 doi:10.20944/preprints202203.0049.v1

Evolutionary and functional characterization of Cytochrome P450 in cotton reveal the role of subgenome expression of GhCYP78A for cotton fiber initiation

Priti Prasad[1,2], Rishi Kumar Verma[1,2,$], Uzma Khatoon[1,3,$], Samir V Sawant[1,2,#] , Sumit K Bag[1,2,#]

[1] Molecular Biology and Biotechnology Division, CSIR-National Botanical Research Institute, Rana Pratap Marg, Lucknow – 226001, India.

[2] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad - 201002, India.

[3] Department of Botany, University of Lucknow - 226001, India.

[$] Contributed equally

[#] **Corresponding Author**

Dr. Sumit K Bag

Principal Scientist

CSIR-National Botanical Research Institute

Lucknow- 226001, India.

Email: sumit.bag@nbri.res.in

Ph: +91-522-2297914/930 (O)

Dr. Samir V. Sawant

Chief Scientist

CSIR-National Botanical Research Institute

Lucknow- 226001, India.

Email: samirsawant@nbri.res.in

Ph: +91-522-2297947 (O)

 doi:10.20944/preprints202203.0049.v1

Abstract

Cytochrome P450 (CYPs) is a functionally diversified third-largest gene family that exploded in the plant kingdom. Their role in different organ development has been illustrated by the intervention of phytohormone. Cotton is a model organism for cell differentiation and cell elongation. To decipher the participation of CYPs in different cotton fiber developmental stages, we identified and characterized 2460 CYPs in three diploids and two allotetraploid cottons. Furthermore, In-silico expression and cluster analysis of cotton CYPs was conducted to distinguish the fiber stage-specific clusters that have the determining role in different stages of fiber development. The subgenome expression of two conserved *Gossypium hirsutum* CYPs, namely, GhCYP78A197 and GhCYP78A198 contributed to fiber initiation at an early stage of fiber development, governed by the co-occurrence of TATA and MYB TFs binding sites. Coexpression network partners of these two GhCYP78A annotated as auxin, kinases, chromatin remodeler, epigenetic regulator and cyclin-related genes that possibly induce the endoreduplication and cell proliferation for fiber cell initiation to define the high yield and biomass.

Keywords

Cotton, Fiber initiation, Cell differentiation, CYP78A, Endoreduplication, Biomass

Introduction

The cytochrome P450s (CYPs) are the largest gene families in Arabidopsis after the F-box and receptor like kinases encoded gene family [1]. Naturally, CYPs exist in a broad range from simple prokaryotes to complex eukaryotes [2]. Diversification of CYPs leads to catalyze diverse processes like cellular components, hormone biosynthesis, fatty acids biosynthesis, secondary metabolites, UV protectants and in defense compounds. Thus they are the fastest evolving ancient heme-thiolate enzyme with high genetic diversity, diverse functionality, large substrate specificity and different catalytic variability in all domains of life [3]. Their copy number is exploded in the plant kingdom by gene duplication event during the evolution [4]. They are membrane-localized monooxygenase enzymes to catalyze extremely diverse processes by activation and heterolytic cleavage of oxygen into the water molecule. The reduced P450 enzyme has a light absorption spectrum at 450nm where carbon monoxide binds to the haeme moiety [5]. Lengths of the CYPs are in the range of 475-550 amino acids and their nomenclatures in plant are based on the protein sequence identity, phylogeny and gene

 doi:10.20944/preprints202203.0049.v1

organization. The 40% identity criteria have been set up for the categorization under the same family while at least 55% identity is required for subfamily assignment [5].

Structurally, CYP protein sequences consist of four high conserved signature motifs; haeme binding motif (CXG) in which C acts as a catalytic center, AGxD/ET motif in I-helix for oxygen binding and activation, EXXR and PERF motif that together forms the E-R-R triad to stabilize the core structure. Out of these four motifs, only cysteine residues in the haeme binding motif and ERR triad are conserved among the plant kingdom. Basically, plant CYPs have been classified into two major clades; A-type and non-A-type. A type clade is the largest P450 clan catalyzing the secondary metabolites in plants while non-A-type clade divergent into 10 different clan that shows their involvement in sterol, lipid oxygenation and hormone biosynthesis [6]. CYP746, a non-A-type single family clan is present only in green algae and moss whose functions are still unknown. Notably, some CYPs are evolved only in angiosperms such as CYP99 and CYP723 were specific to grasses, CYP719 to Ranunculales and Piperales order, CYP726 to Euphorbiaceae whereas CYP702, CYP705 and CYP708 are Brassicales specific families [1]. All these CYPs were expected to execute some family-specific biochemical or molecular functions.

In Arabidopsis, approximately 20% of 249 identified full length CYPs were functionally characterized and their extensive co-expression network analysis has been conducted [7,8]. This microarray based co-expression network analysis leads to the identification of neofunctionalization of AtCYP98A8 and AtCYP98A9 in the phenolic pathway for pollen development [9]. Several CYPs were also validated for different organ development in the plant for instance CYP715A1 shows tissue-specific expression in tapetum regulates flower maturation and synchronization of the petal [10]. Another tapetum expressed CYP703A3 has an important role in male fertility through the development of anther cuticle and sporopollenin production [11]. Overexpression of CYP78A9 in Arabidopsis and loss of function mutant analyses of CYP78B5 in rice illustrates its correlation in embryo development [12,13] whereas ZmCYP78A1 increases the biomass and seed yield by prolonged cell division duration [14]. Furthermore, loss of function analyses of CYP51G1 shows stunted growth and seedling lethality due to improper sterol composition while overexpression of CYP74A1 in the Arabidopsis mutant line can restore the male sterility by limiting the Jasmonate level [15,16]. The involvement of CYP92A6 was also reported to regulate the etiolated seedling growth in pea via cross-talk between Brassinosteroids and light [17]. For root and nodule development in soybean, CYP728H1 shows their involvement in isoflavone metabolism whereas

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

OsCYP87A6 regulates the auxin signaling in coleoptile growth [18,19]. All these studies, delineate the pivotal role of CYPs in different organ development for plant kingdom.

Cotton (Gossypium sp.) belong to Gossypium genus of the Malvaceae family that providing the natural fiber and seed oil. It is composed of 45 diploid (2n = 2x = 26) and seven tetraploid species (2n = 4x = 52) [20] having distinct morphological diversifications, plant architectures and different phenotypic variations in fiber and non-fiber tissues as well [21]. Cotton fiber is single cell elongated structures originates from the outer integument of ovule which undergone through the four distinguishing and overlapping developmental stages viz., initiation, elongation, secondary cell wall (SCW) synthesis and fiber maturation [22]. Initiation of cotton fiber governs the total biomass while elongation and secondary cell wall deposition correlates with the fiber length and strength [23]. Since fiber development is regulated by many genetic and epigenetic factors that shows their involvement in hormones, signal transduction and secondary metabolism related pathways [22].

For better understanding of fiber development, the genome sequences were refined over the year for diploid and allotetraploid cotton species [24–31]. The refined reference genome gives an opportunity to conduct the genome wide comprehensive study for fiber related genes or pathways. Considering CYPs as a functionally diversified gene family, their possible role for the commitment of fiber cells needs to be explored. Therefore, firstly we mined the third plant largest gene family in diploid [(*Gossypium raimondii,* Gr $(D)_5$, *Gossypium arboreum,* Ga $(A)_2$ and *Gossypium herbaceum,* Ghe $(A)_1$] and allotetrapaloid [*Gossypium hirsutum,* Gh $(AD)_1$ and *Gossypium barbadense,* Gb $(AD)_2$] cotton at whole genome scale and conducted the structural, functional and evolutionary analysis for 2460 identified CYPs. We further executed the cluster analysis of fiber expressed CYPs at different fiber developmental stages to identify the stage specific *Gossypium hirsutum* CYPs. Furthermore, we annotated and validated GhCYPs for each cluster at different fiber developmental stages wherein conserved GhCYP78A emerged as a putative candidate for the fiber cell commitment at -3 DPA fiber developmental stage. The At-subgenome expression bias of two GhCYP78A namely GhCYP78A197_A and GhCYP78A198_A have been observed under the regulation of coexpression of TATA-box and MYB Transcription Factor. Coexpression network analysis of these two GhCYPs discloses their role in fiber cell differentiation and endoreduplication along with the expression of auxin, brassinosteroids, SNF1 and HMGB like chromatin remodeler, histone deacetylase, cell wall and cell cycle related gene.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

Results

Genome-wide Identification, Nomenclature and Characterization of the selected Diploid and Allotetraploid Cotton CYPs

CYP450 domain search was conducted against the three diploid [*Gossypium raimondii,* Gr $(D)_5$, *Gossypium arboreum*, Ga $(A)_2$ and *Gossypium herbaceum,* Ghe $(A)_1$] and two allotetraploid cotton species [*Gossypium hirsutum*, Gh $(AD)_1$ and *Gossypium barbadense,* Gb $(AD)_2$] in order to identify the CYPs which was further scrutinized for the presence of haeme binding motif (CXG) and E-R-R (EXXR and PERF) motif. Eventually, a total of 2460 CYPs (382 GrCYPs (*Gossypium raimondii)*, 357 GheCYPs (*Gossypium herbaceum)*, 336 GaCYPs [(*Gossypium arboreum)*, 698 GhCYPs (*Gossypium hirsustum)* and 687 GbCYPs (*Gossypium barbadense*)] were identified in cotton species (**Fig. 1, Supplementary Dataset 1-5**). In accordance with the standard nomenclature procedure, 13 GrCYPs, 12 GaCYPs, 9 GherCYPs, 28 GhCYPs and 1 GbCYPs were probable the pseudogenes either because of presence of indels, gap or missing the N-term and C-Term. In allotetraploid cotton species (Gh, Gb), the At and Dt subgenome were depicted with "_A" and "_D" suffix, respectively. Generally, CYP protein length is in the range of 475-600 amino acids (aa). The identified cotton CYP proteins whose sequence length is greater than the 600 aa would be the fusion protein that has been split further and assigned their name based on the sequence identity. For example, Gh_A11G006600.1 has a sequence length of 13kb which was split into the two GhCYPs, namely GhCYP706Q4_A and GhCYP706Q10_A with sequence length of 339 and 446 bp, respectively. The transcripts of some CYPs were renamed with their isoform number, for instance the isoforms of GrCYP707A143 were named as GrCYP707A143.1 and GrCYP707A143.2.

The haeme motif and ERR triad sequences of each CYPs were mentioned with their corresponding position in all five selected cotton genomes **(Supplementary Dataset 1-5)** except GaCYP74B25 having lack of haeme binding motif. This is a peculiar feature of CYP74 that utilizes oxygenated substrates for catalysing the process and so it does not need to interact with the molecular oxygen [5]. Identified cotton CYPs were predicted for transmembrane helices wherein 24.68%, 60.99%, 9.54% and 0.4% of GrCYPs, 24.64%, 65.54%, 8.68% and 1.12% of GheCYPs, 19.34%, 71.42%, 8.33%, and 0.59% of GaCYPs, 23.63%, 67.62%, 8.16% and 0.57% of GhCYPs, 29.40%, 62.15%, 7.86% and 0.43% of GbCYPs dominated with zero, one, two and three transmembrane domains, respectively. Only 0.29% of GaCYPs and 0.14%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

of GbCYPs possessed four TM domains in the identified cytochromes **(Supplementary Dataset 6).** The chromosomal localization of CYPs on *G.raimondii, G.herabceum, G.arboreum, G.hirsutum* and *G.barbadense* genome demonstrate the compactness of CYPs at the end of the each chromosome. Distribution of cotton CYPs were comparatively more compressed in *Gossypium raimondii* genome might be due to its smaller genome size (**Fig. S1**).

Phylogenetic clustering of Cotton CYPs

To understand the molecular relationships between the amino acid sequences of cotton CYPs, hybrid likelihood method was deployed for phylogenetic clustering. With 1000 bootstrap value, cotton CYPs were also classified into 10 different clans, including six single families (CYP51, CYP74, CYP97, CYP710, CYP711 and CYP727) and three multiple family clans (CYP85, CYP72 and CYP86) in non A-type clade (**Fig. 1A-E**). A-type clade (CYP71 clan) having the largest family member CYP, including 20 CYP families in five selected cotton genome (*G.raimondii, G.herabceum, G.arboreum, G.hirsutum* and *G.barbadense)*. CYP71 clan itself makes an individual cluster in the phylogenetic tree analysis (**Fig. 1, Fig. S2B**). CYP51 is the oldest clan that acts as an outlier for GheCYP, GaCYP, GhCYPs and GbCYPs tree except GrCYPs. In *G.raimondii* phylogenetic tree, GrCYP51 clustered close to the GrCYP74, GrCYP10 and GrCYP85 clan (**Fig. 1A**).

In order to investigate the evolution and occurrence of CYP450 in *Gossypium* species, a comparative view of CYPs with other angiosperm species were drawn (**Fig. S2A**). For the comprehensive study, 20 plant lineages including *Musa accuminata (Ma), Oryza sativa (Os), Triticum aestivum (Ta), Zea mays (Zm), Vitis vinifera (Vv)*, *Theobroma cacao (Tc), Gossypium ramondii (Gr), Gossypium arboreum (Ga), Gossypium herbaceum (Ghe), Gossypium hirsutum (Gh), Gossypium barbadense (Gb)*, *Arabidopsis thaliana (At)*, *Carica papaya (Cp)*, *Citus clementia (Cc)*, *Cucumis sativus (Cs)*, *Glycine max (Gm)*, *Malus domestica (Md)*, *Ricinus communis (Rc)*, *Populus trichocarpa (Pt)* and *Solanum lycopersicum (Sl)* based CYPs were illustrated with A-type and non A-type clade **(Fig. S2B,C)**. The highest number of CYP genes was classified as CYP749 (18 GrCYPs, 30 GaCYPs, 34 GheCYPs, 63 GhCYPs, 38 GbCYPs and 11 TcCYPs) after the CYP71 clan (47 GrCYPs, 39 GaCYPs, 41 GheCYPs, 61 GhCYPs, 78 GbCYPs and 41 TcCYPs) for all Malvaceae family genomes. Surprisingly, CYP749 clan was not present in the *Arabidopsis* genome[5]. In case of cotton, no any gene was classified as CYP709, CYP720 and CYP729 while their single copy gene was annotated for its ancestral

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

genome (*Theobroma cacao*, TcCYP). In corresponding to the AtCYPs (*Arabidopsis thaliana)* CYP705, CYP702 and CYP708 were not present in the Malvaceae family genome while CYP92, CYP736, CYP749, CYP728, CYP729, CYP733 and CYP727 might be not evolved in Brasicaceae family genome (*Arabidopsis*).

Gene expression profiling of cotton CYPs for different fiber developmental stages

To decipher the biological functions of evolving CYPs in diploid and allotetraploid cotton species for different fiber development stages, publicly available RNAseq data were mined and quantified with log2 transformed FPKM (Fragments per kilobase per million reads). In *Gossypium hirsutum,* 567 GhCYPs have the log2 (FPKM + 1) > 0 the expression of A-type and non A-type CYPs were displayed individually for At and Dt subgenome at -3, -1, 0, 1 and 3 DPA of ovule and 5, 10, 20, 25 DPA of fiber tissue (**Fig. 2A-D**). In total, 567 out of the 698 GhCYPs (~81 % of total CYPs) showed their expression value with the log2FPKM > 0. Identified GhCYPs clan expressed at different fiber developmental stages signify their possible role in fiber commitment (-3 to 0 DPA), fiber initiation (1 to 3 DPA), elongation (5 to 10 DPA) and SCW synthesis (25 DPA). Although, some CYPs were relatively highly expressed throughout the fiber developmental stages, for instance GhCYP701A18_A, GhCYP81Q6_A **(Fig. 2A)**, GhCYP736A194_D, GhCYP736A195_D, GhCYP82C59_D **(Fig. 2B)**, GhCYP51G1_A, GhCYP749A70_A **(Fig. 2C)**, GhCYP735A40_D and GhCYP749A_D **(Fig. 2D)** showed their ubiquitous expression throughout the selected developmental stages.

Additionally, the CYP gene expression level of GrCYPs, GheCYPs, GaCYPs and GbCYPs were also investigated to comprehend the evolutionary processes at their expression level **(Fig. 3)**. To check the expression profiling of Gossypium raimondii specific CYPs (GrCYPs), we processed the 10 DPA and 20 DPA fiber developmental stages. Out of the 382 GrCYPs, ~50% cytochromes (192 GrCYPs) were expressed either of 10 DPA and 20 DPA with the log2 (FPKM + 1) value > 0. Moreover, GrCYP706, GrCYP71 and GrCYP92 were relatively highly expressed for 10 and 20 DPA fiber developmental stages. GrCYP82D220, GrCYP82L17 and GrCYP87B31 have a higher expression at 0 DPA while GrCYP734A56, GrCYP749A199, GrCYP78BQ6 and GrCYP97A50.4 showed the higher expression at 10 DPA **(Fig. 3A**). CYPs gene expression was also checked for *G. arboreum*, *G. herabceum* and *Gossypium barbadense*, for 0 DPA, 10 DPA and 20 DPA fiber developmental stages. With the expression value, log2 (FPKM +1) > 0, 320 out of 357 GheCYPs (~90%), 317 out of 336 GaCYPs (~94%) and 509

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

out of 687 GbCYPs (~74%) CYPs were expressed and depicted in **Fig. 3B, 3C and 3D** respectively.

Cluster analysis and functional validation of fiber clustered genes for *G. hirsutum* species

To intensify the biological role of identified CYPs in *Gossypium hisustum*, cluster analysis was performed based on the raw gene expression value. Eventually, all identified GhCYPs were normalized and clustered into six major clusters namely C1 to C6 (**Fig. 4A**) for different fiber development stages and eight clusters namely C1 to C8 for the different stress conditions (Control, Cold, Hot, Salt and PEG) **(Fig. S3)**. Out of the 698 identified GhCYPs, 49, 19, 11, 41, 36 and 18 CYP genes were clustered in C1, C2, C3, C4, C5 and C6 cluster, respectively. The C1 cluster shows their higher expression at -3DPA to 0DPA fiber developmental stages whose expression was continuously decline till the 25 DPA in fiber tissue. The expression level of C2 containing GhCYP genes have relatively basal level of expression from -3 to 3 DPA which show some steepness at 5DPA and 10DPA that abruptly decline at 20 and 25 DPA (**Fig. 4A**). Similarly, C3 clustered GhCYP genes have high expression peak at 5 DPA and 10DPA while C5 and C6 clustered have high expression peak at 20 DPA and 25 DPA respectively. 41 GhCYP genes of C4 cluster demonstrate the lower expression level at -3 to 0 DPA which gradually showed the increasing expression pattern from 0 DPA to 25 DPA fiber developmental stages.

Functional enrichment of each identified cluster uncovered the putative function for fiber development at different developmental stages **(Fig. 4B, Fig. S4)**. C1 cluster, whose expression is significantly high at -3 DPA to 3 DPA of ovule tissue, shows their involvement in different floral organ development (anther wall tapetum formation, anther, stamen and stigma development) Cell differentiation of tapetum was also over-represented in this specific cluster. C1 clustered GhCYP genes indulged in Brassinosteroid (BR) biosynthesis process while interestingly, it over-represented for the abscisic acid catabolic process. Furthermore, other hormone biosynthesis process, regulation of hormone levels and lipid metabolic and biosynthesis processes were significantly enriched in same cluster. Regulation of hormone levels was enriched for C3 and C4 clusters, as well. KEGG pathway enrichment analysis also suggested the involvement C1, C2, C3 and C4 in Brassinosteroid biosynthesis pathway while C5 and C6 clusters involved in cutin, suberine and wax biosynthesis related pathway **(Fig. S5)**. Lipid metabolic process were proportionally enriched for C3, C4, C5 and C6 cluster whereas all six clusters (C1-C6) containing GhCYPs gene were associated with the phenylpropanoid

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

biosynthesis process. Normalized expression value of C5 and C6 clusters significantly increased at SCW. Additionally, C6 clustered genes were remarkably enriched in lipid hydroxylation and fatty acid metabolic process **(Fig. 4B)**. Callose deposition in cell wall, cell wall thickening process and sporopollenin biosynthesis processes were also enriched for this cluster that sequentially leads to fiber maturation at 25 DPA onwards.

Real time expression analysis at different fiber developmental stages (-3, 0, 10, 21 and 30 DPA) and in control sample (leaf) validate the expression of fiber clustered GhCYPs (C1 to C6) through its different representative genes **(Fig. 5)**. GhCYP78A198_A, GhCYP74971_D, GhCYP83F48_A and GhCYP94B70_D of C1 cluster displayed relatively higher expression at -3 DPA which was continuously decreasing till the latter stage of fiber development (30 DPA) except GhCYP78A198_A **(Fig. 5A-D)**. Comparison to leaf internal control, GhCYP78A198_A of C1 and GhCYP87A54_D of C3 cluster showed relatively higher expression indicated its major role in fiber initiation as well as elongation and SCW stage respectively **(Fig. 5A, 5E)**. The representative of GhCYPs of C4 (GhCYP74A1_D), C5 (GhCYP706B28P_A) and C6 (GhCYP84A80_D) cluster showed an increased expression pattern specifically for the SCW synthesis and maturation stages **(Fig. 5F-5H)**. In leaf tissue, all validated GhCYPs were expressed at the basal level underlining the fiber specific role for each clustered CYP gene.

Evolutionary interpretation of CYP78A among the flowering plant

Evolving CYPs with diversified functionality make it is a good standard for conducting the evolutionary study in cotton for fiber development. Gene Ontology (GO) enrichment analysis of different clusters suggesting the involvement of C1 clustered GhCYPs in different floral organ development and one of its members, GhCYP78A showed the significantly higher normalized value at -3 to 0 DPA **(Fig. 4B)**. Similarly, in our previous study, CYP78A was identified as a central hub gene in commitment specific module [32]. Cumulatively these two studies illuminate the importance of GhCYP78A for the fiber commitment processes on or before the day of anthesis (-3 to 0 DPA). At genome wide scale, thirteen CYP78A were identified for diploid cotton [*Gossypium raimondii* $(D)_5$, Gossypium arboreum $(A)_2$ and *Gossypium herbaceum* $(A)_1$], while it copies get doubled (26 CYP78A) in allotetraploid cotton [*Gossypium hirsutum* $(AD)_1$ and *Gossypium barbadense* $(AD)_2$]. Correspondingly, to interpret the evolution of GhCYP78A, we performed the phylogenetic clustering of GhCYP78A with other angiosperm species using hybrid likelihood method with 1000 bootstrap value. We

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

retrieved the CYP78A protein sequences from 13 different plant lineages including Musaceae, Poaceae, Vitaceae, Malvaceae, Brassicaceae, Caricaeae, Rutaceae, Cucurbiatceae, Fabaceae, Rosaceae, Euphorbiaceae, Saliaceae and Solanaceae that comprised of total 205 CYPs of twenty different angiosperm species **(Fig. 6)**. Consistent with the previous report [1], CYP78A was conserved in all angiosperm plant, which was ultimately clustered into eight major groups (I to VIII) where cotton CYP78A was grouped into the Group I, II, IV, V and VIII. Group I and IV were further divided into two and three subgroups, respectively that were named as (a), (b) and (c). Based on previous study, KLUH, KLUH like (EDO) and PLA1 clade was grouped in Group VII, I and IV respectively [14]. Group III, VI and VII were specific for the monocot plant lineages. Furthermore, Group VI and VII clustered CYP78A were assigned only for Poaceae and Musaceae specific family, respectively, suggesting its specific role for monocot specific plant groups. GhCYP78A197 (At and Dt) and GhCYP78A198 (At and Dt) of Group I and GhCYP78A200 (At and Dt) and GhCYP78A202 (At and Dt) of Group VIII have equal branch length structure in the phylogenetic tree clustering representing that they were simultaneously duplicated from the closest ancestral gene named GhCYP78A199 and GhCYP78A192 respectively **(Fig. S6)**.

In silico expression analysis of GhCYP78A in different fiber developmental stages

CYP78A was conserved in all six selected cottons. To interrogate the putative role of cotton CYP78A, *In silico* validation were carried out in *Gossypium raimondii, Gossypium herbaceum, Gossypium arboreum, Gossypium barbadense* and *Gossypium hirsutum* at different fiber developmental stages **(Fig. 7)**. For the validation process, online submitted RNAseq data were mined and checked their transcriptional activity at fiber initiation (0DPA), elongation (10DPA) and SCW (20DPA) stages. Consequently, many of the identified GrCYP78A have zero expression value, while GrCYP78A193 and GrCYP78A196 display the higher expression at 10 DPA and 20 DPA, respectively (**Fig. 7A**). Contrastingly, almost all identified GheCYP78A and GaCYP78A were transcriptionally active either at 0 DPA, 10 DPA or 20 DPA fiber developmental stages in which GheCYP78A198 and GaCYP78A198 were relatively highly expressed at 0 DPA and 10 DPA, respectively (**Fig. 7B, 7C**). In *Gossypium barbadense* and *Gossypium hirsutum*, not all identified CYPs were validated through the *In-silico* expression profiling at 0 DPA, 10 DPA and 20 DPA fiber developmental stages. In spite of these, some GbCYP78A for instance GbCYP78A193_D, GbCYP78A194_A, GbCYP78A200_A and GbCYP78A200_D have an exceptionally higher expression at 0 DPA (**Fig. 7D**). This result indicated that expressed GbCYP78A were functionally selected to execute the biological

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

processes related to cotton fiber initiation. Likewise, some GhCYP78A for instance GhCYP78A190_A, GhCYP78A193_A, GhCYP78A197_A, GhCYP78A198_A, GhCYP78A199_D, GhCYP78A200_A and GhCYP78A202_D were comparatively highly expressed at 0 DPA in compare to 10 DPA and 20 DPA. The expression value of GhCYP78A197_A and GhCYP78A198_A were exceptionally high at 0 DPA suggested its determined contribution in cotton fiber initiation (**Fig. 7E, Fig. S7**). Real time expression validation of GhCYP78A197_A also supplements its functional role in before the day of anthesis (-3 DPA) (**Fig. 5A**). Contrastingly, the Dt subgenome counterpart of these two GhCYP78A (GhCYP78A197_A, GhCYP78A198_A) have nill expression value (zero FPKM) at cotton initiation stage. In comparison to At and Dt subgenome expression level of GhCYP78A reflected the expression biasness either for At or Dt subgenome which was consistent with the previous results [21].

Enrichment of CRE and identification of coexpression network partners of GhCYP78A197_A and GhCYP78A198_A

The pronounced differences between the expression level of GhCYP78A197_A and GhCYP78A198_A was seen through *In-silico* expression analysis. This expression differences might be governed by the presence of CREs of respective GhCYP78A genes. Promoter analysis of 1000 bp upstream regions to the Transcriptional Start Site (TSS) of 26 GhCYP78A genes highlighted the abundance of development related (CAAT box, MYB), hormone related (SARE, AuRE, MeJARE, ERE, GRE) and the light responsive related CREs (GATA motif and I-box) **(Fig. 8A)**. Core promoter element (TATA-box) was also present in the close vicinity of TSS for 60% of identified GhCYP78A genes (17 out of 26) (**Fig. S8**). Whereas the well-studied MYB binding element was also abundantly present in 23 GhCYPs except GhCYP78A190_D, GhCYP78A192_D, GhCYP78A197_D. In the highly expressed GhCYP78A (GhCYP78A197_A and GhCYP78A198_A), both TATA-box and MYB binding element were enriched that might be regulate the fiber initiation at the different transcriptional level.

Along with the expression of GhCYP78A197_A and GhCYP78A198_A, coexpression partner were identified. For GhCYP78A197_A, 87 and 7 genes were positively (+vely) and negatively (-vely) coexpressed with r value ≥ 0.95 and ≤ -0.95, respectively (**Fig. 8B**). Similarly, 234 and 16 genes were positively and negatively co-expressed along with the GhCYP78A198_A, respectively **(Fig. 8C)**. In the +vely co-expressed, many genes were annotated as different transcriptional factors, kinases and as a sucrose transporter (**Fig. 7B and 8C**). Many epigenetic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

regulators and auxin responsive genes were also annotated for +vely co-expressed partner of GhCYP78A197_A and GhCYP78A198_A expressed gene at 0 DPA. Accordingly, the +vely and –vely co-expressed genes of GhCYP78A197_A were categorized as the Group I and II while GhCYP78A198_A were categorized as a Group III and IV respectively.

Mapman pathway enrichment of the +vely (Group I and III) and –vely (Group II and IV) co-expressed genes of both the highly expressed GhCYP78A (GhCYP78A197_A and GhCYP78A198_A) reflected their involvement in the different biological processes (**Fig. 8D to 8O**). The Group III genes were participating in chromatin organisation by the histone modifications activity with the help of HD1 and HD2 histone deactylases (**Fig. 8D**). They were also involved in cell cycle regulation by the DNA replication, recombination and cyclin dependent cell differentiation machinery (**Fig. 8H**). Many transcriptional related genes were also co-regulated that was responsible for making the pre-initiation complex for the RNA polymerase II-dependent transcription with the co-expressed TFIID and TAF14 gene component. Many TFs belongs to the MYB, Homoebox, B3 domain, ARF, GRF and TCP transcription factor were +vely expressed with both the GhCYP78A (Group I and III) (**Fig. 8E**). Protein modification related gene for instance MAPK, AGC Kinase superfamily and CK, CAMK, SNF1 related kinase complex were annotated for group I and group III respectively (**Fig. 8F**). Salicylic acid, stringolactone related phytohormones, solute transporter, redox homoestasis maintaining gene were positively coexpressed with the GhCYP78A197_A **(Fig. 8G, 8I, 8K)**. The group I and II genes were also involved in cell wall related processes through expansins and photosynthesis related mechanisms **(Fig. 8L-8N)**. Vesicle trafficking through the golgi apparatus were also enriched with the group III and IV that displayed the negative correlation with these two GhCYP78A **(Fig. 8O)**.

Discussion

The Cytochrome P450s is the biological macromolecule responsible for the NADPH or O2 dependent hydroxylation processes in many organisms. It showed the diverse role in plants including fatty acid metabolism [33,34], xenobiotic metabolism [35], antioxidant biosynthesis [36], secondary metabolite biosynthesis [37,38], hormone regulation [18,19,39], organ development [12–14] and in defense regulation [34]. To illustrate the possible role of CYPs in cotton fiber development specifically for fiber initiation, we conducted the genome wide comprehensive study of CYPs in the *Gossypium* species. With the availability of diploid and allotetraploid cotton genome, we identified a total of 382, 357, 336, 698 and 687 CYPs in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

Gossypium raimondii; Gr $(D)_5$, *Gossypium herbaceum*; Ghe $(A)_1$, *Gossypium arboreum*; Ga $(A)_2$, *Gossypium hirsutum*; Gh $(AD)_1$ and *Gossypium barbadense*; Gb $(AD)_2$ respectively that abundantly present throughout the genome **(Fig. 1, Fig. S1)**. By using the high contiguity assembled genome, we identified increased numbers of CYPs in *Gossypium raimondii, Gossypium arboreum, Gossypium hirsutum* in compare to previous study [40]. In an increasing order, lower number of CYPs was identified in A2 genome followed by the A1 and D5 genome. This differences is might be due to the expansion of insertion and deletion events in A2 genome and transposons elements bursts in the A1 and A2 in compared to the D5 genome [41]. Maximum number of identified CYPs were seen in Gossypium hirsutum indicating their involvement in fiber related features and to cope up with different stresses **(Fig. 4, Fig. S9)** [31]. A total of 2460 CYPs were identified in *Gossypium* species that possibly proportionate to the ploidy level of organisms likewise the reported CYPs in hexaploid wheat [42], palaeploid soybean [19] and in mesopolyploid brassica [43] (**Fig. S2**).

Phylogenetic clustering of identified CYPs in all five cotton species reveals the presence of ten different clan in which CYP71 was emerged as a largest one **(Fig. 1)**, underlies their structural diversification. During the cotton speciation event, CYP71 expressed to varying degree in all five cottons, consistent with their diverse nature **(Fig. 2, 3)**. In a comparative view, CYP749 was abundantly present in Malvaceae family that was reported only for Rosids, Asetrids and Ranunculales while completely absent in Arabidopsis **(Fig. S2)** [1]. *In-silico* expression of CYP749 suggested their pivotal role at different developmental stages (**Fig. 2, Fig. 3)** which was further validated through the real time expression analysis at -3, 0, 10, 20 and 30 DPA **(Fig. 5B)**. Some other CYPs for instance CYP702, CYP705, CYP708, CYP709, CYP720 and CYP729 were completely lost in *Gossypium* species along with the other taxon, justifies the previous assumption for loss of CYP subfamilies during the course of evolution (**Fig. S2**)[1]. Whereas, one of the primitive CYP, CYP51 was conserved in all five cotton species to maintained the membrane integrity **(Fig. 2, 3)** [16]. Low percentage of GaCYPs expression (50 %) at different DPA could be correlated with the negligible fiber structures in contrast to expression percentage of other selected cotton.

Considering *Gossypium hirsutum* as a high yielding cotton genome, 74% of expressed GhCYPs were furthered clustered into six distinct classes (C1-C6) to recognize their determined role at different fiber developmental stages **(Fig. 4)**. Before the day of anthesis (-3 to 0 DPA), C1 clustered GhCYPs for instance GhCYP51, GhCYP71, GhCYP74, GhCYP79, GhCYP92 and GhCYP78A were highly expressed in ovule tissues **((Supplementary Dataset 7))**. During this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

stage, BR and lipid homoeostasis level were might be maintained by the GhCYP51 and GhCYP92 that worked upstream to the BR synthesis pathway [17,44]. It also aids in floral organ development and cell differentiation **(Fig. 4B)** with upregulation of GhCYP71 and GhCYP78A that were closely associated with the plant hormone metabolism [12,13]. Other member of C1 containing GhCYPs such as GhCYP85A and GhCY90B also regulates the BR biosynthesis pathway in Oryza and tomato whereas GhCYP79 and GhCYP83, GhCYP74 and GhCYP94, and GhCYP88 were involved in the Indole Acetic Acid (IAA), Jasmonic acid and Gibberellin acid biosynthesis process in Arabidopsis, respectively [45,46]. Taken together, all these phytohormones positively regulating the commitment of fiber cells under the effect of different GhCYPs **(Fig. 5A-5D)** [47]. Contrastingly, GhCYP707 modulates the Abscisic acid concentration that negatively regulates the cotton fiber initiation at -3 to 3 DPA fiber developmental stages **(Supplementary Dataset 7)** [48,49]. To reduce the toxicity caused by the excessive amounts of BR, the expressed GhCYP734 inactivates the BR via C-26 hyrodxylation which subsequently altered the expression of calcium, protodermal factor 1 (GbPDF1) and Very Long Chain Fatty Acid (VLCFA) biosynthesis to promote the fiber initiation and elongation respectively [50,51].

The expression of CYP78A members (GhCYP78A190_A, GhCYP78A197_A, GhCYP78A198_A, and GhCYP78A202_D) was also observed at C1 cluster (**Fig. 4A, Fig. 7E)** where they were involved in different floral organ development **(Fig. 4B)**. Other two GhCYP78A, namely GhCYP78A194_A and GhCYP78A197_D displayed high normalized value in C6 cluster representing their crucial role in SCW synthesis through the fatty acid elongation process [52]. Quantitative trait loci demonstrated the role of GhCYP78A194_A in seed development as well as in high seed oil content in cotton [53]. The dual functionality of GhCYP78A has been validated through one of its orthologs, GhCYP78A198_A, whose expression was relatively high at -3 DPA and 21 DPA (**Fig. 5A**) also reproducing our previous result [54]. In plants, different CYP78A viz., CYP78A5 (KLUH), CYP78A6 (EOD3) and CYP78A7 (PLA1) were functionally characterized for different organ development, cell fate determining and differentiation processes that possibly worked downstream to the GA [12–14,52,55–60]. In cotton, the GhKLUH-like (EOD3), GhKLUH and GhPLA1 clades were clustered under group I, group IV and group VIII respectively that postulated their functional conservation **(Fig. 6)** throughout the angiosperm **(Fig. 7A-E)**.

Genome wide expression analysis of GhCYP78A in *Gossypium hirsutum* highlights an interesting point where they prominently contributed by only one homoeolog either At or Dt

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

subgenome **(Fig. 7E, Fig. S7)**. The phenomenon of subgenome expression biasness is might be due to the combinatorial effects of TATA-box and MYB transcription binding sites on their upstream region (**Fig. 8A**), corroborating with previous report [61]. The co-expression network partners of two highly expressed GhEOD3 (GhCYP78A197_A and GhCYP198_A) belongs to different transcription factors (MYB, bZIP, GRF, GATA, TAF14, TFIID, ARF), sucrose synthatase and auxin responsive genes whose functions were well documented in cotton fiber initiation (**Fig. 8B, 8C, 8E**) [62–65]. Many epigenetic regulators like HD1 and HD2 type histone deacetylase were +vely coexpressed with the GhCYP78A198_A **(Fig. 8D)** suggested their involvement in the cotton fiber initiation from ovule epidermal cells [66]. Different phytohormones, kinases (Histidine, MAP, AGC, PIP5K), chromatin remodelers (HMGB, SNF1) and cyclin related genes was also co-expressed with GhEOD3 that links a connecting point for cell proliferation and differentiation (**Fig. 8D, 8F-8H**) [12,13]. For instance, SNF1 and HMGB induces the expression of cell cycle related genes (CYCs/CDKs) in leaf growth of Arabidopsis and somatic embryo in cotton, respectively [60,67]. Their expression (SNF1, HMGB and CDK3) were significantly expressed at -1 and 0 DPA (**Fig. S10**) that might induces DNA endoreduplication in cotton, possibly regulated by the GhCYP78A198_A [12,64]. CYP78A causing endoreduplication and cell proliferation associated with the fruit size and seed size and their weight which ultimately increases the biomass and yield in many crops [12–14,52,55,56,59].

In cotton, no such studies were reported to elucidate the role of different GhCYPs by focusing on fiber cell differentiation or commitment that defines the fiber cell number and thus higher yield. Taken together, several genes mediate the fiber commitment, dependently or independently by GhCYP78A (GhEOD3) **(Fig. 9)**. GhEOD3 was expressed with At-subgenome biased way where they were co-expressed with different other biological macromolecules at before the day of anthesis. Therefore, this study provided a cue for crucial involvement of GhCYPs in cotton fiber development precisely for fiber commitment that eventually decided the fiber number thus increases the high fiber yield.

Materials and Methods

Identification of Cytochrome P450 gene family in Cotton species

For identification of CYPs in cotton, all protein sequences and gene annotation files were downloaded from the cottongen database [68]. Firstly, we aligned the arabidopsis CYPs protein against the protein sequences of five cotton species viz., *Gossypium raimondii* $(D)_5$, *Gossypium*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

arboreum $(A)_2$ and *Gossypium herbaceum* $(A)_1$, *Gossypium hirsutum* $(AD)_1$ and *Gossypium barbadense* $(AD)_2$ [24,25,30,31] through the HMMER search profiling [69]. The CYP450 conserved protein domain (PF00067) was also searched in all cotton genomes with the expectation value < 0.05. Further we conducted the MEME search on the P450 domain containing proteins to look into the haeme binding motif and E-R-R motif sequences, since these two motifs essential for the CYPs functioning. For proper nomenclature based on sequence similarity, we submitted the identified and filtered CYPs sequence to Cytochrome P450 Nomenclature Committee (David Nelson; drnelson1@gmail.com). To predict the transmembrane (TM) helices in identified CYPs protein sequences of all selected cotton genome, we used TMHMM server (http://www.cbs.dtu.dk/services/TMHMM/) for each dataset separately. The proper positioning of identified CYPs in all selected cotton genome were visualized though the TBtools application [70].

Phylogenetic tree clustering for defining the different CYP clan

The multiple sequence alignment of CYP protein sequences of all selected cotton species were carried out through the clustalo program in default parameters [71]. To identify the molecular relationship between the amino acid pattern in the protein domain of CYP450, the hybrid likelihood method was employed using IQ-TREE v 1.7[72]. Phylogenetic analysis was conducted for five selected cotton species and their tree generation and phylogenetic analysis was performed separately for each of them. The model was selected for P450 protein sequences using the inbuild model finder utility of IQ-TREE. The 'JTT+F+R7' model was selected as a best fit on the basis of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) score for cotton CYPs in *Gosssypium raimondii* and *Gossypium barbadense* with ultrafast bootstrap value of 1000. The best fit model 'JTT+F+R8', 'JTT+F+R9' and 'JTT+F+R10' was selected for the CYP protein sequences in *Gossypium hirsustum*, *Gossypium hebacuem* and *Gossypium arboreum* respectively. Further, the constructed phylogenetic tree was uploaded on iTOL web server to delineate the different CYP clan [73].

RNAseq analysis of cotton CYP genes during the different fiber developmental stages

The stage specific expression pattern of Cotton CYPs was analyzed through the publicly available high-throughput RNA sequencing data. For *G.hirsutum*, we downloaded and processes the SRA data with the accession numbers starting from SRR1695181 to SRR1695185 for "-3 DPA", "-1DPA" , "0DPA", "1 DPA" ,"3 DPA" of ovule, respectively and "5 DPA", "10 DPA","20 DPA" and "25 DPA" of fiber tissues with the accession numbers

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

from SRR1695191-SRR1695194, respectively. The high quality filtered reads (Q > 30) were mapped on the hisat2 aligner [74] with default parameters for each individual dataset. The transcript assembly and quantification of uniquely mapped reads was carried out through the StringTie assembler [75] and calculated the FPKM value (fragments per kilobase per million mapped reads). We also scrutinized the expression pattern of identified CYPs in *G.raimondii*, *G.arboreum*, *G.herbaceum* and *G.barbadense* cotton species at 0 DPA, 10 DPA and 20 DPA fiber developmental stages with the accession number listed in **Table S1**. Along with the estimation of expression level of different fiber developmental stages in *G. hirsutum*, we also examined the expression level of GhCYPs in different stress condition for 3hr. Control, Cold, Hot, Salt and PEG with the SRA accession number SRR1768522, SRR1768505, SRR1768509, SRR1768513 and SRR1768517 respectively. respectively.

Cluster profiling of GhCYPs for fiber tissues development and in stress conditions

The expression level of *Gossypium hirsutum* identified CYPs (GhCYPs) for different fiber development tissues and in stress conditions were used to identify the cluster. The biological expectations of co-expressed GhCYPs genes were clustered by the use of clust application from the heterogenous datasets of different fiber developmental stages ("-3 DPA", "-1DPA" , "0DPA", "1 DPA" ,"3 DPA" of ovule, and "5 DPA", "10 DPA","20 DPA" and "25 DPA" of fiber tissues) and in stress conditions (3hr. Control, Cold, Hot, Salt and PEG), separately [76]. The optimized consensus clustering was performed by taking the RNAseq FPKM value of GhCYPs as an input with the normalization parameter of "101 3 4".

Evolutionary relationship of CYP78A in angiosperm species

To reciprocate the evolutionary relationship of CYP78A in vascular plants, we extracted the protein sequences from different angiosperm species. In our study we selected 20 species that represents the 13 different angiosperm family including monocots and dicots. We retrieve the CYP78A protein sequences from the publicly available and our identified CYPs that belongs to Musaceae (*Musa accuminata),* Poaceae (*Oryza sativa, Triticum aestivum, Zea mays*), Vitaceae (*Vitis vinifera*), Malvaceae (*Theobroma cacao, Gossypium ramondii, Gossypium arboreum, Gossypium herbaceum, Gossypium hirsutum, Gossypium barbadense*), Brassicaceae (*Arabidopsis thaliana*), Caricaeae (*Carica papaya*), Rutaceae (*Citus clementia*), Cucurbiatceae (*Cucumis sativus*), Fabaceae (*Glycine max*), Rosaceae (*Malus domestica*), Euphorbiaceae (*Ricinus communis*), Saliaceae (*Populus trichocarpa*) and Solanaceae (*Solanum lycopersicum*) family.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2022 doi:10.20944/preprints202203.0049.v1

Multiple sequence alignment was carried out through the extracted CYP78A sequences through the use of clustal omega application [77]. Further phylogenetic tree clustering was conducted on basis of previously described method wherein 'JTT+R7' was chosen as a best model through IQ-TREE software.

In-Silico validation of fiber clustered GhCYP78A

In-silico validation of identified GhCYP78A clusters in different fiber developmental stages has been conducted through the naturally available cotton mutant line (*fiberless, fl*). For this analysis, comparative profiling of high throughput RNAseq data of wild and mutant variety of *Gossypium hirsutum* was accessed. The SRA data with the accession number SRR6466454 and SRR6466461 were processed for Xu-142 (wild type) and Xu-142*fl* (fibreless) cotton cultivar at 0 DPA fiber developmental stage. Raw reads were individually processed as described earlier.

Cis-regulatory element analysis of classified GhCYP clusters

The cis-regulatory element (CRE) analyses of each identified GhCYP clusters were identified through the PlantCARE database [78]. Firstly, the sequences from 1kb upstream of Transcription Start Site (TSS) of each gene in individual cluster were extracted from Gossypium *hirsutum* genome in strand specific manner. These sequences were assigned as the promoter sequences which was further submitted to the PlantCare database for the identification of CRE. The frequency of occurrence of identified CRE were calculated for each cluster, individually and visualized through the Biosequence view plugin in TBtools.

Similar approaches were applied for the GhCYP78A assigned nucleotide sequences, individually and count the frequency of CRE in their promoter regions.

Co-expression network analysis of GhCYP78A for the fiber commitment

Co-expression network analysis of the GhCYP78A was carried out through the transcriptome data of different fiber developmental stages. The gene expression value in log2 (FPKM +1) were utilized for conducting the coexpression network analysis using the "Expression Correlation Networks" plugin of Cytoscape. Based on the gene expression value, the expression correlation network computes the Pearson correlation coefficients value for each gene. The pearson correlation coefficients value (r) $\geq$ 0.95 among the interacting members of

 doi:10.20944/preprints202203.0049.v1

coexpressed network was assigned as the "positively coexpressed genes" while the pearson correlation coefficients value (r) < -0.95 was assigned as the "negatively coexpressed genes".

Functional validation of Identified GhCYP at different fiber developmental stages

We applied different approaches for the functional validation of *Gossypium hirsustum* CYPs at different fiber developmental stages. The fiber specific six clustered CYPs were separately processed for the over-represented gene set enrichment analysis through the ShinyGo [79]. Arabidopsis based significant gene enrichment terms (p value < 0.005) were selected and represented through the gProfiler web server [80].

Mapman tool was used for functional categorization of the identified positively and negatively coexpressed gene sets of GhCYP78A197_A and GhCYP78A198_A [81]. The significant metabolic pathways were enriched through the different datapoints in a BIN wise manner. For the significant enrichment, adjusted p value (False Discovery Rate, FDR) was applied according to the Benjamini and Hochberg method.

RNA extraction and real time expression validation

Total RNA was extracted from the different stages of ovules (-3 and 0 DPA), fibres (10, 21 and 30 DPA) and leaf using SIGMA spectrum^{TM} total RNA kit following manufacturer's protocol. DNase treatment was performed utilizing Ambion TURBO^{TM} DNase kit and RNA integrity was checked by gel electrophoresis. First-strand complementary DNA (cDNA) was synthesized from different stages of 2µg DNase treated total RNA using SuperscriptIII (Invitrogen) according to manufacturer's protocol. The PCR was performed from 10X diluted cDNA on ABI 7500 Fast Real Time PCR Machine (Applied Biosystems, USA) using Fast SYBR^{TM} Green Master Mix (Applied Biosystems). Two independent biological and three technical replicates were taken for the experiment. Relative gene expression was determined using $2^{-\Delta\Delta Ct}$ method [82]. Ghir_D10G001850.1 (UBQ14) was used as an internal reference gene. The primer was designed using software primer express version 3.0.1 and listed in **Table S2.**

Author contribution

Conceptualization, Priti Prasad; Data curation, Priti Prasad; Formal analysis, Priti Prasad; Funding acquisition, Samir Sawant; Investigation, Priti Prasad; Methodology, Priti Prasad; Project administration, Samir Sawant; Resources, Rishi Verma; Software, Priti Prasad;

Supervision, Sumit Bag; Validation, Uzma Khatoon; Visualization, Priti Prasad and Samir Sawant; Writing – original draft, Priti Prasad; Writing – review & editing, Priti Prasad, Rishi Verma, Uzma Khatoon, Samir Sawant and Sumit Bag. All authors have read and agreed to the published version of the manuscript. The manuscript number CSIR-NBRI_MS/2022/01/09 has been assigned to this manuscript.

Competing Interests

The authors declare that they have no conflict of interest.

Funding

This research was granted by the Council of Scientific and Innovative Research.

Acknowledgments

PP specifically thanks David Nelson for assisting in nomenclature. Authors acknowledge all those cotton researchers who submitted the raw transcriptome data to NCBI.

References

1. Nelson, D.; Werck-Reichhart, D. A P450-centric view of plant evolution. *Plant J.* **2011**, *66*, 194–211, doi:10.1111/j.1365-313X.2011.04529.x.

2. Nelson, D.R.; Kamataki, T.; Waxman, D.J.; Guengerich, F.P.; Estabrook, R.W.; Feyereisen, R.; Gonzalez, F.J.; Coon, M.J.; Gunsalus, I.C.; Gotoh, O.; et al. The P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. *DNA Cell Biol.* **1993**, doi:10.1089/dna.1993.12.1.

3. Konstandi, M.; Johnson, E.O.; Lang, M.A. Consequences of psychophysiological stress on cytochrome P450-catalyzed drug metabolism. *Neurosci. Biobehav. Rev.* 2014.

4. Werck-Reichhart, D.; Feyereisen, R. Cytochromes P450: a success story. *Genome Biol.* 2000.

5. Guengerich, F.P. Cytochromes P450. *Metab. Drugs Other Xenobiotics* **2012**, 27–66, doi:10.1002/9783527630905.ch2.

6. Paquette, S.M.; Bak, S.; Feyereisen, R. Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of Arabidopsis thaliana. *DNA Cell Biol.* **2000**, doi:10.1089/10445490050021221.

7. Schuler, M.A.; Duan, H.; Bilgin, M.; Ali, S. Arabidopsis cytochrome P450s through the looking glass: A window on plant biochemistry. *Phytochem. Rev.* 2006.

8. Ehlting, J.; Sauveplane, V.; Olry, A.; Ginglinger, J.F.; Provart, N.J.; Werck-Reichhart, D. An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in Arabidopsis thaliana. *BMC Plant Biol.* **2008**, doi:10.1186/1471-2229-8-47.

9. Matsuno, M.; Compagnon, V.; Schoch, G.A.; Schmitt, M.; Debayle, D.; Bassard, J.E.; Pollet, B.; Hehn, A.; Heintz, D.; Ullmann, P.; et al. Evolution of a novel phenolic pathway for pollen development. *Science (80-.).* **2009**, doi:10.1126/science.1174095.

10. Liu, Z.; Boachon, B.; Lugan, R.; Tavares, R.; Erhardt, M.; Mutterer, J.; Demais, V.; Pateyron, S.; Brunaud, V.; Ohnishi, T.; et al. A Conserved Cytochrome P450 Evolved in Seed Plants Regulates Flower Maturation. *Mol. Plant* **2015**, doi:10.1016/j.molp.2015.09.002.

11. Yang, X.; Wu, D.; Shi, J.; He, Y.; Pinot, F.; Grausem, B.; Yin, C.; Zhu, L.; Chen, M.; Luo, Z.; et al. Rice CYP703A3, a cytochrome P450 hydroxylase, is essential for development of anther cuticle and pollen exine. *J. Integr. Plant Biol.* **2014**, doi:10.1111/jipb.12212.

12. Chen, Y.; Liu, L.; Shen, Y.; Liu, S.; Huang, J.; Long, Q.; Wu, W.; Yang, C.; Chen, H.; Guo, X.; et al. Loss of function of the cytochrome P450 gene CYP78B5 causes giant embryos in rice. *Plant Mol. Biol. Report.* **2015**, doi:10.1007/s11105-014-0731-3.

13. Ito, T.; Meyerowitz, E.M. Overexpression of a Gene Encoding a Cytochrome P450, CYP78A9, Induces Large and Seedless Fruit in Arabidopsis. *Plant Cell* **2000**, doi:10.2307/3871172.

14. Sun, X.; Cahill, J.; Van Hautegem, T.; Feys, K.; Whipple, C.; Novák, O.; Delbare, S.; Versteele, C.; Demuynck, K.; De Block, J.; et al. Altered expression of maize PLASTOCHRON1 enhances biomass and seed yield by extending cell division duration. *Nat. Commun.* **2017**, doi:10.1038/ncomms14752.

15. Park, J.H.; Halitschke, R.; Kim, H.B.; Baldwin, I.T.; Feldmann, K.A.; Feyereisen, R. A knock-out mutation in allene oxide synthase results in male sterility and defective wound signal transduction in Arabidopsis due to a block in jasmonic acid biosynthesis. *Plant J.* **2002**, doi:10.1046/j.1365-313X.2002.01328.x.

16. Kim, H.B.; Schaller, H.; Goh, C.H.; Kwon, M.; Choe, S.; An, C.S.; Durst, F.; Feldmann, K.A.; Feyereisen, R. Arabidopsis cyp51 mutant shows postembryonic seedling lethality associated with lack of membrane integrity. *Plant Physiol.* **2005**, doi:10.1104/pp.105.061598.

17. Kang, J.G.; Yun, J.; Kim, D.H.; Chung, K.S.; Fujioka, S.; Kim, J. Il; Dae, H.W.; Yoshida, S.; Takatsuto, S.; Song, P.S.; et al. Light and brassinosteroid signals are integrated via a dark-induced small G protein in etiolated seedling growth. *Cell* **2001**, doi:10.1016/S0092-8674(01)00370-1.

18. Chaban, C.; Waller, F.; Furuya, M.; Nick, P. Auxin Responsiveness of a Novel Cytochrome P450 in Rice Coleoptiles. *Plant Physiol.* **2003**, doi:10.1104/pp.103.022202.

19. Guttikonda, S.K.; Trupti, J.; Bisht, N.C.; Chen, H.; An, Y.Q.C.; Pandey, S.; Xu, D.; Yu, O. Whole genome co-expression analysis of soybean cytochrome P450 genes identifies nodulation-specific P450 monooxygenases. *BMC Plant Biol.* **2010**, *10*, 1–19, doi:10.1186/1471-2229-10-243.

20. Yuan, D.; Grover, C.E.; Hu, G.; Pan, M.; Miller, E.R.; Conover, J.L.; Hunt, S.P.; Udall, J.A.; Wendel, J.F. Parallel and Intertwining Threads of Domestication in Allopolyploid Cotton. *Adv. Sci.* **2021**, doi:10.1002/advs.202003634.

21. Huang, G.; Huang, J.-Q.; Chen, X.-Y.; Zhu, Y.-X. Recent Advances and Future Perspectives in Cotton Research. *Annu. Rev. Plant Biol.* **2021**, doi:10.1146/annurev-arplant-080720-113241.

22. Yang, Z.; Qanmber, G.; Wang, Z.; Yang, Z.; Li, F. Gossypium Genomics: Trends, Scope, and Utilization for Cotton Improvement. *Trends Plant Sci.* **2020**, *25*, 488–500, doi:10.1016/j.tplants.2019.12.011.

23. Mathangadeera, R.W.; Hequet, E.F.; Kelly, B.; Dever, J.K.; Kelly, C.M. Importance of cotton fiber elongation in fiber processing. *Ind. Crops Prod.* **2020**,

doi:10.1016/j.indcrop.2020.112217.

24. Yang, Z.; Ge, X.; Yang, Z.; Qin, W.; Sun, G.; Wang, Z.; Li, Z.; Liu, J.; Wu, J.; Wang, Y.; et al. Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* **2019**, doi:10.1038/s41467-019-10820-x.

25. Huang, G.; Wu, Z.; Percy, R.G.; Bai, M.; Li, Y.; Frelichowski, J.E.; Hu, J.; Wang, K.; Yu, J.Z.; Zhu, Y. Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. *Nat. Genet.* **2020**, *52*, 516–524, doi:10.1038/s41588-020-0607-4.

26. Sun, Q.; Huang, J.; Guo, Y.; Yang, M.; Guo, Y.; Li, J.; Zhang, J.; Xu, W. A cotton NAC domain transcription factor, GhFSN5, negatively regulates secondary cell wall biosynthesis and anther development in transgenic Arabidopsis. *Plant Physiol. Biochem.* **2020**, *146*, 303–314, doi:10.1016/j.plaphy.2019.11.030.

27. Wang, M.; Tu, L.; Yuan, D.; Zhu, D.; Shen, C.; Li, J.; Liu, F.; Pei, L.; Wang, P.; Zhao, G.; et al. Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. *Nat. Genet.* **2019**, *51*, 224–229, doi:10.1038/s41588-018-0282-x.

28. Wang, K.; Wang, Z.; Li, F.; Ye, W.; Wang, J.; Song, G.; Yue, Z.; Cong, L.; Shang, H.; Zhu, S.; et al. The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.* **2012**, doi:10.1038/ng.2371.

29. Li, F.; Fan, G.; Lu, C.; Xiao, G.; Zou, C.; Kohel, R.J.; Ma, Z.; Shang, H.; Ma, X.; Wu, J.; et al. Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **2015**, doi:10.1038/nbt.3208.

30. Du, X.; Huang, G.; He, S.; Yang, Z.; Sun, G.; Ma, X.; Li, N.; Zhang, X.; Sun, J.; Liu, M.; et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* **2018**, doi:10.1038/s41588-018-0116-x.

31. Hu, Y.; Chen, J.; Fang, L.; Zhang, Z.; Ma, W.; Niu, Y.; Ju, L.; Deng, J.; Zhao, T.; Lian, J.; et al. Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* **2019**, doi:10.1038/s41588-019-0371-5.

32. Prasad, P.; Khatoon, U.; Verma, R.K.; Aalam, S.; Kumar, A.; Mohapatra, D.; Bhattacharya, P.; Bag, S.K.; Sawant, S.V. Transcriptional landscape of cotton fiber development and its alliance with fiber-associated traits. *Front. Plant Sci.* 99.

33. Lü, S.; Song, T.; Kosma, D.K.; Parsons, E.P.; Rowland, O.; Jenks, M.A. Arabidopsis CER8 encodes LONG-CHAIN ACYL-COA SYNTHETASE 1 (LACS1) that has overlapping functions with LACS2 in plant wax and cutin synthesis. *Plant J.* **2009**, doi:10.1111/j.1365-313X.2009.03892.x.

34. Pinot, F.; Beisson, F. Cytochrome P450 metabolizing fatty acids in plants: Characterization and physiological roles. *FEBS J.* 2011.

35. Xiang, W.; Wang, X.; Ren, T. Expression of a wheat cytochrome P450 monooxygenase cDNA in yeast catalyzes the metabolism of sulfonylurea herbicides. *Pestic. Biochem. Physiol.* **2006**, doi:10.1016/j.pestbp.2005.09.001.

36. Rao, M.J.; Xu, Y.; Tang, X.; Huang, Y.; Liu, J.; Deng, X.; Xu, Q. CSCYT75B1, a citrus CYTOCHROME P450 gene, is involved in accumulation of antioxidant flavonoids and induces drought tolerance in transgenic arabidopsis. *Antioxidants* **2020**, doi:10.3390/antiox9020161.

37. Schuhegger, R.; Nafisi, M.; Mansourova, M.; Petersen, B.L.; Olsen, C.E.; Svatoš, A.; Halkier, B.A.; Glawischnig, E. CYP71B15 (PAD3) catalyzes the final step in camalexin biosynthesis. *Plant Physiol.* **2006**, doi:10.1104/pp.106.082024.

38. Jabran, K.; Mahajan, G.; Sardana, V.; Chauhan, B.S. Allelopathy for weed control in agricultural systems. *Crop Prot.* 2015.

39. Heitz, T.; Widemann, E.; Lugan, R.; Miesch, L.; Ullmann, P.; Désaubry, L.; Holder, E.; Grausem, B.; Kandel, S.; Miesch, M.; et al. Cytochromes P450 CYP94C1 and CYP94B3 catalyze two successive oxidation steps of plant hormone jasmonoyl-isoleucine for catabolic turnover. *J. Biol. Chem.* **2012**, doi:10.1074/jbc.M111.316364.

40. Magwanga, R.O.; Lu, P.; Kirungu, J.N.; Dong, Q.; Cai, X.; Zhou, Z.; Wang, X.; Hou, Y.; Xu, Y.; Peng, R.; et al. Knockdown of cytochrome P450 genes Gh_D07G1197 and Gh_A13G2057 on chromosomes D07 and A13 reveals their putative role in enhancing drought and salt stress tolerance in Gossypium hirsutum. *Genes (Basel).* **2019**, *10*, doi:10.3390/genes10030226.

41. Huang, G.; Wu, Z.; Percy, R.G.; Bai, M.; Li, Y.; Frelichowski, J.E.; Hu, J.; Wang, K.; Yu, J.Z.; Zhu, Y. Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. *Nat. Genet.* **2020**, doi:10.1038/s41588-020-0607-4.

42. Li, Y.; Wei, K. Comparative functional genomics analysis of cytochrome P450 gene superfamily in wheat and maize. *BMC Plant Biol.* **2020**, *20*, 1–22, doi:10.1186/s12870-020-2288-7.

43. Yu, J.; Tehrim, S.; Wang, L.; Dossa, K.; Zhang, X.; Ke, T.; Liao, B. Evolutionary history and functional divergence of the cytochrome P450 gene superfamily between Arabidopsis thaliana and Brassica species uncover effects of whole genome and tandem duplications. *BMC Genomics* **2017**, doi:10.1186/s12864-017-4094-7.

44. Wang, M.; Li, P.; Ma, Y.; Nie, X.; Grebe, M.; Men, S. Membrane sterol composition in arabidopsis thaliana affects root elongation via auxin biosynthesis. *Int. J. Mol. Sci.* **2021**, *22*, 1–20, doi:10.3390/ijms22010437.

45. Xu, J.; Wang, X.Y.; Guo, W.Z. The cytochrome P450 superfamily: Key players in plant development and defense. *J. Integr. Agric.* 2015.

46. Ohnishi, T.; Watanabe, B.; Sakata, K.; Mizutani, M. CYP724B2 and CYP90B3 function in the early C-22 hydroxylation steps of brassinosteroid biosynthetic pathway in tomato. *Biosci. Biotechnol. Biochem.* **2006**, doi:10.1271/bbb.60034.

47. Prakash, P.; Srivastava, R.; Prasad, P.; Kumar Tiwari, V.; Kumar, A.; Pandey, S.; Sawant, S. V Trajectories of cotton fiber initiation: a regulatory perspective. *Preprints* **2020**, doi:10.20944/preprints202011.0060.v1.

48. Davis, L.A.; Addicott, F.T. Abscisic Acid: Correlations with Abscission and with Development in the Cotton Fruit. *Plant Physiol.* **1972**, *49*, 644–648, doi:10.1104/pp.49.4.644.

49. Gilbert, M.K.; Bland, J.M.; Shockey, J.M.; Cao, H.; Hinchliffe, D.J.; Fang, D.D.; Naoumkina, M. A Transcript Profiling Approach Reveals an Abscisic Acid-Specific Glycosyltransferase (UGT73C14) Induced in Developing Fiber of Ligon lintless-2 Mutant of Cotton (Gossypium hirsutum L.). *PLoS One* **2013**, doi:10.1371/journal.pone.0075268.

50. Deng, F.; Tu, L.; Tan, J.; Li, Y.; Nie, Y.; Zhang, X. GbPDF1 is involved in cotton fiber initiation via the core cis-element HDZIP2ATATHB2. *Plant Physiol.* **2012**, *158*, 890–904, doi:10.1104/pp.111.186742.

51. Yang, Z.; Zhang, C.; Yang, X.; Liu, K.; Wu, Z.; Zhang, X.; Zheng, W.; Xun, Q.; Liu, C.; Lu, L.; et al. PAG1, a cotton brassinosteroid catabolism gene, modulates fiber elongation. *New Phytol.* **2014**, *203*, 437–448, doi:10.1111/nph.12824.

52. Li, Q.; Chakrabarti, M.; Taitano, N.K.; Okazaki, Y.; Saito, K.; Al-Abdallat, A.M.; Van Der Knaap, E. Differential expression of SlKLUH controlling fruit and seed weight is associated with changes in lipid metabolism and photosynthesis-related genes. *J. Exp. Bot.* **2021**, doi:10.1093/jxb/eraa518.

53. Zhu, D.; Le, Y.; Zhang, R.; Li, X.; Lin, Z. A global survey of the gene network and key genes for oil accumulation in cultivated tetraploid cottons. *Plant Biotechnol. J.* **2021**, doi:10.1111/pbi.13538.

54. Prasad, P.; Khatoon, U.; Verma, R.K.; Kumar, A.; Mohapatra, D.; Bhattacharya, P.; Bag, S.K.; Sawant, S. V Unravelling cotton RNAseq repositories to the fiber development specific modules and their alliance with the fiber-related traits. *bioRxiv* **2021**, 2021.02.13.431059.

55. Katsumata, T.; Fukazawa, J.; Magome, H.; Jikumaru, Y.; Kamiya, Y.; Natsume, M.; Kawaide, H.; Yamaguchi, S. Involvement of the CYP78A subfamily of cytochrome P450 monooxygenases in protonema growth and gametophore formation in the moss Physcomitrella patens. *Biosci. Biotechnol. Biochem.* **2011**, *75*, 331–336, doi:10.1271/bbb.100759.

56. Ma, M.; Wang, Q.; Li, Z.; Cheng, H.; Li, Z.; Liu, X.; Song, W.; Appels, R.; Zhao, H. Expression of TaCYP78A3, a gene encoding cytochrome P450 CYP78A3 protein in wheat (Triticum aestivum L.), affects seed size. *Plant J.* **2015**, doi:10.1111/tpj.12896.

57. Khan, M.H.U.; Hu, L.; Zhu, M.; Zhai, Y.; Khan, S.U.; Ahmar, S.; Amoo, O.; Zhang, K.; Fan, C.; Zhou, Y. Targeted mutagenesis of EOD3 gene in Brassica napus L. regulates seed production. *J. Cell. Physiol.* **2020**, doi:10.1002/jcp.29986.

58. Zhang, H.; Han, W.; Wang, H.; Cong, L.; Zhai, R.; Yang, C.; Wang, Z.; Xu, L. Downstream of GA4, PbCYP78A6 participates in regulating cell cycle-related genes

and parthenogenesis in pear (Pyrus bretshneideri Rehd.). *BMC Plant Biol.* **2021**, *21*, 1–13, doi:10.1186/s12870-021-03098-z.

59. Poretska, O.; Yang, S.; Pitorre, D.; Poppenberger, B.; Sieberer, T. AMP1 and CYP78A5/7 act through a common pathway to govern cell fate maintenance in Arabidopsis thaliana. *PLoS Genet.* **2020**, doi:10.1371/journal.pgen.1009043.

60. Vercruysse, J.; Baekelandt, A.; Gonzalez, N.; Inzé, D. Molecular networks regulating cell division during Arabidopsis leaf growth. *J. Exp. Bot.* 2021.

61. Zhao, B.; Cao, J.F.; Hu, G.J.; Chen, Z.W.; Wang, L.Y.; Shangguan, X.X.; Wang, L.J.; Mao, Y.B.; Zhang, T.Z.; Wendel, J.F.; et al. Core cis-element variation confers subgenome-biased expression of a transcription factor that functions in cotton fiber elongation. *New Phytol.* **2018**, doi:10.1111/nph.15063.

62. Wang, L.; Cook, A.; Patrick, J.W.; Chen, X.Y.; Ruan, Y.L. Silencing the vacuolar invertase gene GhVIN1 blocks cotton fiber initiation from the ovule epidermis, probably by suppressing a cohort of regulatory genes via sugar signaling. *Plant J.* **2014**, doi:10.1111/tpj.12512.

63. Zhang, M.; Zeng, J.Y.; Long, H.; Xiao, Y.H.; Yan, X.Y.; Pei, Y. Auxin regulates cotton fiber initiation via GHPIN-mediated auxin transport. *Plant Cell Physiol.* **2017**, doi:10.1093/pcp/pcw203.

64. Wu, Y.; Machado, A.C.; White, R.G.; Llewellyn, D.J.; Dennis, E.S. Expression profiling identifies genes expressed early during lint fibre initiation in cotton. *Plant Cell Physiol.* **2006**, doi:10.1093/pcp/pci228.

65. Zhao, L.; Cai, H.; Su, Z.; Wang, L.; Huang, X.; Zhang, M.; Chen, P.; Dai, X.; Zhao, H.; Palanivelu, R.; et al. KLU suppresses megasporocyte cell fate through SWR1-mediated activation of WRKY28 expression in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, doi:10.1073/pnas.1716054115.

66. Kumar, V.; Singh, B.; Singh, S.K.; Rai, K.M.; Singh, S.P.; Sable, A.; Pant, P.; Saxena, G.; Sawant, S. V. Role of GhHDA5 in H3K9 deacetylation and fiber initiation in Gossypium hirsutum. *Plant J.* **2018**, *95*, 1069–1083, doi:10.1111/tpj.14011.

67. Hu, L.; Yang, X.; Yuan, D.; Zeng, F.; Zhang, X. GhHmgB3 deficiency deregulates

proliferation and differentiation of cells during somatic embryogenesis in cotton. *Plant Biotechnol. J.* **2011**, doi:10.1111/j.1467-7652.2011.00617.x.

68. Yu, J.; Jung, S.; Cheng, C.H.; Ficklin, S.P.; Lee, T.; Zheng, P.; Jones, D.; Percy, R.G.; Main, D. CottonGen: A genomics, genetics and breeding database for cotton research. *Nucleic Acids Res.* **2014**, doi:10.1093/nar/gkt1064.

69. Johnson, L.S.; Eddy, S.R.; Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **2010**, doi:10.1186/1471-2105-11-431.

70. Chen, C.; Chen, H.; Zhang, Y.; Thomas, H.R.; Frank, M.H.; He, Y.; Xia, R. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* **2020**, doi:10.1016/j.molp.2020.06.009.

71. Sievers, F.; Higgins, D.G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **2018**, doi:10.1002/pro.3290.

72. Nguyen, L.T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274, doi:10.1093/molbev/msu300.

73. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, doi:10.1093/nar/gkz239.

74. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915, doi:10.1038/s41587-019-0201-4.

75. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. RNA-seq experiments with HISAT , StringTie and Ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667, doi:10.1038/nprot.2016-095.

76. Abu-Jamous, B.; Kelly, S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biol.* **2018**, doi:10.1186/s13059-018-1536-8.

77. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-

quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, doi:10.1038/msb.2011.75.

78. Lescot, M.; Déhais, P.; Thijs, G.; Marchal, K.; Moreau, Y.; Van De Peer, Y.; Rouzé, P.; Rombauts, S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* **2002**, doi:10.1093/nar/30.1.325.

79. Ge, S.X.; Jung, D.; Jung, D.; Yao, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **2020**, doi:10.1093/bioinformatics/btz931.

80. Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, H.; Vilo, J. G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **2019**, doi:10.1093/nar/gkz369.

81. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939, doi:10.1111/j.1365-313X.2004.02016.x.

82. Schmittgen, T.D.; Livak, K.J. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **2001**.

Figure Legends

Fig. 1 Phylogenetic clustering of CYP450 in three diploid and two allotetraploid cotton species. **A.** Representation of CYP450 in 10 different clan in *Gossypium raimondii* (D5) **B.** *Gossypium herbaceum* (A1) **C.** *Gossypium arboreum* (A2) **D.** *Gossypium hirsutum* and (AD1) and **E.** *Gossypium barbadense* (AD2) cotton genome. Ten different CYP clan has been represented in different color while the size of the triangle on each branch represents the bootstrap value in percentage.

Fig. 2 Genome wide whole genome expression profiling of identified CYPs in *Gossypium hirsutum* at different fiber developmental stages viz., -3 DPA, -1 DPA, 0 DPA, 1 DPA, 3 DPA of ovule and 5 DPA, 10 DPA, 20 DPA and 25 DPA of fiber tissue for A-type clade in **A**. At subgenome **B.** Dt subgenome and for non A-type clade in **C**. At subgenome and **D**. Dt subgenome of cotton. Each row represents the log2 transformed FPKM value (Fragments per kilobase per million reads) of GhCYPs with the color code ranging from the 0 to 4. The higher

 doi:10.20944/preprints202203.0049.v1

expression value depicted in red color while blue color denoted the lower expression of GhCYP in different fiber developmental stages.

Fig. 3 Genome wide whole genome expression profiling of CYPs for **A.** *Gossypium raimondii* at 10 DPA and 20 DPA fiber developmental stages and at 0 DPA, 10 DPA, 20 DPA fiber developmental stages for **B.** *Gossypium herbaceum* **C.** *Gossypium arboreum* and **D**. **(i)** At subgenome and **(ii)** Dt subgenome of *Gossypium barbadense*. Expression value is represented in log2 scale of FPKM where red colored cell shows the higher expression value while blue colored cell reflects the lower expression of CYP at different fiber developmental stages.

Fig. 4 Functional characterization of fiber clustered *Gossypium hirsutum* CYPs. **A.** Line graph representing the normalized expression value of GhCYPs into six distinct clusters namely C1, C2, C3, C4, C5 and C6 consists of 49, 19, 11, 41, 36 and 18 GhCYPs respectively. Clustering is based on their raw expression value in different fiber developmental stages. **B.** Gene Ontology (GO) enrichment of each cluster with the significance level (FDR value) < 0.05. The most relevant GO term of each cluster has been indicated through orange bar.

Fig. 5 Real time expression validation of different members of fiber cluster in leaf and different stages of fiber development viz., at -3 DPA, 0 DPA, 10 DPA, 21 DPA and 30 DPA. Y axis represents the relative expression of gene in reference to the UBQ as an internal control. Expression validation of **A.** GhCYP78A198_A, **B.** GhCYP749A71_D, **C.** GhCYP83F48_A and **D.** GhCYP94B70_D of C1 showed relatively higher expression at early stage of fiber development while **E.** GhCYP87A54_D of C3 **F.** GhCYP74A1_D of C4 **G.** GhCYP706B28P_A of C5 and **H.** GhCYP84A80_D of C6 cluster showed the higher expression at later stage of fiber development.

Fig. 6 Evolutionary interpretation of CYP78A in twenty angiosperm species grouped into eight distinct group namely Group I to VIII. Group I, IV and VIII were annotated as EOD3, PLA1 and KLUH clade. CYP78A of different angiosperm plant was represented with different color while turquoise branch color was used for the Malvaceae family CYPs.

Fig. 7 In silico expression analyses of cotton CYP78A for different fiber developmental stages. Bar plot represents the log2 (FPKM + 1) expression value at **A.** 10 DPA and 20 DPA of ovule tissue in *Gossypium raimondii.* 0 DPA, 10 DPA and 20 DPA of fiber developmental stages in **B**. *Gossypium herbaceum* **C.** *Gossypium arboreum* ***D.*** *Gossypium hirsutum and* **E.** *Gossypium barbadense*. **F.** Box plot represented the GhCYP78 expression in wild (Xu142) and fiberless

mutant (Xu142*fl*) of cotton at 0 DPA fiber developmental stages. ns used for the nonsignificant differences with the p value > 0.05.

Fig. 8 Comprehensive study of GhCYP78A_197 and GhCYP78A_198. **A.** Cis-Regulatory Element (CRE) analysis of GhCYP78A in 1000 bp upstream to Transcription Start Site. Y axis represents the percentage of occurrence of classified CRE, visualized in different color. Coexpression network interpretation of **B.** GhCYP78A197_A and **C.** GhCYP78A198_A with positive correlation ($r \geq 0.95$) and negative correlation ($r \leq -0.95$) at early stage of fiber development. Different colored genes highlight the literature based important gene, TFs and phytohormone responsible for the fiber initiation. **D-O**. Mapman pathway enrichment (of positively (I and III) and negatively (II and IV) coexpressed genes of GhCYP78A197_A and GhCYP78A198_A.

Fig. 9 Working model representing the mode of action of subgenome expression biased of GhCYP78A197_A and GhCYP78A198_A in downregulation to Gibberellic Acid (GA). Upstream region of At subgenome of these two genes are enriched with the MYB and TATA box motif that induces the expression at transcript level which ultimately activate and coexpressed with other phytohormones, chromatin remodeler and cell cycle related genes that undergone the endoreduplication at early stages of fiber development to define the fiber development and thus increases the high fiber yield. GA- Gibberellic Acid; TSS – Transcription Start Site; CDK – Cyclin Dependent kinase; BR- Brassinosteroid.

Supplementary materials

Fig. S1 Chromosome localization of identified CYPs in 13 chromosome of diploid cotton **A.** *Gossypium raimondii* (D5) **B.** *Gossypium herbaceum* (A1) **C**. *Gossypium arboreum* (A2) and 26 chromosome of **D**. *Gossypium hirsutum* and (AD1) and **E.** *Gossypium barbadense* (AD2) cotton genome. Yellow chromosome bar was used for the A genome and At subgenome whereas purple chromosome bar represented for D and Dt subgenome of cotton.

Fig. S2

Comparative illustration of CYPs in 20 different plant lineages. **A.** Time scale of selected 20 plant lineages in MYA (**M**illion **Y**ear **A**go) scale. Classified members of CYP450 in 20 plants for **B.** A type clan and **C.** non A type clan. Higher number of CYPs were represented through the red cell and lower number of CYPs were represented through the blue colored cell. Yellow

 doi:10.20944/preprints202203.0049.v1

and orange colored bar showed the "0' members of respective CYPs in cotton species. Red colored bar of CYP749 was used for the highest number CYP in all selected Malvaceae family.

Fig. S3 Functional characterization of *Gossypium hirsutum* CYPs in different stress condition namely Cold, Hot, Salt and PEG in compare to Control. Line graph representing the normalized expression value of GhCYPs into eight distinct clusters viz., C1 to C8.

Fig. S4 Significant Gene Ontology enrichment of the fiber cluster GhCYPs (C1 to C6) for the Molecular Function (MF) and Cellular Component (CC). FDR value less than 0.05 was considered for the significant GO Term.

Fig. S5 Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment of the six fiber cluster (C1 to C6) of GhCYP with significant FDR < 0.05.

Fig. S6 Phylogenetic clustering of GhCYP78A with 1000 bootstrap value. Clade positioning of GhCYP78A197 and GhCYP78A198 indicated that they are parallelly diverged from the GhCYP78A199.

Fig. S7 Illustration of uniquely mapped read intensity for At and Dt subgenome of **A.** GhCYP78A197 and **B.** GhCYP78A198 at -3 DPA, -1 DPA, 0 DPA, 1 DPA, 3 DPA and 5 DPA. Red line was used for the estimation of basal level expression in compare to the 5 DPA expression level. Dt subgenome of both genes at different fiber development stage showed no mapped read on *Gossypium hirsutum* genome.

Fig. S8 Cis-Regulatory Element (CRE) representation of GhCYP78A in 1000bp upstream region of Transcription Start Site (TSS). Horizontal scale of each GhCYP represented the promoter region where colored vertical bar was used for different CRE. Width of the vertical bar highlighted the size of CRE motif on the genome. 3'end of the genomic scale representing the TSS site.

Fig. S9 Venny representation of stress clustered and fiber clustered GhCYPs in which 61 CYPs was commonly clustered for both the condition. The expression level of commonly clustered GhCYPs was further illustrated through the heatmap where they were categorized according to different fiber cluster. C1 fiber clustered GhCYP have higher number of CYPs with higher expression level (log2 (FPKM+ 1)) in stress condition and early stage of fiber development.

Fig. S10 Illustration of uniquely mapped read intensity on *Gossypium hirsutum* genome for cyclin related gene (CYCD3) and chromatin remodeler (SNF1 and HMGB) at -3 DPA, -1 DPA,

 doi:10.20944/preprints202203.0049.v1

0 DPA, 1 DPA, 3 DPA and 5 DPA. Red line was used for the estimation of basal level expression in compare to the 5 DPA expression level.

Table S1 SRA accession numbers that were accessed for the different fiber developmental stages of different cotton species

Table S2 Primer used for this study along with the internal control primer sequence of the UBQ14.

Supplementary Dataset 1 – Identification and characterization of CYP450 in *Gossypium raimondii* genome with haeme binding and ERR triad motif position.

Supplementary Dataset 2 – Identification and characterization of CYP450 in *Gossypium herbaceum* genome with haeme binding and ERR triad motif position.

Supplementary Dataset 3 – Identification and characterization of CYP450 in *Gossypium arboreum* genome with haeme binding and ERR triad motif position.

Supplementary Dataset 4 – Identification and characterization of CYP450 in *Gossypium hirsutum* genome with haeme binding and ERR triad motif position.

Supplementary Dataset 5 – Identification and characterization of CYP450 in *Gossypium barbadense* genome with haeme binding and ERR triad motif position.

Supplementary Dataset 6 – Transmembrane helixes predication of identified CYPs in three diploid and allotetraploid cotton.

Supplementary Dataset 7 – Normalized log2 transformed expression value of each fiber cluster viz., C1 to C6 at different fiber developmental stages namely -3 DPA, -1 DPA, 0 DPA, 1 DPA, 3 DPA, 5 DPA, 10 DPA, 20 DPA and 20 DPA.