Genetics and Molecular Research

Comparative sequence and expression analysis of tapetum specific male sterility related genes in *Medicago truncatula*

L.H. Shao, X.W. Zheng, D.X. Yi and C. Li

Institute of Animal Sciences, the Chinese Academy of Agricultural Sciences, Beijing, China

Corresponding author: C. Li
E-mail: licong0520@sina.com

Genet. Mol. Res. 15 (2): gmr.15028323
Received December 21, 2015
Accepted March 11, 2016
Published July 15, 2016
DOI http://dx.doi.org/10.4238/gmr.15028323

ABSTRACT. Heterosis, or enhancement through outbreeding, is one of the most promising approaches for increasing crop yield. Male sterility (MS), which promotes heterosis, has been widely applied in hybrid crop production. *Medicago truncatula* is a model legume species and is closely related to *M. sativa*, an important legume forage plant. Although the molecular mechanisms of MS in *M. truncatula* and *M. sativa* remain unclear, several studies of MS have been conducted in *Arabidopsis thaliana*. Previous research has shown that MS is associated with the destruction of tapetal cell layers. Disruption of tapetum developmental processes may result in pollen abortion. In an effort to identify genes useful for breeding in *M. sativa*, we identified MS related genes in *M. truncatula* using BLAST and homology to *A. thaliana* genes. In this study, we identified 63 tapetum specific male sterility (TSMS) related genes. The length of TSMS genes varied from 225 to 3747 bp. We identified 15 conserved domains and 8 *cis*-elements associated with TSMS related genes. Analysis of the phylogenetic relationships among these genes allowed them to be classified into three groups, MtTsms A, MtTsms B, and MtTsms C. Expression analyses revealed that these

genes may be involved in developmental processes and response to abiotic stress.

Key words: *Medicago truncatula*; Male sterility; Genetic analysis; Tapetum

INTRODUCTION

Male sterility (MS) occurs in plants when the female organ is normal but the male organ is abnormal and is unable to produce pollen or successfully pollinate. This is a common phenomenon in the plant kingdom and is important for both plant reproductive biology and crop-breeding research because it promotes heterosis or outbreeding (Wu et al., 2015). Kaul (1988) reported that MS had been observed in 43 families, 162 genera, 320 species, and 617 varieties or hybrids. According to the stage at which abortion occurs during anther development, MS can be classified into several types, including degradation of pollen mother cells, degeneration of tapetum cells, critical chemical changes in pollen wall development, and failure of anther dehiscence (Liu et al., 2014). For the promotion of heterosis, it is critical to obtain MS lines that interrupt the development of the anther and/or pollen (Zhang et al., 2006).

Plant anther tapetal cells comprise the interior layer of cells adjacent to the pollen mother cell in the anther clinandrium. The tapetum is of great importance for pollen development, and any process that blocks tapetal development will lead to MS. The tapetum has been studied in many plants using MS strains (Ma et al., 2015). Tapetal cells are essential for the formation of functional pollen grains; therefore, many MS mutations have been acquired through interference with tapetal cell differentiation and/or function degeneration (Li et al., 2015). Many studies have identified tapetum specific (TS) genes by screening MS mutants in *Arabidopsis thaliana* (Honys and Twell, 2003; Hsu et al., 2014) and transgenic MS plants have been produced through the degradation of tapetal cells.

Model plants are useful for translating genomic information between species. This cross-species translation is largely based on orthologous genes, which share a common evolutionary origin (Hyung et al., 2014). *A. thaliana* was the first plant species to have its genome fully sequenced, and among all model plants, the most comprehensive genomic studies have been performed on this species. *Medicago truncatula* is a model legume with a relatively small genome and short life cycle. It therefore serves as a central model system for translational genomics based studies in plants (Town et al., 2006). Although many studies have specifically studied MS in *A. thaliana*, *Brassica oleracea*, and *Oryza sativa* (Honys and Twell, 2003; Qu et al., 2008; Ji et al., 2011; Zhang and Zhang, 2014), few studies have assessed MS in forage plants. In this study, we attempted to identify tapetum specific male sterility (TSMS) related genes in *M. truncatula*, via comparisons with *A. thaliana*, as a source of genomic information. Information obtained from *M. truncatula* may provide critical data for the breeding of new germplasms in *M. sativa*.

MATERIAL AND METHODS

Relational database

M. truncatula genome sequence data were downloaded from JCVI (Mt4.0v1, http://

www.jcvi.org/medicago/). Expression data for *M. truncatula* TSMS related genes were downloaded from the *M. truncatula* Gene Expression Atlas (http://mtgea.noble.org/v3/). Gene sequence alignments and promoter analyses were conducted using NCBI (http://www.ncbi.nlm.nih.gov/) and the PLACE (http://www.dna.affrc.go.jp/PLACE/) software.

Genome-wide screening of TSMS related genes and identification of cross-family orthologous genes

Available data sources were used to identify TSMS related genes, including published literature, gene expression profiles, and the *Arabidopsis* database (TAIR, http://www.arabidopsis.org). We used key words relevant to TSMS to identify related genes in the *Arabidopsis* database, and each resulting gene description was investigated to confirm its functional relevance to TSMS. Genes identified from databases were combined into a single Excel file and organized according to locus ID along with other related information, such as location, cDNA length, exon number, intron number, coding sequence (CDS) length, and annotation (Table 1).

To identify orthologous genes in *M. truncatula* from *A. thaliana*, we used NCBI's BLASTp homolog search tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). We considered the E-value ≤ 1e-10, identity ≥30%, coverage ≥70%, and the cumulative total length of alignment when identifying orthology between pairwise aligned genes through visual inspection (Hsu et al., 2014).

Gene structure and chromosomal location

Protein domain structure for domains with the lowest E-values was analyzed by the SMART online tool (http://smart.embl-heidelberg.de/). Information regarding cDNA sequences and chromosomal location of genes was retrieved from NCBI (http://www.ncbi.nlm.nih.gov/), and the location of each TSMS gene was then marked by the Mapinspect software (http://mapinspect.apponic.com/) (Ma et al., 2013).

Phylogenetic analysis

For phylogenetic analysis, we used the Clustal W2 software for multiple alignment of TSMS protein sequences from *M. truncatula* and *A. thaliana*. MEGA5.0 was then used to create phylogenetic trees using the neighbor-joining algorithm and a Poisson-corrected model with gaps/missing data subject to pairwise deletion and 1000 bootstrap test replications (Honys and Twell, 2003).

Cis-element analysis

The PLACE database (http://www.dna.affrc.go.jp/PLACE/) was used to search for *cis*-elements in the promoter regions of homologous TSMS related genes in *M. truncatula*. We selected a 2000-bp promoter region upstream of the start codon of each gene to search for possible *cis*-elements (Manimaran et al., 2015).

Gene expression analysis

Gene expression data were downloaded from the *M. truncatula* Gene Expression Atlas

of the Samuel Roberts Noble Foundation. Transcriptome data included the analysis of root, stem, leaf, vegetative bud, flower, pod, and seed-coat tissue under various stress conditions, including NH_4, NO_3, N_2, 1-Naphthaleneacetic acid (NAA), and *Phymatotrichum* infection. Data on TSMS gene expression were downloaded using the MeV (Ver.4.6.1) software (http://www.tm4.org/mev/), and were formatted as log2 fold-change and graphically presented in a heat-map using the Cimminer software (http://discover.nci.nih.gov/cimminer/oneMatrix.do) (Ma et al., 2013).

RESULTS

Identification of TSMS related genes

To identify the full complement of TSMS related genes in *M. truncatula*, known TSMS genes from *A. thaliana* and the BLASTp algorithm were used to search the *M. truncatula* genome for homologs. A total of 63 putative *M. truncatula* TSMS related genes were identified (Table 1). Gene structure analyses revealed that these genes were characterized by highly variable exon and intron structures: 4 of the putative TSMS related genes contained no introns, 27 genes contained one intron, 7 genes contained two introns, 4 genes contained three introns, 12 genes contained four introns, 3 genes contained five introns, and 6 genes contained six or more introns. As intron patterns can shed light on phylogenetic relationships (Liu et al., 2014), these patterns indicate a significant degree of divergence among the putative TSMS genes (Ma et al., 2015). These 63 putative TSMS genes are located on eight *M. truncatula* chromosomes, ranging from three genes on chromosome 3 to 17 genes on chromosome 4 (Figure 1). The longest cDNA nucleotide sequence was 11.6 kb and included 20 exons, while the shortest cDNA sequence was 384 bp, comprising only one exon (Table 1).

Structure of tapetum specific male sterility related proteins in *M. truncatula*

Using the online program SMART, we accessed information on the functional domains of *M. truncatula* TSMS genes. Fifteen conserved domains (AAA, AAI, BBOX, HLH, HSP, Knot1, Pept_C1, Pfam, REC, RING, S_TKc, SANT, SERPIN, transmembrane region, and UBCc) were identified by sequence alignment and protein position, and were renamed MtTsms1-63, clusters of orthologous groups A, B, and C, representing the three groups of MtTsms genes identified by the phylogenetic tree (Table 2). Among these, the conserved motif encoding the TSMS domain was found within all 63 proteins. Six genes (MtTsms1-6) contained the AAA domain, five genes (MtTsms7-11) contained the AAI domain, 37 genes contained Pfam domains, three genes (MtTsms57-59) contained SANT domains, and two genes (MtTsms62-63) contained UBCc domains, with the final 10 genes containing additional domains.

The presence of the AAA domain defines the AAA family of proteins, which often perform chaperone-like functions to assist in the assembly, operation, or disassembly of proteins; these proteins exhibit high ATPase activity and oligomerization (Ye et al., 2015). The AAI domain is present in several crucial proteins involved in pollen development (Tian et al., 2009). Although other identified conserved domains such as Pfam and the transmembrane region have not previously been associated with MS, our results suggest that the functions of these genes require further exploration.

Table 1. Tapetum specific male sterility related genes in *Medicago truncatula*.

Arabidopsis thaliana	*Medicago truncatula*	E-value	Location	cDNA length	Exon number	Intron number	CDS length	Annotation
AT1G01280	Medtr7g087520	0	chr7:34077176-34078819	1644	2	1	1566	cytochrome P450 family protein
AT1G03390	Medtr1g053315	0	chr1:22378573-22379970	1398	1	0	1398	HXXXD-type acyl-transferase family protein
AT1G27040	Medtr5g038380	0	chr5:16824210-16830540	6331	5	4	1749	peptide/nitrate transporter plant
	Medtr4g107670	0	chr4: 44609687-44613526	3840	5	4	1734	peptide/nitrate transporter
	Medtr8g469310	0	chr8:25169829-25178038	8210	5	4	1734	peptide/nitrate transporter plant
AT1G36340	Medtr3g062450	1.00E-48	chr3:28197947-28201491	3545	4	3	447	ubiquitin-conjugating enzyme E2
AT1G52560	Medtr4g010320.1	2.00E-57	chr4:2335470-2337224	1755	2	1	696	hsp20/alpha crystallin family protein
AT1G62940	Medtr2g105570	4.00E-139	chr2:45520733-45525707	4975	5	4	1656	4-coumarate:CoA ligase-like protein
AT1G74310	Medtr4g109450	0	chr4:45465442-45469308	3867	7	4	2739	ATP-dependent chaperone ClpB
AT1G06170	Medtr1g072320	5.00E-10	chr1:32096191-32103626	7436	6	5	1971	bHLH transcription factor
AT1G26780	Medtr4g082040	3.00E-74	chr4:31789875-31791702	1828	3	2	1284	myb transcription factor
AT1G61070	Medtr8g095375	3.00E-21	chr8:39904663-39905235	573	2	1	225	defensin MtDef4.7
AT1G06260	Medtr4g079440	1.00E-130	chr4:30712622-30713897	1276	2	1	1038	papain family cysteine protease
AT1G06990	Medtr5g074270	1.00E-115	chr5:31527639-31530478	2840	4	3	1101	GDSL-like lipase/acylhydrolase
AT1G09550	Medtr7g111280	2.00E-158	chr7:45666166-45669433	3268	11	10	1260	pectin acetylesterase family protein
AT1G61110	Medtr2g080010	1.00E-108	chr2:33761666-33764021	2356	3	2	1065	NAC transcription factor-like protein
AT1G66570	Medtr1g096910	0	chr1:43663044-43666502	3459	4	3	1578	sucrose proton symporter, putative
	Medtr4g131920	0	chr4:55076087-55079116	3030	5	4	1536	sucrose transporter 2
	Medtr6g043880	0	chr6:15122970-15125916	2974	5	4	1527	sucrose transporter
AT1G68190	Medtr3g082630	7.00E-29	chr3:37271508-37275432	3925	6	5	1236	B-box type zinc finger protein
AT1G71160	Medtr1g110930	8.00E-162	chr1:50086559-50088597	2039	1	1	1491	3-ketoacyl-CoA synthase-like protein
AT1G74540	Medtr8g020950	1.00E-95	chr8:7422331-7426441	4111	2	1	1545	cytochrome P450 family protein
AT1G75930	Medtr8g075230	8.00E-74	chr8:31794869-31797757	2889	5	4	1077	GDSL-like lipase/acylhydrolase
AT1G30860	Medtr1g090803	3.00E-22	chr1:40822523-40829027	6505	8	7	2556	RING/U-box protein
AT2G16910	Medtr7g083900	2.00E-28	chr7:32316153-32319262	3027	4	3	1428	basic helix loop helix protein BHLH23
AT2G14540	Medtr7g451890	2.00E-106	chr7:17697421-17699702	2282	2	1	1073	serpin-like protein
AT2G26150	Medtr3g104550	3.00E-119	chr3:48170509-48171887	1379	2	1	1137	heat shock transcription factor A3
AT2G31210	Medtr8g024790	6.00E-12	chr8:9212700-9214341	1642	1	1	1395	basic helix loop helix (bHLH) family transcription factor
AT2G38240	Medtr7g068650	2.00E-156	chr7:25175337-25179838	4502	3	2	1095	leucoanthocyanidin dioxygenase-like protein
AT2G42940	Medtr5g011520	3.00E-101	chr5:39959479-39960243	765	1	0	765	AT hook motif DNA-binding family protein
AT2G45630	Medtr8g033400	2.00E-115	chr8:12885180-12886877	1698	2	1	1002	D-isomer specific 2-hydroxyacid dehydrogenase domain protein
AT2G03850	Medtr2g008610	1.00E-06	chr2:1552950-1554469	1520	1	2	786	light-harvesting complex I chlorophyll A/B-binding protein
AT2G23800	Medtr8g078070	3.00E-144	chr8:33275574-33277440	1867	1	1	1119	geranylgeranyl pyrophosphate synthase
AT2G47040	Medtr7g110490	2.00E-129	chr7:45293122-45295123	2002	2	1	1731	pectinesterase inhibitor domain protein
AT3G05780	Medtr1g084020	0	chr1:37455374-37466754	11381	20	19	2934	Lon-related ATP-dependent protease, putative
AT3G09640	Medtr4g061140	1.00E-146	chr4:22580311-22583517	3615	14	9	753	L-ascorbate peroxidase
AT3G13220	Medtr1g054935	0	chr1:24118450-24130074	11625	11	10	2229	white-brown-complex ABC transporter family protein
AT3G48540	Medtr4g012000	1.00E-10	chr4:3190971-3199290	8320	6	5	3747	tRNA-specific adenosine deaminase
	Medtr4g103300.1	8.00E-122	chr4:42684135-42689049	4915	9	8	699	tRNA-specific adenosine deaminase
AT3G52130	Medtr7g065160	1.00E-10	chr7:23754468-23755721	1254	2	1	354	lipid transfer protein
AT3G13890	Medtr2g099740	4.00E-78	chr2:42787607-42789410	1804	2	1	1206	myb transcription factor
AT3G51590	Medtr4g027800	2.00E-28	chr4:9811604-9812808	1205	2	1	351	lipid transfer protein
AT3G56380	Medtr5g036480	4.00E-54	chr5:15918294-15920037	1744	5	4	714	two-component response regulator ARR3-like protein
AT4G14080	Medtr7g053500	0	chr7:18857464-18859936	2473	3	2	1392	glucan endo-1,3-beta-glucosidase-like protein
AT4G20800	Medtr4g091200	0	chr4:36107336-36109152	1817	1	1	1587	FAD-binding berberine family protein
	Medtr6g017205	0	chr6:6838164-6839789	1626	1	0	1626	FAD-binding berberine family protein
	Medtr4g088355	0	chr4:34970644-34972263	1620	1	0	1620	FAD-binding berberine family protein
AT4G34850	Medtr4g078730	0	chr4:30428531-30429812	1282	2	1	1170	chalcone synthase family protein, putative
AT5G02490	Medtr7g099680	0	chr7:39984070-39987331	3262	2	1	1950	heat shock cognate 70 kDa protein
	Medtr7g024580	0	chr7:8101201-81045340	3334	2	1	1950	heat shock cognate 70 kDa protein
	Medtr7g024390	0	chr7:8016048-8019129	3082	2	1	1947	heat shock cognate 70 kDa protein
AT5G09970	Medtr8g027040	5.00E-156	chr8:9421677-9424147	2471	2	1	1641	cytochrome P450 family 78 protein
AT5G59720	Medtr6g061940	6.00E-83	chr6:24040583-24041280	698	1	1	480	17.6 kDa class I heat shock protein
AT5G03800	Medtr7g103230	0	chr7:41735375-41738881	3507	3	2	2913	PPR containing plant-like protein
AT5G07230	Medtr1g101360	2.00E-18	chr1:45783546-45783929	384	2	1	294	lipid transfer protein (LIM1/2/3/MEN-8)
AT5G15250	Medtr7g010800	0	chr7:2723939-2728362	4424	5	4	2014	ATP-dependent zinc metalloprotease FTSH protein
	Medtr2g438140	0	chr2:15376886-15381581	4696	5	4	2073	ATP-dependent zinc metalloprotease FTSH protein
	Medtr4g094662	0	chr4:38677988-38681608	3621	5	4	2085	ATP-dependent zinc metalloprotease FTSH protein
AT5G56110	Medtr6g009430	7.00E-66	chr6:2790201-2792486	2286	3	2	1011	myb transcription factor
AT5G59330	Medtr4g027840	1.00E-10	chr4:9831935-9832433	499	2	1	420	lipid transfer protein
AT5G66110	Medtr4g073040	6.00E-72	chr4:27630362-27631603	1242	2	1	471	heavy metal transport/detoxification superfamily protein
AT5G14980	Medtr1g095110	9.00E-166	chr1:42811469-42812783	1315	1	1	978	alpha/beta hydrolase family protein
AT5G38160	Medtr6g082480	9.00E-19	chr6:30850320-30851288	969	1	1	279	chitinase/hevein/PR-4/wheatwin2
AT5G57670	Medtr4g128990	0	chr4:53672087-53675818	3732	10	9	1728	receptor-like kinase

CDS means coding sequence.

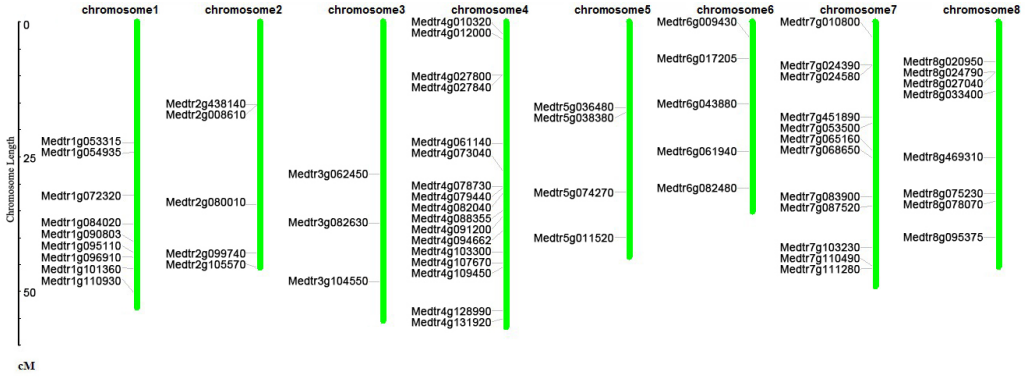

Figure 1. Chromosomal locations of tapetum specific male sterility related genes in *Medicago truncatula.*

Table 2. Conserved domains in tapetum specific male sterility related genes of *Medicago truncatula.*

Gene name	Gene ID	Predicted domains	Clusters of orthologous groups	Start	End	E-value
MtTsms1	Medtr1g084020	AAA	B	458	603	7.54e-12
MtTsms2	Medtr1g054935	AAA	B	19	210	5.41e-13
MtTsms3	Medtr2g438140	AAA	C	261	400	6.06e-25
MtTsms4	Medtr4g094662	AAA	C	264	403	4.28e-25
MtTsms5	Medtr4g109450	AAA	A	601	772	8.08e-12
MtTsms6	Medtr7g010800	AAA	C	241	380	8.7e-26
MtTsms7	Medtr1g101360	AAI	C	34	94	1.14e-8
MtTsms8	Medtr4g027840	AAI	A	27	111	6.77e-16
MtTsms9	Medtr4g027800	AAI	A	27	111	6.77e-16
MtTsms10	Medtr6g082480	AAI	A	27	92	0.0013
MtTsms11	Medtr7g065160	AAI	C	29	111	0.00000614
MtTsms12	Medtr3g082630	BBOX	C	4	47	1.5e-8
MtTsms13	Medtr7g083900	HLH	A	288	337	N/A
MtTsms14	Medtr3g104550	HSF	C	43	136	2.65e-58
MtTsms15	Medtr8g095375	Knot1	A	29	74	6.37e-11
MtTsms16	Medtr4g079440	Pept_C1	A	128	344	6.58e-118
MtTsms17	Medtr8g033400	Pfam:2-Hacid_dh_c	C	57	333	8.9e-19
MtTsms18	Medtr1g095110	Pfam:Abhydrolase_5	B	54	298	2.7e-56
MtTsms19	Medtr2g105570	Pfam:AMP-binging	C	32	451	1.5e-107
MtTsms20	Medtr8g024790	Pfam:Bhlh-MYC_N	B	26	181	4.5e-41
MtTsms21	Medtr1g072320	Pfam:Bhlh-MYC_N	C	11	188	2.9e-56
MtTsms22	Medtr2g008610	Pfam:Chloroa_b-bind	C	72	250	4.2e-48
MtTsms23	Medtr4g012000	Pfam:dCMP_cyt	A	1045	1146	3.7e-26
MtTsms24	Medtr7g068650	Pfam:DIOX_N	A	54	166	1.4e-31
MtTsms25	Medtr5g011520	Pfam:DUF296	C	128	241	1.2e-30
MtTsms26	Medtr7g103230	Pfam:DYW_deaminase	C	741	865	2.3e-33
MtTsms27	Medtr4g091200	Pfam:FAD_binding4	C	74	212	1.4e-27
MtTsms28	Medtr4g088355	Pfam:FAD_binding4	A	81	219	1.7e-25
MtTsms29	Medtr6g017205	Pfam:FAD_binding4	A	74	212	2e-26
MtTsms30	Medtr4g078730	Pfam:FAE_CUT1_RppA	A	81	260	4.5e-8
MtTsms31	Medtr7g053500	Pfam:Glyco_hydro_17	A	26	348	7.9e-102
MtTsms32	Medtr4g073040	Pfam:HMA	C	31	89	1.6e-16
MtTsms33	Medtr4g010320	Pfam:HSP20	C	125	231	4.2e-26
MtTsms34	Medtr6g061940	Pfam:HSP20	C	55	158	7.5e-35
MtTsms35	Medtr7g099680	Pfam:HSP70	A	9	618	6.5e-269
MtTsms36	Medtr7g024390	Pfam:HSP70	A	9	618	1.9e-269
MtTsms37	Medtr7g024580	Pfam:HSP70	A	9	618	1.9e-269
MtTsms38	Medtr5g074270	Pfam:Lipase_GDSL	A	34	348	1.7e-20
MtTsms39	Medtr8g075230	Pfam:Lipase_GDSL	A	37	349	6.6e-13
MtTsms40	Medtr1g096910	Pfam:MFS_2	A	37	521	2e-9
MtTsms41	Medtr2g080010	Pfam:NAM	C	16	142	3.3e-33
MtTsms42	Medtr8g027040	Pfam:p450	C	85	532	4.1e-82
MtTsms43	Medtr8g020950	Pfam:p450	A	39	500	7.4e-97
MtTsms44	Medtr7g087520	Pfam:p450	A	38	509	1.6e-96
MtTsms45	Medtr7g111280	Pfam:PAE	B	50	400	2.9e-164
MtTsms46	Medtr7g110490	Pfam:Pectinesterase	A	258	555	1.7e-122
MtTsms47	Medtr4g061140	Pfam:peroxidase	C	8	227	5.4e-47
MtTsms48	Medtr8g078070	Pfam:polyprenyl_synt	C	110	357	3.4e-58
MtTsms49	Medtr5g038380	Pfam:PTR2	A	100	536	1.7e-109
MtTsms50	Medtr4g107670	Pfam:PTR2	B	96	533	7e-105
MtTsms51	Medtr8g469310	Pfam:PTR2	A	98	533	7e-105
MtTsms52	Medtr3g098980	Pfam:Transferase	A	12	433	8.8e-93
MtTsms53	Medtr1g053315	Pfam:Transferase	A	16	457	1.8e-75
MtTsms54	Medtr5g036480	REC	B	16	147	2.27e-18
MtTsms55	Medtr1g090803	RING	B	799	837	0.249
MtTsms56	Medtr4g128990	S_TKc	C	262	531	6.27e-16
MtTsms57	Medtr6g009430	SANT	B	13	63	8.69e-14
MtTsms58	Medtr4g082040	SANT	C	10	21	N/A
MtTsms59	Medtr2g099740	SANT	B	13	63	4.92e-13
MtTsms60	Medtr7g451890	SERPIN	C	24	399	2.93e-90
MtTsms61	Medtr4g131920	transmembrane region	B	37	56	N/A
MtTsms62	Medtr3g062450	UBCc	A	4	147	5.22e-81
MtTsms63	Medtr1g110930	UBCc	A	4	147	5.22e-81

Phylogeny of TSMS related genes in *M. truncatula*

In order to compare the potential functions of TSMS related genes in *M. truncatula* and *A. thaliana*, we conducted a phylogenetic analysis with all 63 genes from *M. truncatula*

and 53 homologous genes from *A. thaliana*. Based on phylogenetic relationships and domain organization, genes could be separated into three groups (MtTsms A, MtTsms B, and MtTsms C), each of which contains genes encoding proteins with the same or similar domain organizations, as shown in Figure 2. A total of 28 TSMS genes from *M. truncatula* and 17 TSMS genes from *A. thaliana* were included in MtTsms A, while 11 TSMS genes from *M. truncatula* and 16 TSMS genes from *A. thaliana* were included in MtTsms B, and MtTsms C contained 24 TSMS genes from *M. truncatula* and 20 TSMS genes from *A. thaliana*. These results indicate that there is high homology among most TSMS related genes in *M. truncatula*. Interestingly, some proteins carrying the same domain were occasionally separated into different groups, such as MtTsms1 and 2; MtTsms5, 7, and 11; and MtTsms8, 9, and 10 (Table 2). This implies that the evolutionary histories of proteins with unknown C-terminal and N-terminal domains may be complex.

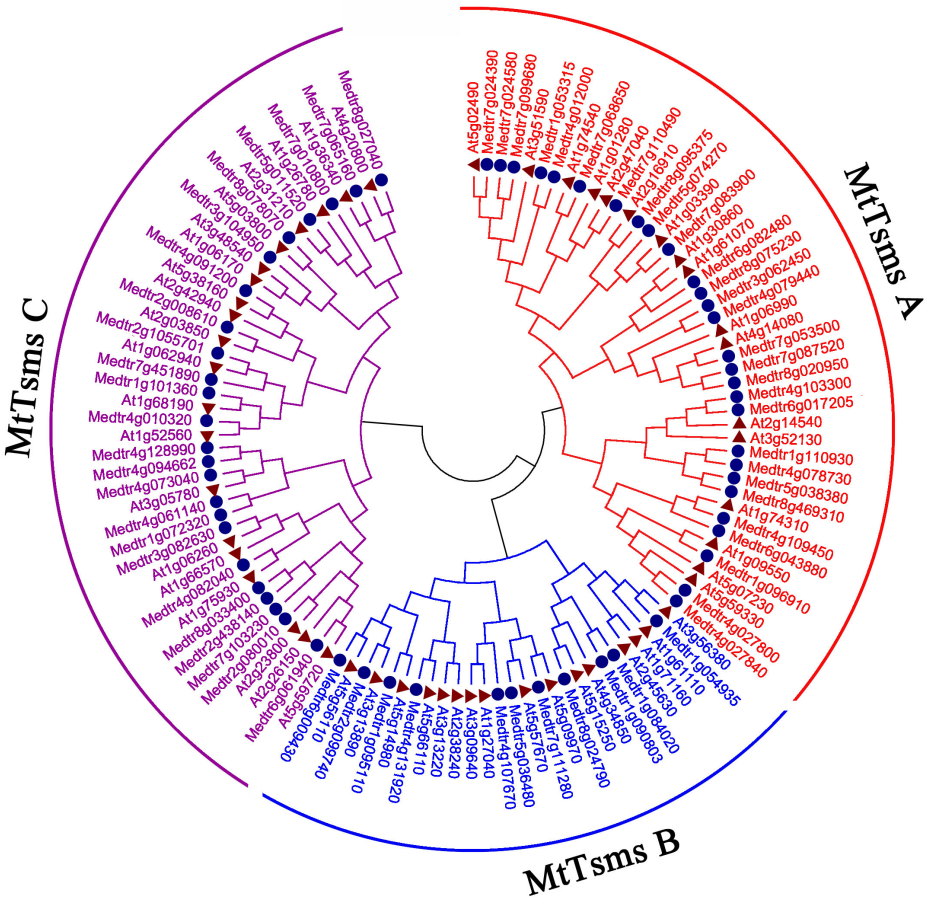

Figure 2. Neighbor-joining tree of tapetum specific male sterility related genes in *Medicago truncatula* and *Arabidopsis thaliana*. Tsms A genes are shown in red, Tsms B genes are shown in blue, and Tsms C genes are shown in purple.

Cis-acting regulatory elements of TSMS related genes in *M. truncatula*

Promoters, which are comprised of various *cis*-regulatory elements, control the development and physiology of plants by regulating gene expression (Pauls et al., 2015). To identify *cis*-regulatory elements that potentially mediate the transcription of TSMS genes, promoter regions of the TSMS genes were analyzed using the PLACE database. In this study, eight *cis*-elements (ABRELATERD1, ARR1AT, CAAT box, CACTFTPPCA1, GTGANTG10, POLLENLELAT52, TATA box, and WRKY71OS) were identified in the promoters of putative TSMS genes (Table 3). ABRELATERD1 is a hormonal-response element, while ARR1AT is involved in cytokinin responsiveness (Li et al., 2015). The CAAT box is generally found upstream of the TATA box; it can be combined with one or more transcription factors and may be related to the structure of RNA polymerase II (Van Opijnen et al., 2004). CACTFTPPCA1 enhances the expression of light reaction components (Wilson and Zhang, 2009). GTGANTG10 and POLLENLELAT52 are enhancer elements of LAT52 and LAT56, and are essential for high expression levels in pollen (Manimaran et al., 2015). The TATA box is an essential element of the eukaryotic promoter and controls the accuracy and frequency of transcription (Van Opijnen et al., 2004). WRKY71OS is involved in components of defense reactions and related disease-resistance (Zhang et al., 2004).

Table 3. Characteristics of *cis*-motifs of male sterility genes identified using the PLACE database.

PLACE *cis*-motif	Sequence	Function	No. of genes
ABRELATERD1	ACGTG	hormone-responsive elements	47
ARR1AT	NGATT	element involved in cytokinin responsiveness	63
CAATBOX1	CAAT	defense-responsive element	63
CACTFTPPCA1	YACT	light-responsive elements	55
GTGANTG10	GTGA	pollen	62
POLLEN1LELAT52	AGAAA	pollen	62
TATABOX	TATAAAT	TATA box	63
WRKY71OS	TGAC	defense-responsive element	60

Expression profiles of TSMS related genes in *M. truncatula*

To investigate the expression patterns of TSMS related genes in different tissues, RNA-seq data for 58 genes (all existing data) were downloaded from the *M. truncatula* Gene Expression Atlas of the Samuel Roberts Noble Foundation. Three-color clustering was used to indicate the strength of gene expression, with bright red representing the strongest signal, bright green representing the weakest signal, and black representing an average signal. The results (Figure 3) illustrated differential transcript abundances of the 58 genes in seven tissues, including root, stem, leaf, vegetative bud, flower, pod, and seed coat. Four genes (MtTsms7, 16, 30, and 57) were expressed at high levels across all seven tissues, indicating a role in constitutive transport processes throughout the plant (Paul et al., 1992). Nine genes (MtTsms5, 14, 27, 32, 33, 34, 43, 49, and 51) were found to be expressed at low levels across all tissues. The expression profiles of these genes diverged, implying that they contribute to functional maintenance through regulatory subfunctionalization or neofunctionalization (Qu et al., 2008). MtTsms11 and MtTsms25 were highly expressed in root tissue, while MtTsms19 and MtTsms56 were highly expressed in stem tissue. MtTsms2 was highly expressed in leaves, and MtTsms8, 9, 35, 46, 48, and 63 were highly expressed in flowers. To identify genes with

potential roles in *M. truncatula* abiotic and biotic stress responses, the expression patterns of all 58 genes were investigated in plants exposed to high salinity, NH_4, NO_3, N_2, NAA, and *Phymatotrichum* infection. MtTsms19, 20, 27, 32, 40, 49, and 52 were upregulated in root tissues under salt stress, indicating that these genes may be involved in the salt stress response. MtTsms2, 3, 13, 63, and 48 were upregulated in leaves in the presence of the growth regulator NAA, indicating that these genes may be involved in plant growth associated with an NAA response. MtTsms10, 11, 25, 43, and 47 were upregulated in roots, but downregulated in shoots under nitrogen stress. Only MtTsms47 was upregulated in roots under *Phymatotrichum* infection.

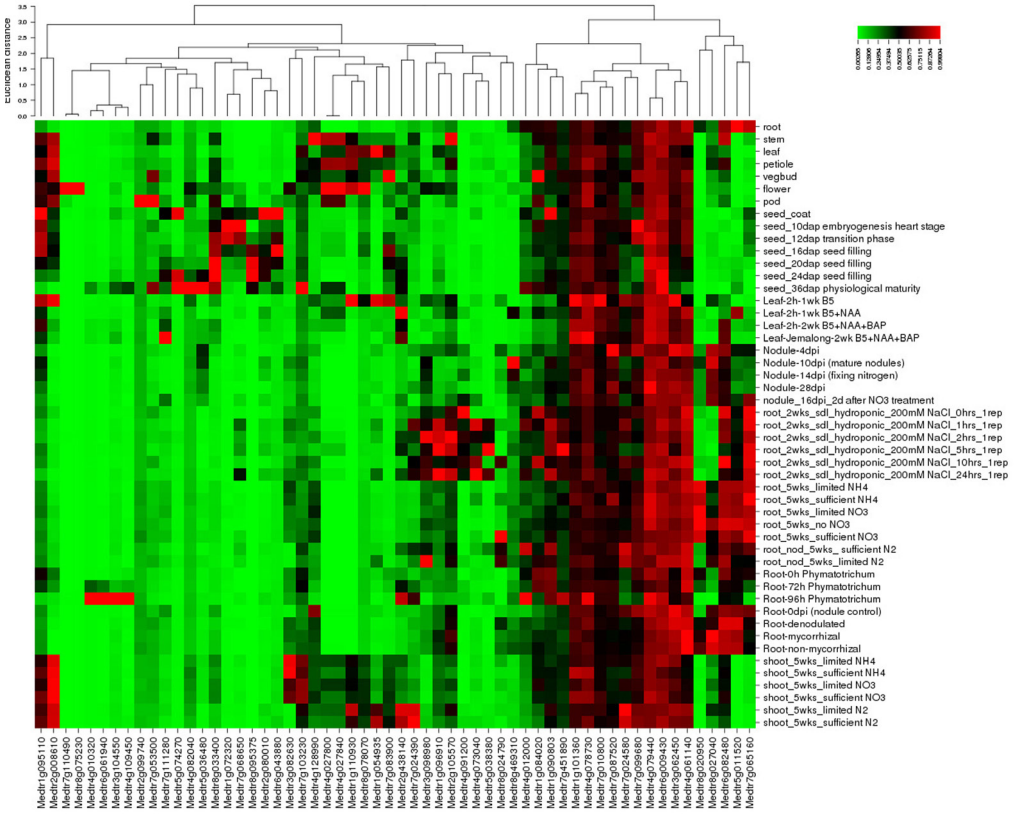

Figure 3. Expression profiles of TSMS related genes in *Medicago truncatula*. Three-color clustering indicates the strength of gene expression, with bright red representing the strongest signal, bright green representing the weakest signal, and black representing an average level of expression.

DISCUSSION

Following the determination of the *Medicago* and *Arabidopsis* genome sequences, data became available for the comparative study of MS related genes in the two plants. The results of this study may provide novel insights into the evolution and development of TSMS genes in *M. truncatula*. We identified 63 putative TSMS related genes in *M. truncatula* and

investigated their structure and chromosomal organization. In previous studies, *PUB4*, *A6*, *LTP12*, *LAP5*, and *TAP44* were identified in *Arabidopsis* and were found to play an important role in early tapetum formation and floral organ differentiation (Liu et al., 2009; Kim et al., 2010; Matsuo et al., 2013; Wang et al., 2013). In our study, homologs of these genes in *M. truncatula* were all found to contain a domain associated with sterility. In addition, a large percentage of TSMS related genes contained very few introns (49.20% with 0-1 intron), similar to those of *A. thaliana* and *O. sativa* (Zhang et al., 2006; Matsuo et al., 2013).

Based on a phylogenetic tree and the analysis of conserved domains, we found that *Medicago* and *Arabidopsis* genes clustered closely together, supporting the close evolutionary relationship of the two species (Liu et al., 2009). Both species are dicotyledonous rosids, a group that appeared 108-117 million years ago (Wikström et al., 2001). We found similar results when BLAST was used to identify *M. truncatula* homologs of *A. thaliana* TSMS related genes. Through analysis of conserved domains in the resulting proteins, we discovered that TSMS related genes often contain domains associated with sterility traits, including the AAA, AAI, BBOX, HLH, Knot1, RING, and SANT domains. The AAA domain confers ATPase activity, and correlates with abnormal development in plants (Ye et al., 2015). The AAI domain is found in the BcMF15 gene of *B. oleracea*, which is mainly expressed in the stamen, indicating that the AAI domain may participate in fertility regulation (Tian et al., 2009). The BBOX protein is involved in the development of light signal conditioning in *A. thaliana* (Kumagai et al., 2008). The HLH domain is found in the gene OsHLH164 and is expressed in the boundary region between the apical meristem and the new formative meristem; this gene is the main regulatory factor controlling the formation of axillary bud primordia in *O. sativa* and contributes to plant growth and development (Zhu et al., 2005). The defensin gene Bnfef of *Brassica napus* contains the Knot1 domain, which plays a significant role in the plant's defense response. The expression of the defensin gene is highest in flower buds, suggesting that it may also be related to fertility (Zheng et al., 2015). AtDUO1 and AtDUO3 from *A. thaliana* both contain SANT domains and participate in male reproductive cell division and differentiation, influencing sporophyte and male gamete development (Brownfield et al., 2009). REC and S_TKc domains are correlated with disease resistance. In a study of *O. sativa*, the REC domain was found to be associated with phosphoric acid signal recognition receptors, and the S_TKc domain was found to be associated with serine and threonine protein kinase family; genes containing these two domains are believed to be involved in disease resistance (Manimaran et al., 2015). UBCc, a ubiquitin combined with an enzyme E2 catalytic domain structure, was found in a cloned ZmPHO2 protein in inbred maize lines and is related to the dynamic balance of phosphorus in plants. The HSP domain is found in heat-shock factors, which are less common in plants (Pant et al., 2015). The binding site of the *O. sativa* gene OsHsp90 is located behind an ATPase domain and plays an important role in proteins related to resistance (Raman and Suguna, 2015). The Pept_C1 domain has been extensively studied in animals, where it is expressed at the highest level in the kidneys (Funkelstein et al., 2008); however, it has not been previously reported in plants.

Although the function of most gene products is dependent on both transcriptional and post-transcriptional regulation, the spatial and temporal expression patterns and transcript levels of genes are primarily regulated by the *cis*-elements present in promoters (Qu et al., 2008). Analysis of the *cis*-elements present among *M. truncatula* TSMS related genes identified eight different *cis*-elements (ABRELATERD1, ARR1AT, CAATBOX1, CACTFTPPCA1, GTGANTG10, POLLEN1LELAT52, TATABOX, and WRKY71OS). In a

previous study, ABRELATERD1 was found to be a hormone-responsive element related to the binding of MYB and zinc finger motif MADS-box transcription factors to promoter regions (Mahajan et al., 2015). ARR1AT, GTGANTG10, WRKY71OS, and TATABOX5 elements are strong candidates for transcriptional enhancers (Pauls et al., 2015). POLLEN1LELAT52 and GTGANTG10 are pollen-specific *cis*-elements, identified in bicellular and tricellular pollen enriched genes, respectively, among rice TSMS genes (Wei et al., 2010). The WRKY71OS element, which is highly represented among *M. truncatula* TSMS genes, contains an OsWRKY71 binding site, suggesting that WRKY71OS may participate in gibberellin signal repression (Zhang et al., 2004).

Our transcriptome analysis revealed that four genes (MtTsms7, 16, 30, and 57) were expressed among all seven tissues analyzed, while nine genes (MtTsms5, 14, 27, 32, 33, 34, 43, 49, and 51) were found to have low expression among all tissue types, and six genes (MtTsms8, 9, 35, 46, 48, and 63) demonstrated high expression in flowers. These patterns indicate that gene expression during different developmental stages can result in abnormal downstream or upstream reactions (Fujii et al., 2010), and that many genetic pathways are associated with the tapetum developmental network (Wilson and Zhang, 2009). Construction of a heat-map showing gene expression under various osmotic-related stresses in different tissues/organs (Figure 3) revealed that these genes play important roles in regulating the stress response (Qu et al., 2008). Taken together, our comparative analysis data of TSMS related genes showed that these genes vary in terms of their expression patterns in *M. truncatula*.

In summary, we identified 63 putative TSMS related genes in *M. truncatula*. We performed comprehensive analyses of gene structure, conserved domains, *cis*-elements, phylogeny, and expression profiling in various tissues, and under abiotic stresses. Based on structural characteristics and a comparison of the phylogenetic relationships between *M. truncatula* and *A. thaliana* homologs, the genes were classified into three groups. Conserved domains related to sterility (AAA, AAI, BBOX, HLH, Knot1, RING, and SANT) and pollen-specific *cis*-elements (GTGANTG10 and POLLEN1LELAT52) were identified. Expression analyses revealed that putative TSMS related genes may be involved in developmental processes and abiotic stress responses. This study provides an overview of these genes in *M. truncatula* and provides insights into MS, although the mechanisms underlying the participation of these genes require further study.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Science Foundation of China (#31372362).

REFERENCES

Brownfield L, Hafidh S, Durbarry A, Khatab H, et al. (2009). Arabidopsis DUO POLLEN3 is a key regulator of male germline development and embryogenesis. *Plant Cell* 21: 1940-1956. http://dx.doi.org/10.1105/tpc.109.066373

Fujii S, Yamada M, Fujita M, Itabashi E, et al. (2010). Cytoplasmic-nuclear genomic barriers in rice pollen development revealed by comparison of global gene expression profiles among five independent cytoplasmic male sterile lines.

Plant Cell Physiol. 51: 610-620. http://dx.doi.org/10.1093/pcp/pcq026

Funkelstein L, Toneff T, Hwang SR, Reinheckel T, et al. (2008). Cathepsin L participates in the production of neuropeptide Y in secretory vesicles, demonstrated by protease gene knockout and expression. *J. Neurochem.* 106: 384-391. http://dx.doi.org/10.1111/j.1471-4159.2008.05408.x

Honys D and Twell D (2003). Comparative analysis of the *Arabidopsis* pollen transcriptome. *Plant Physiol.* 132: 640-652. http://dx.doi.org/10.1104/pp.103.020925

Hsu SW, Liu MC, Zen KC and Wang CS (2014). Identification of the tapetum/microspore-specific promoter of the pathogenesis-related 10 gene and its regulation in the anther of *Lilium longiflorum. Plant Sci.* 215-216: 124-133. http://dx.doi.org/10.1016/j.plantsci.2013.11.006

Hyung D, Lee C, Kim JH, Yoo D, et al. (2014). Cross-family translational genomics of abiotic stress-responsive genes between Arabidopsis and *Medicago truncatula. PLoS One* 9: e91721. http://dx.doi.org/10.1371/journal.pone.0091721

Ji RQ, Song Q, Xin XF, Zhou X, et al. (2011). Isolation of fertility-related genes of multiple-allele-inherited male sterility in *Brassica rapa* ssp *pekinensis* by cDNA-AFLP. *Genet. Mol. Res.* 10: 4073-4083. http://dx.doi.org/10.4238/2011.December.5.8

Kaul ML (1988). Male sterility in higher plants. Springer-Verlag Press, Berlin.

Kim SS, Grienenberger E, Lallemand B, Colpitts CC, et al. (2010). LAP6/POLYKETIDE SYNTHASE A and LAP5/POLYKETIDE SYNTHASE B encode hydroxyalkyl α-pyrone synthases required for pollen development and sporopollenin biosynthesis in *Arabidopsis thaliana. Plant Cell* 22: 4045-4066. http://dx.doi.org/10.1105/tpc.110.080028

Kumagai T, Ito S, Nakamichi N, Niwa Y, et al. (2008). The common function of a novel subfamily of B-Box zinc finger proteins with reference to circadian-associated events in *Arabidopsis thaliana. Biosci. Biotechnol. Biochem.* 72: 1539-1549. http://dx.doi.org/10.1271/bbb.80041

Li L, Li Y, Song S, Deng H, et al. (2015). An anther development F-box (ADF) protein regulated by tapetum degeneration retardation (TDR) controls rice anther development. *Planta* 241: 157-166. http://dx.doi.org/10.1007/s00425-014-2160-9

Liu J, Pang C, Wei H, Song M, et al. (2014). Proteomic analysis of anthers from wild-type and photosensitive genetic male sterile mutant cotton (*Gossypium hirsutum* L.). *BMC Plant Biol.* 14: 390. http://dx.doi.org/10.1186/s12870-014-0390-4

Liu X, Huang J, Parameswaran S, Ito T, et al. (2009). The SPOROCYTELESS/NOZZLE gene is involved in controlling stamen identity in *Arabidopsis. Plant Physiol.* 151: 1401-1411. http://dx.doi.org/10.1104/pp.109.145896

Ma LC, Wang YR, Liu WX and Liu ZP (2013). Expression analysis of seed-specific genes in four angiosperm species with an emphasis on the unconserved expression patterns of homologous genes. *Seed Sci. Res.* 23: 223-231. http://dx.doi.org/10.1017/S0960258513000305

Ma Y, Kang J, Wu J, Zhu Y, et al. (2015). Identification of tapetum-specific genes by comparing global gene expression of four different male sterile lines in *Brassica oleracea. Plant Mol. Biol.* 87: 541-554. http://dx.doi.org/10.1007/s11103-015-0287-0

Mahajan NS, Dewangan V, Lomate PR, Joshi RS, et al. (2015). Structural features of diverse Pin-II proteinase inhibitor genes from *Capsicum annuum. Planta* 241: 319-331. http://dx.doi.org/10.1007/s00425-014-2177-0

Manimaran P, Raghurami Reddy M, Bhaskar Rao T, Mangrauthia SK, et al. (2015). Identification of cis-elements and evaluation of upstream regulatory region of a rice anther-specific gene, OSIPP3, conferring pollen-specific expression in *Oryza sativa* (L.) ssp. *indica. Plant Reprod.* 28: 133-142. http://dx.doi.org/10.1007/s00497-015-0264-4

Matsuo Y, Arimura SI and Tsutsumi N (2013). Distribution of cellulosic wall in the anthers of *Arabidopsis* during microsporogenesis. *Plant Cell Rep*. 32: 1743-1750.

Pant BD, Burgos A, Pant P, Cuadros-Inostroza A, et al. (2015). The transcription factor PHR1 regulates lipid remodeling and triacylglycerol accumulation in Arabidopsis thaliana during phosphorus starvation. *J. Exp. Bot.* 66: 1907-1918. http://dx.doi.org/10.1093/jxb/eru535

Paul W, Hodge R, Smartt S, Draper J, et al. (1992). The isolation and characterisation of the tapetum-specific *Arabidopsis thaliana* A9 gene. *Plant Mol. Biol*. 19: 611-622. http://dx.doi.org/10.1007/s12298-012-0120-z

Pauls S, Goode DK, Petrone L, Oliveri P, et al. (2015). Evolution of lineage-specific functions in ancient cis-regulatory modules. *Open Biol.* 5: 150079. http://dx.doi.org/10.1098/rsob.150079

Qu Q, Xing YP, Liu WX, Xu XP, et al. (2008). Expression pattern and activity of six glutelin gene promoters in transgenic rice. *J. Exp. Bot.* 59: 2417-2424. http://dx.doi.org/10.1093/jxb/ern110

Raman S and Suguna K (2015). Functional characterization of heat-shock protein 90 from *Oryza sativa* and crystal structure of its N-terminal domain. *Acta Crystallogr. F Struct. Biol. Commun.* 71: 688-696. http://dx.doi.org/10.1107/S2053230X15006639

Tian A, Cao J, Huang L, Yu X, et al. (2009). Characterization of a male sterile related gene BcMF15 from *Brassica campestris* ssp. chinensis. *Mol. Biol. Rep.* 36: 307-314. http://dx.doi.org/10.1007/s11033-007-9180-5

Town CD, Cheung F, Maiti R, Crabtree J, et al. (2006). Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* 18: 1348-1359. http://dx.doi.org/10.1105/tpc.106.041665

van Opijnen T, Kamoschinski J, Jeeninga RE and Berkhout B (2004). The human immunodeficiency virus type 1 promoter contains a CATA box instead of a TATA box for optimal transcription and replication. *J. Virol.* 78: 6883-6890. http://dx.doi.org/10.1128/JVI.78.13.6883-6890.2004

Wang H, Lu Y, Jiang T, Berg H, et al. (2013). The *Arabidopsis* U-box/ARM repeat E3 ligase AtPUB4 influences growth and degeneration of tapetal cells, and its mutation leads to conditional male sterility. *Plant J.* 74: 511-523. http://dx.doi.org/10.1111/tpj.12146

Wei LQ, Xu WY, Deng ZY, Su Z, et al. (2010). Genome-scale analysis and comparison of gene expression profiles in developing and germinated pollen in *Oryza sativa*. *BMC Genomics* 11: 338. http://dx.doi.org/10.1186/1471-2164-11-338

Wikström N, Savolainen V and Chase MW (2001). Evolution of the angiosperms: calibrating the family tree. *Proc. Biol. Sci.* 268: 2211-2220. http://dx.doi.org/10.1098/rspb.2001.1782

Wilson ZA and Zhang DB (2009). From *Arabidopsis* to rice: pathways in pollen development. *J. Exp. Bot.* 60: 1479-1492. http://dx.doi.org/10.1093/jxb/erp095

Wu Y, Min L, Wu Z, Yang L, et al. (2015). Defective pollen wall contributes to male sterility in the male sterile line 1355A of cotton. *Sci. Rep.* 5: 9608. http://dx.doi.org/10.1038/srep09608

Ye Q, Rosenberg SC, Moeller A, Speir JA, et al. (2015). TRIP13 is a protein-remodeling AAA+ ATPase that catalyzes MAD2 conformation switching. *eLife* 4: e07367.

Zhang XL and Zhang LG (2014). Molecular cloning and expression of the male sterility-related CtYABBY1 gene in flowering Chinese cabbage (Brassica campestris L. ssp chinensis var. parachinensis). *Genet. Mol. Res.* 13: 4336-4347. http://dx.doi.org/10.4238/2014.June.10.1

Zhang ZL, Xie Z, Zou X, Casaretto J, et al. (2004). A rice WRKY gene encodes a transcriptional repressor of the gibberellin signaling pathway in aleurone cells. *Plant Physiol.* 134: 1500-1513. http://dx.doi.org/10.1104/pp.103.034967

Zhang ZS, Lu YG, Liu XD, Feng JH, et al. (2006). Cytological mechanism of pollen abortion resulting from allelic interaction of F1 pollen sterility locus in rice (*Oryza sativa* L.). *Genetica* 127: 295-302. http://dx.doi.org/10.1007/s10709-005-4848-z

Zheng XM, Guo N, Gao TS and Zhang T (2015). Cloning and expression analysis of defensin genes from *Brassica napus*. *Acta Agron. Sin.* 41: 725-732. http://dx.doi.org/10.3724/SP.J.1006.2015.00725

Zhu H, Qian W, Lu X, Li D, et al. (2005). Expression patterns of purple acid phosphatase genes in Arabidopsis organs and functional analysis of AtPAP23 predominantly transcribed in flower. *Plant Mol. Biol.* 59: 581-594. http://dx.doi.org/10.1007/s11103-005-0183-0