The Cellulose Synthase Superfamily[1]

Todd A. Richmond* and Chris R. Somerville

Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, California 94305 (T.A.R., C.R.S.); and Department of Biological Sciences, Stanford University, Stanford, California 94305 (C.R.S.)

The availability of a nearly complete genome sequence for Arabidopsis has created many novel opportunities to identify, by computational methods, the genes that encode enzymes, which have been difficult to characterize by conventional means. We have used this approach to identify a large family of genes of unknown function that show sequence similarity to cellulose synthase. Our working hypothesis is that these genes encode enzymes that catalyze the synthesis of non-cellulosic polysaccharides (Cutler and Somerville, 1997).

A recent breakthrough in research concerning the biogenesis of plant cell walls was the identification, by genomic methods, of genes encoding cellulose synthase in cotton fibers (Pear et al., 1996; Delmer, 1999). The cotton cellulose synthase genes, now termed *CesA1* and *CesA2*, were identified in a collection of expressed sequence tag (EST) sequences on the basis of weak sequence similarity to genes for cellulose synthase from bacteria. In addition, the genes were expressed at high levels in cotton fibers at the onset of secondary wall synthesis and a purified fragment of one of the corresponding proteins was shown to bind UDP-Glc, the proposed substrate for cellulose biosynthesis. The conclusion that the cotton *CesA* genes are cellulose synthases is supported by results obtained with two cellulose-deficient Arabidopsis mutants, *rsw1* (Arioli et al., 1998) and *irx3* (Turner and Somerville, 1997; Taylor et al., 1999). The genes corresponding to the *RSW1* and *IRX3* loci exhibit a high degree of sequence similarity to the cotton *CesA* genes and are considered orthologs. Ten full-length *CesA* genes have been sequenced from Arabidopsis, and there is a genome survey sequence that may indicate one additional family member (Fig. 1).

It is not known at this time whether other polypeptides are also required for cellulose synthase activity (i.e. the CesA polypeptides may be a component of a multisubunit enzyme complex). Until this matter is resolved we consider it expedient to simply refer to the CesA family members as cellulose synthase. The observation that *IXR3* (AtCesA7), which is required for secondary wall cellulose synthesis, is in a different branch of the *CesA* tree than *RSW1* (AtCesA1), which is required for primary wall synthesis (Fig. 1), may indicate that there is sequence divergence between the enzymes involved in primary and secondary wall synthesis.

Reiterative database searches using the Arabidopsis Rsw1 (AtCesA1) and the cotton CesA polypeptide sequences as the initial query sequences revealed a large superfamily of at least 41 *CesA*-like genes in Arabidopsis. Based on predicted protein sequences, we have grouped these genes into seven clearly distinguishable families (Fig. 1): the *CesA* family, which includes *RSW1* and *IRX3* (*AtCesA7*), and six families of structurally related genes of unknown function designated as the "cellulose synthase-like" genes (*CslA*, *CslB*, *CslC*, *CslD*, *CslE*, and *CslG*). The nomenclature for these families is still under discussion (http://mbclserver.rutgers.edu/CPGN/CelluloseWeb/CesA.proposal.html), so the *Csl* designation for these genes should be considered temporary and may be revised as the enzymatic function of the members of each family is determined.

All of the members of the cellulose synthase superfamily appear to be integral membrane proteins, with three to six transmembrane domains in the carboxy terminal region of the protein and one or two transmembrane domains in the amino terminal region. It is thought that the CesA proteins are located in the plasma membrane (Delmer, 1999). If the Csl proteins participate in the synthesis of non-cellulosic polysaccharides, they would be expected to be located in the Golgi apparatus. Preliminary analysis of CslB, CslG, and CslE fusions to green fluorescent protein appear to localize to the Golgi (T. Richmond and C. Somerville, unpublished data). Also, immunolocalization studies with an antibody to the CslA protein indicates that this family is localized to the cytoplasm (i.e. the Golgi apparatus) rather than the plasma membrane (N. Sprenger and C. Somerville, unpublished data).

Intron-exon organization is conserved among the *CesA*, *CslB*, *CslG*, and *CslE* gene families, but not the *CslA*, *CslC*, or *CslD* families (Fig. 2). However, the C-terminus of a subset of the *CslD* genes is congruent with this organization as well. The *CslD* gene family is the most similar of the *Csl* gene families to the *CesA* family (approximately 45% identical at the amino acid level). The gene structure for this family is unusual in that the seven genes for which complete genomic sequence information is available have four

[1] This work was supported in part by the U.S. Department of Energy (grant no. DOE–FG02–00ER20133).

* Corresponding author; e-mail todd@andrew2.stanford.edu; fax 650–325–6857.

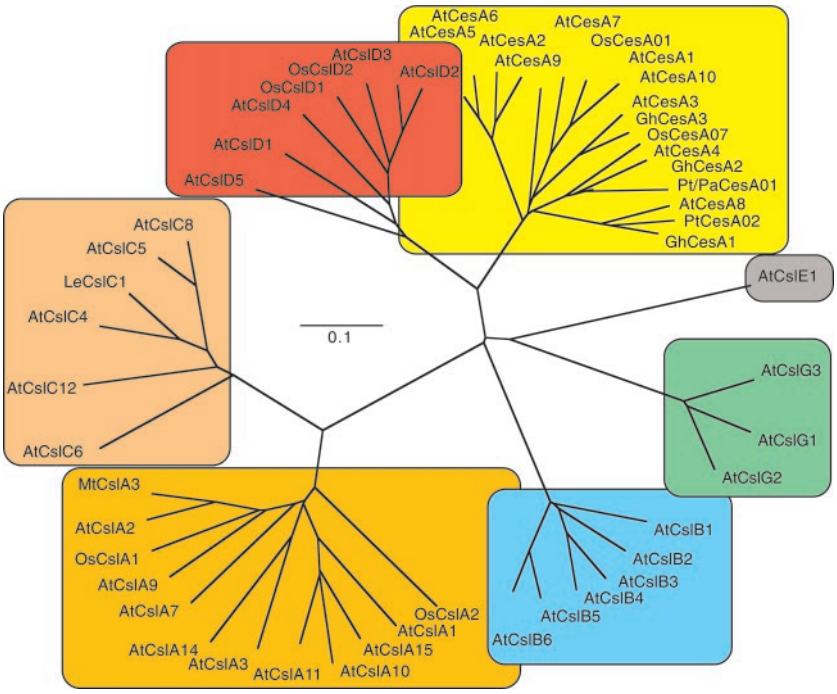

Figure 1. Unrooted, bootstrapped tree of the *CesA* superfamily. ClustalX (version 1.8) was used to create an alignment of the full-length, publicly available protein sequences that was then bootstrapped ($n = 5{,}000$ trials) to create the final tree. Subfamilies are boxed. At, Arabidopsis; Gh, cotton; Le, tomato; Mt, *Medicago truncatula*; Os, rice; Pt, *Populus tremuloides*; Pt/Pa, *Populus tremula* × *Populus alba*.

different patterns of intron-exon organization. Based on recent thinking about the evolution of intron/exon structure (de Souza et al., 1998), the small number of introns in this family, and their divergent nature, would seem to suggest that this gene family is the oldest in the cellulose synthase superfamily and may predate the *CesA* family.

All members of the *CesA* family contain a putative LIM-like Zn-binding domain/RING finger domain in the N-terminal region, which is similar to several putative plant Leu zipper transcription factors (Kawagoe and Delmer, 1997a, 1997b; Arioli et al., 1998). LIM domains are known to mediate protein-to-protein interactions (Bach, 2000), whereas RING finger domains are thought to play a role in ubiquitin-mediated proteolysis (Freemont, 2000). These domains may play a role in mediating CesA function via protein partners or targeted degradation. All of the *Csl* proteins lack this amino terminus extension, including the *CslD* family, which contains proteins similar in size to the CesAs.

Although the various CesA and Csl proteins vary in their degree of sequence similarity to one another (Table I), they share several features that have been proposed to be indicative of processive glycosyltransferases (Saxena et al., 1995). All of the *CesA* and *Csl* gene products contain a D,D,D,QxxRW motif (Fig. 2), which has been proposed to define the nucleotide sugar-binding domain and the catalytic site of these enzymes. Based on this motif, the proposed topology of these proteins (discussed above), and sequence-based classification, the various members of the Arabidopsis cellulose synthase superfamily appear to belong to family 2 of the inverting nucleotide-

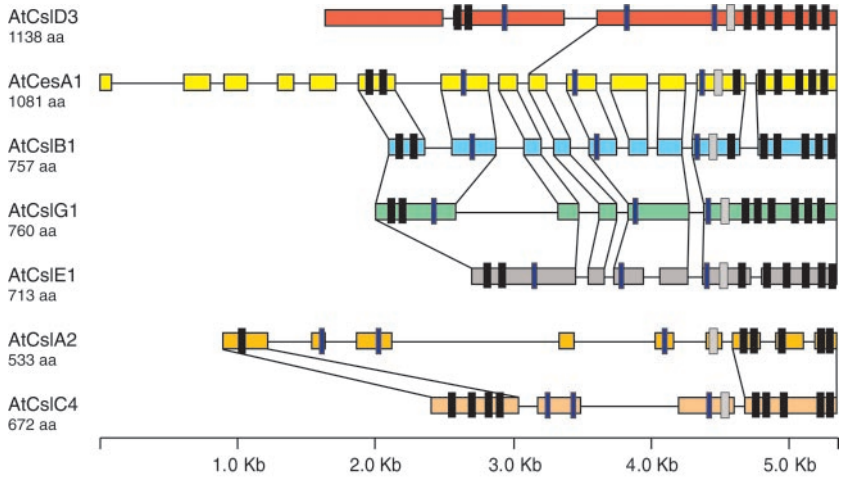

Figure 2. Comparison of the gene structure of representative genes of the Arabidopsis *CesA* superfamily. Colored boxes represent exons and the lines connecting them denote introns. Thick vertical black bars indicate predicted transmembrane domains as predicted by HMMTOP (http://www.enzim.hu/hmmtop/). Thin blue bars represent conserved Asp residues, and the thicker gray bar represents the QxxRW domain. Thin lines connecting different genes indicate conserved intron-exon junctions.

Table I. *Identity/similarity matrix for selected members of the CesA superfamily*

Identity	Similarity						
	AtCesA1	AtCslD3	AtCslB1	AtCslG1	AtCslE1	AtCslA2	AtCslC4
AtCesA1	–	48.1	31.3	29.3	30.7	13.1	14.3
AtCslD3	37.1	–	28	27.7	28.3	12.2	14
AtCslB1	22.1	18.9	–	37.4	41.1	17.4	18.6
AtCslG1	21.2	18.4	25.4	–	48.7	16.3	17.3
AtCslE1	21.4	18.9	30.1	34.4	–	16.6	18.2
AtCslA2	7.1	6.3	9.1	8.4	8.7	–	44.8
AtCslC4	8.2	6.7	9.3	8.7	9.1	31.9	–

diphospho-sugar glycosyltransferases (Campbell et al., 1997) that synthesize repeating β-glycosyl unit structures. To date, this family includes over 500 putative members, including cellulose synthase, chitin synthase, hyaluronan synthase, β-1,3-glucan synthase, and a number of uncharacterized genes from many organisms (Campbell et al., 1997; http://afmb.cnrs-mrs.fr/~pedro/CAZY/gtf_2.html). The function of the various *Csl* families is not known, but speculation is that they are responsible for producing some of the other polysaccharides found in plant cell walls and in secretions such as root cap or stylar mucilage (Cutler and Somerville, 1997). Although the D,D,D,QxxRW motif is thought to be indicative of processive β-glycosyltransferases, there is no comparative sequence data available on processive α-glycosyltransferases. Therefore we cannot rule out the possibility that some of these enzymes produce polysaccharides with α-linkages, such as rhamnogalacturonan I or rhamnogalacturonan II. It is possible that linkage specificity is determined by subtle features in the active site of the proteins (Stasinopoulos et al., 1999) and that members of the Arabidopsis cellulose synthase superfamily make polysaccharides with both β- and α-linkages.

DISCUSSION

With six families of *Csl* genes and six major noncellulosic polysaccharides in Arabidopsis (i.e. callose, xyloglucan, glucuronoarabinoxylan, homogalacturonan, rhamnogalacturonan I, and rhamnogalacturonan II), it is tempting to speculate that each family is responsible for the biosynthesis of one of the principal polysaccharides of the cell wall. Although we consider it possible that the gene superfamily described here encodes enzymes that catalyze the synthesis of different polymers, there is at present no evidence for this other than the observation that sequence divergence is frequently associated with functional divergence. It is also possible that there are additional functional divisions within the gene families that are not evident from our analysis. Recent results concerning the relationship between enzyme structure and function, such as experiments showing that as few as four amino acid changes can alter the catalytic outcome of an enzymatic reaction from desaturation to hydroxylation (Broun et al., 1998), emphasize the need for caution in using sequence similarity to infer function based on sequence.

The amount of plant genome sequence and EST information in the public sequence databases is expanding rapidly. At present there are more than 900,000 plant ESTs and genome survey sequences in GenBank, most of which are from 35 species. In the first 8 months of the year 2000, more than 516,000 new ESTs and genome survey sequences from 16 plant species were deposited. Thus except for species such as Arabidopsis, which will soon be completely sequenced, any attempt at a comprehensive compilation of CesA-related sequence information represents a continuing challenge. To facilitate research on these genes, we have established a website (http://cellwall.stanford.edu) that summarizes the ever-increasing number of cellulose synthase and cellulose synthase-like genes. At present, there are more than 1,250 *CesA* and *Csl* sequences, from 29 different plant species in GenBank. Although the most extensive information available is for Arabidopsis where there are more than 330 partial or complete gene sequences, there is also a significant amount of information available for several other species, especially rice, maize, soybean, and tomato. A crude estimate of the relative abundance of

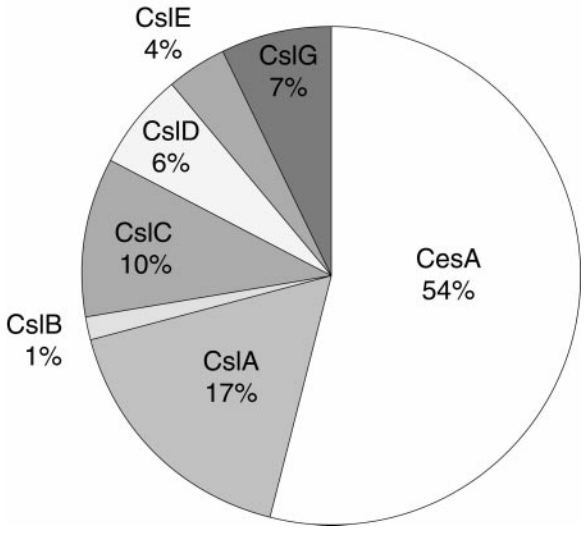

Figure 3. Relative abundance of EST sequences for members of the *CesA* and *Csl* families in GenBank.

mRNA for the various family members can be calculated from the frequency with which each gene family is represented by EST sequences in the public databases (Fig. 3).

Polysaccharides found in other plant species, but not in Arabidopsis (Zablackis et al., 1995), such as mixed linkage xylans, mannans, or arabinans, may be synthesized by genes that are not represented by orthologs in Arabidopsis. A number of gene sequences from plants in GenBank show limited similarity (<50% identity) to the members of the various *Csl* families in Arabidopsis. This and other issues will undoubtedly become more transparent when the function of the *Csl* genes in Arabidopsis is known from direct experimental evidence. Our laboratory, along with others, is examining the patterns of gene expression and protein localization of the Arabidopsis *Csl* genes, and attempting to characterize their enzymatic function using reverse genetics. We are confident that in the next several years the function of these genes will be understood and it will then be possible to begin to unravel the challenge of understanding how cell wall composition and deposition is controlled.

Received May 25, 2000; accepted July 7, 2000.

LITERATURE CITED

Arioli T, Peng L, Betzner AS, Burn J, Wittke W, Herth W, Camilleri C, Hofte H, Plazinski J, Birch R, Cork A, Glover J, Redmond J, Williamson RE (1998) Molecular analysis of cellulose biosynthesis in *Arabidopsis*. Science **279:** 717–720

Bach I (2000) The LIM domain: regulation by association. Mech Dev **91**: 5–17

Broun P, Shanklin J, Whittle E, Somerville CR (1998) Catalytic plasticity of fatty acid modification enzymes underlying chemical diversity of plant fatty acids. Science **282:** 1315–1317

Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem J **326:** 929–939

Cutler S, Somerville C (1997) Cellulose synthase: cloning by in silico. Curr Biol **7:** R108–R111

Delmer DP (1999) Cellulose biosynthesis: exciting times for a difficult field of study. Annu Rev Plant Physiol Plant Mol Biol **50:** 245–276

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA **95:** 5094–5099

Freemont PS (2000) Ubiquitination: ring for destruction? Curr Biol **10:** R84–R87

Kawagoe Y, Delmer DP (1997a) Cotton CelA1 has a LIM-like Zn binding domain in the N-terminal cytoplasmic region (abstract no. 337). Plant Physiol **114:** S–85

Kawagoe Y, Delmer DP (1997b) Pathways and genes involved in cellulose biosynthesis. Genet Eng **19:** 63–87

Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM (1996) Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. Proc Natl Acad Sci USA **93:** 12637–12642

Saxena IM, Brown RM Jr, Fevre M, Geremia RA, Henrissat B (1995) Multidomain architecture of β-glycosyl transferases: implications for mechanism of action. J Bacteriol **177:** 1419–1424

Stasinopoulos SJ, Fisher PR, Stone BA, Stanisich VA (1999) Detection of two loci involved in (1→3)-β-glucan (curdlan) biosynthesis by *Agrobacterium* sp. ATCC31749, and comparative sequence analysis of the putative curdlan synthase gene. Glycobiology **9:** 31–41

Taylor NG, Scheible WR, Cutler S, Somerville CR, Turner SR (1999) The *irregular xylem 3* locus of Arabidopsis encodes a cellulose synthase gene required for secondary cell wall synthesis. Plant Cell **11:** 769–780

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

Zablackis E, Huang J, Müller B, Darvill AG, Albersheim P (1995) Characterization of the cell wall polysaccharides of *Arabidopsis thaliana* leaves. Plant Physiol **107:** 1129–1138