PLAMED-2010-04-0432-OP.R1

Supporting Information

Transcriptome analysis of *Taxus cuspidata* needles based on 454 pyrosequencing

Qiong Wu[1], Chao Sun[1], Hongmei Luo[1], Ying Li[1], Yunyun Niu[1], Yongzhen Sun[1], Aiping Lu[2], Shilin Chen[1]

Affiliation

[1] Institute of Medicinal Plant Development (IMPLAD), Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[2] Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China

Correspondence

Prof. Dr. Shilin Chen

Institute of Medicinal Plant Development (IMPLAD)

Chinese Academy of Medical Sciences & PekingUnion Medical College

No. 151, Malianwa North Road, HaiDian District

Beijing 100193

People's Republic of China

Tel.: +86/10/628/99700

Fax: +86/10/628/99776

slchen@implad.ac.cn

Text 1

Plant material

Taxus cuspidata Sieb. et Zucc. was collected from cultivated fields (116°16′E, 40°01′N) in the Institute of Medicinal Plant Development (IMPLAD), Beijing, China, in April 2009. The specimen was identified by Prof. Zhao Zhang of IMPLAD. A voucher specimen (HDS20090417) has been deposited in the herbarium of IMPLAD. Needles were collected from a field-growing female plant, which was initiated from a seedling through an approximately 15-years growth period. After cleaning, needles were immediately frozen in liquid nitrogen and stored at -80° C until use.

Text 2

RNA extraction and cDNA library construction

Total RNA was extracted from 3 g needles of *T. cuspidata* following the standard protocol of RNeasy Plant Kit (BioTeke). Approximately 2.52 μg poly(A)$^+$ RNA was isolated using Oligotex○,R mRNA Midi Kit (Qiagen), while cDNA was synthesized using the SMART™ PCR cDNA Synthesis Kit (Clontech). First-strand cDNA synthesis was performed with 3′ SMART CDS Primer II A, as described in the provided protocol, using 0.9 μg purified poly(A)$^+$ RNA (Clontech). Double-stranded cDNA was prepared from 2 μL of the first strand by polymerase chain reaction (PCR) with 5′ PCR Primer II A in a 100 μL reaction and amplified using PCR Advantage II polymerase. The following thermal profile was applied: 1 min at 95° C followed by 13 cycles of 95° C for 15 s, 65° C for 30 s, and 68° C for 6 min (Clontech). Amplified cDNA product was purified with the PureLink™ PCR Purification Kit (Invitrogen) using Buffer HC to remove fragments less than 300 bp.

Text 3

454 Library preparation and sequencing

Approximately 5 μg amplified cDNA were sheared by nebulization to produce random fragments of about 500 bp in length for 454 sequencing. The FLX-specific adapters, Adapter A (GCCTCCCTCGCGCCATCAG) and Adapter B (GCCTTGCCAGCCCGCTCAG), were ligated to the fragmented cDNA samples resulting in Adapter A-DNA fragment-Adapter B constructs. The fragment samples were denatured to generate single-stranded DNA amplified by emulsion PCR for the construction of the libraries. A 454-GS FLX Titanium sequencing platform and the GS FLX Titanium Kit (Roche Diagnostic) were used for pyrosequencing.

Text 4

Sequence assembly and annotation

Reads generated by the GS FLX Titanium sequencer were trimmed of low quality, low complexity [poly (A/T)], and adaptor sequences using 454 commercial software utilities (Roche Diagnostic). Sequences shorter than 50 bp were removed from the high-quality sequences before assembly. Derived high-quality reads were assembled into unique sequences using the Newbler Assembler software v2.0.01.14 (Roche Diagnostic), with a quality score threshold set at 40. All obtained unique sequences involving contigs and singletons were subjected to BLASTX comparisons with the SwissProt protein database (released on June 19, 2009) and the non-redundant (Nr) protein database (released on June 23, 2009), respectively, in local servers [1]. Subject sequences with the best scores and a maximum e-value cutoff of 1.0e-5 were used to annotate the unique sequences.

Text 5

Sequence comparisons with other species

Similarity searches were performed at different e-value cutoffs with the tBLASTX programs against the Dana-Farber Cancer Institute gene indices available for oryza (OGI17.0), pine (PGI7.0), poplar (PPLGI4.0), and spruce (SGI3.0), retrieved from the Dana-Farber Cancer Institute Web site (http://compbio.dfci.harvard.edu/tgi/) [1,2]. BLASTX searches were conducted against the SwissProt protein database (released on June 19, 2009) and the *A. thaliana* proteome database (version TAIR9) (http://www.arabidopsis.org) at different stringent cutoffs. All sequences were retrieved with the newest version from the respective databases and calculated in local servers.

Text 6

Gene ontology (GO) analysis and metabolic pathway assignment using the Kyoto Encyclopedia of Genes and Genomes (KEGG)

To correlate the yew unique sequences to GO controlled vocabularies, the annotations of homologous *Arabidopsis* proteins and cDNA sequences (TAIR9) were analyzed. Each of the unique sequences was assigned GO terms based on the top BLAST hit for queries in the *Arabidopsis* proteins and cDNA sequences (TAIR9), and the GO slim categories of *Arabidopsis* Information Resource were used for GO classifications [3]. The GO provided a systematic and consistent description of gene attributes in three key biological domains: molecular function, biological process, and cellular component.

Metabolic pathway assignments were carried out according to the KEGG resource (version KEGG 50). The enzyme commission (EC) number was assigned to unique sequences based on BLASTX similarity search with an e-value cutoff of 1.0e-5. Sequences were mapped to KEGG biochemical pathways according to the EC distribution in the pathway databases [4].

Text 7

Identification of simple sequence repeats (SSR) and Sanger sequencing validation

The search for the presence of SSRs was performed with Perl script SSR identification tool (http://www.gramene.org/db/markers/ssrtool). In this study, the search was restricted to motifs having at least 14 bp in length. The criteria for selection of a minimum of seven repeats for dinucleotide motifs, five repeats for trinucleotide motifs, four repeats for tetranucleotide motifs, and three repeats for pentanucleotide and hexanucleotide motifs were used. For validation of the SSRs, we randomly selected 10 SSRs, and primer sets were designed according to the adjacent sequences of the selected motif using Primer Premier 5. The PrimeScript^{TM} 1st Strand cDNA Synthesis Kit (TaKaRa) was used with approximately 1 μg of the total RNA to synthesize a single-strand cDNA. The reaction mixture (25 μL) containing 10×LA PCR Buffer II (Mg^{2+} Plus) 2.5 μL, dNTP mixture (2.5 mM each) 4μL, TaKaRa LA Taq (5U/μL) 0.25 μL (TaKaRa), 1 μL each of forward and reverse primers (10 μM), and 1 μL template cDNA, was added into sterilized water of up to 25 μL. PCR amplification was performed under the following conditions: 95° C for 1 min, followed by 35 cycles of 94° C for 30 s, 55° C 30 s and then 72° C for 1 min. The PCR products of interest were gel-purified using the TIANquick Midi

Purification Kit (Tiangen Biotech). Isolated fragments were ligated into pGEM®,R T Easy vectors (Promega), and then cloned into *Escherichia coli* strain Trans5*α* (TransGen Biotech). After blue/white screening of recombinants, the interest clones were cultured and Sanger sequencing from both strands was performed.

(A)

(B)

Fig. 1S *Taxus cuspidata* with needles and fruits. (a) Long shot. (b) Close shot.

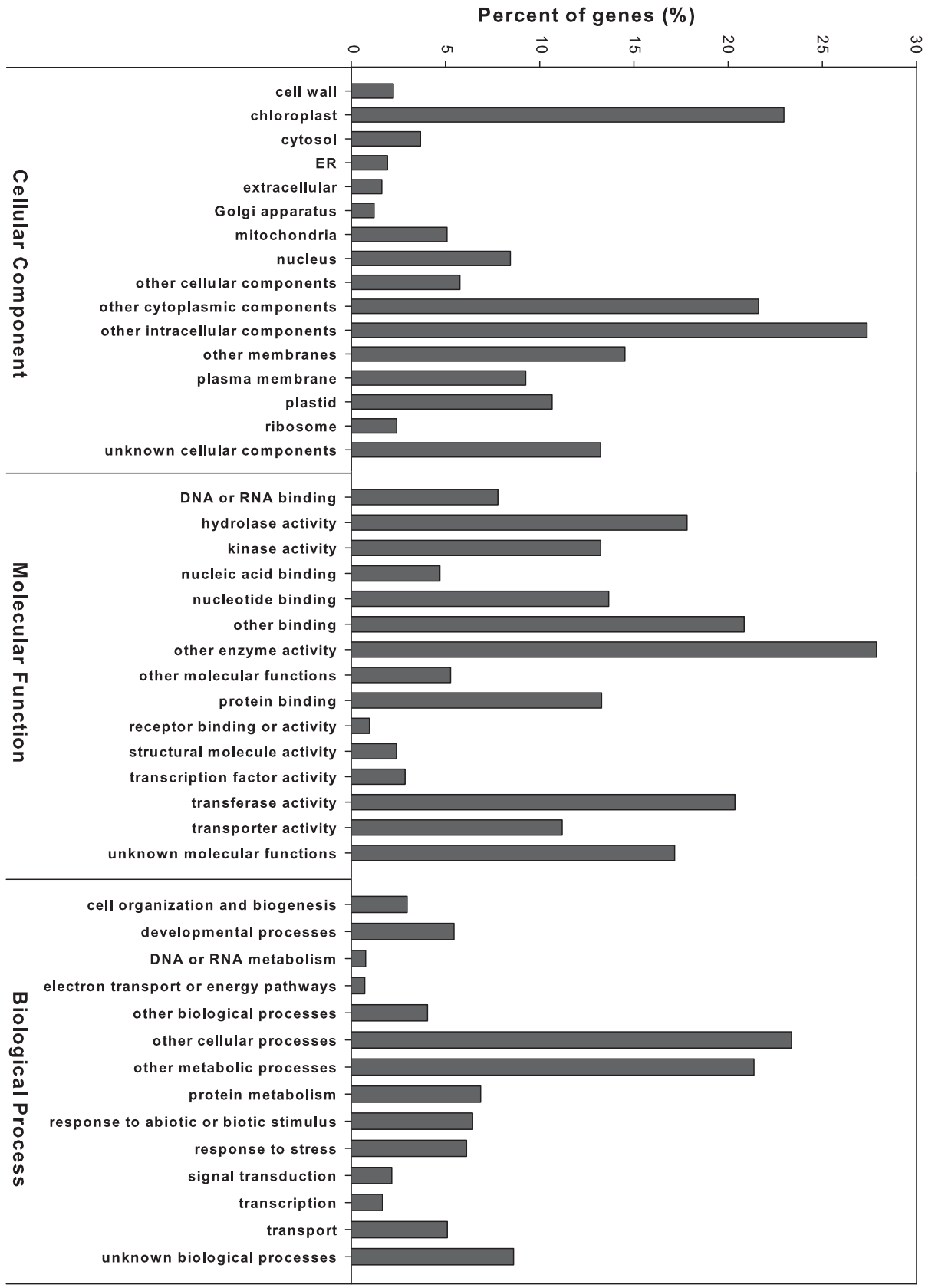

Fig. 2S Gene ontology (GO) analysis of *T. cuspidata* unique sequences based on cellular component, molecular function, and biological process.

Table 1S Predominant transcripts in the needles of *T. cuspidata*

Unigene	Putative Identity	Organism	E-value	Reads	Abundance (%)	Acc. No.
contig07549	Ribulose bisphosphate carboxylase small chain	*Capsicum annuum*	9.00E-30	1280	1.60	O65349
contig07488	Ribulose bisphosphate carboxylase small chain	*Larix laricina*	1.00E-14	1072	1.34	P16031
contig07477	Glycogenin-2	*Homo sapiens*	1.00E-09	758	0.95	O15488
contig00485	Ribulose bisphosphate carboxylase/oxygenase activase	*Malus domestica*	0	668	0.83	Q40281
contig00054	unknown	*Picea sitchensis*	2.00E-43	445	0.56	gb|ABR17981.1
contig07473	Photosystem I reaction center subunit XI	*Cucumis sativus*	5.00E-57	437	0.55	Q39654
contig00952	Glutamine synthetase cytosolic isozyme	*Pinus sylvestris*	0	387	0.48	P52783
contig00021	Caffeoyl-CoA O-methyltransferase	*Eucalyptus globulus*	8.00E-42	383	0.478	O81185
contig00390	Granule-bound starch synthase 1	*Manihot esculenta*	1.00E-161	376	0.47	Q43784
contig00906	L-ascorbate peroxidase, cytosolic	*Pisum sativum*	1.00E-108	343	0.428	P48534
contig00305	Chlorophyll a-b binding protein 6A	*Solanum lycopersicum*	3.00E-92	323	0.403	P12360
contig00147	Sedoheptulose-1,7-bisphosphatase	*Triticum aestivum*	1.00E-157	310	0.387	P46285
contig07480	Malate dehydrogenase	*Citrullus lanatus*	1.00E-150	297	0.371	P19446
contig07496	Carbonic anhydrase 2	*Arabidopsis thaliana*	3.00E-76	259	0.324	P42737
contig02104	Ribulose-phosphate 3-epimerase	*Spinacia oleracea*	1.00E-125	255	0.319	Q43157
contig00643	Photosystem II 22 kDa protein	*Solanum lycopersicum*	2.00E-80	250	0.312	P54773
contig00222	Inositol-3-phosphate synthase	*Sesamum indicum*	0	248	0.310	Q9FYV1
contig00943	Serine--glyoxylate aminotransferase	*Arabidopsis thaliana*	1.00E-176	248	0.307	Q56YA5
contig07424	Cathepsin B	*Rattus norvegicus*	2.00E-57	246	0.307	P00787
contig01106	1-aminocyclopropane-1-carboxylate oxidase	*Musa acuminata*	3.00E-90	237	0.296	Q9FR99
contig02928	Metallothionein-like protein EMB30	*Picea glauca*	7.00E-17	219	0.274	Q40854
contig01059	Probable peroxisomal (S)-2-hydroxy-acid oxidase	*Arabidopsis thaliana*	1.00E-166	216	0.270	Q9LRR9
contig00331	Chlorophyll a-b binding protein CP26	*Arabidopsis thaliana*	1.00E-51	215	0.269	Q9XF89
contig00392	Probable granule-bound starch synthase 1	*Arabidopsis thaliana*	8.00E-70	214	0.267	Q9MAQ0
contig00812	Fructose-bisphosphate aldolase	*Oryza sativa subsp. japonica*	1.00E-170	212	0.265	Q40677
contig00332	Chlorophyll a-b binding protein CP26	*Arabidopsis thaliana*	7.00E-16	208	0.260	Q9XF89
contig02089	Glyceraldehyde-3-phosphate dehydrogenase A	*Spinacia oleracea*	1.00E-161	203	0.254	P19866

Table 2S Mapping of *T. cuspidata* unique sequences to KEGG biochemical pathways

KEGG pathway	No. of sequences	Percent (%)
Metabolism		
Amino acid metabolism	457	11.96
Biosynthesis of polyketides and nonribosomal peptides	6	0.16
Biosynthesis of secondary metabolites	159	4.16
Carbohydrate metabolism	527	13.79
Energy metabolism	350	9.16
Glycan biosynthesis and metabolism	118	3.09
Lipid metabolism	244	6.39
Metabolism of cofactors and vitamins	129	3.38
Metabolism of other amino acids	107	2.80
Nucleotide metabolism	86	2.25
Xenobiotics biodegradation and metabolism	111	2.90
Genetic information processing		
Folding, sorting and degradation	214	5.60
Replication and repair	85	2.22
Transcription	59	1.54
Translation	236	6.18
Environmental information processing		
Membrane transport	56	1.47
Signal transduction	145	3.79
Signaling molecules and interaction	2	0.05
Cellular processes		
Behavior	4	0.10
Cell communication	57	1.49
Cell growth and death	79	2.07
Cell motility	20	0.52
Development	14	0.37
Endocrine system	111	2.90
Immune system	51	1.33
Nervous system	23	0.60
Sensory system	1	0.03
Human diseases		
Cancers	86	2.25
Immune disorders	20	0.52
Infectious diseases	40	1.05
Metabolic disorders	12	0.31
Neurodegenerative diseases	212	5.55
Unclassified	220	5.76
Unassigned	8,668	42.2[c]
Unannotated	9,266	45.1

Total	20,557

a. Of the 2,623 assigned to KEGG, only 2,403 were mapped to at least one KEGG pathway (note that individual KEGG categories may have multiple mappings); 220 were unclassified and only general functions were predicted. b. Account in a total of 3821 KEGG pathways. c. Account in total unique sequences.

Table 3S Candidate genes with functions similar to CYP450, epoxidase, CoA ligases, and *N*-benzoyltransferase in the taxol biosynthesis

Candidates	Organism	E-value	Unigene	Accession No.
Zeaxanthin epoxidase	*Solanum lycopersicum*	3.00E-27	FXAT9O003CW5TV	P93236
Zeaxanthin epoxidase	*Prunus armeniaca*	0	contig00328	O81360
Zeaxanthin epoxidase	*Nicotiana plumbaginifolia*	3.00E-15	contig01021	Q40412
Zeaxanthin epoxidase	*Prunus armeniaca*	1.00E-10	contig01640	O81360
Zeaxanthin epoxidase	*Prunus armeniaca*	9.00E-06	contig05732	O81360
Zeaxanthin epoxidase	*Nicotiana plumbaginifolia*	1.00E-06	contig05942	Q40412
Zeaxanthin epoxidase	*Prunus armeniaca*	4.00E-08	contig07494	O81360
Cytochrome P450 87A3	*Oryza sativa* subsp. *japonica*	2.00E-33	FXAT9O003C10S7	Q7XU38
Cytochrome P450 720B2	*Pinus taeda*	2.00E-07	FXAT9O003C2OPM	Q50EK5
Cytochrome P450 97B1	*Pisum sativum*	3.00E-61	FXAT9O003C4NRI	Q43078
Cytochrome P450 97B2	*Glycine max*	2.00E-19	FXAT9O003C4YE9	O48921
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	2.00E-06	FXAT9O003C6QYG	P37116
Cytochrome P450 77A3	*Glycine max*	2.00E-29	FXAT9O003C6VS3	O48928
Cytochrome P450 716B1	*Picea sitchensis*	5.00E-36	FXAT9O003C6Z98	Q50EK1
Cytochrome P450 716B2	*Picea sitchensis*	2.00E-31	FXAT9O003C72PY	Q50EK0
Cytochrome P450 750A1	*Pinus taeda*	3.00E-22	FXAT9O003C7NC4	Q50EK4
Cytochrome P450 720B2	*Pinus taeda*	3.00E-10	FXAT9O003C87J3	Q50EK5
Cytochrome P450 93A3	*Glycine max*	9.00E-32	FXAT9O003C8J5A	O81973
Cytochrome P450 86A1	*Arabidopsis thaliana*	6.00E-27	FXAT9O003C8Z50	P48422
Cytochrome P450 750A1	*Pinus taeda*	3.00E-37	FXAT9O003C98FI	Q50EK4
Cytochrome P450 750A1	*Pinus taeda*	1.00E-25	FXAT9O003C9GNJ	Q50EK4
Cytochrome P450 94A1	*Vicia sativa*	1.00E-42	FXAT9O003CVK2X	O81117
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	3.00E-47	FXAT9O003CVXTH	P37116
Cytochrome P450 71A14	*Arabidopsis thaliana*	3.00E-07	FXAT9O003CWYTD	P58045
Cytochrome P450 750A1	*Pinus taeda*	2.00E-48	FXAT9O003CY6DK	Q50EK4
Cytochrome P450 750A1	*Pinus taeda*	3.00E-41	FXAT9O003CZFIH	Q50EK4
Cytochrome P450 90C1	*Arabidopsis thaliana*	4.00E-34	FXAT9O003DAFHX	Q9M066
Cytochrome P450 86A2	*Arabidopsis thaliana*	2.00E-21	FXAT9O003DAP5M	O23066
Cytochrome P450 89A2	*Arabidopsis thaliana*	5.00E-26	FXAT9O003DBXJN	Q42602
Cytochrome P450 750A1	*Pinus taeda*	1.00E-07	FXAT9O003DCERL	Q50EK4
Cytochrome P450 720B2	*Pinus taeda*	7.00E-48	FXAT9O003DDAQ8	Q50EK5
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	2.00E-50	FXAT9O003DDLOI	P37116
Cytochrome P450 71A9	*Glycine max*	1.00E-25	FXAT9O003DG9QA	O81970
Cytochrome P450 82A1	*Pisum sativum*	3.00E-07	FXAT9O003DGDX0	Q43068
Cytochrome P450 750A1	*Pinus taeda*	5.00E-27	FXAT9O003DGJW8	Q50EK4
Cytochrome P450 750A1	*Pinus taeda*	2.00E-24	FXAT9O003DGKMA	Q50EK4
Cytochrome P450 87A3	*Oryza sativa subsp. japonica*	9.00E-29	FXAT9O003DH9OW	Q7XU38
Cytochrome P450 85A1	*Arabidopsis thaliana*	2.00E-19	FXAT9O003DHHLV	Q9FMA5
Cytochrome P450 750A1	*Pinus taeda*	6.00E-25	FXAT9O003DIXND	Q50EK4
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	3.00E-33	FXAT9O003DJ4XN	P37116

NADPH--cytochrome P450 reductase	*Phaseolus aureus*	3.00E-45	FXAT9O003DJFCD	P37116
Cytochrome P450 90A1	*Arabidopsis thaliana*	6.00E-13	FXAT9O003DJP82	Q42569
Cytochrome P450 716B1	*Picea sitchensis*	1.00E-16	FXAT9O003DJVN1	Q50EK1
Cytochrome P450 750A1	*Pinus taeda*	2.00E-26	FXAT9O003DLAAL	Q50EK4
Cytochrome P450 86A2	*Arabidopsis thaliana*	1.00E-20	FXAT9O003DMH6G	O23066
Cytochrome P450 98A2	*Glycine max*	3.00E-29	FXAT9O003DNJ7B	O48922
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	7.00E-32	FXAT9O003DNQ5M	P37116
Cytochrome P450 71A2	*Solanum melongena*	1.00E-24	FXAT9O003DNR6P	P37118
Cytochrome P450 720B2	*Pinus taeda*	2.00E-33	FXAT9O003DPH4W	Q50EK5
Cytochrome P450 716B2	*Picea sitchensis*	7.00E-25	FXAT9O003DPQ9P	Q50EK0
Cytochrome P450	*Populus trichocarpa*	3.00E-15	FXAT9O003C4X4P	ref|XP_002303546.1
Cytochrome P450 CYP735A2	*Arabidopsis thaliana*	2.00E-36	FXAT9O003C594U	ref|NP_176882.1
Cytochrome P450 CYP735A2	*Arabidopsis thaliana*	1.00E-15	FXAT9O003DJN4L	ref|NP_176882.1
cytochrome P450 probable 14-α-demethylase	*Populus trichocarpa*	2.00E-18	FXAT9O003DLYNB	ref|XP_002303744.1
cytochrome P450	*Panax ginseng*	1.00E-46	contig01909	dbj|BAD15331.1
cytochrome P450	*Ricinus communis*	1.00E-24	contig03727	gb|EEF49098.1
cytochrome P450 obtusifoliol 14-α-demethylase	*Populus trichocarpa*	1.00E-82	contig05061	ref|XP_002299356.1
cytochrome P450	*Populus trichocarpa*	1.00E-06	contig06619	ref|XP_002322452.1
nadph-cytochrome P450 oxydoreductase	*Populus trichocarpa*	2.00E-30	contig07608	ref|XP_002305157.1
Cytochrome P450 85A	*Phaseolus vulgaris*	6.00E-22	contig00009	Q69F95
Cytochrome P450 85A1	*Solanum lycopersicum*	2.00E-45	contig00012	Q43147
Cytochrome P450 87A3	*Oryza sativa subsp. japonica*	5.00E-71	contig00297	Q7XU38
Cytochrome P450 87A3	*Oryza sativa subsp. japonica*	2.00E-15	contig00306	Q7XU38
Cytochrome P450 87A3	*Oryza sativa subsp. japonica*	1.00E-09	contig00307	Q7XU38
Cytochrome P450 97B3	*Arabidopsis thaliana*	7.00E-22	contig00365	O23365
Cytochrome P450 90B1	*Arabidopsis thaliana*	2.00E-55	contig00506	O64989
Cytochrome P450 90A1	*Arabidopsis thaliana*	5.00E-98	contig00527	Q42569
Cytochrome P450 716B1	*Picea sitchensis*	5.00E-73	contig00626	Q50EK1
Cytochrome P450 716B2	*Picea sitchensis*	1.00E-143	contig00673	Q50EK0
Cytochrome P450 4X1	*Homo sapiens*	3.00E-11	contig00698	Q8N118
Cytochrome P450 76C2	*Arabidopsis thaliana*	4.00E-63	contig01487	O64637
Cytochrome P450 720B2	*Pinus taeda*	1.00E-144	contig01537	Q50EK5
Cytochrome P450 716B1	*Picea sitchensis*	9.00E-43	contig01839	Q50EK1
Cytochrome P450 750A1	*Pinus taeda*	6.00E-41	contig01910	Q50EK4
Cytochrome P450 750A1	*Pinus taeda*	1.00E-38	contig01911	Q50EK4
Cytochrome P450 704C1	*Pinus taeda*	1.00E-124	contig01982	Q50EK3
Cytochrome P450 720B2	*Pinus taeda*	1.00E-127	contig02015	Q50EK5
Cytochrome P450 716B2	*Picea sitchensis*	1.00E-148	contig02144	Q50EK0
Cytochrome P450 94A1	*Vicia sativa*	6.00E-59	contig02510	O81117
Cytochrome P450 98A1	*Sorghum bicolor*	1.00E-47	contig02606	O48956
Cytochrome P450 71D11	*Lotus japonicus*	1.00E-14	contig02795	O22307
Cytochrome P450 90B1	*Arabidopsis thaliana*	1.00E-41	contig02797	O64989
Cytochrome P450 98A2	*Glycine max*	1.00E-159	contig03018	O48922
Cytochrome P450 90C1	*Arabidopsis thaliana*	1.00E-104	contig03159	Q9M066
Cytochrome P450 71A1	*Persea americana*	1.00E-54	contig03417	P24465
Cytochrome P450 94A1	*Vicia sativa*	2.00E-24	contig03531	O81117
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	1.00E-57	contig03666	P37116
NADPH--cytochrome P450 reductase	*Phaseolus aureus*	3.00E-50	contig04519	P37116
Cytochrome P450 76A1	*Solanum melongena*	5.00E-14	contig05066	P37121
Cytochrome P450 97B2	*Glycine max*	1.00E-07	contig06776	O48921
Cytochrome P450 89A2	*Arabidopsis thaliana*	5.00E-09	contig06815	Q42602
Cytochrome P450 720B2	*Pinus taeda*	3.00E-21	contig07187	Q50EK5
Anthranilate N-benzoyltransferase protein 1	*Dianthus caryophyllus*	3.00E-11	FXAT9O003CXPKG	O24645
Anthranilate N-benzoyltransferase protein 1	*Dianthus caryophyllus*	4.00E-07	FXAT9O003DFMBT	O24645

Anthranilate N-benzoyltransferase protein 1	*Dianthus caryophyllus*	2.00E-15	FXAT9O003DLKSA	O24645
Succinyl-CoA ligase subunit beta	*Oryza sativa*	1.00E-71	FXAT9O003C63HZ	Q6K9N6
4-coumarate--CoA ligase-like 7	*Arabidopsis thaliana*	9.00E-16	FXAT9O003DCT8H	Q9M0X9
Long-chain-fatty-acid--CoA ligase	*Xenopus laevis*	2.00E-07	FXAT9O003DELJD	Q7ZYC4
4-coumarate--CoA ligase-like 4	*Arabidopsis thaliana*	1.00E-11	FXAT9O003DHE09	P0C5B6
4-coumarate--CoA ligase-like 3	*Arabidopsis thaliana*	6.00E-16	FXAT9O003DLVCG	Q3E6Y4
Succinyl-CoA ligase subunit beta	*Oryza sativa*	6.00E-42	FXAT9O003DNNNR	Q6K9N6
4-coumarate--CoA ligase-like 7	*Arabidopsis thaliana*	1.00E-166	contig00517	Q9M0X9
Long-chain-fatty-acid--CoA ligase 4	*Rattus norvegicus*	5.00E-55	contig01108	O35547
Probable succinyl-CoA ligase subunit alpha	*Oryza sativa*	1.00E-55	contig01369	Q6ZL94
2-succinylbenzoate--CoA ligase	*Bacillus subtilis*	1.00E-23	contig01468	P23971
Succinyl-CoA ligase subunit beta	*Oryza sativa*	0	contig02932	Q6K9N6
4-coumarate--CoA ligase-like 7	*Arabidopsis thaliana*	3.00E-26	contig03005	Q9M0X9
Long-chain-fatty-acid--CoA ligase 1	*Cavia porcellu*	6.00E-29	contig03398	Q9JID6
4-coumarate--CoA ligase-like 5	*Arabidopsis thaliana*	3.00E-34	contig04038	Q84P21
Long-chain-fatty-acid--CoA ligase 4	*Homo sapiens*	1.00E-42	contig04730	O60488
Long-chain-fatty-acid--CoA ligase 4	*Homo sapiens*	6.00E-37	contig05866	O60488
Long-chain-fatty-acid--CoA ligase 5	*Mus musculus*	2.00E-39	contig06089	Q8JZR0
Long-chain-fatty-acid--CoA ligase 5	*Homo sapiens*	3.00E-13	contig06961	Q9ULC5

Table 4S Unique sequences with putative functions similar to transcription factor family members

Putative homologs of TF family	**Unique sequence**	**E-value**	**Reads**
Zinc finger CCCH domain-containing protein	contig02422, contig01696, contig01074, contig00487, contig04628, contig02483, contig03455, contig07393, contig03973, contig05603, contig05637, contig06018, contig05201, contig06574, FXAT9O003C0I9W, FXAT9O003C4Y72, FXAT9O003C558W, FXAT9O003C7Y8O, FXAT9O003CWL3Y, FXAT9O003CXUKY, FXAT9O003CYSAF, FXAT9O003DA3PA, FXAT9O003DD79B, FXAT9O003DDNKD, FXAT9O003DDSM6, FXAT9O003DDZIT, FXAT9O003DII56, FXAT9O003DOM2F	4.00E-98, 7.00E-51, 6.00E-50, 1.00E-73, 4.00E-43, 4.00E-08, 1.00E-83, 5.00E-16, 2.00E-54, 1.00E-20, 1.00E-24, 5.00E-14, 3.00E-20, 6.00E-18, 1.00E-19, 2.00E-14, 3.00E-19, 2.00E-17, 2.00E-14, 2.00E-17, 2.00E-27, 4.00E-16, 3.00E-45, 2.00E-24, 2.00E-07, 2.00E-26, 4.00E-14, 9.00E-06	45, 9, 6, 5, 5, 4, 4, 4, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
Zinc finger CCHC domain-containing protein	contig02563, FXAT9O003C5IGL, FXAT9O003C9VZP, FXAT9O003CZ9FK	3.00E-20, 6.00E-09, 1.00E-15, 4.00E-57	6, 1, 1, 1
Zinc finger protein CONSTANS-LIKE	contig01284, contig01278, contig02915, contig06382, FXAT9O003CXDY0, FXAT9O003DBHN3, FXAT9O003DD8ZK, FXAT9O003DFG6T	2.00E-32, 3.00E-25, 5.00E-08, 1.00E-11, 9.00E-06, 7.00E-17, 2.00E-11, 2.00E-06	46, 25, 17, 4, 1, 1, 1, 1
Dof zinc finger protein	contig00628, contig05584, FXAT9O003DE5UE, FXAT9O003DERSJ, FXAT9O003DFJV4	4.00E-15, 9.00E-27, 4.00E-24, 2.00E-09, 4.00E-18	8, 3, 1, 1, 1
Zinc finger A20 and AN1 domain-containing stress-associated protein	contig00640, contig01102, contig05931, FXAT9O003DFYFG, FXAT9O003C3KYT	2.00E-13, 4.00E-26, 3.00E-27, 7.00E-14, 1.00E-44	7, 8, 4, 1, 1

Zinc finger protein MAGPIE	contig02458, FXAT9O003C3LWB, FXAT9O003C70KH, FXAT9O003DQ2IN	1.00E-55, 7.00E-29, 6.00E-31, 2.00E-63	2, 1, 1, 1
RING finger and CHY zinc finger domain-containing protein	contig01589, contig05569, contig05098, contig01588	1.00E-29, 6.00E-58, 2.00E-12, 9.00E-06	10, 7, 4, 2
Zinc finger protein	contig00093, contig02314, contig01550, contig06570, contig02804, contig02926, contig01647, contig01655, contig02383, contig02567, contig03841, contig04373, contig02481, contig00113, contig05560, contig07347, FXAT9O003DHXDM, FXAT9O003DLCIV, FXAT9O003C1Q0V, FXAT9O003DIVK7, FXAT9O003DNSSF, FXAT9O003DKOGZ	6.00E-13, 3.00E-82, 6.00E-28, 5.00E-93, 3.00E-17, 2.00E-12, 2.00E-13, 5.00E-37, 2.00E-17, 5.00E-11, 8.00E-13, 5.00E-13, 2.00E-09, 1.00E-09, 2.00E-07, 1.00E-12, 4.00E-10, 2.00E-19, 4.00E-17, 2.00E-06, 3.00E-17, 4.00E-14	75, 19, 17, 13, 11, 10, 9, 7, 4, 4, 4, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1
WD repeat-containing protein	contig01048, contig02139, contig02228, contig02143, contig01764, contig04626, contig05273, contig01638, contig04503, contig04602, contig05065, contig05088, contig06340, contig05432, contig05485, contig06688, contig05537, FXAT9O003C1AI6, FXAT9O003C1FD5, FXAT9O003C2OKZ, FXAT9O003C57IM, FXAT9O003C642U, FXAT9O003C7L6Z, FXAT9O003C7UT4, FXAT9O003C9881, FXAT9O003CWRRL, FXAT9O003CXPWQ, FXAT9O003DBCG4, FXAT9O003DDWK8, FXAT9O003DEICU, FXAT9O003DES1I, FXAT9O003DGERX, FXAT9O003DI596, FXAT9O003DJQ12, FXAT9O003DKBOL, FXAT9O003DKQ3X, FXAT9O003DLTKG, FXAT9O003DODOD, FXAT9O003DPD06, FXAT9O003DPP8J, FXAT9O003DQQKF, FXAT9O003DQVWU	3.00E-27, 1.00E-54, 1.00E-22, 4.00E-14, 9.00E-21, 7.00E-06, 7.00E-21, 2.00E-41, 4.00E-11, 5.00E-23, 6.00E-29, 7.00E-29, 2.00E-06, 4.00E-14, 2.00E-14, 3.00E-26, 3.00E-13, 2.00E-25, 4.00E-16, 6.00E-15, 4.00E-38, 1.00E-19, 1.00E-17, 6.00E-14, 3.00E-06, 1.00E-15, 2.00E-09, 5.00E-12, 1.00E-20, 4.00E-24, 7.00E-13, 2.00E-13, 3.00E-11, 4.00E-45, 2.00E-12, 5.00E-12, 2.00E-24, 2.00E-07, 2.00E-14, 2.00E-10, 1.00E-28, 4.00E-69	13, 7, 6, 5, 4, 4, 4, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
WD40 repeat-containing protein	contig05583, contig03308, contig03308, FXAT9O003C30WY, FXAT9O003DN6AM, FXAT9O003DNRWJ, FXAT9O003CXAO6, FXAT9O003DLPQV, FXAT9O003DMUFC, FXAT9O003DN6AM, FXAT9O003DNRWJ	2.00E-19, 9.00E-58, 9.00E-58, 3.00E-36, 1.00E-32, 3.00E-12, 4.00E-39, 5.00E-43, 5.00E-40, 1.00E-32, 3.00E-12	3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1
Homeobox-leucine zipper protein, Homeobox protein	contig01125, contig02249, contig03510, contig02175, contig05404, contig02451, contig03463, contig04750, contig03952, contig06769, FXAT9O003C05IM, FXAT9O003C0N4K, FXAT9O003C4DHL, FXAT9O003C70KM, FXAT9O003CWYLB,	3.00E-40, 4.00E-34, 4.00E-17, 1.00E-36, 9.00E-35, 6.00E-06, 3.00E-09, 9.00E-53, 3.00E-06, 1.00E-08, 2.00E-29, 7.00E-25, 4.00E-23, 1.00E-62, 2.00E-60, 6.00E-31, 1.00E-22, 1.00E-11,	30, 12, 9, 8, 5, 3, 3, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

	FXAT9O003DAUC2, FXAT9O003DBQI7, FXAT9O003DC6Y3, FXAT9O003DEEA2, FXAT9O003DJCEP, FXAT9O003DL1SI, FXAT9O003DOOCF	1.00E-08, 1.00E-30, 3.00E-21, 2.00E-06	
Probable WRKY transcription factor	contig00605, contig05955, contig03492, contig06591, FXAT9O003C0LAL, FXAT9O003C3HRB, FXAT9O003C86JV, FXAT9O003C8XP5, FXAT9O003CWY38, FXAT9O003CX3ET, FXAT9O003CZ7WS, FXAT9O003DA1EN, FXAT9O003DCVV0, FXAT9O003DDH47, FXAT9O003DDM4Y, FXAT9O003DFEI5, FXAT9O003DIUIC, FXAT9O003DLUTK	4.00E-20, 4.00E-48, 9.00E-06, 2.00E-06, 4.00E-35, 4.00E-42, 1.00E-23, 1.00E-14, 6.00E-32, 1.00E-10, 7.00E-24, 4.00E-17, 8.00E-52, 7.00E-21, 3.00E-23, 5.00E-19, 9.00E-13, 3.00E-44	14, 5, 4, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
Myb family transcription factor	contig01613, contig01356, contig00310, contig06625, contig03848, FXAT9O003C39S1, FXAT9O003C3VFF, FXAT9O003C7TX6, FXAT9O003DC4SW, FXAT9O003DDNND, FXAT9O003DET0E, FXAT9O003DKMI7, FXAT9O003C69DQ, FXAT9O003DMT8W	2.00E-17, 4.00E-07, 7.00E-07, 1.00E-10, 6.00E-19, 7.00E-08, 5.00E-31, 8.00E-10, 2.00E-32, 2.00E-33, 1.00E-39, 4.00E-15, 7.00E-21, 1.00E-20	34, 7, 6, 5, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
basic helix-loop-helix protein	contig04648, contig06082, FXAT9O003CWLKF, FXAT9O003CXO9O, FXAT9O003DIBV7, FXAT9O003C18TI, FXAT9O003C5884, FXAT9O003C9UZN, FXAT9O003CYDEF, FXAT9O003DCBGI, FXAT9O003DLTDB, FXAT9O003DOM7V, FXAT9O003DIC1R	1.00E-19, 7.00E-13, 3.00E-10, 2.00E-06, 3.00E-26, 7.00E-09, 5.00E-35, 1.00E-20, 8.00E-31, 6.00E-21, 2.00E-36, 3.00E-17, 2.00E-51	4, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
Auxin response factor	contig00560, contig05659, FXAT9O003C1VW5, FXAT9O003C618E, FXAT9O003C2YVS, FXAT9O003CWZOS, FXAT9O003DGYSK, FXAT9O003DIHSI, FXAT9O003DMU7V, FXAT9O003DOYWB	3.00E-38, 1.00E-34, 4.00E-26, 6.00E-66, 3.00E-68, 1.00E-46, 4.00E-19, 7.00E-07, 4.00E-08, 1.00E-34	7, 2, 1, 1, 1, 1, 1, 1, 1, 1
NAC domain-containing protein	contig02417, contig02617, contig05341, contig02894, FXAT9O003C2N5J, FXAT9O003C6NVQ, FXAT9O003C7G68, FXAT9O003CY0C3, FXAT9O003DKCE4, FXAT9O003DP3A9	3.00E-59, 4.00E-25, 3.00E-18, 1.00E-39, 3.00E-24, 1.00E-34, 2.00E-16, 1.00E-56, 1.00E-08, 9.00E-30	6, 3, 3, 2, 1, 1, 1, 1, 1, 1
Ethylene-responsive transcription factor	contig01626, contig02694, contig03412, contig00507, contig03043, FXAT9O003C0JI1, FXAT9O003C6TFV, FXAT9O003DC1EF, FXAT9O003DL90V	1.00E-20, 8.00E-30, 5.00E-14, 2.00E-14, 1.00E-21, 1.00E-12, 2.00E-17, 1.00E-09, 1.00E-14	14, 9, 6, 5, 2, 1, 1, 1, 1
Nuclear transcription factor	contig00186, contig04096, contig03092, FXAT9O003C7ATZ, FXAT9O003CX06D, FXAT9O003DAIWT, FXAT9O003DG8E7, FXAT9O003DKJ0Q, FXAT9O003DQCTJ	1.00E-50, 1.00E-51, 5.00E-54, 1.00E-17, 1.00E-32, 2.00E-24, 7.00E-11, 3.00E-08, 2.00E-32	24, 18, 3, 1, 1, 1, 1, 1
homeodomain protein	contig03746, contig01380, contig02973,	3.00E-19, 2.00E-30, 3.00E-33,	7, 4, 3, 2, 1,

	contig06668, FXAT9O003DAXY0, FXAT9O003DBFSK	2.00E-06, 9.00E-25, 7.00E-22	1
Heat stress transcription factor	contig04922, contig06654, FXAT9O003C8KLL, FXAT9O003CZXY9, FXAT9O003DFQJ1	1.00E-27, 1.00E-07, 2.00E-06, 1.00E-21, 4.00E-33	5, 2, 1, 1, 1
MADS-box transcription factor	contig02579, contig05856, FXAT9O003C4ZZB, FXAT9O003DGAYV	2.00E-30, 6.00E-35, 1.00E-08, 3.00E-20	7, 3, 1, 1
GATA transcription factor	contig03576, contig07003, contig06329, FXAT9O003C6GOZ	4.00E-50, 6.00E-14, 4.00E-11, 1.00E-29	4, 3, 2, 1
general transcription factor IIF, General transcription factor IIH	FXAT9O003C82C8, FXAT9O003DJFJT, contig02502, FXAT9O003DCGWS	5.00E-12, 1.00E-17, 2.00E-29, 1.00E-07	1, 1, 3, 1
Probable transcription factor PosF21	FXAT9O003C5GAV, FXAT9O003DD9GO, FXAT9O003DGTNT	3.00E-38, 5.00E-20, 2.00E-32	1, 1, 1
Transcription factor FER-LIKE IRON DEFICIENCY-INDUCED	contig00263, contig00268, contig03534	5.00E-15, 4.00E-17, 4.00E-18	21, 9, 57
Transcription factor ILR3	contig01366, contig05825	4.00E-35, 3.00E-12	7, 5
Transcription factor BTF3	contig03420, contig01036	5.00E-29, 2.00E-22	11, 9
Transcription factor HBP-1b	contig05891, contig03773	2.00E-66, 3.00E-29	3, 4
Scarecrow-like transcription factor PAT1	FXAT9O003C4RD9	1.00E-57	1
TFIIH basal transcription factor complex p47 subunit	FXAT9O003DGFK6	2.00E-22	1
Transcription factor 25	FXAT9O003C4X33	2.00E-23	1
Transcription factor CPC	FXAT9O003CWVLQ	5.00E-10	1
Transcription factor Dp-1	FXAT9O003C3AXW	3.00E-11	1
Transcription factor E2F5	contig05666	2.00E-16	5
Transcription factor IIIA	contig02191	8.00E-08	3
Transcription factor IWS1	FXAT9O003DF84I	7.00E-13	1
Transcription factor MYC2	contig02043	9.00E-24	2
Transcription factor PIF3	FXAT9O003DDLFI	4.00E-12	1
Transcription factor SCREAM2	FXAT9O003C49BE	3.00E-21	1
Transcription factor TCP13	FXAT9O003C27GN	5.00E-22	1
Transcription factor TFIIIB component B	contig03901	1.00E-17	7
Transcription factor TT2	FXAT9O003C16H	3.00E-33	1
Transcription factor TT8	FXAT9O003CYDAT	8.00E-17	1
Transcription factor TT8	FXAT9O003DI2ZW	4.00E-10	1
Transcription factor UNE10	contig02549	1.00E-25	20
Basic leucine zipper and W2 domain-containing protein 2	contig01325	3.00E-57	102
Fork head transcription factor 1	contig06873	7.00E-06	3

Table 5S Summary of SSR motifs in unique sequences of *T. cuspidata*. The number of repeated units was classed into three groups (units repeated at < 10 times, 11–20 times, and 21–50 times)

	Number of SSR repeat units				
SSR motif	**< 11 repeat units**	**11-20 repeat units**	**21-50 repeat units**	**Total**	**Percent (%)**
TA/TA	18	8	6	32	4.25
GAA/TTC	32			32	4.25
AGA/TCT	26		1	27	3.59
TAAAA/TTTTA	26			26	3.45
AG/CT	16	5		21	2.79
AAG/CTT	18	1		19	2.52
GCA/TGC	19			19	2.52
CAG/CTG	17	1		18	2.39
GA/TC	12	4		16	2.12
TAAA/TTTA	16			16	2.12
AC/GT	11	2	2	15	1.99
CTC/GAG	14			14	1.86
ATG/CAT	13			13	1.73
TAA/TTA	12	1		13	1.73
AGC/GCT	12			12	1.59
AGG/CCT	11			11	1.46
GGA/TCC	11			11	1.46
ATC/GAT	10			10	1.33
AAT/ATT	9			9	1.20
AT/AT	5	2	1	8	1.06
CA/TG	5	3		8	1.06
ATA/TAT	5		1	6	0.80
GAGGAA/TTCCTC	6			6	0.80
ACC/GGT	5			5	0.66
CTCTTC/GAAGAG	5			5	0.66
GGAAGA/TCTTCC	5			5	0.66
Other type	375	1		376	49.93
Total	714	28	11	753	

Table 6S Primers used in simple sequence repeat sequence validation through Sanger methods

Primer	**Oligo sequence**
SSR1f	AACTCTGCTGCGTTTGCGAC

SSR1r	CCGTTTTTGGGGGGATTTTC
SSR2f	ATGGGTAAGACTTTTTGATTGT
SSR2r	TTTAGTAGTATCCCTCCAAGAC
SSR3f	GCAGAGATGGTAAATCCCGACA
SSR3r	TGGAGCGAGAGAAAAAACAAGA
SSR4f	CGTTCTATTGCTTCATTATCTA
SSR4r	AGTGTATTCCTCCATTTGTGTA
SSR5f	CTATCGCCCTCTTCAAATGCTG
SSR5r	CCAGATACACCACTCCCCTTCC
SSR6f	CAACTGCTCCTCCAACAAAC
SSR6r	TCCTCCGTCTATCATCTCTCC
SSR7f	AGGTTTGGTCTGAGGGTGAA
SSR7r	AGGCTGGTTAGGATGTCTTTG
SSR8f	TGCGGAGATTCGTTGTTGAC
SSR8r	AATGGTGGCGACTGTGAGAG
SSR9f	CCAAAGCAGATTCCAGATTACAGG
SSR9r	GTCCACAAGACTACCACAACAGAT
SSR10f	CTACGGAATGCTCAGACACG
SSR10r	TCTTCACCCCAGAACTCAAA

References

1 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215: 403-410

2 Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 2000; 29: 159-164

3 Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G, Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. Plant Physiol 2004; 135: 745-755

4 Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006; 34: D354-D357