Supporting Material of "Identification of novel proteins involved in plant cell-wall synthesis based on protein-protein interaction data"

This supplemental material includes the following:

1. Predicted structures and structure-based alignment of two selected CWSR candidates

2. Probability estimation for the subcellular localization and co-evolution

1. Predicted structures and structure-based alignment of two selected CWSR candidates

We selected two candidate CWSR proteins (AT3G55830 and AT1G25460) that only interacted with CWSR proteins for further detailed functional analysis by structure-based comparison and phylogeny evidence. We firstly performed a structural prediction for the catalytic domain of AT3G55830 (from Tyr-74 to Met-323) and that of AT1G25460 (from Val-6 to Leu-320) using the I-TASSER server (Zhang, 2007; Zhang, 2008) (see Figure 1 and 3) and then studied the conserved motifs through structure-based sequence alignment (see Figure 2 and 4) to reveal their possible functional sites. From now on, we count the 1st residue in the catalytic domain of AT3G55830 and AT1G25460 as Tyr-1 and Val-1, respectively. And all of the other residues in AT3G55830 follow this counting rule in this paper.

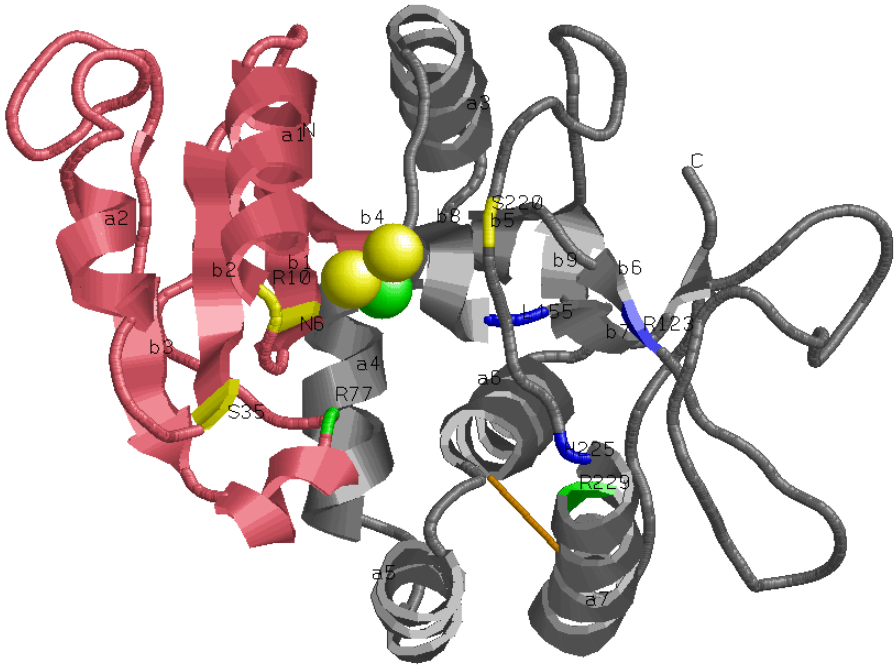
AT3G55830 was found to have a possible role in pectin biosynthesis of cell wall synthesis, as discussed in this paper and literature (Singh et al., 2005). Then, based on detailed structure-based alignment analysis (Figure 1 and 2), we further found that AT3G55830 has a signature sequence motif for UDP-sugar-dependent glycosyltransferases, the DXD motif (from Asp-93 to Asp-95) (Wiggins & Munro, 1998; Pedersen *et al.*, 2003), which aligned well with the DXD motif of 1OMXA (from Asp-114 to Asp116) (in triangle in Figure 2). It implies that this motif in AT3G55830 may interact with the Mn^{2+} and UDP-sugar donor as it in 1OMXA

(Pedersen et al., 2003). We also found several potential binding sites in AT3G55830 since they are conserved in AT3G55830 and 1OMXA (highlighted in color in Figure 1 and 2). These potential binding sites are Asn-6, Arg-10, Ser-35, Arg-77, Asp-93, Asp-94, Asp-95, Arg-123, Leu-155, His-225, Arg-229, Ser-220. Among them, Arg-77, Asp-93 and Arg-229 (in *green* in Figure 1 and their corresponding conserved binding sites in 1OMXA colored in *green* in Figure 2) may interact with the donor sugar of the UDP-donor. Arg-123, Leu-155 and His-225 (in *blue* in Figure 1, and their corresponding conserved binding sites in 1OMXA in *blue* in Figure 2) of AT3G55830 may be the acceptor substrate binding sites. Asn-6, Arg-10, Ser-35, Asp-94 and Asp-95 of AT3G55830 (in *yellow* in Figure 1, and their corresponding conserved binding sites in 1OMXA in *orange* in Figure 2) may involve in binding UDP.

Besides, two conserved Cys residues in AT3G55830 may form a disulfide bond and tether alpha-helices 6 and 7 (in *orange stick* in Figure 1, highlighted in star in Figure 2). Similar with the structure of 1OMXA，the UDP binding subdomain of AT3G55830 also consists of a Rossmann-like fold binding motif with four parallel beta-strands, ordered β3-β2-β1-β4, each separated by a helix (in *pink* Figure 1).

AT1G25460 has been proposed to be related to lignin pathway of cell wall synthesis, as a DFR-like protein in our paper. Based on the conserved binding sites in the structure-based alignment (Figure 3 and 4), we further found several potential binding sites in AT1G25460, i.e., Gly-3, Thr-5, Phe-7, Ile-8, Arg-28, Asp-29, Lys-35, Asp-55, Leu-56, Ala-76, Ser-77, Ser-120, Ser-121, Tyr-154, Lys-158, Pro-181, and Ser-196 (in color in Figure 3). Among them, Gly-3, Thr-5, Phe-7, Ile-8, Arg-28, Asp-29, Lys-35, Asp-55, Leu-56, Ala-76, Ser-77, Lys-158, Pro-181, and Ser-196 are potential NADP+ binding sites (in *yellow* in Figure 3, their conserved binding sites in 2C29D in *yellow* in Figure 4). Ser-120 and Ser-121 are potential substrate binding sites (in *dark-blue* in Figure 3, their conserved binding sites in 2C29D in *dark-blue* in Figure 4) The residue Tyr-154 is both NADP+ binding and substrate binding (in *green* in Figure 3, and its conserved site in 2C29D also in *green* in Figure 4). Besides, Ser-120, Tyr-154,

and Lys-158 are the conserved catalytic triad between AT1G25460 and 2C29D (highlighted in *ball* in Figure 3, and in *red circle* in Figure 4).



SFigure 1. Structure of the catalytic domain (Tyr-74-Met-323) of AT3G55830. The UDP-donor binding subdomain is colored in *pink* and the acceptor binding subdomain is in *grey*. The DXD motif for interacting with the Mn2+ metal ion and UDP-sugar donor is displayed in *ball*. We labeled the alpha-helix with the letter '*a*' and the beta-sheets with the letter '*b*'. The disulfide bond linking Cys-187 and Cys-232 is pictured in *orange stick* and it tethers alpha-helix 6 and 7. We colored the possible UDP-donor sugar binding sites in green (Arg-77, Asp-93, Arg-229) in *green* and the potential UDP-binding sites in *yellow* (Asn-6, Arg-10, Ser35, Asp-94, Asp-95, Ser-220). The residues interacting with the acceptor substrate are displayed in *blue* (Arg-123, Leu-155, His-225). The big pocket for UDP-sugar and acceptor substrate binding is surrounded by the DXD motif and eight beta-sheets (b3-b2-b1-b4 in the UDP-donor binding subdomain and b8-b5-b9-b6 in the acceptor binding subdomain) as well as alpha-helix 4,5,6,7. This structure is predicted by the current best rated

protein structure prediction tool I-TASSER (Zhang, 2007; Zhang, 2008) and is drawn using Rasmol (Sayle & Milner-White, 1995). Note: we count the 1st residue Tyr in this catalytic domain as No. 1 in this figure and whole paper.

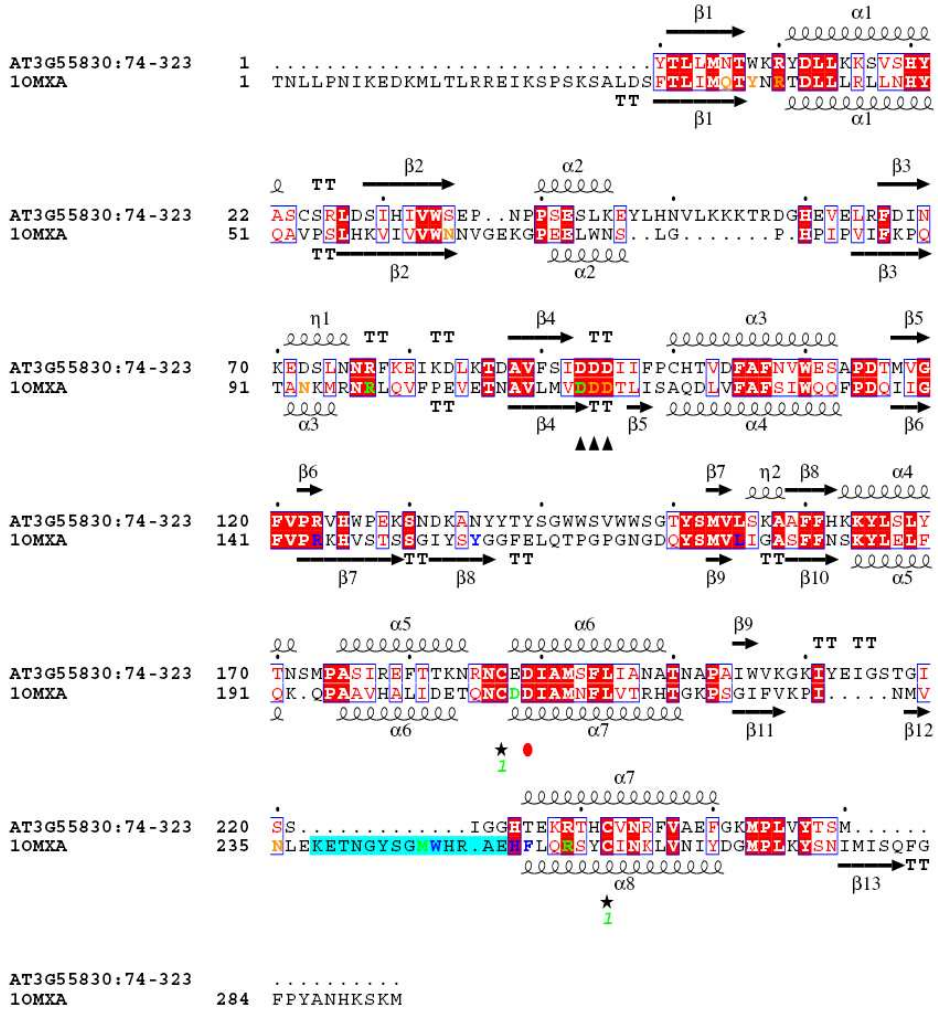


Figure 2. Structure-based sequence alignment between AT3G55830 and 1OMXA. Secondary structure elements of AT3G55830 are displayed in the top and secondary structure elements of 1OMXA (strand A of 1OMX, a mouse EXTL2) are displayed in the bottom (Pedersen et al., 2003). The sequence of 1OMXA was retrieved from NCBI and lacks the head 37 residues of the sequence used in the 1OMXA crystal structure literature (Pedersen et al., 2003). Here, we only aligned the catalytic domain of AT3G55830, ranging from residue 74 to residue 323, and thus take the residue 74

Deleted: Figure

in the full length of AT3G55830 as residue 1 when referring the position of residues. The residue pairs in blue box are these having more than 70% similarity of physico-chemical properties. Identity residue pairs are shown in red background. The DXD motif for UDP-sugar-dependent glycosyltransferases is marked with triangles. The two conserved Cysteines involved in a disulfide bond are labeled with black star in the bottom and Asp-209, a proposed catalytic residue in 1OMXA, is marked with red circle. Residues in *orange* are involved in binding UDP and those in *green* interact with the donor sugar of the UDP-donor in 1OMXA. Residues in *blue* are the acceptor substrate binding site in 1OMXA. The segment shaded with *yellow background* in 1OMXA forms a loop ordered upon acceptor substrate binding. This structure alignment was created using TM-align (Zhang & Skolnick, 2005) and most of residue pairs are less than 5 Å in the alignment. This graphical display using the ESPript webserver (Gouet et al., 1999).

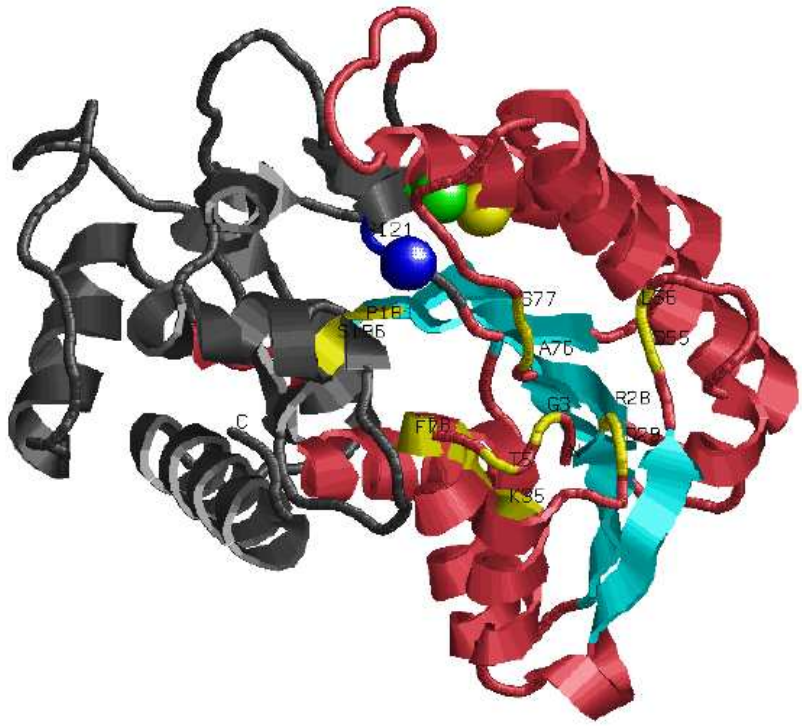


Figure 3. Structure of the catalytic domain (from Val-6 to Leu-320) of AT1G25460.

Deleted: Figure

The NADPH binding domain (N-terminal domian) is composed of a six-strand β-sheet (in *skyblue*) separated byα-helices (in *pink*). The C-terminal domain is colored in *grey*. The catalytic triad (Ser-120, Tyr-154, and Lys-158) is displayed in *ball*. We colored the potential NADP+ binding sites in *yellow* (Gly-3, Thr-5, Phe-7, Ile-8, Arg-28, Asp-29, Lys-35, Asp-55, Leu-56, Ala-76, Ser-77, Lys-158, Pro-181, and Ser-196) and the potential substrate binding sites in *dark-blue* (Ser-120 and Ser-121). The residue Tyr-154 is in *green* since it is both NADP+ binding and substrate binding. These binding sites are conserved ones which match the binding sites in their structure template 2C29D (highlighted in the same color scheme in Figure 4). This structure is predicted by the current best rated protein structure prediction tool I-TASSER (Zhang, 2007; Zhang, 2008) and is drawn using Rasmol (Sayle & Milner-White, 1995). Note: we count the 1st residue val in this catalytic domain as No. 1 in this figure and whole paper.

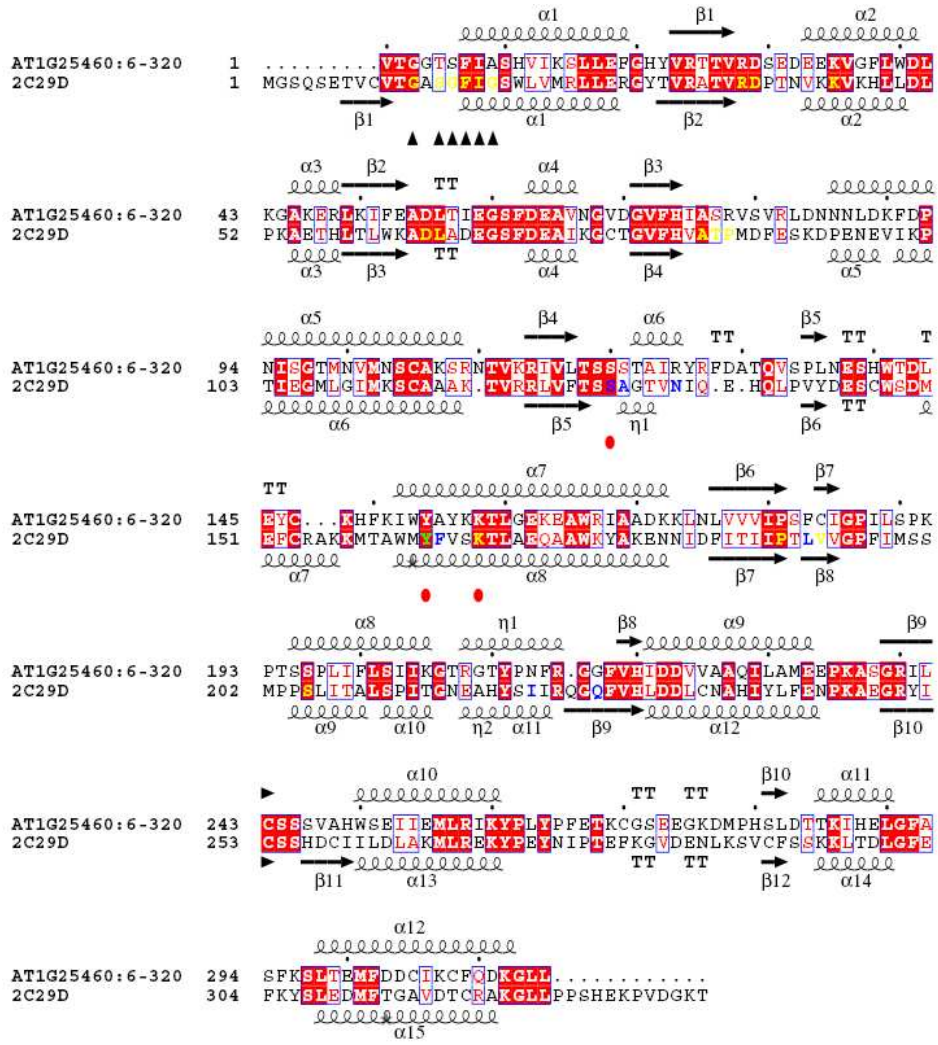


Figure 4. Structure-based sequence alignment between AT1G25460 and 2C29D. Secondary structure elements of AT1G25460 are displayed in the top and secondary structure elements of 2C29D (strand D of 2C29, a grape dihydroflavonol 4-reductase) are displayed in the bottom (Petit et al., 2007). Here, we only aligned the catalytic domain of AT1G25460, ranging from residue 6 to residue 320, and thus take the residue 6 in the full length of AT1G25460 as residue 1 when referring the position of residues. The glycine-rich motif constituting the NADPH binding sites in 2C29D is highlighted in triangles. The conserved catalytic triad (Ser-120, Tyr-154, and Lys-158) is marked in red circle. The blue box and red background have the same code as in the legend of Figure 2. The tools used for alignment and display are the same as in the

Deleted: Figure

legend of Figure 2.

2. Probability estimation for the subcellular localization and co-evolution

In the co-evolution analysis and subcellular localization analysis, we estimated the probability of two randomly selected proteins having the same subcellular localization or coevolved from one operon as a control, respectively.

According to the combinatorics theory and probability theory, the probability of two proteins P_1 and P_2 randomly selected from whole *Arabidopsis* proteins pool which have the same subcellular localization is:

$$\Pr ob(\mathrm{P_1 \text{ and } P_2 \text{ have the same subcellular localization} \mid \text{randomly selecting } P_1 \text{ and } P_2 \text{ in } Arabidopsis}) = \frac{\sum_{i=1}^{s} \binom{m_i}{1}\binom{m_i - 1}{1}/2}{\binom{N}{1}\binom{N-1}{1}/2}$$

, where m_i is the number of proteins residing in i-th subcellular component if we assigned a digital label to all of the subcellular components, i.e., nucleus, ribosome, Golgi, ER and so on, and N is the number of all proteins in *Arabidopsis*.

Similarly, the probability of two proteins P_1 and P_2 randomly selected from whole *Arabidopsis* proteins pool which coevolved from orthogous genes within one operon in bacteria genomes can be calculated from the following formula:

$$\Pr ob(\mathrm{P_1 \text{ and } P_2 \text{ coevolved from one operon in bacteria} \mid \text{randomly selecting } P_1 \text{ and } P_2 \text{ in } Arabidopsis}) = \frac{\text{number of protein pairs coevloved from one operon}}{\text{number of all protein pairs}}$$

, where the number of all protein pairs is $\binom{N}{1}\binom{N-1}{1}/2$ and the number of protein pairs coevolved from one operon was calculated based on the real data as described in Section 4.4.

References

Gouet P, Courcelle E, Stuart DI, Metoz F. 1999. Espript: Analysis of multiple sequence alignments in postscript. *Bioinformatics* **15**(4): 305-308.

Pedersen LC, Dong J, Taniguchi F, Kitagawa H, Krahn JM, Pedersen LG, Sugahara K, Negishi M. 2003. 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* **278**(16): 14420-14428.

Petit P, Granier T, d'Estaintot BL, Manigand C, Bathany K, Schmitter JM, Lauvergeat V, Hamdi S, Gallois B. 2007. Crystal structure of grape dihydroflavonol 4-reductase, a key enzyme in flavonoid biosynthesis. *J Mol Biol* **368**(5): 1345-1357.

Sayle RA, Milner-White EJ. 1995. Rasmol: Biomolecular graphics for all. *Trends Biochem Sci* **20**(9): 374.

Singh SK, Eland C, Harholt J, Scheller HV, Marchant A. 2005. Cell adhesion in arabidopsis thaliana is mediated by ectopically parting cells 1--a glycosyltransferase (gt64) related to the animal exostosins. *Plant J* **43**(3): 384-397.

Wiggins CA, Munro S. 1998. 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* **95**(14): 7945-7950.

Zhang Y. 2007. Template-based modeling and free modeling by i-tasser in casp7. *Proteins* **69 Suppl 8**: 108-117.

Zhang Y. 2008. I-tasser server for protein 3d structure prediction. *BMC Bioinformatics* **9**: 40.

Zhang Y, Skolnick J. 2005. Tm-align: A protein structure alignment algorithm based on the tm-score. *Nucleic Acids Res* **33**(7): 2302-2309.