De novo Sequencing and Native Mass Spectrometry Reveals Hetero-Association of Dirigent Protein Homologs and Potential Interacting Proteins in *Forsythia* × *intermedia*

Mowei Zhou,[1] * Joseph A. Laureanti,[2] Callum J. Bell,[3] Mi Kwon,[4] Qingyan Meng,[4] Irina V. Novikova,[1] Dennis G. Thomas,[5] Carrie D. Nicora,[5] Ryan L. Sontag,[5] Diana L. Bedgar,[4] Isabelle O'Bryon,[6] Eric D. Merkley,[6] Bojana Ginovska,[2] John R. Cort,[4,5] Laurence B. Davin, and [4] Norman G. Lewis[4]

1. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA
2. Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, USA
3. National Center for Genome Resources, Santa Fe, New Mexico, USA
4. Institute of Biological Chemistry, Washington State University, Pullman, WA, USA
5. Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA
6. National Security Division, Pacific Northwest National Laboratory, Richland, Washington, USA

Corresponding email: mowei.zhou@pnnl.gov

Abstract: The discovery of dirigent proteins (DPs) and their functions in plant phenol biochemistry was made over two decades ago with *Forsythia* × *intermedia.* Stereo-selective, DP-guided, monolignol-derived radical coupling *in vitro* was then reported to afford the optically active lignan, (+)-pinoresinol from coniferyl alcohol, provided one-electron oxidase/oxidant capacity was present. It later became evident that DPs have several distinct sub-families. In vascular plants, DPs hypothetically function, along with other essential enzymes/proteins (e.g. oxidases), as part of lignin/lignan forming complexes (LFCs). Herein, we used an integrated bottom-up, top-down, and native mass spectrometry approach to detect potential interacting proteins in a DP-enriched solubilized protein fraction from *Forsythia* × *intermedia,* via adaptation of our initial report of DP solubilization and purification. Because this hybrid species lacks a published genome, *de novo* sequencing was performed using publicly available transcriptome and genomic

data from closely related species. We detected and identified two new DP homologs, which appear to form hetero-trimers. Molecular dynamics simulations suggest that similar hetero-trimers were possible between Arabidopsis DP homologs with comparable sequence similarity. Other identified proteins in the DP-enriched preparation were putatively associated with DP function or the cell wall. Although their co-occurrence after extraction and chromatographic separation is suggestive for components of a protein complex *in vivo*, none were found to form stable complexes with DPs under the specific experimental conditions we have explored. Nevertheless, our integrated mass spectrometry method development helps prepare for future investigations directed to detect hypothetical LFCs and other related complexes isolated from plant biomass fractionation.

Keywords: dirigent protein, native mass spectrometry, top-down mass spectrometry, protein complex, proteomics, structural biology, *de novo* sequencing, plant biology, lignans, lignins, cell walls

Introduction

Dirigent proteins (DPs) were discovered in *Forsythia* × *intermedia* over two decades ago, with the first example stipulating stereoselective coupling of two *E*-coniferyl alcohol molecules to give the lignan (+)-pinoresinol.[1] This (+)-pinoresinol-forming DP was proposed to capture *E*-coniferyl alcohol-derived free-radical intermediates (produced by oxidases), and position them in the DP active site to form a product with specific stereochemistry. In contrast, oxidase-engendered coupling of *E*-coniferyl alcohol only yielded various racemic products. This stereoselectivity in coupling thus suggested the requirement of at least two interacting proteins (a DP and an oxidase). The gene encoding the (+)-pinoresinol-forming DP from *Forsythia* × *intermedia*, named as FiDir, was obtained using cDNA methods prior to complete sequencing of any plant genome (Table S1).[2] Its recombinant form was expressed in a *Spodoptera frugiperda*/baculovirus system, and it had the same activity as the native DP.[2] DPs highly homologous to FiDir were also obtained from pea (PsDRR206)[3] and *Arabidopsis thaliana* (AtDir6)[4] and their X-ray structures have been determined. PsDRR206 (PDB: 4REV) is a (+)-pinoresinol-forming DP,[3] whereas AtDir6 engenders formation of the opposite antipode to produce (-)-pinoresinol (PDB: 5LAL).[4,5]

In addition to pinoresinol-forming DPs, other DPs with different substrate specificities have been reported. For example, GhDir4 from cotton (*Gossypium hirsutum*) mediates (+)-gossypol formation, an aromatic diterpenoid,[6] and GePTS1 from licorice (*Glycyrrhiza echinata*) and PsPTS1 from pea (PDB: 6OOD) are pterocarpan-forming DPs.[7] A dirigent-domain containing protein, trivially named enhanced suberin1 (ESB1), was shown to be involved in formation of the lignified Casparian strip of *Arabidopsis thaliana* roots, as its constitutive lignin deposition was interrupted when specific DPs were knocked out.[8] Interestingly, ESB1 was also apparently co-localized with CASP proteins in the lignified Casparian strip of the Arabidopsis roots using in vivo fluorescence imaging. More recently, a Dir-a sub-family member (AtDir12), has been implicated in the formation of 8-O-4′-linked feruloylcholine- and sinapoylcholine (SC)-conjugated lignans.[9]

DP sub-family homologs are found throughout the plant kingdom, but are absent in algae and cyanobacteria.[1,7,10] Multiple DP genes in different sub-families are found in all vascular plant species studied, even though *circa* 95% of DPs currently have no known biochemical function. For example, in the *Arabidopsis thaliana* genome, 24 genes were assigned as encoding DP or DP-like proteins (named as AtDirs).[11] Bioinformatic analysis of their expression levels suggested distinct physiological/biochemical functions of different DP homologs, ranging from various stress responses, hormonal regulations, to developmental processes.[11] Therefore, it is reasonable to expect sophisticated genetic control of DP expression, function, and localization *in vivo*. It is also noteworthy that all DP sub-family members of known biochemical functions are gateway entry points to different plant phenol metabolic classes,[6,7] including Casparian strip lignin formation.[8]

Lignin biosynthesis requires many enzymatic steps to biosynthesize lignin monomers in the cytoplasm beginning from phenylalanine, as well as to transport monolignol monomers to the apoplast for polymerization.[12] While it is frequently viewed that the polymerization step of lignification is a chemically controlled abiotic process, others have indicated that a protein guided assembly mechanism is more likely involved,[13–15] such as formation of specialized forms of lignin where tight control of localization/cell wall thickness is required (e.g., Casparian strip lignin).[8] The biological functions of DPs led to the hypothesis that DPs may be part of macromolecular assembly that is involved in lignin biosynthesis, i.e. given their highly specialized lignified cell wall structure. This lignin-forming complex (LFC) was thus hypothesized to be a membrane-anchored protein complex,[8] likely containing DPs (e.g. ESB1), CASP domains, oxidases, and other proteins (Figure 1a). Biochemical proof supporting this hypothesis, however, is not yet reported. Nevertheless, at minimum, we can hypothesize that both DPs and one or more oxidases could be part of putative LFCs or could provide clues to the architecture of LFCs involving other DP homologs.

From a biochemical mechanistic perspective, it was concluded that all DPs of known biochemical function share common quinone methide intermediate-binding/stabilizing functions.[7] Their detailed DP structures, including flexible loops and termini, apparently evolved for diverse substrate specificity, and in possibly

binding other proteins for function or localization.[7] Indeed, such substrate versatility may help to design biotechnological routes to produce pharmaceuticals difficult to make by conventional methods.[6,11,16] Given the potential roles of DPs in lignin biosynthesis, a comprehensive understanding of their structure and function is considered necessary to optimally balance more facile degradability with the structural role of lignin needed *in situ*. Accordingly, since both the *Forsythia* (+)-pinoresinol forming DP and oxidase(s) are apparently solubilized from the cell wall/membrane enriched fraction of its stem tissue, we speculated they may be part of a membrane-anchored protein complex, perhaps somewhat similar to the hypothetical Casparian strip LFC. In this study, we thus revisited our previous work on *Forsythia* × *intermedia* DPs[1] using integrated mass spectrometry (MS) analysis to identify other DP homologs present, as well as other major protein components that were released together with the DPs engendering (+)-pinoresinol forming activity.

Experimental

Protein extraction and purification

Stem tissues were harvested from mature *Forsythia* × *intermedia* plants grown at Washington State University (Pullman, WA, USA). Solubilization of cell wall proteins, partial purification of DP-containing fractions, and activity assays were carried out as described in Davin et al.[1] The final fractions are buffer exchanged into MES-HEPES-sodium acetate buffer (pH 5). Analysis of native proteins were performed with fresh samples stored at 4 °C. Denaturing LCMS were performed from frozen aliquots.

Native mass spectrometry

Protein samples from above were buffer exchanged into either 100 mM ammonium formate (pH 5) or 100 mM ammonium acetate (pH 6.8) using Zeba Spin size exclusion desalting columns (7 kDa cutoff, ThermoScientific, Catalog 89877). The buffer-exchanged protein solutions were then injected into an electrospray glass capillary (tip size 1~5 μm) made from borosilicate glass (O.D. 1 mm, I.D. 0.78 mm, 10

cm length with filament, part number: BF100-78-10, Sutter Instrument) using a P-1000 micropipette puller (Sutter Instrument, Novato, CA, USA). A platinum wire was inserted into the capillary to supply a 1 kV voltage for electrospray. Mass spectra were collected on a Waters Synapt G2s-i mass spectrometer. Source temperature was 30 °C, the cone voltage was 50 V for ion mobility mode (for maintaining folded structures), and 150 V was used for TOF mode (for best mass resolution). Trap gas (Argon) was 3 mL/min. Other tuning voltages were kept at default values. Peaks were assigned manually, or automatically using UniDec.[17] Mass values were calibrated using cesium iodide clusters up to ~6000 *m/z* and extrapolated to 14000 *m/z*.

Top-down LC/MS of intact proteins

Reversed phase separation of denatured intact proteins was performed on a Waters NanoAcquity liquid chromatography (LC) system, equipped with a trap column for online desalting (in-house packed, 5 cm, inner diameter 150 μm, outer diameter 360 μm, C2 reversed phase, MEB2-3-300, Separation Methods Technologies) and an analytical column with C2 stationary phase (in-house packed, 50 cm, inner diameter 100 μm, outer diameter 360 μm, same packing material as the trap column). The binary solvents were 0.2% formic acid in water (A) and 0.2% formic acid in acetonitrile (B), with a linear gradient running from 5-50% solvent B in A over 100 min. MS was operated under "intact protein mode" on a Thermo Fusion Orbitrap Lumos. Electron transfer dissociation (ETD, 25 ms), higher-energy collisional dissociation (HCD, 25%+/-10%), and EThcD (20 ms ETD supplemented by 15% HCD) spectra were collected on the same precursor. Resolution was 120K or 7500 for MS1, and 120K for MS2.

Bottom-up LC/MS of digested peptides

Urea was added to the protein solutions and then water bath sonicated into solution at room temperature to a final urea concentration of 8M. Dithiothreitol (DTT) was added to a concentration of 5 mM and the samples were reduced and denatured at 60 °C for 30 minutes with constant shaking at 800 rpm and then allowed to cool. Samples were next diluted 8-fold for preparation for digestion with 100 mM NH_4HCO_3, 1

mM $CaCl_2$ and sequencing-grade modified porcine trypsin (Promega, Madison, WI), this being added to all protein samples at a 1:50 (w/w) trypsin-to-protein ratio for 3 hours at 37 °C with shaking at 500 rpm. Digested samples were then acidified with 1% trifluoroacetic acid (TFA) and desalted using a 4-probe positive pressure Gilson GX-274 ASPEC™ system (Gilson Inc., Middleton, WI) with Discovery C18 50 mg/1 mL solid phase extraction tubes (Supelco, St.Louis, MO), via the following protocol: MeOH (3 mL) was added for conditioning, followed by 0.1% TFA in H_2O (3 mL). Samples were then loaded onto each column followed by 4 mL of (95:5 v/v) H_2O:MeCN containing 0.1% TFA. Samples were eluted with 1 mL (80:20 v/v) MeCN:H_2O containing 0.1% TFA, and concentrated down to ~50 µL using a Speed Vac. A bicinchoninic acid (BCA) assay (Thermo Scientific, Waltham, MA USA) was performed to determine peptide concentrations. Samples were diluted to 0.10 µg/µL with nanopure H_2O and stored at -20 °C until MS analysis.

Peptide separation was performed on the same LC/MS system as top-down, but with C18 stationary phase (3 µm, 300 Å pore size, Phenomenex, Terrence, USA). The binary solvents were 0.1% formic acid in water (A) and 0.1% formic acid in MeCN (B). Peptides (0.5 ug) were injected onto the trap column for 10 min for online desalting, then injected into the analytical column. Separations were performed with a gradient of 5-35% B in A over 100 min. Data dependent acquisition was used on the MS with 3 s cycle time. HCD (collision energy 35% ± 5%) was used for MS2. When common glycan oxonium ions were detected in HCD, collision induced dissociation (CID) in ion trap and ETD (calibrated charge dependent reaction time) were triggered on the same precursor. Resolution was 120K for MS1 and 60K for MS2.

De novo *sequencing and sequence assembly assisted by transcripts*

Bottom-up LC/MS data for the tryptic peptides was first analyzed using PEAKS Studio to generate *de novo* peptide sequences. Mass tolerance was 20 ppm for MS1 and 0.02 Da for MS2. The *de novo* sequenced peptides (Average Local Confidence, ALC score >= 75%) were used to assemble transcript reads as described below. Protein annotations for *Olea europaea* var. *sylvestris* v1.0, a lignan rich plant species,[18] were downloaded from Phytozome. Two *Forsythia koreana* transcriptome data sets were also downloaded

from the NCBI Sequence Read Archive SRR2075824, consisting of 41.3 million paired end 101 bp reads derived from leaf tissue, and SRR2075825 consisting of 47.8 million paired end 101 bp reads from callus tissue. The data sets were converted to FASTA format, combined into forward and reverse sets, normalized using bbnorm (https://sourceforge.net/projects/bbmap/), and assembled with Trinity[19] using the --no_normalize_reads command line option.

In order to classify peptide fragments derived from the PEAKS Studio analysis in the proteomics experiments (default search settings with score filtering as described above), the fragments were arranged in FASTA format and searched against *O. europaea* protein annotations using BLASTP, and against the *Forsythia koreana* assembled transcriptome data using TBLASTN. The short nature of the peptide fragments meant that even perfect matches resulted in relatively high E-values. Accordingly, BLAST parameters were set to report 50 alignments, which were all inspected manually to identify potential good hits. TBLASTN hits to *Forsythia koreana* transcriptome entries were further investigated by translating target RNA sequences in the appropriate reading frame and running BLASTP against the *O. europaea* proteins. RNA-Seq support for transcripts of interest was evaluated by aligning the reads back to the assembled transcripts using GSNAP,[20] loading the resulting BAM file into the Integrative Genomics Viewer[21], and examining the read alignment depth.

Assembled sequences were used as a custom protein sequence database FASTA. Based on target masses of interest observed in native MS, top-down MS2 spectra were analyzed manually to find terminal sequence tags. The tags were then used as a proxy to find candidate protein sequences in the custom FASTA, allowing several small proteins (< 30 kDa) to be confidently identified. In addition, the custom FASTA was used in Byonic on the bottom-up LCMS peptide data to identify proteins with high sequence coverage, allowing for larger protein (> 30 kDa) identification as above. Target proteins with high sequence coverage were manually selected and saved into a "focused" FASTA for additional analysis. The peptide data were then re-processed with Byonic (mass error tolerance 10 ppm, FDR 1%) for post-translational modification (PTM) profiling. Plant N-glycans, 6 common *O*-glycans, methionine oxidation, protein N-terminal acetyl,

and asparagine/glutamine deamidation were included in the dynamic modifications during the search. Top-down data were re-processed using TopPIC[22] (mass error tolerance 15 ppm, FDR 1%), and manually analyzed/visualized in LcMsSpectator.[23] The major proteins identified were also searched (BLASTP) against the recently published *Forsythia suspensa* genome[24] (https://www.ncbi.nlm.nih.gov/assembly/GCA_013103335.1) and verified.

Homology models and docking of DP trimers

Experimental sequences determined from *de novo* sequencing were submitted to I-TASSER[25] for generating homology models of DP monomers. Structures were visualized in VMD. The experimentally determined AtDir6 homo-trimer structure (DP from *Arabidopsis thaliana*, PDB: 5LAL)[4] was used as template for generating homology models of homo-trimeric DPs using Swiss-Model.[26] Hetero-trimers were built by docking clipped dimers with another monomer unit (mixed with 1:2 and 2:1 stoichiometry, no cross-species hetero-trimers were analyzed) using PyDockWEB (https://life.bsc.es/pid/pydockweb).

Molecular dynamics

Molecular dynamics simulations of homo-trimeric and hetero-trimeric FiDir and AtDir systems in aqueous solution were performed starting from the homology models. Amber13 forcefield parameters were used for all residues.[27] All simulations were performed with the GROMACS simulation package,[28] using the following protocol: (1) initial geometry of the system optimized using a conjugate gradient approach; (2) optimized structure was gradually heated by carrying out 100–250 ps equilibrations at increasingly higher temperatures from 0 K to 300 K in increments of 100 K, followed by a 20 ns equilibration at 300 K; (3) trajectories were collected for 180 to 200 ns. All simulations were run at constant pressure (1 atm) and temperature (300 K), with a time step of 2 fs. All water molecules are explicit. Coordinates were saved every 10 ps, providing ~20,000 snapshots for analysis.

Hydrogen bonding analysis was completed using the Visual Molecular Dynamics (VMD) hydrogen bonding plug-in. The distance cutoff between the heavy donor-acceptor atoms was set to 3.5 Å, and the

cutoff was set to 60 degrees for the donor-proton-acceptor atom angle. Hydrogen bond occupancy was calculated only for polar/charged atoms and unique residues (if the residue had more than one polar atom, all hydrogen bonds were counted together). Occupancy > 100% represents a residue with more than one hydrogen bond. The average distance of each residue from every residue on the opposite chain was calculated using GROMACS, by using the center of mass of the side-chains and calculating the average distance throughout the last 80% of the trajectories.

Results & Discussion

Analysis of DP-enriched fraction from Forsythia × intermedia

We followed our published protocol for solubilizing DPs from *Forsythia* × *intermedia* plant stem tissues.[1] Crude protein extract, following ammonium sulfate precipitation, was subjected to Mono S cation exchange chromatography (Figure 1b), with active fractions being eluted with 233 mM and 333 mM Na_2SO_4. Both fractions gave similar bands on denaturing gels, with a representative gel for the 333 mM Na_2SO_4 fraction shown in Figure 1c. Fractions with enriched (+)-pinoresinol forming activity were pooled, with products examined using chiral chromatographic separations[1] (Figure S1). At this initial stage of DP purification, the oxidase(s) could potentially be complexed with the DP or be present as a separate entity to produce coniferyl alcohol radical intermediates needed to afford (+)-pinoresinol via DP involvement.

DP-enriched fractions were electro-sprayed under non-denaturing conditions (pH 5, 100 mM ammonium formate), as the (+)-pinoresinol forming DP was previously shown to have highest activity between pH 4.25-6.0.[29] Native MS spectra taken with 100 mM ammonium acetate (pH 6.8, commonly used in other native MS studies) were similar to those in ammonium formate (pH 5), suggesting that assembly states of the proteins in the sample were not significantly affected by pH (data not shown). A representative ion mobility (IM) – mass spectrum is shown in Figure 1d. IM separates ions based on their shape and charge, helping to better resolve overlapping entities in the *m/z* dimension. Several major proteinaceous species were observed at 9.4 kDa, 34 kDa, ~58 kDa, and 70 kDa, respectively. Each protein species was isolated

by its *m/z* value and activated via gas collisions (collision induced dissociation, CID). Bound ligands and protein sub-units, if any, could be released during CID to infer the composition of any non-covalent complex/interaction present. The ~58 kDa species was confirmed as trimers of 18-19 kDa monomers, and assigned as a DP given its known trimer structure (discussed in more detail later). The other species (9.4 kDa, 34 kDa, and 70 kDa) were assigned as monomeric proteins that were also observed on the gel bands. Protein extraction was performed on several sets of biological replicates. The major species detected in the purified fractions was qualitatively consistent. Significantly, no higher mass complexes were detected above 70 kDa when using native MS analyses, indicative that any putative DP-containing LFCs were disassembled under the experimental conditions employed. If so, this might suggest that either full assembly of the hypothetical LFC was transient, or unable to survive the partial purification and MS experimental conditions employed. The purification was initially developed for purification of soluble DP fractions with pinoresinol-forming activity; however, if the putative complex is membrane anchored, it is possible that the complex is no longer intact at the time of MS analysis.

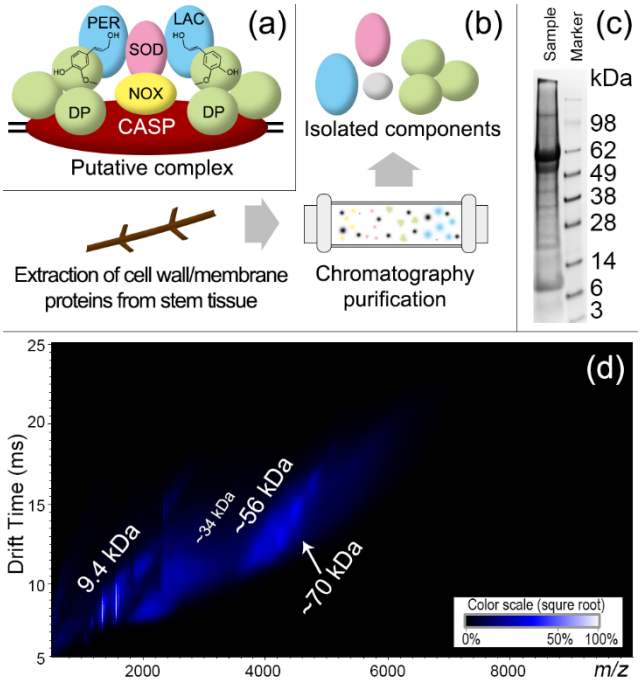

Figure 1. (a) Hypothetical lignin forming complex (LFC) cartoon in Arabidopsis based on published experimental data. PER – peroxidase; LAC – laccase; SOD – superoxide dismutase; NOX – NADPH oxidase; CASP – Casparian strip domain protein; DP – dirigent protein. (b) Simplified representation of the extraction method of DP from *Forsythia* × *intermedia* stem tissue and partially purified by chromatographic steps. The hypothetical LFC in *Forsythia* × *intermedia*, perhaps equivalent to that in Arabidopsis, may have been disassembled during purification, as only isolated non-interacting protein components were detected in DP-enriched fractions. (c) Denaturing gel showing *Forsythia* × *intermedia* proteins that co-eluted with DPs after MonoS column chromatography. The right lane is the molecular weight marker. (d) Ion mobility mass spectrum of the DP-enriched fraction under non-denaturing conditions. The x axis shows the *m/z*, the y axis shows the drift time in milliseconds by ion mobility. The color represents relative intensity, with the scale bar in the bottom right corner. The major resolved species are labeled with their masses in kDa. No significant amount of higher mass complexes was detected above ~70 kDa.

De novo *sequencing identified two new DP homologs in* Forsythia × intermedia

Forsythia × *intermedia* does not yet have a sequenced genome, whereas that of the closely related *Forsythia suspensa* has only recently been published.[24] In our earlier work, we used complementary DNA sequencing to yield the amino acid sequences (Table S1) of two DP homologs, FiDir1 and FiDir2,[2] and a laccase (oxidase), FiLaccase.[30] Both DPs were heterologously expressed in a *Spodoptera frugiperda*/baculovirus system as ~ 78 kDa trimers, and both had the expected (+)-pinoresinol forming activity in the presence of exogenously provided oxidase or oxidant and substrate.[2] Since other *Forsythia* × *intermedia* proteins were not sequenced or characterized in those studies, we attempted instead to *de novo* sequence high abundance proteins in DP-enriched preparations to identify potential LFC components. Although sequence tags can be generated by fragmentation data of the intact proteins, < 30 kDa, their coverage is incomplete to define the full sequences. Many proteins were also glycosylated, further complicating analysis. We thus complemented a top-down analysis with a bottom-up proteomics strategy.

To maximize coverage, following ammonium sulfate precipitation, the samples were subjected to sequential cation exchange chromatography (MonoS and PoroS SP columns), with the resulting eluate pooled into 4 fractions (F1 -F4) for analysis (denaturing gel of all fractions shown in Figure S2). Native MS of F1 -F4 Figure S3) detected similar major protein species as seen in the sample shown in Figure 1d.

F1-F4 were also treated with trypsin, and subjected to LCMS for *de novo* peptide sequencing, with identified peptide sequences used to identify matching genes in assembled transcriptome sequencing reads from the closely related plant species, *Forsythia koreana*.[31] Putative protein sequences were generated and used in a "pilot" analysis of the bottom-up proteomics data. Target proteins with high sequence coverages were then manually selected into a target protein database for processing intact protein and peptide LCMS data.

In these fractions (and at this early stage of (+)-pinoresinol forming DP purification), both FiDir1/FiDir2, and the previously sequenced laccase were detected but with very limited sequence coverages (Figure S4). They likely had low abundances and were heterogeneously modified (glycosylation, etc). Additionally, we identified 14 other proteins with high sequence coverage and good quality spectra based on analysis of the tryptic peptide and/or top-down data (Table S2). Many of these proteins had near complete sequence coverage and were also mapped to the recently published *Forsythia suspensa* genome[24] with near 100% sequence identities.

Of these, we provisionally identified two new DP homologs *circa* 18.6 and 19.8 kDa, tentatively named as FiDir18 and FiDir19 (following their nominal molecular weights). Their protein sequences were confirmed based on bottom-up and top-down data (Figure 2a-b). Peptide coverage was near complete (full coverage maps in Figures S5-6), with top-down data having high coverage near the N-termini (annotated spectra in Figures S7-8). We also confirmed two and three N-glycosylation sites for FiDir18 and FiDir19, respectively. The high coverage confirms that the two species are distinct protein homologs, but not the same protein with different post-translational modifications (PTMs, e.g., glycosylation). Conversely, the C-termini sequences were uncertain in both bottom-up and top-down data due to lack of coverage (especially for FiDir19). We evaluated the de novo sequences by TBLASTN against the genome of *Forsythia suspensa*.[24], and found an exact match of the first 127 residues for FiDir19 whereas no exact match was found for FiDir18. Interestingly, we also only found an exact match for FiDir2, but not for FiDir1, in the *Forsythia suspensa* genome. Because *Forsythia* × *intermedia* is a hybrid of *Forsythia*

suspensa and *Forsythia viridissima*, we suspect *Forsythia* × *intermedia* inherited FiDir19 and FiDir2 from *Forsythia suspensa*, thereby potentially explaining the absence of FiDir18 and FiDir1 in the *Forsythia suspensa* genome. Additionally, the full protein sequence mapped to FiDir19 in the *Forsythia suspensa* genome has a different and longer C-terminus from the one we predicted from the *de novo* analysis (Table S2). The FiDir19 sequence from *Forsythia suspensa* offered a better fit (Figures S6, S8), and was used in Figure 2b and the following discussions.

The main species detected at the intact protein level were reasonably uniform, with only 3-4 proteoforms (i.e. unique protein species carrying specific PTMs) for each protein (Figure 2c). Variations in proteoform masses can be explained by different combinations of PTMs (almost exclusively from glycans). However, the experimental masses of the intact proteins were larger than the sequences of FiDir18 and FiDir19, Figure 2a and 2b, after adding the masses of the PTMs, this resulting in products larger by 387.2 Da and 556.3 Da for FiDir18 and FiDir19, respectively. C-terminal truncations alone did not explain the experimental intact masses. These unexplained mass shifts may represent a combination of different amino acid sequences and PTMs, which cannot currently be verified due to limited sequence coverage in this region.

We generated homology models of trimeric DPs in I-TASSER[25] and Swiss-Model[26] (Figure 2d). The unconfirmed C-termini were in the flexible region outside the core and were not expected to significantly impact the inter-subunit interfaces. The identified N-glycan sites were all in loops of the structural models. N17 is close to the interface of another monomer in the complex, and other two sites are also facing outside. The equivalent of N88 in FiDir19 is absent in FiDir18 (K88).

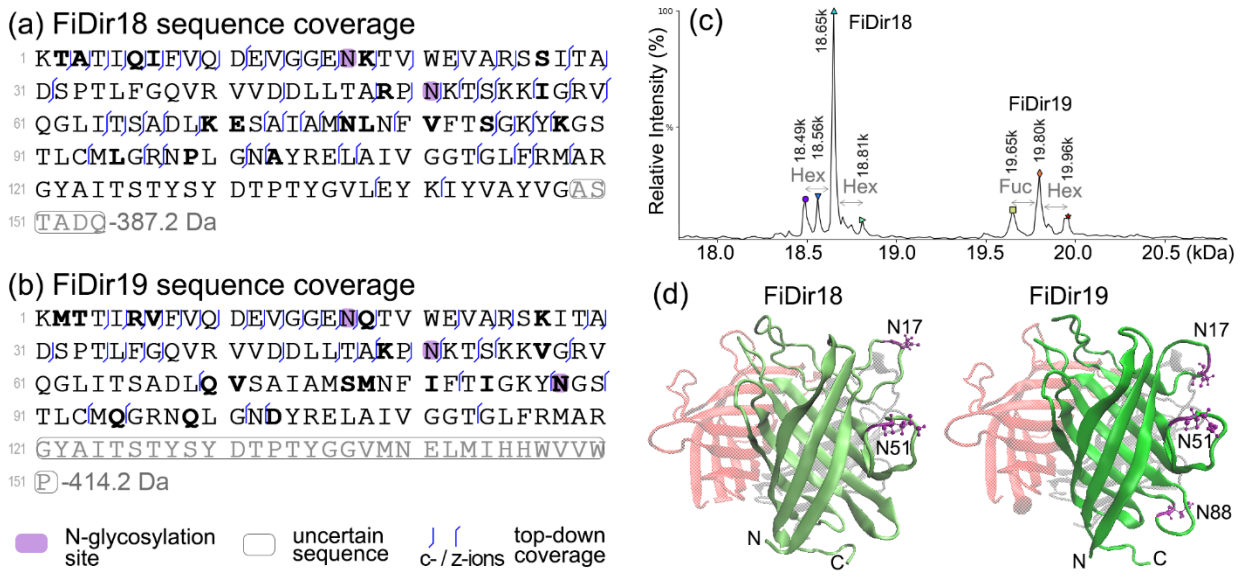

Figure 2. Sequence coverage maps for (a) FiDir18 and (b) FiDir19. Gray letters with rectangles indicate no peptide coverage and sequences were not confirmed in the C-terminal regions. Blue wedges represent top-down sequence coverage at the intact protein level. N-glycosylation sites are labeled in purple. Uncertain regions of the sequences are labeled in gray boxes, with unknown mass shifts written at the end of the sequences. Residues in bold are different between FiDir18 and Fir19. The sequence coverage at the unique residues confirms the two species are distinct protein homologs. (c) Deconvoluted intact mass distribution of FiDir18 and FiDir19 in denaturing LCMS. Several minor forms of each protein can be explained by variation in glycosylation. (d) Homology homo-trimer models of FiDir18 and FiDir19. Each subunit is a different color (green, red, and gray) with the green subunit highlighted to show structural details and post-translational modifications (PTMs). N-/C-termini are labeled with letter N/C. Glycosylated Asn residues are highlighted as purple bond structures in the green subunit. The major N-glycans identified were HexNAc(2)Hex(3)Fuc(1)Pent(1). All glycan sites are in loop regions facing outwards.

Two newly discovered Forsythi*a DP homologs formed hetero-trimers*

In native MS analysis, both FiDir18 and FiDir19 were detected as ~58 kDa trimers, i.e. as for other DPs of known biochemical function.[3] Interestingly, FiDir18 and FiDir19 not only formed homo-trimers, but also hetero-trimers (Figure 3a). The stoichiometry was further confirmed by performing MS2 on these species via CID. Homo-trimers of FiDir18 and FiDir19 only yielded one protein species (FiDir18 and FiDir19 monomers, respectively). Hetero-trimers were confirmed by the presence of both protein species in the released monomers. In essence, the ratio of released FiDir18 to FiDir19 monomers correlated directly with

their trimer stoichiometry (Figure S9). As an example, CID of the mass-isolated 57.1 kDa hetero-trimer species (FiDir18:FiDir19 = 1:2) released both FiDir18 and FiDir19 monomers (Figure 3b), confirming the trimer contained both DP monomers., Both hetero-trimers (FiDir18:FiDir19 = 1:2, or 2:1) were reproducibly detected in three biological replicates (Figure S10).

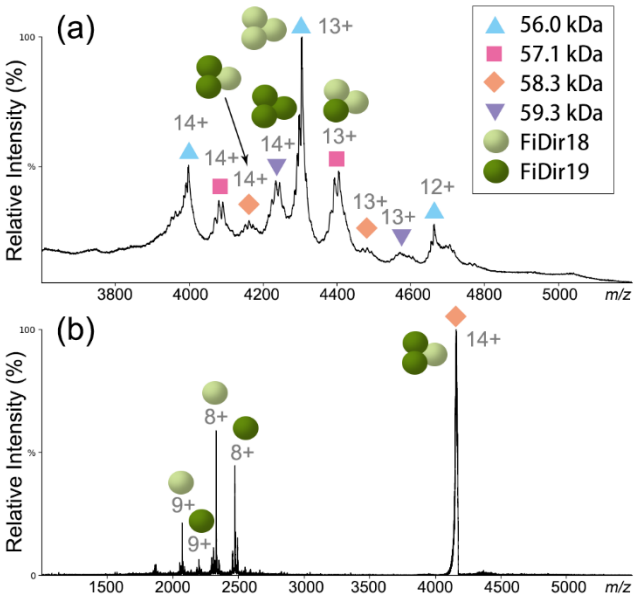

Figure 3. (a) Native MS spectrum zoomed into the *m/z* range showing hetero-complexes of the two DP homologs FiDir18 and FiDir19. Peak assignments are shown with colored symbols with the keys in the box to the right. Charge states of the assigned peaks are labeled in gray. Based on the mass of the assigned species, they can be fitted to different combinations of trimers between the two 18.6 kDa and 19.8 kDa DP homologs (annotated as light and dark green spheres, respectively). (b) CID of the mass isolated FiDir18:FiDir19 1:2 complex. Both FiDir18 and FiDir19 monomers were released around *m/z* of 2000-2500. Their masses matched to the 18.6 kDa and 19.8 kDa DP homologs as identified by top-down in Figure 2, thus confirming the assignment of a hetero-complex. The intensity of the released monomer peaks also correlated with the stoichiometry of the complex as shown in Figure S9.

Hetero-association of dirigent protein homologs may have implications to their underexplored functions

The 8-stranded β-barrel structure seen in several crystal structures of dirigent proteins is the highly conserved fold of all known DPs. In order to better understand the possible function of the two *Forsythia* × *intermedia* DP homologs above, we compared their sequences with known DPs from other plant species

and placed them into a phylogenetic tree (Figure 4). DPs in Arabidopsis and characterized DP homologs with published reports[1,3,4,6–9,32–36] are highlighted in the tree, with more details in Table S3. FiDir18 and FiDir19 localize to the broad Dir-b/d subfamily, which is distinct from the Dir-a subfamily to which the (+)-pinoresinol forming FiDir1/FiDir2 belong. However, FiDir18 and FiDir19 are sufficiently distant from most other Dir-b/d sequences that it may be problematic to place them accurately in the overall DP phylogenetic tree. All homologs with sequence identity greater than 50-55% are apparently exclusive from plants in the order Lamiales. Thus, FiDir18 and FiDir19 may have novel biochemical functions given their low sequence identity to DPs of known biochemical function. However, FiDir18 and FiDir19 have 87% identity to each other, with only a single mutation within the putative active site in the core of the β-barrel, likely suggesting similar substrate specificity between them. Their shared sequence identities may even help explain the hetero-association observed in Figure 3. While the reason for the hetero-trimer assemblies is not clear, it may simply reflect the presence of two alleles from the contributing parent genomes to the hybrid species. On the other hand, many non-hybrid plants have pairs of similar DPs with high sequence identities, possibly due to polyploidy or recent gene duplications. It may be worth considering whether hetero-association of DPs with high sequence similarity has any functional relevance beyond being a consequence of gene duplication.

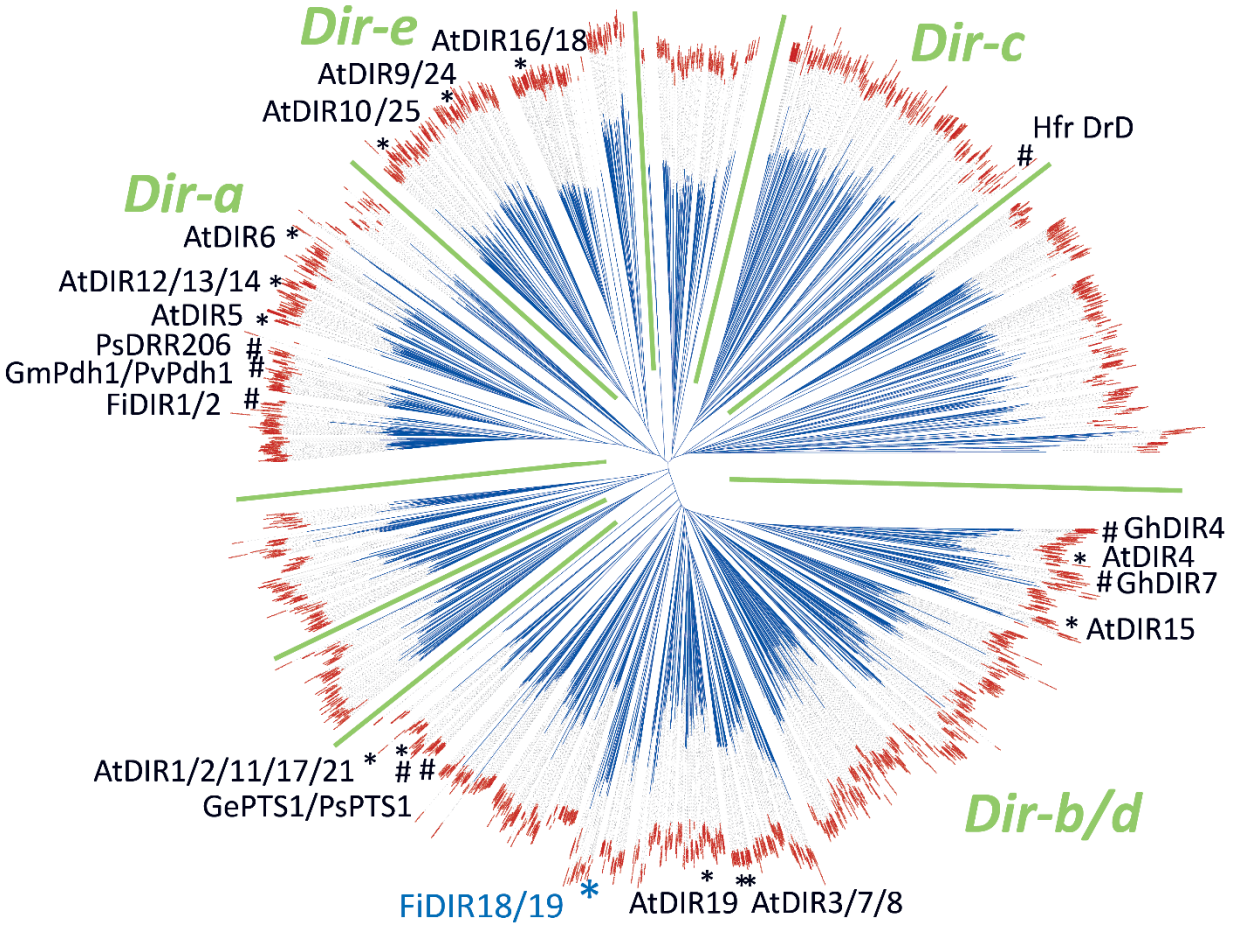

Figure 4. Unrooted phylogenetic tree of the dirigent/dirigent-like protein superfamily constructed from several thousand sequences. Major subfamily divisions are indicated with green lines, and the Dir-a, Dir-b/d, Dir-c, and Dir-e are identified by name. Key landmark sequences are indicated with asterisks (for Arabidopsis DPs) and hash symbols (for other characterized DPs). More details of the characterized DPs are included in see Table S3. The new *Forsythia* dirigent proteins identified in this work are indicated in blue. The tree was rendered with iTOL[37] using a sequence alignment and tree generated with Clustal Omega.[38]

We used molecular dynamics (MD) simulations to understand whether DP heterotrimers are hypothetically more universally plausible beyond the FiDir18/FiDir19 pair detected by native MS in this study, i.e., by analyzing putative hydrogen bonds and salt-bridges, as well as unresolvable steric clashes at the interfaces between the monomers in the trimer. In homology models of FiDir18 and FiDir19 homotrimers, based on the AtDir6 structure template (5LAL), one monomer was removed and replaced by docking with its homologous counterpart, to form the corresponding heterotrimer. MD was then performed to allow side-

chains to adjust and resolve clashes. Because our homology models were built using AtDir6 as a template, we also examined close homologs of AtDir6 in Arabidopsis. AtDir5 has sequence identity of 72% to AtDir6 (78% when ignoring the predicted signal peptides, and versus 87% between FiDIR18/19). We thus performed the same MD analysis on the AtDir5-AtDir6 pair to explore the possibility of hetero-association of other DP homologs with slightly less similarity.

All homo-trimers (FiDir18, FiDir19, AtDir6, and AtDir5) and hetero-trimers (mixed FiDir18-FiDir19, and mixed AtDir5-AtDir6) examined showed no dissociation for the duration of the simulation (180-200 ns), suggesting that all the trimeric species investigated form stable systems, and that AtDir5-AtDir6 hetero-trimers should have similar stability to FiDir18-FiDir19 hetero-trimers, at least within the time frame examined. The aligned sequences by Clustal Omega[38] are shown in Figure 5a, with some structural features highlighted. We examined side-chain to side-chain distances between each pair of subunits of all the trimers (Figure S11 and Figure S12 for FiDir18/FiDir19 and AtDir5/AtDir6 systems, respectively), and all showed an identical pattern, suggesting highly similar inter-subunit contacts. The interfacial residues (distance < 10 Å) are highlighted in yellow and are in the highly homologous/conserved beta-sheet regions. We also defined presumed hydrogen bonds with >20% occupancy over the period of the MD simulation as potentially important interactions for stabilizing the interface. Residues consistently seen in hydrogen bonds or as salt-bridges at both homo- and hetero-interfaces are highlighted in red on the aligned protein sequences in Figure 5a. Again, putative hydrogen bonds and salt-bridges at the interfaces are largely conserved, but with small variations when comparing FiDir18-FiDir19 and AtDir5-AtDir6 trimers. Several hydrogen bonds and salt-bridges unique to hetero-trimeric systems were highlighted in cyan. Interestingly, the interaction patterns observed are different between the AtDir and FiDir systems. In AtDir6/AtDir5, the interactions mapped to the interface facing the solvent (Figure 5b). Instead, those in FiDir18/FiDir19 were seen at the inner side (in the center of the three subunits, Figure 5c). The results indicated small changes in structure and dynamics may occur in hetero-trimers. Experimental structural studies could potentially unveil their functional implications in the future.

Our MD analysis showed that formation of stable hetero-complexes is plausible among other close DP homologs, such as AtDir5 and AtDir6. The putative hydrogen bonding and salt-bridging patterns at the subunit interfaces are similar between FiDir18-FiDir19 and AtDir5-AtDir6 systems, but the details are different due to distinct amino acids at equivalent locations. Therefore, DP homologs that have higher sequence dissimilarity may not form hetero-trimers, if key interfacial interactions are disrupted. Additional MD simulations may provide insight on the minimum amount of conserved interface residues for stable hetero-association. Our initial analysis of interfaces also suggested subtle changes of structure and dynamics upon hetero-trimer formation.

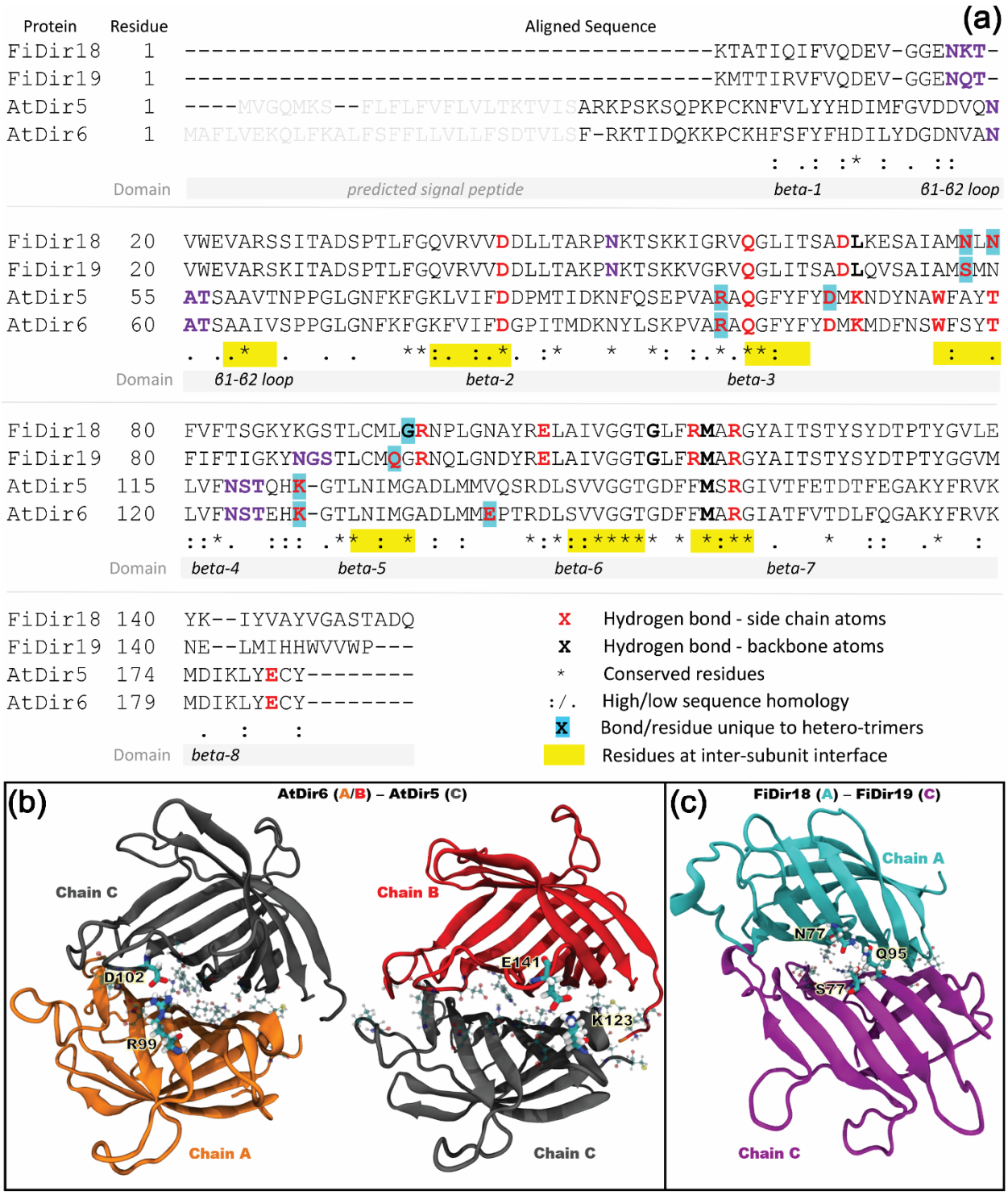

Figure 5. (a) Multi-sequence alignment of FiDir18, FiDir19, AtDir5, and AtDir6. Structural features are annotated following the format described in the legend in the bottom right corner. N-glycosylation sites are colored in purple. (b) Snapshot from MD simulations of heterotrimeric AtDir6/AtDir5 and (c) FiDir18/FiDir19 showing residues experiencing putative hydrogen bonding interactions (>20% occupancy) and salt bridges. Heterotrimeric AtDir6/AtDir5 was composed of two monomers of AtDir6

(orange and red cartoon ribbon structures) and one monomer of AtDir5 (grey cartoon ribbon structure). The FiDir18/FiDir19 hetero-trimer is composed of two monomers of FiDir18 (cyan cartoon ribbon structure) and one monomer of FiDir19 (purple cartoon ribbon structure). The two AtDir6/AtDir5 interfaces are displayed in the same orientation. The FiDir18/FiDIr19 orientation is displaying the inside interface (orientation rotated ~180 degrees from AtDir6/AtDir5 interfaces). The residues experiencing putative hydrogen bonding interactions (>20% occupancy) are shown as transparent ball-and-sticks or licorice representations where the carbon, hydrogen, nitrogen, oxygen, and sulfur atoms are cyan, white, blue, red, and yellow, respectively. Residues involved in interactions unique to hetero-trimers are opaque, while other interactions are transparent.

The functional roles of such hetero-associations are unknown. In the case of the FiDir proteins, heterozygosity of the *Forsythia* x *intermedia* hybrid would be expected to permit such hetero-oligomers to assemble and function normally. Because Arabidopsis is not a hybrid species, the predicted hetero-association between AtDir5 and AtDir6 should not be from heterozygosity. One hypothesis for the hetero-trimer of the *Forsythia* DPs could be coupling of different substrates (if formed among homologs with different activities). Such functional roles have been suggested for hetero-dimers between Golgi N-glycotransferases.[39] Those enzymes have strict Golgi localization and sequential order of function. Therefore, the hetero-dimer may be involved in specialized, ordered processing of N-glycans *in vivo*. However, predicted substrate binding pockets in pinoresinol-forming DPs are deeply buried within the barrel of each monomer. Two substrate radicals can be bound within one subunit for coupling.[4] Therefore, transfer of substrates between two subunits are less likely, at least for pinoresinol-forming DP homologs. The perceivable function of a hetero-complex is to bring different products in close proximity, possibly allowing them to be used by other downstream reactions.

Another possible function of the hetero-trimers is fine regulation of interactions with other molecules (e.g. enzymes, scaffold proteins, cell wall structures). Such a mechanism has been described for many pseudo-enzymes, which are typically defined as catalytically deficient homologs of canonical enzymes.[40] Some known pseudo-enzymes do not have enzymatic activity, but serve as a scaffold to mediate protein-protein interactions. For example, the pseudo-enzyme of human epidermal growth factor receptor (EGFR3, or

HER3) lost its canonical kinase activity, but is involved in altering signaling pathways that lead to evasion of cancer treatments.[41] Most DPs have a conserved domain to sustain the basic trimeric scaffold, but have high variability on the terminal sequences. Both N-/C-termini appear to be flexible and are not fully resolved in crystal structures. Therefore, the terminal regions may be very dynamic and involved in molecular interactions. Taken together, the function for mediating molecular interactions seems to be a plausible hypothesis for the hetero-trimer of *Forsythia* DPs, although this needs to be tested with additional experimental data.

Other proteins in the DP enriched fraction are generally associated with plant cell walls

We examined other major protein components in the DP-enriched fraction. Although they were not directly associated with DPs, they may be weakly associated and co-eluted under the conditions employed. The 9.4 kDa species was identified as a small non-specific lipid transfer protein (nsLTP, top-down data in Figure S13), whereas the 34 kDa species was a peroxidase. The major species at 60-70 kDa was assigned to a *beta*-fructofuranosidase (invertase, peptide mapping data in Figure S14), which appears to be very heterogeneous both on the gel and not well resolved in the top-down data (data not shown). Invertases are known to be important cell wall proteins in plant metabolism and in defense responses[42], raising a possibility that they may be associated with DPs. Additional proteins in the bottom-up data had sequence mass in the ~60k range (Table S2), most of which were detected with glycosylation and were not individually resolved in top-down and native MS. Their functional roles are not clear, but one possibility is that they may be weakly associated with DPs. The laccase we previously identified (FiLaccase)[30] had expected mass ~60 kDa,.but only showed a few peptides hits in the bottom-up data likely due to low concentration and/or resistance to trypsin digestion.

The 34 kDa peroxidase may also be involved in a hypothetical LFC, because oxidases are required for pinoresinol-forming DP function. Its sequence is similar to peroxidase 4-like of vascular plants, and showed multiple forms of glycosylation (Figure 6a). The major peaks were spaced by different combinations of known glycan masses. The heterogeneity of glycosylation spans ~ 2000 Da (Figure 6b), these being mainly

on two N-glycan sites (Figure S15). Under denaturing conditions (Figure 6c, with the organic solvent acetonitrile in LC), the charge states of the proteins significantly increased from the that of the aqueous native condition (Figure 6a). The deconvoluted intact mass profile in Figure 6d had a very similar distribution to Figure 6b, but masses were shifted lower by 600-700 Da. To investigate potential non-covalent ligands, we mass isolated the 32.98 kDa peroxidase under native conditions and activated it via gas collisions (i.e., CID). A peak at 616.2 Da emerged in the low *m/z* region (Figure 6e). The accurate mass and the unique isotope distribution from Fe confirm that heme-Fe (III) was non-covalently bound to the peroxidase, consistent with its expected cofactor (Figure 6f). The information about ligand binding obtained from native MS can potentially be used to infer functions of unknown proteins.

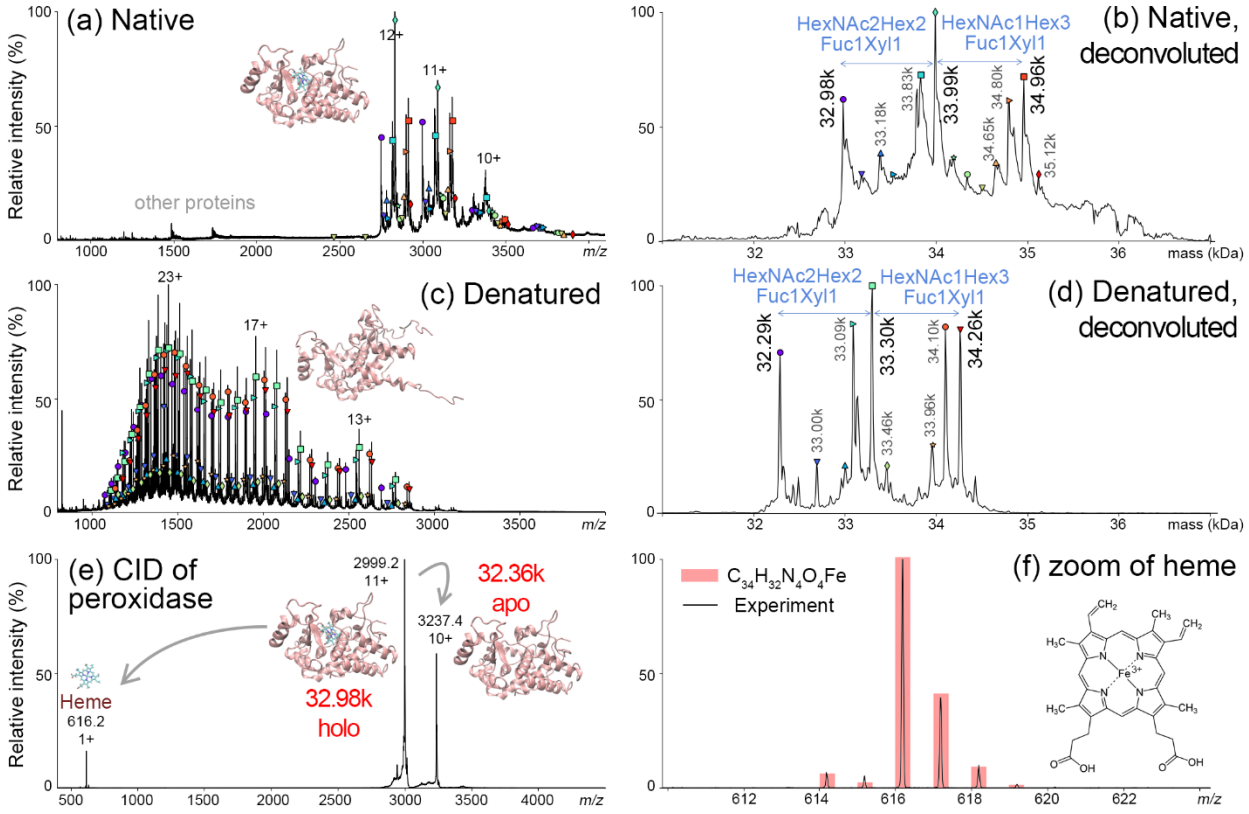

Figure 6. Peroxidase MS data: (a) MS spectrum and (b) deconvoluted mass distribution of peroxidase under native conditions. (c) MS spectrum and (d) deconvoluted mass distribution of peroxidase under denaturing conditions. Deconvoluted masses for major species are annotated in (b) and (d). Symbols in (b) and (d) match to (a) and (c), respectively. Each species in (b) and (d) correspond to multiple peaks at different charge states in (a) and (c), where several major charge states are labeled. (e) Collision induced dissociation (CID) of isolated 11+ peroxidase of 32.98 kDa. After activation, the heme group is released from the holo-

protein, leaving behind the 32.36 kDa apoprotein. (f) Zoom-in view of heme peak released from the holo-peroxidase at *m/z* 616.2 in (e). The isotopic distribution matches well to the theoretical distribution shown in red bars. Homology model of the peroxidase and assigned heme structure (Fe-protoporphyrin IX) are shown as inserts.

Several other low abundance species were also detected by integration of *de novo* peptide sequencing, top-down LCMS, and native MS. A copper binding protein at 10.4 kDa was identified as a member of the cupredoxin family (Figure S16), which is known to be involved in electron transfer and may be a putative SOD as in the model in Figure 1a. Two germin-like proteins were also identified, with one identified by bottom-up data (named as germin-like protein 1, coverage map in Figure S17) and another by native MS (named as germin-like protein 2). Peptide mapping did not yield sufficient coverage for the germin-like protein 2 homolog, but several backbone fragments directly released from the hexamer by native top-down helped map it to the transcript data (Figure S18). Native MS also suggests this protein binds Mn (Figure S19), consistent with known germin homologs.[43]

We performed co-expression analysis based on published database (*Populus trichocarpa* v 3.0, and *A. thaliana* TAIR10, from https://phytozome.jgi.doe.gov) to understand potential correlations with DPs and the co-eluting proteins. In absence of a full *Forsythia* × *intermedia* genome, we chose to examine homologs of the identified FiDir18, nsLTP, and germin-like protein 1 in the Arabidopsis and poplar genomes. Interestingly, many were positively correlated with expression of essential genes that are involved in vascular bundle development in poplar. For example, the poplar homolog of *Forsythia* nsLTP was highly co-expressed with VRLK1 (Vascular-Related RLK1) in poplar, which is a leucine-rich repeat transmembrane protein kinase. The Arabidopsis VRLK1 homolog is involved in switching between cell elongation and secondary cell wall thickening in Arabidopsis.[44] Given its predicted function of transferring lipids, nsLTP is perhaps involved in either membrane localization or restructuring. However, we did not observe well defined complexes of nsLTP with other major proteins in the sample, although it appeared to form multimers and may form higher-order complexes (F4 in Figure S3).

The known structure of germin (PDB: 1FI2), a homo-hexamer with a six-fold rotational symmetry, hypothetically suggest that a head-to-head interaction along the same symmetry axis with a DP might be possible. Germins in cereals are known to have oxalate oxidase activity, generating hydrogen peroxide from oxalate. They are also known to be associated with cell walls and involved in plant development and defense.[45] Given the known function and structure of germin family proteins, we provisionally hypothesize that they may generate hydrogen peroxide, which could then presumably be used by peroxidase to oxidize monolignols. Interestingly, Arabidopsis germin-like protein 10, a homolog of *Forsythia* germin, was co-expressed with several cellulose synthases including CesA4 and cellulose synthase-like C6 as well as Pinoresinol Lariciresinol Reductatse 1 (PLR1), a downstream lignan biosynthetic enzyme. This finding possibly implicates the *Forsythia* germin to a role in either cell wall biosynthesis and/or in defense responses.

Characterization of unknown DPs is challenging and demands novel analytical techniques

Earlier studies on *Forsythia* and *Schizandra* DPs showed that glycosylation was necessary for activity.[46] A previous study on AtDir6 used the *Pichia pastoris* expression system, and deglycosylation resulted in loss of activity.[47] Other reports of FiDir1/FiDir2,[2] *Arabidopsis thaliana*, and *Schizandra* DPs[5] were also in eukaryotic cell lines with glycosylation machinery. Conversely, several recent studied DPs[7] were successfully expressed as active enzymes from *E. coli*. Because many factors can affect protein yield in cell-based systems, the role of glycosylation on DP structure and function is not yet fully established in terms of DPs as different gateway entry points to distinct classes of plant phenols. We attempted to synthesize the identified *Forsythia* proteins using the wheat germ cell-free expression system we developed previously,[48] but only had very limited success with the germin proteins. While the DPs can be translated, they were apparently not able to fold correctly (Figure S20). If no specific chaperone is required, the cell-free results suggest that glycosylation is likely essential for maintaining proper fold and/or prevent aggregation for these DPs, possibly by changing the folding energy landscape.[49] Further optimization of cell-free and heterologous expression with glycosylation is out of the scope of this study. More robust

expression systems are beneficial for further *in vitro* characterization of such proteins, especially those that can faithfully reproduce plant glycosylation (and other PTMs).

Another common challenge for plant research is the lack of completely sequenced and annotated genomes beyond the most studied model systems. Unlike most other organisms, plant genomes are often polyploid, meaning each cell has more than two pairs of homologous chromosomes. Polyploidy in plants makes them more difficult to sequence, resulting in fewer published, fully assembled genomes.[50] However, biochemical experiments may reveal enzymes with novel functions from plant extracts without the need of a genome. As shown by the results presented here, *de novo* sequencing using mass spectrometry data and genomic data from homologous organisms can help provide essential information for further structural and functional analysis. Similar *de novo* strategies have been applied in forensics, archaeology, venomics, etc.[51]

For the current study, initial attempts to directly map the peptides to the olive tree (*Olea europaea* var. *sylvestris*) FASTA did not generate high quality sequences to easily match to top-down MS data. Therefore, we assembled the transcriptome data from *Forsythia koreana*,[31] which allowed us to pick up candidate protein sequences and manually examined these with mass spectrometry data. The recently published genome of *Forsythia suspensa*[24] aided our verification of the *de novo* sequencing results and further extended coverage of target proteins. However, complete *de novo* characterization, especially sequence regions that are variable among the *Forsythia* species (e. g. termini of DPs), remains challenging even with extensive manual analysis due to lack of 100% sequence coverage in the MS data. In addition, PTMs such as glycosylation change the fragmentation behavior of peptides/proteins and complicates the *de novo* analysis. Nonetheless, protein sequences and their PTMs can be identified and confirmed with manual analysis. Further improvements of the workflow by implementing of other existing *de novo* sequencing tools[51–53] and the *Forsythia suspensa* genome will expand our knowledge of the *Forsythia* × *intermedia* proteome for discovering the full components of the hypothetical LFC.

Conclusion

Herein we used integrated mass spectrometry analysis at the tryptic peptide, intact protein, and native protein complex level to identify potential components of a hypothetical LFC in *Forsythia* × *intermedia*. We incorporated published, but incomplete, transcriptomic data from closely related plant species to "*de novo*" sequence proteins in the DP-enriched samples. Sequences were verified by the recently published genome of a similar plant, *Forsythia suspensa*,[24] confirming the feasibility of this approach and its potential for other plant systems. The hypothetical LFC (or part of the LFC) may have survived earlier stages but disassembled into smaller components after extensive purification steps. We plan to improve isolation protocols to better preserve the putative LFCs in the near future.

Nevertheless, we focused on identification of previously uncharacterized protein components in the fraction associated with (+)-pinoresinol forming activity, including the expected oxidase(s). We identified a peroxidase, a beta-fructofuranosidase (invertase), a non-specific lipid transfer protein (nsLTP), and two new DP homologs that were considerably different from the DPs first discovered. Other proteins detected in low abundance included germin-like proteins and a Cu binding protein of the cupredoxin family. Some of them may be weakly associated with DPs with functional or structural roles in the putative LFC, but this remains to be proven. In future work, it will be thus instructive to determine whether these identified proteins are part of a hypothetical LFC or not, and can serve as targets for studying DPs and LFCs in other plants with sequenced genomes in the future.

Interestingly, we also identified hetero-trimers of two DP homologs using native MS. MD simulations showed that stable hetero-trimers of DPs are also possible between close homologs in Arabidopsis. Such hetero-association among enzyme homologs may be more commonly present but largely undiscovered and unexplored. Although their functional biochemical roles are unclear, the results demonstrate the power of native MS for identifying heterogeneity and hetero-association of proteins in a discovery mode directly from native plant protein extracts, even when a complete and annotated genome is not yet available.

Author Contributions

Conceptualization: MZ, JRC, LBD, NGL. Formal analysis: MZ, JAL, CJB, MK, BG, LBD, IO, NGL. Investigation: MZ, JAL, IVN, QM, CDN, RLS, DLB, LBD, NGL. Project administration: MZ. Resources: LBD, NGL. Writing – original draft: MZ, JAL, JRC, CJB, MK. Writing – review & editing: LBD, NGL, IVN, EDM, BG.

Conflict of Interest

There are no conflicts to declare.

Acknowledgement

We thank Ronald Moore, Thomas Fillmore, and Jared Shaw for helping with MS experiments. We also thank Prof. Xiaowen Liu for discussion of top-down data analysis using TopPIC. The research was supported by the Intramural program at EMSL (grid.436923.9), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and operated under Contract No. DE-AC05-76RL01830. Contributions from the Lewis laboratory are supported the USDA National Institute of Food and Agriculture, Hatch umbrella project #1015621 as well as by the Arthur and Katie Eisig-Tode endowment.

Footnotes

Electronic supplementary information (ESI) available: Table S1-S3 for protein sequences. Figure S1-S20 for supporting data.

References

1 L. B. Davin, H. Bin Wang, A. L. Crowell, D. L. Bedgar, D. M. Martin, S. Sarkanen and N. G. Lewis, *Science (80-.).*, 1997, **275**, 362–366.

2 D. R. Gang, M. A. Costa, M. Fujita, A. T. Dinkova-Kostova, H.-B. Wang, V. Burlat, W. Martin, S. Sarkanen, L. B. Davin and N. G. Lewis, *Chem. Biol.*, 1999, **6**, 143–151.

3 K. W. Kim, C. A. Smith, M. D. Daily, J. R. Cort, L. B. Davin and N. G. Lewis, *J. Biol. Chem.*, 2015, **290**, 1308–1318.

4 R. Gasper, I. Effenberger, P. Kolesinski, B. Terlecka, E. Hofmann and A. Schaller, *Plant Physiol.*, 2016, **172**, 2165 LP – 2175.

5 K.-W. Kim, S. G. A. Moinuddin, K. M. Atwell, M. A. Costa, L. B. Davin and N. G. Lewis, *J. Biol. Chem.* , 2012, **287**, 33957–33972.

6 I. Effenberger, B. Zhang, L. Li, Q. Wang, Y. Liu, I. Klaiber, J. Pfannstiel, Q. Wang and A. Schaller, *Angew. Chemie Int. Ed.*, 2015, **54**, 14660–14663.

7 Q. Meng, S. G. A. Moinuddin, S. J. Kim, D. L. Bedgar, M. A. Costa, D. G. Thomas, R. P. Young, C. A. Smith, J. R. Cort, L. B. Davin and N. G. Lewis, *J. Biol. Chem.*, 2020, **295**, 11584–11601.

8 P. S. Hosmani, T. Kamiya, J. Danku, S. Naseer, N. Geldner, M. L. Guerinot and D. E. Salt, *Proc. Natl. Acad. Sci.*, 2013, **110**, 14498–14503.

9 K. Yonekura-Sakakibara, M. Yamamura, F. Matsuda, E. Ono, R. Nakabayashi, S. Sugawara, T. Mori, Y. Tobimatsu, T. Umezawa and K. Saito, *Plant Cell*, 2021, **33**, 129–152.

10 C. Corbin, S. Drouet, L. Markulin, D. Auguin, É. Lainé, L. B. Davin, J. R. Cort, N. G. Lewis and C. Hano, *Plant Mol. Biol.*, 2018, **97**, 73–101.

11 C. Paniagua, A. Bilkova, P. Jackson, S. Dabravolski, W. Riber, V. Didi, J. Houser, N. Gigli-Bisceglia, M. Wimmerova, E. Budínská, T. Hamann and J. Hejatko, *J. Exp. Bot.*, 2017, 68, 3287–3301.

12 Q. Liu, L. Luo and L. Zheng, *Int. J. Mol. Sci.*, 2018, 19.

13 R. A. Dixon and J. Barros, *Open Biol.*, 2019, **9**, 190215.

14 L. B. Davin and N. G. Lewis, *Curr. Opin. Biotechnol.*, 2005, **16**, 407–415.

15 L. B. Davin, M. Jourdes, A. M. Patten, K.-W. Kim, D. G. Vassão and N. G. Lewis, *Nat. Prod. Rep.*, 2008, **25**, 1015–1090.

16 B. J. Schultz, S. Y. Kim, W. Lau and E. S. Sattely, *J. Am. Chem. Soc.*, 2019, **141**, 19231–19235.

17 M. T. Marty, A. J. Baldwin, E. G. Marklund, G. K. A. Hochberg, J. L. P. Benesch and C. V. Robinson, *Anal. Chem.*, 2015, **87**, 4370–4376.

18 T. Unver, Z. Wu, L. Sterck, M. Turktas, R. Lohaus, Z. Li, M. Yang, L. He, T. Deng, F. J. Escalante, C. Llorens, F. J. Roig, I. Parmaksiz, E. Dundar, F. Xie, B. Zhang, A. Ipek, S. Uranbey, M. Erayman, E. Ilhan, O. Badad, H. Ghazal, D. A. Lightfoot, P. Kasarla, V. Colantonio, H. Tombuloglu, P. Hernandez, N. Mete, O. Cetin, M. Van Montagu, H. Yang, Q. Gao, G. Dorado and Y. Van de Peer, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, E9413–E9422.

19 M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev, *Nat. Biotechnol.*, 2011, **29**, 644–652.

20 T. D. Wu, J. Reeder, M. Lawrence, G. Becker and M. J. Brauer, *Methods Mol. Biol.*, 2016, **1418**, 283–334.

21 J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov, *Nat. Biotechnol.*, 2011, **29**, 24–26.

22 Q. Kou, L. Xun and X. Liu, *Bioinformatics*, 2016, **32**, btw398.

23 J. Park, P. D. Piehowski, C. Wilkins, M. Zhou, J. Mendoza, G. M. Fujimoto, B. C. Gibbons, J. B.

Shaw, Y. Shen, A. K. Shukla, R. J. Moore, T. Liu, V. A. Petyuk, N. Tolić, L. Paša-Tolić, R. D. Smith, S. H. Payne and S. Kim, *Nat. Methods*, 2017, **14**, 909–914.

24 L.-F. Li, S. A. Cushman, Y.-X. He and Y. Li, *Hortic. Res.*, 2020, **7**, 130.

25 J. Yang and Y. Zhang, *Nucleic Acids Res.*, 2015, **43**, W174–W181.

26 A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore and T. Schwede, *Nucleic Acids Res.*, 2018, **46**, W296–W303.

27 D. S. Cerutti, J. E. Rice, W. C. Swope and D. A. Case, *J. Phys. Chem. B*, 2013, **117**, 2328–2338.

28 R. Kumari, R. Kumar and A. Lynn, *J. Chem. Inf. Model.*, 2014, **54**, 1951–1962.

29 S. C. Halls, L. B. Davin, D. M. Kramer and N. G. Lewis, *Biochemistry*, 2004, **43**, 2587–2595.

30 N. G. Lewis, L. B. Davin and S. Sarkanen, *The nature and function of lignins*, Elsevier Science, 1999.

31 A. Shiraishi, J. Murata, E. Matsumoto, S. Matsubara, E. Ono and H. Satake, *PLoS One*, 2016, **11**, e0164805.

32 S. Subramanyam, C. Zheng, J. T. Shukle and C. E. Williams, *Arthropod. Plant. Interact.*, 2013, **7**, 389–402.

33 M.-K. Tan, M. El-Bouhssini, L. Emebiri, O. Wildman, W. Tadesse and F. C. Ogbonnaya, *Mol. Breed.*, 2015, **35**, 216.

34 K. Uchida, T. Akashi and T. Aoki, *Plant Cell Physiol.*, 2017, **58**, 398–408.

35 H. Funatsuki, M. Suzuki, A. Hirose, H. Inaba, T. Yamada, M. Hajika, K. Komatsu, T. Katayama, T. Sayama, M. Ishimoto and K. Fujino, *Proc. Natl. Acad. Sci.*, 2014, **111**, 17797 LP – 17802.

36 T. A. Parker, J. C. B. Mier y Teran, A. Palkovic, J. Jernstedt and P. Gepts, *bioRxiv*, 2019, 517516.

37 I. Letunic and P. Bork, *Nucleic Acids Res.*, 2021, **49**, W293–W296.

38 F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson and D. G. Higgins, *Mol. Syst. Biol.*, , DOI:10.1038/msb.2011.75.

39 A. Hassinen, A. Rivinoja, A. Kauppila and S. Kellokumpu, *J. Biol. Chem.* , 2010, **285**, 17771–17777.

40 A. J. M. Ribeiro, S. Das, N. Dawson, R. Zaru, S. Orchard, J. M. Thornton, C. Orengo, E. Zeqiraj, J. M. Murphy and P. A. Eyers, *Sci. Signal.*, 2019, **12**, eaat9797.

41 R. Mishra, H. Patel, S. Alanazi, L. Yuan and J. T. Garrett, *Oncol. Rev.*, 2018, **12**, 355.

42 R. K. Proels and R. Hückelhoven, *Mol. Plant Pathol.*, 2014, **15**, 858–864.

43 E. J. Woo, J. M. Dunwell, P. W. Goodenough, A. C. Marvier and R. W. Pickersgill, *Nat. Struct. Biol.*, 2000, **7**, 1036–1040.

44 C. Huang, R. Zhang, J. Gui, Y. Zhong and L. Li, *Plant Physiol.*, 2018, **177**, 671–683.

45 D. Patnaik and P. Khurana, *Indian J. Exp. Biol.*, 2001, **39**, 191–200.

46 L. B. Davin and N. G. Lewis, *Phytochem. Rev.*, 2003, **2**, 257.

47 C. Kazenwadel, J. Klebensberger, S. Richter, J. Pfannstiel, U. Gerken, B. Pickel, A. Schaller and B. Hauer, *Appl. Microbiol. Biotechnol.*, 2013, **97**, 7215–7227.

48 I. V. Novikova, N. Sharma, T. Moser, R. Sontag, Y. Liu, M. J. Collazo, D. Cascio, T. Shokuhfar, H. Hellmann, M. Knoblauch and J. E. Evans, *Adv. Struct. Chem. Imaging*, 2018, **4**, 13.

49 M. M. Chen, A. I. Bartlett, P. S. Nerenberg, C. T. Friel, C. P. R. Hackenberger, C. M. Stultz, S. E.

Radford and B. Imperiali, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 22528–22533.

50 M. Kyriakidou, H. H. Tai, N. L. Anglin, D. Ellis and M. V Strömvik, *Front. Plant Sci.* , 2018, 9, 1660.

51 I. O'Bryon, S. C. Jenson and E. D. Merkley, *Protein Sci.*, 2020, **29**, 1864–1878.

52 X. Han, L. He, L. Xin, B. Shan and B. Ma, *J. Proteome Res.*, 2011, **10**, 2930–2936.

53 F. V. Leprevost, R. H. Valente, D. B. Lima, J. Perales, R. Melani, J. R. Yates, V. C. Barbosa, M. Junqueira and P. C. Carvalho, *Mol. Cell. Proteomics*, 2014, **13**, 2480–2489.