research articles Journal of proteome research

Identification of Novel Proteins Involved in Plant Cell-Wall Synthesis Based on Protein−Protein Interaction Data

Chan Zhou,[†,‡] Yanbin Yin,[†,‡] Phuongan Dam,[†,‡] and Ying Xu*[,†,‡,§]

Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, Georgia 30602, BioEnergy Science Center (http://genomicsgtl.energy.gov/centers/), Oak Ridge, Tennessee 37831, and College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China

Received April 4, 2010

The plant cell wall is mainly composed of polysaccharides, representing the richest source of biomass for future biofuel production. Currently, the majority of the cell-wall synthesis-related (CWSR) proteins are unknown even for model plant *Arabidopsis thaliana*. We report a computational framework for predicting CWSR proteins based on protein−protein interaction (PPI) data and known CWSR proteins. We predict a protein to be a CWSR protein if it interacts with known CWSR proteins (seeds) with high statistical significance. Using this technique, we predicted 100 candidate CWSR proteins in *Arabidopsis thaliana*, 8 of which were experimentally confirmed by previous reports. Forty-two candidates have either independent supporting evidence or strong functional relevance to cell-wall synthesis and, hence, are considered as the most reliable predictions. For 33 of the predicted CWSR proteins, we have predicted their detailed functional roles in CWS, based on analyses of their domain architectures, phylogeny, and current functional annotation in conjunction with a literature search. We present the constructed PPIs covering all the known and predicted CWSR proteins at http://csbl.bmb.uga.edu/~zhouchan/CellWallProtein/. The 42 most reliable candidates provide useful targets to experimentalists for further investigation, and the PPI data constructed in this work provides new information for cell-wall research.

Keywords: plant cell-wall synthesis • protein function prediction • protein−protein interaction • *Arabidopsis thaliana;* biofuel • computational biology

Introduction

The plant cell wall plays various roles, such as determining the overall cell shape, providing support and mechanical strength, preventing the cell membrane from bursting in a hypotonic medium, controlling the rate and direction of cell growth and regulating cell volume, ultimately being responsible for the plant architectural design and controlling plant morphogenesis, as well as in response to environmental and pathogen-induced stresses. The plant-cell wall is made of cellulose, hemicellulose, pectin, lignin, structural proteins and aromatic substances.[1] The current understanding is that the cell-wall synthesis process consists of six major components: (1) substrate generation, (2) polysaccharide synthesis and modification, (3) secretion and targeting pathways, (4) assembly, architecture and growth, (5) differentiation and secondary wall formation, and (6) signaling and response during the wall development and disassembly.[1] This synthesis process has been extensively studied,[2–5] yet the detailed understanding is still rather limited about this process at the molecular and the cellular level due to its high complexity:[6] the vast majority of the CWSR genes are yet to be identified and the detailed functions of many known CWSR genes remain to be elucidated. An improved understanding of plant-cell wall synthesis could lead to a much improved capability in converting plant cell walls into biofuel[7] through transgenic approaches to lessening lignin's impact on the recalcitrance of biomass.[8] Identification and functional characterization of proteins involved in plant cell-wall synthesis represent the first key step in deciphering the molecular and cellular mechanisms of cell-wall synthesis.

The current estimation is that over 2000 proteins are involved in the cell-wall synthesis in *A. thaliana*.[1] This is far more than the number of proteins already known to be involved in this process. As of now, the Cell Wall Navigator (CWN) database has ~600 *A. thaliana* proteins believed to be involved in sugar substrate generation and primary cell wall metabolism.[9] The Purdue Cell Wall database contains ~1000 *A. thaliana* genes annotated to be relevant to cell-wall synthesis, which includes genes that encode precursor proteins, structural proteins, and enzymes involved in polysaccharide modification and depolymerization though only a small portion of them has been experimentally validated.[10] There is a clear gap between what we already know about cell-wall synthesis proteins and what we need to know so we can more effectively utilize the biomass

* To whom correspondence should be addressed. Ying Xu, Phone: 706-542-9779. Fax: 706-542-9751. Email: xyn@bmb.uga.edu.
† University of Georgia.
‡ BioEnergy Science Center.
§ Jilin University.

10.1021/pr100249c

available from plant cell walls. The focus of this study is on identification of novel proteins that are involved in the cell-wall synthesis encoded in the genome of *A. thaliana*.

Elucidation of cell-wall synthesis-related (CWSR) proteins using pure experimental approaches without target candidates has proved to be a challenging task.[11] We believe that a bioinformatics approach could play a useful role in suggesting candidate proteins, based on integration and application of multiple sources of information derived from plant genomic, transcriptomic and proteomic data in conjunction with computational prediction of protein functions, subcellular localization and phylogeny, as demonstrated previously.[11] Several bioinformatics studies have been conducted to identify proteins involved in cell-wall synthesis. The Carbohydrate-Active enZyme (CAZy) database represents the most comprehensive database for enzymes relevant to building and degrading complex carbohydrates and glycoconjugates, for example, polysaccharide lyases, carbohydrate esterases, glycoside hydrolases and glycosyltransferases (GTs).[12] Using a bioinformatics approach, Egelund et al. identified 27 new GTs in *A. thaliana*[13] and Hansen et al. predicted 20 putative GTs in *A. thaliana*.[14] Brown et al. (2005) and Persson et al. (2005) have identified novel genes involved in secondary cell-wall formation and cellulose synthesis, independently, both based on coexpression data analyses of plant cell walls.[15,16] In addition, a few glycine-rich protein (GRP) genes have been found to play important roles in the development of xylem secondary cell walls using a combination of gene expression data analyses and reverse genetics.[17] Also, several candidate proteins for the synthesis and feruloylation of arabinoxylan have been identified using bioinformatics approaches.[18]

Computational prediction of protein−protein interaction (PPI) networks represents an effective technique for deriving useful information regarding physical and functional interactions among otherwise unrelated proteins.[19] Such predictions are typically done based on experimental PPI data generated using "high-throughput" techniques such as yeast two-hybrid[20] or protein affinity chromatography.[21] In this paper, we present a computational approach based on PPIs to infer novel proteins that may be involved in cell-wall synthesis in *A. thaliana* but are not present in the CWN and Purdue Cell Wall databases. We predicted 92 candidate proteins. Among them, 42 (in Table 1) are considered to be the most reliable predictions because there are independent supporting evidence or functional related information for them.

Materials and Methods

Data. We integrated four PPI data sets of *A. thaliana* in this study. Two of them were downloaded from the TAIR database (http://www.arabidopsis.org/) and the IntAct database[22] (http://www.ebi.ac.uk/intact), respectively, and the other two data sets, consisting of computationally predicted PPIs, were obtained from *Arabidopsis* predicted interactome[23] and *Arabidopsis thaliana* Protein Interactome database.[24] We listed the data information in Supplemental Table S1 (Supporting Information).

To select the most reliable cell-wall synthesis related proteins as seeds in our prediction, we used 572 CWSR proteins, among which 571 are covered by both the CWN database[9] and the Purdue Cell Wall database[10] (http://cellwall.genomics.purdue.edu) and one CWSR protein (AT5G15630) was supported by published literature.[25]

Our microarray gene-expression data came from AtGenExpress[26] at the EBI site (ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/AFMX/E-AFMX-9/). We used the expression data of 63 samples from normal plant developmental stages among a total of 79 samples under different physiological or treatment conditions, after excluding data collected under conditions of special treatments or mutagenesis. This microarray data contains the expression data of 21 021 *A. thaliana* genes, including 497 known CWSR genes and 88 predicted CWSR candidates. Hence, the expression data covers 80 interaction pairs between 38 known CWSR genes and 70 candidates, and 5 pairs of proteins sharing common interaction partners between 5 known CWSR genes and 4 candidates.

For our coevolutionary analysis, we downloaded 597 bacteria genomes (version 11/19/2007) as well as their gene annotations from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria).

The *A. thaliana* subcellular localization data were downloaded from the *Arabidopsis* subcellular database SUBA.[27] If a protein was annotated to reside in multiple subcellular compartments, we assigned multiple locations to that protein. We downloaded the Carbohydrate-Active EnZymes database (CAZy)[12] released on 10/20/2008 from http://www.cazy.org/ in this study.

Constructing the Positive and Negative Training Sets for PPI Prediction. The positive set (PS) was composed of validated PPIs as reported in the published literature. To construct the negative set (NS), we collected all protein pairs in which one protein was annotated to be located in the plasma membrane and the other in the nucleus according to the GO annotation.[28] Although plasma membrane proteins could occasionally interact with nuclear proteins, we considered that such cases were extremely rare and hence could be ignored, and hence we essentially assumed that all such pairs would not have protein−protein interactions.

Integrating PPI Data Using a Naïve Bayesian Approach. For each of the four PPI data sets, we employed a naïve Bayesian approach[29] to derive a likelihood score for each PPI in different data sets, representing the probability of the two corresponding proteins interacting with each other. Then we took the maximum of the different likelihood scores for each predicted PPI based on different data sets as the final likelihood score for that PPI. The likelihood score is the probability of observing the values "*f*" in the predicted data set given that a pair of proteins interacts divided by the probability of observing the values "*f*" given that the pair does not interact. That is, $L = P(f \mid PS)/P(f \mid NS)$. We referred the reader to refs 19, 24, and 29 for the detailed calculation procedure for combining different evidence for PPI prediction. We deposited the integrated *A.thaliana* PPI network with likelihood score into our pDAWG database.[30]

Coevolutionary Analysis. We used the phylogenetic profile approach[31] to predict coevolutionary relationships among genes. To build the phylogenetic profile for each of 100 predicted CWSR candidate genes and the 597 known CWSR genes, we mapped each of them to its homologue in each of the 597 bacteria genomes (NCBI release of 11/19/2007) using Blast[32] with *E*-value $< 1 \times 10^{-4}$ as the cutoff.[33] The *phylogenetic profile* of a gene is defined as a binary string with the *i*th bit being 1 if the gene has a homologue in the *i*th (reference) genome among the 572, otherwise with the bit being 0. We defined the similarity level between each pair of phylogenetic profiles as the total number of their corresponding positions both having 1. To assess the significance of a particular

Table 1. Most Reliable 42 CWSR Candidate Proteins[a]

gene ID	annotation	Pfam domain	possible pathway/ role in cell-wall synthesis	independent evidence	CWSR protein interacting partners	EST	reference
AT3G55830	Transferase, transferring glycosyl groups	EXTL2	Glycosyl transferase family 64	Golgi	AT5G25310; AT5G20260	45	48
AT1G25460	Putative dihydroflavonol 4-reductase[b]	Epimerase	Related to lignin and flavonoid pathways	Co-evolution (gene neighborhood; phylogenetic profile)	AT1G17890	2[c]	NA
AT5G24318	**Hydrolase, hydrolyzing O-glycosyl compounds**	**Glyco_hydro_17, X8**	**Glycosyl hydrolase family 17 protein, involved in the assembly process of cell wall.**	**NA**	**AT2G39640**	**64**	**NA**
AT3G08980	**Signal peptidase I family protein**	**Peptidase_S24**	**Removing the signal peptide of enzymes involved in cell wall biosynthesis, targeting pathways**	**NA**	**AT1G05570**	**1**	64
AT1G24280	G6PD3 (glucose-6-phosphate dehydrogenase 3)	G6PD_N, G6PD_C	Glycoside hydrolase involved in cell wall reassembly process	Coexpression (positive) and coevolution (gene neighborhood; phylogenetic profile)	AT1G73250	34	65
AT5G25760	PEX4 (Peroxin 4); ubiquitin-protein ligase	UQ_con	–	Coexpression (positive)	AT1G05570	39	66,67
AT5G41700	UBC8 (Ubiquitin conjugating enzyme 8)	UQ_con	–	Coexpression (positive)	AT1G05570	326	66,68
AT3G01090	AKIN10 (*Arabidopsis* SNF1 kinase homologue 10)	Protein kinase domain, UBA/TS-N domain, Kinase associated domain 1	–	Coexpression (positive)	AT1G05570	81	69
AT5G24940	Putative protein phosphatase 2C	PP2C	Protein phosphatase 2C	Coexpression (negative)	AT1G05570	3	70
AT3G19980	ATFYPP3 (protein serine/threonine phosphatase)	Metallophos	–	Coexpression (negative)	AT2G39640	34	71,72
AT4G03560	ATTPC1 (two-pore channel 1); calcium channel/ voltage-gated calcium channel	Two Ion transport protein domains, one EF hand domain	Endocytosis process of secretary pathway	Coexpression (positive)	AT1G05570	97	73
AT1G17070	D111/G-patch domain-containing protein	G-patch domain, Tuftelin interacting protein 11	Secretion and targeting pathway	Coexpression (positive)	AT1G72990	30	74
AT3G18430	Calcium-binding EF hand family protein	EF hand	Endocytosis, involved in the secretion of polysaccharide-laden vesicles to their extracellular location	Coexpression (positive)	AT1G05570	30	75
AT5G54750	Transport protein particle (TRAPP) component Bet3, putative	TRAPP_Bet3	Secretion and targeting pathway (transporter)	Coexpression (positive)	AT1G05570	64	76
AT2G16950	Protein transporter	Four HEAT repeat domains and one Importin-beta N-terminal domain	Secretion and targeting pathway	Coexpression (negative)	AT2G39640	32	77
AT2G01320	ABC transporter family protein	ABC_tran, ABC2_membrane	Secretion and targeting pathway	Coexpression (negative) and coevolution (phylogenetic profile)	AT3G29360	70	78
AT3G12110	ACT11 (ACTIN-11); structural constituent of cytoskeleton	Actin	Cytoskeleton-associated proteins involved in secretion and targeting pathways	Coexpression (positive)	AT1G11730	41	79
AT3G22845	**Emp24/gp25L/p24 protein-related**	**EMP24_GP25L**	**Vesicle trafficking process in secretion and targeting pathways**	**NA**	**AT1G05570**	**44**	80

Table 1. Continued

gene ID	annotation	Pfam domain	possible pathway/ role in cell-wall synthesis	independent evidence	CWSR protein interacting partners	EST	reference
AT1G15130	Hydroxyproline-rich glycoprotein family protein	BRO1-like domain	Structural protein of cell wall	Coexpression (positive)	AT1G05570	44	NA
AT5G19680	Leucine-rich repeat family protein	NA	Glycosylphosphatidylinositol (GPI)-anchored proteins, involved in signaling and response mechanism	Coexpression (positive)	AT1G17890	28	NA
AT3G21340	**Leucine-rich repeat protein kinase, putative**	**Pkinase**	**Signaling and response mechanisms**	**NA**	**AT2G44540**	**2**	81
AT3G59510	**Leucine-rich repeat family protein**	**LRRNT_2**	**GPI-anchored proteins, involved in the signaling and response mechanisms**	**NA**	**AT1G30620**	**1**	82
AT1G01960	EDA10 (embryo sac development arrest 10); guanyl-nucleotide exchange factor	Domain of unknown function (DUF1981), Sec7 domain	–	Coexpression (positive)	AT1G05570	53	83
AT3G09550	Ankyrin repeat family protein	Three Ankyrin repeat domains	–	Coexpression (negative)	AT5G44480; AT4G12250	3	84
AT4G35450	AKR2 (Ankyrin repeat-containing protein 2); protein binding	Two Ankyrin repeat domains	–	Coexpression (negative) and coevolution (phylogenetic profile)	AT5G59290	331	85
AT1G03000	PEX6 (PEROXIN6); ATPase	AAA, AAA	–	Coexpression (negative)	AT5G24090	16	86
AT1G04510	Transducin family protein/WD-40 repeat family protein	Four WD40 domains,Prp19		Coexpression (positive)	AT1G78570	33	87
AT1G54080	UBP1A; mRNA 3′-UTR binding	3 RRM_1 domains	Nucleotide-sugar interconversion pathways	Coexpression (positive)	AT3G53520; AT4G00110	179	88
AT1G63210	RNA binding/hydrolase, acting on ester bonds	NA	Nucleotide-sugar interconversion pathways	Coexpression (positive) and coevolution (phylogenetic profile)	AT1G12780	0	89
AT1G68980[d]	Pentatricopeptide (PPR) repeat-containing protein	2 PPR domains	–	Coexpression (positive)	Having exactly the same interacting partners as CWSR protein: AT3G06440	6	NA
AT1G72340	Eukaryotic translation initiation factor 2B family protein/eIF-2B family protein	IF-2B	Reassembly	Coexpression (positive)	AT2G39640	49	NA
AT2G43770	Transducin family protein/WD-40 repeat family protein	7 WD40 domains	Nucleotide-sugar interconversion pathways	Co-evolution (phylogenetic profile)	AT2G28760	27	90
AT2G47470	ATPDIL2−1/MEE30/UNE5 (PDI-LIKE 2−1, maternal effect embryo arrest 30, unfertilized embryo sac 5); thiol-disulfide exchange intermediate	2 Thioredoxin domains	Nucleotide-sugar interconversion pathways	Co-evolution (phylogenetic profile)	AT5G66280; AT1G73250	299	91
AT3G09440	Heat shock cognate 70 kDa protein 3 (HSC70−3) (HSP70−3)	HSP70	Nucleotide-sugar interconversion pathways	Co-evolution (phylogenetic profile)	AT5G28840	162	92
AT4G26840	SUM1 (SMALL UBIQUITIN-LIKE MODIFIER 1)	Ubiquitin	Nucleotide-sugar interconversion pathways; Assembly, Architecture, and Growth	Coexpression (positive; negative)	AT1G02460; AT5G39320	192	93
AT4G29880	Protein binding	NA	Nucleotide-sugar interconversion pathways	Co-evolution (phylogenetic profile)	AT3G51160	1	94

Table 1. Continued

gene ID	annotation	Pfam domain	possible pathway/ role in cell-wall synthesis	independent evidence	CWSR protein interacting partners	EST	reference
AT4G34460	AGB1 (GTP BINDING PROTEIN BETA 1); nucleotide binding	5 WD40 domains	Nucleotide-sugar interconversion pathways; Assembly, Architecture, and Growth	Subcellular localization (Golgi)	AT5G28840; AT5G24090	93	95
AT4G36960	RNA recognition motif (RRM)-containing protein	3 RRM_1 domains	Nucleotide-sugar interconversion pathways	Coexpression (positive)	AT4G20460	27	96
AT5G17770	ATCBR (NADH:CYTOCHROME B5 REDUCTASE 1)	NAD_binding_1, FAD_binding_6	Polysaccharide synthesis	Coexpression (positive)	AT1G05570	188	97
AT5G22060	ATJ2 (*Arabidopsis thaliana* DnaJ homologue 2)	DnaJ_CXXCXGXG, DnaJ, DnaJ_C	Polysaccharide synthesis	Coexpression (positive)	AT1G05570	107	98
AT5G41180[d]	Leucine-rich repeat protein kinase, putative	LRRNT_2, Pkinase_Tyr	Signaling and Response Mechanisms	Coexpression (positive)	Having exactly the same interacting partners as CWSR protein: AT3G14790	5	99
AT5G59210	Myosin heavy chain-related	NA	Nucleotide-sugar interconversion pathways	Coexpression (positive)	AT1G02000	26	100

[a] These candidates are not documented in CWN[9] and the Purdue Cell Wall database,[1] and some of them (AT3G55830 and AT5G24318) have consistent annotations in CAZy.[12] The proteins in bold are classified to be reliable predictions based on their domain architectures and functional annotations instead of independent evidence, while the others have independent evidence in addition to protein−protein interaction data. We retrieved the Number of Expressed Sequences Tags (EST) from TAIR. The proposed possible pathway/role in cell-wall synthesis was based on the candidates' Pfam domains, functional annotation or interacting CWSR partner's role in cell-wall synthesis. We used the Hmmpfam program[101] to search a protein against the pfam profile (version 21.0) and assigned the best non-overlapped Pfam domains matching with *E*-value $< 1 \times 10^{-2}$ to the query protein. NA, not available. [b] Annotation of AT1G25460 is based on our functional analysis and the others are obtained from TAIR. [c] That number of EST was obtained from NCBI Blast search against the ESTdb. [d] These candidates were identified by sharing common interacting partners with known CWSR proteins.

similarity level, we used the following to calculate the *P*-value for achieving a particular similarity:

$$\mathrm{Prob}_n(k, d) = \sum_{k=k_0}^{n} \left(\sum_{d=d_0}^{k} p_b^d (1 - p_b)^{k-d} \binom{k}{d} \right) p_a^k (1 - p_a)^{n-k} \binom{n}{k}$$

where n is the number of reference genomes (here $n = 597$), k is the number of corresponding positions in the two phylogenetic profiles having identical values, and d is the similarity level; and p_a and p_b are the probabilities for having k and d such positions as defined above. So $p_a = x^2 + (1 - x)^2$ and $p_b = x^2$, where x is the probability of any one protein having homologues in a reference genome. We used a cutoff *P*-value $< 1 \times 10^{-55}$ to identify two coevolved proteins, which was selected based on our preliminary study.

Phylogeny Reconstruction and Analysis. To perform the needed phylogeny reconstruction in our study, we performed multiple sequence alignments (MSAs) using the MAFFT program[34] with the L-INS-I model, which was considered to be one of the most accurate MSA methods.[35] Maximum likelihood trees were built using PhyML[36] using the JTT model, 100 replicates of bootstrap analyses, estimated proportion of invariable sites, four rate categories, estimated gamma distribution parameter, and optimized starting BIONJ tree.

Results and Discussion

Computational Prediction Procedure of CWSR Proteins. We have implemented and employed the following computational procedure for prediction of proteins that are possibly involved in plant cell-wall synthesis but are not included in Cell Wall Navigator[9] and Purdue Cell Wall database.[10] The key steps of the procedure are outlined in Figure 1.

The first step of the procedure is to construct a global PPI network, based on available experimental and computationally predicted PPI data, which are noisy, using a Naïve Bayesian approach (see Materials and Methods). We used four sets of public PPI data of *A. thaliana* to build the network and used 572 cell-wall synthesis proteins of *A. thaliana* that are present in both the CWN and the Purdue Cell Wall databases as the seeds for our prediction. The procedure then identifies subnetworks (also called *clusters* depending on the context) from the global PPI network, each of which contains mostly proteins working in the same biological process among all annotated proteins and is highly enriched with CWSR proteins based on Gene Ontology (GO).[28] Then it predicts proteins found in such subnetworks as candidate CWSR proteins. The rationale is that since the majority of the annotated proteins in each such subnetwork work in the same biological process and most of them are CWSR proteins, it is likely that the remaining (uncharacterized) proteins also work in the same or highly related biological process, and hence are related to cell-wall synthesis. Besides, we also predict a protein as a candidate, if it either interacts with *isolated CWSR proteins* (*Isolated CWSR proteins* are defined as those that do not form clusters/subnetworks with other known CWSR proteins and only interact with uncharacterized proteins in our global PPI network. Each such interaction network is labeled as *discrete_X* where *X* is the index in our webpage.) in the global PPI network or shares exactly the same interacting partners with a known CWSR protein. Our predictions are validated based on known CWSR proteins, as well as against (independent) experimental

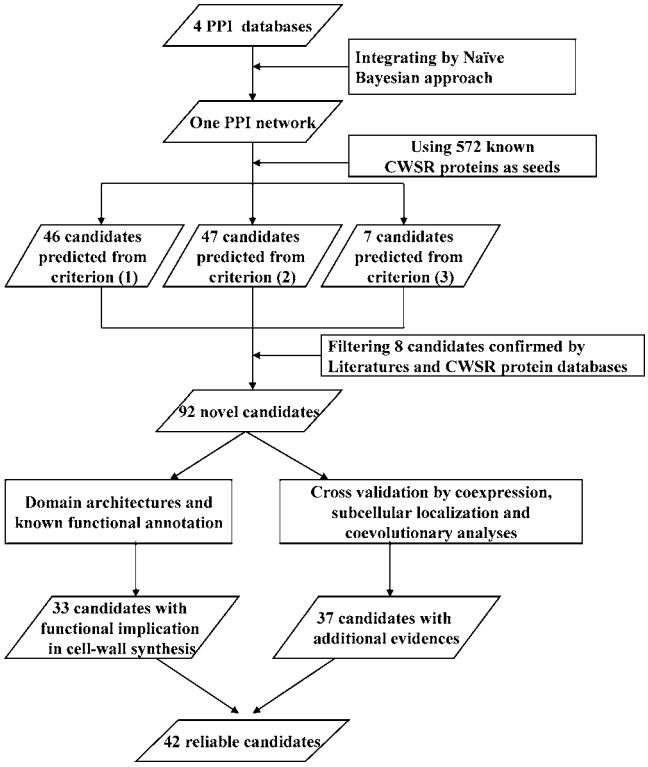

Figure 1. Flowchart for identifying novel CWSR proteins. The three criteria for identification are: (1) the candidates are part of a CWSR protein-containing subnetwork; (2) the candidates interact with isolated CWSR proteins; and (3) the candidates share exactly the same interacting partners with some CWSR protein.

data and associated computational analyses. Overall, 42 out of 100 candidate CWSR proteins (Figure 1 and Table 1) were considered to be the most reliable predictions, because they have independent supporting evidence or have strong functional relevance to cell wall synthesis. As the last step, the procedure gives functional analyses of the predicted cell-wall proteins.

Computational Prediction of a Global PPI Network. We have constructed a global PPI network for *A. thaliana*, based on four sets of protein−protein interaction data (see Materials and Methods). In this network, 9.1% of the PPIs have experimental evidence (from The *Arabidopsis* Information Resource (TAIR) and IntAct database[22] and the remaining PPI were predicted based on computational biological approaches.[23,24] For each individual PPI in the network, we assign a likelihood score to measure the confidence level of that PPI link, based on the reliability scores associated with the four underlying data sets (see Materials and Methods). The overall likelihood score distribution for the whole network is given in Supplemental Figure S1 (Supporting Information). To assess the effectiveness of this likelihood score as a confidence indicator for our predicted PPIs, we have collected two additional sets of protein pairs, one containing 4085 experimentally validated PPIs in *A. thaliana* as the positive set and the other containing 484 806 protein pairs that will not have interactions as the negative set (see Materials and Methods). The likelihood scores of PPIs in the positive set are the largest among all scores whereas about 99.9% of PPIs in the negative set have the scores equal to zero. In addition, Supplemental Figure S1 (Supporting Information) shows that all the interactions have likelihood scores larger than 29, far higher than the scores for the negative set. All these indicate that our likelihood score provides a good measure for the confidence of our PPI predictions; thus, it is reasonable to use all predicted interactions from different databases in construction of our global PPI network.

The resulting PPI network consists of 45 058 interactions among 13 347 proteins, covering ~50% of the ~27 000 protein-coding genes in *A. thaliana* and also covering 340 out of the 572 known CWSR proteins. This network is not completely connected, consisting of 1617 connected subnetworks, of which 1113 consists of only one pair of interacting proteins in each subnetwork, and the largest one contains 8798 proteins. The detailed information of the predicted network is given in Supplementary Table S1 (Supporting Information). Our analysis indicates that this PPI network follows a truncated power-law distribution (data not shown), which is consistent with previous findings.[37]

PPI Subnetworks Containing Known CWSR Proteins. Among all the 1617 connected subnetworks derived in the above section, 89 contain at least one known CWSR protein, and the largest one contains 75 CWSR proteins. We have made the following observations with respect to the 89 CWSR protein-containing subnetworks.

A. The size distribution of the 89 subnetworks is given in Supplemental Figure S2(a) (Supporting Information), and the percentage of known CWSR proteins within each subnetwork is shown in Supplemental Figure S2(b) (Supporting Information). Among all CWSR protein-containing subnetworks, either 100% of their annotated proteins are CWSR proteins or at least 50% of the proteins are known CWSR proteins, except for the largest one as we can see from Supplemental Figure S2(b);

B. For 65 of the 89 subnetworks excluding the largest one, all proteins with GO biological-process annotations[28] in each of these subnetworks share one common biological process. For four of the remaining 23 subnetworks, all proteins with GO biological-process annotations in each subnetwork are in two related biological processes. One of the remaining subnetworks contains nine xyloglucan endotransglycosylase/hydrolase proteins, seven of which are involved in a glucan metabolic process, and the other two are relevant to cell wall biogenesis and gibberellic acid mediated signaling, respectively; we note that these three biological processes are related.[38,39] The remaining 18 subnetworks do not have GO annotations;

C. A total of 25 different biological processes are covered by the 88 CWSR protein-containing subnetworks (excluding the largest subnetwork), indicating that some subnetworks share common or related biological processes. Those subnetworks are probably connected when more complete PPI data becomes available. Then in the best scenario, proteins in different subnetworks may in general work in different biological processes;

D. The largest subnetwork contains 75 known CWSR proteins out of its 8798 proteins. Sixty-six of the 75 CWSR proteins fall into 10 natural clusters (subsubnetworks), each containing CWSR proteins and a few additional proteins, mostly from the same biological process. The remaining 9 CWSR proteins do not link to any other known CWSR proteins, but link with a number of uncharacterized proteins. Hence these 9 known CWSR proteins are called *isolated CWSR proteins*. Detailed biological processes associated with the 89 subnetworks are given in Supplementary Table S2 (Supporting Information); and

E. We found cases where two known CWSR proteins share exactly the same interacting partners in the global PPI network,

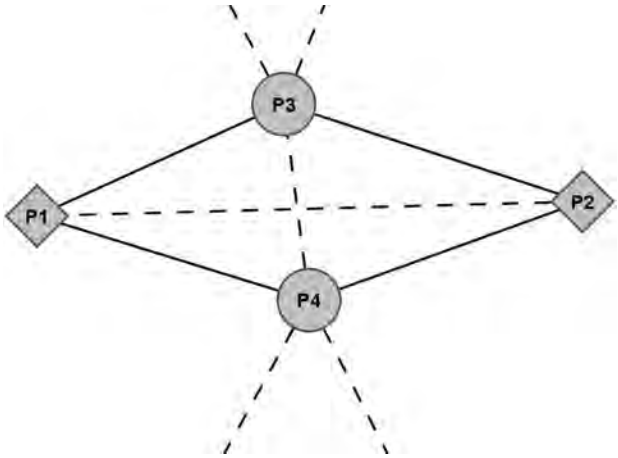

Figure 2. Illustration of two proteins P1 and P2 sharing exactly the same interacting partners P3 and P4. P1 and P2 may or may not interact with each other. If P1 is a known CWSR protein, then P2 is also likely to be related to cell-wall synthesis. The known interactions are shown as solid lines while the possible interactions are shown as dashed lines.

and they do not have any other interacting partners (see Figure 2). There are 175 such cases involving 107 CWSR proteins. Mostly, two CWSR proteins in one such case share a common or related biological process, which is consistent with a previous study in yeast.[40]

All 88 of the predicted CWSR subnetworks and the CWSR clusters in the largest subnetwork are given at http://csbl.bmb.uga.edu/~zhouchan/CellWallProtein. Supplementary Table S3 (Supporting Information) gives the detailed information about the percentages of the CWSR proteins within each of the 89 CWSR protein-containing networks.

Prediction of New CWSR Proteins. We predict a protein to be a candidate CWSR protein if it meets one of the following criteria: (1) it is part of a CWSR protein-containing subnetwork in which the majority of the annotated proteins share one common biological process or are in related biological processes; (2) it interacts with an isolated CWSR protein; or (3) it shares exactly the same interacting partners with a known CWSR protein (see Figure 2). Note that these three criteria are complementary with each other for novel CWSR protein prediction; and similar criteria have been successfully used in other studies for protein function prediction.[40,41] Using these prediction criteria, we have predicted 100 new candidate CWSR proteins (see Table 1), of which 46 are identified from the subnetworks (criterion (1)), 47 are identified from the direct interaction with isolated CWSR proteins (criterion (2)), and 8 are found because they share common interacting partners with known CWSR proteins (criterion (3)). Among all the candidates, one protein (AT3G55830) is identified by both criteria (1) and (3).

Assessing Predicted CWSR Proteins through Cross Validation. Several lines of evidence supported the 100 predicted CWSR proteins. While none of the supporting evidence can be considered as conclusive by itself except for the ones in the following section, the Supporting Information from an independent source does boost our confidence of the predicted CWSR proteins.

Predicted CWSR Proteins with Information from Literature and Public Databases. One predicted CWSR protein (AT1G05560) has been experimentally confirmed to be a UDP-glucose transferase involved in cell-wall synthesis,[42] which was not included in our seed CWSR protein, and thus the information was not available to us at the time we made the prediction. Seven of the predicted CWSR proteins are confirmed directly by one of the two public cell-wall databases, which are absent from our seed list of CWSR proteins since they were not included in one of the two cell-wall gene databases that we used. These missing CWSR proteins were treated like other uncharacterized proteins in our prediction. Among the seven candidate proteins confirmed by public cell wall databases (in red in Supplementary Table S5, Supporting Information), five of them (two GH17 proteins: AT1G11820 and AT2G05790; one GH28 protein: AT1G17150; one expansin-like B1 precursor: AT4G17030; and one cellulose synthase-like D2: AT5G16910) are in the CWN database but not in the Purdue database, while two (one calcium-binding EF hand family protein: AT1G21630; associated protein 19: AT2G17380) are in the Purdue database but not in the CWN database.

Predictions Supported by Coexpression Data. It is generally known that interacting proteins are more likely to be coexpressed presumably to maintain the proper stoichiometry among interacting partners.[43] We used the Spearman rank correlation test to check whether the 92 predicted CWSR proteins are coexpressed or not with their interacting CWSR partners (or their corresponding CWSR proteins which share common interacting proteins with the candidates). Our coexpression analysis indicates that 23 out of the 92 predicted CWSR proteins have positively correlation coefficients with their interacting CWSR partners, and 10 are negatively correlated with their interacting CWSR partners (Supplementary Table S4, Supporting Information), with P-value <0.05. Interestingly, one predicted CWSR protein (small ubiquitin-like modifier 1, AT4G26840) is positively and negatively correlated with two different CWSR proteins, respectively. Hence, we consider that these 32 predicted CWSR proteins have supporting evidence from gene expression data. Further details are given in the Supplementary Table S4 (Supporting Information).

Predictions Supported by Subcellular Localization Analysis. We have checked if our predicted interacting CWSR proteins are colocalized in the same subcellular localization as true interacting proteins should be.[44] We have examined our predictions against experimentally validated subcellular localization data for *A. thaliana* proteins from the SUBA database.[27] Out of the 92 candidate CWSR proteins, 33 are included in SUBA, among which 20 have their interacting partners also in the SUBA database. Among these 20 interacting pairs, two candidates (AT3G55830 and AT4G34460) are found to reside in Golgi (in purple in Supplementary Table S5, Supporting Information). As a control, we estimated the probability for two randomly selected proteins sharing a common subcellular location using SUBA data, which is 0.016 (see Supporting Information). Hence our above supporting data is statistically significant. So overall, two candidate CWSR proteins (AT3G55830 and AT4G34460) have supporting evidence from subcellular localization prediction. We believe that the low coverage of our predicted CWSR proteins by the SUBA database is due to the very limited data that the database has.

Predictions Supported by Coevolutionary Analysis. Proteins that have coevolved can provide another piece of supporting evidence for proteins working in the same cellular process. We have carried out the following coevolutionary analysis among the predicted CWSR proteins and their interacting CWSR partners. We first mapped all the predicted CWSR proteins

along with their interacting CWSR partners to their orthologs in bacterial genomes, and checked if the orthologs of the interacting pairs share common operons since operon-sharing proteins generally work in the same cellular process.[24,45]

We found that two pairs of candidate proteins (AT1G25460 and AT1G24280) and their interacting CWSR partners (GER2, AT1G17890 and GER1, AT1G73250, respectively) have their bacterial orthologs in the same operons (see Supplementary Table S6 and Methods, Supporting Information). As a control, we estimated that the probability for two randomly selected proteins having orthologous genes in a bacterium which localize in the same operon is 0.005378, indicating that our operon data is statistically significant. So this information provides supporting data for two predicted CWSR proteins (AT1G25460 and AT1G24280).

We have also used the phylogenetic profile method[31] to examine whether the predicted CWSR proteins and their interacting CWSR partners (or CWSR proteins sharing common interacting proteins) have coevolved. The basic idea of the approach is that genes whose homologues are of copresence and coabsence in a number of genomes, that is, with highly similar phylogenetic profiles, are in general functionally associated.[31] Nine out of the 92 CWSR candidates were found to have similar phylogenetic profiles with their CWSR interacting partners with P-values $< 1 \times 10^{-55}$ (data in Supplementary Table S5, Supporting Information). These nine predicted CWSR proteins include the two proteins (AT1G25460 and AT1G24280) validated by the above gene neighborhood method. Overall, nine candidates were supported to be related to cell-wall biosynthesis by coevolutionary evidence.

By combining all the above analyses, 37 out of 92 predicted CWSR proteins have additional evidence to support their possible roles in cell wall biosynthesis.

Functional Prediction of CWSR Candidates. To examine the specific roles of the predicted CWSR proteins in cell-wall biosynthesis, we have carried out a functional assignment based on protein structure and phylogeny information for two out of the 37 predicted CWSR proteins with supporting evidence as discussed above. These two candidate CWSR proteins (AT3G55830 and AT1G25460) were found to interact only with CWSR proteins (see Figure 3 for the interacting subnetworks/clusters containing them). For the remaining predicted CWSR proteins, we have mapped them onto well-characterized Pfam domains to derive some level of new functional information.

Using a protein threading server LOMETS,[46] we found that the best structural template for AT3G55830 in PDB[47] is 1OMXA, a mouse murine α, 4-*N*-acetylhexosaminyltransferase (EXTL2). Phylogenetic analysis also suggests that AT3G55830 is closer to 1OMXA than to the other EXTL2 domain-containing proteins (Figure 4). 1OMXA is known to catalyze the transfer reaction of GlcNAc and GalNAc from the respective UDP-sugars to the nonreducing end of [glucuronic acid]beta1−3[galactose]beta1-*O*- naphthalenemethanol, an acceptor substrate analog of the natural common linker of various glycosylaminoglycans. All of these indicate that AT3G55830 has a possible role in transferring UDP-GlcNAc and UDP-GalNAc to acceptor substrate as a glycosyltransferase. This is consistent with the annotation of AT3G55830 in the CAZy database[12] as a potential member of GT64 as well as with a previous report.[48]

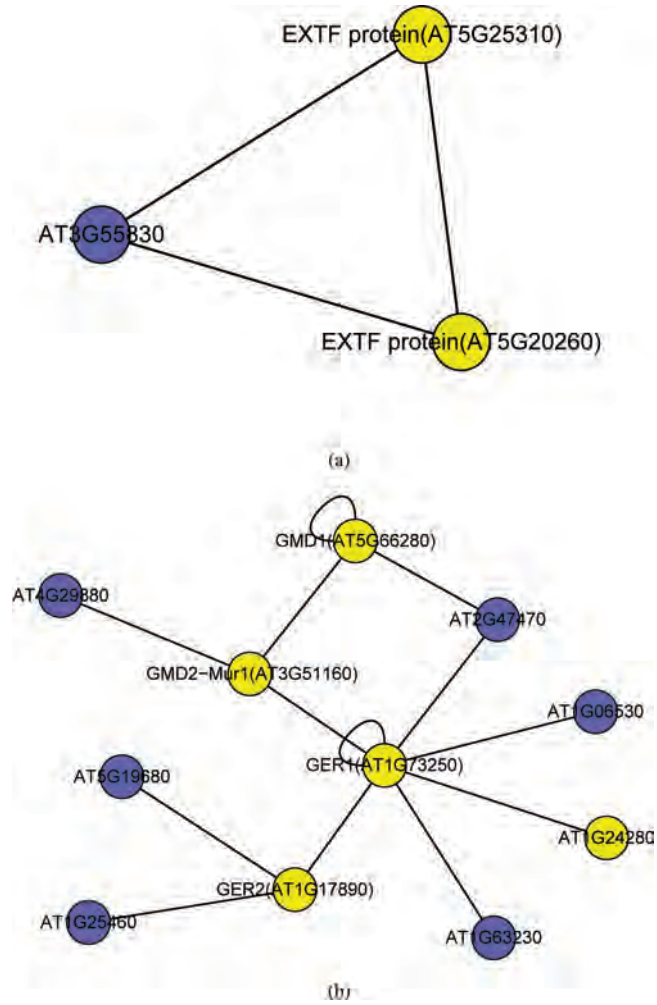

Figure 3. Two subnetworks that contain known CWSR proteins. (a) Subnetwork containing the candidate AT3G55830. (b) Cluster containing the candidate AT1G25460. Both of these two candidates only interact with known CWSR proteins. In both graphs, the yellow nodes represent the known CWSR proteins and the blue ones represent the uncharacterized candidate CWSR proteins.

We have modeled the tertiary structure of AT3G55830 based on 1OMXA (see Supporting Information). From the predicted structure, we found that AT3G55830 has a signature sequence motif for UDP-sugar-dependent glycosyltransferases, the DXD motif,[49,50] which aligned well with the DXD motif of 1OMXA. We refer the reader to the Supporting Information for the predicted structure of AT3G55830 and its other potential substrate binding sites.

The second CWSR candidate AT1G25460 has an epimerase domain and coevolved with GER2 (AT1G17890), suggesting that it may be an NADH-dependent oxidoreductase family protein. Using the LOMETS server, we have found that the best structural template for AT1G25460 is 2C29D, a grape dihydroflavonol-4-reductase (DFR).[51] In addition, our phylogenetic analysis suggests that AT1G25460 and 2C29D along with other *A. thaliana* DFR proteins are clustered together (Figure 5). All these suggest that AT1G25460 may be a DFR-like protein. DFR is the NADPH-dependent reductase that catalyzes a key step in the biosynthesis of anthocyanins and condensed tannins (two flavonoid classes),[51] where flavonoid is believed to be related to lignin synthesis.[52] Hence we suspect that AT1G25460 is related to lignin synthesis as a DFR-like protein.

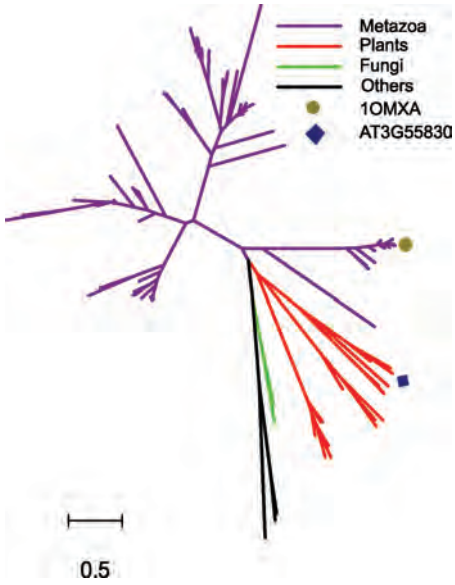

Figure 4. Maximum likelihood phylogeny of 215 GT64 domain-containing proteins. The sequences are the conserved Pfam GT64 domains of proteins in the nr database. Branches with different colors indicate proteins from different organismic groups. AT3G55830 and its closest homologue 1OMXA in PDB are highlighted with a blue diamond and brown circle, respectively.

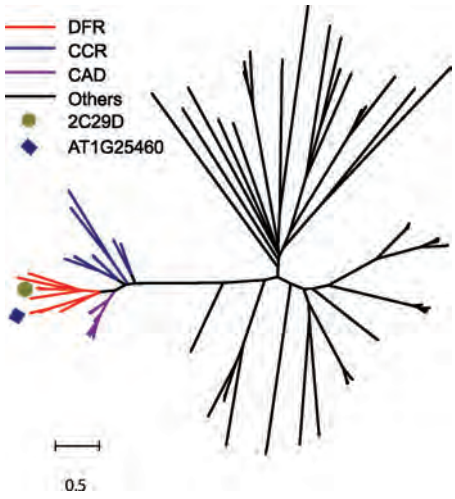

Figure 5. Maximum likelihood phylogeny of epimerase domain-containing proteins in *A. thaliana* and one grape DFR (PDB id: 2C29D). Different colors depict proteins from different families. AT1G25460 and 2C29D are highlighted with a blue diamond and brown circle, respectively.

For the other candidate CWSR proteins, we have carried out functional prediction based on their domain architectures and TAIR annotation. Consequently, 27 of 35 proteins with independent supporting evidence were assigned with putative functions (indicated on the fourth column of Table 1). And five additional candidate proteins (AT5G24318, AT3G08980, AT3G21340, AT3G22845, and AT3G59510) without additional supporting evidence were also given functional prediction (in bold in Table 1). The functional information of these proteins is listed in Table 1.

It is possible that some of our predicted CWSR proteins may not play any direct role in cell-wall synthesis; instead, they may have biological functions only indirectly related to cell-wall synthesis. On the other hand, some CWSR candidates with generic functions based on their names may actually play important roles in cell wall synthesis, such as Golgi transporters[5] as alteration in the function of such transporters may result in serious effects on cell wall structure.

Among the CWSR candidates without additional supporting data, five candidates (AT5G24318, AT3G08980, AT3G21340, AT3G22845, and AT3G59510) have domain information that indicates their possible role in cell-wall synthesis. AT5G24318 has both the GH17 catalytic domain and the X8 domain.[53] Hence it may be involved in the assembly of cell-wall biosynthesis. AT3G08980 has one Peptidase S24-like domain (Pfam acc: PF00717) and was annotated as signal peptidase I family protein in TAIR, so it is probably involved in removing the signal peptide of enzymes which are required in cell-wall biosynthesis.[54] AT3G21340 has a pkinase domain and is annotated as a putative leucine-rich repeat protein kinase in TAIR. Thus, it may function similarly like two recently discovered leucine-rich repeat receptor kinases that regulate cell-wall biosynthesis.[55] AT3G22845 has the EMP24_GP25L domain, suggesting that it may be a member of the emp24/gp25L/p24 family, which is known to participate in the transportation of cargo molecules from the endoplasmic reticulum to the Golgi apparatus.[56] So it may be a Golgi transporter,[5] involved in the vesicle trafficking process of secretion and targeting pathways in cell-wall biosynthesis as suggested in Purdue Cell Wall database.[1] The leucine rich repeats in AT3G59510 suggest it may be involved in signaling and response mechanisms like the leucine-rich repeat protein AT1G80080 in Purdue Cell Wall database.

Therefore, we predicted 92 candidate CWSR proteins and 42 of them have functional assignments at the molecular level, which we consider as reliable predictions. Among them, two (AT3G55830 and AT5G24318) are included in CAZy[12] as putative GT64 and GH17 proteins, respectively.

Biological Pathway Analyses on the 42 Most Reliable Candidates. We have examined the 42 proteins against annotated biological pathways in *Arabidopsis* Reactome,[57] derived from AraCyc[58] and KEGG,[59] which covers ~8% of the proteins in *A. thaliana*. Three of the 42 genes (AT1G24280, AT1G25460 and AT5G17770) are covered by these pathways. According to the assigned pathways, AT1G24280 is involved in the oxidative branch of the pentose phosphate pathway, which may affect cell-wall synthesis through decreasing the level of galactose and other related sugars.[60] AT1G25460 was found to be involved in anthocyanin biosynthesis, as part of the lignin synthesis. As discussed earlier, AT1G25460 may be a DFR-like protein, with DFR having been found to participate in anthocyanin biosynthesis.[61] Hence, the pathway information of AT1G25460 further supports our prediction of its role as DFR-like protein involved in lignin synthesis. AT5G17770 may be involved in aminosugars metabolism, related to cell-wall synthesis through amino sugars such as glucosamine (GlcN) and galactosamine (GalN). It is known that as part of plant cell wall, arabinogalactan proteins may contain GlcNAc or GlcN,[62] indicating that the previous knowledge supports our prediction of the roles of the three proteins in cell wall synthesis. It should be noted that some of the 42 proteins may have multiple functional roles as documented in TAIR, which is consistent with our general understanding about eukaryotic proteins.[63]

CWSR proteins are classified into six categories in the Purdue Cell Wall database as discussed earlier. The 572 CWSR proteins we used in this study as seeds cover only three of the six

categories (i.e., (i) substrate generation; (ii) polysaccharide synthases and glycosyl transferases, (iv) assembly, architecture and growth)), specifically, 35 in category (i), 163 in category (ii) and 374 in category (iv). This reflects our (low) knowledge level about these six categories of proteins. An interesting observation is that though the known CWSR proteins do not cover categories (iii), (v) and (vi) (i.e., (iii) secretion and targeting pathways, (v) differentiation and secondary wall formation, and (vi) signaling and response during the wall development and disassembly), our predictions do have proteins in categories (iii) and (vi) in addition to categories (i), (ii) and (iv). For example, our predictions include proteins involved in secretion and targeting pathways, which highlights the strength of protein−protein interaction based approach for CWSR protein prediction as these proteins interact with proteins in categories (i), (ii) and (iv).

We have developed a webpage (http://csbl.bmb.uga.edu/~zhouchan/CellWallProtein/) to display all the CWSR seed proteins and predicted candidates. It allows for searching related subnetwork/clusters and interacting partners of a protein by clicking on its ID. We plan to update this Web site on a regular basis as more and better experimental evidence for the interactome and protein functions become available.

Limitations. It has been previously estimated that over 2000 proteins are related to plant cell-wall synthesis.[1] We predicted 100 additional ones in this study. The remaining large gap could be due to a number of factors: (1) the current PPI network covers less than 50% of genes in *A. thaliana* so not all the known CWSR proteins have interaction information; and (2) in the current PPI data for *A. thaliana,* 86.18% of known CWSR proteins interact with only known CWSR proteins. While the predicted subnetworks are required to contain proteins involved in the same biological process/function roles, making our prediction more reliable, this requirement also limits our ability to identify proteins involved at the interfaces between different biological processes, and hence might have reduced the coverage of our prediction protocol. We expect that more CWSR proteins will be predicted as the PPI data have a higher coverage in terms of the protein−protein interactions relevant to cell wall synthesis, which may take place over the next few years.

Conclusion

Using the PPI network and known CWSR proteins, we predicted 100 candidate CWSR proteins, of which 92 are found to be not present in the current cell-wall protein databases. Among the 92 candidates, 42 are considered to be the most reliable predictions because independent supporting evidence or strong functional relevance information are available for them. Validations of our predictions are done based on analyses of publicly available microarray gene expression data, subcellular localization analyses and coevolutionary analyses. Although the CWSR proteins used as seeds in this study cover only three of the six functional categories of CWSR proteins, the newly identified candidate CWSR proteins cover two additional categories, showing the power of our PPI-based CWSR protein prediction.

Abbreviations: CWN, Cell Wall Navigator; CWSR, cell wall synthesis related; GH, glycosyl hydrolase; GT, glycosyl transferases; GRP, glycine-rich protein; PPI, protein−protein interaction.

Acknowledgment. We thank all of the members of the biofuel group, especially Wen-chi Chou, in the Computational Systems Biology Laboratory of UGA for helpful discussions and suggestions. We also thank Dr. Victor Olman for discussion about the statistical methods, Dr. Huiling Chen for helpful discussion about structure prediction, and Dr. Xizheng Mao for discussion about network analysis. The work of Y.Y., P.D., and Y.X. was supported, in part, by the DOE grant for the BioEnergy Science Center. The BioEnergy Science Center (BESC) is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. The work of P.D. and Y.X. was also supported by NSF (DBI-0354771, ITR-IIS-0407204, CCF-0621700, DBI-0542119). The work of C.Z. and Y.Y. was also supported in part by NSF/DEB 0830024.

Supporting Information Available: Predicted structures and structure-based alignment of two selected CWSR candidates and probability estimation for the subcellular localization and coevolution. Figures S1 and S2. Table S1: PPI data used to construct the interaction network. Table S2: (a)The biological processes associated with the 88 CWSR protein containing subnetworks and (b) the CWSR containing clusters in the largest subnetworks. Table S3: The percentage of the CWSR proteins within each of the 89 networks. Table S4: The significant coexpressed interaction pairs between candidate proteins and CWSR proteins, or pairs of proteins sharing common interacting partners. Table S5: The list of all candidate proteins and their additional information. Table S6: Pairs of candidate proteins and their interacting partners which coevolved from the orthologous genes within one common operon in bacteria. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Carpita, N.; Tierney, M.; Campbell, M. Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. *Plant Mol. Biol.* **2001**, *47* (1−2), 1–5.
(2) Gilbert, H. J.; Stalbrand, H.; Brumer, H. How the walls come crumbling down: recent structural biochemistry of plant polysaccharide degradation. *Curr. Opin. Plant Biol.* **2008**, *11* (3), 338–48.
(3) Knox, J. P. Revealing the structural and functional diversity of plant cell walls. *Curr. Opin. Plant Biol.* **2008**, *11* (3), 308–13.
(4) Popper, Z. A. Evolution and diversity of green plant cell walls. *Curr. Opin. Plant Biol.* **2008**, *11* (3), 286–92.
(5) Reyes, F.; Orellana, A. Golgi transporters: opening the gate to cell wall polysaccharide biosynthesis. *Curr. Opin. Plant Biol.* **2008**, *11* (3), 244–51.
(6) Geisler, D. A.; Sampathkumar, A.; Mutwil, M.; Persson, S. Laying down the bricks: logistic aspects of cell wall biosynthesis. *Curr. Opin. Plant Biol.* **2008**, *11* (6), 647–52.
(7) Rubin, E. M. Genomics of cellulosic biofuels. *Nature* **2008**, *454* (7206), 841–5.
(8) Chapple, C.; Ladisch, M.; Meilan, R. Loosening lignin's grip on biofuel production. *Nat. Biotechnol.* **2007**, *25* (7), 746–8.
(9) Girke, T.; Lauricha, J.; Tran, H.; Keegstra, K.; Raikhel, N. The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. *Plant Physiol.* **2004**, *136* (2), 3003–8, and discussion 3001.
(10) Yong, W.; Link, B.; O'Malley, R.; Tewari, J.; Hunter, C. T.; Lu, C. A.; Li, X.; Bleecker, A. B.; Koch, K. E.; McCann, M. C.; McCarty, D. R.; Patterson, S. E.; Reiter, W. D.; Staiger, C.; Thomas, S. R.; Vermerris, W.; Carpita, N. C. Genomics of plant cell wall biogenesis. *Planta* **2005**, *221* (6), 747–51.
(11) Perrin, R.; Wilkerson, C.; Keegstra, K. Golgi enzymes that synthesize plant cell wall polysaccharides: finding and evaluating candidates in the genomic era. *Plant Mol. Biol.* **2001**, *47* (1), 115–30.
(12) Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. The Carbohydrate-Active EnZymes

database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **2009**, *37* (Database issue), D233–8.
(13) Egelund, J.; Skjot, M.; Geshi, N.; Ulvskov, P.; Petersen, B. L. A complementary bioinformatics approach to identify potential plant cell wall glycosyltransferase-encoding genes. *Plant Physiol.* **2004**, *136* (1), 2609–20.
(14) Hansen, S. F.; Bettler, E.; Wimmerova, M.; Imberty, A.; Lerouxel, O.; Breton, C. Combination of several bioinformatics approaches for the identification of new putative glycosyltransferases in *Arabidopsis*. *J. Proteome Res.* **2009**, *8* (2), 743–53.
(15) Brown, D. M.; Zeef, L. A.; Ellis, J.; Goodacre, R.; Turner, S. R. Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **2005**, *17* (8), 2281–95.
(16) Persson, S.; Wei, H.; Milne, J.; Page, G. P.; Somerville, C. R. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (24), 8633–8.
(17) Ryusuke Yokoyama, K. N. Identification and characterization of Arabidopsis thaliana genes involved in xylem secondary cell walls. *J. Plant Res.* **2006**, *119*, 189–94.
(18) Mitchell, R. A.; Dupree, P.; Shewry, P. R. A novel bioinformatics approach identifies candidate genes for the synthesis and feruloylation of arabinoxylan. *Plant Physiol.* **2007**, *144* (1), 43–53.
(19) Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N. J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J. F.; Gerstein, M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **2003**, *302* (5644), 449–53.
(20) Parrish, J. R.; Gulyas, K. D.; Finley, R. L., Jr. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **2006**, *17* (4), 387–93.
(21) Miller, K. G.; Field, C. M.; Alberts, B. M.; Kellogg, D. R. Use of actin filament and microtubule affinity chromatography to identify proteins that bind to the cytoskeleton. *Methods Enzymol.* **1991**, *196*, 303–19.
(22) Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; Kohler, C.; Khadake, J.; Leroy, C.; Liban, A.; Lieftink, C.; Montecchi-Palazzi, L.; Orchard, S.; Risse, J.; Robbe, K.; Roechert, B.; Thorneycroft, D.; Zhang, Y.; Apweiler, R.; Hermjakob, H. IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.* **2007**, *35* (Database issue), D561–5.
(23) Geisler-Lee, J.; O'Toole, N.; Ammar, R.; Provart, N. J.; Millar, A. H.; Geisler, M. A predicted interactome for Arabidopsis. *Plant Physiol.* **2007**, *145* (2), 317–29.
(24) Cui, J.; Li, P.; Li, G.; Xu, F.; Zhao, C.; Li, Y.; Yang, Z.; Wang, G.; Yu, Q.; Shi, T. AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology. *Nucleic Acids Res.* **2007**, *36* (Database issue), D999–1008.
(25) Bischoff, V.; Cookson, S. J.; Wu, S.; Scheible, W. R. Thaxtomin A affects CESA-complex density, expression of cell wall genes, cell wall composition, and causes ectopic lignification in Arabidopsis thaliana seedlings. *J. Exp. Bot.* **2009**, *60* (3), 955–65.
(26) Schmid, M.; Davison, T. S.; Henz, S. R.; Pape, U. J.; Demar, M.; Vingron, M.; Scholkopf, B.; Weigel, D.; Lohmann, J. U. A gene expression map of Arabidopsis thaliana development. *Nat. Genet.* **2005**, *37* (5), 501–6.
(27) Heazlewood, J. L.; Verboom, R. E.; Tonti-Filippini, J.; Small, I.; Millar, A. H. SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res.* **2007**, *35* (Database issue), D213–8.
(28) Harris, M. A.; Clark, J.; Ireland, A. etc., The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32* (Database issue), D258–61.
(29) Rhodes, D. R.; Tomlins, S. A.; Varambally, S.; Mahavisno, V.; Barrette, T.; Kalyana-Sundaram, S.; Ghosh, D.; Pandey, A.; Chinnaiyan, A. M. Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* **2005**, *23* (8), 951–9.
(30) Mao, F.; Yin, Y.; Zhou, F.; Chou, W.-C.; Zhou, C.; Chen, H.; Xu, Y. pDAWG: an Integrated Database for Plant Cell-Wall Genes. *BioEnergy Res.* **2009**, *2*, 209–216.
(31) Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D.; Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (8), 4285–8.
(32) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–402.
(33) Rujan, T.; Martin, W. How many genes in Arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet.* **2001**, *17* (3), 113–20.
(34) Katoh, K.; Kuma, K.; Toh, H.; Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **2005**, *33* (2), 511–8.
(35) Nuin, P. A.; Wang, Z.; Tillier, E. R. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinform.* **2006**, *7*, 471.
(36) Guindon, S.; Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **2003**, *52* (5), 696–704.
(37) Ma, S.; Gong, Q.; Bohnert, H. J. An Arabidopsis gene network based on the graphical Gaussian model. *Genome Res.* **2007**, *17* (11), 1614–25.
(38) Gibeaut, D. M.; Carpita, N. C. Tracing Cell Wall Biogenesis in Intact Cells and Plants: Selective Turnover and Alteration of Soluble and Cell Wall Polysaccharides in Grasses. *Plant Physiol.* **1991**, *97* (2), 551–61.
(39) Lev-Yadun, S.; Beharav, A.; Di-nur, R.; Abbo, S. Gibberellic acid (GA) increases fibre cell differentiation and secondary cell-wall deposition in spring wheat (Triticum aestivum L.) culms. *Plant Growth Regul.* **1999**, *27* (3), 161–65.
(40) Samanta, M. P.; Liang, S. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (22), 12579–83.
(41) Sharan, R.; Ulitsky, I.; Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **2007**, *3*, 88.
(42) Hong, Z.; Zhang, Z.; Olson, J. M.; Verma, D. P. A novel UDP-glucose transferase is part of the callose synthase complex and interacts with phragmoplastin at the forming cell plate. *Plant Cell* **2001**, *13* (4), 769–79.
(43) Fraser, H. B.; Hirsh, A. E.; Wall, D. P.; Eisen, M. B. Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (24), 9033–8.
(44) Citovsky, V.; Lee, L. Y.; Vyas, S.; Glick, E.; Chen, M. H.; Vainstein, A.; Gafni, Y.; Gelvin, S. B.; Tzfira, T. Subcellular localization of interacting proteins by bimolecular fluorescence complementation in planta. *J. Mol. Biol.* **2006**, *362* (5), 1120–31.
(45) Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G. D.; Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (6), 2896–901.
(46) Wu, S.; Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **2007**, *35* (10), 3375–82.
(47) Sussman, J. L.; Lin, D.; Jiang, J.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr., D: Biol. Crystallogr.* **1998**, *54* (Pt 6 Pt 1), 1078–84.
(48) Singh, S. K.; Eland, C.; Harholt, J.; Scheller, H. V.; Marchant, A. Cell adhesion in Arabidopsis thaliana is mediated by ECTOPICALLY PARTING CELLS 1--a glycosyltransferase (GT64) related to the animal exostosins. *Plant J.* **2005**, *43* (3), 384–97.
(49) Pedersen, L. C.; Dong, J.; Taniguchi, F.; Kitagawa, H.; Krahn, J. M.; Pedersen, L. G.; Sugahara, K.; Negishi, M. Crystal structure of an alpha 1,4-N-acetylhexosaminyltransferase (EXTL2), a member of the exostosin gene family involved in heparan sulfate biosynthesis. *J. Biol. Chem.* **2003**, *278* (16), 14420–8.
(50) Wiggins, C. A.; Munro, S. Activity of the yeast MNN1 alpha-1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (14), 7945–50.
(51) Petit, P.; Granier, T.; d'Estaintot, B. L.; Manigand, C.; Bathany, K.; Schmitter, J. M.; Lauvergeat, V.; Hamdi, S.; Gallois, B. Crystal structure of grape dihydroflavonol 4-reductase, a key enzyme in flavonoid biosynthesis. *J. Mol. Biol.* **2007**, *368* (5), 1345–57.
(52) Markham, K. R.; Ryan, K. G.; Gould, K. S.; Rickards, G. K. Cell wall sited flavonoids in lisianthus flower petals. *Phytochemistry* **2000**, *54* (7), 681–7.
(53) Henrissat, B.; Davies, G. J. Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiol.* **2000**, *124* (4), 1515–9.
(54) van Roosmalen, M. L.; Geukens, N.; Jongbloed, J. D.; Tjalsma, H.; Dubois, J. Y.; Bron, S.; van Dijl, J. M.; Anne, J. Type I signal peptidases of Gram-positive bacteria. *Biochim. Biophys. Acta* **2004**, *1694* (1−3), 279–97.
(55) Xu, S. L.; Rahman, A.; Baskin, T. I.; Kieber, J. J. Two Leucine-Rich Repeat Receptor Kinases Mediate Signaling, Linking Cell Wall Biosynthesis and ACC Synthase in Arabidopsis. *Plant Cell* **2008**, *20* (11), 3065–79.

(56) Anantharaman, V.; Aravind, L. The GOLD domain, a novel protein module involved in Golgi function and secretion. *Genome Biol.* **2002**, *3* (5), research0023.
(57) Tsesmetzis, N.; Couchman, M.; Higgins, J.; Smith, A.; Doonan, J. H.; Seifert, G. J.; Schmidt, E. E.; Vastrik, I.; Birney, E.; Wu, G.; D'Eustachio, P.; Stein, L. D.; Morris, R. J.; Bevan, M. W.; Walsh, S. V. Arabidopsis reactome: a foundation knowledgebase for plant systems biology. *Plant Cell* **2008**, *20* (6), 1426–36.
(58) Mueller, L. A.; Zhang, P.; Rhee, S. Y. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.* **2003**, *132* (2), 453–60.
(59) Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28* (1), 27–30.
(60) Baxter, C. J.; Redestig, H.; Schauer, N.; Repsilber, D.; Patil, K. R.; Nielsen, J.; Selbig, J.; Liu, J.; Fernie, A. R.; Sweetlove, L. J. The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. *Plant Physiol.* **2007**, *143* (1), 312–25.
(61) Weisshaar, B.; Jenkins, G. I. Phenylpropanoid biosynthesis and its regulation. *Curr. Opin. Plant Biol.* **1998**, *1* (3), 251–7.
(62) van Hengel, A. J.; Tadesse, Z.; Immerzeel, P.; Schols, H.; van Kammen, A.; de Vries, S. C. N-acetylglucosamine and glucosamine-containing arabinogalactan proteins control somatic embryogenesis. *Plant Physiol.* **2001**, *125* (4), 1880–90.
(63) Albenne, C.; Canut, H.; Boudart, G.; Zhang, Y.; Clemente, H. S.; Pont-Lezica, R.; Jamet, E. Plant Cell Wall Proteomics: Mass Spectrometry Data, a Trove for Research on Protein Structure/Function Relationships. *Mol. Plant* **2009**, 1–13.
(64) Tripathi, L. P.; Sowdhamini, R. Cross genome comparisons of serine proteases in Arabidopsis and rice. *BMC Genomics* **2006**, *7*, 200.
(65) Wakao, S.; Benning, C. Genome-wide analysis of glucose-6-phosphate dehydrogenases in Arabidopsis. *Plant J.* **2005**, *41* (2), 243–56.
(66) Kraft, E.; Stone, S. L.; Ma, L.; Su, N.; Gao, Y.; Lau, O. S.; Deng, X. W.; Callis, J. Genome analysis and functional characterization of the E2 and RING-type E3 ligase ubiquitination enzymes of Arabidopsis. *Plant Physiol.* **2005**, *139* (4), 1597–611.
(67) Zolman, B. K.; Monroe-Augustus, M.; Silva, I. D.; Bartel, B. Identification and functional characterization of Arabidopsis PEROXIN4 and the interacting protein PEROXIN22. *Plant Cell* **2005**, *17* (12), 3422–35.
(68) Stone, S. L.; Hauksdottir, H.; Troy, A.; Herschleb, J.; Kraft, E.; Callis, J. Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. *Plant Physiol.* **2005**, *137* (1), 13–30.
(69) Hrabak, E. M.; Chan, C. W.; Gribskov, M.; Harper, J. F.; Choi, J. H.; Halford, N.; Kudla, J.; Luan, S.; Nimmo, H. G.; Sussman, M. R.; Thomas, M.; Walker-Simmons, K.; Zhu, J. K.; Harmon, A. C. The Arabidopsis CDPK-SnRK superfamily of protein kinases. *Plant Physiol.* **2003**, *132* (2), 666–80.
(70) Yazaki, J.; Shimatani, Z.; Hashimoto, A.; Nagata, Y.; Fujii, F.; Kojima, K.; Suzuki, K.; Taya, T.; Tonouchi, M.; Nelson, C.; Nakagawa, A.; Otomo, Y.; Murakami, K.; Matsubara, K.; Kawai, J.; Carninci, P.; Hayashizaki, Y.; Kikuchi, S. Transcriptional profiling of genes responsive to abscisic acid and gibberellin in rice: phenotyping and comparative analysis between rice and Arabidopsis. *Physiol. Genomics* **2004**, *17* (2), 87–100.
(71) Blakeslee, J. J.; Zhou, H. W.; Heath, J. T.; Skottke, K. R.; Barrios, J. A.; Liu, S. Y.; DeLong, A. Specificity of RCN1-mediated protein phosphatase 2A regulation in meristem organization and stress response in roots. *Plant Physiol.* **2008**, *146* (2), 539–53.
(72) Kim, D. H.; Kang, J. G.; Yang, S. S.; Chung, K. S.; Song, P. S.; Park, C. M. A phytochrome-associated protein phosphatase 2A modulates light signals in flowering time control in Arabidopsis. *Plant Cell* **2002**, *14* (12), 3043–56.
(73) Peiter, E.; Maathuis, F. J.; Mills, L. N.; Knight, H.; Pelloux, J.; Hetherington, A. M.; Sanders, D. The vacuolar Ca2+-activated channel TPC1 regulates germination and stomatal movement. *Nature* **2005**, *434* (7031), 404–8.
(74) Nakamura, M.; Satoh, T.; Tanaka, S.; Mochizuki, N.; Yokota, T.; Nagatani, A. Activation of the cytochrome P450 gene, CYP72C1, reduces the levels of active brassinosteroids in vivo. *J. Exp. Bot.* **2005**, *56* (413), 833–40.
(75) Day, I. S.; Reddy, V. S.; Shad Ali, G.; Reddy, A. S. Analysis of EF-hand-containing proteins in Arabidopsis. *Genome Biol.* **2002**, *3* (10), RESEARCH0056.
(76) Panjabi, P.; Jagannath, A.; Bisht, N. C.; Padmaja, K. L.; Sharma, S.; Gupta, V.; Pradhan, A. K.; Pental, D. Comparative mapping of Brassica juncea and Arabidopsis thaliana using Intron Polymorphism (IP) markers: homoeologous relationships, diversification and evolution of the A, B and C Brassica genomes. *BMC Genomics* **2008**, *9*, 113.
(77) Bollman, K. M.; Aukerman, M. J.; Park, M. Y.; Hunter, C.; Berardini, T. Z.; Poethig, R. S. HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development* **2003**, *130* (8), 1493–504.
(78) Sanchez-Fernandez, R.; Davies, T. G.; Coleman, J. O.; Rea, P. A. The Arabidopsis thaliana ABC protein superfamily, a complete inventory. *J. Biol. Chem.* **2001**, *276* (32), 30231–44.
(79) Huang, S.; An, Y. Q.; McDowell, J. M.; McKinney, E. C.; Meagher, R. B. The Arabidopsis ACT11 actin gene is strongly expressed in tissues of the emerging inflorescence, pollen, and developing ovules. *Plant Mol. Biol.* **1997**, *33* (1), 125–39.
(80) Carter, C.; Pan, S.; Zouhar, J.; Avila, E. L.; Girke, T.; Raikhel, N. V. The vegetative vacuole proteome of Arabidopsis thaliana reveals predicted and unexpected proteins. *Plant Cell* **2004**, *16* (12), 3285–303.
(81) Lall, S.; Nettleton, D.; DeCook, R.; Che, P.; Howell, S. H. Quantitative trait loci associated with adventitious shoot formation in tissue culture and the program of shoot development in Arabidopsis. *Genetics* **2004**, *167* (4), 1883–92.
(82) Kobe, B.; Kajava, A. V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **2001**, *11* (6), 725–32.
(83) Pagnussat, G. C.; Yu, H. J.; Ngo, Q. A.; Rajani, S.; Mayalagu, S.; Johnson, C. S.; Capron, A.; Xie, L. F.; Ye, D.; Sundaresan, V. Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. *Development* **2005**, *132* (3), 603–14.
(84) Bock, K. W.; Honys, D.; Ward, J. M.; Padmanaban, S.; Nawrocki, E. P.; Hirschi, K. D.; Twell, D.; Sze, H. Integrating membrane transport with male gametophyte development and function through transcriptomics. *Plant Physiol.* **2006**, *140* (4), 1151–68.
(85) Kuhlmann, M.; Horvay, K.; Strathmann, A.; Heinekamp, T.; Fischer, U.; Bottner, S.; Droge-Laser, W. The alpha-helical D1 domain of the tobacco bZIP transcription factor BZI-1 interacts with the ankyrin-repeat protein ANK1 and is important for BZI-1 function, both in auxin signaling and pathogen response. *J. Biol. Chem.* **2003**, *278* (10), 8786–94.
(86) Baker, A.; Charlton, W.; Johnson, B.; Lopez-Huertas, E.; Oh, J.; Sparkes, I.; Thomas, J. Biochemical and molecular approaches to understanding protein import into peroxisomes. *Biochem. Soc. Trans.* **2000**, *28* (4), 499–504.
(87) Zhang, Y.; Feng, S.; Chen, F.; Chen, H.; Wang, J.; McCall, C.; Xiong, Y.; Deng, X. W. Arabidopsis DDB1-CUL4 ASSOCIATED FACTOR1 forms a nuclear E3 ubiquitin ligase with DDB1 and CUL4 that is involved in multiple plant developmental processes. *Plant Cell* **2008**, *20* (6), 1437–55.
(88) Lambermon, M. H.; Simpson, G. G.; Wieczorek Kirk, D. A.; Hemmings-Mieszczak, M.; Klahre, U.; Filipowicz, W. UBP1, a novel hnRNP-like protein that functions at multiple steps of higher plant nuclear pre-mRNA maturation. *EMBO J.* **2000**, *19* (7), 1638–49.
(89) Zhang, Y.; Wu, Y.; Liu, Y.; Han, B. Computational identification of 69 retroposons in Arabidopsis. *Plant Physiol.* **2005**, *138* (2), 935–48.
(90) Gepstein, S.; Sabehi, G.; Carp, M. J.; Hajouj, T.; Nesher, M. F.; Yariv, I.; Dor, C.; Bassani, M. Large-scale identification of leaf senescence-associated genes. *Plant J.* **2003**, *36* (5), 629–42.
(91) Houston, N. L.; Fan, C.; Xiang, J. Q.; Schulze, J. M.; Jung, R.; Boston, R. S. Phylogenetic analyses identify 10 classes of the protein disulfide isomerase family in plants, including single-domain protein disulfide isomerase-related proteins. *Plant Physiol.* **2005**, *137* (2), 762–78.
(92) Reddy, V. S.; Ali, G. S.; Reddy, A. S. Genes encoding calmodulin-binding proteins in the Arabidopsis genome. *J. Biol. Chem.* **2002**, *277* (12), 9840–52.
(93) Colby, T.; Matthai, A.; Boeckelmann, A.; Stuible, H. P. SUMO-conjugating and SUMO-deconjugating enzymes from Arabidopsis. *Plant Physiol.* **2006**, *142* (1), 318–32.
(94) Forsthoefel, N. R.; Cutler, K.; Port, M. D.; Yamamoto, T.; Vernon, D. M. PIRLs: a novel class of plant intracellular leucine-rich repeat proteins. *Plant Cell Physiol.* **2005**, *46* (6), 913–22.
(95) Wang, S.; Narendra, S.; Fedoroff, N. Heterotrimeric G protein signaling in the Arabidopsis unfolded protein response. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (10), 3817–22.
(96) Kawaguchi, R.; Girke, T.; Bray, E. A.; Bailey-Serres, J. Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana. *Plant J.* **2004**, *38* (5), 823–39.
(97) Fukuchi-Mizutani, M.; Mizutani, M.; Tanaka, Y.; Kusumi, T.; Ohta, D. Microsomal electron transfer in higher plants: cloning and heterologous expression of NADH-cytochrome b5 reductase from Arabidopsis. *Plant Physiol.* **1999**, *119* (1), 353–62.

(98) Zhou, R.; Kroczynska, B.; Hayman, G. T.; Miernyk, J. A. AtJ2, an arabidopsis homolog of Escherichia coli dnaJ. *Plant Physiol.* **1995**, *108* (2), 821–2.
(99) Dievart, A.; Clark, S. E. Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. *Curr. Opin. Plant Biol.* **2003**, *6* (5), 507–16.
(100) Iwase, A.; Ishii, H.; Aoyagi, H.; Ohme-Takagi, M.; Tanaka, H. Comparative analyses of the gene expression profiles of Arabidopsis intact plant and cultured cells. *Biotechnol. Lett.* **2005**, *27* (15), 1097–103.
(101) Krogh, A.; Brown, M.; Mian, I. S.; Sjolander, K.; Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **1994**, *235* (5), 1501–31.

PR100249C