

Contents of Supplementary Data:

1. Pipeline-Annotated “Pollen Allergens” Not Specifically Expressed in Rice Pollen
2. Ole e 1-related allergens (as characterized by Jiang et al., 2005)
3. Ory s 1 (Further Details)
4. Ory s 2 (Further Details)
5. Additional Potential Group I, II/III Allergens (Further Details)
6. Profilin A (Ory s 12) (Further Details)
7. Ory s 23 (Further Details)
8. Other Suspected Allergens (Further Details)
 - a. EF Hand
 - b. Pectate Lyase
 - c. Glycoside hydrolase family 28, Exopolygalacturonase
 - d. FAD Binding

1. Pipeline-Annotated “Pollen Allergens” Not Specifically Expressed in Rice Pollen

Pipeline annotation of the rice genome includes loci annotated in the rice genome as “pollen allergens” particularly those from dicots (annotated as being similar to “Alt a 7”, “Ole e 8”, and “Bet v 1-D/H”) and more distant gymnosperms (“Jun a 1”) were poor candidates and were not found in rice. “Bet v 1-D/H”-like proteins, abundantly expressed in seeds and sporophytic tissues (supported by microarray data, digital Northern, and ESTs), are potential allergens, but not pollen allergens. Not abundant, but nonetheless produced mainly in pollen were putative Ory s 4, (pipeline annotated as “pollen allergen Phl p 4,” LOC_Os05g28490) and a putative Ory s 3 (“Group 3 pollen allergen, putative, expressed,” LOC_Os06g45210).

2. Ole e 1-related allergens (as characterized by Jiang et al., 2005)

Numerous Ole e I related sites were documented in the genomic assay of Jiang et al. (2005), but as is the case for other dicot allergens, in fact only two loci were highly activated and upregulated in pollen grains—LOC_Os06g36240.1, pollen allergen Ory s 11 (pipeline annotated as Phl p 11), and LOC_Os09g39950.1, pollen-specific protein C13 precursor (Supplementary Table S1). The Ory s 11 locus encodes a protein with a signal

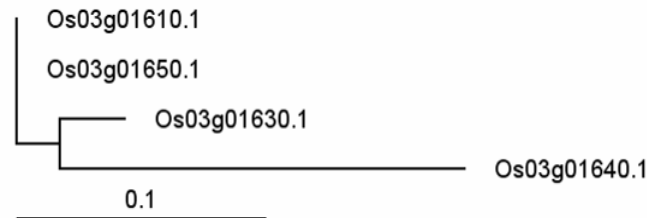
peptide (27 AVA-TA) with 40% identity and 64% positives with an *Arabidopsis* gene (At1g29140, pollen Ole e 1 allergen and extensin family protein, $p < 9e^{-29}$); this protein is also expressed in MPSS at an intensity rank of 246 and has proteome support (Dai et al., 2007), indicating translation and reasonable abundance. The pollen-specific protein C13 precursor locus also encodes a protein with a signal peptide (27 AVA-TA) and has 38% identity and 53% positives with an *Arabidopsis* gene (At4g08685, SAH7, $p < 2e^{-20}$); it is 63rd in intensity among MPSS signatures and is represented in three matches to rice proteins in two proteome studies, reflecting high abundance in mature pollen (Dai et al., 2006, 2007). Both contain a predicted transmembrane domain enclosing a signal peptide region at the C-end, suggesting a hydrophobic terminus. *Arabidopsis* homologues both reflect annotation in TAIR8 as pollen Ole e 1 allergens and extensin family proteins, indicating retention of allergic activity across dicot and monocot classes; these are presumed to share a similar role in cell wall modification.

Supplementary Table S1. Pollen expression microarray data on sequences with “Ole e 1” motif. (Predicted MWs 18.86 to 17,73 KDa, respectively).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (with Pfams, orientation and introns)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
45.	Os.53412.1.S1_at	Os06g36240.1 (11)	LOC_Os06g36240 Pollen allergen Ory s 11, putative, expressed (Pfam:PF01190) (+ with 1 intron)	8.858	12.285	3.18E-12	ESTs: LOC_Os06g36240 (9)	GATCTGTTTTGTAC GTA 95 GATCTGCGGGGGGC GCC 661	9b Dai et al., 2007
200.	Os.50339.1.S1_at	Os09g39950.1 (11)	LOC_Os09g39950 Pollen-specific protein C13 precursor, putative, expressed (Pfam:PF01190) (+ with no introns)	7.375	10.885	4.68E-08	ESTs: LOC_Os09g39950 (15)	GATCGGCGGAGGG GACG 2307	BAD54680, 10b, 11b Dai et al., 2006, 2007

3. Ory s 1 (Further Details)

Outside of the coding region, there are minor differences in the different isoforms of LOC_Os03g01610 and LOC_Os03g01650, including a slightly shorter 5'-UTR region for LOC_Os03g01650 (22 bp vs. 35 bp) and a slightly longer 3'UTR region (400 versus 388 bp); both are attributed with ESTs. The most similar gene to the former two is LOC_Os03g01630, which possesses 96% identity (before removal of its 117 bp intron), whereas LOC_Os03g01640 is the least similar, displaying a 82% identity (Supplementary Figure S1). The presence of an intron in LOC_Os03g01630 is unusual. Genomic information from *Oryza sativa* ssp. *indica* indicates similar sequence patterns that are conserved in other *Oryza* species.



Supplementary Figure S1. Phylogenetic diagram of relationships of *Oryza sativa* loci using sequence of putative proteins. LOC_Os03g01610.1 is identical to LOC_Os03g01650.1 though the latter is on the (-) strand, whereas the former is in opposite orientation on the (+) strand. Note grouping of loci in chromosome 4 match locus on (+) strand with locus on (-) strand. Chromosome 6 sequences are also largely grouped in triplets. Note significant divergence of amino acid sequences.

According to probe set Os.2402.1.S1_at (which matches LOC_Os03g01610 and LOC_Os03g01650), *Oryza sativa* 1 was the second most upregulated transcript in rice pollen, displaying 9.18 log₂ intensity of label, upregulated over the seedling control by 13.185 log₂ (>×8000). Probe set Os.2405.1.S1_at ranked 30th among activated probes (full match for LOC_Os03g01640), displaying an intensity of 7.805 log₂ upregulated by 12.447 log₂ over the seedling control (Supplementary Table S2). Abundance of *Oryza sativa* 1 transcripts in pollen is also strongly supported by MPSS (<http://mpss.udel.edu/rice/>, dataset NPO; Nobuta et al., 2007) with *Oryza sativa* 1-related signatures in the top 85 signatures in mature pollen. Proteome support indicates abundant translation products, coinciding most closely with LOC_Os03g01610 and LOC_Os03g01650 sequences, and accounting for nine reported proteins from five proteome studies to date on rice (Kerim et al., 2003a, 2003b; Imin et al., 2004; Dai et al., 2006, 2007). *Oryza sativa* 1 members appear to be exclusively expressed in pollen, so the number of recovered ESTs has been relatively low; pollen has not been a direct target of EST studies, unlike MPSS and proteome studies. As indicated in the text, the four *Oryza sativa* 1 loci are phylogenetically closely related to one another.

The upstream promoter region of these four genes had several areas of local sequence conservation, with the most prevalent area of conservation (-68 to -51) coinciding with the TAAATA sequence noted as a putative TATA box by Xu et al. (1999). Typical pollen vegetative cell promoter motifs, for example those of LAT52 (Bate and Twell, 1998) and g10 (Rogers et al., 2001) promoter motifs were not noted within positions coinciding with promoter activity in those prior studies. Sequence information and alignment are in another supplementary data file.

Supplementary Table S2. Pollen expression microarray data on Ory s 1. (Predicted molecular weights 24.63 to 28.94 KDa, measured weights have varied from 30-34 KDa).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (with Pfams, strand orientation and introns)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
2	Os.2402.1.S1_at	LOC_Os03g01610 (11) LOC_Os03g01630 (3) LOC_Os03g01650 (11)	[LOC_Os03g01610 ¹ , Major pollen allergen Ory s 1 precursor, putative, expressed (Pfam:PF01357, PF03330)] (+ with no intron), [LOC_Os03g01630 ² , Major pollen allergen Ory s 1 precursor, putative (Pfam:PF01357,PF03330) (+ with 1 intron)], [LOC_Os03g01650 ³ , Major pollen allergen Ory s 1 precursor, putative, expressed (Pfam:PF01357, PF03330)] (- with no introns)	9.814	13.185	5.29E-08	Xu et al., 1995, 1999 ESTs: Os03g01610 (28) Os03g01630 (0) Os03g01650 (18)	GATCAAGTGCTC CAAGC ¹⁻³ 77 GATCACCTCCAC ATCG ¹⁻³ 52 GATCATCGCCGA GGACG ¹⁻³ 1097 GATCCGGCGCGG CTGCC ^{1,3} 1772	^{1,3} Kerim et al., 2003a, 2003b; Imin et al., 2004; Dai et al., 2006, 2007 ² LOC_Os03g01630 Kerim et al., 2003b; Dai et al., 2006
30	Os.2405.1.S1_at	LOC_Os03g01640 (11)	LOC_Os03g01640 Major pollen allergen Ory s 1 precursor, putative, expressed (Pfam:PF01357,PF03330) (+ with no introns)	7.805	12.447	4.49E-08	Xu et al., 1995, 1999 ESTs: Os03g01640 (32)	GATCCAAGTCGA CTAAG 367 GATCGAGTCTTC AATTC 76 GATCAGAAAATT TTGAT 18	Kerim et al., 2003b; Dai et al., 2006, 2007

4. Ory s 2 (Further Details)

MPSS supports that these Phl p 2 pollen allergen homologues are transcriptionally very highly upregulated. There are ten MPSS signatures—seven, corresponding to the loci listed below, are in the top 25 MPSS for rice pollen.

7. LOC_Os04g26220 (Ory s 2-C)
8. LOC_Os04g26230 (Ory s 2-B)
10. LOC_Os06g45290 (Ory s 2-A)
12. LOC_Os06g45180 (Ory s 2-A)
17. LOC_Os04g25160 (Ory s 2-C)
18. LOC_Os06g44470 (Ory s 2)
23. LOC_Os04g25150 (Ory s 2-B)

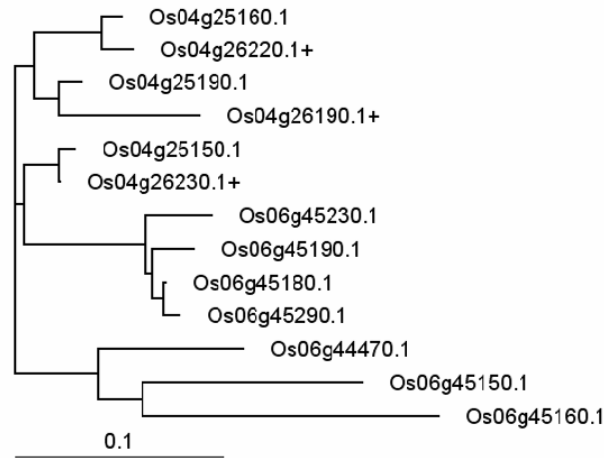
Less abundant transcripts are LOC_Os06g45150 (Ory s 2), which is the 78th most highly upregulated locus, and LOC_Os06g45160 (Ory s 2), which is 104th. LOC_Os06g45190 (Ory s 2) is present at lower levels but is clearly also transcribed (<http://mpss.udel.edu/rice/>, dataset NPO; Nobuta et al., 2007).

Abundance of translated products is supported by proteome data (Kerim et al., 2003b; Dai et al., 2006, 2007), the latter studies supporting loci LOC_Os04g25160, LOC_Os04g25190, LOC_Os06g44470, and LOC_Os06g45180. Thus, high expression of Ory s 2 in rice pollen is supported by transcriptome and proteome data. In Kerim et al. (2003b), sequences corresponding to these were designated Ory s 2-C.

The genomic arrangement of the Ory s 2 family suggests that its loci natively lack introns; intron-bearing loci appear to represent altered copies that may be functionless pseudogenes (LOC_Os04g26190 and LOC_Os06g45200). There appear to be four major groups of Ory s 2 genes. Two groups are located on chromosome 4 and two on chromosome 6, with triplicate pairings a common pattern. On chromosome 6, all copies were on the (-) strand. Of these, LOC_Os06g45190, LOC_Os06g45180, LOC_Os06g45230, and LOC_Os06g45290 were most similar; LOC_Os06g44470, LOC_Os06g45150, LOC_Os06g45160 formed a second grouping that was nested within nearer to some chromosome 4 sequences (see Figure 2, Supplementary Table S3).

In contrast, loci on chromosome 4 are found in groups of three and each of these shares strong sequence similarity with loci on the opposite strand and oriented in the opposite direction. The (-) strand of chromosome 4 has three loci, LOC_Os04g25150, LOC_Os04g25160, and LOC_Os04g25190. These sequences are most closely paired with LOC_Os04g26230, LOC_Os04g26220 and LOC_Os04g26190, respectively. Thus, this sequence appears to represent a reversed insertion of a locus triplet on the reverse strand that also seems to include interstitial loci as well (Fig. 2).

The remaining seven related sequences are found on chromosome 6 in tandem and non-tandem groupings and all on this chromosome are coded on the (-) strand. Predicated peptides display sufficient variability that less than half of the coded amino acid sequences are identical (51 of the 117), with slightly greater conservation of sequence in the “pollen allergen group I” motif region. The most divergent genes are LOC_Os06g45200, which has two introns, and LOC_Os04g26190, which is a truncated version. Both of these excise the signal peptide region and have not been recovered. LOC_Os04g26190 encodes essentially only the core pollen allergen group I motif (PF01357) from position 2 to 75, a 90 amino acid sequence that is a close match for Lol p 2-A in organization and sequence, though it seems to have arisen from Ory s 2 (see Supplementary Figure S2). Kerim et al. (2003b) named Ory s 2-A based on SWISS PROT P83466 (LOC_Os06g45180 and LOC_Os06g45290), Ory s 2-B based on CAD40510 (LOC_Os04g26230) and Ory s 2-C based on CAD40509 (LOC_Os04g25160).



Supplementary Figure S2. Phylogenetic diagram of relationships of Ory s 2 loci using sequence of putative proteins (omitting LOC_Os06g45200, which is the most divergent). Note grouping of loci in chromosome 4 match locus on (+) strand with locus on (-) strand. Chromosome 6 sequences are also largely grouped in triplets. Note significant divergence of amino acid sequences.

Supplementary Table S3. Pollen expression microarray data on Ory s 2. (Predicted molecular weights 12.30 to 12.46 KDa, near measured weight on 2D).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (with Pfams, orientation and introns)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
32	Os.9665.1.S1_at	LOC_Os04g25190.1 (11); LOC_Os04g26190.1 (5)	[LOC_Os04g25190 ¹ , Pollen allergen Ory s 2 precursor, putative, expressed (Pfam:PF01357) (- with no introns)], [LOC_Os04g26190 ² , Pollen allergen Lol p 2-A, putative (Pfam:PF01357)] (+ with 1 intron)	8.222	12.427	1.47E-10	ESTs: LOC_Os04g25190 (16) LOC_Os04g26190 (0)	^{1,2} GATCGTGTTCATGTTTGA (11946)	XP_471811 Dai et al. 2006, 2007
75	Os.7428.1.S1_at	LOC_Os06g44470.1 (11)	LOC_Os06g44470 Pollen allergen Ory s 2 precursor, putative, expressed (Pfam:PF01357) (- with no introns)	7.103	12.011	6.69E-06	ESTs: LOC_Os06g44470 (20)	GATCTGTTTGTCTTTG (6635)	BAD37571 Dai et al. 2006, 2007
116	OsAffx.15933.1.S1_s_at	LOC_Os06g45150.1 (11)	LOC_Os06g45150 Pollen allergen Ory s 2 precursor, putative, expressed (Pfam:PF01357) (- with no introns)	7.718	11.718	3.02E-07	ESTs: LOC_Os06g45150 (5)	GATCAATGATGTTTCATC (244) GATCGAGTCTGCCTGTG (943)	
140	Os.25407.1.S1_at	LOC_Os04g25160.1 (10); LOC_Os04g26220.1 (2)	[LOC_Os04g25160 ¹ , Pollen allergen Ory s 2-C precursor, putative, expressed (Pfam:PF01357) (- with no introns)], [LOC_Os04g26220 ² , Pollen allergen Ory s 2-C precursor, putative, expressed (Pfam:PF01357) (+ with no introns)]	7.994	11.358	2.11E-08	ESTs: LOC_Os04g25160 (7) LOC_Os04g26220 (4)	¹ GATCACTCCATGTTTCGA (953) ² GATCGCTCCATGTTTCGA (15264)	CAD40508 Dai et al. 2006, 2007 Kerim et al., 2003b

148	Os.22589.1.S1_at	LOC_Os06g45160.1 (11)	LOC_Os06g45160 Pollen allergen Ory s 2 precursor, putative, expressed (Pfam:PF01357) (- with no introns)	7.998	11.303	2.14E-09	ESTs: (7)	GATCAAGGAGAAGGGTG (3203)	
149	Os.12319.1.S1_x_at	LOC_Os04g25150.1 (9); LOC_Os04g26230.1 (11)	[LOC_Os04g25150, Pollen allergen Ory s 2-B precursor, putative, expressed (Pfam:PF01357) (- with no introns)], [LOC_Os04g26230, Pollen allergen Ory s 2-B precursor, putative, expressed (Pfam:PF01357) (+ with no introns)]	9.334	11.303	2.02E-10	ESTs: LOC_Os04g25150 (5) LOC_Os04g26230 (29)	¹ GATCACGCCATGTTGA (5449) ² GATCACGCCACGTTGA (13806) ^{1,2} GATCAAGGAGAAGGGTG (3203)	CAD40509 Kerim et al., 2003b
906	OsAffx.28114.3.A1_at	LOC_Os06g45180.1 (1); LOC_Os06g45190.1 (11)	[LOC_Os06g45180 ¹ , Pollen allergen Ory s 2-A precursor, putative, expressed (Pfam:PF01357) (- with no introns)], [LOC_Os06g45190 ² , Pollen allergen Ory s 2 precursor, putative, expressed (Pfam:PF01357) (- with no introns)]	5.204	6.463	1.06E-05	ESTs: LOC_Os06g45180 (6) LOC_Os06g45190 (3)	¹ GATCAAGTCTGCATTG (6545) ¹ GATCACGCTACGCGTGA (8787) ² GATCAAGTCTGCATTG (2191)	¹ BAD45861 Dai et al. 2006, 2007 Kerim et al., 2003b
1010	OsAffx.28114.2.A1_at	LOC_Os06g45180.1 (3); LOC_Os06g45290.1 (11)	[LOC_Os06g45180 ¹ , Pollen allergen Ory s 2-A precursor, putative, expressed (Pfam:PF01357) (- with no introns)], [LOC_Os06g45290 ² , Pollen allergen Ory s 2-A precursor, putative, expressed (Pfam:PF01357) (- with no introns)]	5.413	6.082	6.24E-08	ESTs: LOC_Os06g45180 (6) LOC_Os06g45290 (24)	¹ GATCAAGTCTGCATTG (6545) ¹ GATCACGCTACGCGTGA (8787) ² GATCATGCTATGCGTGA (10885)	¹ BAD45861 Dai et al. 2006, 2007 Kerim et al., 2003b

5. Additional Potential Group I, II/III Allergens (Further Details)

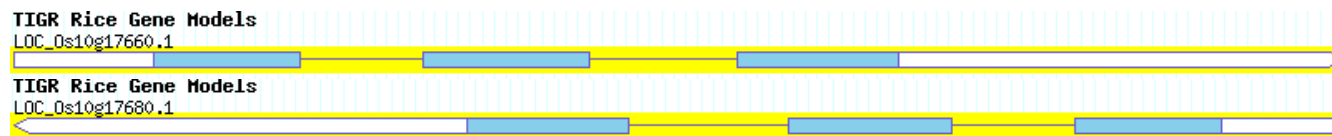
The first additional allergen is LOC_Os10g40090, which is annotated as a putative, expressed β -expansin-1a precursor with a signal peptide at position 24 (GSC-GP), an RlpA-like double-psi beta-barrel (DPBB_1) at positions 80 to 159 and “pollen_allerg_1” motif at positions 172 to 253). This is ranked 3rd among transcripts in the 57K microarray, the 34th most highly upregulated sequence in MPSS and has proteome support. The second is LOC_Os08g44790, which is a putative, expressed α -expansin-3 precursor with signal peptide at position 25 (GDA-AP) and a leading (mostly cleaved) transmembrane domain, a DPBB_1 motif at positions 77 to 164 and “pollen_allerg_1” motif at positions 175 to 253. This is ranked 54th among transcripts (5th in dChip) in the 57K microarray and is the 70th most abundant transcript sequence in MPSS. The sequence has many EST matches, but no principal matches among reported proteome products. The third is LOC_Os12g36040, which is a putative, expressed α -expansin-9 precursor with signal peptide at position 29 (SFA-AD), a DPBB_1 motif at positions 116 to 186 and “pollen_allerg_1” motif at positions 197 to 275. This is ranked 301st among transcripts in the 57K microarray (Supplementary Table S4).

Supplementary Table S4. Pollen expression microarray data on highly pollen-expressed expansins. (Predicted MWs 28.78, 28.99, 30.05 KDa).

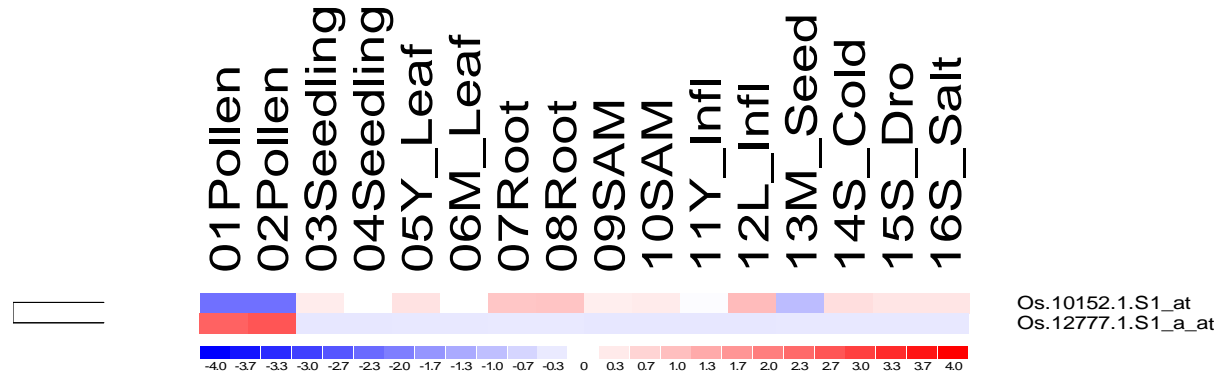
Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (and Pfams)	Intensity (log ₂)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
3.	Os.12697.1.S1_at	Os10g40090.1 (11)	Beta-expansin 1a precursor, putative, expressed (Pfam:PF01357,PF03330) (+ strand with 1 intron)	9.693	13.172	9.62E-10	ESTs: 177	GATCGCTCACCACCAAC (56) GATCAAGAACGTCAACC (18) GATCTCCTGCGGCAACG (243) GATCCTCCTAATTATT(2107)	14a, 15b Dai et al. 2007
54	Os.53024.1.S1_at	Os08g44790.1 (11)	Alpha-expansin 3 precursor, putative, expressed (Pfam:PF01357,PF03330) (+ strand with 2 introns)	7.964	12.214	2.46E-10	ESTs: 18	GATCAGGTACACGATAA (49) GATCATACTAGCTACAA (135)	
301	Os.20191.1.S1_at	Os12g36040.1 (11)	Alpha-expansin 9 precursor, putative, expressed (Pfam:PF01357,PF03330) (+ strand with 1 intron)	8.293	9.908	2.31E-10	ESTs: 7	GATCGATTTTTTCAAC (113)	

6. Profilin A (*Ory s 12*) (Further Details)

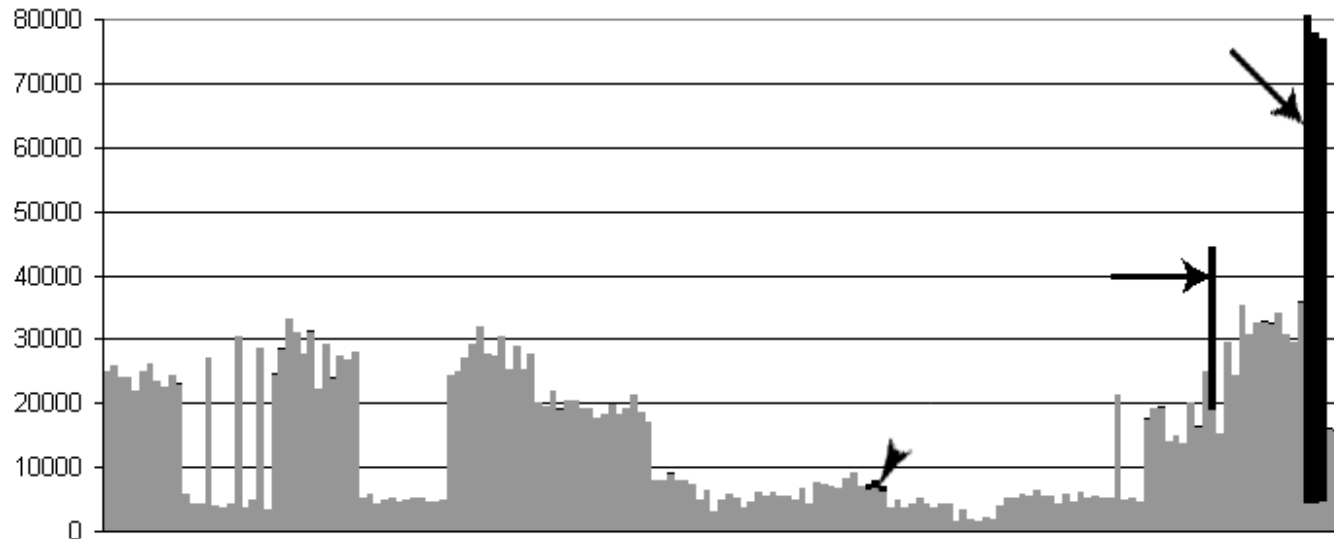
In rice pollen, profilin A transcripts appear to be highly expressed (intensity 8.120 log₂) and upregulated by 11.74 log₂ times (>×2000) above seedlings, according to microarray results. MPSS indicates that transcript abundance is among the top 30 signatures, of which two loci were specific for pollen-expressed profilin. In rice, the two loci active exclusively in pollen, LOC_Os10g17660 and LOC_Os10g17680, are on opposite strands and each of the loci contains a short 5'-UTR (116 bp for LOC_Os10g17660 and 104 bp for LOC_Os10g17680) and a longer 3'-UTR (371 bp for LOC_Os10g17660 and 181 bp for LOC_Os10g17680) with similar overall construction (Supplementary Figure S3) and high sequence similarity. Although six nucleotides differ in their coding sequences, they encode identical polypeptide sequences. Their 1k upstream region displays less conservation than other inverted near duplicates, with the exception of the 0 to -100 presumed promoter region. Strong proteome support exists for translation in pollen grains (Kerim et al., 2003a; Dai et al., 2006) and post-germination pollen tubes (Dai et al., 2007).



Supplementary Figure S3. Diagrammatic representation of TIGR gene models of LOC_Os10g17660 and LOC_Os10g17680 show mirror image organizations and very high sequence similarities. These include extensive similarities in non-coding regulatory sequences.



Supplementary Figure S4. Heat map of proflin (Pfam: PF00235) based on dChip (Li and Wong, 2003; <http://www.biostat.harvard.edu/complab/dchip/>) analysis of CEL data with pollen and seedling data supplemented by GEO data. Genomic data on rice indicate two major types of proflin: two pollen loci of proflin A and a single locus of proflin-2 apparently expressed in all other tissues.



Supplementary Figure S5. Expression of proflin using GEO data from the Affymetrix 57K rice genomic microarray platform. Stacked bars show proflin A in black (probe set Os.12777.1.S1_a_at, representing paired loci LOC_Os10g17660 and LOC_Os10g17680) and proflin-2 in gray (probe set Os.10152.1.S1_at, representing the largely sporophytically-expressed LOC_Os06g05880). Proflin A is detectable in mature inflorescences (arrowhead), and abundantly in collections of anthers at anthesis (horizontal arrow) and especially in pollen (diagonal arrow).

In the rice genome there are only three loci devoted to profilin, of which two seem only to be expressed in pollen (Supplementary Figure S4; Supplementary Table S5). All other sporophytic expression seems to correspond to a remaining profilin-2 (LOC_Os06g05880, Supplementary Figure S5). Profilins are typically involved in the sequestration of actin monomers and are highly expressed in pollen, presumably as a protein that modulates polymerization and the dynamic nature of the elaborate actin cytoskeleton of elongating pollen tubes.

Supplementary Table S5. Pollen expression microarray data on *Ory s 12* (predicted molecular weight 14.25 KDa).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (and Pfams)	Intensity (log ₂)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
113	Os.12777.1.S1_a_at	LOC_Os10g17660.1 (11); LOC_Os10g17680.1 (11)	[LOC_Os10g17660 ¹ , Profilin A, putative, expressed (Pfam: PF00235) (+ strand with 2 introns)], [LOC_Os10g17680 ² , Profilin A, putative, expressed (Pfam:PF00235) (- strand with 2 introns)]	8.120	11.737	7.42E-10	ESTs: LOC_Os10g17660 (16) LOC_Os10g17660 (13)	¹ GATCACCGTCGCCATCA (16) ¹ GATCACCTCTGTACGT (172) ¹ GATCCATTCTTCATG (5760) ² GATCACTCCTGCAGTG (73) ² GATCACATCCATTGTCG (63) ² GATCAGGAGGCATCAC A (5)	^{1,2} NP_920667, 38b Dai et al. 2006, 2007 Q9FUD1 Kirim et al. 2003

7. *Ory s 23* (Further Details)

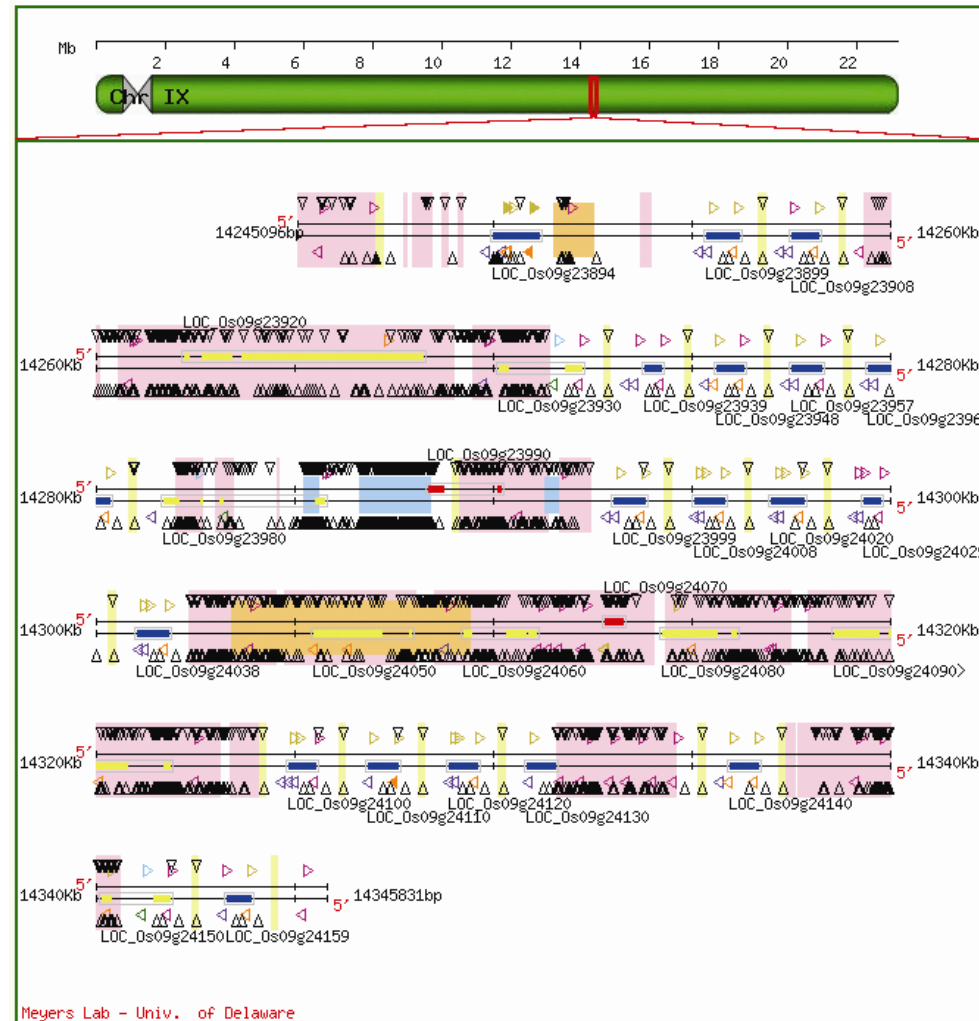
The peptide sequences of *Ory s 23* consist of two distinct forms, differing at positions 48, 79, 81, 82, 128, and 130. The two alternative peptide sequences are highlighted as yellow and light blue in Figure 3. Six single nucleotide polymorphisms are present that code for three single amino acid substitutions in LOC_Os09g24020, LOC_Os09g24100, and LOC_Os09g24130, each of which is annotated as an expressed product. The divergence is narrow (Supplementary Figure S6). All are represented by one probe set which is highly upregulated (Supplementary Table S6).

Interestingly, sequence synonymy also extends to upstream sequences, expanding sometimes to the full 1.4 kbp distance between coding regions of adjacent loci; these typically differ by only a few basepairs despite that this is a non-coding region, with the exception of three upstream sequences that have apparently been displaced. Alignments of coding and non-coding regions are presented in supplementary data. The low sequence variability expressed by *Ory s 23* suggests either a very short history for this gene at its current copy number, or a mechanism that has extraordinarily strict sequence conservation. Given that non-coding upstream and UTR sequences provide an almost complete match, with

The presence of multiple repetitive elements in broad intervals between tandem Ory s 23 sequences, and short repetitive elements between Ory s 23 sequences in non coding regions suggests that this region may be a target of TE insertion. The alternative of DNA repair genes selecting this region as a target remains a possibility but the presence of multiple TE loci between Ory s 23 tandem repeats implicates action by TEs in their origin. The presence of such sequences may represent pseudogenes which may have been inserted by retrotransposable elements. Wang et al. (2006), for example, unexpectedly observed that retroposition was involved in generating large numbers of intronless genes in rice and provided strong evidence for involvement of RTEs as a mechanism of insertion. RTEs, representing a “copy and paste” transposition, can rapidly increase gene copy numbers and plant genome size, as well as providing, over time, a source of paralog formation (Bennetzen, 2007). The rapid and highly related sequences of Ory s 23 noted here, along with LTR-repeat linked variants make RTE insertion an attractive mechanism for amplification of transcription of this sequence pair.

Supplementary Table S6. Pollen expression microarray data on Ory s 23. (Predicted molecular weight 13.99 KDa).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (with Pfams, orientation) (all are on – strand and intronless)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
6.	Os.17958.1.S1_at	LOC_Os09g23899.1 (10); LOC_Os09g23908.1 (3); LOC_Os09g23930.1 (1); LOC_Os09g23939.1 (1); LOC_Os09g23948.1 (7); LOC_Os09g23957.1 (1); LOC_Os09g23966.1 (8); LOC_Os09g23980.1 (1); LOC_Os09g23999.1 (8); LOC_Os09g24008.1 (8); LOC_Os09g24020.1 (5); LOC_Os09g24029.1 (1); LOC_Os09g24038.1 (6); LOC_Os09g24100.1 (1); LOC_Os09g24110.1 (8); LOC_Os09g24120.1 (8); LOC_Os09g24130.1 (1); LOC_Os09g24140.1 (4); LOC_Os09g24150.1 (1); LOC_Os09g24159.1 (1)	[LOC_Os09g23899, LOC_Os09g23939, LOC_Os09g23948, LOC_Os09g24029, Pollen allergen Ory s 23, putative], [LOC_Os09g23908, LOC_Os09g23957, LOC_Os09g23966, LOC_Os09g23999, LOC_Os09g24008, LOC_Os09g24020, LOC_Os09g24038, LOC_Os09g24100, LOC_Os09g24110, LOC_Os09g24120, LOC_Os09g24130, LOC_Os09g24140, LOC_Os09g24159, Pollen allergen Ory s 23, putative, expressed], [LOC_Os09g23930, Retrotransposon protein, putative, unclassified], [LOC_Os09g23980, Retrotransposon protein, putative, unclassified (Pfam:PF03578[x2])], [LOC_Os09g24150, Transposon protein, putative, unclassified]	9.749	13.068	1.87E-07	ESTs: LOC_Os09g23899 (0) LOC_Os09g23908 (3) LOC_Os09g23930 (0) LOC_Os09g23939 (0) LOC_Os09g23948 (0) LOC_Os09g23957 (2) LOC_Os09g23966 (26) LOC_Os09g23980 (0) LOC_Os09g23999 (5) LOC_Os09g24008 (2) LOC_Os09g24020 (5) LOC_Os09g24029 (0) LOC_Os09g24038 (8) LOC_Os09g24100 (2) LOC_Os09g24110 (5) LOC_Os09g24120 (18) LOC_Os09g24130 (10) LOC_Os09g24140 (9) LOC_Os09g24150 (0) LOC_Os09g24159 (8)	GATCCTTGCACTTGGC (30302) (all) GATCCCTGTCCAGTTCA (16697) (all) GATCGGACACGTACCGA (2346) (blue group)	LOC_Os09g24100: BAD36285 Dai et al. 2006 LOC_Os09g24120: 59b, 60b Dai et al. 2006, 2007 LOC_Os09g24130: BAD36288 Dai et al. 2006



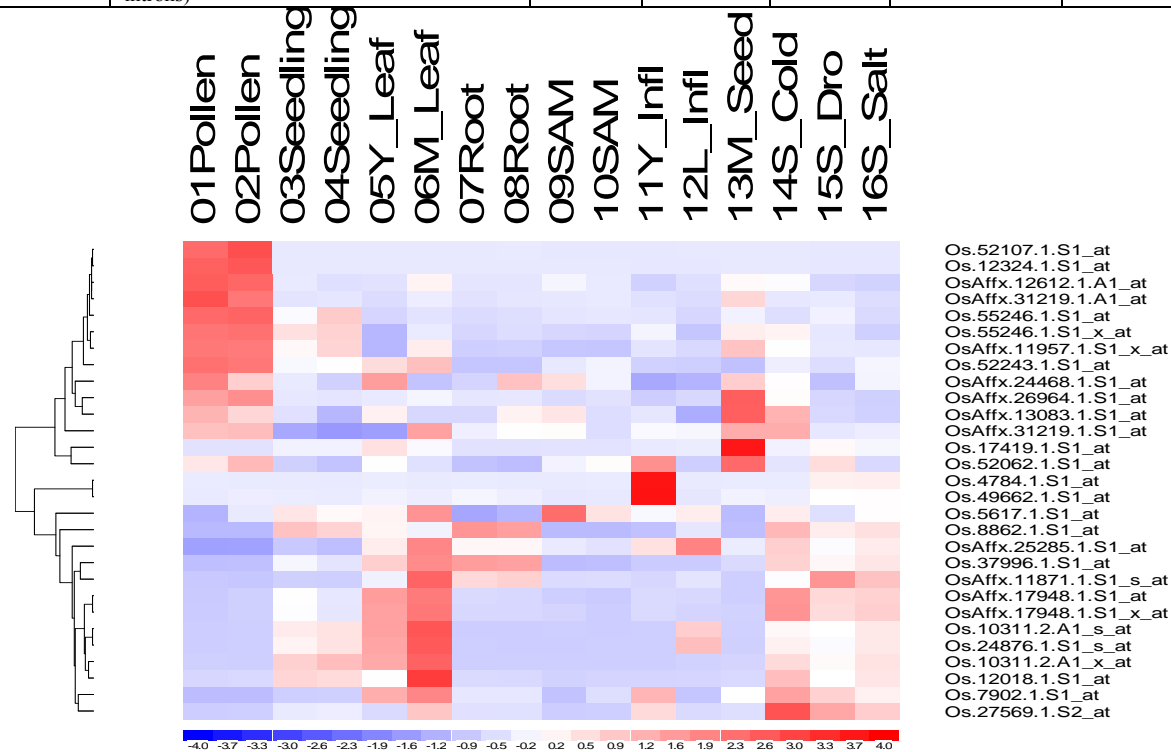
Supplementary Figure S7. Data from Rice Massively Parallel Signature Sequencing (MPSS) and TIGR sequences (Nobuta et al., 2007), showing segment of chromosome 9 containing all 20 copies of the *Ory s 23* sequence. Blue regions on the Crick strand represent *Ory s 23* loci. Note that *Ory s 23* sequences are interspersed with equally spaced light yellow repetitive sequences. Transposable elements are indicated in yellow and orange shaded regions. Upward and downward pointing arrows represent small RNA signatures. (Image downloaded and modified from Meyers Lab site, <http://www.mpss.udel.edu/rice/>).

8. Other Suspected Allergens (Further Details)

a. EF hand family protein Ory s 7

Supplementary Table S7. Pollen expression microarray data on EF hand family protein Ory s 7 (predicted MWs 31.44, 20.09, 18.58 KDa, respectively).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (and Pfams)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
21	Os.12324.1.S1_at	LOC_Os08g44660.1 (4)	EF hand family protein, expressed (Pfam:PF00036 [x2]) (+ strand with 2 introns)	7.894	12.526	2.53E-08	ESTs: 0	GATCGACACCGACGGCG (152)	
331	Os.52107.1.S1_at	LOC_Os12g12730.1 (11)	EF hand family protein, expressed (Pfam:PF00036 [x4]) (- strand and no introns)	7.538	9.671	7.42E-08	ESTs: 2	GATCAGCATGGTGGACG (327)	
4597	Os.52243.1.S1_at	LOC_Os04g45180.1 (11)	EF hand family protein, expressed (Pfam:PF00036, PF02809) (+ strand w/5 introns)	8.664	2.087	6.57E-06	ESTs: 15		

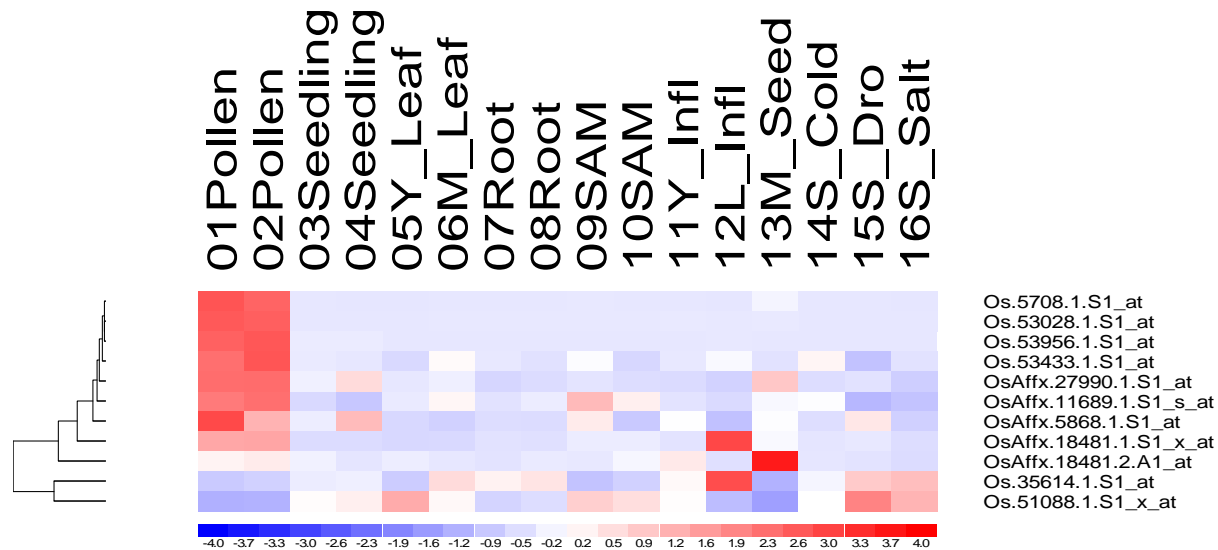


Supplementary Figure S8. Heat map of EF hand-related motifs (Pfam: PF00036) generated based on dChIP analysis supplemented by GEO data.

b. Pectate lyase

Supplementary Table S8. Pollen expression microarray data on Pectate lyase (predicted molecular weight 49.23, 50.32, 50.22, 39.07, 50.27, 50.27 KDa, respectively).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (and Pfams)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
85	Os.18224.1.S1_at	LOC_Os06g38510.1 (11)	Pectate lyase precursor, putative, expressed (Pfam: PF00544, PF04431) (+ strand with 2 introns)	8.135	11.923	2.51E-08	ESTs: 87	GATCATCAGCCAGGGGA (2284) GATCACGAAAGAGCCAA (1211)	
209	Os.53956.1.S1_at	LOC_Os02g12300.1 (11)	Pectate lyase precursor, putative, expressed (Pfam: PF00544, PF04431) (+ strand with 3 introns)	8.336	10.754	6.87E-11	ESTs: 55	GATACCAATGACCAAA (76) GATCCGCCGCGAGATG (14) GATCACCGTGCGTTCA (32)	
219	Os.53028.1.S1_at	LOC_Os06g05260.1 (11)	Pectate lyase precursor, putative, expressed (PF00544, PF04431) (+ strand with 1 intron)	7.870	10.673	6.86E-09	ESTs: 31	GATCATCAGCCAGGGGA (2284)	
230	Os.5708.1.S1_at	LOC_Os06g05209.1 (11); LOC_Os06g05272.1 (9)	[LOC_Os06g05209, Pectate lyase precursor, putative, expressed (Pfam: PF00544, PF04431)] (- with 1 intron), [Os06g05272, Pectate lyase precursor, putative (Pfam: PF00544)] (+ strand with 1 intron)	7.375	10.536	2.01E-07	ESTs: Os06g05209 (14); Os06g05272 (0)	¹ GATCCGCGACTCCAAGC (20) ¹ GATCGTGAACCACAACA (29) ² GATCCGCGACTCCAAGC (16) ² GATCACCGTGCGTTCA (32) ^{1,2} GATCATCAGCCAGGGGA (2284)	

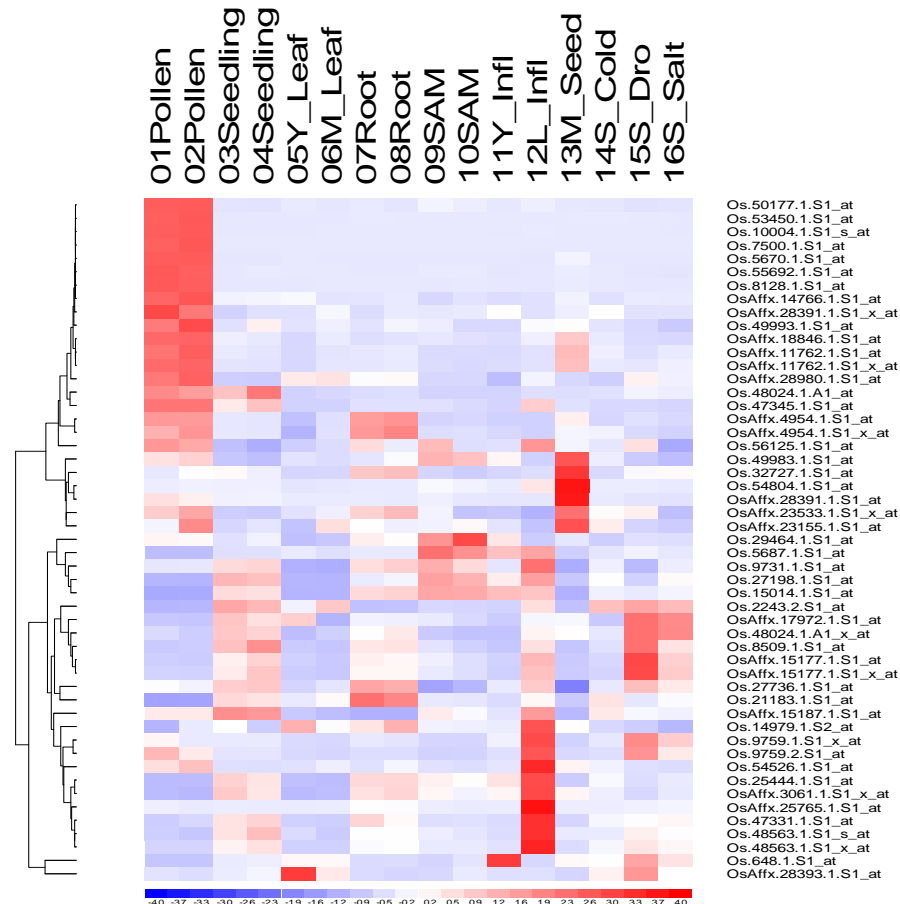


Supplementary Figure S9. Heat map of pectate lyase-related motifs (Pfam: PF00544, PF04431) based on dChip analysis supplemented by GEO data.

c. Glycoside hydrolase family 28: Exopolygalacturonase

Supplementary Table S9. Pollen expression microarray data on Glycoside hydrolase family 28, Exopolygalacturonase (predicted MWs 43.00, 46.34, 35.12, 43.12, 43.12. 44.61, 43.50 KDa, respectively).

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (and Pfams)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
35	Os.7500.1.S1_at	LOC_Os06g35300.1 (2); LOC_Os06g35320.1 (11); LOC_Os06g35370.1 (4)	[LOC_Os06g35300 ¹ , Exopolygalacturonase precursor, putative (Pfam:PF00295)] (- w/1 intron), [LOC_Os06g35320 ² , Exopolygalacturonase precursor, putative, expressed (Pfam:PF00295)] (- w/1 intron), [LOC_Os06g35370 ³ , Exopolygalacturonase precursor, putative, expressed (Pfam:PF00295)] (- w/1 intron)	8.672	12.387	3.21E-09	ESTs: LOC_Os06g35300 (0) LOC_Os06g35320 (27) LOC_Os06g35370 (93)	^{2,3} GATCTGCACTGCCAACG (59) ^{2,3} GATCAACCCACGCTACA (4351) ^{2,3} GATCAAGTCCTATGAGG (40)	1b, 2b Dai et al., 2007
63	Os.10004.1.S1_s_at	LOC_Os02g10300.1 (11)	Exopolygalacturonase precursor, putative, expressed (Pfam:PF00295) (+ w/2 introns)	8.431	12.099	7.03E-08	ESTs: LOC_Os02g10300 (9)	GATCGACAAGGTGACCA (43) GATCAAGTCGTACGAGG (6) GATCATCATCGACCAGA (4194)	XP_464471 13a, 14a Dai et al., 2006, 2007
121	Os.8128.1.S1_at	LOC_Os01g33300.1 (11)	Exopolygalacturonase precursor, putative, expressed (Pfam:PF00295) (- w/2 introns)	7.986	11.691	2.98E-09	ESTs: LOC_Os01g33300 (10)	GATCCATCCATCGTCGA (771) GATCAGCGTGGGGTGCC (11) GATCCCAAGGGAGAAT (50)	
125	Os.53450.1.S1_at	LOC_Os06g40890.1 (11)	Exopolygalacturonase precursor, putative, expressed (Pfam:PF00295) (+ w/1 intron)	8.577	11.557	1.02E-08	ESTs: LOC_Os06g40890 (21)	GATCAGCGCGTGACCA(20) GATCACCATCGCCGCCA (12) GATCAAGTCGTACGAGG (6)	
800	Os.5670.1.S1_at	LOC_Os08g23790.1 (11)	Exopolygalacturonase precursor, putative, expressed (Pfam:PF00295) (+ w/4 introns)	5.775	6.822	3.13E-06	ESTs: LOC_Os08g23790 (77)	GATCGTTAATTTTTTGC (433)	BAD03446 Dai et al., 2006

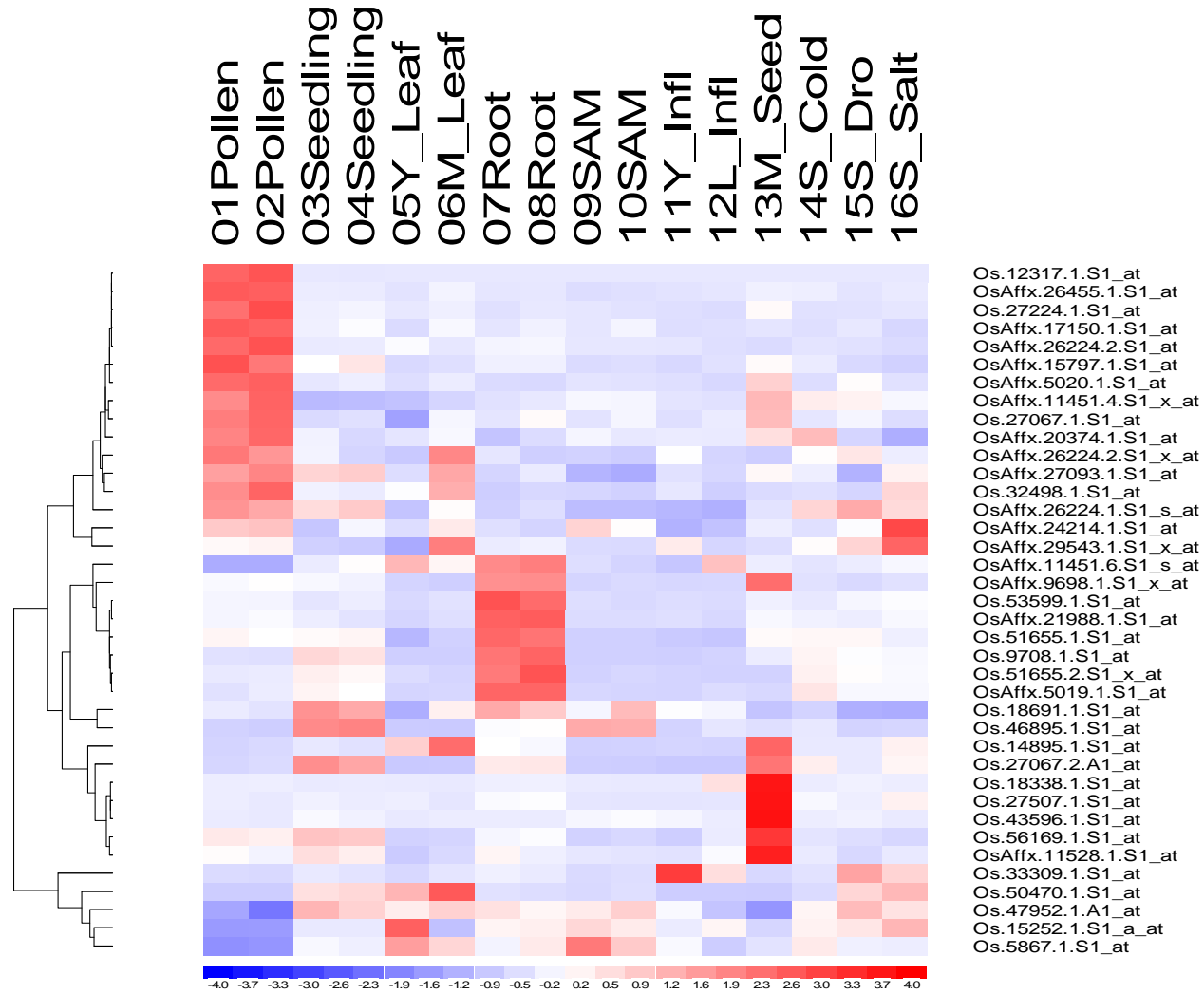


Supplementary Figure S10. Heat map of glycosyl hydrolase family 28 proteins (PF00295), most abundantly represented in rice by exopolysaccharidase-related motifs. Image based on dChip (Li and Wong, 2003; <http://www.biostat.harvard.edu/complab/dchip/>) analysis of CEL data with pollen and seedling data supplemented by GEO data.

d. FAD-binding motif (PF01565)

Supplementary Table S10. Pollen expression microarray data on FAD-binding motif (PF01565) (predicted molecular weight 58.43 KDa),

Rank	Probe Set ID	TIGR v5 locus (probe match)	Description (and Pfams)	Intensity (log2)	Log ₂ upreg >seedling	p datasets	Experimental confirmation		
							Gene	MPSS (TPM)	Proteome
33	Os.12317.1.S1_at	LOC_Os06g35590.1 (11)	Reticuline oxidase precursor, putative, expressed (Pfam:PF01565,PF08031) (-strand with no introns)	8.508	12.414	1.94E-08	ESTs LOC_Os06g35590 (12)	GATCAACCAACTAA CGT (804)	BAD54133 Dai et al., 2006



Supplementary Figure S11. Heat map of FAD-binding motif (PF01565), based on dChip (Li and Wong, 2003; <http://www.biostat.harvard.edu/complab/dchip/>) analysis of CEL data with pollen and seedling data supplemented by GEO data.

REFERENCES (ONLY APPEARING IN THIS DATA SUPPLEMENT)

- Bate, N., and Twell, D.** (1998). Functional architecture of a late pollen promoter: Pollen-specific transcription is developmentally regulated by multiple stage-specific and co-dependent activator elements. *Plant Mol. Biol.* **37**, 885-896.
- Rogers, H.J., Bate, N., Combe, J., Sullivan, J., Sweetman, J., Swan, C., Lonsdale, D.M., and Twell, D.** (2001). Functional analysis of cis-regulatory elements within the promoter of the tobacco late pollen gene g10. *Plant Mol. Biol.* **45**, 577-585.