

Supplementary Table 1. Coverage of annotated gene models by the EST sequences generated from sequencing of leaf transcriptome in *P. deltoides*. The 100 gene models encoding small proteins (<200AA in length) were randomly selected from the *P. trichocarpa* genome annotation v1.1. The sequence coverage information was obtained from *Populus* genome browser (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html), in which the ~634,000 ESTs from one of the six leaf samples were mapped onto the genome sequence of *P. trichocarpa*. “NA” represents no EST mapped; “Full” represents complete coverage; “Alternative” represents alternative splicing; and “Partial” represents partial coverage.

Gene model	Protein length, AA	Coverage by EST
fgenes4_pg.C_LG_XVIII000639	107	NA
eugene3.01780018	59	NA
eugene3.00010568	63	NA
eugene3.00061765	100	NA
eugene3.00700078	177	Full
eugene3.00002665	52	Full
fgenes4_pg.C_LG_I001604	147	NA
eugene3.15770001	174	NA
estExt_fgenes4_pm.C_LG_II0207	135	Full
eugene3.13070001	101	NA
eugene3.190770001	65	NA
eugene3.00101285	95	NA
eugene3.27800003	126	NA
fgenes4_pg.C_scaffold_119000049	161	NA
fgenes4_pg.C_scaffold_3615000001	109	NA
eugene3.00150970	169	Full
eugene3.10910001	156	NA
grail3.0101004301	81	Full
estExt_fgenes4_pg.C_LG_XV0532	127	Full
eugene3.00030963	75	NA
eugene3.00061340	136	Partial
estExt_Genewise1_v1.C_LG_XIII0862	129	Partial
estExt_fgenes4_pm.C_1290009	174	Partial
fgenes4_pg.C_scaffold_162000018	117	NA
eugene3.00180409	125	NA
grail3.0445000101	58	NA
eugene3.28760003	53	NA
eugene3.11070002	174	NA
eugene3.00130614	190	Full
fgenes4_pg.C_LG_XVI000527	124	NA
estExt_fgenes4_pg.C_LG_V1665	187	Full
eugene3.00141145	194	NA
fgenes4_pg.C_scaffold_66000143	147	NA

eugene3.00190786	121	NA
grail3.0019030801	104	Full
eugene3.00410107	112	NA
grail3.0021029701	96	Full
estExt_fgenes4_pg.C_LG_II1142	124	Full
eugene3.00110148	70	Alternative
fgenes4_pg.C_LG_XIV000628	141	NA
eugene3.00570218	91	NA
eugene3.200290001	76	NA
eugene3.00290069	142	NA
estExt_Genewise1_v1.C_LG_XIV2676	104	NA
fgenes4_pg.C_LG_II000694	190	NA
grail3.0022010101	180	NA
grail3.16813000101	161	NA
grail3.0002045101	92	Full
eugene3.219150001	197	NA
eugene3.00161257	59	NA
eugene3.01810028	50	NA
eugene3.00111138	133	NA
eugene3.79760001	79	NA
eugene3.45990001	105	NA
eugene3.00062018	112	NA
eugene3.108760001	54	NA
eugene3.00102388	71	NA
eugene3.00040638	69	NA
eugene3.00150749	64	NA
fgenes4_pg.C_LG_X001082	131	NA
fgenes4_pg.C_LG_XVIII001167	122	NA
eugene3.00161291	73	NA
fgenes4_pg.C_scaffold_165000007	176	NA
eugene3.00021489	153	NA
eugene3.02280026	80	NA
fgenes4_pg.C_scaffold_40000001	165	NA
eugene3.00060759	144	NA
grail3.0001129601	122	Partial
eugene3.75900001	117	NA
estExt_fgenes4_pg.C_LG_XVI0536	190	Full
eugene3.12050002	116	NA
eugene3.00110911	145	NA
eugene3.02350013	127	NA
grail3.0040012101	182	NA
fgenes4_pg.C_LG_IX001097	122	NA
estExt_Genewise1_v1.C_LG_XV0145	111	NA
fgenes4_pg.C_LG_XIX000150	142	NA

fgenes4_pg.C_LG_XV000725	187	NA
fgenes4_pg.C_LG_XI001122	141	NA
fgenes4_pg.C_LG_VI000303	125	Alternative
fgenes4_pm.C_LG_V000320	151	NA
eugene3.01340005	120	Partial
fgenes4_pg.C_LG_III000865	161	NA
eugene3.00010416	107	NA
fgenes4_pg.C_LG_X000888	137	NA
eugene3.02000007	122	NA
eugene3.00290181	165	Full
eugene3.00170448	93	NA
fgenes4_pg.C_LG_X001405	181	NA
eugene3.13770002	159	NA
grail3.1481000101	81	NA
fgenes4_pg.C_LG_VI000745	154	Full
eugene3.13940002	94	NA
estExt_Genewise1_v1.C_LG_X5756	162	Alternative
fgenes4_pg.C_scaffold_5441000001	190	NA
fgenes4_pm.C_LG_I000862	191	NA
grail3.0069006401	67	NA
eugene3.00101430	50	Full
eugene3.00150636	63	Full
eugene3.28340001	148	NA

Supplementary Table 2. The sORF genes and gene-specific primers used for full-length cDNA cloning by Rapid Amplification of cDNA Ends.

ID	Accession ^a	Protein length (AA)	Direction	Gene specific primer pair (5' - 3')
PDSP05367	HP451655	187	Forward	TGCATCGTGATCTTAAGCCCCAGCAA
			Reverse	GCAACTCCGGTGCTCTGTACCAACGA
PDSP09810	HP451656	87	Forward	CTGCCAGGCGGAGAACCTGTGATGGA
			Reverse	TGCCGTGAGCCAGGCTGCAATTGAGTTC
PDSP00948	HP451657	160	Forward	GAGAGGGGAAAAGGGGATCGGAGATGGA
			Reverse	GCCTGCAGGAGGAATTGGGCATGCATCT
PDSP00613	HP451658	135	Forward	CCTGTTGGGCGTATCCACCGCCTTCTG
			Reverse	CAGCTGCCGTTGCACCCACTCTTCCA
PDSP01297	HP451659	135	Forward	CCTGTCGGGCGTATCCACCGCCTTCT
			Reverse	CGGCTGCTGTTGCTCCCACTCTTCCA
PDSP06085	HP451660	135	Forward	GAGCTGGCCGGAATGCAAGCAAGGA
			Reverse	GGACCGGGGGCCAATTATCACCCAATGT
PDSP05937	HP451661	107	Forward	CCCTGGCGGCAACATGCTTCGATTCT
			Reverse	GGGCGAGATCTTCAAGCCCGGCGTATC
PDSP06903	HP451662	104	Forward	GGGAGGAGCCAAGAGGCATCGCAAGGT
			Reverse	GGCAGGCTTGGTGATGCCCTGGATGT
PDSP05999	HP451663	104	Forward	GGGAGGTGCAAAGAGGCACAGGAAGGTG
			Reverse	TCCTGTGCCTCTTTGCACCTCCCTTTCC
PDSP04099	HP451664	195	Forward	CAAGACCTCATCGCTTGGGGTCGAGGAA
			Reverse	TCGCCAAAGCTTCAGGGAAACCCATCTC
PDSP03630	HP451665	166	Forward	GTGCAGCCAGTTCCCTCGCTCCCAAA
			Reverse	CGGAGGCCCTTCCAGTCCTTGGCAGTT
PDSP01844	HP451666	183	Forward	TTGGGATCTCGGTGGGCAGACAAGCA
			Reverse	CGCCTCAGTCACAGCCGCATCATCAA
PDSP11241	HP451667	175	Forward	TGCAGTTGGAGGTCTGCCTGGTCAA
			Reverse	TTGCCTGGCATGGAAGGAAGGATCA
PDSP02892	HP451668	123	Forward	TTCGCAGATACGGGCTCCGATACGATGA
			Reverse	CGTCGACGATCTCTTGGGAGCCGATT
PDSP11512	HP451669	197	Forward	GCGGTACTGGTGGCGGTGGTGAAAGTG
			Reverse	ACGGGGTCCAACCGGTCCACAAAAGG

^aGenBank Transcriptome Shotgun Assembly (TSA) accession numbers.

Supplementary Table 3. Summary of the sORF-encoded proteins identified by mass spectrometry. The data is in file titled “Supplementary Table 3_Protein MS master_table.xls” The LocusID column lists the 1,158 sORF-encoded proteins for which peptide identifications were obtained. For each of the 27 LC-MS-MS analyses performed, the table contains four columns:

Peps: Peptides – the number of tryptic peptides from the LocusID protein identified by mass spectrometry. Note that for cases in which Peps = 1, independent identification was obtained from multiple charge states of that peptide.

SC: Spectrum Count - the number of tandem mass spectra that were identified for peptides in the protein.

Cov: % Sequence Coverage - The percentage of amino acid residues in the protein belonging to peptides identified by mass spectrometry.

Uniq: Unique Peptides – Number of tryptic peptides identified for the protein in the LocusID column that are not contained in the sequence of any other protein in the database used for Sequest identification.

These four columns are repeated for each of the LC-MS-MS analyses performed, which are listed below. See Methods in the main text for further details on leaf, xylem, and phloem samples. MS data for xylem samples containing “Kalluri” in the identifiers below were obtained previously (Kalluri et al. 2009), and re-analyzed using Sequest along with the protein database containing sORF-encoded proteins in addition to the current *Populus* annotation (see Methods in main text). TR = technical replicate.

Leaf Proteomes	Xylem and Phloem Proteomes	Published Xylem Proteomes (Kalluri et al. 2009)
leaf_pool1_TR1	phloem_pool1_Pellet	xylem_Crude1_Kalluri
leaf_pool1_TR2	phloem_pool1_Crude	xylem_Crude2_Kalluri
leaf_pool1_TR3	xylem_pool1_Pellet	xylem_Nuclei1_Kalluri
leaf_pool2_TR1	xylem_pool1_Crude	xylem_Nuclei2_Kalluri
leaf_pool2_TR2	phloem_pool2_Pellet	xylem_Pellet1_Kalluri
leaf_pool2_TR3	phloem_pool2_Crude	xylem_Pellet2_Kalluri
leaf_pool3_TR1	xylem_pool2_Pellet	
leaf_pool3_TR2	xylem_pool2_Crude	
leaf_pool3_TR3	phloem_pool3_Pellet	
	phloem_pool3_Crude	
	xylem_pool3_Pellet	
	xylem_pool3_Crude	

The final four columns at the right hand side of the table provide the number of amino acids (L), the calculated molecular mass (MW), the isoelectric point (pI), and number of LC-MS-MS

analyses in which the LocusID protein was detected with Parsimony status of DS, DF, or df (see
Supplementary Table 4).

Supplementary Table 4. Peptides identified by mass spectrometry from sORF-encoded proteins.

The data is in file titled “Supplementary Table 4_Protein MS_peptides.txt”. Columns in this tab-delimited text file are as follows:

Run: LC-MS-MS experiment, as listed above for Supplementary Table 3.

LocusID: sORF-encoded protein identifier

SpectrumCount: the number of tandem mass spectra that were identified for peptides in the LocusID protein.

Coverage: % Sequence Coverage - The percentage of amino acid residues in the protein belonging to peptides identified by mass spectrometry.

IdentifiedPeptides: the number of tryptic peptides from the LocusID protein identified by mass spectrometry. Note that for cases in which IdentifiedPeptides = 1, independent identification was obtained from multiple charge states of that peptide.

Peptide: amino acid sequence of the identified peptide

Unique: column contains an asterisk for unique peptides (i.e. those only occurring in a single protein in the database used for Sequest searches), or is blank otherwise.

Parsimony: Summary of peptides shared between the LocusID protein and other proteins in the sORF and previous *Populus* annotations. The codes in the column are as follows (Yang et al. 2004):

DS: Distinct – All tryptic peptides identified by MS in the LocusID protein are unique to the LocusID protein in the database used for Sequest searches.

DF: Differentiable – One or more (but not all) tryptic peptides identified by MS in the LocusID protein are unique to the LocusID protein in the database used for Sequest searches.

df: differentiable among identified – One or more tryptic peptides identified by MS in the LocusID protein is not found in any other MS-identified protein (but is present in another protein in the database).

I: Indistinguishable - The LocusID protein shares all identified peptides with one or more other proteins in the database used for Sequest searches; proteins indistinguishable from the LocusID protein are listed in parentheses in the parsimony code: I (list)

SE: Subset – All identified peptides in the LocusID protein are a subset of the identified peptides in one or more other proteins in the database used for Sequest searches; the proteins for which the LocusID protein is a subset are listed in parentheses: SE (list)

SU: Subsumable – All the identified peptides in the LocusID protein can be found across two or more other proteins in the database used for Sequest searches. Proteins into which the LocusID protein is subsumable are listed in parentheses: SU (list)

O: Other – The LocusID protein shares peptides with other proteins, but not in the relationships listed above. For example, a protein in the O category may be a subset of the LocusID protein.

Note that the parsimony code may include more than one of the above categories. For example, a protein might be indistinguishable from some proteins, but also a subset of other proteins, leading to a code such as I (list) SE (list).

Supplementary Table 5. The sORF candidates in the *highest-confidence sORF candidate set*. Only the sORF candidates encoding protein sequences new to the *P. trichocarpa* genome annotation v2.0 (<http://www.phytozome.net/>) were listed.

ID	Accession ^a	Protein length AA	ID	Accession ^a	Protein length AA
PDSP11241	HP451667	174	PDSP03194	HP451697	134
PDSP02892	HP451668	122	PDSP04937	HP451698	151
PDSP11512	HP451669	196	PDSP01087	HP451699	117
PDSP12097	HP451670	154	PDSP04965	HP451700	66
PDSP07460	HP451671	169	PDSP01042	HP451701	165
PDSP04086	HP451672	136	PDSP04021	HP451702	138
PDSP02620	HP451673	148	PDSP03443	HP451703	113
PDSP11368	HP451674	97	PDSP05683	HP451704	49
PDSP10717	HP451675	152	PDSP08829	HP451705	108
PDSP06394	HP451676	109	PDSP00825	HP451706	174
PDSP00395	HP451677	123	PDSP08147	HP451707	147
PDSP01365	HP451678	86	PDSP10387	HP451708	148
PDSP09877	HP451679	140	PDSP08858	HP451709	191
PDSP04469	HP451680	83	PDSP03833	HP451711	159
PDSP05530	HP451681	142	PDSP00288	HP451712	119
PDSP02581	HP451682	155	PDSP08675	HP451713	199
PDSP10365	HP451683	177	PDSP08721	HP451714	158
PDSP08992	HP451684	199	PDSP00601	HP451715	171
PDSP12811	HP451685	182	PDSP07864	HP451716	181
PDSP11910	HP451686	170	PDSP07655	HP451717	159
PDSP08449	HP451687	109	PDSP06405	HP451718	163
PDSP10762	HP451690	178	PDSP11415	HP451719	160
PDSP03795	HP451691	80	PDSP03404	HP451720	104
PDSP06922	HP451692	91	PDSP05895	HP451721	170
PDSP11000	HP451693	192	PDSP03310	HP451722	154
PDSP04206	HP451694	86	PDSP09876	HP451723	170
PDSP02066	HP451695	111	PDSP01487	HP451724	108
PDSP08503	HP451696	151	PDSP03924	HP451725	191

^aGenBank Transcriptome Shotgun Assembly (TSA) accession numbers.

References:

Kalluri UC, Hurst GB, Lankford PK, Ranjan P, Pelletier DA. 2009. Shotgun proteome profile of *Populus* developing xylem. *Proteomics* **9**(21): 4871-4880.

Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, Epstein J, Chen X, Markey SP, Kowalak JA. 2004. DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* **3**(5): 1002-1008.