., 2018)
*Arabidopsis thaliana**	167 TAIR10	27,416	5	(Arabidopsis-Genome-Initiative, 2000)
*Vitis vinifera**	v145 Genoscope.12X	26,346	19	(Jaillon et al., 2007)
*Amborella trichopoda**	AmTr v1.1	26,846	13	(Albert et al., 2013)
*Aquilegia coerulea**	v3.1	30,023	7	(Filiault et al., 2018)
*Selaginella moellendorffii**	v1.0	22,285	10	(Banks et al., 2011)

3.1.2 Prediction and classifications of transcription factors (TFs)

First, predicted proteins with less than 50 amino acids or containing premature stop codons or more than 20% ambiguous amino acids were excluded. To remove redundancy due to splicing isoforms and incomplete gene predictions, we removed nearly identical sequences using BLASTCLUST (Altschul et al., 1997) as previously described (parameters: -S 1.89 -L 0.9 -b F) (Gossani et al., 2014; Vidal et al., 2016). We adopted the TF family classification scheme of plantTFDB (Zhang et al., 2011; Jin et al., 2017). We created a local database of protein domains by combining all HMM profiles from PFAM-A (Release 31.0) (Finn et al., 2016) and 13 plant specific TF HMM profiles downloaded from PlantTFDB (**Table 3**). Protein sequences were searched for conserved domains using HMMER 3.0 (http://hmmer.org) (domain e-value cutoff < 0.01). TFs were classified in 58 families according to their DBD.

Table 3: Conserved domains and rules used for identifying transcription factors. A transcription factor was expected to have at least one DNA binding domain and may contain auxiliary domains. Forbidden domains were used to eliminate false positives.

Superfamily	Family	Description	DBD* domain [# of domains] (Pfam accession)	Auxiliary domain	Forbidden domain
AP2/ERF	AP2	APETALA 2	AP2[>=2] (PlantTFDB)		
	ERF	Ethylene Responsive Factor	AP2[1] (PlantTFDB)		
	RAV	Related to ABI3/VP1	AP2[1] (PlantTFDB) and B3(PF02362)		
B3-superfamily	B3	B3	B3(PF02362)		
	ARF	Auxin response factor	B3(PF02362)	Auxin_resp (PF06507)	
BBR-BPC	BBR-BPC	Barley B Recombinant (BBR) - BASIC PENTACYSTEINE1 (BPC1)	GAGA_bind (PF06217)		
BES1	BES1	BRI1-EMS-SUPPRESSOR	BES1_N (PF05687)		Glyco_hydro_14 (PF01373)
bHLH	bHLH	basic helix loop helix	HLH (PF00010)		
bZIP	bZIP	Basic leucine zipper	bZIP_1 (PF00170)		
C2C2	Dof	DNA binding with one finger	Zf-Dof (PF02701)		
	GATA	GATA	GATA-zf (PF00320)		
	CO-like	CONSTANS-like	Zf-B_box (PF00643)	CCT (PF06203)	
	YABBY	YABBY	YABBY (PF04690)		
	LSD	LESION SIMULATING DISEASE 1, LSD1	Zf-LSD1 (PF06943)		Peptidase_C14 (PF00656)
	C2H2	CCHH (Zn)	Zf-C2H2 (PF00096)		Exonuc_X-T(PF00929)

C3H	C3H	CCCH (Zn)	Zf-CCCH (PF00642)		RRM_1 (PF00076) or Helicase_C (PF00271)
CAMTA	CAMTA	Calmodulin binding transcription factors	CG1 (PF03859)		
CPP	CPP	Cystein-rich polycomb-like protein	CXC (PF03638)		
DBB	DBB	Double B-box zinc finger	zf-B_box[>=2] (PF00643)		
E2F/DP	E2F/DP	E2 factor protein and DP protein	E2F_TDP (PF02319)		
EIL	EIL	Ethylene-Insentive3 (EIN3)-like protein3 (EIL3)	EIN3 (PF04873)		
FAR1	FAR1	FAR-RED IMPAIRED RESPONSE1	FAR1 (PF03101)		
FLO	LFY	LEAFY	FLO_LFY (PF01698)		
GARP	G2-like	Golden2 (G2)-like	G2-like (PlantTFDB)	Response_reg (PF00072)	
	ARR-B	Type-B phospho-accepting response regulator (ARR) family	G2-like (PlantTFDB)		
GeBP	GeBP	GLABROUS1 enhancer-binding protein (GeBP)	DUF573 (PF04504)		
GRAS	GRAS	GAI, RGA, and SCR	GRAS (PF03514)		
GRF	GRF	GROWTH-REGULATING FACTOR (GRF)	WRC (PF08879)	QLQ (PF08880)	
HB	HD-ZIP	HD-Zip	Homeobox (PF00046)	HD-ZIP_I/II (PlantTFDB) or SMART (PF01852)	
	TALE	Three Amino acid Loop Extension	Homeobox (PF00046)	BELL (PlantTFDB) or ELK (PF03789)	
	WOX	WUS homeobox-containing protein family	Homeobox (PF00046)	Wus type homeobox (PlantTFDB)	
	HB-PHD	HB-PHD finger	Homeobox (PF00046)	PHD (PF00628)	
	HB-other	HB-other	Homeobox (PF00046)		
HRT-like	HRT-like	Hairy-Related transcription-factor-like	HRT-like (PlantTFDB)		
HSF	HSF	Heat shock factor	HSF_dna_bind (PF00447)		
LBD (AS2/LOB)	LBD	ASYMMETRIC LEAVES2/LATERAL ORGAN BOUNDARIES	LOB (PF03195)		
MADS	M-type	MADS-type I	SRF-TF (PF00319)		
	MIKC	MADS-type II	SRF-TF (PF00319)	K-box (PF01486)	
MYB superfamily	MYB	Myb proto-oncogene protein	Myb_dna_bind [1] (PF00249)	SWIRM (PF04433)	
	MYB_related	Myb-related	Myb_dna_bind [>=2] (PF00249)	SWIRM (PF04433)	
NAC	NAC	NAM, ATAF1, 2 and CUC2	NAM (PF02365)		
NF-X1	NF-X1	Nuclear factor, X-box binding 1	Zf-NF-X1 (PF01422)		
NF-Y	NF-YA	Nuclear factor Y subunit A	CBFB_NFYA (PF02045)		

	NF-YB	Nuclear factor Y subunit B	NF-YB (PlantTFDB)
	NF-YC	Nuclear factor Y subunit C	NF-YC (PlantTFDB)
Nin-like	Nin-like	NODULE INCEPTION	RWP-RK (PF02042)
NZZ/SPL	NZZ/SPL	SPOROCYTELESS/NOZZLE	NOZZLE (PF08744)
S1Fa-like	S1Fa-like	S1Fa-like	S1FA (PF04689)
WRKY	WRKY	WRKY	WRKY (PF03106)
Trihelix	Trihelix	Trihelix	Trihelix (PlantTFDB)
TCP	TCP	TEOSINTE-LIKE1, CYCLOIDEA, and PROLIFERATING CELL FACTOR1	TCP (PF03634)
ZF-HD	ZF-HD	Zinc finger homeodomain protein	ZF-HD_dimer (PF04770)
SBP	SBP	SQUAMOSA promoter binding protein	SBP (PF03110)
SRS	SRS	SHI RELATED SEQUENCE	DUF702 (PF05142)
SAP	SAP	STERILE APETALA	SAP (PlantTFDB)
Whirly	Whirly	Whirly	Whirly (PF08536)
STAT	STAT	Signal Transducers and Activators of Transcription	STAT (PlantTFDB)
VOZ	VOZ	Vascular plant One-Zinc finger	VOZ (PlantTFDB)

3.1.3 Identification of syntenic blocks and analysis of synonymous substitution rate (Ks)

We compared the homologous gene order on two chromosomes to identify the syntenic blocks. At first, we identified the homologous genes by an all-vs-all BLASTP search. From these we selected bidirectional best BLASTP hits (e-value ≤ $1e^{-10}$, 35% minimum identity, 50% minimum query coverage) and annotated with their genomic coordinates. We identified segmental duplications using DAGCHAINER (version r02062008) (Haas et al., 2004). A minimum of four collinear genes were required to identify a syntenic block (DAGCHAINER, parameter -A 4), as previously used in soybean (Severin et al., 2011). Tandem duplicates were also identified using DAGCHAINER (parameters -T -A 2). Tandem or segmental gene pairs had their non-synonymous (Ka) and synonymous (Ks) mutation rates estimated using the *bp_pairwise_kaks* script, distributed with BioPerl (v5.22.1) (Stajich et al., 2002).

3.1.4 Orthologous groups (OGs) and TF paralog identification

We clustered the predicted proteins on the basis of the pairwise sequence similarity of their longest protein products, which was computed with BLAST (e-value ≤ $1e^{-5}$) (Altschul et al., 1997). Sequence pairs with percentage identity of at least 35% and query coverage of at least 50% were used for Markov clustering using mclblastline

(v. 12-068; Inflation parameter: 1.5) (Enright et al., 2002). From these OGs identified from all proteins from all the 20 species we selected only those OGs with TFs. We classified the TF paralog pairs according to their predicted modes of duplication as described previously (Proulx et al., 2011; Qiao et al., 2018) (**Figure 4**). The categories and classification priority were as following: TF paralogs within collinear regions were classified as large segmental duplicates (SD). Those present tandemly were called tandem duplicates (TD). Pairs with paralogs separated by one to five intervening genes were called proximal duplicates (PD). Distant single gene transposition duplication may happen by RNA- or DNA-based mechanisms (Cusack and Wolfe, 2007). DNA- transposons like MULE, helitrons, CACTA elements may relocate genes to a distance place on the genome. RNA- transposons (retro-transposons) insert RNA molecule in to a novel location in the genome. To identify such transposon mediated duplicates we first checked that one gene out of a paralog pairs was located in its ancestral position by comparing syntenic regions in closest outgroup species. Further, if one paralog from a pair has a single exon, while the other copy had at least two exons, then they could have originated through the action of RNA-based transposon activity and were classified as retro-transposons (rTE). If only one member of the duplicate pair is in its original ancestor loci while the other copy was found in a distant location, it was termed as DNA-transposon mediated duplication (dTE). The remaining duplicates were called dispersed duplicates (DD). The order to assign the mode was SD>TD>PD>rTE>dTE>DD.

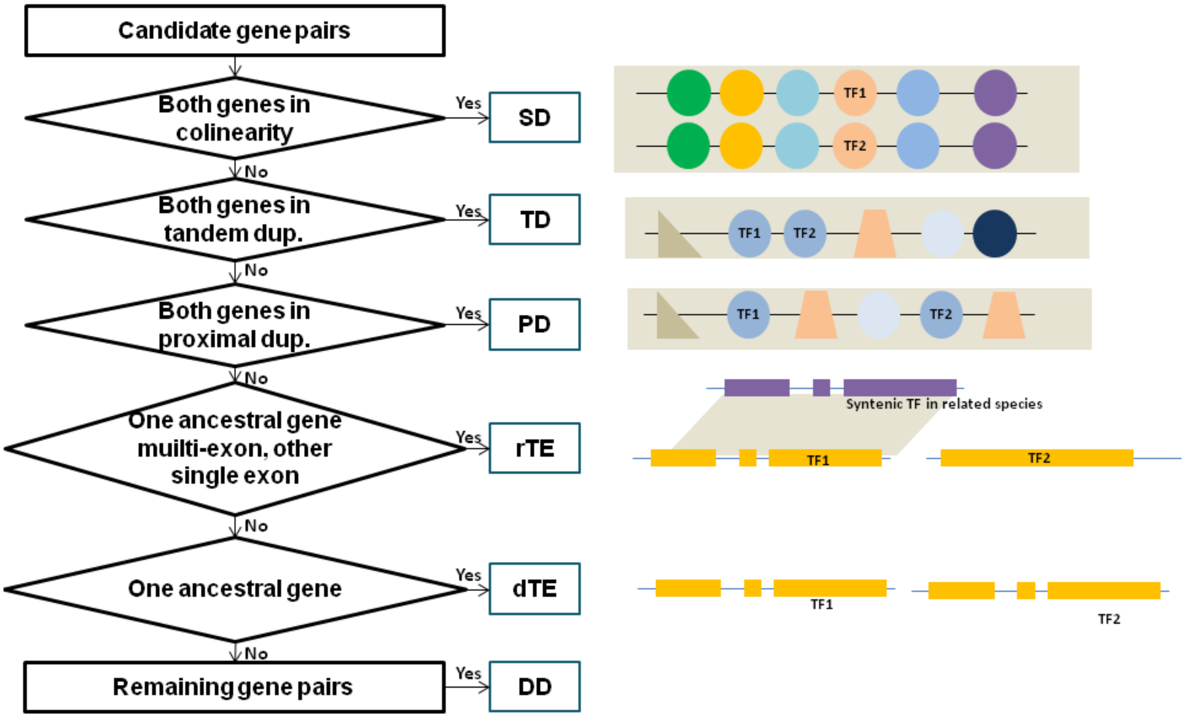
Candidate gene pairs
Both genes in colinearity
Yes
SD
No
Both genes in tandem dup.
Yes
TD
No
Both genes in proximal dup.
Yes
PD
No
One ancestral gene muilti-exon, other single exon
Yes
rTE
No
One ancestral gene
Yes
dTE
No
Remaining gene pairs
Yes
DD
TF1
TF2
Syntenic TF in related species

Figure 4: Rules used for classifying transcription factor paralogs. The flowchart represents the overall pipeline used to determine mode TF paralog origin. The schematics on the right hand side described the gene arrangements for each duplicate category. SD:Segmental duplicates; TD: Tandem duplicates; PD: Proximal duplicates; rTE: Retrotransposon mediated duplicates; dTE: Transposon mediated duplicates; DD: Dispersed duplicates. Adapted from (Proulx et al., 2011).

3.1.5 Species phylogeny

We reconstructed a phylogenetic tree using low copy orthologs present in all 20 species. To obtain the single copy orthologs, we clustered the predicted proteins on the basis of the pairwise sequence similarity of their longest protein products, which was computed with BLAST (e-value ≤ 1e^{-5}) (Altschul et al., 1997). Sequence pairs with percentage identity of at least 35% and query coverage of at least 50% were used for Markov clustering using mclblastline (v. 12-068; Inflation parameter: 1.5) (Enright et al., 2002). Clusters containing up to 22 genes with at least one gene from each species were used. If a species had paralogous genes, the paralog with greater identity to orthologs from other species was used. Amino acid sequence alignment was performed using DECIPHER (Wright, 2015) and cDNA alignment performed with PAL2NAL (Suyama et al., 2006). We concatenated the codon alignments of these genes to create a super-alignment. Next, phangorn (Schliep, 2011) was used to estimate the best substitution model for the phylogenetic reconstruction, which was performed using RAxML (v8.2.11; model: GTRGAMMAIG4, bootstrap: 1000) (Stamatakis, 2014). The phylogram and sequence alignment were used in relTime-ML (implemented in MEGA-X, v.10.0.1) (Tamura et al., 2018) to generate an ultrametric tree. We used the TimeTree database (Kumar et al., 2017) to retrieve the divergence times of Fabaceae and *Vi. vinifera* (110 mya) and of *Ph. vulgaris* and *Gl. max* (24 mya), which were used as references.

3.1.6 Estimation of expansions and contractions in TF families

We used CAFE (v4.2) (Han et al., 2013) to assess the evolution of TF family sizes. The time-calibrated species tree and TF OG compositions were given as inputs to CAFE. We used the *cafeerror.py* script, available in the CAFE package, to model error rates that might have been introduced in gene family sizes, particularly by species with more fragmented genome assemblies (e.g. *Lu. angustifolius*) (Han et al., 2013). This error model was used adjust family sizes.

The analysis with CAFE involves three steps: i. Estimating gene-birth (λ)/ gene-death (μ) rate parameters by running CAFE for multiple times and selecting the parameters that gave the best maximum likelihood estimate. ii. Use these rate parameters to estimate OG sizes at ancestor nodes and to predict rapidly evolving OGs (p-value < 0.05), which are those that significantly gained or lost genes. iii. Interpret the overall evolution from the net change in gene family size. The net change in gene family size at each node on the species tree was expressed as:

$$\text{The average expansion on node } \boldsymbol{A_m} = \frac{\sum_{i=1}^{n}(\boldsymbol{M_i} - \boldsymbol{X_i})}{\boldsymbol{n}}$$

where n is total number of OGs, (M_i−X_i) is the difference in OG size between node M and its parent node X for a given OG i. A negative or positive A_m value stands for contraction or expansion of the OG, respectively. Some remarkably expanded or contracted TF OG had their phylogenies reconstructed with RAxML (v8.2.11; model:

GTRGAMMAAUTO, bootstrap: 1000) and visualized using Figtree (v.1.4.3) (http://tree.bio.ed.ac.uk/software/figtree/).

3.1.6.1 Estimating multiple gene-birth (λ)/ gene-death (μ) rates for different parts of the species tree

We searched for optimal rate parameters (λ/μ) based on the maximum likelihood score using CAFE. Due to frequent polyploidization events in angiosperms, we believed that use of single global rate parameters (λ/μ) would result in erroneous results. Assuming different gene gain and loss rates both preceding and following a WGD event, we decided to estimate separate rate parameters for different lineages throughout the species tree. Considering known polyploidization events, we used 10 separate lineages (**Figure 5A**) for rate estimation. *Se. moellendorffii* , *Am. trichopoda, Aq. coerulea, Vi. vinifera* and *Ar. thaliana* were assigned as distinct a priori rates. All legumes except *Lu. angustifolius*, *Gl. max* and *Gl. soja*, were considered to evolve at similar rate. Then, we repeated estimating λ and μ by running CAFE for 50 times and selected the those parameters that gave the maximum likelihood score (**Figure 5B**). Such a priori parameter setting allowed CAFE to estimate a separate λ/μ for each lineage and clades evolved after WGD events. Although not a perfect solution, this strategy is valid given the limitations of existing methods to explicitly incorporate WGD information.

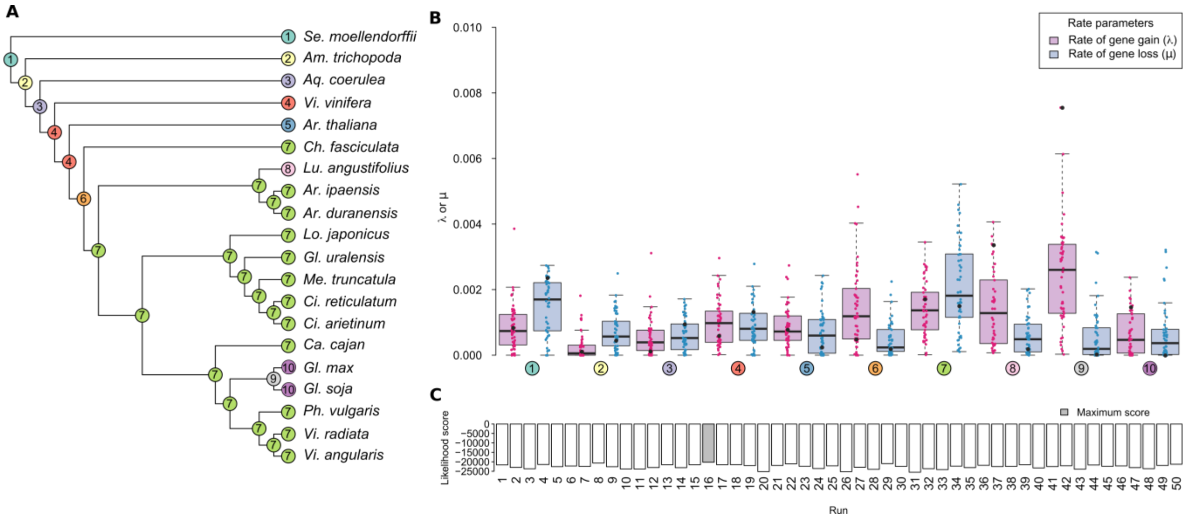

Figure 5: Various steps and parameters used for searching optimal gene birth (λ) and death (μ) rates. A. The tree topology showing 10 clades groups. Species sharing common polyploidy events have been grouped together. We assumed that each of these clades evolves at different rates and instructed CAFE to assign 10 different a priori λ/μ values to them. Nodes with similar color codes and numeric labels were assumed to evolve at similar rates. For example, the *Glycine* common ancestor created after ~13Mya WGD was labeled as 9, *Gl. max* and *Gl. soja* were labeled as 10 to allow CAFE to estimate a separate λ/μ for each species after the WGD events. B. Box plots showing λ-μ values obtained from multiple runs, in different branches on the species tree, as defined in A. The larger points represent those rate values that yielded maximum likelihood score (shown as grey bar in C) among the 50 repetitions. C. Histogram showing likelihood scores for different runs 50 λ and μ rate estimation steps. The grey bar represents the run yielding the maximum likelihood score (using larger points from B) among the 50 runs.

3.1.7 Gene expression data

Ar. thaliana and *Ph. vulgaris* normalized gene expression data were obtained from ArrayExpress (Liu et al., 2012) and *Pv*GEA (O'Rourke et al., 2014), respectively. Four additional RNAseq datasets were downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/). The first two datasets comprise two soybean transcriptome studies (Bioproject PRJNA208048, PRJNA79597) (Libault et al., 2010b; Severin et al., 2010). The third dataset includes *Me. truncatula* transcriptomes (Boscari et al., 2013) in the following conditions and tissues: nitrogen-starving roots, roots inoculated with *Sinorhizobium meliloti*, and root nodules (BioProject PRJNA79233). We also downloaded an additional *Me. truncatula* RNAseq data covering 7 different tissues (BioProject PRJNA80163).

RNAseq reads were mapped on each species genome using STAR v2.5.3a (Dobin et al., 2013) and normalized gene expression values were estimated with StringTie v1.3.4d (Pertea et al., 2015), both with default parameters.

3.1.8 Tissue specific expression

Normalized gene expression estimates from *Ar. thaliana, Ph. vulgaris,* two soybean transcriptome studies and *Me. truncatula* transcriptomes were obtained as described in previous chapter. Expression values lower than 1 were converted to 0 and considered not expressed. We added 1 to all values, which were then $\log_2$ transformed. To determine tissue-preferential expression, we transformed the gene expression values in a transformed z-score index (Kryuchkova-Mostacci and Robinson-Rechavi, 2016). Depending on the highest expression in a given tissue, genes with transformed z-score index > 0.9 were considered as preferentially expressed.

3.1.9 Microsyntenic regions between *Gl. max*, *Gl. soja*, and *Ph. vulgaris*

We used DAGCHAINER output files to identify the microsyntenic regions in *Gl. max, Gl. soja*, and *Ph. vulgaris*. In particular, we queried the genes from OGs with significantly larger (as predicted by CAFE) sizes in *Gl. max* in comparison to the *Glycine* node. For each *Gl. max* gene, we considered only one collinear region from *Gl. soja* and *Ph. vulgaris*. When more than one collinear region was detected, we selected that with the highest DAGCHAINER alignment score. We visualized the microsynteny regions using Genome Context Viewer available on Legume Information System (Cleary et al., 2017).

3.1.10 QTL intervals

We obtained the chromosomal coordinates of 150 QTLs significantly associated with 57 soybean traits (Fang et al., 2017). Chromosomal coordinates of soybean genes were mapped to these QTL regions using bedtools v2.26.0 (Quinlan and Hall, 2010).

4. RESULTS AND DISCUSSION

4.1.1 Systematic identification of transcription factors

We used a set of diagnostic specific DNA binding domains and forbidden domains (**Table 3**) to identify TFs in the genomes of 20 plant species (**Table 2**). We predicted a total of 37,008 TFs (**Appendix A.1**), which were classified in 58 broad families (**Table 4**). A total of 31,111 TFs were predicted in the 15 legume genomes. We benchmarked our pipeline by comparing the detected TFs with those previously predicted in *Ar. thaliana*. Out of 1,713 *Ar. thaliana* TFs available in PlantTFDB, 1,673 (98%) were correctly predicted. Further, 59 TFs were exclusively predicted by our pipeline, out of which 40 were annotated as TFs in the TAIR database (https://www.arabidopsis.org) (**Table 5**).

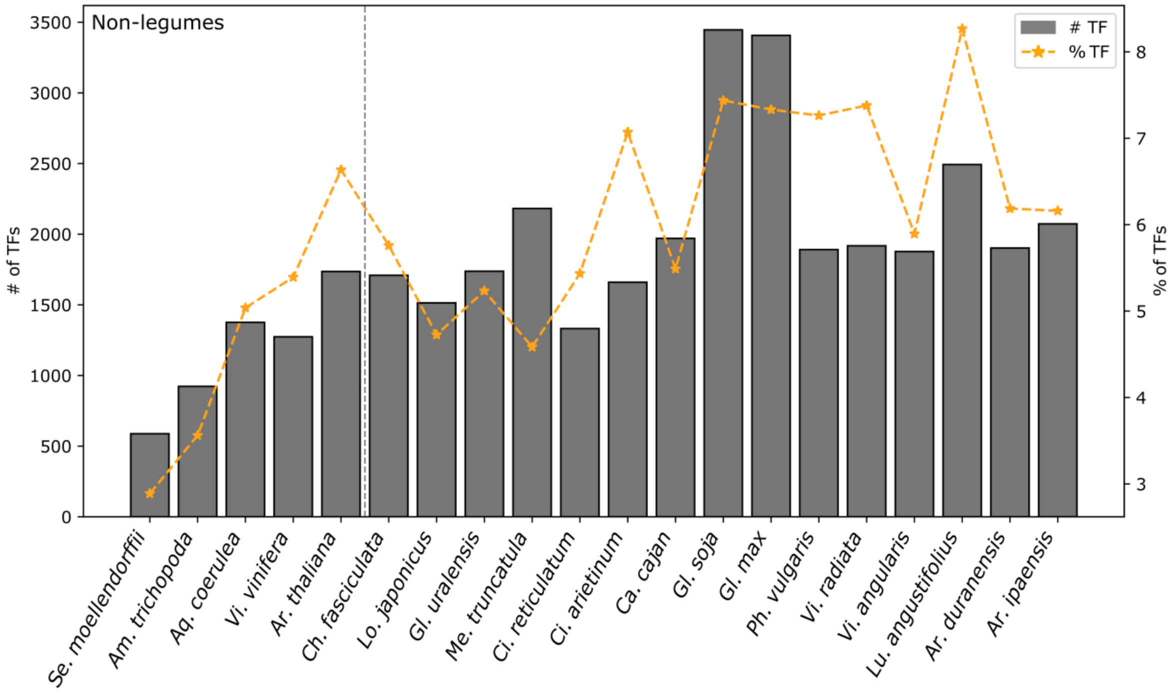

Figure 6: Absolute and relative number of transcription factors in each species. Grey bars and the orange line represent the absolute number and percentage of transcription factors in each species, respectively. Legumes and non-legumes are separated by a dotted vertical line.

Table 4: Number of transcription factors identified in each species.

Super family	TF family	*Ca. cajan*	*Ph. vulgaris*	*Vi. radiata*	*Vi. angularis*	*Gl. max*	*Gl. soja*	*Ci. reticulatum*	*Ci. arietinum*	*Me. truncatula*	*Gl. uralensis*	*Lo. japonicus*	*Lu. angustifolius*	*Ar. ipaensis*	*Ar. duranensis*	*Ch. fasciculata*	*Ar. thaliana*	*Vi. vinifera*	*Aq. coerulea*	*Am. trichopoda*	*Se. moellendorffii*
AP2/ERF	AP2	26	31	24	26	49	52	24	25	31	24	28	38	30	24	22	18	20	16	12	16
	ERF	14	149	156	161	290	279	115	129	186	112	107	189	143	130	131	122	72	78	64	35
	RAV	2	3	3	3	5	5	2	3	3	3	3	6	3	3	4	6	3	4	1	2
	BBR-BPC	5	4	4	6	4	5	2	3	2	4	5	10	5	6	4	7	5	4	4	1
	BES1	6	6	6	8	10	8	7	6	7	6	5	12	9	8	7	8	6	4	5	5
B3	ARF	23	27	28	27	42	45	25	25	38	11	25	43	31	30	22	22	18	13	13	7
	B3	49	48	36	33	77	79	23	35	132	49	61	52	69	68	29	73	29	104	18	19
	C3H	44	44	42	43	73	80	45	50	58	40	54	65	48	45	37	50	47	40	32	29
C2C2 Zn-finger	CO-like	10	13	12	10	22	22	10	10	11	11	8	20	11	11	10	17	6	8	5	4
	Dof	37	42	42	40	74	73	34	37	40	41	30	67	39	38	39	36	22	29	18	11
	GATA	33	32	29	27	61	63	26	27	42	30	20	45	25	25	26	30	20	29	21	6
	LSD	7	4	4	4	5	4	4	4	4	2	5	7	4	4	3	3	3	3	2	2
	YABBY	9	8	10	9	12	13	7	8	7	8	7	14	7	8	9	6	7	5	6	0
	C2H2	22	134	128	128	219	218	78	108	111	108	93	183	136	127	157	104	64	87	85	35
	CAMTA	10	8	11	6	14	15	7	7	8	8	7	9	8	7	6	6	5	5	3	5
	CPP	8	9	10	7	16	15	7	8	12	9	10	13	14	13	9	10	8	7	6	5
	DBB	11	12	14	13	16	16	6	7	8	8	8	13	7	8	10	8	8	5	4	4
	E2F/DP	7	7	7	9	12	13	6	6	6	10	7	12	10	9	6	8	7	7	5	4
MADS	M-type	48	44	23	31	77	80	34	43	101	29	34	38	28	23	32	65	17	50	19	12
	MIKC	23	34	47	27	75	71	16	51	38	14	21	27	44	48	28	42	35	24	15	3
	EIL	6	7	4	5	10	12	6	7	13	9	7	9	7	6	17	6	2	2	2	6
	FAR1	49	25	67	20	68	79	17	37	76	79	29	10	298	198	41	17	19	92	10	0
	LFY	1	1	1	1	3	2	1	1	1	2	0	3	2	2	1	1	1	1	1	1
GARP	ARR-B	15	15	16	18	31	34	13	14	28	14	10	19	12	11	9	15	12	13	7	6
	G2-like	44	50	52	46	96	100	36	41	44	49	39	78	49	46	47	42	39	34	27	20
	GRAS	57	55	58	58	108	111	47	46	66	53	62	54	49	48	52	34	43	36	44	47
	GRF	10	10	9	8	20	20	8	8	8	10	9	15	11	11	10	9	8	7	6	4
	GeBP	7	5	8	5	9	11	7	8	7	5	4	12	4	7	4	23	1	13	6	1
Homeobox	HB-PHD	2	3	3	3	6	6	2	2	2	3	3	2	4	3	3	2	2	2	2	1
	HD-ZIP	51	55	54	54	89	92	45	49	57	54	40	82	45	43	45	48	33	30	22	9
	TALE	34	32	31	28	57	61	22	24	23	32	25	45	28	30	25	21	22	17	12	7
	WOX	18	18	20	61	32	33	14	18	19	17	14	31	15	15	13	16	10	10	9	8
	HB-Other	6	7	8	7	15	13	8	8	7	8	9	12	9	7	6	6	7	6	5	4
	HRT-like	1	1	1	1	1	1	0	2	3	1	1	1	1	1	1	2	1	1	1	2
	HSF	27	29	32	32	46	45	20	22	28	32	10	33	22	22	32	24	18	15	12	6
	LBD	52	48	47	49	75	75	35	47	59	53	41	70	52	49	52	41	43	29	23	13
MYB	MYB	17	170	172	169	294	288	80	132	162	146	100	210	136	135	152	146	138	97	61	21
	MYB related	84	82	73	75	162	165	53	62	93	95	62	107	76	70	62	72	57	55	36	34
	NAC	91	90	92	90	142	139	61	77	97	72	81	117	85	84	88	112	70	76	45	21
	NF-X1	1	2	2	2	3	3	3	3	3	3	3	2	2	2	2	2	3	3	2	2
NF-Y	NF-YA	11	9	9	9	16	18	7	8	8	9	8	13	12	11	7	10	7	5	5	1
	NF-YB	23	19	23	18	36	36	19	22	24	23	18	25	16	14	20	13	17	14	9	7
	NF-YC	14	15	14	14	22	22	11	11	15	14	9	16	13	12	10	14	8	10	8	5
	NZZ/SPL	1	3	2	3	2	2	2	2	1	1	2	3	4	3	2	1	2	2	1	3
	Nin-like	13	12	11	12	27	26	10	9	13	12	9	15	16	12	11	14	8	9	7	7
	S1Fa-like	2	4	3	3	4	4	3	4	5	3	3	6	2	2	5	3	3	1	1	0
	SAP	3	5	3	1	5	5	2	2	3	3	4	5	5	5	1	4	3	3	2	1
	SBP	24	23	22	23	42	40	17	19	24	25	16	35	19	18	19	16	18	13	12	9
	SRS	11	10	11	10	22	22	8	8	11	10	11	12	11	10	10	11	6	4	6	4
	STAT	1	1	1	1	1	1	1	1	1	1	1	1	2	1	0	2	1	1	1	2
	TCP	25	27	27	27	55	56	19	24	21	24	24	42	30	23	23	24	15	16	14	4
	Trihelix	38	41	41	42	70	74	28	38	36	36	32	64	40	43	39	30	26	33	30	34
	VOZ	2	3	3	3	4	5	2	2	2	3	2	4	2	2	2	2	2	2	1	0
	WRKY	96	90	95	94	171	170	59	81	104	79	72	112	85	84	71	73	59	38	31	12
	Whirly	3	3	3	5	7	7	3	3	3	2	3	4	2	3	3	3	2	3	2	1
	ZF-HD	18	19	18	18	41	45	14	15	17	15	20	26	13	15	31	17	10	16	9	7
	bHLH	15	164	161	159	321	320	110	141	162	141	129	207	150	145	137	142	106	99	72	45
	bZIP	72	79	85	85	141	146	61	70	91	73	64	128	73	74	65	77	50	46	41	28

Table 5: *Arabidopsis* genes TFs exclusively predicted by our pipeline.

Category	Gene Description	Genes
Genes annotated as transcription factors	C2H2 and C2HC zinc fingers superfamily protein	AT1G13400, AT1G68360, AT1G68480, AT3G01030, AT3G23140, AT4G04404, AT5G54340
	B3 domain protein	AT1G50220, AT1G51970, AT1G78640, AT4G03160, AT4G05630, AT5G54067
	Zinc finger protein 622	AT2G24500, AT4G31420
	C2H2-like zinc finger protein	AT5G48890, AT5G54360
	Zinc finger C-x8-C-x5-C-x3-H type family protein	AT1G21580, AT5G66270
	SET domain-containing protein	AT2G23380, AT4G02020
	Homeodomain-like superfamily protein	AT2G21650, AT3G53440
	response regulator 23	AT5G62120
	B3 DNA-binding domain protein	AT1G10455
	bZIP family transcription factor	AT1G35490
	Basic-leucine zipper (bZIP) transcription factor family protein	AT2G21235
	bZIP protein	AT5G04840
	C2H2-type zinc finger family protein	AT1G02040
	C2H2 type zinc finger transcription factor family	AT1G49900
	zinc finger protein 6	AT1G67030
	zinc finger (C2H2 type) family protein	AT2G36930
	CCCH-type zinc finger family protein	AT2G02160
	DNA-binding storekeeper protein transcriptional regulator-	AT2G20805
	Homeodomain-like protein with RING/FYVE/PHD-type zinc	AT1G01150
	ALWAYS EARLY 2	AT5G27610
	ADA2 2A	AT3G05380
	ADA2 2B	AT3G07740
	Protein ALWAYS EARLY 1	AT4G16420
	WRKY family transcription factor	AT3G32090
Genes not annotated as transcription factors	F-box associated ubiquitination effector family protein	AT2G21920
	Telomerase activator1	AT3G09290
	Protein arginine methyltransferase 3	AT3G12270
	Protein CTF7	AT4G31400
	D111/G-patch domain-containing protein	AT5G26610
	Bromodomain 4	AT1G61215
	ELM2 domain-containing protein	AT2G03470
	Histone H2A deubiquitinase (DUF3755)	AT3G07565
	Duplicated homeodomain-like superfamily protein	AT3G12560
	TRF-like 9	AT3G46590
	TRF-like 1	AT4G11400
	ARID/BRIGHT DNA-binding , ELM2 domain and myb-like	AT5G03780
	TRF-like 10	AT5G13820
	Telomeric DNA binding protein 1	AT5G59430
	Telomeric repeat binding protein 1	AT2G46280
	TGF-beta receptor interacting protein 1	AT2G46290
	Transducin/WD40 repeat-like superfamily protein	AT5G01770
	Regulatory-associated protein of TOR 2 (RAPTOR2)	AT5G51800
	Protein kinase superfamily protein	AT2G21920

The percentage of TFs across genomes ranged from 5 to 8%, which is in line with a previous estimation from 95 eudicot species (Jin et al., 2017). *Gl. soja* and *Se. moellendorffii* showed the highest and lowest number of TFs, respectively (**Figure 6**). Legumes typically showed greater number of TFs than non-legumes (**Figure 6**), although the variation in these fractions indicates that some TF expansions play specific roles in particular lineages.

To better understand the different proportion of TFs across genomes, we compared TF family sizes between pairs of species. All except six TF families (i.e. AP2, GRAS, B3, Nin-like, HRT-like, and Trihelix) expanded in the basal angiosperm *Am. trichopoda* in comparison to the lycophyte *Se. moellendorffii*. Although MADS TFs tightly regulate flower development, their diversification has been proposed to predate the origin of angiosperms (Albert et al., 2013). We found twice more MADS genes in *Am. trichopoda* (n=34) than in *Se. moellendorffii* (n=15). In particular, the MIKC-type MADS subfamily (type II) alone has increased by five-fold, in spite of the higher rate of gene birth/death of the M-type MADS subfamily (type-I) (Nam et al., 2004; Kumpeangkeaw et al., 2019). By analyzing TF clusters (described in Chapter 3), we observed that genes from two M-type MADS clusters are exclusively present in *Am. trichopoda*, probably as a result of a lineage-specific expansion. We also observed the expansion of the GRAS family in *Am. trichopoda* (n= 44) as compared to the basal dicot *Aq. coerulea* (n= 36), which happened via lineage-specific tandem duplications (10 genes) in the former (**Figure 7**). These 10 genes belong to a single OG that does not have orthologs from other dicots except one from the basal eudicot *Aq. coerulea*. In addition, we found one more *Am. trichopoda* specific OG consisting of two GRAS genes (scaffold00007.332 and scaffold00007.335). Further, there are some remarkable expansions in few TF families in *Am. trichopoda* in comparison to *Se. moellendorffii* (e.g. HD-Zip, NAC, TCP, GATA, expanded by more than two fold), Fewer families such as Trihelix, GRAS, AP2 had slightly higher number of genes in *Se. moellendorffii* than *Am. trichopoda* suggesting a potential contribution of the zeta WGD, shared by all seed plants (Albert et al., 2013) (**Table 6**). Together, these results support a growth of the TF repertoire early in the diversification of angiosperms.

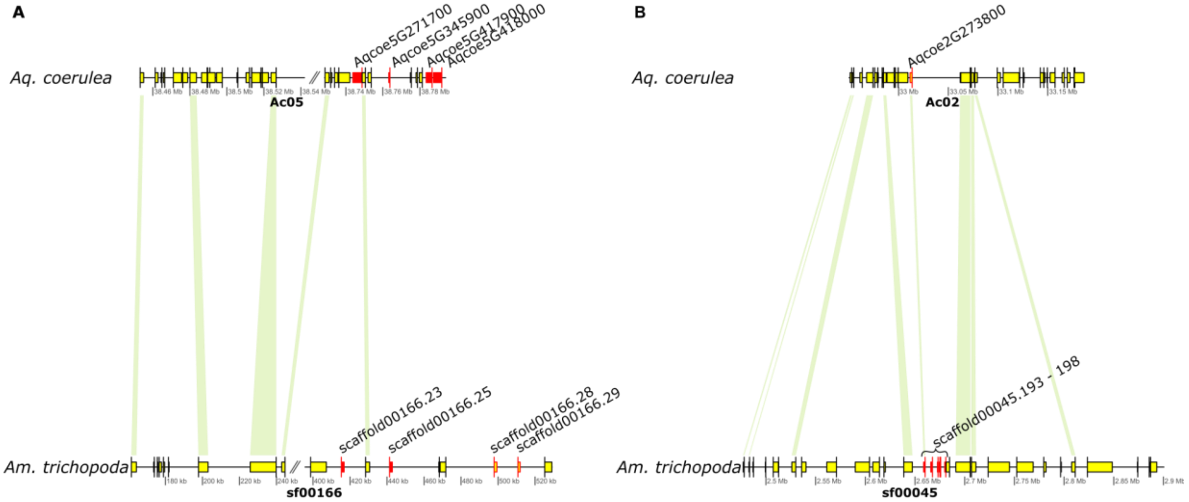

Figure 7: Schematic representation of collinear regions in *Aquilegia coerulea* and *Amborella trichopoda*, showing the expansion of GRAS TFs in the latter. In both panels (A and B), the upper and lower bars represent pseudo-chromosomes/contigs from *Aq. coerulea* and *Am. trichopoda*, respectively*.* Genes are represented by yellow arrows. Green shades connect homologous genes in the two species. Locally duplicated genes are shown as red arrows. GRAS genes are labeled with gene names and have red borders. A. *Aq. coerulea* (Ac05:38.45Mb-38.73Mb) versus *Am. trichopoda* (Sf00166:0.16Mb-0.52Mb), B. *Aq. coerulea* (Ac02:32.95Mb-33.18Mb) versus. *Am. trichopoda* (Sf0045:2.47Mb-2.90Mb).

Aq. coerulea is an ancient tetraploid and this tetraploidy was likely an important first step towards the *gamma* hexaploidy (4n+2n) that is shared by all core eudicots (Aköz and Nordborg, 2019). Nevertheless, we found some TF families that are remarkably larger in *Aq. coerulea* than in *Vi. vinifera*, such as FAR1 (Aco: 92, Vvi: 19), B3 (Aco: 104, Vvi:29), GeBP (Aco: 13, Vvi: 1), and M-Type MADS (Aco: 50, Vvi: 17) (**Figure 8; Table 4**). After the *gamma* hexaplodization event, *Vi. vinifera* has not undergone any large scale duplication event, making it a suitable reference for comparative analysis with other core eudicots (Jaillon et al., 2007; Severin et al., 2011; Wang et al., 2017). Most large families are expanded in *Ar. thaliana* and legumes in comparison to *Vi. vinifera* (**Figure 8; Table 4**). There are also some notable species-specific expansions in legumes, such as that of FAR1, B3, and M-Type MADS in *Me. truncatula* (**Figure 8**). FAR1 has also expanded 10 times in the peanuts *Ar. ipaensis* and *Ar. duranensis*. This TF family has been linked with skotomorphogenesis and photomorphogenesis in higher plants and its expansion might be related with the fructification process in peanuts (Chen et al., 2016; Lu et al., 2018). Unlike the above-mentioned expansions of MADS TFs in *Am. trichopoda,* only M-type (type-I) MADS had large expansions in all legumes as compared to *Vi. Vinifera*, as previously discussed (Nam et al., 2004; Feil et al., 2013; Kumpeangkeaw et al., 2019).

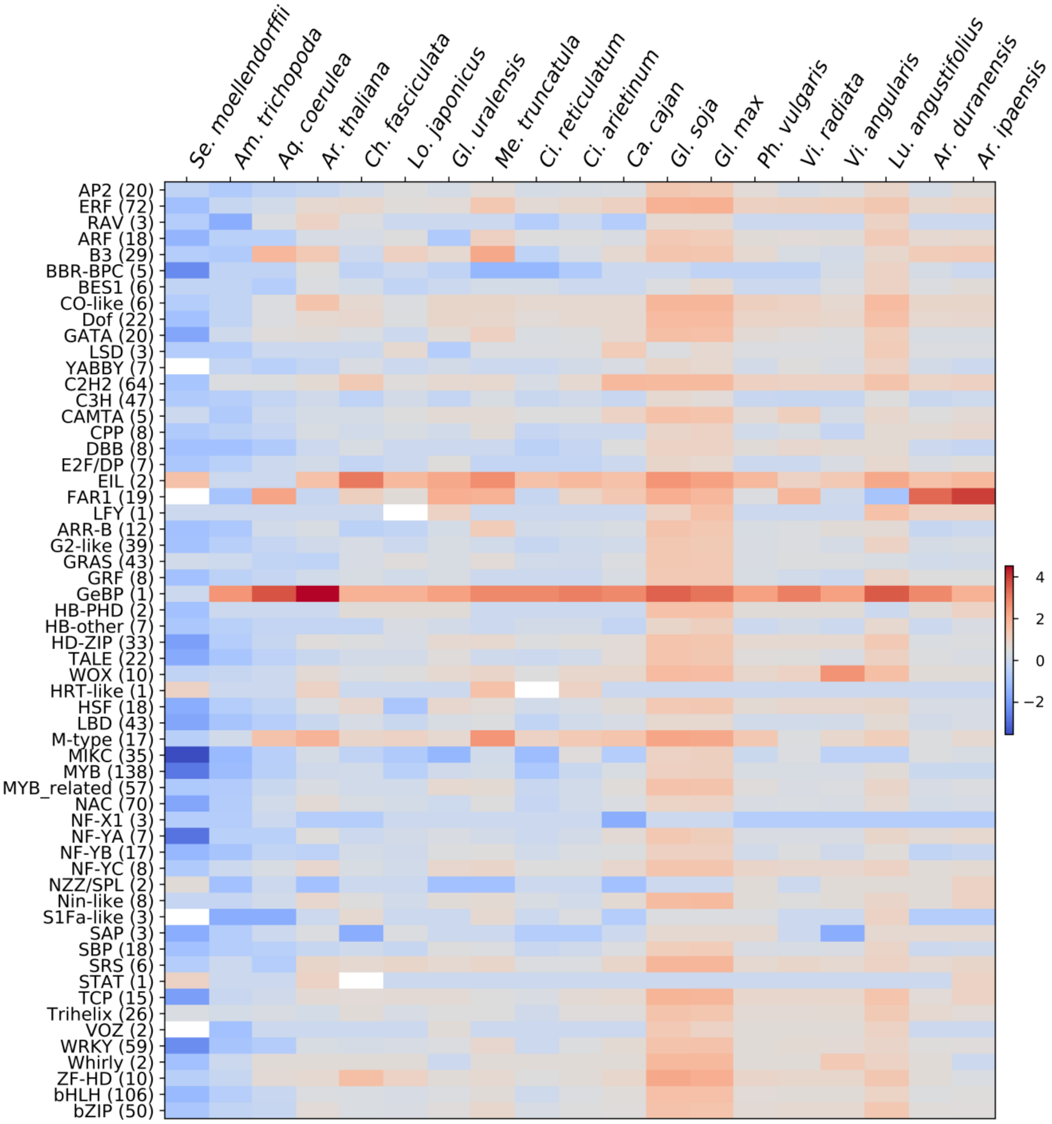

Figure 8: Ratio (in $\log_2$ scale) of the sizes of each transcription factor family each species in relation to *Vi. vinifera*. Values greater or smaller than zero represent transcription factor families that are relatively larger or smaller in a given species (in columns) in comparison to *Vi. vinifera*, respectively. The numbers in parentheses stand for the absolute size of that particular family in *Vi. vinifera*.

The *Glycine* genus has a more recent WGD that is not shared with *Phaseolus*. Accordingly, we found an approximate ratio of 1:2 between *Ph. vulgaris* and *Gl. max* in 90% (52/58) of TF families, implying that the *Glycine* WGD has strongly contributed to the soybean TF repertoire. Nevertheless, there are also deviations from this trend, such as the NAC (*Gl. max*: 142 and *Ph. vulgaris*: 90) and

NOZZLE/SPL (*Gl. max*: 2 and *Ph. vulgaris:* 3) families. Of the 42 NAC OGs with genes from *Ph. vulgaris* and *Gl. max*, 9 had identical number of genes, indicating that there are subfamilies that rapidly reverted back to their configuration before the *Glycine* WGD, probably due to gene dosage sensitivity. We also noticed that *Gl. soja* has 38 more TFs than *Gl. max*. While sixteen TF families had identical number of genes in both species, others such as ERF (*Gl. max*: 290, *Gl. soja*: 279) and FAR1 (*Gl. max*: 68, *Gl. soja*:79) showed significant variation between cultivated and wild soybeans.

4.1.2 Segmental duplication is the prevalent mode of duplication

We used the priority order SD>TD>PD>rTE>dTE>DD to assign a single duplication mode to each pair. In legumes, more than 70% of the TFs have at least one paralog (**Table 6**). Further, it is clear that SD is the main duplication mode, supporting their origin through large scale duplication, as previously reported (Lehti-Shiu et al., 2017). *Gl. max* and *Gl. soja* are the species with the greatest number SD TFs, which comprise 77.7% of the TF repertoire in the former. Local duplications (i.e. TD and PD) had also significantly contributed to TF repertoires, particularly in *Me. truncatula*, which has 11% (241/1752) of the duplicate TFs classified as TD and PD, especially in the ERF, WRKY, and B3 families (**Appendix A.2**). Interestingly, the prevalence of TD pairs in *Me. truncatula* has also been reported in genes related to other regulatory roles, such as in the F-box family (Bellieny-Rabelo et al., 2013). Further, in *Ar. duranensis, Ar. ipaensis*, and *Vi. radiata*, nearly 7% of the duplicated TFs are derived from TD, whereas dTE duplications account for 29.8% of the duplicated TFs in *Ca. fasciculata*, especially in the MYB, NAC, and bHLH families (**Table 6; Appendix A.2**). There are also some notable differences in the prevalence of modes of duplication between closely related species. For example, local TF duplications are more frequent in *Ar. ipaensis* than in *Ar. duranensis* (**Table 6**).

Between 5.5% (188/3407, in *Gl. max*) and 60.7% (560/923, in *Am. trichopoda*) of the TFs were classified as singletons (**Table 6**). While in *Ar. thaliana* 22.58% (392/1736) of the TFs were singletons, in legumes this number ranges between 5.5 and 29% (**Table 6**). Importantly, a large fraction of these singletons remain syntenic to a reference outgroup species (**Table 7**). In *Ph. vulgaris* and *Me. truncatula*, syntenic singleton TFs were significantly more expressed than their non-syntenic counterparts (**Figure 9A**), suggesting that their greater functional conservation is associated with their genomic context. Most SD TFs were also found to be syntenic in their closest outgroup species (**Table 7**).

Table 6: Prevalence of different modes of duplication of transcription factors.

Species	Total	Singletons (%)	Duplicates (%)	SD	TD	PD	rTE	dTE	DD
Ca. cajan	1970	314 (15.94)	1656 (84.06)	548	90	21	40	397	560
Ph. vulgaris	1891	257 (13.59)	1634 (86.41)	914	99	8	16	244	353
Vi. radiata	1918	253 (13.19)	1665 (86.81)	743	115	9	24	319	455
Vi. angularis	1877	307 (16.36)	1570 (83.64)	842	74	9	14	220	411
Gl. max	3407	188 (5.52)	3219 (94.48)	2645	79	14	28	242	211
Gl. soja	3445	224 (6.5)	3221 (93.5)	2648	85	13	24	231	220
Ci. reticulatum	1332	365 (27.4)	967 (72.6)	350	45	3	24	207	338
Ci. arietinum	1660	339 (20.42)	1321 (79.58)	521	101	7	17	267	408
Me. truncatula	2182	430 (19.71)	1752 (80.29)	648	209	32	27	235	601
Gl. uralensis	1738	403 (23.19)	1335 (76.81)	482	36	3	19	281	514
Lo. japonicus	1514	428 (28.27)	1086 (71.73)	195	39	6	25	345	476
Lu. angustifolius	2493	223 (8.95)	2270 (91.05)	1654	51	0	0	0	565
Ar. ipaensis	2073	365 (17.61)	1708 (82.39)	336	113	21	32	434	772
Ar. duranensis	1902	363 (19.09)	1539 (80.91)	410	95	5	27	373	629
Ch. fasciculata	1709	426 (24.93)	1283 (75.07)	186	36	2	36	383	640
Ar. thaliana	1736	392 (22.58)	1344 (77.42)	707	81	13	15	152	376
Vi. vinifera	1274	478 (37.52)	796 (62.48)	303	82	13	5	193	200
Aq. coerulea	1376	552 (40.12)	824 (59.88)	130	86	14	0	0	594
Am. trichopoda	923	560 (60.67)	363 (39.33)	14	33	4	0	0	312
Se. moellendorffii	588	278 (47.28)	310 (52.72)	79	10	4	0	0	217

Abbreviations: Segmental duplicates (SD); Tandem duplicates (TD); Proximal duplicates (PD); Retrotransposon mediated duplicates (rTE); Transposon mediated duplicates (dTE); Dispersed duplicates (DD).

Table 7: Percentage of singletons within collinear regions in a reference species.

Species	Reference outgroup species	Singletons	Singletons in collinear regions	SD	SD in collinear regions in the reference outgroup
Ca. cajan	*Vi. vinifera*	314	132 (42%)	548	430 (78%)
Ph. vulgaris	*Vi. vinifera*	257	142 (55%)	914	750 (82%)
Vi. radiata	*Vi. vinifera*	253	120 (47%)	743	595 (80%)
Vi. angularis	*Vi. vinifera*	307	161 (52%)	842	659 (78%)
Gl. max	*Ph. vulgaris*	188	93 (49%)	2645	2546 (96%)
Gl. soja	*Ph. vulgaris*	224	105 (47%)	2648	2546 (96%)
Ci. reticulatum	*Vi. vinifera*	365	199 (55%)	350	286 (81%)
Ci. arietinum	*Vi. vinifera*	339	201 (59%)	521	414 (79%)
Me. truncatula	*Vi. vinifera*	430	184 (43%)	648	520 (80%)
Gl. uralensis	*Vi. vinifera*	403	170 (42%)	482	376 (78%)
Lo. japonicus	*Vi. vinifera*	428	177 (41%)	195	154 (78%)
Lu. angustifolius	*Vi. vinifera*	223	73 (32%)	1654	1114 (67%)
Ar. ipaensis	*Vi. vinifera*	365	171 (47%)	336	275 (81%)
Ar. duranensis	*Vi. vinifera*	363	152 (42%)	410	327 (79%)
Ch. fasciculata	*Vi. vinifera*	426	144 (34%)	186	95 (51%)
Ar. thaliana	*Vi. vinifera*	392	226 (58%)	707	512 (72%)
Vi. vinifera	*Aq. coerulea*	478	295 (62%)	303	285 (94%)
Aq. coerulea	*Am. trichopoda*	552	177 (32%)	130	65 (50%)
Am. trichopoda	*Se. moellendorffii*	560	0 (0%)	14	0 (0%)

The significant number of syntenic SD TFs derived mostly by gene retention after successive WGD events (**Table 6**). As discussed above, gene duplicability, the ability of a duplicate pair to remain duplicate, is non-random and biased towards specific gene families, including TFs (Lynch and Conery, 2000; Davis and Petrov, 2004; Li et al., 2016). We analyzed TF duplicability using *Gl. max* TFs from syntenic blocks that survived the 58 mya and the 13 mya WGD events. We used intra-species collinear blocks to identify *Gl. max* SD TFs that correspond to single syntenic regions in *Ph. vulgaris*. We used a maximum Ks threshold of 0.4 to filter the *Gl. max* SD pairs that likely emerged in the 13 mya WGD (Schmutz et al., 2010). Nearly 81% (1808/2230) of the *Gl. max* SD TFs within that Ks range had a syntenic gene in *Ph. vulgaris*. Further, 75% (676/904) of such *Ph. vulgaris* orthologs had a single *Vi. vinifera* syntenic ortholog. In both cases we found that bHLH family had the highest number of syntenic gene pairs (*Gl. max-Ph. vulgaris:* 99 pairs and *Ph. vulgaris-Vi. Vinifera:* 43 pairs). Conversely, only 16% (15/93) of the *Gl. max* syntenic singleton TFs correspond to single genomic regions in *Ph. vulgaris* and *Vi vinifera*. We hypothesize that these genes do not only depend on the conservation of a local genomic context, but are also sensitive to gene dosage. Our results clearly illustrate the high duplicability of most TF families in soybean and further support the impact of two WGD events that account for a prominent fraction of the TF repertoire of this species.

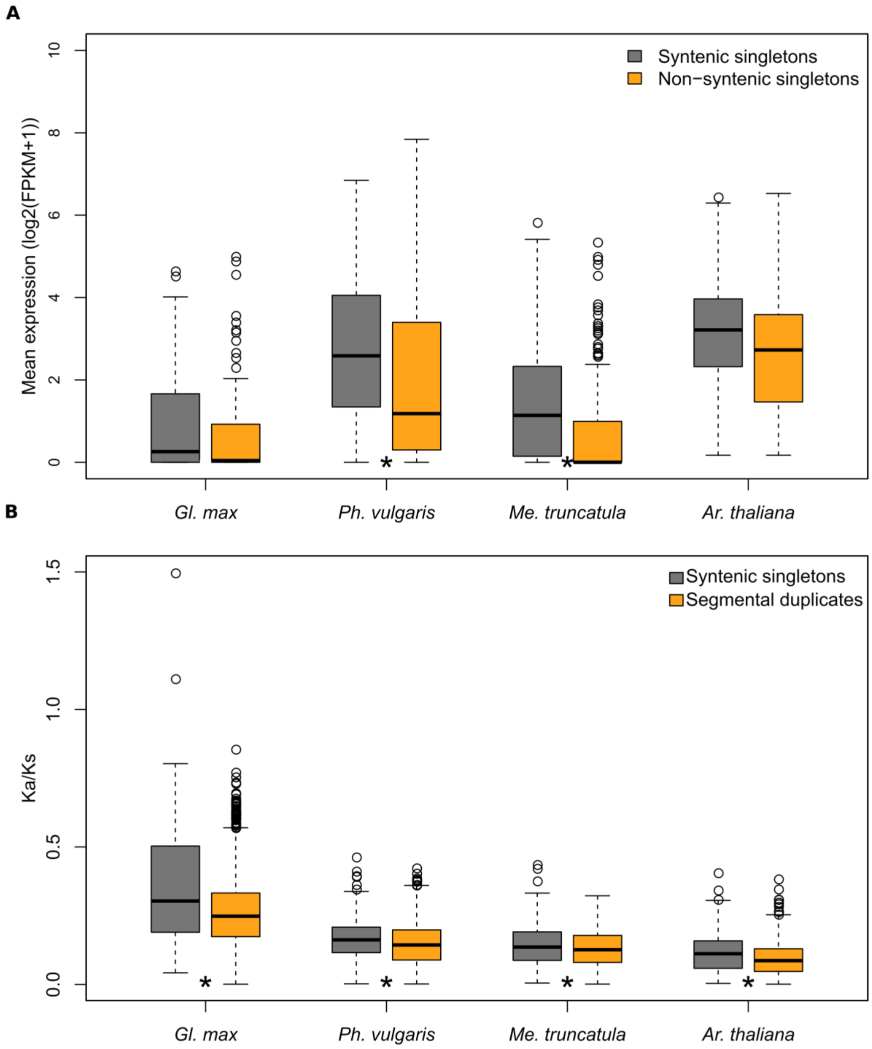

Figure 9: Expression levels and Ka/Ks ratio of singleton transcription factors. A. Expression (in FPKM) of syntenic- and non-syntenic singletons in three legume species and in Arabidopsis thaliana. B. Ka/Ks distribution of syntenic singletons and syntenic segmental duplicates. Syntenic singletons are transcription factors genes, without a close paralog, that are located in a syntenic region in a reference outgroup. Segmental duplicates are paralogous transcription factors with preserved synteny in the same genome, as well as in the genome of a reference outgroup. *Phaseolus vulgaris* was used as reference for *Glycine max* and *Vitis vinifera* was used as reference for the other three species. Statistical significance test was performed using the Mann-Whitney U test and asterisk (*) mark indicates p-value < 0.05.

4.1.3 Selection pressure on TF duplicates

We also estimated non-synonymous/synonymous mutation ratios (Ka/Ks) between singleton and SD TFs with preserved synteny in an outgroup species. Orthologs from SD pairs had significantly lower Ka/Ks value than the singleton orthologs (**Figure 9B**), leading us to hypothesize that these genes are under strong purifying selection due to their involvement in intricate regulatory systems emerging from the WGD events. Similar observations on the strong negative selection of duplicated genes have been also reported in other species (Davis and Petrov, 2004; Jordan et al., 2004).

4.1.4 Diversity in TF OGs

Many TF families explored here are broad and diversified, often comprising multiple sub-groups, such as bHLH (Pires and Dolan, 2010), MYB (Du et al., 2012), and ERF (Nakano, 2006). To obtain an overview of the diversification of plant TF families, we assigned them to OGs by using all-vs-all reciprocal BLASTP search, followed by Markov clustering. For example, AP2 had 28 clusters, labeled as AP:1 to AP:28. We found 1557 TF OGs from the 58 TF families reported above. Nearly 9% (144/1557) of these OGs had no members from legume species, whereas 29% (452/1557) were legume-specific, and 43% (672/1557) had genes from at least 10 species. Expectedly, larger families had more OGs, such as bHLH, C2H2, and MYB, with more than 100 OGs each. Conversely, a few families diverted from this trend, such as SAP (65 genes and 7 OGs) and EIL (143 members and 13 OGs) (**Figure 10**).

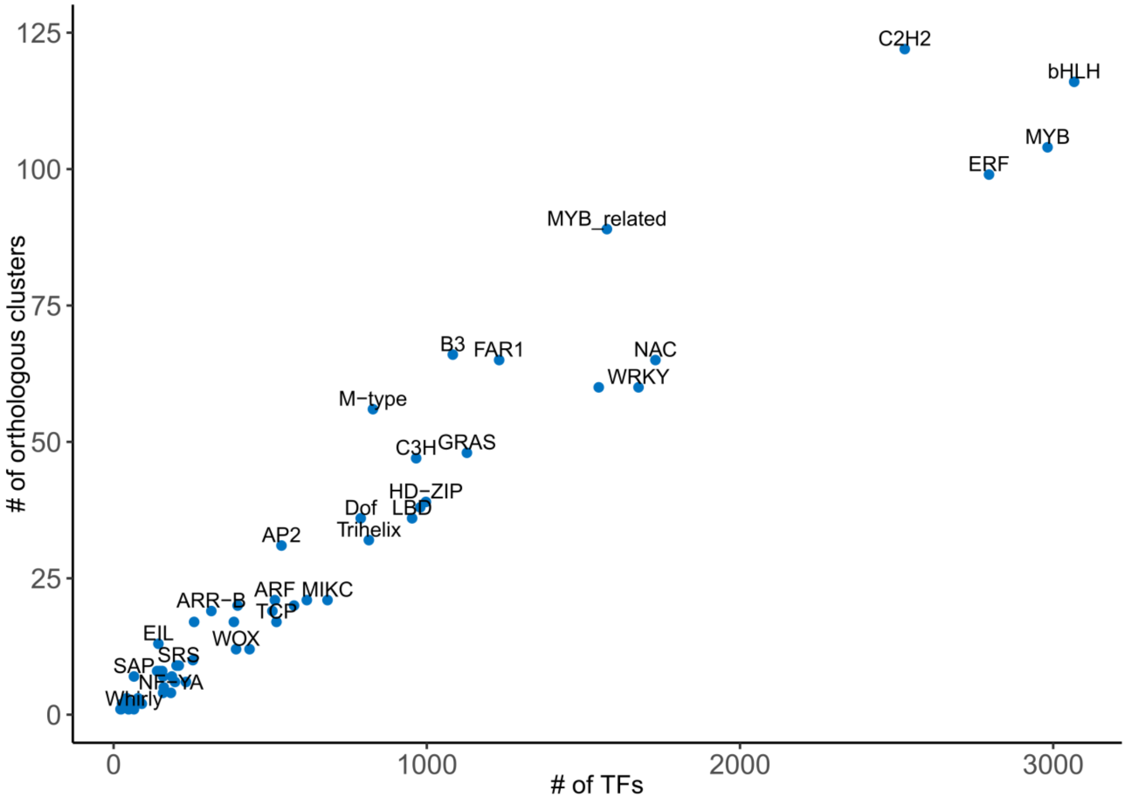

Figure 10: Number of orthologous groups and number of members in each transcription factor family.

4.1.5 Large scale duplication events correlate with increase in TF copy number

To investigate the evolution of TF families in more detail, we analyzed the number of genes per OG in each species using CAFE (v.4.2) (Bie et al., 2006; Han et al., 2013), as described above. We used 672 TF OGs with sufficient variation in number of genes per species (statistical variance ≥ 0.5) and containing genes from at least 10 species. For example, 10 out of 31 AP2 clusters were used for rate estimation. The results obtained with CAFE largely confirm the general trend for TF gain upon WGD (**Figure 11**), which is in line with the previous observation on correlation between SD, the retention of paralogous TF pairs, and intraspecies synteny. This trend can be exemplified by the nodes representing the legume and *Glycine* ancestors, which have a high number of expanded TF families (**Figure 11**).

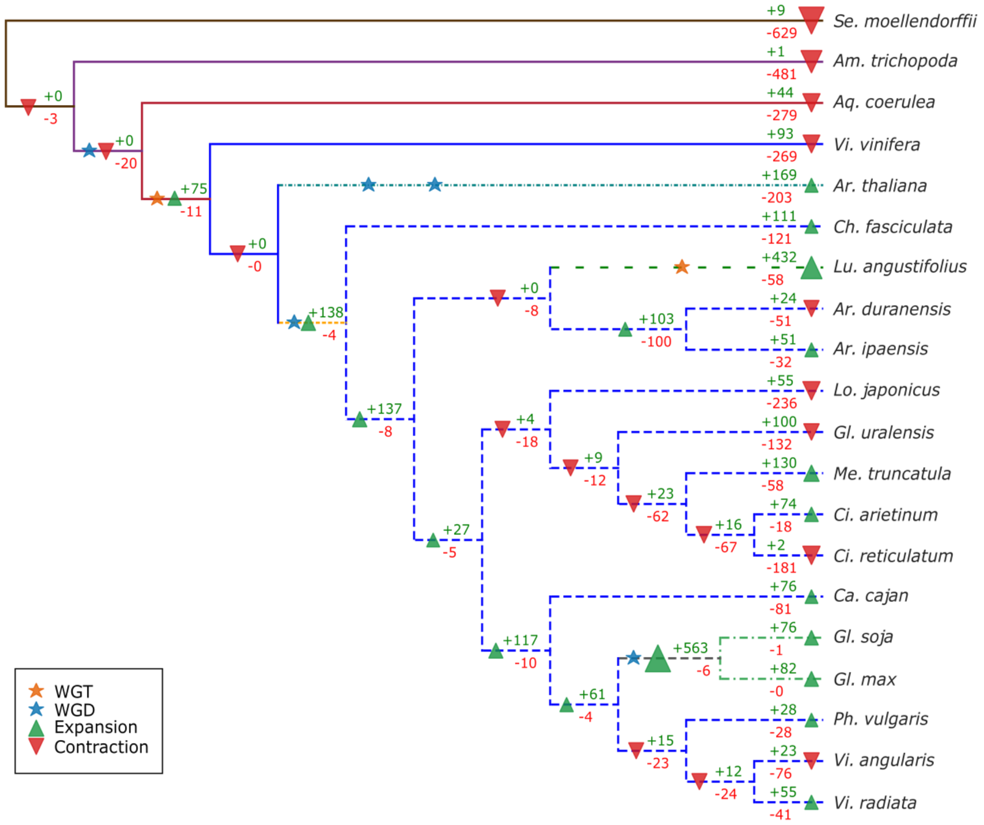

Figure 11: Species tree showing number of transcription factor orthologous groups that gained or lost genes. We used different rates of evolution in different lineages, which are represented as branch styles/colors. Known polyploidization events are marked with stars. Green and red triangles refer to nodes with more expansions and contractions, respectively. Numbers of expanded and contracted orthologous groups are shown in green and red, respectively.

To better understand the nature of TF gains, we analyzed the impact of the legume and *Glycine* WGDs in the TF repertoires of *Gl. max* and *Gl. soja*. Firstly, we analyzed the 138 OGs (from 34 TF families) that expanded in legumes in comparison to non-legumes (**Figure 11**). If all 1,557 OGs are considered, an average of 0.21 genes were gained per OG in legumes, in contrast to 1.09 genes in the 138 expanded OGs. Secondly, we identified rapidly evolving OGs (10 of 138; 7.25%) (**Table 8**), which are those with significant gene gain or loss rate (p-value < 0.05) (**Table 9**). In these 10 OGs, the average rate of gene gain was found to be 2.0, nearly twice of that observed in the 138 expanded OGs and 10 times of that observed for the complete set of 1,557 OGs.

Table 8: Number of genes and prevalence of modes of duplication in TF orthologous groups that expanded in legumes.

Species	Total	SD	TD	PD	rTE	dTE	DD
Ca. cajan	38	14	3	0	0	7	14
Ph. vulgaris	36	14	6	1	0	7	8
Vi. radiata	36	15	4	1	0	12	4
Vi. angularis	38	14	9	0	0	9	6
Gl. max	67	48	5	3	0	10	1
Gl. soja	69	50	2	3	0	13	1
Ci. reticulatum	22	6	0	0	0	7	9
Ci. arietinum	24	11	0	2	0	1	10
Me. truncatula	39	11	4	2	0	12	10
Gl. uralensis	31	12	0	0	0	4	15
Lo. japonicus	28	6	2	0	0	13	7
Lu. angustifolius	60	42	6	0	0	0	12
Ar. ipaensis	31	11	4	1	0	9	6
Ar. duranensis	32	9	5	1	0	10	7
Ch. fasciculata	34	2	6	0	0	13	13

Abbreviations: Segmental duplicates (SD); Tandem duplicates (TD); Proximal duplicates (PD); Retrotransposon mediated duplicates (rTE); Transposon mediated duplicates (dTE); Dispersed duplicates (DD).

Table 9: Orthologous groups with significantly (p-value < 0.05) rapid expansion in legumes.

Description	ARF:1	ARF:2	bHLH:5	bHLH:12	ERF:10	LBD:8	M-type:1	MYB:1	MYB:25	NAC:3
Ca. cajan	4	2	3	3	4	3	7	5	3	4
Ph. vulgaris	3	3	4	2	4	3	5	6	2	4
Vi. radiata	4	3	4	2	3	4	4	5	3	4
Vi. angularis	3	3	4	2	4	3	5	8	3	3
Gl. max	5	6	8	6	5	4	14	9	5	5
Gl. soja	6	7	8	5	6	4	14	9	5	5
Ci. reticulatum	3	3	2	4	3	1	4	1	1	3
Ci. arietinum	3	3	2	4	3	2	4	1	1	3
Me. truncatula	3	3	3	4	2	2	7	5	7	3
Gl. uralensis	0	3	3	2	4	4	6	2	3	4
Lo. japonicus	3	3	3	3	3	2	4	4	1	3
Lu. angustifolius	5	5	6	5	5	6	18	2	3	5
Ar. ipaensis	4	3	4	2	3	3	4	3	1	5
Ar. duranensis	4	3	4	2	2	2	5	3	2	5
Ch. fasciculata	3	3	4	3	3	3	3	4	5	3
Ar. thaliana	1	0	0	0	1	1	0	0	0	1
Vi. vinifera	1	1	1	1	0	2	1	1	0	2
Aq. coerulea	1	1	1	1	1	1	0	0	0	1
Am. trichopoda	0	0	0	1	0	0	0	0	0	0
Se. moellendorffii	0	0	0	0	0	0	0	0	0	0

We analyzed the most prevalent modes of duplication in the 138 OGs that expanded in legumes and found that a major fraction of them emerged via SD. While SDs comprise more than 80% of the TFs in species with more recent WGDs (i.e. *Gl. max, Gl. soja*, and *Lu. angustifolius*) (**Figure 12**), several SD pairs might have lost collinearity after the legume WGD and were assigned as DDs. When inspecting the Ks distributions of the paralogous pairs from the 138 OGs that expanded in legumes, we found SD Ks distributions corresponding to both, the legume and *Glycine* WGDs (**Figure 12**) (Schmutz et al., 2010; Cannon, 2013), suggesting that a fraction of the TFs that expanded in the legume WGD subsequently duplicated for a second time in the *Glycine* WGD. Deviations from this range were observed for *Me. truncatula, Arachis* spp., *Cicer* spp. and, *Ch. fasciculata,* as previously reported (Cannon et al., 2010; Varshney et al., 2013; Tang et al., 2014; Chen et al., 2016). DD paralogs have a more dispersed Ks distribution than that SD, although their Ks distributions also indicate that several DD pairs were likely generated by SD with subsequent loss of collinearity (**Figure 12**). Collectively, these results support the association between the expansion of legume-specific TF expansions and the WGD event that took place 58 mya.

To further explore the functional relevance of legume TF expansions, we analyzed gene expression patterns across multiple tissues from *Me. truncatula*, *Ph. vulgaris*, and *Gl. max* (**Figure 13; Figure 14**). Strikingly, the 138 OGs that expanded in legumes are enriched in genes with preferential expression in nodules (Fisher's Exact Test, p-values = 1.7×10^{-4} and 1.4×10^{-5} for *Me. truncatula* and *Gl. max*, respectively) and roots (Fisher's Exact Test, p-values = 1.4×10^{-9} and 5.9×10^{-3} for *Me. truncatula* and *Ph. vulgaris*, respectively). These results indicate that the recruitment of these genes predate the emergence of nodulation in legumes and might have played roles in the root physiology involved with this process.

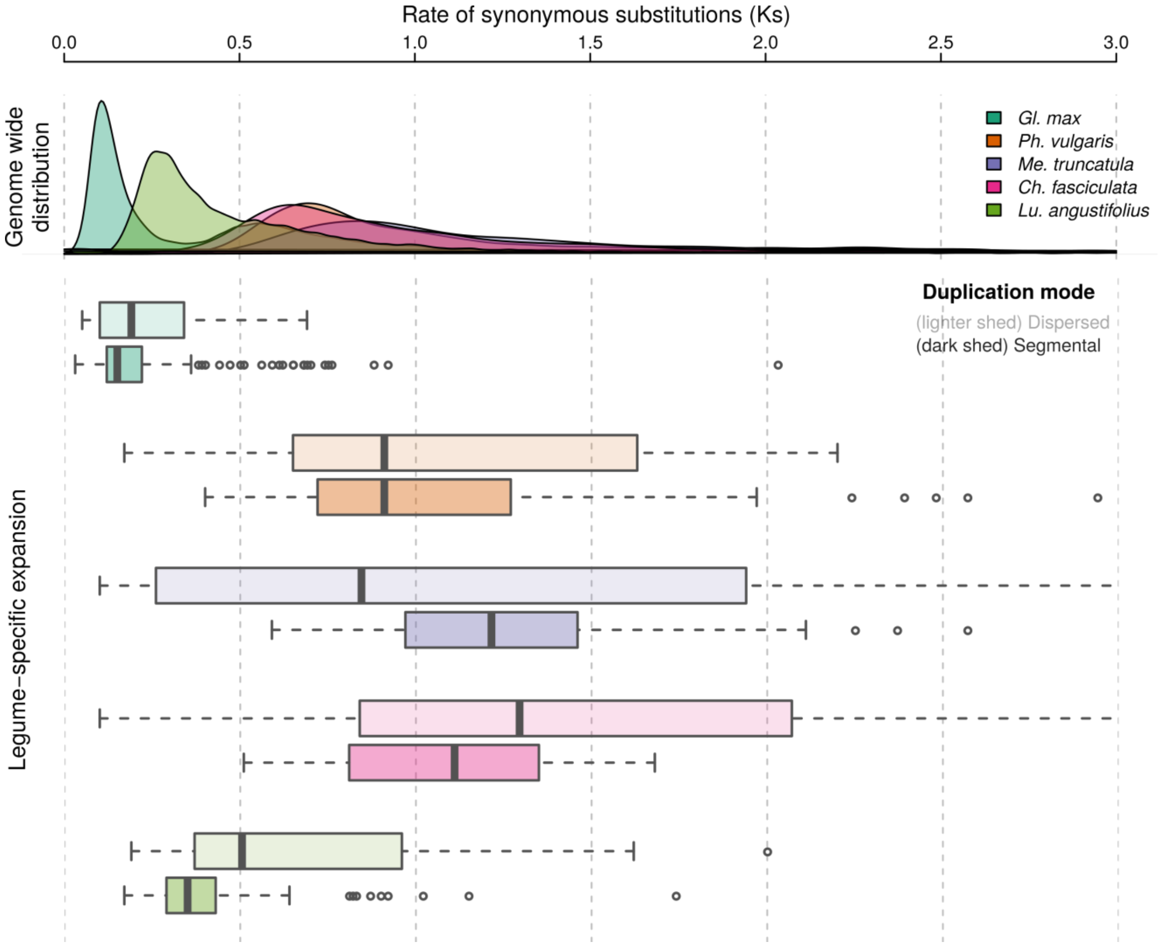

Figure 12: Distribution of synonymous substitution rates (Ks) of 138 orthologous groups with gene gain in legumes. Genome-wide Ks distributions are shown as density plots on the top panel. The bottom panel shows Ks distributions of segmental and dispersed duplicate gene pairs.

Next, we integrated phylogenetic reconstructions of the 10 rapidly expanded OGs with gene expression data and found three very interesting groups (i.e. bHLH:12, M-type:1, ERF:10) (**Table 9**). The bHLH:12 OG showed significantly higher expression during nodule development in *Me. truncatula, Ph. vulgaris*, and *Gl. max* (**Figure 13B** and **Figure 14**). This OG includes two SD pairs of *Me. truncatula* bHLHs, Medtr4g087920-Medtr2g015890 and Medtr4g079760-Medtr2g091190 with Ks values of 1.0523 and 0.8292, respectively. *Ph. vulgaris* and *Gl. max* orthologs of these genes were also more expressed in roots and nodules than in other tissues (**Figure 13 A**).

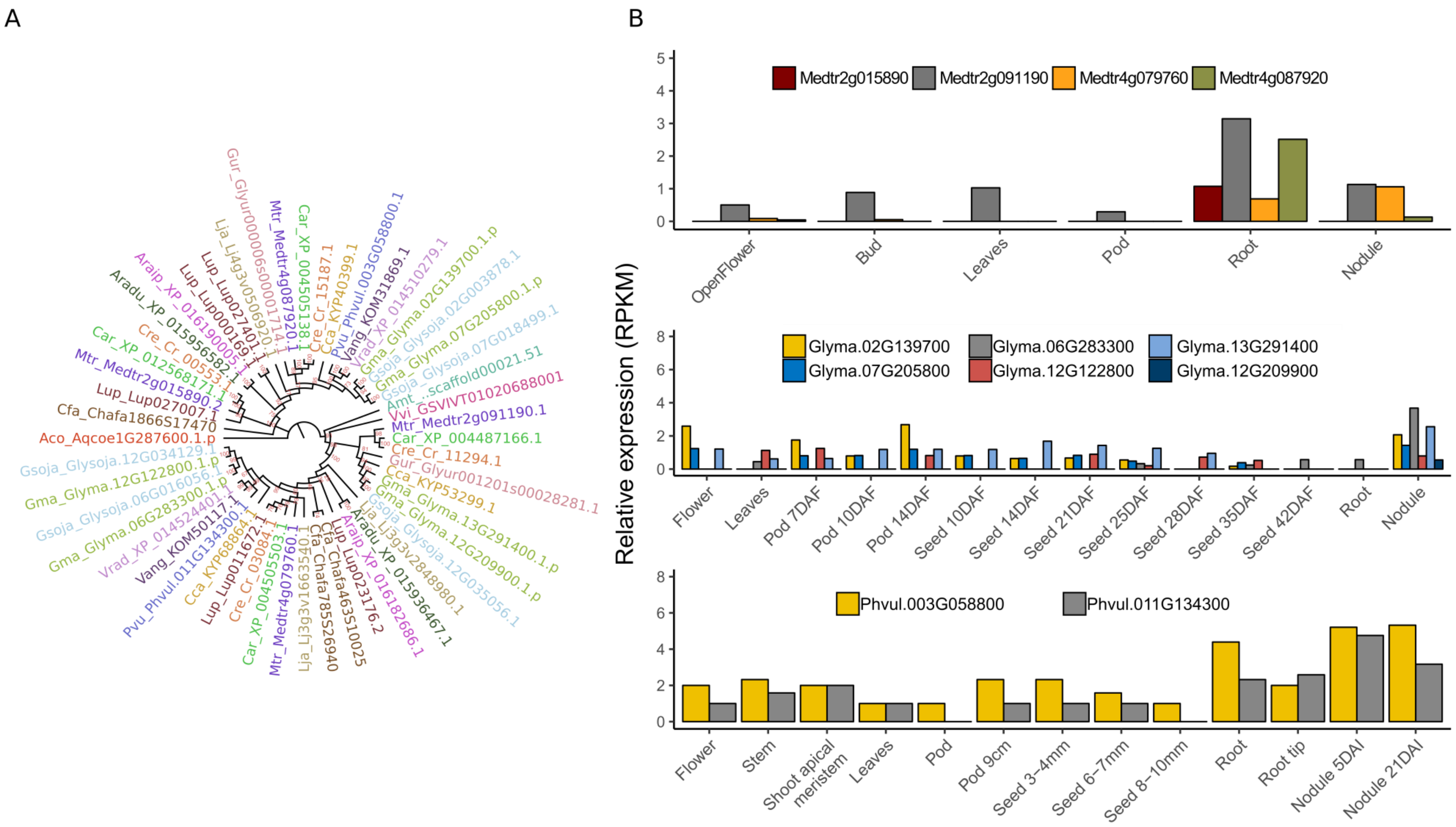

Figure 13:Lineage specific expansion of bHLH:12 OG in legumes and their expression patterns. A. Phylogenetic reconstruction of the orthologous group bHLH:12, which is expanded in legumes. B. Gene expression patterns of bHLH:12 genes in *Me. truncatula* (BioProject: PRJNA80163), *Gl. max* (Libault et al., 2010b) and *Ph. vulgaris* (O'Rourke et al., 2014), showing a trend for greater expression in roots and nodules.

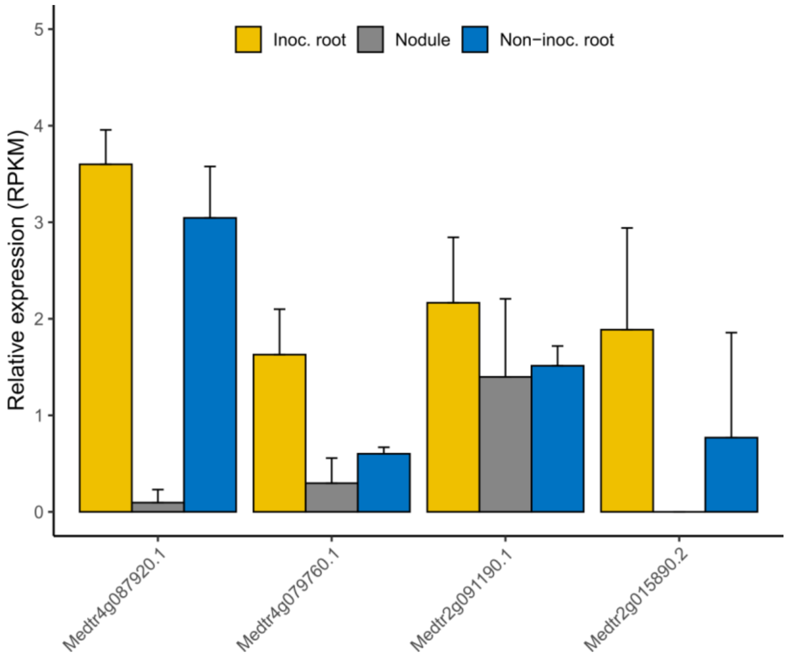

Figure 14: Normalized expression of genes from the bHLH:12 orthologous group. Gene expression patterns of four *Me. truncatula* bHLH genes. RNA-seq data were obtained from a previous study (Boscari et al., 2013). Non-inoc. Root: nitrogen-starving roots; Inoc.-root: roots inoculated with *Sinorhizobium meliloti*; Nodule: root nodules.

Another interesting OG encodes TFs from the M-type MADS family (M-Type:1). This OG has independently expanded in different species (**Figure 15 A**), including a major expansion in *Lu. angustifolius*. Nine of 14 *Gl. max* genes in this cluster emerged within the *Glycine* genus, including one gene (Glyma.03G083700.1) with preferential expression in seeds and flowers (**Figure 16**). The two *Ph. vulgaris* orthologs (Phvul.006G077700 and Phvul.006G077800.1) showed seed-specific expression, suggesting their importance in seed development (**Figure 16**). Interestingly, these *Ph. vulgaris* genes originated from ancestral tandem duplication, as this organization is also found in other legumes (**Figure 15 B**). Although this OG lacked an *Ar. thaliana* member, the closest *Ar. thaliana* homologs include *AT5G27810*, *AT1G22590* (AGAMOUS-LIKE 87, *AGL87*), and *AT5G48670* (AGAMOUS-LIKE80, *AGL80*). Importantly, *AGL80* has been shown to be responsible for central cell and endosperm development in *Arabidopsis* (Portereiko et al., 2006).

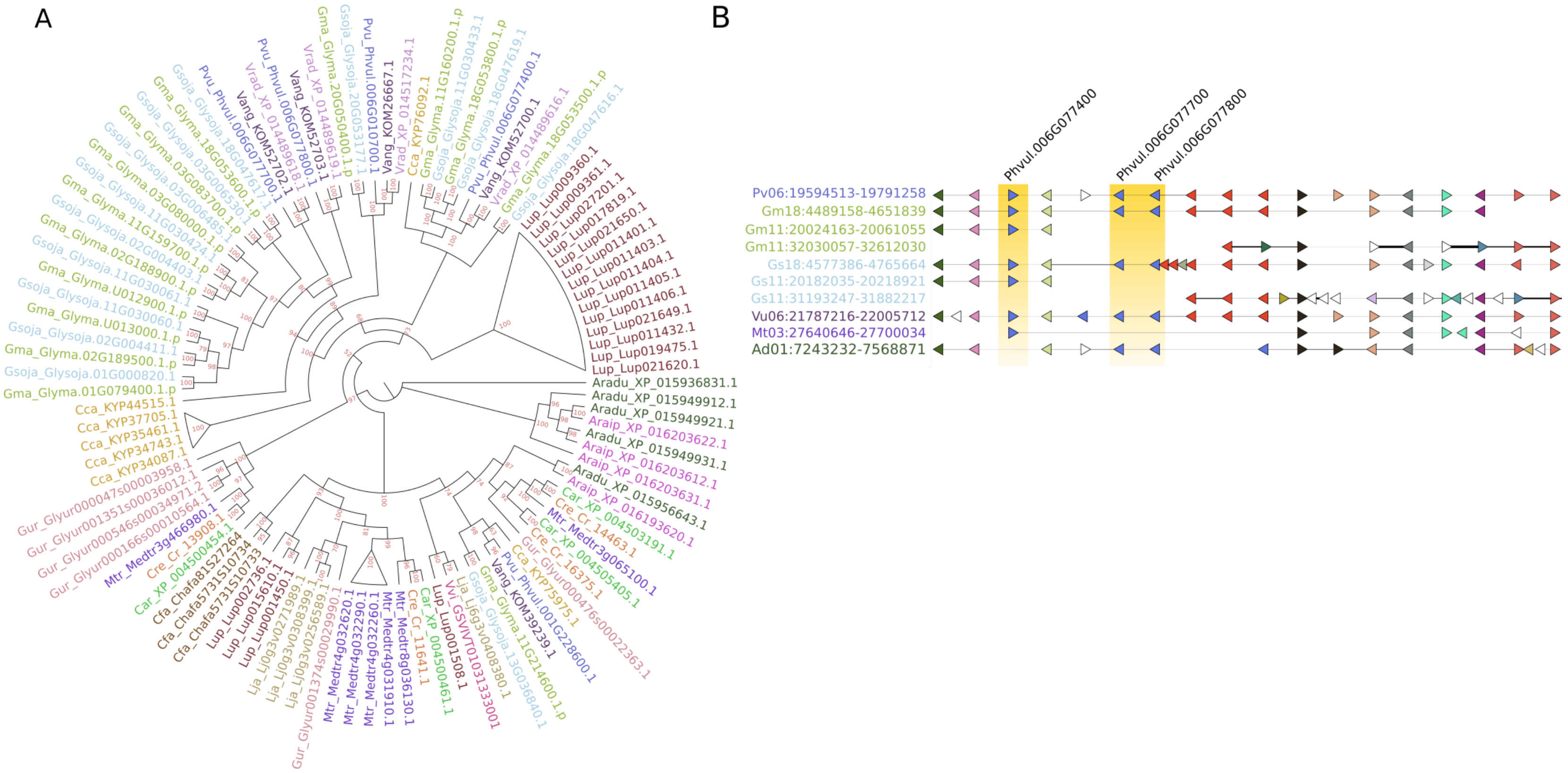

Figure 15: Phylogenetic tree of the M-type:1 orthologous group. Genes from same species are labeled with similar colors. Clades comprising genes from the same species were collapsed. The Vigna species used in our analyses are not available in Genome Context Viewer at Legume Information System (Cleary et al., 2017). Hence, we used a close species, *Vigna unguiculata*, to illustrate the syntenic relationships.

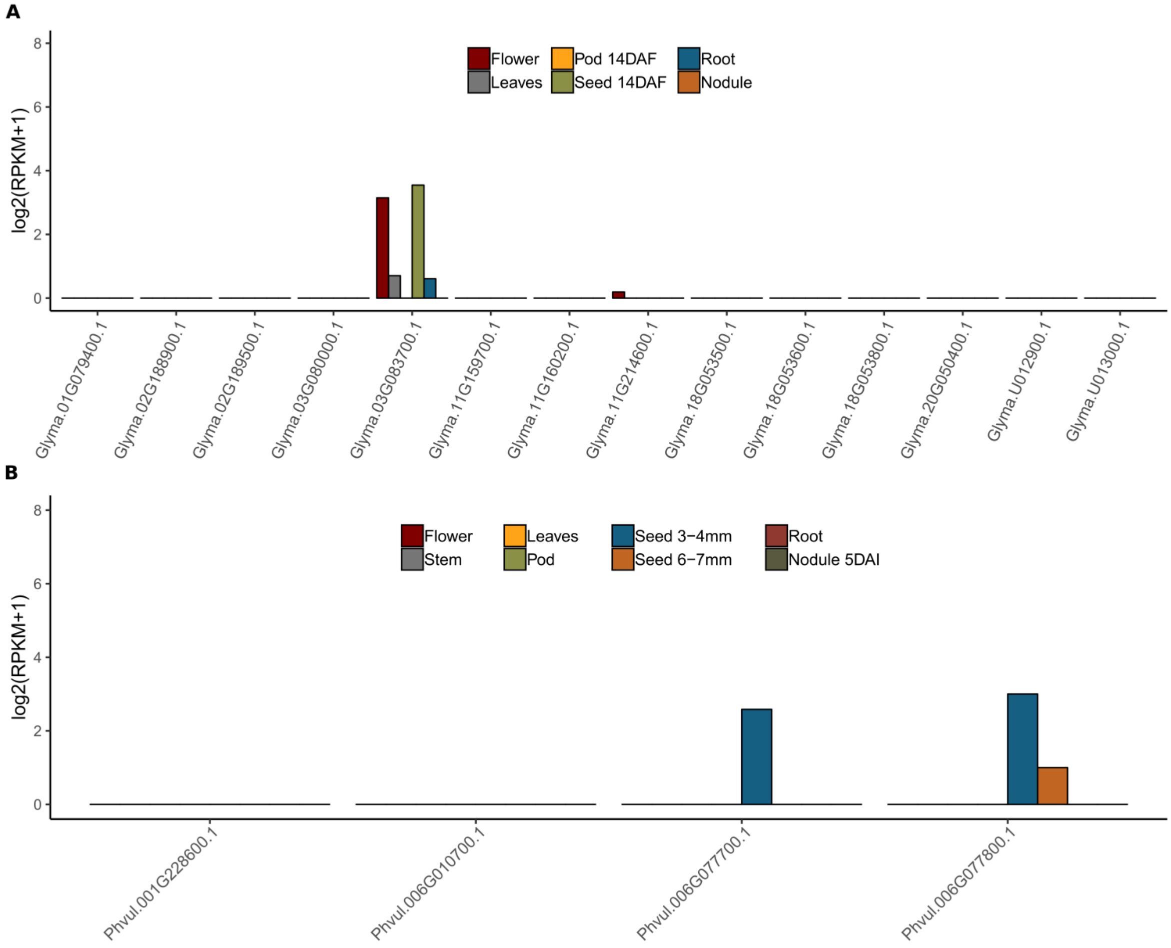

Figure 16: Seed specific expression levels of M-type:1 genes. A and B represent expression levels of M-type:1 genes across various tissues in *Gl. max* as observed in (Severin et al., 2010) and *Ph. vulgaris* members (O'Rourke et al., 2014), respectively.

We also analyzed two ERF (*ethylene response factor*) OGs (ERF:10 and ERF:18) containing genes playing critical roles in nodulation. Of these two, only ERF:10 was among the 10 rapidly expanded OGs. Manual curation revealed that ERF:10 and ERF:18 comprise ERF *required for nodule differentiation* (EFD) and ERF *required for nodulation* (ERN) genes, respectively. ERN and EFD genes regulate nodulation in *Me. truncatula* (Vernie et al., 2008; Young et al., 2011; Cerri et al., 2012). Three of the four *MtERNs* (i.e. Medtr7g085810.1, Medtr6g029180.1, and Medtr8g085960.1) had relatively higher expression after inoculation than in roots or nodules, supporting their critical role in nodule development (**Figure 17A**). The biased expression towards nodules and root tissues are also observed in *Ph.*

vulgaris and *Gl. max* orthologs (**Figure 17A**). Of the two *MtEFDs*, Medtr4g008860 and Medtr3g106290 were more expressed in nodules and inoculated root hairs, respectively. As observed in the previous OGs, *Ph. vulgaris* and *Gl. max* ERNs are also more expressed in roots and nodules than in aerial tissues (**Figure 17B-C**).

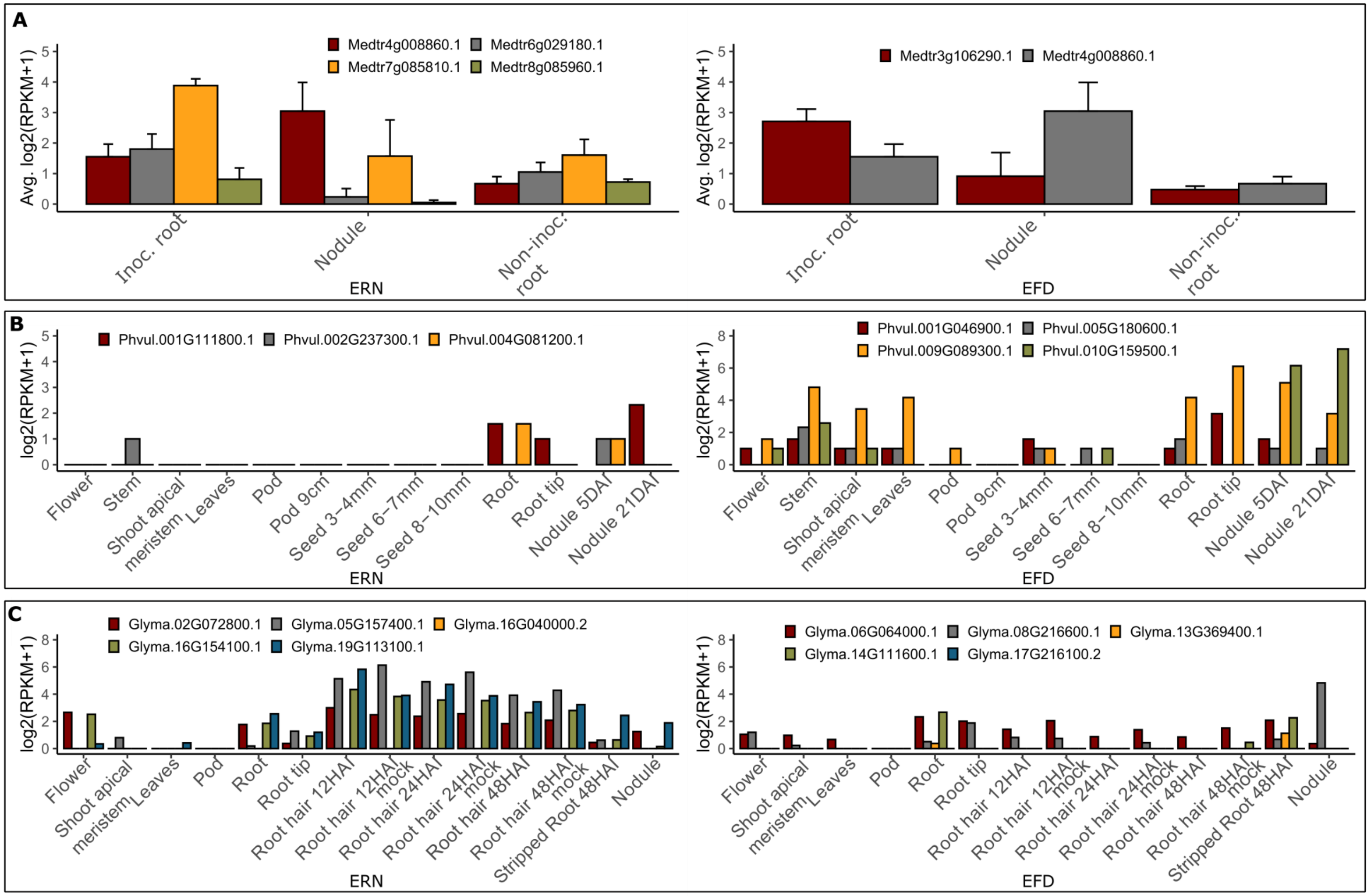

Figure 17: Expression levels of ERF required for nodulation (ERN) and nodule differentiation (EFD) genes in *Me. truncatula.* (A)*, Phaseolus vulgaris* (B) and *Gycine. max* (C). In all three panels, left and right images represent ERN and EFD genes, respectively.

4.1.6 Association between quantitative traits and *Glycine max* specific TFs

The *Glycine* node had the largest number of expanded OGs, with 36% (563/1557) (**Figure 11**) of them expanding by an average rate of 1.67 genes per OG. Out of these, 57 had rapidly expanded (p-value < 0.05) with an average rate of 2 genes per OG. In comparison to the *Glycine* node, 76 and 82 OGs expanded in *Gl. soja* and *Gl. max*, respectively. Of the OGs that expanded in *Gl. max*, 79% (65/82) showed significant expansions (p-value < 0.05), with average rate of 1.44. Among these families, ERF (6 OGs), MYB (5 OGs), MYB_related (8 OGs), bHLH (7 OGs), and C2H2 (7 OGS) TFs gained more than 5 genes per OG. Interestingly, 59% (202/341) of the genes from expanded OGs lie within SD regions, supporting the importance of the *Glycine* WGD in shaping these families in *Gl. max*.

We explored whether some of these OGs could be related with important soybean agronomic traits. We searched the *Gl. soja* syntenic regions corresponding to the 341 *Gl. max* TFs from the 65 rapidly expanded OGs. We identified 50 TFs without a homeolog in *Gl. soja* (**Appendix A.3**), out of which only five had a syntenic ortholog in *Ph. vulgaris*. Since multiple genome assemblies are available for both *Gl. soja* and *Gl. max* (Valliyodan et al., 2019) we decided to compare other soybean assembles available on SynMap tool to verify whether the loss of these 50 TF in wild soybean is real or simply due to differences in assembly completeness or annotation quality. Comparing with the genome assembly of southern *Gl. max* line Lee (version: glyma.Lee.gnm1), we found all the 50 genes present in the genome. Comparing with SynMap results of syntenic blocks between genome assembly of *Gl. soja* W05 (used in main analysis) and *Gl. soja* accession PI 483463 (version: glyso.PI483463.gnm1) we found only 48% (24/50) genes were missing while remaining 26 genes present. Interestingly, two ERF (Glyma.20G115300, Glyma.14G161900) and one SBP (Glyma.06G205700) TFs are within previously reported chromosomal regions associated with important quantitative traits (**Figure 18**) (Fang et al., 2017). The ERF Glyma.20G115300 was located within a region associated with overall leaf size and average number of seeds per pod. The second ERF, Glyma.14G161900, is within a region associated with FA18 content and ratio in mature seeds. Finally, the SBP TF Glyma.06G205700 is within a region regulating branch density (i.e. ratio of branch number and plant height) and beginning bloom date. (**Appendix A.3**).

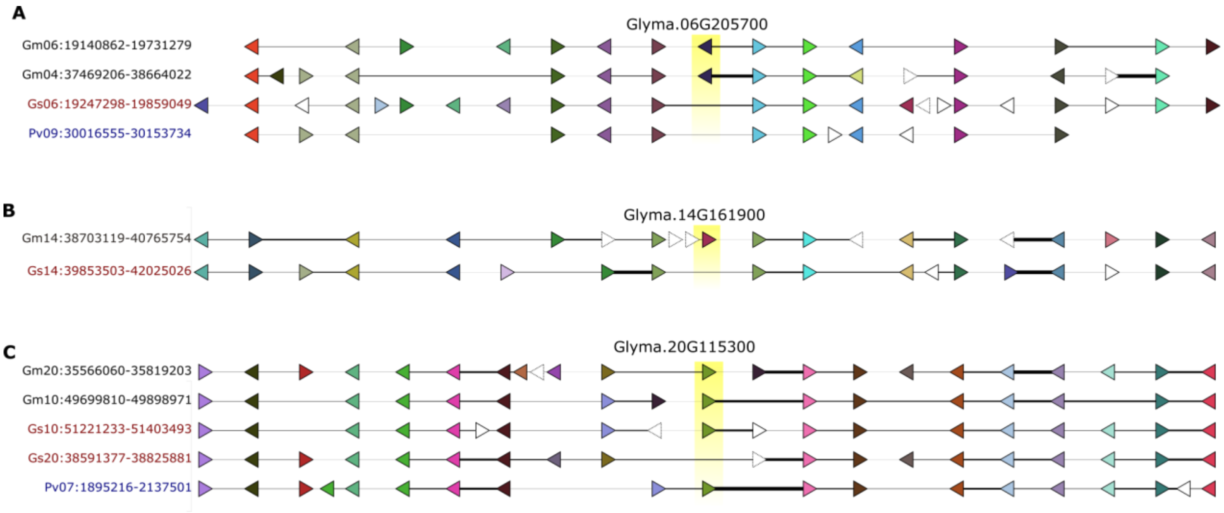

Figure 18: Microsynteny of *Gl. max* genes located within chromosomal regions associated with phenotypic traits. The genomic context was rendered on legumeinfo.org. The queried *Gycine. max* genes are on the top rows, labeled and highlighted with yellow background. Homologous genes are colored with similar colors. Singletons are in white. The thicknesses of the horizontal lines are proportional to the intergenic distances. A. *Glyma.06G205700* is located in a region associated with branch density (ratio of branch number and plant height).B. *Glyma.14G161900* is within a region associated with linolenic fatty acid (FA18) content in mature seeds. (C) *Glyma.20G115300* is within a region associated with phenotypes such as overall leaf length, shape, width, Number of four seed per pod, Ratio of four seed per pod and Ratio of two seed per pod.

5. CONCLUSION

The overall objective of this study was to access the impact of polyploidization events on the expansion of TF families throughout legumes, to identify potential association of such expansions with important traits such as nitrogen fixation and seed development.

Gene duplication events, particularly polyploidization, are thought to be an important factor of evolutionary innovation and phenotypic diversification; hence, the mechanisms governing the evolutionary fate of gene duplicates have been studied intensively (Lynch and Conery, 2000; 2003; Van de Peer et al., 2009). Because of multiple rounds of polyploidizations in past angiosperm genomes contain redundant copies of genes. In each round, most common fate of newly created duplicate copy is pseudogenization, however TF families are retained leading to their expansion various lineages. We aimed to systematically screen TF DNA-binding domains in genomes of 15 legume and 5 non-legume species. Our results suggest that the percentage of TFs ranged from 3-8% of the gene complements. Overall comparison revealed that, legumes typically has greater number of TFs than non-legumes.

Between 5.5% (*Gl. max*) and 60.7% (*Am. trichopoda*) of the TFs observed to loss their duplicate copy and as singletons. A significant fraction of the singletons remain syntenic to a reference outgroup species. Using gene expression data, in *Ph. vulgaris* and *Me. truncatula*, we found that syntenic singleton TFs were significantly more expressed than their non-syntenic counterparts. This suggests that greater functional conservation of the syntenic singletons is associated with their local genomic context. Further, comparing the Ka/Ks ratio between singleton and SD TFs with preserved synteny in an outgroup species orthologs we observed that TFs from SD are under strong purifying selection.

Results from our phylgenomic analysis unveil a profound impact of polyploidization events on the expansion of TF families throughout legumes. Expansions of major TF families are strongly associated with known WGD events in the legume (~58 mya) and *Glycine* (~13 mya) lineages, which account for a large fraction of the *Ph. vulgaris* and *Gl. max* TF repertoires. Genes from OGs showing legume WGD (~58 mya) specific expansions are preferentially expressed in roots and nodules, supporting their importance in the evolution of nodulation. If this hypothesis holds valid, it has a much wider contribution in evolution of all other organism on earth. In the context of Cr-Pg mass extinction event, the climatic condition must be very hares and soil is relatively infertile. Evolution of nodulation not only helped to fix atmospheric nitrogen to provide these legume species to survive, it also helped other plant species, in general, to survive. Development of green lineage significantly affected the composition of atmosphere making it more suitable for lineages such as primate evolution.

Further, TF expansions that happened at the *Glycine* WGD (~13 mya) include genes that were subsequently lost in the wild soybean, *Gl. soja*, including TF genes that are within *Gl. max* QTLs associated with leaf shape, area and width, proportion of FA18 in seeds, and branch density. We envisage that many more of such TFs will be associated with important traits, which could be revealed in by a more comprehensive work integrating QTL information from other genotypes and studies with our phylogenomic results.

6. REFERENCES

Aköz, G., and Nordborg, M. (2019). The Aquilegia genome reveals a hybrid origin of core eudicots. *bioRxiv*. doi: 10.1101/407973.

Albalat, R., and Cañestro, C. (2016). Evolution by gene loss. *Nature reviews. Genetics* 17**,** 379-391. doi: 10.1038/nrg.2016.39.

Albert, V.A., Barbazuk, W.B., DePamphilis, C.W., Der, J.P., Leebens-Mack, J., Ma, H., et al. (2013). The Amborella Genome and the Evolution of Flowering Plants. *Science* 342**,** 1241089-1241089. doi: 10.1126/science.1241089.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215(3)**,** 403-410. doi: 10.1016/S0022-2836(05)80360-2.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17)**,** 3389-3402.

Arabidopsis-Genome-Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408**,** 796-815. doi: 10.1038/35048692.

Azani, N., Babineau, M., Bailey, C.D., Banks, H., Barbosa, A., Pinto, R.B., et al. (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG). *Taxon* 66**,** 44-77. doi: 10.12705/661.3.

Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., dePamphilis, C., et al. (2011). The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332(6032)**,** 960-963. doi: 10.1126/science.1203810.

Bellieny-Rabelo, D., De Oliveira, E.A., Ribeiro, E.S., Costa, E.P., Oliveira, A.E., and Venancio, T.M. (2016). Transcriptome analysis uncovers key regulatory and metabolic aspects of soybean embryonic axes during germination. *Sci Rep* 6**,** 36009. doi: 10.1038/srep36009.

Bellieny-Rabelo, D., Oliveira, A.E., and Venancio, T.M. (2013). Impact of whole-genome and tandem duplications in the expansion and functional diversification of the F-box family in legumes (Fabaceae). *PLoS One* 8(2)**,** e55127. doi: 10.1371/journal.pone.0055127.

Bertioli, D.J., Cannon, S.B., Froenicke, L., Huang, G., Farmer, A.D., Cannon, E.K., et al. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat Genet* 48(4)**,** 438-446. doi: 10.1038/ng.3517.

Bie, T.D., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE : a computational tool for the study of gene family evolution. 22**,** 1269-1271. doi: 10.1093/bioinformatics/btl097.

Birchler, J.A., and Veitia, R.A. (2007). The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. *THE PLANT CELL ONLINE* 19**,** 395-402. doi: 10.1105/tpc.106.049338.

Birchler, J.A., and Veitia, R.A. (2011). Protein–protein and protein–DNA dosage balance and differential paralog transcription factor retention in polyploids.

Frontiers in Plant Genetics and Genomics 2**,** 64. doi: 10.3389/fpls.2011.00064.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* 12(7)**,** 1093-1101.

Blanc, G., and Wolfe, K.H. (2004a). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant cell* 16**,** 1679-1691. doi: 10.1105/tpc.021410.

Blanc, G., and Wolfe, K.H. (2004b). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* 16**,** 1667-1678. doi: 10.1105/tpc.021345.

Boscari, A., Del Giudice, J., Ferrarini, A., Venturini, L., Zaffini, A.L., Delledonne, M., et al. (2013). Expression dynamics of the Medicago truncatula transcriptome during the symbiotic interaction with Sinorhizobium meliloti: which role for nitric oxide? *Plant Physiol* 161(1)**,** 425-439. doi: 10.1104/pp.112.208538.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422**,** 433-438. doi: 10.1038/nature01521.

Cannon, S.B. (2013). The model legume genomes. *Methods Mol Biol* 1069**,** 1-14. doi: 10.1007/978-1-62703-613-9_1.

Cannon, S.B., Ilut, D., Farmer, A.D., Maki, S.L., May, G.D., Singer, S.R., et al. (2010). Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS One* 5(7)**,** e11630. doi: 10.1371/journal.pone.0011630.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol* 4**,** 10. doi: 10.1186/1471-2229-4-10.

Cardoso, D., de Queiroz, L.P., Pennington, R.T., de Lima, H.C., Fonty, E., Wojciechowski, M.F., et al. (2012). Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. *American Journal of Botany* 99**,** 1991-2013. doi: 10.3732/ajb.1200380.

Cerri, M.R., Frances, L., Laloum, T., Auriac, M.C., Niebel, A., Oldroyd, G.E., et al. (2012). Medicago truncatula ERN transcription factors: regulatory interplay with NSP1/NSP2 GRAS factors and expression dynamics throughout rhizobial infection. *Plant Physiol* 160(4)**,** 2155-2172. doi: 10.1104/pp.112.203190.

Chan, C., Qi, X., Li, M.-W., Wong, F.-L., and Lam, H.-M. (2012). Recent developments of genomic research in soybean. *Journal of genetics and genomics = Yi chuan xue bao* 39**,** 317-324. doi: 10.1016/j.jgg.2012.02.002.

Chen, X., Li, H., Pandey, M.K., Yang, Q., Wang, X., Garg, V., et al. (2016). Draft genome of the peanut A-genome progenitor (Arachis duranensis) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc Natl Acad Sci U S A* 113(24)**,** 6785-6790. doi: 10.1073/pnas.1600899113.

Cleary, A., Farmer, A., and Hancock, J. (2017). Genome Context Viewer: visual exploration of multiple annotated genomes using microsynteny. *Bioinformatics* 34(9)**,** 1562-1564. doi: 10.1093/bioinformatics/btx757.

Crepet, W.L. (2013). "Origin and Diversification of Angiosperms", in: *Encyclopedia of Biodiversity.* Elsevier).

Cusack, B.P., and Wolfe, K.H. (2007). Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* 24(3)**,** 679-686. doi: 10.1093/molbev/msl199.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488**,** 213-217. doi: 10.1038/nature11241.

Davis, J.C., and Petrov, D.A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* 2(3)**,** E55. doi: 10.1371/journal.pbio.0020055.

De Bodt, S., Maere, S., and Van De Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology and Evolution* 20**,** 591-597. doi: 10.1016/j.tree.2005.07.008.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1)**,** 15-21. doi: 10.1093/bioinformatics/bts635.

Dodsworth, S., Chase, M.W., and Leitch, A.R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* 180**,** 1-5. doi: 10.1111/boj.12357.

Doebley, J., and Lukens, L. (1998). Transcriptional regulators and the evolution of plant form. *The Plant cell* 10**,** 1075-1082. doi: 10.1105/tpc.10.7.1075.

Dong, Y., Yang, X., Liu, J., Wang, B.H., Liu, B.L., and Wang, Y.Z. (2014). Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. *Nat Commun* 5**,** 3352. doi: 10.1038/ncomms4352.

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

Endress, P.K., and Doyle, J.A. (2009). Reconstructing the ancestral angiosperm flower and its initial specializations. *American Journal of Botany* 96**,** 22-66. doi: 10.3732/ajb.0800047.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7)**,** 1575-1584.

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology* 18(1). doi: 10.1186/s13059-017-1289-9.

Fawcett, J.A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A* 106**,** 5737-5742. doi: 10.1073/pnas.0900906106.

Feil, R., Yoshida, T., and Kawabe, A. (2013). Importance of Gene Duplication in the Evolution of Genomic Imprinting Revealed by Molecular Evolutionary Analysis of the Type I MADS-Box Gene Family in Arabidopsis Species. *PLoS One* 8(9), e73588. doi: 10.1371/journal.pone.0073588.

Filiault, D.L., Ballerini, E.S., Mandakova, T., Akoz, G., Derieg, N.J., Schmutz, J., et al. (2018). The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* 7. doi: 10.7554/eLife.36426.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1), D279-285. doi: 10.1093/nar/gkv1344.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* 60, 433-453. doi: 10.1146/annurev.arplant.043008.092122.

Freeling, M., Scanlon, M.J., and Fowler, J.E. (2015). Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Curr Opin Genet Dev* 35, 110-118. doi: 10.1016/j.gde.2015.11.002.

Gonzalez, D.H. (2016). Introduction to Transcription Factor Structure and Function. 3-11. doi: 10.1016/b978-0-12-800854-6.00001-4.

Gossani, C., Bellieny-Rabelo, D., and Venancio, T.M. (2014). Evolutionary analysis of multidrug resistance genes in fungi - impact of gene duplication and family conservation. *FEBS J* 281(22), 4967-4977. doi: 10.1111/febs.13046.

Graham, P.H., and Vance, C.P. (2003). Legumes: importance and constraints to greater use. *PLANT PHYSIOLOGY* 131, 872-877. doi: 10.1104/pp.017004.

Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., et al. (2007). Eukaryotic genome size databases. *Nucleic Acids Res* 35(Database), D332-D338. doi: 10.1093/nar/gkl828.

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B., et al. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* 361(6398). doi: 10.1126/science.aat1743.

Gupta, S., Nawaz, K., Parween, S., Roy, R., Sahu, K., Kumar Pole, A., et al. (2017). Draft genome sequence of Cicer reticulatum L., the wild progenitor of chickpea provides a resource for agronomic trait improvement. *DNA Res* 24(1), 1-10. doi: 10.1093/dnares/dsw042.

Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: A tool for mining segmental genome duplications and synteny. *Bioinformatics* 20, 3643-3646. doi: 10.1093/bioinformatics/bth397.

Han, M.V., Thomas, G.W., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* 30(8), 1987-1997. doi: 10.1093/molbev/mst100.

Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D.A., Yang, Z., et al. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytol* 209(2), 871-884. doi: 10.1111/nph.13626.

Hane, J.K., Ming, Y., Kamphuis, L.G., Nelson, M.N., Garg, G., Atkins, C.A., et al. (2017). A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant-microbe interactions and legume evolution. *Plant Biotechnol J* 15(3)**,** 318-330. doi: 10.1111/pbi.12615.

Hartman, G.L., West, E.D., and Herman, T.K. (2011). Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. *Food Security* 3**,** 5-17. doi: 10.1007/s12571-010-0108-x.

Hood, L., and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome Medicine* 5(9)**,** 79. doi: 10.1186/gm483.

Jackson, S.A., Iwata, A., Lee, S.-H., Schmutz, J., and Shoemaker, R. (2011). Sequencing crop genomes: approaches and applications. *New Phytologist* 191(4)**,** 915-925. doi: 10.1111/j.1469-8137.2011.03804.x.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449**,** 463-467. doi: 10.1038/nature06148.

Jain, M., Misra, G., Patel, R.K., Priya, P., Jhanwar, S., Khan, A.W., et al. (2013). A draft genome sequence of the pulse crop chickpea (Cicer arietinum L.). *Plant J* 74(5)**,** 715-729. doi: 10.1111/tpj.12173.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 45(D1)**,** D1040-D1045. doi: 10.1093/nar/gkw982.

Jordan, I.K., Wolf, Y.I., and Koonin, E.V. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol* 4**,** 22. doi: 10.1186/1471-2148-4-22.

Kang, Y.J., Kim, S.K., Kim, M.Y., Lestari, P., Kim, K.H., Ha, B.K., et al. (2014). Genome sequence of mungbean and insights into evolution within Vigna species. *Nat Commun* 5**,** 5443. doi: 10.1038/ncomms6443.

Kang, Y.J., Satyawan, D., Shim, S., Lee, T., Lee, J., Hwang, W.J., et al. (2015). Draft genome sequence of adzuki bean, Vigna angularis. *Sci Rep* 5**,** 8069. doi: 10.1038/srep08069.

Kellogg, E.A. (2016). Has the connection between polyploidy and diversification actually been tested? *Curr Opin Plant Biol* 30**,** 25-32. doi: 10.1016/j.pbi.2016.01.002.

Kim, J., Yang, J., Yang, R., Sicher, R.C., Chang, C., and Tucker, M.L. (2016). Transcriptome Analysis of Soybean Leaf Abscission Identifies Transcriptional Regulators of Organ Polarity and Cell Fate. *Frontiers in Plant Science* 7. doi: 10.3389/fpls.2016.00125.

Kim, M.Y., Lee, S., Van, K., Kim, T.H., Jeong, S.C., Choi, I.Y., et al. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc Natl Acad Sci U S A* 107(51)**,** 22032-22037. doi: 10.1073/pnas.1009526107.

Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform*, bbw008. doi: 10.1093/bib/bbw008.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* 34(7), 1812-1819. doi: 10.1093/molbev/msx116.

Kumpeangkeaw, A., Tan, D., Fu, L., Han, B., Sun, X., Hu, X., et al. (2019). Asymmetric birth and death of type I and type II MADS-box gene subfamilies in the rubber tree facilitating laticifer development. *PLoS One* 14(4), e0214335. doi: 10.1371/journal.pone.0214335.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* 42, 1053-1059. doi: 10.1038/ng.715.

Lee, C., Yu, D., Kim, H.-k.C.R.W., and Kim, R.W. (2017). Reconstruction of a composite comparative map composed of ten legume genomes. *Genes & Genomics* 39, 111-119. doi: 10.1007/s13258-016-0481-8.

Lehti-Shiu, M.D., Panchy, N., Wang, P., Uygun, S., and Shiu, S.H. (2017). Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. *Biochim Biophys Acta Gene Regul Mech* 1860(1), 3-20. doi: 10.1016/j.bbagrm.2016.08.005.

Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* 12(7), 1048-1059. doi: 10.1101/gr.174302.

Levin, D.A. (1983). Polyploidy and Novelty in Flowering Plants. *The American Naturalist* 122, 1-25. doi: 10.1086/284115.

Lewis, G.P. (2005). *Legumes of the World.* Royal Botanic Gardens, Kew.

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., and De Smet, R. (2016). Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. *Plant Cell* 28(2), 326-344. doi: 10.1105/tpc.15.00877.

Libault, M., Farmer, A., Brechenmacher, L., Drnevich, J., Langley, R.J., Bilgin, D.D., et al. (2010a). Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to Bradyrhizobium japonicum infection. *Plant Physiol* 152(2), 541-552. doi: 10.1104/pp.109.148379.

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R.J., Franklin, L.D., et al. (2010b). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J* 63(1), 86-99. doi: 10.1111/j.1365-313X.2010.04222.x.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., et al. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* 24(11), 4333-4345. doi: 10.1105/tpc.112.102855.

Lu, Q., Li, H., Hong, Y., Zhang, G., Wen, S., Li, X., et al. (2018). Genome Sequencing and Analysis of the Peanut B-Genome Progenitor (Arachis ipaensis). *Frontiers in Plant Science* 9. doi: 10.3389/fpls.2018.00604.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290(5494)**,** 1151-1155.

Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3(1-4)**,** 35-44.

Maldonado Dos Santos, J.V., Valliyodan, B., Joshi, T., Khan, S.M., Liu, Y., Wang, J., et al. (2016). Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. *BMC genomics* 17**,** 110. doi: 10.1186/s12864-016-2431-x.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452**,** 991-996. doi: 10.1038/nature06856.

Mochida, K., Sakurai, T., Seki, H., Yoshida, T., Takahagi, K., Sawai, S., et al. (2017). Draft genome assembly and annotation of Glycyrrhiza uralensis, a medicinal legume. *Plant J* 89(2)**,** 181-194. doi: 10.1111/tpj.13385.

Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K., and Tran, L.-S.P. (2010). LegumeTFDB: an integrative database of Glycine max, Lotus japonicus and Medicago truncatula transcription factors. *Bioinformatics (Oxford, England)* 26**,** 290-291. doi: 10.1093/bioinformatics/btp645.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., et al. (2015). Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biology* 16**,** 262. doi: 10.1186/s13059-015-0814-y.

Murat, F., Xu, J.H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., et al. (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research* 20**,** 1545-1557. doi: 10.1101/gr.109744.110.

Nagata, T., Hosaka-Sasaki, A., and Kikuchi, S. (2016). The Evolutionary Diversification of Genes that Encode Transcription Factor Proteins in Plants. 73-97. doi: 10.1016/b978-0-12-800854-6.00005-1.

Nakano, T. (2006). Genome-Wide Analysis of the ERF Gene Family in Arabidopsis and Rice. *PLANT PHYSIOLOGY* 140(2)**,** 411-432. doi: 10.1104/pp.105.073783.

Nam, J., Kim, J., Lee, S., An, G., Ma, H., and Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proceedings of the National Academy of Sciences* 101(7)**,** 1910-1915. doi: 10.1073/pnas.0308430100.

Nei, M., and Nozawa, M. (2011). Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. *Genome Biol Evol* 3**,** 812-829. doi: 10.1093/gbe/evr028.

O'Rourke, J.A., Iniguez, L.P., Fu, F., Bucciarelli, B., Miller, S.S., Jackson, S.A., et al. (2014). An RNA-Seq based gene expression atlas of the common bean. *BMC genomics* 15(1)**,** 866. doi: 10.1186/1471-2164-15-866.

Ohno, S. (1970). Evolution by gene duplication.

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. (2016). Evolution of Gene Duplication in Plants. *Plant Physiol* 171(4)**,** 2294-2316. doi: 10.1104/pp.16.00523.

Parween, S., Nawaz, K., Roy, R., Pole, A.K., Venkata Suresh, B., Misra, G., et al. (2015). An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). *Sci Rep* 5**,** 12806. doi: 10.1038/srep12806.

Paterson, a.H., Bowers, J.E., and Chapman, B.a. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A* 101**,** 9903-9908. doi: 10.1073/pnas.0307901101.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33(3)**,** 290-295. doi: 10.1038/nbt.3122.

Pires, N., and Dolan, L. (2010). Early evolution of bHLH proteins in plants. *Plant Signal Behav* 5(7)**,** 911-912. doi: 10.4161/psb.5.7.12100.

Portereiko, M.F., Lloyd, A., Steffen, J.G., Punwani, J.A., Otsuga, D., and Drews, G.N. (2006). AGL80 is required for central cell and endosperm development in Arabidopsis. *Plant Cell* 18(8)**,** 1862-1872. doi: 10.1105/tpc.106.040824.

Proulx, S.R., Wang, Y., Wang, X., Tang, H., Tan, X., Ficklin, S.P., et al. (2011). Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms. *PLoS One* 6(12)**,** e28150. doi: 10.1371/journal.pone.0028150.

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., and Zhang, S. (2018). Different Modes of Gene Duplication Show Divergent Evolutionary Patterns and Contribute Differently to the Expansion of Gene Families Involved in Important Fruit Traits in Pear (Pyrus bretschneideri). *Front Plant Sci* 9**,** 161. doi: 10.3389/fpls.2018.00161.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6)**,** 841-842. doi: 10.1093/bioinformatics/btq033.

Riaño-Pachón, D.M., Ruzicic, S., Dreyer, I., and Mueller-Roeber, B. (2007). PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* 8**,** 42. doi: 10.1186/1471-2105-8-42.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., et al. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science (New York, N.Y.)* 290**,** 2105-2110. doi: 10.1126/science.290.5499.2105.

Roulin, A., Auer, P.L., Libault, M., Schlueter, J., Farmer, A., May, G., et al. (2013). The fate of duplicated genes in a polyploid plant genome. *Plant J* 73(1)**,** 143-153. doi: 10.1111/tpj.12026.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., et al. (2008). Genome structure of the legume, Lotus japonicus. *DNA Res* 15(4)**,** 227-239. doi: 10.1093/dnares/dsn008.

Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4)**,** 592-593. doi: 10.1093/bioinformatics/btq706.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., et al. (2004). Mining EST databases to resolve evolutionary events in major crop species. *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada* 47**,** 868-876. doi: 10.1139/g04-047.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278)**,** 178-183. doi: 10.1038/nature08670.

Schmutz, J., McClean, P.E., Mamidi, S., Wu, G.A., Cannon, S.B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet* 46(7)**,** 707-713. doi: 10.1038/ng.3008.

Sedivy, E.J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytologist* 214(2)**,** 539-553. doi: 10.1111/nph.14418.

Sémon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. *Current Opinion in Genetics and Development* 17**,** 505-512. doi: 10.1016/j.gde.2007.09.007.

Severin, A.J., Cannon, S.B., Graham, M.M., Grant, D., and Shoemaker, R.C. (2011). Changes in twelve homoeologous genomic regions in soybean following three rounds of polyploidy. *Plant Cell* 23(9)**,** 3129-3136. doi: 10.1105/tpc.111.089573.

Severin, A.J., Woody, J.L., Bolon, Y.T., Joseph, B., Diers, B.W., Farmer, A.D., et al. (2010). RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol* 10**,** 160. doi: 10.1186/1471-2229-10-160.

Shiu, S.-H. (2005). Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals. *PLANT PHYSIOLOGY* 139**,** 18-26. doi: 10.1104/pp.105.065110.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 99**,** 13627-13632. doi: 10.1073/pnas.212522399.

Song, L., Prince, S., Valliyodan, B., Joshi, T., Maldonado dos Santos, J.V., Wang, J., et al. (2016). Genome-wide transcriptome analysis of soybean primary root under varying water-deficit conditions. *BMC genomics* 17**,** 57. doi: 10.1186/s12864-016-2378-y.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12(10)**,** 1611-1618. doi: 10.1101/gr.361602.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9)**,** 1312-1313. doi: 10.1093/bioinformatics/btu033.

Stebbins Jr, C.L. (1950). *Variation and evolution in plants.*: Oxford University Press (Geoffrey Cumberlege).

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34(Web Server issue)**,** W609-612. doi: 10.1093/nar/gkl315.

Tamura, K., Tao, Q., Kumar, S., and Russo, C. (2018). Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. *Mol Biol Evol* 35(7)**,** 1770-1782. doi: 10.1093/molbev/msy044.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. *Science* 320(5875)**,** 486-488. doi: 10.1126/science.1153917.

Tang, H., Bowers, J.E., Wang, X., and Paterson, A.H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* 107**,** 472-477. doi: 10.1073/pnas.0908007107.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014). An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC genomics* 15**,** 312. doi: 10.1186/1471-2164-15-312.

Thomas, B.C. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* 16**,** 934-946. doi: 10.1101/gr.4708406.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, N.Y.)* 313**,** 1596-1604. doi: 10.1126/science.1128691.

Valliyodan, B., Cannon, S.B., Bayer, P.E., Shu, S., Brown, A.V., Ren, L., et al. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant Journal* 100(5)**,** 1066-1082. doi: 10.1111/tpj.14500.

Van de Peer, Y. (2004). Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* 5(10)**,** 752-763. doi: 10.1038/nrg1449.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. *Trends in Plant Science* 14**,** 680-688. doi: 10.1016/j.tplants.2009.09.001.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Research* 24**,** 1334-1347. doi: 10.1101/gr.168997.113.

Varshney, R., Graner, A., and Sorrells, M. (2005). Genomics-assisted breeding for crop improvement. *Trends in Plant Science* 10(12)**,** 621-630. doi: 10.1016/j.tplants.2005.10.004.

Varshney, R.K., Chen, W., Li, Y., Bharti, A.K., Saxena, R.K., Schlueter, J.A., et al. (2011). Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat Biotechnol* 30(1)**,** 83-89. doi: 10.1038/nbt.2022.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., et al. (2013). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat Biotechnol* 31(3)**,** 240-246. doi: 10.1038/nbt.2491.

Vernie, T., Moreau, S., de Billy, F., Plet, J., Combier, J.P., Rogers, C., et al. (2008). EFD Is an ERF transcription factor involved in the control of nodule number and differentiation in Medicago truncatula. *Plant Cell* 20(10)**,** 2696-2713. doi: 10.1105/tpc.108.059857.

Vidal, N.M., Grazziotin, A.L., Iyer, L.M., Aravind, L., and Venancio, T.M. (2016). Transcription factors, chromatin proteins and the diversification of Hemiptera. *Insect Biochem Mol Biol* 69**,** 1-13. doi: 10.1016/j.ibmb.2015.07.001.

Wang, J., Sun, P., Li, Y., Liu, Y., Yu, J., Ma, X., et al. (2017). Hierarchically Aligning 10 Legume Genomes Establishes a Family-Level Genomics Platform. *Plant Physiol* 174(1)**,** 284-300. doi: 10.1104/pp.16.01981.

Wang, L., Cao, C., Ma, Q., Zeng, Q., Wang, H., Cheng, Z., et al. (2014). RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications. *BMC plant biology* 14**,** 169. doi: 10.1186/1471-2229-14-169.

Wang, Z., Libault, M., Joshi, T., Valliyodan, B., Nguyen, H.T., Xu, D., et al. (2010). SoyDB: A knowledge database of soybean transcription factors. *BMC plant biology* 10**,** 14. doi: 10.1186/1471-2229-10-14.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2**,** 333-341. doi: 10.1038/35072009.

Wright, E.S. (2015). DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* 16**,** 322. doi: 10.1186/s12859-015-0749-z.

Xie, M., Chung, C.Y.-L., Li, M.-W., Wong, F.-L., Wang, X., Liu, A., et al. (2019). A reference-grade wild soybean genome. *Nat Commun* 10(1). doi: 10.1038/s41467-019-09142-9.

Yang, K., Tian, Z., Chen, C., Luo, L., Zhao, B., Wang, Z., et al. (2015). Genome sequencing of adzuki bean (Vigna angularis) provides insight into high starch and low fat accumulation and domestication. *Proc Natl Acad Sci U S A* 112(43)**,** 13213-13218. doi: 10.1073/pnas.1420949112.

Young, N.D., and Bharti, A.K. (2012). Genome-enabled insights into legume biology. *Annu Rev Plant Biol* 63**,** 283-305. doi: 10.1146/annurev-arplant-042110-103754.

Young, N.D., Debelle, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., et al. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480(7378)**,** 520-524. doi: 10.1038/nature10625.

Zhang, H., Jin, J., Tang, L., Zhao, Y., Gu, X., Gao, G., et al. (2011). PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res* 39(suppl_1)**,** D1114-D1117. doi: 10.1093/nar/gkq1141.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* 33(4)**,** 408-414. doi: 10.1038/nbt.3096.

7. APPENDICES

Appendix A. Supplemental data files

The supplementary data file named *KcMoharana_Tables_S1-3.xls* at following website link:

https://figshare.com/s/d3cc998c72d2a23a1924

Appendix A.1: List of identified transcription factors (Table S1).

Appendix A.2: Number of genes duplicated by local (tandem or proximal) or segmental duplication across species and TF families (Table S2).

Appendix A.3: Genes from orthologous groups (OGs) expanded in *Glycine max* with corresponding syntenic orthologs in *Glycine soja* and *Phaseolus vulgaris* (Table S3).

Appendix B. List of publications

1. Passarelli-Araujo, H., Palmeiro, J., **Moharana, K.**, Pedrosa-Silva, F., Dalla-Costa, L. M., & Venancio, T. (2018). Molecular epidemiology of 16S rRNA methyltransferase in Brazil: RmtG in Klebsiella aerogenes ST93 (CC4). Annals of the Brazilian Academy of Sciences, 90(3 (Supl.1)).
2. Gazara, R. K., **Moharana, K. C.,** Bellieny-Rabelo, D., & Venancio, T. M. (2018). Expansion and diversification of the gibberellin receptor GIBBERELLIN INSENSITIVE DWARF1 (GID1) family in land plants. Plant Molecular Biology, 97(4–5), 435–449. https://doi.org/10.1007/s11103-018-0750-9
3. Danilevicz, M., **Moharana, K**., Venancio, T., Franco, L., Cardoso, S., Cardoso, M., … Ferreira, P. (2018). Copaifera langsdorffii Novel Putative Long Non-Coding RNAs: Interspecies Conservation Analysis in Adaptive Response to Different Biomes. Non-Coding RNA, 4(4), 27. https://doi.org/10.3390/ncrna4040027
4. **Moharana, K. C.**, & Venancio, T. M. (2019). Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae). BioRxiv, 849778. https://doi.org/10.1101/849778.
5. Machado*, F. B., **Moharana*, K. C.**, Almeida-Silva, F., Gazara, R. K., Pedrosa-Silva, F., Coelho, F. S., … Venancio, T. M. (2019). Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine* max) expression atlas. BioRxiv, 2019.12.23.886853. https://doi.org/10.1101/2019.12.23.886853; * equal contributors.
6. Oliveira, T.R., Aragão, VPM, **Moharana, K.C.,** Amaral F.P., Pluciani F.,, Fedosejevs E., Thelen, J.J, Venancio T.M., Silveira V., Santa-Catarina C., The light spectra affect the *in vitro* shoot development of *Cedrela fissilis* Vell. (Meliaceae) changing endogenous polyamine levels and induced a differential accumulation of proteins. Under review.

7. Franco* L. O.; **Moharana* K. C.**, Gazara R. K., Oliveira Eduardo A. G.; Cardoso M. A.; Cardoso S. R. S.; Hemerly A S.; Venancio T. M.; Ferreira, P. C. G., Whole transcriptome assembly and SNP identification in *Copaifera langsdorffii* Desf. from environmentally contrasting populations. * equal contributors. Under review.