Scientific Correspondence

Cellulose Synthase-Like Genes of Rice[1]

Samuel P. Hazen, John S. Scott-Craig, and Jonathan D. Walton*

Department of Energy-Plant Research Laboratory, Michigan State University, East Lansing, Michigan 48824

Identification of the biosynthetic enzymes involved in cell wall biosynthesis remains one of the major unsolved problems of plant biology. Of the major polysaccharides of the plant cell wall, pectins and hemicelluloses are synthesized in the Golgi, and callose and cellulose are synthesized at the plasma membrane. The evidence is now quite extensive that the catalytic subunits of cellulose synthase are encoded by members of the large *CESA* gene family (Arioli et al., 1998; Fagard et al., 2000; Holland et al., 2000; Taylor et al., 2000). With a few exceptions, however, the genes for the enzymes of pectin and hemicellulose biosynthesis have not been identified (Edwards et al., 1999; Perrin et al., 1999). Nothing is currently known about the genes encoding the enzymes that catalyze the synthesis of the hemicellulose backbones.

The primary cell walls of all higher plants contain large amounts of cellulose in their walls, and, consistent with this, *CESA* genes are found throughout the plant kingdom (Richmond, 2000; Richmond and Somerville, 2000). In contrast, the hemicelluloses of dicotyledons and graminaceous monocotyledons (cereals) are distinct. Whereas dicots contain large amounts of pectin and xyloglucan, cereals contain low amounts of pectin and xyloglucan, large amounts of glucuronoarabinoxylan, and, at least in some tissues, the cereal-specific polymer (1–3),(1–4)-β-D-glucan (also known as mixed-linked glucan) (Carpita and Gibeaut, 1993; Carpita, 1996). On the basis of these structural differences, it would be expected that dicots and cereals would have a distinct panoply of hemicellulose biosynthetic enzymes.

Plants contain a superfamily of genes, called *CSL* (cellulose synthase-like), whose amino acid sequences are related to the *CESA* genes. The Csl proteins are predicted to be integral membrane proteins and contain a sequence, the "D,D,D,QXXRW" motif, that seems to be characteristic of processive glycosyl transferases (Saxena and Brown, 1995). On these grounds, it has been proposed that the *CSL* genes encode the catalytic subunits of the enzymes that synthesize the hemicellulose backbones (Richmond and Somerville, 2000, 2001).

Although no biochemical function has yet been elucidated for any *CSL* gene, three studies implicate them in wall biosynthesis. Root hairs of Arabidopsis plants that are mutated in *AtCSLD3* are defective, apparently because of abnormal cell walls (Favery et al., 2001; Wang et al., 2001). A gene (*NaCSLD1*) that is highly expressed in *Nicotiana alata* pollen tubes, whose walls are composed almost entirely of callose and cellulose, has been proposed to encode a pollen-specific cellulose synthase (Doblin et al., 2001). Arabidopsis mutants in *AtCSLA9* have increased resistance to *Agrobacterium tumefaciens*, which binds to plant cell walls at an early stage of infection (Nam et al., 1999).

With the completion of the Arabidopsis genome, every *CSL* gene in this plant has been identified (Richmond and Somerville, 2001). The rice (*Oryza sativa*) genome is expected to be complete by the end of 2002, and currently, approximately 50% of the rice genome is available either publicly in GenBank or through Monsanto's password-protected web site (http://www.rice-research.org). Approximately 80,000 rice expressed sequence tags (ESTs) and the actual corresponding cDNAs are also in the public domain.

We present here an analysis of the *CSL* genes present in the available rice sequence databases. We have identified 37 *CSL* genes and have deduced full-length protein coding sequences for 23 of them (Table I). The genes were identified by BLAST searches of GenBank (nonredundant and dbEST) and the Monsanto database using the Arabidopsis CesA and Csl proteins as queries. Richmond's web page (http://cellwall.stanford.edu) served as a very useful starting point for the analysis. cDNAs corresponding to all *OsCSL* ESTs were obtained from the appropriate sources and sequenced completely. Most of the cDNAs came from the Rice Genome Research Program (http://rgp.dna.affrc.go.jp). The Rice Genome Research Program cDNA clones were of high quality; all but one were viable and accurately annotated. The one exception, D22177, was chimeric, containing *OsCSLA2* at one end and a predicted DNA-binding protein at the other. For all sequences, the corresponding proteins were deduced using gene prediction software from GeneMark (Atlanta; http://opal.biology.gatech.edu/GeneMark) and Softberry, Inc. (White Plains, NY; http://www.softberry.com), and by manual alignment with the Arabidopsis Csl

[1] This work was supported in part by the U.S. Department of Energy, by the Division of Energy Biosciences, and by the National Science Foundation Plant Genome Program (to Natasha Raikhel [MSU-DOE], Ken Keegstra [MSU-DOE], and J.D.W.).

* Corresponding author; e-mail walton@msu.edu; fax 517–353–9168.

www.plantphysiol.org/cgi/doi/10.1104/pp.010875.

Table I. *The CSL superfamily of rice*

Sequences are available at www.prl.msu.edu/walton.

No.	Gene Name[a]	Monsanto/GenBank Accession Nos.[b]	EST Accession No. (Size in kb)	Protein Size (Amino Acids)	Full Length?[c]	Chromosome	GenBank Accession No. from This Paper	
							Genomic Sequence[d]	cDNA Sequence
1	*CSLA1*	OSM12487, AP000367		521	Yes	2	BK000080	
2	*CSLA2*	AC021893	D22177 (1.1)	524	Yes	10	BK000092	AF435640
3	*CSLA3*	AP003509		551	Yes	6	BK000081	
4	*CSLA4*[e]	OSM11235, AC073556		602	Yes	3	BK000082	
5	*CSLA5*	OSM13798, OSM13800, AC084766		574	Yes	3	BK000083	
6	*CSLA6*	OSM15467	AA749881 (0.7) AU166554 (NS)[f]	574	Yes		AF432498	AF435648
7	*CSLA7*		AU093819 (1.9) (=C71923)[g] BE040507 (NS)[f]	479	No			AF435643
8	*CSLA8*[h]	OSM150433		429	No			
9	*CSLA9*	OSM145719		527	Yes		AF432499	
10	*CSLA10*[h]	OSM124376		435	No			
11	*CSLC1*	OSM15560, AP003377		690	Yes	1	BK000086	
12	*CSLC2*	OSM129292	AI978402 (1.8)	698	Yes		BK000087	AF435650
13	*CSLC3*	OSM13550, AP004013		745	Yes	8	BK000088	
14	*CSLC4*	OSM15738		159	No			
15	*CSLC5*	OSM1603		123	No			
16	*CSLC6*	OSM15729		155	No			
17	*CSLC7*	OSM13738	C74862 (1.1)	572	No			AF435642
18	*CSLC8*	OSM146469		210	No			
19	*CSLC9*	OSM133403	AU068180 (2.0)	595	Yes		AF435652[i] AF435653	AF435641
20	*CSLD1*	OSM13541, AC027037		1127	Yes	10	BK000089	
21	*CSLD2*	OSM14185, AP001552	AA753599 (0.6)	1170	Yes	6	BK000090	AF435649
22	*CSLD3*[h]	AC091687		1148	Yes	9	BK000093	
23	*CSLD4*		AU078363 (0.4) (=AU082165)[g] AU082190 (1.2) (=AU082189)[g]	399	No			AF435644
24	*CSLE1*	OSM151624, OSM151625	AU068392 (1.1) (=AU166543)[g]	730	Yes		AF432500	AF435647
25	*CSLE2*	OSM147124, OSM147116		745	Yes		AF432501	
26	*CSLE3*	OSM16239		173	No			
27	*CSLE4*[h]	OSM133730		135	No			
28	*CSLE5*[h]	OSM151623		623	No			
29	*CSLF1*	OSM14797, OSM151757, OSM151758, AP004261		860	Yes	7	AF432502	
30	*CSLF2*	OSM151759, OSM14795, AP004261	C98682 (1.6) (=AU101138)[g]	889	Yes	7	AF432503	AF435651
31	*CSLF3*	OSM151756, OSM14798, OSM14796, AP004261		868	Yes	7	AF432504	
32	*CSLF4*	OSM151756, OSM14798, OSM14796, AP004261		889	Yes	7	AF432505	
33	*CSLF5*[h]	OSM151760		330	No			
34	*CSLF6*		D40419 (2.0)	560	No			AF435645
35	*CSLF7*	OSM16238, AC090441		830	Yes	10	BK000091	
36	*CSLH1*[h]	OSM16234	AU085988 (2.4)	750	Yes		BK000084	AF435646
37	*CSLH2*	OSM13388		762	Yes		BK000085	

[a] To the extent possible, the gene nomenclature has been made consistent with that of Richmond (http://cellwall.stanford.edu). [b] OSM indicates a Monsanto database accession number; all other accession numbers refer to GenBank. Multiple OSM contigs for a single gene indicate that the contigs overlap; OSM151756, OSM14798, and OSM14796 overlap to form one contig containing two *CSLF* genes, which are also present on AP004261 along with *OsCSLF1* and *OsCSLF2*. [c] Indicates whether a full-length protein can be deduced with reasonable confidence. [d] Accession numbers starting with AF are standard GenBank entries. Numbers starting with BK are in the GenBank Third Party Annotation database. [e] There appear to be three frameshifts within an ~80-bp region of *CSLA4*. Two apparently independent genomic sequences containing this gene, one from Monsanto (OSM11235) and the other from The Institute for Genomic Research (TIGR) (GenBank AC073556), are identical. The sequence covering this region in AC073556 is of "very high quality" (Robin Buell, TIGR, personal communication). Therefore, *CSLA4* is probably a pseudogene. [f] NS, not sequenced. The sequence of AU166554 did not correspond to the published EST sequence; the source of this discrepancy has not been determined. [g] the "equals" sign indicates that the two accession numbers represent two EST sequences from the same cDNA clone, confirmed by complete sequencing of the cDNA. [h] These DNA sequences were concluded to contain the following errors: three frame shifts in *OsCSLA8*; one frame shift in *OsCSLA9*; one frame shift and one in-frame stop codon in *OsCSLA10* (in addition, OSM124376 is probably chimeric); two nucleotide omissions in the genomic sequence of *OsCSLH1* (OSM16234), which were identified by comparison to the cDNA sequence of AU085988; an intron start of GC instead of GT in *OsCSLD3*; one frame shift in *OsCSLE4*; five frame shifts and an in-frame stop codon in *OsCSLE5*; a frame shift and two in-frame stop codons in *OsCSLF5*. If any of these assumed errors are real, then the corresponding genes might be pseudogenes. [i] The sequence of OSM133403 is interrupted by a string of undefined nucleotides (NNNN...). It has therefore been submitted to GenBank as two sequences. The undefined sequences occur within an intron, which has been established using the sequence of an overlapping cDNA, and therefore do not affect the deduced protein sequence.

proteins and with each other. The sequences were aligned with Clustal X and presented with TreeView (Glasgow, UK) and CorelDraw (Ottawa, ON, Canada) (Thompson et al., 1994; Page, 1996; Jeanmougin et al., 1998).

Like the Arabidopsis Csl proteins, all of the rice Csl proteins are predicted to be integral membrane proteins. All except two have the QXXRW motif (Saxena and Brown, 1995). The exceptions are OsCslA10, which has RXXRW, and OsCslE2, which has LXXRW, at the equivalent positions. All of the OsCsl proteins have a DXD motif approximately 120 to 250 amino acids upstream of QXXRW.

The results indicate that there are both striking similarities as well as differences between the *CSL* genes of rice and Arabidopsis, which may reflect the similarities and differences in the hemicellulose composition of dicots and graminaceous monocots. Arabidopsis and rice both contain members of the *CSLA*, *CSLC*, *CSLD*, and *CSLE* families with no consistent distinctions between the two species (Fig. 1). However, the rice and Arabidopsis sequences differ in at least three respects.

First, rice has a group of *CSL* genes, the products of which are related to CesA and CslD but nonetheless form a distinct group separate from either of these families (Fig. 1). These proteins are also significantly shorter than the CesA or CslD proteins because of truncation at their N termini (Fig. 1). On these grounds, we propose that these genes constitute a new cereal-specific family, for which we propose the name *CSLF*. (As with earlier classifications of the *CSL* genes [Richmond and Somerville, 2001], the family designations are solely for nomenclatural convenience and do not necessarily reflect any underlying functional relationships).

The products of *OsCSLF1* and *OsCSLF2* have >98% amino acid identity but are clearly two different genes based on a number of nucleotide differences in their 5′- and 3′-untranslated regions. *OsCSLF1*, *OsCSLF2*, *OsCSLF3*, and *OsCSLF4* are physically linked within an approximately 49-kb region on PAC AP004261. Consistent with this, *OsCSLF3* and *OsCSLF4* are on the same overlapping Monsanto contigs (Table I). It is not yet known if any of the other *OsCSL* genes are clustered, although some are on the same chromosomes (Table I).

Some doubt remains about the accuracy of the deduced amino acid sequence of *OsCSLF7*. It appears to be both the most divergent and the shortest of the

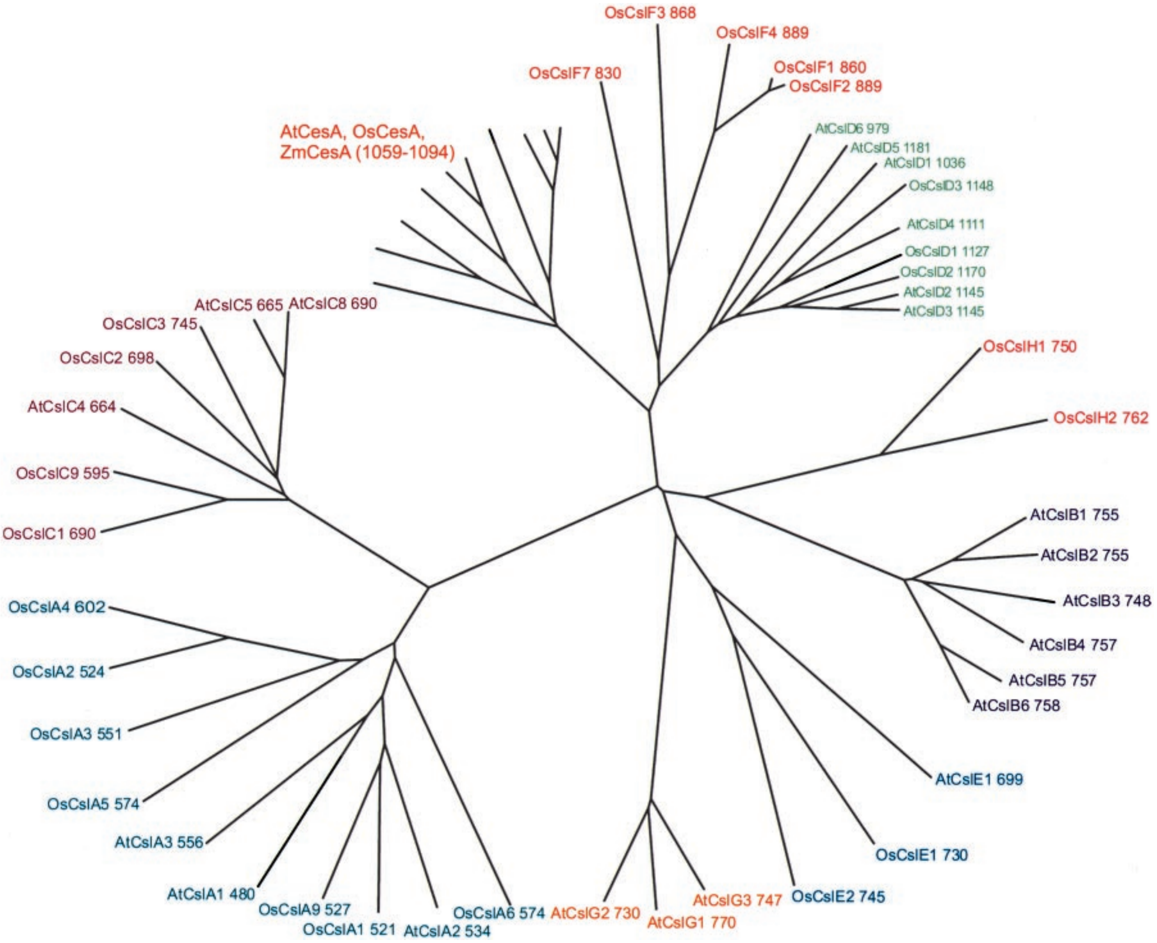

Figure 1. Unrooted phylogenetic tree of Csl proteins from rice and Arabidopsis. Only the deduced full-length rice Csl (OsCsl) proteins are included. The Arabidopsis Csl coding sequences were deduced by the same criteria used for the rice proteins and the sizes of many of the AtCsl proteins differ slightly from those given by Richmond (http://cellwall.stanford.edu). All of the Arabidopsis CslB, CslD, CslE, and CslG proteins are included, but for clarity only three of nine AtCslA, three of five AtCslC, and a sampling of maize (*Zea mays*), rice, and Arabidopsis CesA proteins are shown; inclusion of the others did not significantly change any of the relationships. The lengths of each deduced protein in number of amino acids are indicated after the protein names.

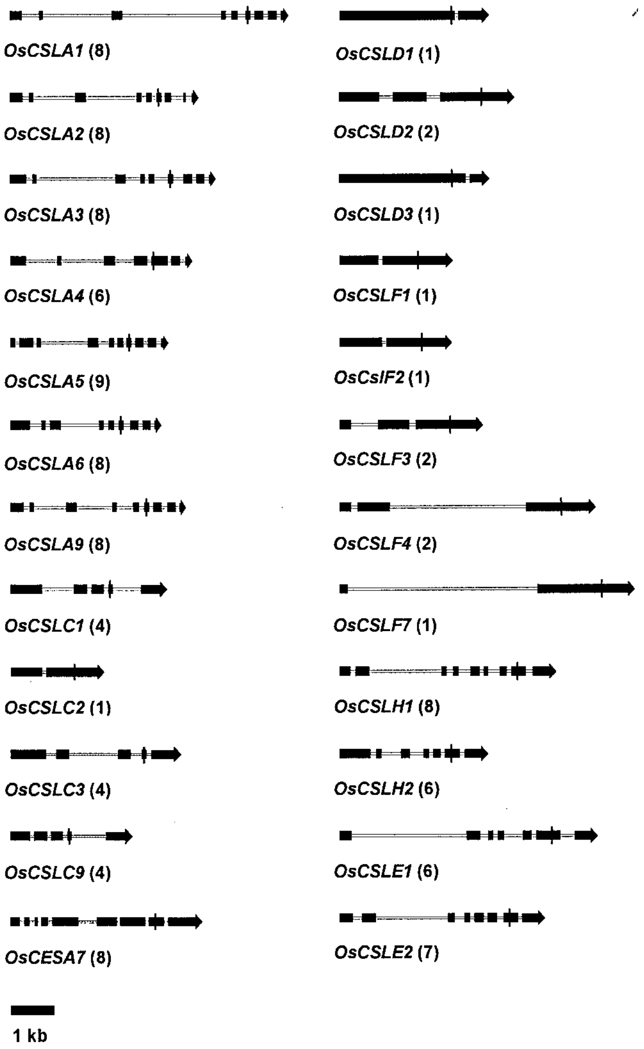

Figure 2. Intron/exon structures of the full-length rice *CSL* genes. Exons are indicated by solid boxes and introns by white boxes. Vertical black lines indicate the position of the QxxRW motif. The number of introns for each gene is indicated in parentheses after the gene name. The genes are drawn to scale; the bar in the lower left indicates 1 kb.

OsCSLF family (Fig. 1). The structure of *OsCSLF7*, with a short N-terminal exon followed by a large (4 kb) intron (Fig. 2), is one that in our experience is particularly hard for gene prediction programs to call correctly. The structure of *OsCSLF7* should be considered tentative until a full-length cDNA is sequenced.

Full-length coding sequences for *OsCSLF5* and *OsCSLF6* are not available, and the two deduced partial proteins do not overlap. Therefore, it is possible that these two proteins are from the same gene.

A second major difference between Arabidopsis and rice is the deep branching between their respective members in the CslB family. All six Arabidopsis CslB proteins form one cluster, whereas the two rice CslB-like proteins form a related but distinct branch. No rice proteins cluster tightly with the AtCslB sequences. In contrast to the OsCslF proteins, the deduced CslB-like proteins of the two species are similar in size (Fig. 1). We attempted to analyze other CslB and CslB-like proteins, based on EST sequences, from other dicots and cereals to see if the dichotomy shown in Figure 1 would hold up. Two partial *Sorghum bicolor* CslB-like proteins could be reliably assembled from public ESTs, and both of these (SbCslB2 accession nos. A286049 and BE594529; SbCslB3 nos. BE597410 and BG463462; see http://cellwall.stanford.edu) aligned more closely with the rice CslB-like proteins than with the AtCslB family (data not shown). This supports the hypothesis that the cereal CslB-like proteins constitute a distinct family, and we therefore propose the name *CSLH* for the rice *CSLB*-like genes.

A third salient feature of the tree (Fig. 1) is that rice apparently lacks any *CSLG* family, members of which are widespread in dicots and have not been found so far in any monocot. This observation was made earlier by Richmond and Somerville (2001).

Arabidopsis is predicted to have 30 *CSL* genes (Richmond and Somerville, 2001), whereas rice has at least 37 (Table I). A number of the rice genome survey sequences predict the existence of additional *OsCSL* genes (see http://cellwall.stanford.edu), but because of their short lengths, unavailability for further sequencing, and lack of utility for predicting intron/exon structure, they have not been included in the current analysis. Rice and Arabidopsis differ in the number of predicted genes in each of the "common" families. Arabidopsis and rice have nine and 10 *CSLA* genes, five and nine *CSLC* genes, six and four CSLD genes, and one and five *CSLE* genes, respectively.

Intron/exon structures were deduced for all of the full-length *OsCSL* genes (Fig. 2). The *OsCESA*, *OsCSLA*, *OsCSLH*, and *OsCSLE* families tend to have more introns compared with *OsCSLD*, *OsCSLC*, and *OsCSLF*. In Arabidopsis, the *AtCSLD* family has the fewest introns (Richmond and Somerville, 2000). Intron number also tends to be conserved within a family (Fig. 2).

Genes in the *CSL* superfamily are currently the most promising candidates for encoding the glycosyl synthases that make the hemicellulose backbones of plant cell walls (Richmond and Somerville, 2001). Although all plant cell walls have similarities in their polysaccharide composition, the hemicelluloses of dicots and cereals show marked differences (Carpita, 1996). This dimorphism is expected to be reflected in distinct patterns of wall biosythetic enzymes and hence encoding genes. Consistent with both the similarities and differences between the walls of dicots and cereals, the *CSL* gene superfamily shows both degrees of conservation and degrees of differences between Arabidopsis and rice.

ACKNOWLEDGMENTS

We thank Robin Hawley (Michigan State University-Department of Energy [MSU-DOE]) for technical assistance and Todd Richmond (NimbleGen Systems, Inc., Madison, WI) for useful discussions. We thank Weiqing Zeng (MSU-DOE) for sharing his analyses of the Arabidopsis *CSL* genes. For cDNA clones, we thank the Rice Genome Research Program of the National Institute of Agrobiological Resources (Tsukuba, Japan); the Department of Cytogenetics (National Institute of Agricultural Sciences and Technology, Suwon City, Korea); the Department of Plant Breeding (Cornell University, Ithaca, NY); and Christine Michalowski (University of Arizona, Tucson). All cDNA clones mentioned in this paper are available to nonprofit researchers directly from the original source or, with their written permission, from the corresponding author.

Received September 25, 2001; returned for revision September 29, 2001; accepted November 2, 2001.

LITERATURE CITED

Arioli T, Peng LC, Betzner AS, Burn J, Wittke W, Herth W, Camilleri C, Höfte H, Plazinski J, Birch R et al. (1998) Science **279:** 717–720

Carpita NC (1996) Annu Rev Plant Physiol Plant Mol Biol **47:** 445–476

Carpita NC, Gibeaut DM (1993) Plant J **3:** 1–30

Doblin MS, De Melis L, Newbigin E, Bacic A, Read SM (2001) Plant Physiol **125:** 2040–2052

Edwards ME, Dickson CA, Chengappa S, Sidebottom C, Gidley MJ, Reid JSG (1999) Plant J **19:** 691–697

Fagard M, Desnos T, Desprez T, Goubet F, Refregier G, Mouille G, McCann M, Rayon C, Vernhettes S, Höfte H (2000) Plant Cell **12:** 2409–2423

Favery B, Ryan E, Foreman J, Linstead P, Boudonck K, Steer M, Shaw P, Dolan L (2001) Genes Dev **15:** 79–89

Holland N, Holland D, Helentjaris T, Dhugga KS, Xoconostle-Cazares B, Delmer DP (2000) Plant Physiol **123:** 1313–1323

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Trends Biochem Sci **23:** 403–405

Nam J, Mysore KS, Zheng C, Knue MK, Matthysse AG, Gelvin SB (1999) Mol Gen Genet **261:** 429–438

Page RDM (1996) Comp Appl Biosci **12:** 357–358

Perrin RM, DeRocher AE, Bar-Peled M, Zeng WQ, Norambuena L, Orellana A, Raikhel NV, Keegstra K (1999) Science **284:** 1976–1979

Richmond T (2000) Genome Biol **1:** 3000.1–3000.6.

Richmond TA, Somerville CR (2000) Plant Physiol **124:** 495–498

Richmond TA, Somerville CR (2001) Plant Mol Biol **47:** 131–143

Saxena IM, Brown RM Jr (1995) J Bacteriol **177:** 5276–5283

Taylor NG, Laurie S, Turner SR (2000) Plant Cell **12:** 2529–2539

Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res **11:** 4673–4680

Wang X, Cnops G, Vanderhaeghen R, De Block S, Van Montagu M, Van Lijsebettens M (2001) Plant Physiol **126:** 575–586