# Transcription factors and molecular markers revealed asymmetric contributions between allotetraploid Upland cotton and its two diploid ancestors

Seyed Mehdi Jazayeri[1]* (iD) , Ronald Oswaldo Villamar-Torres[2,3] (iD) , Cristian Zambrano-Vega[4] (iD) ,
Luz Cecilia García Cruzatty[5] (iD) , Byron Oviedo-Bayas[4] (iD) , Milena do Amaral Santos[6] (iD) ,
Naga Raju Maddela[7] (iD) , Seyed Mohammad Hossein Ale Seyed Ghafoor[8] (iD) , Christopher Viot[9] (iD)

1. Universidad Nacional da Colombia - Facultad de Ciencias - Departamento de Biología - Bogotá, Colombia
2. Universidad Técnica Estatal de Quevedo - Facultad de Ciencias Pecuarias - Carrera de Ingeniería Agropecuaria – Quevedo (Los Ríos), Ecuador
3. Instituto Superior Tecnológico "Ciudad de Valencia" – Tecnología en Producción Agrícola y Tecnología en Procesamiento de Alimentos - Quevedo (Los Ríos), Ecuador
4. Universidad Técnica Estatal de Quevedo - Facultad de Ciencias de la Ingeniería - Carrera de Ingeniería en Sistemas - Quevedo (Los Ríos), Ecuador
5. Universidad Técnica de Manabí - Facultad de Ingeniería Agronómica - Instituto de Investigación - Portoviejo (Manabí), Ecuador.
6. Universidade Estadual de Santa Cruz - Programa de Pós-Graduação em Genética e Biologia Molecular - Ilhéus (BA), Brasil.
7. Universidad Técnica de Manabí - Facultad de Ciencias de la Salud - Instituto de Investigación - Portoviejo (Manabí), Ecuador.
8. Arak University - Faculty of Engineering (Daneshkadeh Fani Mohandesi) - Informatic Department, Arak (Markazi) Iran
9. Centre de Coopération - Internationale en Recherche Agronomique pour le Développement - Département Bios - Montpellier, France.

**ABSTRACT:** Three *Gossypium* species have been used to breed cotton as they vary in their fiber production and resistance to stresses. Transcription factors (TFs) mostly are present in different copies or isoforms by which they conduct their regulation. Their copy number can determine organism behavior to a cue. Simple sequence repeats (SSRs) are one of the most informative and versatile molecular markers. Transcription factors of three *Gossypium* species were compared *in silico*. Seventy eight percent of TFs were common between the three species. Single copy for each species were 6057 TF. *Gossypium hirsutum* and *G. raimondii* shared the most common interspecific TF. *Gossypium arboreum* species-specific TF were the least. MYB TF family with its subfamilies is the most abundant followed by bHLH and AP2/ERF family. *Gossypium hirsutum* generally possesses more TF copies compared to other two species. The 2109 single-copy clusters indicate that *G. hirsutum* has received one copy from only one parent. The five most abundant SSR markers of TF were dinucleotides AT, TA, TC, CT and TG belonging to *G. raimondii*. For *G. arboreum* and *G. hirsutum* they were trinucleotides CAA, CGA, TGA, GAA (CAT: *G. hirsutum*) and TCA. The findings suggest that there is regulatory difference between the three *Gossypium* species for fiber production and insect attack response. The differences may be due to some adaptive deletion events during speciation of *G. hirsutum* from its parents *G. arboreum* and *G. raimondii*.

**Key words:** fiber, insect attack, *Gossypium*, regulation.

# INTRODUCTION

Cotton plant belonging to *Gossypium* genus is one of the principal providers of cellulosic fiber and of great economic importance. In addition, it is one of the most important oleaginous crops worldwide because of cottonseed, which makes it an economically and agronomically important crop, mainly for the textile industry where it is usually utilized as raw material for the manufacture of garments. According to an estimated approximation concerning the world production of cotton, 80% of cotton fiber comes from Brazil, China, India, Pakistan, Turkey, USA and Uzbekistan. This crop represents a crucial percentage to the gross national product in many countries (Zahid et al. 2016). Cotton is a differentiated epidemic cell produced by *Gossypium* species (John and Crow 1992). However, cotton production is usually vulnerable to biotic and abiotic stresses. Therefore, breeding for stress-resistant cotton is a fundamental issue for plant biotechnology programs. This importance makes it mandatory to advance the application of biotechnological tools for cotton improvement (Rathore et al. 2015) in order to provide solutions to common problems that cotton producers are facing quotidianly.

The temperature ranges for cotton production are around 20 and 30 °C (Reddy et al. 1991). The effects of temperature on some developmental stages (germination, seedling growth, vegetative production, morphological development and maturity attributes) are very important for ensuring a correct boll production. For this, it is well known that cotton is enormously susceptible to temperatures throughout all developmental stages. However, flowering stage is when cotton plant shows more vulnerability, mainly to high temperature. Regarding biotic stress, cotton plant is mainly attacked by pests, particularly arthropods. These herbivores are housed mainly on foliage, flower buds and capsules (Peterson et al. 2016; Renou et al. 2012). However, cotton also suffers in lesser proportion from the attack of diseases and pests housed on soil, in which the nematodes are included (Greenberg et al. 2012). Cotton is considered a very high consumer of synthetic products, especially pesticides, which translates into high production costs. All these characteristics make it a crop with great difficulties, which endanger harvests, human health and the environment.

The genome of three species including *G. raimondii* (Wang et al. 2012), *G. hirsutum* (Li et al. 2015) and *G. arboretum* (Li et al. 2014) have been published. Genomic gene-rich regions responsible for fiber development have been disclosed (Xu et al. 2008) on four chromosomes involved in initiation (chr5), early, middle and late elongation (chr10 and 14), secondary cell wall deposition (SCWD), maturation (chr15) and finally verification (chr10 and 14). Transcription factor (TF) importance for cell wall process involved in cotton fiber production has been revealed recently (MacMillan et al. 2017).. However, due to their impacts on cotton production, the genes that are involved in fiber and cellulose production as well as biotic and abiotic stresses is under more consideration.

Transcription factor (TF) is one of the most important regulatory elements among plants. They can control the whole plant ontogeny and change expression of other key genes when needed like different stages of growth and development, response to stresses, production and reproduction. They can make networks collaborating with other TF or genes to control a mechanism, pathway, or production and help plant respond to stresses (Jazayeri et al. 2018; 2019). Transcription factor importance for cell wall process involved in cotton fiber production has been recently revealed (MacMillan et al. 2017).

Rice produces terpenes for defense against pathogens and herbivores. *OsWRKY45* TF was identified as a regulator in the production of diterpenoid phytoalexins like momilactone, phytocassane and oryzalexin by priming the expression of biosynthetic genes (Akagi et al. 2014). The expression of *HbWRKY1* has been associated with increased biosynthesis of natural rubber, a polyisoprenoid derived from wounding the bark of the tropical tree *Hevea brasiliensis* (Zhang et al. 2012. WRKY TFs family directly or indirectly regulate plant defense response altering the secondaries metabolites biosynthesis (Schluttenhofer and Yuan 2015). All these advantages have been taken into account by breeders in other crops due to the inherent role that TF play on these biomolecular processes. However, despite their importance, they have not well been studied in *Gossypium* and are still under development.

Molecular markers are DNA fragments that colligate with a particular region of the genome. Molecular markers in sequence motifs repeat in tandem include from 1 to 6 base pairs generally. They are grouped in different classes as microsatellites or short tandem repeats (STRs), simple sequence repeats (SSRs), simple sequence length polymorphism (SSLP), sequence tagged microsatellite site (STMS) and single nucleotide polymorphism (SNP). Among them, SSRs and SNPs are under

more consideration as they are useful informative markers for various studies on evolution, biodiversity, phylogeny and molecular breeding (MAS: marker-assisted selection) and have importantly been employed for constructing genetic linkage maps. They can be transferred between different species. They are reproducible, economic, polymorphic, heterozygous and allele specific and are known robust markers. Simple sequence repeats are evaluated by PCR due to flanking sequences by primer design. To study them, it is possible to disclose SSRs in a genome wide range and then see if the genes of interest possess any SSRs to use them in subsequent plant selection and breeding program (Jazayeri and Villamar-Torres 2017). Simple sequence repeats are important for plant breeding and selection.

This article aimed to comparatively study TFs in three species of *Gossypium* to reveal how similar and different they are in terms of regulatory system. Transcription factors related to fiber and stress are disclosed. In addition, the SSR markers are reported for TFs of three species. The results of this article provide a base for cotton plant breeding and selection in terms of regulatory system and TFs.

## MATERIAL AND METHODS

The protein sequences of TF families from PlantTFDB (Jin et al. 2017) and iTAK database (Zheng et al. 2016) for three species of *Gossypium* including *G. arboretum*, *G. hirsetum* and *G. raimondii*. PlantTFDB presents 320,370 TFs from 165 green plant species in its last version underlying TF motifs, regulatory elements and literature. iTAK identifies TF based on the consensus rules of TF protein domains of each family using other databases as evidence.

The TF sets from two databases were combined for each species to generate an integrated TF list. The TF number of each family were compared for three species. The combined TF set for each species were subject to cluster using CD-HIT (Xiong et al. 2012) by 90% similarity threshold to remove redundancy. CD-HIT can group the input sequences based on their similarity and overlap. Each cluster is a representative of the sequences according to the threshold similarity. The final clustered TF list for each species was used for subsequent analyses.

In order to extract gene nucleic acid sequences and genome position for the clustered TFs of three species CottonGen (Cotton Database Resources) (Yu et al. 2014) was employed.

Three TF lists were compared by OrthoVenn (Wang et al. 2015). It is a web-based server program and clusters protein sequences up to six lists to find their common and different proteins. It compares different protein sequence datasets, analyzes their annotation and generates Venn diagram and Gene Ontology (GO) enrichment for the clustered proteins. The default parameters were used.

Transcription factors of each species were classified using PlantTFcat (Dai et al. 2013). It groups the input sequences, compares them against its TF database and generates the information about the likely TF found in the sequences such as their family and domain.

GSEAServer was employed to annotate TFs of the three species. It is an online web-based and high-performance pipeline to perform fast functional annotation and gene/transcript set enrichment analysis for *de novo* assembled transcripts from non-model plant. It does blastx-search against nonredundant NCBI protein database based on available plant species and then annotates the blast hits using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes reference databases.

The protein clusters generated by OrthoVenn were visualized by TreeView of Environment for Tree Exploration (ETE) toolkit (Huerta-Cepas et al. 2016). It was also used to construct the gene tree for three Gossypium species TF sets. MEME was applied to find motifs for the selected TF clusters. UniProt was employed to reveal pertinent function of each protein cluster. iTOL (Letunic and Bork 2016) was utilized to construct the phylogenetic tree of three studied species for the chosen TF families. MEGA7.0.21 (Kumar et al., 2016) was employed to align, find and visualize the common sequence domain and repeats.

Simple sequence repeats were identified by the Perl script MISA (MIcroSAtellite identification tool. The parameters established for MISA were set to identify SSRs of di-, tri-, tetra-, penta- and hexanucleotides. These SSRs were selected because of their more polymorphic character than SSRs of 12 to 20 bp. Mononucleotide SSRs were not considered as they

are prone of errors due to assembly and sequencing. A distance of 100 bp between two SSRs was adjusted for compound SSRs (distinct and adjacent SSRs).

GOMapMan (Ramšak et al. 2014) was employed to extract the genes related to stress based on Arabidopsis map. The genes were subject to search in Bulk Data Retrieval in TAIR to generate fasta file containing gene sequences. Blast+ 2.6.0 (Camacho et al. 2009) was used to find homologs for TFs.

# RESULTS

## Transcription factor of *Gossypium* species

Transcription factors for each *Gossypium* species were compiled from two main plant TF databases and clustered to obtain TF species-specific lists (Table 1). *Gossypium hirsutum* possesses more TFs than the other two species comparing the final TF clustered integrated datasets.

**Table 1.** The TF number for each species and a list of genus-specific list. Transcription factors were clustered to obtain the nonredundant representatives.

| Species | PlantTFDB | iTAK* | Total before clustering | Total after clustering |
|---|---|---|---|---|
| *G. arboreum* | 2532 | 2619 | 5151 | 2603 |
| *G. hirsutum* | 5022 | 4424+5205 | 14651 | 3530 |
| *G. raimondii* | 4894 | 5013+2634 | 12541 | 3173 |
| Total | 12448 | 19895 | 32343 | 9306 |
| *Gossypium* | - | - | - | 4260 |

*iTAK Database has two TF lists that were merged as the iTAK TF list in the present study.

The integrated TF list of each species that was subject to group the TF families show that TFs, in terms of family number and member, were most in *G. hirsutum* followed by *G. raimondii* and *G. arboreum*, respectively. The numbers of nonredundant TFs were 2441, 3303 and 2967 for *G. arboretum, G. hirsutum* and *G. raimondii*, respectively. *Gossypium hirsutum* gains 61 TF families followed by *G. raimondii* with 60 and *G. arboreum* with 59 families (Table 2). The remaining proteins were categorized into other group since they are not TFs and belong to other type of regulatory elements.

Based on PlantTFcat classification, the most abundant family is MYB-HB-like followed by AP2-EREBP, bHLH, C2H2, NAM, Homobox-WOX, WRKY and bZIP. If MYB, MYB-HB and MYB/SANT were considered as a whole group, i.e., MYB family, then the family number increases for three species of *Gossypium* species as the biggest family. However, the position of the top families might move a little for the above-mentioned families.

The least abundant families were LFY, NOZZLE, STAT and WD40-like having at least one copy in each species. However, *G. arboreum* lacks two families including BED-type(Zn) and MYB-related and *G. raimondii* lacks MYB-related family based on the classification done by PlantTFcat. These results are not in accordance with PlantTFDB and iTAK databases where MYB-related members are very abundant for all three species (more than 90 members). This is due to the categorization system followed by PlantTFcat that classifies MYB-related TF under MYB family. BED-type(Zn) family has 4 and 9 copies in *G. raimondii* and *G. hirsutum*, respectively.

All total TFs of three species including 9306 proteins were clustered by 90% similarity. A list of 4260 clustered proteins was generated as the reference for the *Gossypium* genus[10] out of which 4112 proteins were annotated and classified by PlantTFcat. They were categorized as TF (4341), chromatin regulator (35), chromatin remodeler (2), chromatin remodeling (71), response regulator (44), transcription regulator (5).

---

[10]List available at https://doi.org/10.13140/RG.2.2.29830.73284

**Table 2**. The TF families of the three *Gossypium* species based on PlantTFcat classification.

| TF family | *G. arboreum* | *G. hirsutum* | *G. raimondii* | TF family | *G. arboreum* | *G. hirsutum* | *G. raimondii* |
|---|---|---|---|---|---|---|---|
| AP2-EREBP | 252 | 295 | 272 | HSF-type-DNA-binding | 41 | 46 | 41 |
| ARF | 30 | 48 | 49 | JmjC | 3 | 6 | 3 |
| AS2-LOB | 66 | 74 | 72 | JmjN | 3 | 5 | 3 |
| B3-Domain | 72 | 118 | 96 | JUMONJI | 3 | 5 | 3 |
| BED-type(Zn) | 0 | 9 | 4 | LFY | 1 | 1 | 1 |
| BES/BZR | 11 | 14 | 16 | MADS-MIKC | 38 | 69 | 59 |
| bHLH | 205 | 272 | 244 | MADS-type1 | 53 | 106 | 91 |
| bZIP | 104 | 138 | 136 | MYB | 30 | 37 | 33 |
| C2C2-CO-like | 18 | 23 | 27 | MYB/SANT | 29 | 35 | 35 |
| C2C2-Dof | 57 | 65 | 62 | MYB-HB-like | 348 | 474 | 417 |
| C2C2-GATA | 45 | 59 | 52 | MYB-related | 0 | 5 | 0 |
| C2C2-YABBY | 10 | 16 | 14 | NAM | 145 | 206 | 189 |
| C2H2 | 173 | 212 | 185 | Nin-like | 17 | 27 | 24 |
| C3H | 83 | 119 | 102 | NOZZLE | 1 | 1 | 1 |
| CG1-CAMTA | 7 | 10 | 15 | RAV | 9 | 12 | 10 |
| CW-Zn-B3/VAL | 4 | 4 | 3 | RR-A-type | 21 | 23 | 20 |
| DDT | 8 | 14 | 10 | RR-B-type | 11 | 12 | 11 |
| E2F-DP | 11 | 22 | 18 | S1Fa-like | 1 | 5 | 4 |
| EIL | 7 | 20 | 11 | SBP | 29 | 40 | 35 |
| FAR | 25 | 15 | 42 | ssDNA-binding-TF | 2 | 4 | 3 |
| GAGA-Binding-like | 7 | 7 | 8 | STAT | 1 | 2 | 1 |
| GARP-G2-like | 20 | 22 | 19 | STY-LRP1 | 13 | 16 | 13 |
| GeBP | 8 | 10 | 8 | TCP | 36 | 43 | 41 |
| GRAS | 82 | 121 | 93 | Tesmin | 11 | 25 | 21 |
| Hap2/NF-YA | 17 | 24 | 21 | TIFY | 27 | 35 | 35 |
| Hap3/NF-YB | 35 | 43 | 41 | TUBBY | 18 | 29 | 21 |
| HD-SAD | 18 | 26 | 25 | WD40-like | 2 | 2 | 2 |
| HD-ZIP | 30 | 39 | 33 | WRKY | 111 | 151 | 133 |
| Homeodomain-TALE-BEL | 20 | 31 | 24 | ZF-HD | 22 | 29 | 24 |
| Homeodomain-TALE-KNOX | 20 | 27 | 25 | Znf-LSD | 6 | 10 | 6 |
| Homobox-WOX | 128 | 173 | 152 | | | | |

## Common and species-specific transcription factors between the three *Gossypium* species

The common TFs between the three species included 7280 proteins out of 9306, which is 78% of all TFs. Between *G. hirsutum* and *G. raimondii* 7857 TFs (84%), between *G. arboreum* and *G. hirsutum* 7447 (80%) and finally between *G. arboreum* and *G. raimondii* 7388 TFs (79%) were found common from the most to the least. *Gossypium arboreum* did not have any specific TF cluster while *G. hirsutum* and *G. raimondii* possess 16 and 12 TF clusters covering 33 and 25 genes, respectively. In addition, 2109 single-copy gene clusters including 6057 TFs (93%) that comprise three copies (one copy for each species) were disclosed. The singleton TFs were 136, 595 and 385 as species-specific for *G. arboreum*, *G. hirsutum* and *G. raimondii*, respectively (Table 3). As shown in Fig. 1, the most abundant clusters belong to *G. hirsutum* followed

by *G. raimondii* and *G. arboreum*. The species-specific singletons belonged to different families and were not completely specific. However, they could not match the clusters due to the blast parameters (more than 90%), although they belong to the families those were clustered[11].

The most abundant TF families between the three species were MYB, bHLH, Homeobox, NAC and WRKY whose clusters possess more than 100 members. Homeobox-leucine zipper protein HOX1 was the first common cluster with 20 proteins between the three species followed by Homeobox-leucine zipper protein HDG9 and Calmodulin-binding transcription activator 2. The phylogenetic tree for the three species, as shown in Fig. 2, demonstrates that the TFs of *G. hirsutum* and *G. raimondii* are closer to each other than to *G. arboreum*.

**Table 3.** The input and cluster and singleton numbers for each species.

| Species | Proteins | Clusters | Singletons |
|---------|----------|----------|------------|
| *G. arboreum* | 2603 | 2400 | 136 |
| *G. hirsutum* | 3530 | 2639 | 595 |
| *G. raimondii* | 3173 | 2603 | 385 |
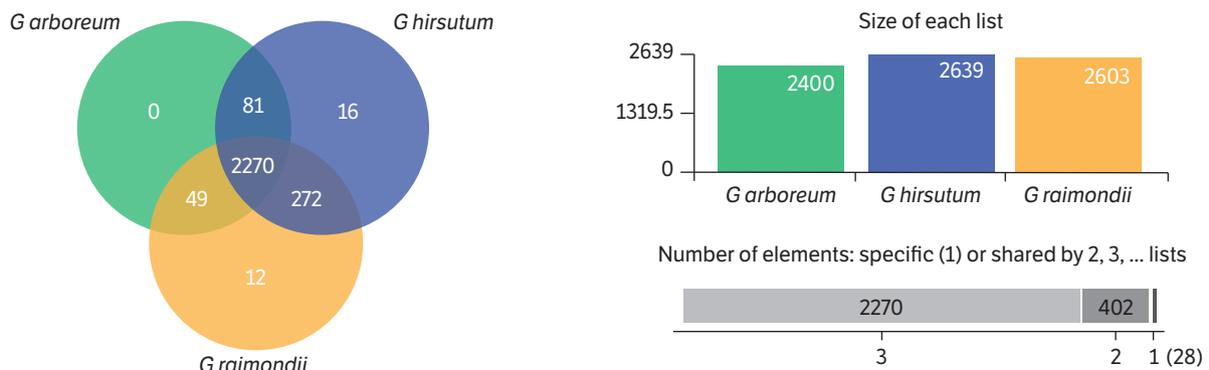


**Figure 1.** The Venn diagram of TFs in the three Gossypium species. The size of each list shows the number of clusters assigned to each species. The number of elements means the clusters between three species (3), two species (2) and specific for one species (1).
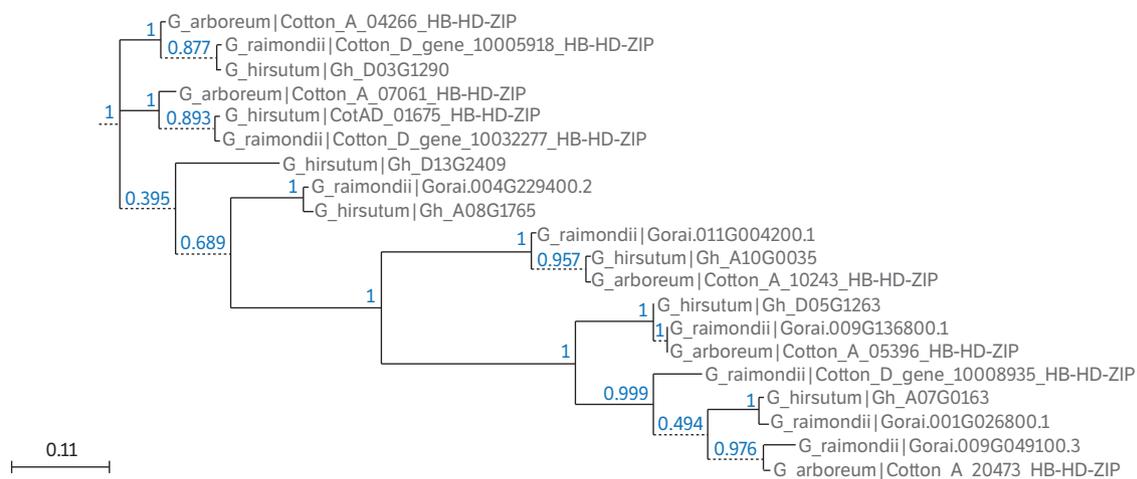


**Figure 2**. HOX1 cluster with the greatest number of common TFs between the three Gossypium species.

Considering other clusters with more than three copies, the authors of this work could generally find that for clusters containing for example 4, 5, 6 proteins, only one copy belongs to *G. arboreum* generally while *G. hirsutum* and *G. raimondii* share the other copies. In the clusters with more TF numbers, this trend is seen that *G. arboreum* possesses less number than the other two. As an example, for BES1/BZR TFs (cluster 237) four proteins were clustered with two copies from *G. arboreum* and only one copy from *G. hirsutum* and one copy from *G. raimondii*.

## Gene Ontology annotation analysis of transcription factors

In order to disclose in a better manner TF functionality of *Gossypium* species and analyze more effective *in silico*, the authors tried to mine information based on Gene Ontology (GO) terms from different sources. First, the clustered TF set of each species was subject to compare with *Arabidopsis* to annotate their function based on *Arabidopsis* annotation and GO terms. More TF genes in KOBAS matched with *Arabidopsis* homologs compared to GSEAServer as shown in Table 4. However, it is not significant and might be due to differences between both programs used.

**Table 4.** The genes revealed by GSEAServer and KOBAS compared with Arabidopsis thaliana to find annotation and GO terms.

| Species | Input genes | Cottongen.org match | GSEAServer* | KOBAS* |
|---|---|---|---|---|
| *G. arboreum* | 2603 | 2603 mRNA | 2561 | 2581 |
| *G. hirsutum* | 3530 | 6040 mRNA | 3426 | 3487 |
| *G. raimondii* | 3173 | 3172 mRNA | 3107 | 3152 |

*These genes were annotated against *Arabidopsis* in the program.

As reported, genes associated with GO terms of "membranes" are involved in fiber initial and development in cotton (Taliercio and Boykin 2007). Based on membrane-related GO terms on GSEAServer, it was found that *G. hirsutum* showed more membrane-related GOs (8411 hits belonging to 68 nonredundant terms) while *G. raimondii* possesses 7553 hits of 65 GO terms and *G. arboreum* 6194 hits of 53 terms. Interestingly, *G. hirsutum* has 3389 membrane-related genes among its TF dataset followed by *G. raimondii* with 3068 genes and *G. arboreum* with 2542 genes. However, the relationship of these TFs with fiber production needs further analyses to disclose.

Cell wall GO terms were analyzed between the three species and *G. hirsutum* showed 1684 cell wall associated GO for its TFs followed by *G. raimondii* and *G. arboreum* with 1522 and 1215 hits, respectively. The cell wall-associated genes were 259, 229 and 180 for *G. hirsutum, G. raimondii* and *G. arboreum*, respectively. Among the TFs related to secondary cell wall (SCW) that are responsible for fiber production in cotton, *G. hirsutum* showed many species-specific TFs. However, the members of WRKY, NAC and MYB for *G. raimondii* and *arboreum* as well as common genes between the three species were found.

Based on OrthoVenn analyses, TFs related to cell wall were mined by pairwise comparisons as they function in fiber production and biosynthesis (MacMillan et al., 2017). Interestingly, between *G. hirsutum* and *G. raimondii* 8 clusters including 17 genes belong to cell wall biogenesis, cell wall modification and other cell wall related processes while for *G. arboreum* vs *G. hirsutum* and *G. arboreum* vs *G. raimondii* there were 2 genes for each comparison. The TFs including NAC43, ROOT HAIR DEFECTIVE 3, MADS-box 21, SOMBRERO, Vegetative cell wall protein gp1, MADS-box 27 were common between *G. hirsutum* and *G. raimondii*.

Among the genes involved in fiber production, like SCW and PCW formation, many are common between *G. hirsutum* and *G. raimondii* and no copy is common between the three of them or between *G. arboreum* and the two other species. Some examples are WUSCHEL, MYB48, WRKY3, WRKY 22, WRKY 29, WRKY 75. These TFs were found as common TFs only between *G. hirsutum* and *G. raimondii*.

As previously revealed, membrane-associated GO include genes involved in fiber initiation and formation. There were 23 genes of membrane GO between *G. hirsutum* and *G. raimondii*. Interestingly, the most abundant common cluster, 60S

ribosomal protein L6-1, was associated with membrane in 7 genes of which 6 belonged to *G. hirsutum* and only one copy was found in *G. raimondii*. The other membrane associated genes were ROOT HAIR DEFECTIVE 3, Transcription repressor OFP7, Auxin response factor 5, Probable inactive receptor kinase At2g26730, NAC69, Thylakoid ADP, ATP carrier protein, L-arabinokinase and Tubby-like F-box protein 7. In the comparison between *G. arboreum* with *G. hirsutum* and *G. raimondii*, there were only two common genes, i.e., Tubby-like F-box protein 3 and Eukaryotic translation initiation factor 3 subunit A.

Due to the importance of climate change and stresses that can have impact on cotton cultivation and production, GO related to response to stress were mined. Stress-related GO terms were distributed on 931, 1234 and 1127 genes for *G. arboreum*, *G. hirsutum* and *G. raimondii*, respectively. Defense related genes were compared between the three species as 412, 532 and 499 *G. arboreum*, *G. hirsutum* and *G. raimondii*, respectively. Comparing three species, 178 clusters were determined as stress-based GO terms. Comparison of *G. hirsutum* and *G. raimondii* showed 19 clusters covering 39 genes that are involved in stress response. These genes belong to the TF families of MYB, WRKY, AP2, DRE, DELLA, protein phosphatase 2C 27, REVEILLE 5 and Homeobox-leucine zipper protein ATHB-40.

For the comparisons between *G. arboreum* and the two other species, 2 and 5 clusters were found under stress-related terms. GO related to response to insects were two members of MYB family including MYB76 and MYB51, that were common between *G. hirsutum* and *G. raimondii*, while no stress-GO term was found in other pairwise comparisons.

In addition, due to its importance and to have a better knowledge on TFs related to insect attack in cotton, genes with insect-related GO were extracted, which were 14, 16 and 15 for *G. arboreum*, *G. hirsutum* and *G. raimondii*, respectively (Table 5). The insect-related TFs of the three species shared MYC2, WRKY70, MYB51, dof zinc

**Table 5**. Genes related to insect GO terms. The name in parenthesis is the name of the gene for each species as appeared in NCBI RefSeq database.

| *G. arboreum* | *G. hirsutum* | *G. raimondii* |
|---|---|---|
| Cotton_A_21424, dof zinc finger protein DOF2.2-like (LOC108468989) | Gh_D12G0862_NBI-AD1_v1.1, dof zinc finger protein DOF5.7-like (LOC107945309) | Gorai.004G129600.1-JGI_221_v2.1, MYC3-like (LOC107911642) |
| Cotton_A_25894, MYC2-like (LOC108479071) | CotAD_36462_BGI-AD1_v1.0, MYC2-like (LOC107933572) | Gorai.010G219200.2-JGI_221_v2.1, WRKY70 (LOC105774341) |
| Cotton_A_07908, MYC2-like (LOC108479862) | Gh_A03G0594_NBI-AD1_v1.1, MYB34 (LOC108481270) | Gorai.008G128800.1-JGI_221_v2.1, MYC2-like (LOC105763947) |
| Cotton_A_32570, MYC2-like (LOC108479648) | CotAD_18444_BGI-AD1_v1.0, dof zinc finger protein DOF2.2-like (LOC107906545) | Gorai.004G004800.1-JGI_221_v2.1, dof zinc finger protein DOF2.2-like (LOC105789974) |
| Cotton_A_04044, WRKY 70 (LOC108451964) | Gh_A02G0029_NBI-AD1_v1.1, WRKY 70 (LOC107935935) | Gorai.001G151700.1-JGI_221_v2.1, MYC2-like (LOC105771003) |
| Cotton_A_06025, WRKY 70 (LOC108459758) | Gh_D13G0071_NBI-AD1_v1.1, WRKY 54 (LOC107940464) | Gorai.004G184800.1-JGI_221_v2.1, MYC2-like (LOC105791930) |
| Cotton_A_26203, MYB34 (LOC108481270) | Gh_D08G1175_NBI-AD1_v1.1, MYC3-like (LOC107911642) | Gorai.008G226300.1-JGI_221_v2.1, MYC2-like (LOC107946831) |
| Cotton_A_15733, WRKY 70 (LOC108463357) | Gh_D03G0882_NBI-AD1_v1.1, MYB34 (LOC105788423) | Gorai.001G109300.1-JGI_221_v2.1, dof zinc finger protein DOF2.2-like (LOC105791205) |
| Cotton_A_36552, MYC2-like (LOC108468801) | Gh_D04G0112_NBI-AD1_v1.1, WRKY transcription factor 70 (LOC107931553) | Gorai.009G292600.1-JGI_221_v2.1, WRKY 70 (LOC105770589) |
| Cotton_A_31334, DOF2.2-like (LOC108481883) | Gh_A12G1040_NBI-AD1_v1.1, MYC2-like (LOC108479071) | Gorai.012G013900.1-JGI_221_v2.1, WRKY 70 (LOC105780920) |
| Cotton_A_07316, MYC4-like (LOC108454458) | Gh_A12G1893_NBI-AD1_v1.1, MYC2-like (LOC108479862) | Cotton_D_gene_10032206-BGI-CGP_v1.0, dof zinc finger protein DOF2.2-like (LOC105770986) |
| Cotton_A_15595, WRKY 70 (LOC108485095) | Gh_D06G1939_NBI-AD1_v1.1, WRKY46 | Gorai.013G008300.1-JGI_221_v2.1, WRKY 54 (LOC105782571) |

**Table 5**. Continuation...

| *G. arboreum* | *G. hirsutum* | *G. raimondii* |
|---|---|---|
| Cotton_A_04316, WRKY 70 (LOC108475703) | Gh_D07G0954_NBI-AD1_v1.1, dof zinc finger protein DOF2.2-like (LOC107954067) | Gorai.006G216700.1-JGI_221_v2.1, MYC4-like (LOC105800310) |
| Cotton_A_06981, DOF2.2-like (LOC108452059) | Gh_A13G0059_NBI-AD1_v1.1, WRKY 70 (LOC107959346) | Gorai.003G097900.1-JGI_221_v2.1, MYB34 (LOC105788423) |
| | Gh_D09G1895_NBI-AD1_v1.1, MYC4-like (LOC107890614) | Gorai.010G219200.1-JGI_221_v2.1, WRKY 70 (LOC105774341) |
| | Gh_D07G1340_NBI-AD1_v1.1, MYC3-like (LOC107956701) | |

finger DOF2.2, bHLH3, bHLH28 genes. *Gossypium arboreum* possesses one C2C2_Dof (Cotton_A_06981_C2C2-Dof), *G. hirsutum* one Znf_Dof (Gh_D12G0862) and *G. raimondii* one MYC2/bHLH and one WRKY70/WRKY46 (Gorai.004G184800.1, Gorai.010G219200.2) as species-specific insect response TFs.

To develop enriching study on stress, stress-related genes from GOMapMan were compared with each TF dataset. Six common TFs of different members of heat shock TF family related to stress for three species were found, including HSFA2 (AT2G26150), HSFA7B (AT3G63350), HSFA1E (AT3G02990), A6Bb (AT3G22830), HSFA1B (AT5G16820) and HSFB4 (AT1G46264). The species-specific TF which belongs to *G. arboreum* was HSFA4A (AT4G18880). For *G. raimondii*, HSFA6B (AT3G22830) was the species-specific stress related TF. For *G. hirsutum* no stress-related TF was found as species-specific TF.

## Simple sequence repeats identifications

Transcription factors dataset of each species was subject to identify simple sequence repeats (SSRs). Interestingly no 4 and 5-unit-size SSRs were found for *G. arboreum* and *G. hirsutum* while *G. raimondii* showed all different unit size (from 2 to 6), as shown in Fig. 3. The SSR markers were distributed on 346, 503 and 1287 TFs of *G. arboreum, G. hirsutum* and *G. raimondii*, respectively. *G. raimondii* possesses the most SSRs. The trinucleotide SSRs were most abundant for *G. arboretum* and *G. hirsutum* while dinucleotide SSRs were most abundant in *G. raimondii* followed by dinucleotide SSRs (Fig. 3). Density of SSRs distributed among TF datasets showed that *G. raimondii* has denser SSR per 1Mbp as 234/Mbp, while for *G. hirsutum and G. arboreum* they were almost close and 153 and 145 SSRs/Mbp, respectively. In terms of density of SSR per TF, *G. raimondii* showed 66.4 SSRs for each 100 TFs, while for *G. arboreum* and *G. hirsutum* it was around one fifth of *G. raimondii* (Fig. 4).
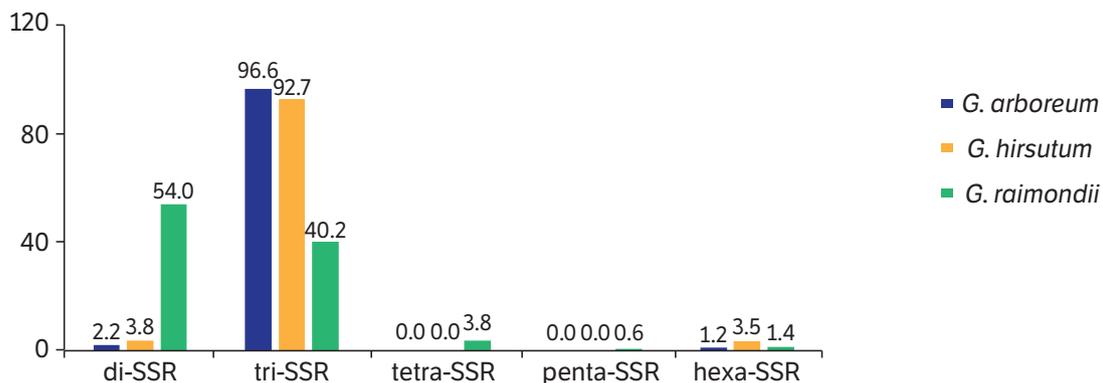


**Figure 3.** SSR categories (%) for three species of **Gossypium**. The categories are from di- to hexamer SSRs.
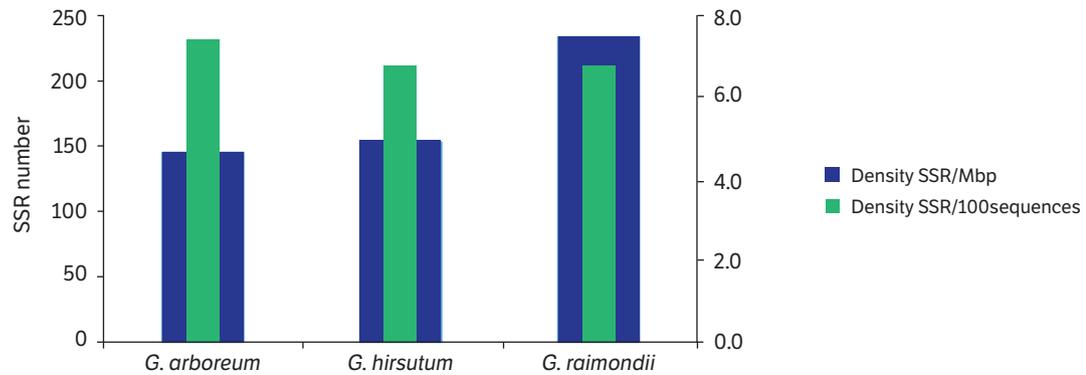
**Figure 4.** The SSR density of TFs of the three Gossypium species.

As shown in Table 6 for 20 the most abundant SSRs for the three *Gossypium* species, CAA, CAG, CAC were three SSR markers at top in *G. arboreum* and *G. hirsutum* while the first one, i.e., CAA, covered 9 and 8% of SSR markers, respectively. CAA covered 2.8% of SSRs in *G. raimondii* ranked in the ninth level. In *G. raimondii* AT, TA, TC, CT, TG were the most abundant markers followed by TTC, AG, TCA, CAA, CTT, ATC, CAG.

**Table 6.** Twenty most abundant SSRs of TFs in three *Gossypium* species.

| G. arboreum | | | G. hirsutum | | | G. raimondii | | |
|---|---|---|---|---|---|---|---|---|
| Repeats | Number | Percent | Repeats | Number | Percent | Repeats | Number | Percent |
| CAA | 37 | 9.02 | CAA | 48 | 7.92 | AT | 352 | 16.69 |
| CAG | 23 | 5.61 | CAG | 47 | 7.76 | TA | 324 | 15.36 |
| CAC | 22 | 5.37 | CAC | 27 | 4.46 | TC | 150 | 7.11 |
| GAA | 19 | 4.63 | CAT | 27 | 4.46 | CT | 101 | 4.79 |
| TCA | 19 | 4.63 | TCA | 27 | 4.46 | TG | 68 | 3.22 |
| CAT | 17 | 4.15 | GAA | 26 | 4.29 | TTC | 60 | 2.84 |
| AAC | 16 | 3.90 | ATC | 25 | 4.13 | AG | 57 | 2.70 |
| ACA | 16 | 3.90 | GAT | 21 | 3.47 | TCA | 48 | 2.28 |
| ATC | 16 | 3.90 | ACA | 20 | 3.30 | CAA | 43 | 2.04 |
| TCT | 16 | 3.90 | CTT | 20 | 3.30 | CTT | 40 | 1.90 |
| TGA | 16 | 3.90 | AAC | 19 | 3.14 | ATC | 39 | 1.85 |
| ACC | 14 | 3.41 | TGA | 19 | 3.14 | CAG | 39 | 1.85 |
| GAT | 14 | 3.41 | ACC | 18 | 2.97 | GA | 35 | 1.66 |
| CTT | 12 | 2.93 | GCA | 18 | 2.97 | TCT | 35 | 1.66 |
| TGG | 12 | 2.93 | TCT | 13 | 2.15 | CAT | 33 | 1.56 |
| AGA | 10 | 2.44 | ATG | 12 | 1.98 | GAA | 31 | 1.47 |
| GCA | 10 | 2.44 | AGA | 11 | 1.82 | GAT | 28 | 1.33 |
| ATG | 9 | 2.20 | GGT | 11 | 1.82 | TGA | 28 | 1.33 |
| GGT | 8 | 1.95 | TGG | 11 | 1.82 | TTA | 28 | 1.33 |
| TTC | 8 | 1.95 | CCA | 10 | 1.65 | AC | 23 | 1.09 |

# DISCUSSION

## Comparison of transcription factors of *Gossypium* species revealed phylogenic relation

*Gossypium hirsutum* possesses more TFs than *G. arboreum* and *G. raimondii* in terms of family number and members. This might be due to the fact that this species is tetraploid, while the other two are diploid. The most abundant family is MYB for all three species, which is in accordance with *Arabidopsis* genus and species (Jazayeri et al., 2018). However, due to varying categorization systems used by different TF sources and databases, some little changes in the TF ranking based on their number can be seen. For example, PlantTFDB and iTAK database classify AP2/ERF as two separate families, i.e., AP2 and ERF, then bHLH is the second most abundant family followed by ERF, NAM, bZIP, C2H2, WRKY. The family members might vary a little, but evidently MYB and bHLH are the most abundant TF families as seen in Arabidopsis (Jazayeri et al., 2018) and almost similar pattern exists between both genera and their species. Interestingly, many TFs members of MYB family as well as WRKY, bHLH and NAC are involved in regulation of cell wall in fiber production (MacMillan et al. 2017). The two most-numbered families, i.e., MYB and bHLH, have physically and functionally interaction to achieve their function (Pireyre and Burow 2015). This makes them important for any kind of interaction involved in biological processes and suggest their abundancy in plants.

Gene ontology terms related to membrane cover almost all TFs in the three species (more than 96%). This result may not be justified by GO terms to elucidate candidate TFs associated to fiber production considering their membrane compartmentalization. However, they might be involved in fiber production mechanisms. This remains to disclose how they might be involved in fiber production.

BED-type (Zn) is the only family that has no member in *G. arboreum.* In *G. hirsutum*, its number is twice as higher as *G. raimondii*. This finding suggests that this family has appeared during the evolution of *G. raimondii* and transferred and duplicated in *G. hirsutum*, bearing in mind that *G. hirsutum* is the progeny of *G. arboreum* and *G. raimondii*. BED finger proteins are known as either transcription activators or repressors. It regulates transcription by changing local chromatin structure on binding to GC-rich sequences. BED-type(Zn) is involved in plant disease resistance, for example blast resistance and bacterial blight in rice (Kroj et al. 2016). It is a domain type (CC)-NBS-LRR found in many R-genes among plants that functions in chromatin-boundary proteins and transposases (Aravind 2000).

Common TFs between the three species and pairwise comparisons showed that *G. raimondii* and *G. arboreum* share less TFs, but *G. hirsutum* has most common TFs with each of them. The most common TFs were found between *G. hirsutum* and *G. raimondii*, while the least common belonged to *G. arboreum* and *G. raimondii.* Additionally, they did not have the same contribution to *G. hirsutum*, especially in fiber- and stress-associated TFs. The findings in this study for TFs are in accordance with another that reported that A and D genomes do not equally contribute and D genome (*G. raimondii*) provides more genes than A genome (*G. arboreum*) (Xu et al. 2015). Therefore, they might share less genes with each other than with *G. hirsutum*. The results of this work suggested that *G. hirsutum* shares more TFs with *G. raimondii* and it might duplicate the *G. raimondii* originated TFs during its evolution. On the other hand, the TFs number of *G. arboreum* is less than in the other two species. These discrepancies and similarities between the three species might occur due to less TFs numbers of *G. arboreum* compared with the other species, causing less shared proteins. In addition, *G. hirsutum* may have received more TF genes from *G. raimondii* than *G. arboreum* or duplicated the genes of *G. raimondii* during its evolution. This is in accordance with a previous report (Shan et al. 2016).

A large number of single-copy genes suggests that *G. hirsutum* have not received these TFs from both parents, as only one copy exists for it. This imply that these genes might have been transmitted from *G. arboreum*, *G. raimondii* or other mechanisms that cause this single-copy occurrence like deletion of TFs transmitted from one parent. For other clusters with four and more copies, *G. hirsutum* possesses generally more copies.

The representative classified TF proteins of genus *Gossypium* showed other function than transcription factor. This indicates that there are some motif and sequence fragments that function differently. This can be one of the advantages that TFs possess by which they can regulate in different manners their target genes. However, to assure their varying functions by their motifs, further studies using functional genomics on each of them are required.

The findings of this study can justify that the principal species of *Gossypium* genus, i.e., *G. hirsutum*, employs more regulatory genes associated to fiber production and stress response than its parents.

The first cluster with the most common number of TFs belongs to Homeobox family. HOX1 is probably a transcription repressor involved in leaf development, it binds to the DNA sequence 5'-CAAT[GC]ATTG-3' and may act as a regulatory switch to specify provascular cell fate. Interestingly, it belongs to HOX family, whose members control fiber elongation in cotton and are expressed in cotton fiber cells (Shan et al. 2014).

## Species-specific genes disclosed regulatory system difference between the three *Gossypium* species

The TFs including NAC43, ROOT HAIR DEFECTIVE 3, MADS-box 21, SOMBRERO, Vegetative cell wall protein gp1, MADS-box 27 were common between *G. hirsutum* and *G. raimondii*. However, it remains to reveal if these TFs are key fiber production regulatory elements in cotton in order to use them in breeding and MAS programs.

For example, WUSCHEL related homeobox 4 is one of the master key TFs involved in vascular differentiation and procambium development (Ji et al., 2010). It functions between IAA and downstream molecules, i.e., jasmonate (JA), salicylic acid (SA) and ethylene to initiate subsequently fiber formation (Gorshkova et al. 2012).

MYB48 is differentially expressed in *G. raimondii* as one of PCW formation genes that is involved in fiber biosynthesis of cotton (MacMillan et al. 2017). WRKY TFs, WRKY6, WRKY75, WRKY3, WRKY29, WRKY22, during SCW and PCW formation comparing *G. hirsutum* with *G. raimondii* showed differential expression profiles as reported by MacMillan et al. (2017).

Basically, this research is an *in silico* study but the findings reported in it are in accordance with genes involved in fiber formation and production confirmed by other studies and used in cotton breeding programs successfully. Thus, they have been validated as candidate genes toward more fiber production in cotton.

## Insect related transcription factors are common between the three species

The common insect related TFs between the three species, including MYC2, WRKY70, MYB51, dof zinc finger DOF2.2, bHLH3 and bHLH28, have been reported that are involved in the processes related to defense against insect attacks. They are related to JA and SA balance. MYC2 regulates jasmonate-dependent functions (Dombrecht et al. 2007). WRKY70, as a crosstalk, acts as a repressor for JA and an activator for SA in favor of modulating signaling network as a defense mechanism (Li et al. 2006). On the other hand, WRKY70 is involved in negatively glucosinolate (GLS) biosynthesis which causes accumulation of GLS (Li et al. 2004). Different members of MYC family, MYC2, MYC3 and MYC4 are involved in GLS biosynthesis (Schweizer et al. 2013) and are classified in bHLH group with bHLH3 and bHLH28 (Niu et al. 2011). Interestingly, they have physical interaction with MYB family in order to regulate biosynthesis of GLS (Frerigmann et al. 2014; Schweizer et al. 2013). However, no species-specific TF was found in *Gossypium* species and the authors of this study were not able to conclude if any species-specific mechanism exists or not.

These TFs function in regulation of GLS biosynthesis while MYB 51 involved in indole GLS biosynthesis and MYB76 involved in aliphatic GLS biosynthesis. They increase insect resistance in crops (Malka and Cheng 2017) and can be considered as two key regulatory TFs in resistance to insect attacks (Gigolashvili et al. 2009) in cotton, as *G. hirsutum* and *G. raimondii* show more resistance to such stress. However, further expression analysis to disclose their function in response to insect attack in cotton species is necessary.

The discrepancies among species-specific TFs between the three *Gossypium* species suggest that the species-specific TFs belong to other isoforms or paralogs of TF families. Further functional genomic analyses are necessary to demonstrate how each species can respond to insect attack considering their specific TFs.

On the other hand, comparative analyses revealed some TFs in response to insect attack. HSFs are involved in herbivory and insect attack mechanism (Kuśnierczyk et al. 2007). *Gossypium raimondii* and *G. arboreum* possess species-specific HSF. These HSF can be used for plant screening with the aim of resistance to insect attack and herbivory.

## Simple sequence repeats are sufficiently different to be used as markers for *Gossypium* species

For *G. raimondii*, the SSR polymorphism trend is in accordance with the whole genome SSR analysis in plants, while for *G. arboretum* and *G. hirsutum* this ranking is different. This might be due to the input data as TFs that are part of the genome and more conservative than other functional genes. Multi-allelic SSRs are good tools to detect genetic structure and can demonstrate any likely relation between two genotypes, species, genera, ecotypes, etc.

*Gossypium hirsutum* is tetraploid and a descendant of *G. arboreum* and *G. raimondii*, therefore, it is assumed that it possesses a complex dataset of SSRs, but the findings of this study are not in accordance with expectation. This suggest that during its evolution *G. hirsutum* has experienced deletion events that has decrease its SSRs and polymorphism level. As previously declared, upland cotton has narrow genetic background toward low polymorphism rate (Khan et al. 2016) and this study confirm it.

The SSRs presented in this study are potential molecular markers in order to use in subsequent breeding program. As this study presented only SSRs within the TF genes, these markers can provide a powerful resource for searching markers associated to known genes and TFs. Due to regulatory function of TFs, these markers can be sued in association mapping research as well as finding gene networks. On the other hand, they can help to understand the likely phylogenetic links among different species of *Gossypium* and be useful for diversity and evolution analyses. They can provide knowledge of how mutation and changes in *Gossypium* species have been occurring over time.

## Simple sequence repeats for the genes involved in stress and insect attack response

The SSRs of the genes of interests in fiber, stress and insect attack were mined. Interestingly, there are SSRs present for the genes of interest that make such SSRs worthy to work on. These SSRs could be used in subsequent studies as a dataset to choose the genes for plant selection and breeding in fiber and defensive response. Among them, CAA and CAG are more abundant for *G. hirsutum* and *G. arboreum*. For *G. raimondii*, AT and TA appear in more numbers. However, considering the genes of interest in fiber production and in stress and insect response, relevant SSRs may attract attention as they share more genes as markers in their sequences. These markers are presented based on *in silico* analysis as useful markers for *Gossypium* species.

# CONCLUSION

Three *Gossypium* species were analyzed based on their TFs. The similarities among TFs showed that *G. hirsutum* and *G. raimondii* share more TFs and, therefore, are closer to each other. *Gossypium hirsutum* is a descendant of the other two, but it seems that its TF set has experienced deletion events, since in terms of number and common TFs it showed different behavior from what expected. The TFs related to SCW, fiber, membrane and response to insects and diseases showed that the members of NAC, WRKY, MYB families are present in more copies in *G. hirsutum* than in the other two species, suggesting its ability to produce more fiber. However, some other TFs like BES1/BZR, that is involved in fiber production via amylase activity, are present in one/two more copies in *G. arboreum*.

Simple sequence repeats markers showed that *G. arboreum* and *G. hirsutum* have more dense number of TFs per bp, while for *G. raimondii* 3-5 times more SSRs were seen for each 100 TFs. Between the three species, *G. raimondii* show more SSRs per Mbp comparing to the other species. However, the average SSR density for TF can be a representative of the whole genome. Comparing with other plants, *Gossypium* show denser SSR distribution.

These findings suggest that *G. raimondii* could cause more variations in genes of interest and is a source of polymorphism markers that can be used in breeding programs. The SSRs of TFs associated to fiber, stress and insect attack should be further studied for plant selection purposes. The details on the functionality of such TFs families and their members remain to disclose.

## AUTHORS' CONTRIBUTION

Conceptualization, Viot C., Jazayeri S. M. and Villamar-Torres R. O.; Methodology, Jazayeri S. M., Villamar-Torres R. O., Ghafoor S. M. H. A. S. and Viot C.; Investigation, Jazayeri S. M., Ghafoor S. M. H. A. S., Zambrano-Vega C., Villamar-Torres R. O., Cruzatty L. C. G., Bayas B. O. and Maddela N. R.; Writing – Original Draft, Jazayeri S. M. and Villamar-Torres R. O.; Writing – Review and Editing, Jazayeri S. M., Villamar-Torres R. O., Santos M. A. and Viot C., Supervision, Jazayeri S. M. and Viot C.

## REFERENCES

Akagi, A., Fukushima, S., Okada, K., Jiang, C.-J., Yoshida, R., Nakayama, A. and Takatsuji, H. (2014). WRKY45-dependent priming of diterpenoid phytoalexin biosynthesis in rice and the role of cytokinin in triggering the reaction. Plant Molecular Biology, 86, 171-183. https://doi.org/10.1007/s11103-014-0221-x

Aravind, L. (2000). The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. Trends in Biochemical Sciences, 25, 421-423. https://doi.org/10.1016/s0968-0004(00)01620-0

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10, 1-9. https://doi.org/10.1186/1471-2105-10-421

Dai, X., Sinharoy, S., Udvardi, M. and Zhao, P. X. (2013). PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics, 14, 1-6. https://doi.org/10.1186/1471-2105-14-321

Dombrecht, B., Xue, G. P., Sprague, S. J., Kirkegaard, J. A., Ross, J. J., Reid, J. B., Fitt, G. P., Sewelam, N., Schenk, P. M., Manners, J. M. and Kazan, K. (2007). MYC2 differentially modulates diverse jasmonate-dependent functions in *Arabidopsis*. The Plant Cell, 19, 2225-2245. https://doi.org/10.1105/tpc.106.048017

Frerigmann, H., Berger, B. and Gigolashvili, T. (2014). bHLH05 is an interaction partner of MYB51 and a novel regulator of glucosinolate biosynthesis in *Arabidopsis*. Plant Physiology, 166, 349-369. https://doi.org/10.1104/pp.114.240887

Gigolashvili, T., Berger, B. and Flügge, U.-I. (2009). Specific and coordinated control of indolic and aliphatic glucosinolate biosynthesis by R2R3-MYB transcription factors in *Arabidopsis thaliana*. Phytochemistry Reviews, 8, 3-13. https://doi.org/10.1007/s11101-008-9112-6

Gorshkova, T., Brutch, N., Chabbert, B., Deyholos, M., Hayashi, T., Lev-Yadun, S., Mellerowicz, E. J., Morvan, C., Neutelings, G. and Pilate, G. (2012). Plant fiber formation: state of the art, recent and expected progress and open questions. Critical Reviews in Plant Sciences, 31, 201-228. https://doi.org/10.1080/07352689.2011.616096

Huerta-Cepas, J., Serra, F. and Bork, P. (2016). ETE 3: reconstruction, analysis and visualization of phylogenomic data. Molecular Biology and Evolution, 33, 1635-1638. https://doi.org/10.1093/molbev/msw046

Jazayeri, S. M., Cruzatty, L. C. G. and Villamar-Torres, R. (2019). Genomic comparison among three *Arabidopsis* species revealed heavy metal responsive genes. The Journal of Animal and Plant Science, 29, 539-548.

Jazayeri, S. M., Pooralinaghi, M., Villamar-Torres, R. O. and Cruzatty, L. C. G. (2018). An *in silico* comparative genomic report on transcription factors in three *Arabidopsis* species. Ciencia y Tecnología, 11, 1-9. https://doi.org/10.18779/cyt.v11i1.185

Jazayeri, S. M. and Villamar-Torres, R. (2017). Genomic and transcriptomic approaches toward plant selection. Journal of Science and Research: Revista Ciencia e Investigación, 2, 54-64. https://doi.org/10.26910/issn.2528-8083vol2iss8.2017pp54-64

Ji, J., Strable, J., Shimizu, R., Koenig, D., Sinha, N. and Scanlon, M. J. (2010). WOX4 promotes procambial development. Plant Physiology, 152, 1346-1356. https://doi.org/10.1104/pp.109.149641

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

John, M. E. and Crow, L. J. (1992). Gene expression in cotton (*Gossypium hirsutum* L.) fiber: cloning of the mRNAs. Proceedings of the National Academy of Sciences of the United States of America, 89, 5769-5773. https://doi.org/10.1073/pnas.89.13.5769

Khan, M. K. R., Chen, H., Zhou, Z., Ilyas, M. K., Wang, X., Cai, X., Wang, C., Liu, F. and and Wang, K. (2016). Genome wide SSR high density genetic map construction from an interspecific cross of *Gossypium hirsutum* × *Gossypium tomentosum*. Frontiers in Plant Science, 7, 1-15. https://doi.org/10.3389/fpls.2016.00436

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X. and Morel, J.-B. (2016). Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. New Phytologist, 210, 618-626. https://doi.org/10.1111/nph.13869

Kumar, S., Stecher, G. and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution, 33, 1870-1874. https://doi.org/10.1093/molbev/msw054

Kuśnierczyk, A., Winge, P., Midelfart, H., Armbruster, W. S., Rossiter, J. T. and Bones, A. M. (2007). Transcriptional responses of *Arabidopsis thaliana* ecotypes with different glucosinolate profiles after attack by polyphagous *Myzus persicae* and oligophagous *Brevicoryne brassicae*. Journal of Experimental Botany, 58, 2537-2552. https://doi.org/10.1093/jxb/erm043

Letunic, I. and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research, 44, W1, W242-W245. https://doi.org/10.1093/nar/gkw290

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., Ma, Z., Shang, H., Ma, X., Wu, J., Liang, X., Huang, G., Percy, R. G., Liu, K., Yang, W., Chen, W., Du, X., Shi, C., Yuan, Y., Ye, W., Liu, X., Zhang, X., Liu, W., Wei, H., Wei, S., Huang, G., Zhang, X., Zhu, S., Zhang, H., Sun, F., Wang, X., Liang, J., Wang, J., He, Q., Huang, L., Wang, J., Cui, J., Song, G., Wang, K., Xu, X., Yu, J. Z., Zhu Y and Yu, X. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nature Biotechnology, 33, 524-530. https://doi.org/10.1038/nbt.3208

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., Zou, C., Chen, W., Liang, X., Shang, H., Liu, W., Shi, C., Xiao, G., Gou, C., Ye, W., Xu, X., Zhang, X., Wei, H., Li, Z., Zhang, G., Wang, J., Liu, K., Kohel, R. J., Percy, R. G., Yu, J. Z., Zhu, Y.-X., Wang, J. and Yu, S. (2014). Genome sequence of the cultivated cotton *Gossypium arboreum*. Nature Genetics, 46, 567-572. https://doi.org/10.1038/ng.2987

Li, J., Brader, G., Kariola, T. and Tapio Palva, E. (2006). WRKY70 modulates the selection of signaling pathways in plant defense. The Plant Journal, 46, 477-491. https://doi.org/10.1111/j.1365-313X.2006.02712.x

Li, J., Brader, G. and Palva, E. T. (2004). The WRKY70 transcription factor: A node of convergence for Jasmonate-mediated and salicylate-mediated signals in plant defense. The Plant Cell Online, 16, 319-331. https://doi.org/10.1105/tpc.016980

Greenberg, S. M., Adamczyk, J. J. and Armstrong, J. S. (2012). Principles and practices of integrated pest management on cotton in the lower Rio Grande Valley of Texas. In M. L. Larramendy and S. Soloneski (Eds.) Integrated pest management and pest control - current and future tactics (p. 4-27). London: IntechOpen Limited. https://doi.org/10.5772/31467

MacMillan, C. P., Birke, H., Chuah, A., Brill, E., Tsuji, Y., Ralph, J., Dennis, E. S., Llewellyn, D. and Pettolino, F. A. (2017). Tissue and cell-specific transcriptomes in cotton reveal the subtleties of gene regulation underlying the diversity of plant secondary cell walls. BMC Genomics, 18, 1-18. https://doi.org/10.1186/s12864-017-3902-4

Malka, S. K. and Cheng, Y. (2017). Possible interactions between the biosynthetic pathways of indole glucosinolate and auxin. Frontiers in Plant Science, 8, 1-8. https://doi.org/10.3389/fpls.2017.02131

Niu, Y., Figueroa, P. and Browse, J. (2011). Characterization of JAZ-interacting bHLH transcription factors that regulate jasmonate responses in *Arabidopsis*. Journal of Experimental Botany, 62, 2143-2154. https://doi.org/10.1093/jxb/erq408

Peterson, J. A., Ode, P. J., Oliveira-Hofman, C. and Harwood, J. D. (2016). Integration of plant defense traits with biological control of arthropod pests: challenges and ppportunities. Frontiers in Plant Science, 7, 1-13. https://doi.org/10.3389/fpls.2016.01794

Pireyre, M. and Burow, M. (2015). Regulation of MYB and bHLH transcription factors: A glance at the protein level. Molecular Plant, 8, 378-388. https://doi.org/10.1016/J.MOLP.2014.11.022

Ramšak, Ž., Baebler, Š., Rotter, A., Korbar, M., Mozetič, I., Usadel, B. and Gruden, K. (2014). GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. Nucleic Acids Research, 42, D1, D1167-D1175. https://doi.org/10.1093/nar/gkt1056

Rathore, K. S., Campbell, L. M., Sherwood, S. and Nunes, E. (2015). Cotton (*Gossypium hirsutum* L.). In K. Wang (Ed.). Agrobacterium Protocols. Methods in molecular biology (p. 11-23). Clifton: Springer. https://doi.org/10.1007/978-1-4939-1658-0_2

Reddy, V. R., Baker, D. N. and Hodges, H. F. (1991). Temperature effects on cotton canopy growth, photosynthesis and respiration. Agronomy Journal, 83, 699-704. https://doi.org/10.2134/agronj1991.00021962008300040010x

Renou, A., Togola, M., Téréta, I. and Brévault, T. (2012). First steps towards "Green" cotton in Mali. Outlooks on Pest Management, 23, 173-176. https://doi.org/10.1564/23aug07

Schweizer, F., Bodenhausen, N., Lassueur, S., Masclaux, F. G. and Reymond, P. (2013). Differential contribution of transcription factors to *Arabidopsis thaliana* defense against *Spodoptera littoralis*. Frontiers in Plant Science, 4, 13, 1-10. https://doi.org/10.3389/fpls.2013.00013

Schweizer, F., Fernandez-Calvo, P., Zander, M., Diez-Diaz, M., Fonseca, S., Glauser, G. and Reymond, P. (2013). *Arabidopsis* basic helix-loop-lelix transcription factors MYC2, MYC3 and MYC4 regulate glucosinolate biosynthesis, insect performance and feeding behavior. The Plant Cell, 25, 3117-3132. https://doi.org/10.1105/tpc.113.115139

Shan, C.-M., Shangguan, X.-X., Zhao, B., Zhang, X.-F., Chao, L., Yang, C.-Q., Wang, L.-J., Zhu, H.-Y., Zeng, Y.-D., Guo, W.-Z., Zhou, B.-L., Hu, G.-J., Guan, X.-Y., Chen, Z. J., Wendel, J. F., Zhang, T.-Z. and Chen. X.-Y. (2014). Control of cotton fibre elongation by a homeodomain transcription factor GhHOX3. Nature Communications, 5, 1-8. https://doi.org/10.1038/ncomms6519

Shan, W., Jiang, Y., Han, J. and Wang, K. (2016). Comprehensive cytological characterization of the *Gossypium hirsutum* genome based on the development of a set of chromosome cytological markers. The Crop Journal, 4, 256-265. https://doi.org/10.1016/J.CJ.2016.04.001

Schluttenhofer, C and Yuan, L. (2015). Regulation of specialized metabolism by WRKY transcription factors. Plant Physiology, 167, 295-306. https://doi.org/10.1104/pp.114.251769

Taliercio, E. W. and Boykin, D. (2007). Analysis of gene expression in cotton fiber initials. BMC Plant Biology, *7*, 1-13. https://doi.org/10.1186/1471-2229-7-22

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., Liu, K., Wang, B., Song, C., Shi, N., Kohel, R. J., Percy, R. G., Yu, J. Z., Zhu, Y.-X., Wang, J. and and Yu, S. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. Nature Genetics, 44, 1098-1103. https://doi.org/10.1038/ng.2371

Wang, Y., Coleman-Derr, D., Chen, G. and Gu, Y. Q. (2015). OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research, 43, W1, W78-W84. https://doi.org/10.1093/nar/gkv487

Xiong, J., Zhang, L., Fu, G., Yang, Y., Zhu, C. and Tao, L. (2012). Drought-induced proline accumulation is uninvolved with increased nitric oxide, which alleviates drought stress by decreasing transpiration in rice. Journal of Plant Research, 1251, 155-164. https://doi.org/10.1007/s10265-011-0417-y

Xu, Z., Kohel, R. J., Song, G., Cho, J., Alabady, M., Yu, J., Koo, P., Chu, J., Yu, S., Wilkins, T. A., Zhu, Y. and Yu, J. Z. (2008). Gene-rich islands for fiber development in the cotton genome. Genomics, 92, 173-183. https://doi.org/10.1016/J.YGENO.2008.05.010

Xu, Z., Yu, J., Kohel, R. J., Percy, R. G., Beavis, W. D., Main, D. and Yu, J. Z. (2015). Distribution and evolution of cotton fiber development genes in the fibreless *Gossypium raimondii* genome. Genomics, 106, 61-69. https://doi.org/10.1016/J.YGENO.2015.03.002

Yu, J., Jung, S., Cheng, C.-H., Ficklin, S. P., Lee, T., Zheng, P., Jones, D., Percy, R. G. and Main, D. (2014). CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Research, 42, D1, D1229-D1236. https://doi.org/10.1093/nar/gkt1064

Zahid, K. R., Ali, F., Shah, F., Younas, M., Shah, T., Shahwar, D., Hassan, W., Ahmad, Z., Qi, C., Lu, Y., Iqbal, A. and Wu, W. (2016). Response and tolerance mechanism of cotton *Gossypium hirsutum* L. to elevated temperature stress: a review. Frontiers in Plant Science, 7, 1-10. https://doi.org/10.3389/fpls.2016.00937

Zhang, Q., Zhu, J., Ni, Y., Cai, Y. and Zhang, Z. (2012). Expression profiling of HbWRKY1, an ethephon-induced *WRKY* gene in latex from *Hevea brasiliensis* in responding to wounding and drought. Trees, 26, 587-595. https://doi.org/10.1007/s00468-011-0623-1

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., Banf, M., Dai, X., Martin, G. B., Giovannoni, J. J., Zhao, P. X., Rhee, S. Y. and Fei, Z. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators and protein kinases. Molecular Plant, 9, 1667-1670. https://doi.org/10.1016/j.molp.2016.09.014