Biochemical Genetics (2019) 57:248–272
https://doi.org/10.1007/s10528-018-9888-z

ORIGINAL ARTICLE

# Characterization of Cellulose Synthase A (CESA) Gene Family in Eudicots

Muhammad Amjad Nawaz[1] · Xiao Lin[2] · Ting-Fung Chan[2] · Muhammad Imtiaz[3] · Hafiz Mamoon Rehman[1] · Muhammad Amjad Ali[4] · Faheem Shehzad Baloch[5] · Rana Muhammad Atif[6] · Seung Hwan Yang[1] · Gyuhwa Chung[1]

Received: 31 January 2018 / Accepted: 20 September 2018 / Published online: 28 September 2018

## Abstract

Cellulose synthase A (CESA) is a key enzyme involved in the complex process of plant cell wall biosynthesis, and it remains a productive subject for research. We employed systems biology approaches to explore structural diversity of eudicot CESAs by exon–intron organization, mode of duplication, synteny, and splice site analyses. Using a combined phylogenetics and comparative genomics approach coupled with co-expression networks we reconciled the evolution of cellulose synthase gene family in eudicots and found that the basic forms of CESA proteins are retained in angiosperms. Duplications have played an important role in expansion of CESA gene family members in eudicots. Co-expression networks showed that primary and secondary cell wall modules are duplicated in eudicots. We also identified 230 simple sequence repeat markers in 103 eudicot CESAs. The 13 identified conserved motifs in eudicots will provide a basis for gene identification and functional characterization in other plants. Furthermore, we characterized (in silico) eudicot CESAs against senescence and found that expression levels of CESAs decreased during leaf senescence.

**Keywords** Cell wall · Cellulose synthase-A · Eudicots · Phylogenetic construction · Comparative co-expression · Senescence

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10528-018-9888-z) contains supplementary material, which is available to authorized users.

✉ Seung Hwan Yang
ymichigan@chonnam.ac.kr

✉ Gyuhwa Chung
chung@chonnam.ac.kr

Extended author information available on the last page of the article

## Introduction

Eudicots are a major group of angiosperms characterized by two cotyledons and tricolpate pollen (Worberg et al. 2007). Being prominent species in sustainable agricultural operations, these land plants are of great economic importance, as they are commonly used for food, feed, fiber, fuel, paper and structural materials. The evolutionary inclusion of cellulose in the cell wall (CW) facilitated colonization of plants to terrestrial environment (Somerville 2006) by enabling them to withstand internal turgor pressure, retain essential water levels for cellular functioning, and control material entering and exiting the cell. The walls of plant cells play essential roles in determining cell size expansion and protecting cells from pathogens (Maleki et al. 2016). Cellulose, which consists of 24–36 β(1–4)-linked glucan residue chains (500 and 14,000 monomers long) interlinked by hydrogen bonds to create a regular crystalline structure that form a stabilized inert core, is the primary load-bearing and organizing structure of the CW and as such is the most abundant biopolymer and renewable carbon source on the planet (Carroll and Specht 2011). Properties of cellulose are determined by coordinated synthesis of multiple glucan chains whose collective characteristics confer higher tensile strength and rigidity to the CW, as well as improve resistance to chemical attacks (Somerville 2006; Olek et al. 2014).

The cellulose synthase (CESA) gene family, which belongs to the cellulose synthase-like superfamily, are responsible for the synthesis of cellulose. This gene family is characterized by Type 1 glycosyltransferases of type-2 glycosyltransferases containing three aspartyl residues, a QxxRW motif and a zinc-finger domain (Kaur et al. 2016). Up to eight transmembrane domains are present in the CESAs of plants, forming a pore across the inner membrane to embed cellulose chain secretion through the CW (Doblin et al. 2002; Kaur et al. 2016; Maleki 2016). These synthases are located in the plasma membrane and cytosol, where they act as a large synthetic complex (Kumar and Turner 2015). Evidence to date has suggested that these complexes form “particle rosettes” made up of hexamers, but the exact number of CESAs involved in the making of these complexes, as well as their stoichiometry, remains unknown (Doblin et al. 2002; Carroll and Specht 2011). Functional analyses of 10 Arabidopsis CESAs have revealed that three different gene products CESA1, CESA3, and CESA6 are essential for cellulose deposition in the primary CW (PCW), whereas CESA4, CESA7 and CESA8 are implicated in cellulose deposition in the secondary CW (SCW) (Taylor et al. 2004; Persson et al. 2007). In addition, CESA2, CESA5, and CESA9 are considered partially redundant with CESA6, as they exhibit non-overlapping expression patterns, and null mutants of each of these genes showed less severe alterations in seed coat development and root/hypocotyl elongation (Carroll and Specht 2011). Recently, CESA6 was proposed as a subfamily with four members, i.e. CESA2, 5, 6, and 9. The functional redundancy between the CESA6 subfamily was reported to be partial and one CESA may substitute for another, but not completely, that concludes that there are functional differences between CESA6 subfamily members (Persson et al. 2007; Ruprecht et al. 2017).

The role of the cellulose biosynthesizing gene family has been extensively studied in bacteria, plants, and algae (Yin et al. 2009; Carroll and Specht 2011; Römling and Galperin 2015; Little et al. 2018). Among land plants, the CESA gene family has been assessed in wheat, barley, maize, pine, rice, popular, Arabidopsis, grapevine, soybean and sorghum (Holland et al. 2000; Appenzeller et al. 2004; Burton et al. 2004; Carroll and Specht 2011; Kaur et al. 2016; Rai et al. 2016; Nawaz et al. 2017a). Considerable progress has been made in identifying the full range of functions and molecular machinery of cellulose biosynthesis in plants (Roberts and Roberts 2004; Kumar and Turner 2015) particularly in model plant Arabidopsis (10 CESAs), popular (17 CESAs) and grapevine (13 CESAs) species (Carroll and Specht 2011). Evolutionary analysis of CESA gene modules based on combined phylostratigraphic analysis and phylogenetic data was recently discussed in angiosperms. Emergence of PCW and SCW modules in angiosperms was studied with the findings that cellulose biosynthesis related modules are well conserved in land plants. Duplication of cellulose biosynthesis modules in bryophyte lineage (splitting of mosses and vascular plants), and later the PCW and SCW duplicated in land plants. This duplication dated back in the ancestors of angiosperms as in angiosperms PCW and SCW modules are well conserved (Ruprecht et al. 2017). Technological improvements in whole-genome sequencing and annotation data of eudicots, including soybean, barrel clover, turnip, cotton, common bean, and red clover, over the past 5 years (2010–2015) calls for updating analyses of the CESA gene family in eudicots (Schmutz et al. 2010, 2014; Wang et al. 2011, 2012a; Young et al. 2011; De Vega et al. 2015).

Molecular evolutionary analysis of a particular gene family can be studied at two levels, i.e., comparative genomics and molecular phylogenetics. In comparative genomics different genomic features can be compared between two or more organisms. Whereas molecular phylogenetics deals with DNA or protein sequences to infer relationships between organisms and genes in the form of phylogenetic tree. Particularly, multispecies phylogenetic trees are of great significance as they improve our understanding about speciation and duplication history of genes within a gene family (Ruprecht et al. 2016, 2017). Based on these two molecular evolutionary approaches one can infer the evolutionary history of gene or gene families; however, knowledge about functional gene modules cannot be studied. Furthermore, knowledge about neo- or sub-functionalization of gene paralogues needs to study gene modules along with molecular evolutionary approaches. Recently, a new approach was introduced to study gene modules with respect to evolution (Proost and Mutwil 2016; Ruprecht et al. 2017). Co-expression networks are conserved across species and even across distinct kingdoms of life clearly indicating cross-kingdom orthologs presence. Co-expression networks have emerged as an important tools to rapidly infer the functional relatedness among genes belonging to similar functions/pathways. Phylogenomics supported with transcriptionally coordinated gene modules can explain very well the emergence of new traits of organisms (Ruprecht et al. 2017). Duplication and/or depletion of these modules have also been observed in some species. To understand evolution, duplication/depletion of cellulose synthase gene family, we studied phylogenomics of CESA gene family supported with gene co-expression networks.

To explore structural diversity of eudicots CESAs, we performed a genome-wide comparative analysis of CESA sequences and constructed the phylogeny of the CESA gene family in 10 eudicot species, including those of Arabidopsis. Domain organization, motif discovery, gene structure, gene duplication, splice site and synteny analyses proved to be highly effective for gaining deeper insights into the evolutionary relationships among the CESAs gene family in eudicots. We extended our analysis to system's characterization of CESAs during senescence by employing available RNA expression data. Gene families controlling CW played crucial roles in the evolution of land plants and continue to play important roles in a broad range of biological processes (Doblin et al. 2002; Rai et al. 2016). CESA gene families have been analyzed earlier (Richmond and Somerville 2000; Somerville 2006; Suzuki et al. 2006; Yin et al. 2009; Carroll and Specht 2011); hence, this study aimed at phylogenomic analysis of CESA gene family in eudicots aided with molecular phylogenetics, comparative genomics and co-expression networks, to better understand their evolution.

## Materials and Methods

### Identification of Eudicot CESA Genes

Putative CESA sequences for nine eudicots (*Brassica rapa*, *Cucumis sativus*, *Glycine max*, *Gossypium raimondii*, *Medicago truncatula*, *Phaseolus vulgaris*, *Populus trichocarpa*, *Trifolium pratense*, and *Vitis vinifera*) were retrieved using 10 Arabidopsis CESAs in a BLASTp 2.2.28+ search of available sequences in the Joint Genome Institute (JGI; https://www.phytozome.net) (Goodstein et al. 2011) (Tables 1 and S1). The amino acids sequences of the putative proteins were formatted into FASTA files where each file contained the sequences belonging to an individual species. In order to define the orthogroups, we performed the OrthoFinder

**Table 1** CESA members of selected eudicots

| Plant Species | Common name | Genome Sequenced reference | Total CESA genes |
|---|---|---|---|
| *Arabidopsis thaliana* | Arabidopsis | The Arabidopsis Initiative (2000) | 10 |
| *Brassica rapa* | Turnip | Wang et al. (2011) | 13 |
| *Cucumis sativus* | Cucumber | Huang et al. (2009) | 8 |
| *Glycine max* | Soybean | Schmutz et al. (2010) | 26 |
| *Gossypium raimondii* | New world cotton | Wang et al. (2012) | 15 |
| *Medicago truncatula* | Barrel clover | Young et al. (2011) | 13 |
| *Phaseolus vulgaris* | Common bean | Schmutz et al. (2014) | 15 |
| *Populus trichocharpa* | Black cotton wood | Tuskan et al. (2006) | 17 |
| *Trifolium pratense* | Red clover | De Vega et al. (2015) | 11 |
| *Vitis vinifera* | Grapevine | Jaillon et al. (2007) | 12 |

(v1.1.8) (Emms and Kelly 2015) analysis using the default parameters. The gene phylogenetic trees were computed visualized using iTOL (https://itol.embl.de/) (Letunic and Bork 2006). The potential domains of CESA family members were identified based on the InterPro v. 60.0 (https://www.ebi.ac.uk/interpro/) (Mitchell et al. 2014) to confirm that each candidate gene was a CESA.

A general BLASTp (2.2.28+) using 10 Arabidopsis CESAs against *Picea abies* (via a spruce genome project available at https://congenie.org/start), *Physcomitrella patens* v. 3.3, and *Selaginella moellendorffii* v. 1.0 (available at https://www.phytozome.net) was performed to retrieve their respective CESA gene sequences. The embryophyte, tracheophyte, and pinophyte plants were included as a means of tracing the evolutionary pathway of the CESA gene family through the eudicots. Furthermore, we also included CESA gene family members of three monocots i.e. sorghum, Rice and *Brachypodium distachyon* (Yin et al. 2014; Schwerdt et al. 2015; Rai et al. 2016).

Gene name, peptide length, isoelctric point/molecular weight, chromosomal location, of CESA genes were obtained from Phytozome 11.0 (https://phytozome.jgi.doe.gov/) and ExPASy (https://web.expasy.org/compute_pi/) (Atrimo et al. 2012).

### Phylogenetic Analysis of Eudicot CESAs

We generated an alignment of all 206 CESA peptides using the alignment program Clustal Omega (Sievers et al. 2011). The evolutionary history was then inferred using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al. 1992). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join (NJ) and BIONJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 523 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016). The tree was visualized and formatted in iTOL (https://itol.embl.de/) (Letunic and Bork 2006).

### Gene Structure Analysis and Motif Identification

We constructed intron–exon maps [35] and classified them according to the structure in the genome, including phase, length, and number based on alignment of coding sequences with genomic sequences as previously described (Barvkar et al. 2012), and developed gene structure diagrams using the web application Gene Structure Display Server (GSDS) (https://gsds.cbi.pku.edu.cn/) (Hu et al. 2014). We determined exon and intron boundaries of all phylogenetic clades using the Analyze feature of PhytoMine, which is available at the JGI website (https://phytozome.jgi.doe.gov/phytomine/begin.do).

The conserved motifs among all the eudicots were manually identified. For this purpose the peptide sequence of all the CESA genes were aligned using Clustal Omega from online webserver (https://www.ebi.ac.uk/Tools/msa/clustalo/). The resultant alignment was then imported to CLC Sequence Viewer (CLC Bio, version 7.5). All the proteins containing truncated and incomplete sequences (that may owe to their coding pseudogenes) were deleted. The motifs sharing high percentage of conservation in all the remaining CESA proteins were identified.

### Gene Duplication and Synteny Analyses

Tandem duplications (TD) and segmental duplications (SD) per phylogenetic group were assessed based on information obtained from the Plant Genome Duplication Database (PGDD; https://chibba.agtec.uga.edu/) (Lee et al. 2012). To further explore the conservation of exon structure, exon maps of duplicated genes in each phylogenetic group were constructed by delineating exon boundaries and counting the number of nucleotides (nt) per exon. Conserved exons of the same length further helped to determine their boundaries as common splice sites (He et al. 2012). To further explore possible conservation of the CESA gene family in eudicots, synteny was visualized with Circoletto for all phylogenetic groups (Darzentas 2010).

### SSR Markers Identification

Genomic sequences of 141 eudicot CESAs were used to identify putative SSRs by employing a high-throughput web tool i.e. BatchPrimer3 (https://batchprimer3.bioinformatics.ucdavis.edu/) (You et al. 2008). We screened the eudicot CESAs for di-nucleotide, tri-nucleotide, tetra-nucleotide, penta-nucleotide and hexa-nucleotide repeats with default settings.

### Co-expression Network Analysis

To determine significant, conserved associations between gene CESA gene labels we used NetworkComparer tool integrated in PlaNet (https://aranet.mpimp-golm.mpg.de/index.html) (Mutwil et al. 2011). As PlaNet hosts only four species under our investigation i.e. Arabidopsis, *M. truncatula*, *G. max* and *P. trichocarpa.* So, we present here the co-expression networks of only four species. Initially, we explored gene co-expression networks of each species separately to explore gene modules related to PCW and SCW. For this we manually assigned genes to PCW and SCW modules based on their grouping in phylogenetic ML tree. Genes with highest label co-occurrence scores were used for visualizing co-expression networks.

### System's Characterization of CESAs Against Senescence

Four set of raw sequencing data were downloaded from SRA, of which two sets were from *A. thaliana* (Brusslan et al. 2015; Woo et al. 2016) (PRJNA186843 and PRJNA280870), two were from *G. max* (Brown and Hudson 2015, 2017) (PRJNA262564 PRJNA339152) and one from *T. pratense* (PRJNA377931). Adapter and quality trimming were performed with Trimmomatic (0.36) (Bolger et al. 2014). Genome sequences and annotation files were downloaded from Ensembl Plant for *A.*

*thaliana* (TAIR10) and Phytozome for *G. max* (275 v2.0) and *T. pratense* (385 v2) (Goodstein et al. 2011; Bolser et al. 2017). HISAT2 (2.1.0) (Kim et al. 2015) was used to map the trimmed data to reference genomes and StringTie (1.3.3b) (Pertea et al. 2015) was used to quantify gene expression levels. Average expression levels of CESA genes for each condition was shown in heatmaps using gplots package in R.

## Results

### Identification of Eudicot CESA Genes

Data mining of the genome databases for nine eudicots (Table 1) led to the identification of 8 to 26 genes containing full protein sequences homologous to 10 Arabidopsis CESA proteins. In total, we identified 140 protein-coding transcripts in ten eudicots via a BLASTp search in Phytozome v. 11.0 (Table S1). The results of this search also included previously published CESAs in the assessed genomes. Collectively, the BLASTp and OrthoFinder results provided a solid foundation for identification of orthologues and construction of phylogenetic tree.

All searched CESA genes contained a CESA-RING-type zinc finger domain and/or a nucleotide-diphosphate-sugar-transferase domain, with a few exceptions. The CESA-RING-type zinc finger domains were typically located between 0 and 250 residues on the N-terminal. The nucleotide-diphospho-sugar-transferase domain was repeated 1–4 times in the CESA genes at variable locations but was generally located in the center and the C-terminal of CESA peptides. Genes from *T. pratense* (*Tp57577_TGAC_v2_gene18720*) and *C. sativus* (Cucsa.212920.1) also contained concanavalin A-like lectin/glucanase domains, and one gene from *G. max* (*Glyma05g29240*) contained a WD40/YVTN repeat-like-containing domain. The presence of these domains suggests that these genes may be involved in a variety of cell functions (Figure S1).

The conserved motifs among all the eudicots were manually identified. For this purpose the peptide sequence of all the CESA genes were aligned using Clustal Omega from online webserver (https://www.ebi.ac.uk/Tools/msa/clustalo/). The resultant alignment was then imported to CLC Sequence Viewer (CLC Bio, version 7.5). All the proteins containing truncated and incomplete sequences (that may owe to their coding pseudogenes) were deleted. The final dataset contained a total of 13 conserved motifs ranging from 12 to 200 amino acids were identified to be strongly conserved among these sequences analyzed (Fig. 1).

### Phylogeny of the CESA Gene Family in Eudicots

To trace the evolutionary path of the CESA gene family throughout the eudicots, we constructed the phylogenetic relationships among 140 identified proteins. First, a ML tree was built to classify eudicot CESA genes; to this aim, the identified protein sequences were aligned with previously characterized Arabidopsis CESA genes.

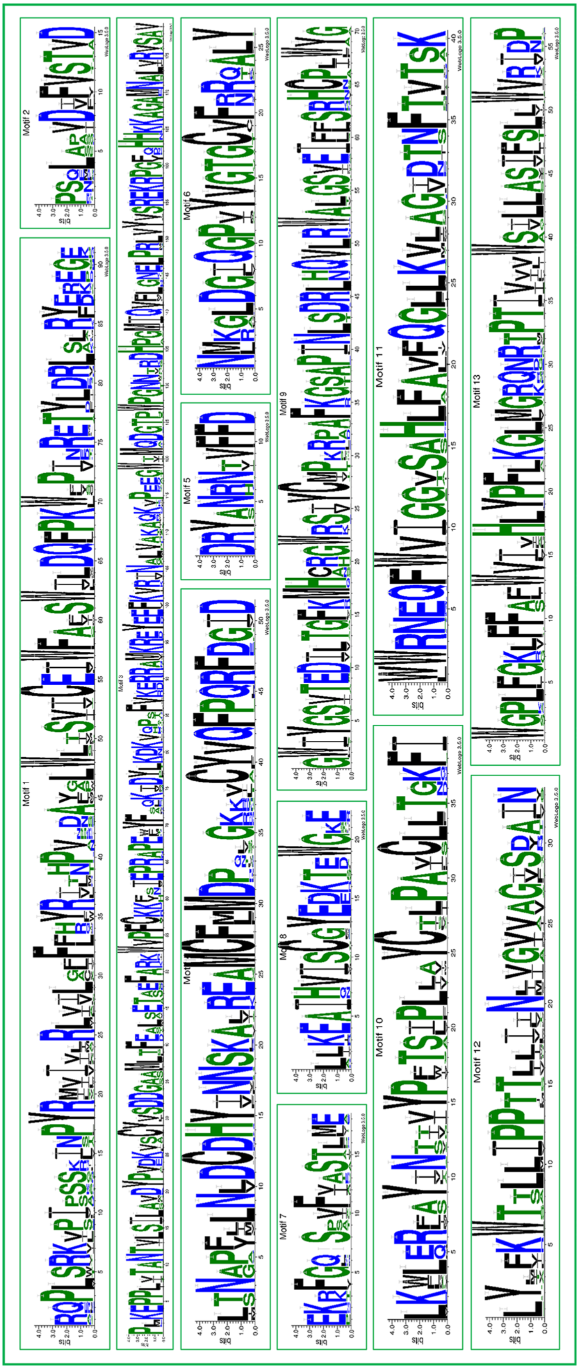

**Fig. 1** Conserved motifs in Cellulose synthase A gene family (Color figure online)

For tracing detailed evolutionary relationships within the eudicots, we added three ancestral species, consisting of an embryophyte (nineteen sequences of *P. patens*), a tracheophyte (eight sequences of *S. moellendorffii*), and a pinophyta (eight sequences of *P. abies*) and three monocots. The resulting ML tree separated the CESAs into PCW and SCW monophyletic groups which further formed CESA specific clads. SCW CESAs clustered according to previous grouping i.e. CESA4, -7 and -8. Paralogues of SCW CESAs in eudicot formed a sub-clad under monocot SCW CESAs with a single gene from *P. abies* in each clad. Three genes from tracheophyte formed separate clad under which SCW, PCW–CESA6 and embryophyte CESAs formed separate monophyletic groups. Three pinophyte CESAs grouped with CESA3 clad while no pinophyte CESA formed group with CESA1. Remaining pinophyte, embryophyte and tracheophyte genes formed a separate clad and didn't group with any of the CESAs (Fig. 2a). The eudicot ML tree excluding all ancestral as well as monocot CESAs is shown in Fig. 2b. The tree clearly differentiated PCW and SCW monophyletic groups and greatly supported previous tree structures. Within each CESA clad the legume CESAs, i.e. soybean, common bean, red clover, and barrel clover, formed sub-clads. Brassica CESAs formed discrete clusters with Arabidopsis CESAs. Among PCW CESAs, CESA6 was the largest clad, i.e., CESA6 sub-family which further formed three subgroups suggesting extensive diversification (Fig. 2b).

## Gene Structure and Splice-Site Analyses

Availability of whole-genome sequences provides an aid in understating gene structure and function. To characterize the structural diversity of CESA genes in eudicots, an exon–intron organization analysis was performed for each clad. The number of introns and exons, their boundaries, and intron phases are presented in Table S2. All genes analyzed in our study contained at least one intron, with 24 introns being the highest. Almost all of the CESA members we examined included three intron phases (0, 1, 2), with a few exceptions (Table S2). Lengths of the CESA genes observed in this study ranged from 508 to 1341 amino acids, and the number of exons ranged from 2 to 25. The majority of the CESA genes in the eudicots included in this study contained 13 exons and the longest gene was Glyma05g29240.1, at 1,337 nt and 25 exons. The CESA genes clustered within each clad shared exons and introns of the same length. Eudicot CESAs containing exons of the same nucleotide lengths were observed in all of the phylogenetic groups. In addition, we found a general trend of the ending exon being relatively longer than other exons.

## Gene Duplication and Synteny Analyses

A total of 73 duplicated genes were found in studied eudicots (Table S3). We further investigated these duplicated eudicot CESAs to find out clade specific conserved exons (Fig. 3). Exons conserved at rates of > 90%, > 80% and < 80% are colored red, honey, and light marigold. Duplicated genes had three (196 nt, 267 nt, and 346 nt), six (126 nt, 138 nt, 213 nt, 203 nt, 346 nt, and 351 nt), six (126 nt, 138 nt, 213 nt, 351 nt, 519 nt, and 613 nt), one (138 nt), one (196 nt), and two (213 nt and 346 nt)

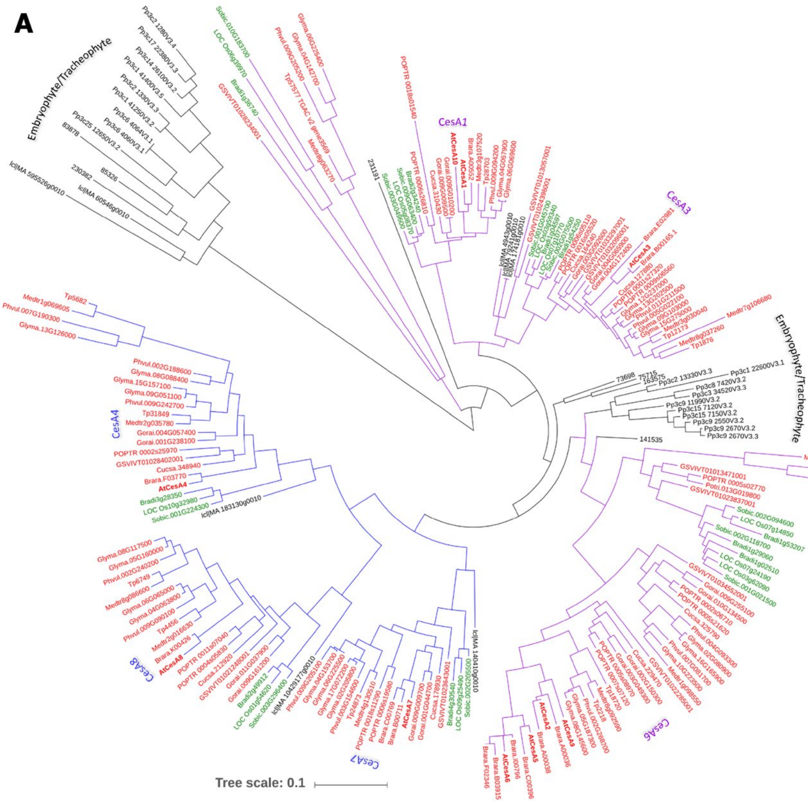

**Fig. 2** Phylogenetic construction of the CESA gene family in eudicots. **a** ML tree of CESA gene family. The evolutionary history was inferred using the Maximum Likelihood method based on the JTT matrix-based model. The tree with the highest log likelihood (38,007.95) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 206 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 523 positions in the final dataset. Evolutionary analyses were conducted in MEGA7. **b** ML tree of eudicot CESAs (Color figure online)

in clades I–VI, respectively. The most conserved exons among the duplicated genes in clades I–VI were 126 nt, 138 nt and 346 nt in length.

Segmental duplication was prevalent in the eudicots with the exception of two TD gene pairs belonging to *G. raimondii* (Gorai.004G065900.3/Gorai.004G172400.5) and *P. trichocarpa* (POPTR_0005s08970.3/POPTR_0005s21620.1) in clades II and IV, respectively, an indication that SD events have likely played significant roles in the expansion of the CESA gene family throughout the eudicots. Single-nucleotide reductions were observed in the exons of many of the duplicated gene pairs (Table S3). When duplications

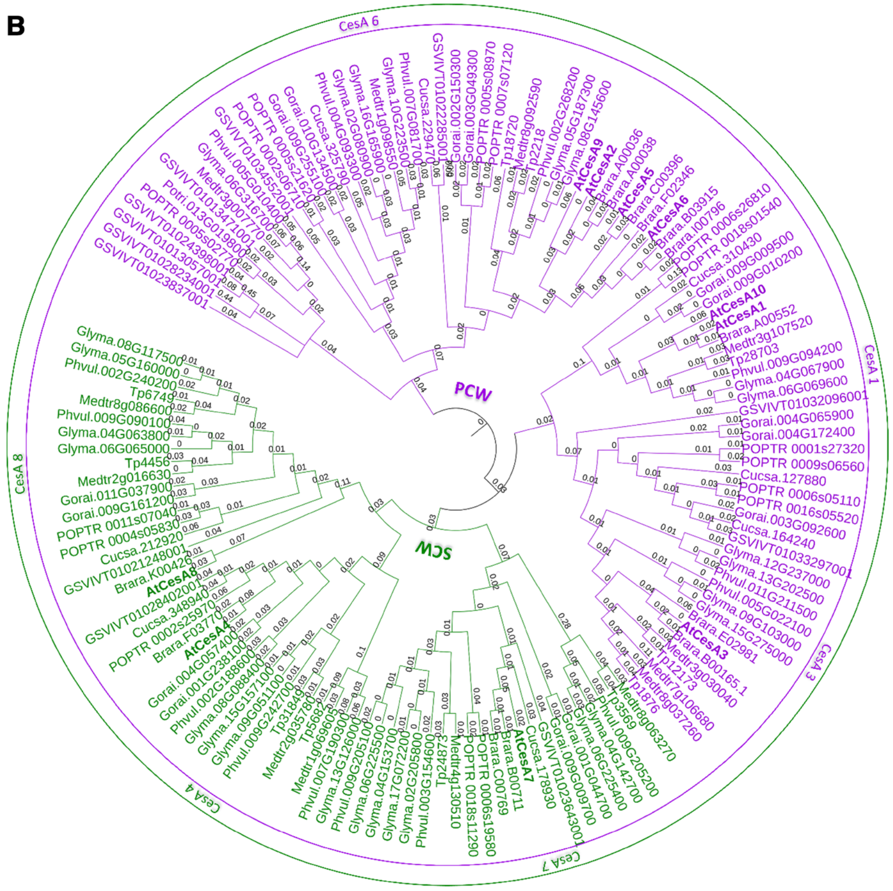

Fig. 2 (continued)

were considered clade-wise an interesting feature was observed that each duplicated gene in CESA6AI had a homologue in CESA6AII suggesting the expansion and diversification is sported by duplication (Fig. 1). Similarly, within each CESA clade (PCW as well as SCW), duplication played roles in expansion of genes number.

As such, we generated a comparative synteny relationship map. The colored connecting ribbons indicate relative syntenic relationships based on bit score (Fig. 4). Analysis of the syntenic relationships among and within the clades provided additional evidence of a high degree of CESA similarity within phylogenetic groups, suggesting the existence of common syntenic blocks among eudicots. All *B. rapa* CESA genes had maximal synteny with Arabidopsis CESAs genes, whereas soybean CESA genes displayed higher syntenic relationships with common bean across all clades (Fig. 4). Synteny analysis also identified two TD gene pairs in *G. raimondii* and *P. trichocarpa*.

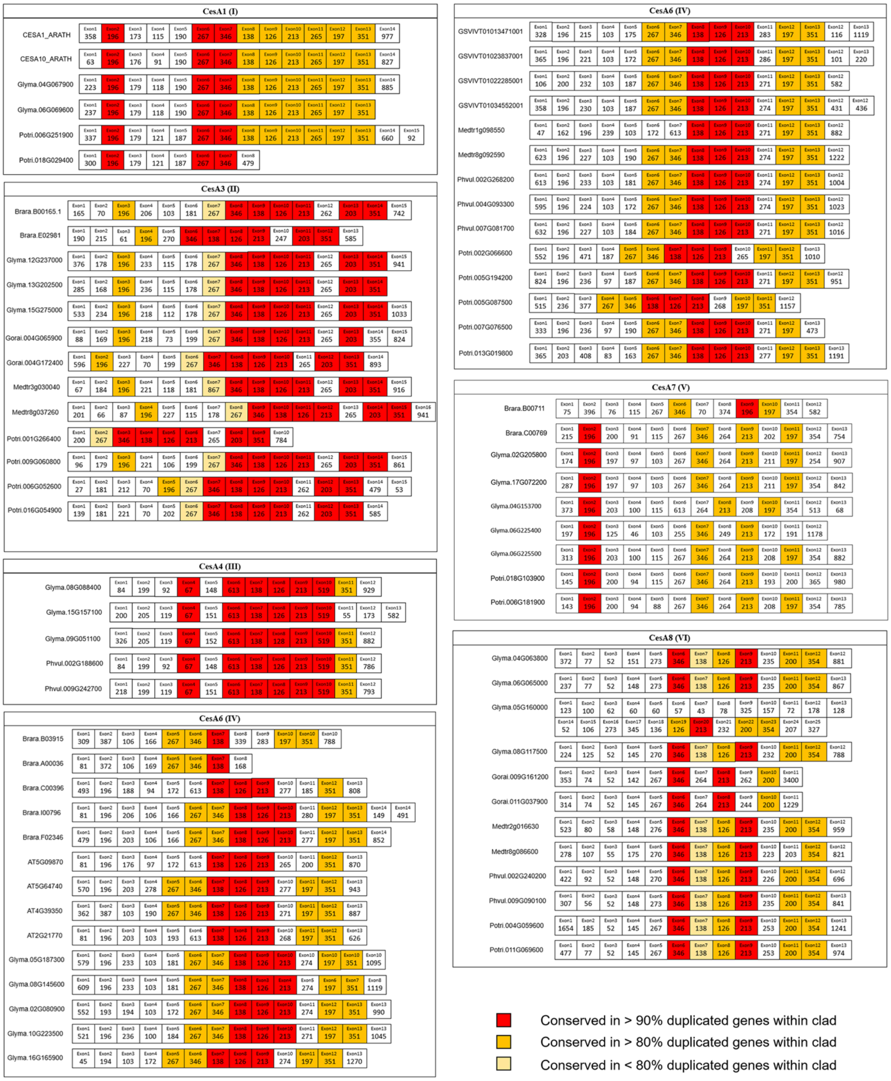

**Fig. 3** Exon structures (5′ to 3′) of each tandemly or segmentally duplicated CESA gene pair in all phylogenetic groups. Numbers in boxes are nucleotide lengths. Exon sizes are not drawn to scale (Color figure online)

## In Silico SSR Marker Identification

To aid studies related to PCW and SCW and for selection of higher cellulose content containing lines/parents, we analyzed 140 eudicot CESAs for the presence of SSR markers. We found 230 SSR distributed on 103 eudicot CESAs. Trinucleotide repeats were the most prevalent while hexanucleotide repeats were the least prevalent in analyzed sequences (Table 2 and Table S4).

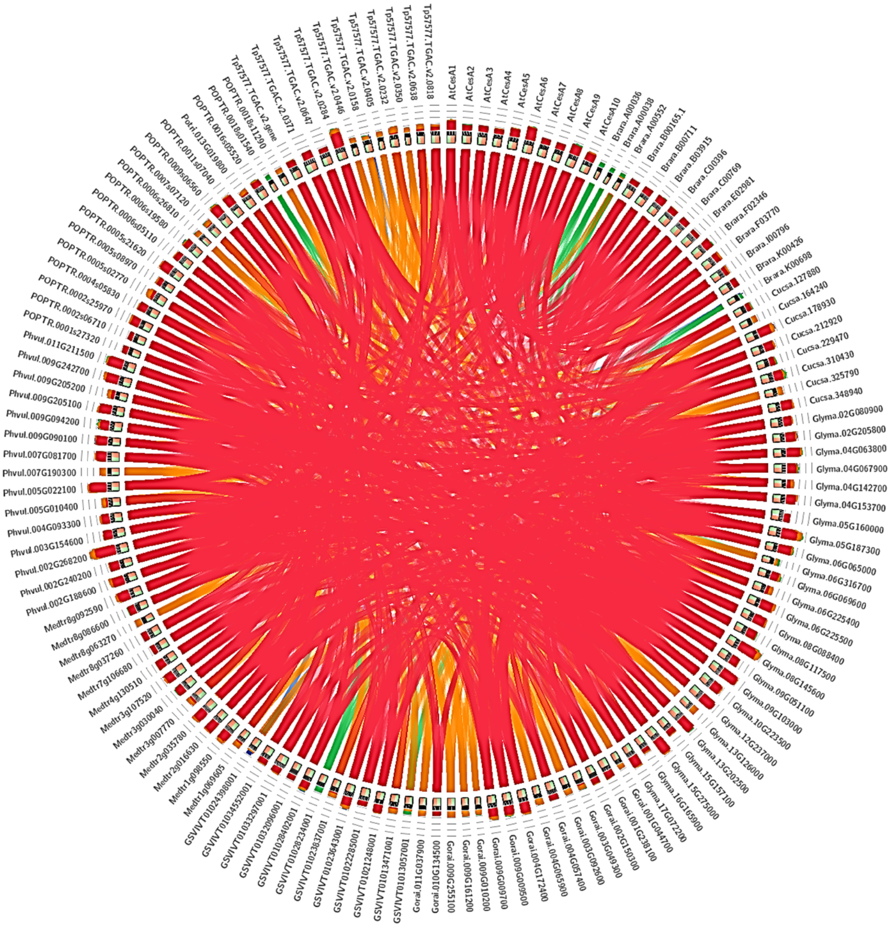

**Fig. 4** Synteny analysis of eudicot CESAs. Figure shows synteny within members of cellulose synthase A gene family members of eudicots. Inside the circle, ribbons represent local alignments based on bit score, red (> 80%), orange (> 60%), green (> 40%) and blue (> 20%). Ribbon width is correlated with % identity. Ribbons representing best hits are outlined and placed on top of all other ribbons. Histogram on the top of the ideograms, shows how many times each color has hit the specific part of the sequence (Color figure online)

## Eudicot CESA Co-expression Networks

Figure 5 is showing the significantly enriched/depleted phylostrata within CESA modules and expansion of CESAs through duplication. The PCW and SCW modules duplicated with the emergence of angiosperms. The CESA gene family members (both PCW and SCW) further expanded within angiosperms as well as in eudicots based on SDs and TDs (Fig. 5). To explore more about the evolution of CESAs in eudicots, we constructed the co-expression networks of four eudicots included in this research. We found that all four co-expression networks have duplication

**Table 2** Summary of identified SSR markers in eudicot CESAs

| Parameters | Details |
|---|---|
| Total no. of sequences examined | 141 |
| Total no. of identified SSRs | 230 |
| No. of SSR containing sequences | 103 |
| No. of sequences containing more than 1 SSR | 72 |
| *Distribution of different repeat type classes* | |
| No. of gene | 103 |
| No. of Markers | 231 |
| Dinucleotide repeats | 58 |
| Trinucleotide repeats | 79 |
| Tetranucleotide repeats | 65 |
| Pentanucleotide repeats | 18 |
| Hexanucleotide repeats | 11 |
| No. of sequences with successful primer pairs | 103 |
| Total primer pairs picked | 203.5 |

modules between PCW and SCW as shown by red edges. The co-expression networks of Arabidopsis showed that Cu-oxidase_2, Cu-oxidase_3, Fasciclin, DUF231, COBRA, glycol_hydro_19, and cytochrome_B561/DUF568 genes co-expressed with CESAs (Fig. 5a). The co-expression networks of barrel clover showed that peptidase_S10/retrotrans_gag was transcriptionally coordinated with CESAs (Fig. 5b). Populous CESAs co-expressed with glycol_hydro_19 and glycol_transf_9 whereas COBRA, Fasciclin, 3 Beta_HSD, TPT, and glycol_hydro_19 genes family members showed transcriptional coordination with soybean CESAs (Fig. 5c, d). Annotation and gene ontology of the terms found in co-expression networks are given in Table S5. Molecular function of co-expressed genes was related to membrane activity and protein binding while the genes were found to be involved in different biological processes such as carbohydrate metabolic processes, cell growth, and growth.

## In Silico Characterization of Eudicot CESAs During Leaf Senescence

Since CESAs were showed involved in cell growth and growth by the co-expression analysis, we looked into how the changes of CESAs expression levels could relate to leaf senescence. In leaf tissues of *A. thaliana*, expression levels of CESA genes decreased during senescence, with *AT4G32410*, *AT5G05170* and *AT5G64740* showing higher expression levels. (Fig. 6a) *AT4G32410* and *AT5G05170* decreased more rapidly, in leaves beyond approximately 28 or 29 days, while *AT4G39350*, *AT5G09870* and *AT5G64740* decreased gradually during senescence (Fig. 6b). In leaf tissues of *T. pratense*, CESA genes were downregulated in the old ones compared to the mature ones (Fig. 6c). In *G. max*, *Glyma.02G08900*, *Glyma.06G065000*, *Glyma.09G051100* and *Glyma.16G165900* were specifically expressed in leaf tissues, not in cotyledons (Fig. 6D). In general, the expression levels of CESA genes decreased during leaf senescence (Fig. 6d, e).

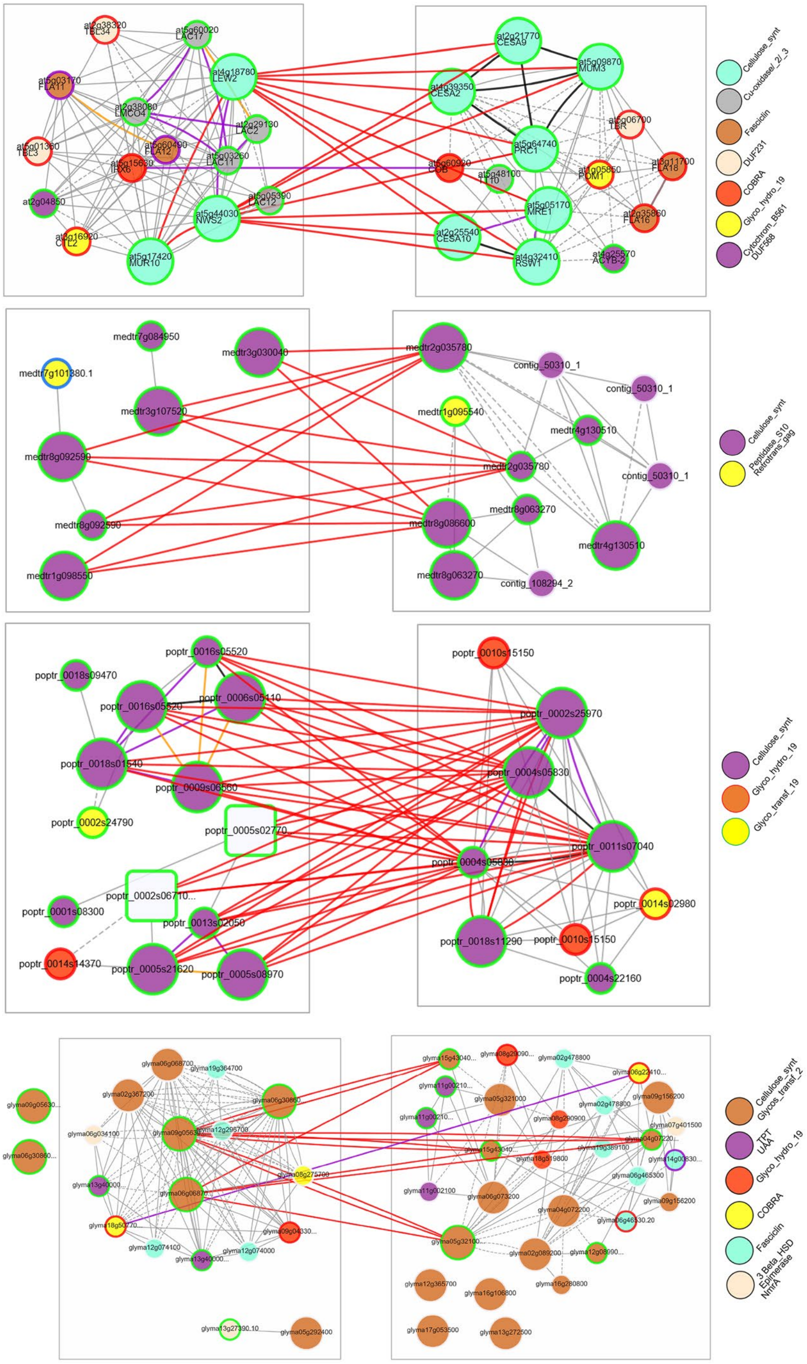

**Fig. 5** Co-expression networks of PCW and SCW CESA modules in **a***Arabidopsis*, **b** Barrel clover, **c***Populous*, and **d** Soybean (Color figure online)

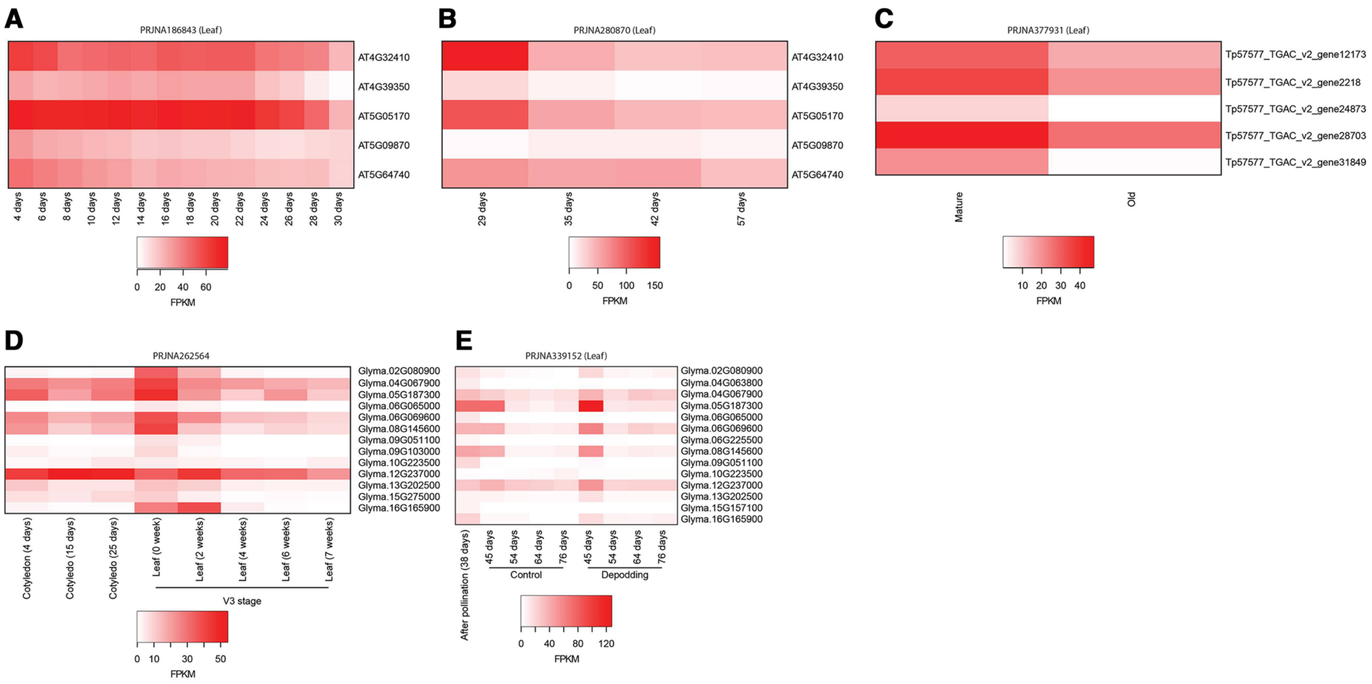

**Fig. 6** Heatmaps of CESA genes expression levels in public data sets. Only genes with maximal expression levels larger than 5 FPKM across all time points are shown. **a** Leaf tissues of *A. thaliana* during senescence (PRJNA186843); **b** leaf tissues of *A. thaliana* during senescence (PRJNA280870); **c** Old and mature leaves of *T. pratense* (PRJNA377931); **d** cotyledons and leaf tissues of *G. max* during senescence (PRJNA262564); **e** control and depodded leaf tissues of *G. max* after pollination during senescence (PRJNA339152) (Color figure online)

## Discussion

Involvement of CW in plant-cell structural integrity, growth, shape and cellular communication processes has made this an important topic of study. Biogenesis of plant CW is under the synergistic control of multiple gene families that regulate the entire process, from biosynthesis to controlled degradation (Rai et al. 2016). Among these families, the cellulose synthase-like gene superfamily is the most commonly studied in land plants, having been described in Arabidopsis, grapevine, sorghum, soybean, and black cottonwood, among others (Suzuki et al. 2006; Carroll and Specht 2011; Yin et al. 2014; Rai et al. 2016; Nawaz et al. 2017a). The continual accumulation of data regarding whole-genome sequences has created numerous opportunities for exploration of the CESA gene family in related plants; for instance, the whole-genome sequences of major eudicots have recently been made available (Table 1). Eudicots are the predominant group of angiosperms and are essential in the global production of food, feed, fiber, fuel, paper and structural materials (Worberg et al. 2007), and thus an extensive and comprehensive survey of the CESA gene family in eudicots is required. As a first step, the phylogenic relationships of the CESA gene family members aided with co-expression networks would enhance our understanding of the evolutionary history of this gene family in eudicots, which may assist in the creation of plants that feature enhanced CW composition for use in crop and industrial applications.

In this study, we focused on CESAs responsible for the biosynthesis of cellulose in eudicots. To obtain a broad perspective of the putative CESAs in eudicots, we selected 10 eudicot species to explore the evolutionary history of the CESA gene family within this group of plants. For comparison, we also analyzed three ancestral plants, including an embryophyte, a tracheophyte, and a pinophyte, along with the 3 monocots. Our results demonstrated that the basic forms of CESA proteins have been retained throughout the angiosperms, which accords with Ruprecht et al. (2017). The number of CESAs in all explored eudicots presented at least six protein types. The eudicot CESA proteins are similar in number to those of the ancestral representative *P. patens*, which reflects the conservation of this gene family during the course of embryophyte to angiosperm evolution. Specialization of the CESA gene family must have occurred during the evolution of lycophtes and gymnosperms (Carroll and Specht 2011; Ruprecht et al. 2017). Ruprecht et al. (2017) reported an independent duplication of cellulose biosynthesis pathways in angiosperms and bryophytes.

The 140 eudicot CESA genes identified here all contained a CESA-RING-type zinc finger domain and/or a nucleotide-diphosphate-sugar-transferase domain (Suzuki et al. 2006; Kumar et al. 2009, 2016). The nucleotide-diphosphate-sugar-transferase domain is responsible for the release of UDP from UDP-glucose and the addition of glucose to cellulose peptides (Carpita 2011; Maleki et al. 2016). Apart from these two conserved domains, other peptide sequences may also be useful for further identification of orthologs common to related species. Sequence similarity is only one of several means of identifying potential orthologs, however, and oftentimes it proves to be insufficient by itself (Kaur et al. 2016). CESA genes are known to contain three aspartyl residues and a QXXRW motif, which was conserved in all eudicot CESAs we examined; in addition, the motifs CXXC and SVICEXWFA are conserved throughout the eudicots. Previous studies on CESA orthologs were truly structured on sequence similarity (Suzuki et al. 2006; Carroll and Specht 2011). Several studies focusing on unraveling evolutionary histories and genome-wide analysis of major gene families have clearly demonstrated that certain domains and motifs are conserved over the course of evolution (Le et al. 2016; Rehman et al. 2016, 2017); here, we identified 13 motifs among the CESA orthologs that have been highly conserved within this gene family (Fig. 1). Despite the variable protein sequences of each member of the CESA family among the orthologs deriving from various species, motif organization remained more or less constant, suggesting the possibility of their involvement in vital cellular functions.

Following identification of the full membership of a gene family, conserved amino acid sequences can be used to assess the evolutionary relationships within a species (homologs) and among related species (orthologs and paralogs) (Kumar et al. 2009). Comparisons of whole-genome-sequence surveys can then be conducted to identify potential orthologs of gene families conserved among closely related species, such as among the eudicots (Caputi et al. 2012). Furthermore, cross-species comparative functional genomes based on knowledge of parallels in related species can yield predictions of functional properties. In our study, several strong associations became apparent when the 206 aligned CESA protein sequences identified from the 10 eudicots, 3 monocots, and the three ancestral species were used to construct a bootstrapped phylogenetic ML tree (Fig. 2a). The eudicot and monocot CESAs formed groups of clearly defined orthologs and paralogs, with a separate clad containing embryophyte and tracheophyte genes. All of the six clades have been previously reported from genome-wide CESA surveys of Arabidopsis, Populus, soybean and grapevine (Suzuki et al. 2006; Carroll and Specht 2011; Nawaz et al. 2017a). The grouping into six clades suggests that early evolutionary duplications must have given rise to these essential clades (Kumar et al. 2009); for instance, CESAs from pine were clustered within clades CESA3 (II), CESA4 (III), CESA7 (V), and CESA8 (VI), strongly suggesting that this gene family has been conserved over the course of land plant evolutionary history. This grouping was different from a recent report showing that Norway spruce only contains cell wall module rather than containing separate modules for PCW and SCW (Proost and Mutwil 2016). Previously, it is known that CESAs from bryophytes do not contain SCW specific CESAs (Roberts and Roberts 2004; Ruprecht et al. 2017). Similar case was observed in our studies. Our ML tree showed that in this specie three unique components of SCW might be present. As far as PCW is concerned three genes grouped with grape wine CESA3 paralogues. The tree structure greatly supports that the PCW and SCW contains three unique components each (Turner and Somerville 1997; Taylor et al. 2004; Persson et al. 2007) (Fig. 2). Cellulose synthase complex responsible for the formation of PCW have two unique components i.e. CESA1 and CESA3, and a small family of CESA6 related proteins provides the third essential component of PCW. Among the members of subfamily of CESA6 genes which have high sequence similarity CESA6, only CESA2 has been described until now that it is functionally redundant with CESA6. The functional redundancy between the CESA6 sub family was recently reported to be partial and one CESA may substitute for another, but not completely, that concludes that there are functional differences between CESA6 family members (Ruprecht et al. 2017). To gain insight into the evolution of cell walls in eudicots, we studied gene co-expression networks of PCW and SCW in four eudicots, i.e. Arabidopsis, soybean, common bean, and popular. Phylogenetic analysis indicated that PCW and SCW modules and duplicated in all four species clearly indicating that some parts of these modules were duplication in this lineage (Fig. 5).

Analysis of the genomes of the eudicot species included in our study revealed that, collectively, copies of CESA genes related to PCW (i.e., in clades 1 [I], 3 [II], and 6 [IV]) processes were more numerous (82) than were copies of CseA genes (58) related to SCW (i.e., in clades 4 [II], 7 [V], and 8 [VI]), a pattern similar to that observed previously in wheat, Arabidopsis, grapevine, *Populus*, soybean and sorghum (Holland et al. 2000; Somerville 2006; Suzuki et al. 2006; Carroll and Specht 2011; Rai et al. 2016; Nawaz et al. 2017a). The CESA6 clade failed to cluster with any pine-derived CESA genes and contained the highest number of CESA copies, suggesting that this position has experienced a greater amount of divergence and diversification than have other CESA positions (Ruprecht et al. 2016, 2017). Further subgrouping of this clad provided further evidence that the divergence of CESA6 is probably due to one or more duplication events that occurred after the evolution of the eudicots; Carroll and Specht (2011) proposed that the extensive diversification of the CESA6 clad signifies that it plays only an indirect or non-essential role in CESA complex formation and/or cellulose biosynthesis.

Interestingly, all *B. rapa* CESAs (clades I–VI) clustered specifically with the CESAs deriving from Arabidopsis and formed discrete clades. This is further confirmation that both of these species share a most recent common ancestor (MRCA). Homologous to 10 Arabidopsis CESAs, 13 *B. rapa* CESAs formed discrete clusters. CESA1, CESA4, and CESA8 had single homologs, whereas in *B. rapa,* CESA3, CESA7, and CESA6 had two, two, and six homologs, respectively. This may be due to the *B. rapa* genome undergoing considerable fractionation in the time since divergence of both *B. rapa* and Arabidopsis from their MRCA and more recent genomic triplication events (Schranz et al. 2006; Wang et al. 2011). All CESAs of the four legumes included in our study (soybean, common bean, barrel clover, and red clover) also formed discrete clades with one another and clustered relatively closer to Arabidopsis CESAs than did the other species (Fig. 2b). Moreover, the high number of soybean CESAs may be because of two distinct genome duplication events. Phylogenetic classification of the 140 CESAs provided a basis for identifying clad-specific conserved motifs, which could prove useful in the demarcation of class specificity (Table S1).

Given that gene structure, boundary determination, and location of splice sites determine gene regulation, more effective demarcation of exon boundaries would aid in the identification of conserved exons, which in turn would improve our understanding of splice-site conservation over the course of evolutionary history (Hu et al. 2014; Wang et al. 2016). Conserved exon structures containing the same number of nucleotides coupled with conserved intron phases represent genetic similarities that provide important clues for understanding evolutionary history (Betts et al. 2001;

Wang et al. 2012b). The average number of exons observed in this study was 13, and clade-wise gene structure analysis revealed that there is a maximum of three intron phases in eudicot CESA members (Table S2 and Fig. 3). This variation in intron phases suggests that the codons of this gene family are not intact in eudicots. In soybean, the presence of the longest gene encountered in this study (*Glyma05g29240*) may be due to two rounds of whole-genome duplication (Schmutz et al. 2010). Exon length was conserved throughout the individual clades, with few exceptions, and therefore variability in the number of exons may be due to deletions or duplications. The presence of exons of similar lengths provides strong evidence that the CESA gene family is conserved in eudicots, and has experienced little or no divergence, and that the splice sites remained conserved as well is an additional indication of functional conservation (Sheth et al. 2006).

Investigating gene duplication aids in determining whether a gene family expanded or remained conserved during evolution (Schmutz et al. 2010; Wang et al. 2011). Gene families expand via three basic mechanisms: TD, SD, and whole-genome duplication (Worberg et al. 2007; Lee et al. 2012). Given that eudicots descended from a single ancestor, large-scale chromosomal rearrangement(s) may have occurred over the course of their evolution; thus, we searched for gene duplication events as a means of gaining further insight into the pattern of expansion of the CESA gene family in eudicots. Gene duplication analysis based on phylogenetic clustering revealed the occurrence of SD and TD events (Fig. 3 and Table S2), with SD the more prevalent duplication in eudicots; this suggests that SD played an important role in the expansion of the CESA gene family. Arabidopsis, new-world cotton, and Populus each had a single pair of TD genes. Such duplication could potentially lead toward functional diversification. The duplicated gene pairs of Arabidopsis validate the previous clustering of CESA1 and CESA10 as one single type of protein, and CES6 and its three homologs (i.e. CESA2, CESA5, and CESA9) as another single type of protein. Similar clustering patterns of duplicated genes were observed in all of the examined eudicots, with the exception of red clover, which contained no duplicated gene pairs. Such clustering demonstrates that the CESA gene family has been highly conserved in the eudicots, and that the duplicated gene pairs do not necessarily represent the development of new protein types or expanded functional diversity. Another way to demonstrate the multitude of plant gene duplication is to consider the paralogs within gene families. Gene-sequence comparisons among related plant genomes and within individual genomes provides sufficient evolutionary evidence to verify these relationships, and such information facilitates rapid examination of the genomes of closely related organisms (Lee et al. 2012; De Vega et al. 2015). Synteny analysis could provide evidence for functional connections between members of a gene family among related species (Panchy et al. 2016). In our study, the presence of synteny in all phylogenetic groups confirmed the evolutionary conservation of the CESA gene family in eudicots, suggesting that CESA-derived products have been largely conserved in the eudicots and that they evolved from a common ancestor (Fig. 4). In order to accelerate genetic mapping, identification of molecular markers such as SSRs is effective (Nawaz et al. 2017c). Identification of SSRs greatly supports investigations related to genetic variations, screening and identification of mutants. SSR markers are distributed throughout the eukaryotic genomes and are considered to display taxon-specific variations in relation to motif structure, genomic location and frequency (Rai et al. 2016; Nawaz et al. 2017b). These markers are efficient in genetic mapping of traits specific loci as well as to detect genetic variation or diversity (Nawaz et al. 2013; Nadeem et al. 2018). We found 230 SSR markers located on 103 eudicot CESAs, with trinucleotide repeats (34.34%) being the most abundant and hexanucleotide repeats (0.04%) being the least abundant (Table 2).

Senescence is an integrated response of plants to multiple signals either endogenous or from external environment. It is an age-dependent programmed degradation and degeneration process of plant cells (tissues, organs or the entire organism) followed by death (Woo et al. 2013). The impact of senescence on lignocellulosic material is consistent with substantial macromolecule degradation. The reduced expression of Arabidopsis CESAs during senescence could be the result of multiple signals ending up with reduced cellulose synthesis. A gradual reduction in expression of AtCESAs (Fig. 6a, b) signifies reduction of biomass during senescence observed in our in silico RNA seq analysis is consistent with previous findings observed in Mandarin fruit where a reduction in and disassembly of the cellulose-hemicellulose network was noted (Li et al. 2016). As previously known that cell walls of the plants undergo precise metabolic changes during senescence-related processes and cellulose, hemicellulose and pectin polysaccharides are targeted during programmed cell death/senescence (Ghosh et al. 2013). Similarly, the downregulation of *T. pretense* CESAs in older leaves as compared to mature leaves suggest cellulose disassembly during leaf senescence (Fig. 6c). The comparative expression level of soybean CESAs in leaves and cotyledons revealed significant differences. The little or no expression of CESAs in soybean cotyledons as compared to relatively higher expression in leaves suggest that leaves respond differently to senescence as compared to cotyledons (Fig. 6d, e). Based on the available datasets we conclude that expression of CESAs is reduced in eudicot during leaf senescence.

**Acknowledgements** This work was carried out with the support of ‘‘Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ01133402)’’ Rural Development Administration, Republic of Korea.

## Compliance with Ethical Standards

**Conflict of interest** We claim that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

Appenzeller L et al (2004) Cellulose synthesis in maize: isolation and expression analysis of the cellulose synthase (CesA) gene family. Cellulose 11:287–299

Artimo P (2012) ExPASy: SIB bioinformatics resource portal. Nucl Acids Res 40:W597–W603

Barvkar VT, Pardeshi VC, Kale SM, Kadoo NY, Gupta VS (2012) Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in *Linum usitatissimum* identified genes with varied expression patterns. BMC Genom 13:175

Betts MJ, Guigó R, Agarwal P, Russell RB (2001) Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J 20:5354–5360

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bolser DM, Staines DM, Perry E, Kersey PJ (2017) Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomic data. In: Plant genomics databases. Springer, Berlin, pp 1–31

Brown AV, Hudson KA (2015) Developmental profiling of gene expression in soybean trifoliate leaves and cotyledons. BMC Plant Biol 15:169

Brown AV, Hudson KA (2017) Transcriptional profiling of mechanically and genetically sink-limited soybeans. Plant Cell Environ 40:2307–2318

Brusslan JA, Bonora G, Rus-Canterbury AM, Tariq F, Jaroszewicz A, Pellegrini M (2015) A genome-wide chronological study of gene expression and two histone modifications, H3K4me3 and H3K9ac, during developmental leaf senescence. Plant Physiol 168:1246–1261

Burton RA, Shirley NJ, King BJ, Harvey AJ, Fincher GB (2004) The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol 134:224–236

Caputi L, Malnoy M, Goremykin V, Nikiforova S, Martens S (2012) A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. Plant J 69:1030–1042

Carpita NC (2011) Update on mechanisms of plant cell wall biosynthesis: how plants make cellulose and other (1→4)-β-D-glycans. Plant Physiol 155:171–184

Carroll A, Specht CD (2011) Understanding plant cellulose synthases through a comprehensive investigation of the cellulose synthase family sequences. Front Plant Sci 2:5

Darzentas N (2010) Circoletto: visualizing sequence similarity with Circos. Bioinformatics 26:2620–2621

De Vega JJ et al (2015) Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Sci Rep 5:17394

Doblin MS, Kurek I, Jacob-Wilk D, Delmer DP (2002) Cellulose biosynthesis in plants: from genes to rosettes. Plant Cell Physiol 43:1407–1420

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Ghosh S et al (2013) Induction of senescence and identification of differentially expressed genes in tomato in response to monoterpene. PLoS One 8:e76029

Goodstein DM et al (2011) Phytozome: a comparative platform for green plant genomics. Nucl Acids Res 40:D1178–D1186

He C, Cui K, Duan A, Zeng Y, Zhang J (2012) Genome-wide and molecular evolution analysis of the Poplar KT/HAK/KUP potassium transporter gene family. Ecol Evol 2:1996–2004

Holland N, Holland D, Helentjaris T, Dhugga KS, Xoconostle-Cazares B, Delmer DP (2000) A comparative analysis of the plant cellulose synthase (CesA) gene family. Plant Physiol 123:1313–1324

Huang S et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1282

Hu B, Jin J, Guo A-Y, Zhang H, Luo J, Gao G ((2014)) GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31:1296–1297

Jaillon O et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449(7161):463–467

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275–282

Kaur S, Dhugga KS, Gill K, Singh J (2016) Novel structural and functional motifs in cellulose synthase (CesA) genes of bread wheat (*Triticum aestivum*, L.). PLoS One 11:e0147046

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357

Kumar M et al (2009) An update on the nomenclature for the cellulose synthase genes in Populus. Trends Plant Sci 14:248–254

Kumar M, Turner S (2015) Plant cellulose synthesis: CESA proteins crossing kingdoms. Phytochemistry 112:91–99

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874

Le B et al (2016) Genome-wide characterization and expression pattern of auxin response factor (ARF) gene family in soybean and common bean. Genes Genom 38:1165–1178

Lee T-H, Tang H, Wang X, Paterson AH (2012) PGDD: a database of gene and genome duplication in plants. Nucl Acids Res 41:D1152–D1158

Letunic I, Bork P (2006) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128

Li T, Zhang J, Zhu H, Qu H, You S, Duan X, Jiang Y (2016) Proteomic analysis of differentially expressed proteins involved in peel senescence in harvested mandarin fruit. Front Plant Sci 7:725

Little A et al (2018) Revised phylogeny of the Cellulose Synthase gene superfamily: insights into cell wall evolution. Plant Physiol 177:1124–1141

Maleki SS, Mohammadi K, Ji KS (2016) Characterization of cellulose synthesis in plant cells. Sci World J 2016:8641373

Mitchell A et al (2014) The InterPro protein families database: the classification resource after 15 years. Nucl Acids Res 43:D213–D221

Mutwil M et al (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell 23:895–910

Nadeem MA et al (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32:261–285

Nawaz MA, Sadia B, Awan FS, Zia MA, Khan IA (2013) Genetic diversity in hyper glucose oxidase producing *Aspergillus niger* UAF mutants by using molecular markers. Int J Agri Biol 15:362–366

Nawaz MA et al (2017a) Genome and transcriptome-wide analyses of cellulose synthase gene superfamily in soybean. J Plant Physiol 215:163–175

Nawaz MA et al (2017b) Systems identification and characterization of cell wall reassembly and degradation related genes in *Glycine max* (L.) Merill, a Bioenergy Legume. Sci Rep 7:10862

Nawaz MA, Yang SH, Rehman HM, Baloch FS, Lee JD, Park JH, Chung G (2017c) Genetic diversity and population structure of Korean wild soybean (*Glycine soja* Sieb. and Zucc.) inferred from microsatellite markers. Biochem Syst Ecol 71:87–96

Olek AT et al (2014) The structure of the catalytic domain of a plant cellulose synthase and its assembly into dimers. Plant Cell 26:2996–3009

Panchy N, Lehti-Shiu MD, Shiu S-H (2016) Gene Duplicates: from origins to implications for plant evolution. Plant Physiol 171:2294–2316

Persson S et al (2007) Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. Proc Natl Acad Sci 104:15566–15571

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290

Proost S, Mutwil M (2016) Tools of the trade: studying molecular networks in plants. Curr Opin Plant Biol 30:143–150

Rai KM, Thu SW, Balasubramanian VK, Cobos CJ, Disasa T, Mendu V (2016) Identification, characterization, and expression analysis of cell wall-related genes in *Sorghum Bicolor* (L.) Moench, a food, fodder, and biofuel crop. Front Plant Sci 7:1287

Rehman HM et al (2016) Genome-wide analysis of Family-1 UDP-glycosyltransferases in soybean confirms their abundance and varied expression during seed development. J Plant Physiol 206:87–97

Rehman HM, Nawaz MA, Shah ZH, Daur I, Khatoon S, Yang SH, Chung G (2017) In-depth genomic and transcriptomic analysis of five K+ transporter gene families in soybean confirm their differential expression for nodulation. Front Plant Sci 8:804

Richmond TA, Somerville CR (2000) The cellulose synthase superfamily. Plant Physiol 124:495–498

Roberts AW, Roberts E (2004) Cellulose synthase (CesA) genes in algae and seedless plants. Cellulose 11:419–435

Römling U, Galperin MY (2015) Bacterial cellulose biosynthesis: diversity of operons, subunits, products, and functions. Trends Microbiol 23:545–557

Ruprecht C et al (2016) FamNet: A framework to identify multiplied modules driving pathway diversification in plants. Plant Physiol 170:1878–1894

Ruprecht C et al (2017) Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. Plant J 90:447–465

Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178

Schmutz J et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707

Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci 11:535–542

Schwerdt JG et al (2015) Evolutionary dynamics of the cellulose synthase gene superfamily in grasses. Plant Physiol 168:968–983

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R (2006) Comprehensive splice-site analysis using comparative genomics. Nucl Acids Res 34:3955–3967

Sievers F et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Somerville C (2006) Cellulose synthesis in higher plants. Annu Rev Cell Dev Biol 22:53–78

Suzuki S, Li L, Sun Y-H, Chiang VL (2006) The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. Plant Physiol 142:1233–1245

Taylor NG, Gardiner JC, Whiteman R, Turner SR (2004) Cellulose synthesis in the Arabidopsis secondary cell wall. Cellulose 11:329–338

The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Turner SR, Somerville CR (1997) Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. Plant Cell 9:689–701

Tuskan GA et al (2006) The genome of black cottonwood *Populus trichocarpa* (Torr. & Gray). Science 313(5793):1596–1604

Wang X et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035

Wang K et al (2012a) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44:1098–1103

Wang Y et al (2012b) PIECE: a database for plant gene structure comparison and evolution. Nucl Acids Res 41:D1159–D1166

Wang N, Xia E-H, Gao L-Z (2016) Genome-wide analysis of WRKY family of transcription factors in common bean, *Phaseolus vulgaris*: chromosomal localization, structure, evolution and expression divergence. Plant Gene 5:22–30

Woo HR, Kim HJ, Nam HG, Lim PO (2013) Plant leaf senescence and death—regulation by multiple layers of control and implications for aging in general. J Cell Sci 126:4823–4833

Worberg A, Quandt D, Barniske A-M, Löhne C, Hilu KW, Borsch T (2007) Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. Org Divers Evol 7:55–77

Yin Y, Huang J, Xu Y (2009) The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biol 9:99

Yin Y, Johns MA, Cao H, Rupani M (2014) A survey of plant and algal genomes and transcriptomes reveals new insights into the evolution and function of the cellulose synthase superfamily. BMC Genom 15:260

You FM et al (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinform 9:253

Young ND et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 480:520

## Affiliations

**Muhammad Amjad Nawaz[1] · Xiao Lin[2] · Ting-Fung Chan[2] · Muhammad Imtiaz[3] · Hafiz Mamoon Rehman[1] · Muhammad Amjad Ali[4] · Faheem Shehzad Baloch[5] · Rana Muhammad Atif[6] · Seung Hwan Yang[1] · Gyuhwa Chung[1]**

[1] Department of Biotechnology, Chonnam National University, Chonnam 59626, Republic of Korea

[2] Center for Soybean Research, State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, SAR, China

[3] School of Environmental Science and Engineering, Guangzhou University, Guangzhou 510275, China

[4] Department of Plant Pathology, University of Agriculture, Faisalabad 38040, Pakistan

[5] Department of Field Crops, Faculty of Agricultural and Natural Science, Abant Izzet Baysal University, 14280 Bolu, Turkey

[6] US-Pakistan Centre for Advanced Studies in Agriculture and Food Security, University of Agriculture, Faisalabad 38040, Pakistan