

Transcriptome-Based Examination of Putative Pollen Allergens of Rice (*Oryza sativa* ssp. *japonica*)

Scott D. Russell^a, Prem L. Bhalla^b and Mohan B. Singh^b

^a Department of Botany and Microbiology, University of Oklahoma, Norman, OK 73019, USA

^b Plant Molecular Biology and Biotechnology Laboratory, Australian Research Council Centre of Excellence for Integrative Legume Research, Faculty of Land and Food Resources, The University of Melbourne, Parkville, Vic. 3010, Australia

ABSTRACT Pollen allergens are among the most abundantly transcribed and translated products in the life history of plants, and particularly grasses. To identify different pollen allergens in rice, putative allergens were identified in the rice genome and their expression characterized using the Affymetrix 57K rice GeneChip microarray. Among the most abundant pollen-specific candidate transcripts were *Ory s 1* beta-expansin, *Ory s 2*, *Ory s 7* EF hand, *Ory s 11*, *Ory s 12* profilin A, *Ory s 23*, glycosyl hydrolase family 28 (polygalacturonase), and FAD binding proteins. Highly expressed pollen proteins are frequently present in multiple copy numbers, sometimes with mirror images located on nearby regions of the opposite DNA strand. Many of these are intronless and inserted as copies that retain nearly exact copies of their regulatory elements. *Ory s 23* reflects low variability and high copy number, suggesting recent gene amplification. Some copies contain pseudogenes, which may reflect their origin through activity of retrotransposition; some putative allergenic sequences bear fusion products with repeat sequences of transposable elements (LTRs). The abundance of nearby repetitive sequences, activation of transposable elements, and high production of mRNA transcripts appear to coincide in pollen and may contribute to a syndrome in which highly transcribed proteins may be copied and inserted with streamlined features for translation, including grouping and removal of introns.

INTRODUCTION

Pollen allergens represent a major human health challenge that is particularly pernicious among wind-pollinated seed plants and particularly the grasses (Mohapatra et al., 2005). Often, these proteins are among the most abundantly produced proteins in their life history. These proteins may play a biological role in facilitating and enabling the elongation of the pollen tube, which may occur through controlling the internal metabolism of the tube and modifying the immediate environment of the tube. In anemophilous plants such as rice, pollen is at least partially desiccated and is protected by a thick layer of proteins and a lipid coat which together provide a barrier effect. Both proteins and lipids in rice may be implicated in allergic response to pollen (Asero et al., 2007; Sen et al., 2003). Not surprisingly, genes that produce these proteins are frequently present in multiple copies, and are frequently closely situated to one another. Grasses have evolved a group of allergens that appear to be largely restricted to species of Poaceae. In maize, the recent proliferation of Group 1 pollen allergens has been well documented, suggesting a rapid diversification of cell wall modifying genes that also occupy a role in human allergy (Valdivia et al., 2007). A similar pattern of expression

was proposed using the rice genome to identify putative pollen allergens with these domains and other motifs (Jiang et al., 2005).

In the rice genome, pollen allergens represent conspicuous and highly transcribed groups of genes with conserved sequences. In order to further understand their relationship and diversification, an investigation was conducted on their genomic sequence and placement, as well as the nature of conservation of coding and regulatory sequences. Given the known doubling of chromosomes in rice and the age of this event (Wang et al., 2007), conservation of pollen allergens could be reflected in copy numbers, positioning and diversification of function of expressed proteins. In general, the pollen transcriptome has been found to be the most highly divergent of flowering plant life history according to microarrays of

¹ To whom correspondence should be addressed. E-mail srussell@ou.edu, fax +1-405-325-7619, tel. +1-405-325-7619.

Arabidopsis (Becker and Feijo, 2007 and references therein) and MPSS (massively parallel signature sequencing of transcripts, Nobuta et al., 2007). Proteomic data indicate that pollen transcripts are translated at high expression levels at anthesis in *Arabidopsis* (Holmes-Davis et al., 2005; Noir et al., 2005; Sheoran et al., 2006), tomato (Sheoran et al., 2007) and in rice (Dai et al., 2006), and that dynamic changes occur during pollen germination and initial elongation of the tube (Dai et al., 2007). We undertook this examination of pollen expression in rice employing the Affymetrix 57K rice genome chip to assess whether suspected allergens indeed express a pollen enhanced profile. Genomic features will then be examined that may contribute to explaining their high level of expression.

RESULTS AND DISCUSSION

Pollen allergens of *Oryza sativa* recognized by the International Union of Immunological Societies (IUIS) official list of allergens (www.allergen.org/) include Ory s 1, which represents a beta-expansin gene; Ory s 7, which represents an EF-hand protein; and Ory s 12, which represents profilin A. Non-IUIS Ory s 2, Ory s 11, and Ory s 23 are largely designated from rice genome annotations of homologues of Phl p 2, Phl p 11, and Cyn d 23. In a survey of the rice genome for potential allergens, Jiang et al. (2005) used homologues of three general pollen allergen motifs: Ole e 1, expansins and ribonucleases, but their occurrence was not validated for pollen expression. There are additional candidate protein families that were not considered, as well. In the current study, Affymetrix rice genomic microarray tests revealed the most highly expressed and up-regulated pollen genes bearing similarities with grass allergens. These findings were confirmed by MPSS signatures of anthesis rice pollen (dataset NPO at www.mpps.udel.edu/rice/), expressed sequence tags (ESTs, at TIGR and NCBI), and proteome data from developing and mature rice anthers (Kerim et al., 2003a, 2003b; Imin et al., 2004; Dai et al., 2006, 2007).

Canonical pollen allergens are characteristically highly expressed in pollen, often largely restricted to gametophytic expression, and involved in unique attributes of pollen physiology. Radauer and Breiteneder (2006) examined pollen allergens known to elicit allergenic response and categorized them into prevalent protein groups, which we then screened to identify candidates based on high expression and up-regulation in pollen.

In the only other genomic survey of rice pollen allergens, Jiang et al. (2005) identified, as putative pollen putative allergens, 45 proteins with 'Ole e 1-like motifs', 65 proteins with Pfam motifs designated as 'Protein_allerg_1' (expansin motif), and seven proteins with Pfam motifs designated as 'Protein_allerg_2' (ribonuclease motif). Once sequences were assigned to TIGR and IRGSP loci, transcriptional expression was determined for each candidate using the Affymetrix chip, MPSS and proteome data. Of Jiang et al. (2005) putative allergens, 10 of the 117 proteins were highly expressed in pollen. Two

candidates of the so-called OsOle series showed pollen expression (OsOle24 and OsOle28), eight candidates of the OsAllerg1 group showed high activity (3 Ory s 1 proteins, 3 Ory s 2 proteins and two expansin candidates), and none of the seven OsAllerg2 group candidates showed high up-regulation in pollen (see supplementary spreadsheet). Specific candidates are further characterized in the supplementary data, and grouped below into major allergen groups.

Pollen Major Group 1 Allergens

Related to the β -expansins, these proteins account for most of the major group I allergens of grasses (Sampedro and Cosgrove, 2005), with which they share high sequence similarity (Jiang et al., 2005). In rice, pollen proteins containing this characteristic protein motif include Ory s 1, which has been validated as an allergen on the basis of its recognition by IgE antibodies from allergic individuals (Xu et al., 1995), and Ory s 2, which represents a smaller protein containing a similar motif. Extensive supplementary data complement this and other topics in the results.

Ory s 1

The Ory s 1 family of pollen allergens is closely related to the β -expansins, major group I allergens of grasses, such as Zea m 1; β -expansin shares high similarity in a number of motifs and its genomic representatives reflect a dynamic history (Sampedro et al., 2005). The sequence of Ory s 1 is consistent with a group I allergen, encoding a putatively secreted water-soluble protein with allergenic properties. Ory s 1-encoded protein sequences contain a signal peptide at position 22 (VSC-GP), a 'rare lipoprotein A-like double-psi beta-barrel' motif (PF03330) at positions 78–157 and a pollen_allerg_I motif (PF01357) from position 170 to 251.

Ory s 1 bears closest sequence relationship in the grasses to Phl p 1 of *Poa*, sharing 74% identity, 84% positives and $p < 1e^{-129}$. In *Oryza sativa ssp. japonica*, Ory s 1 is represented by four gene loci that are in tandem association near the origin of chromosome 3 (Figure 1). The loci range from LOC_Os03g01610 to LOC_Os03g01650 and consist of a coding region that is typically 804 bp, encoding a putative protein of 267 amino acids. Two loci are identical in the coding region (LOC_Os03g01610 and LOC_Os03g01650), but interestingly are encoded in opposite orientations on chromosome 3; LOC_Os03g01610 is on the (+) strand (as are the other two loci) and LOC_Os03g01650 on the (–) strand (Figure 1). They share a similar promoter motif (Xu et al., 1999).

Ory s 2

Ory s 2 is a Group II/III pollen allergen of 117 amino acid length that has a core 'pollen_allerg_I' motif (Pfam: PF01357) at positions 26–102. A signal peptide at position 23 (most often SCATE) is consistent with this gene locus, representing a secreted protein product. Similarly to Ory s 1, the role of Ory s 2 is expected to involve cell wall modification, sharing one major motif held in common with expansins. Unlike Ory s 1 and

canonical expansins, however, Ory s 2 lacks a RlpA-like DPBB_1 motif (Pfam: PF03330). Available data suggest that Ory s 2 is highly transcribed, with eight probe sets of the 57K rice genomic microarray report up-regulation $>6 \log_2$ in pollen—and six of these higher than $11 \log_2$ (>2000 -fold). The high degree of up-regulation of the transcripts of Ory s 2 is even more strongly supported by MPSS, which reports significant Ory s 2 transcription for 10 signature sequences, of which seven are in the top 25 for pollen intensity. Abundance of translated products is supported by proteome data (Dai et al., 2006, 2007). According to Kerim et al. (2003b), Ory s 2 has multiple isoforms displaying some amino acid differences and therefore they named them as three different forms, largely on the basis of their N-terminus.

The genomic arrangement of the Ory s 2 family suggests that its loci natively lack introns. The intron-bearing loci ap-

pear to represent altered copies that may be functionless pseudogenes (see LOC_Os04g26190 and LOC_Os06g45200, supplementary data). 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 being a common pattern (Figure 2). Chromosome 4 repeats the theme of sharing strong sequence similarity with inverted loci on the opposite strand. Additional data are available as supplementary data.

Additional Potential Group I, II/III Allergens

In addition to Ory s 1 and Ory s 2, there are three additional very highly up-regulated expansin-related proteins in pollen. Each has an expansin motif, as well as a rare lipoprotein A (RlpA)-like double-psi beta-barrel motif (PF03330), as in Ory s 1. One includes LOC_Os10g40090.1, which is a Beta-expansin

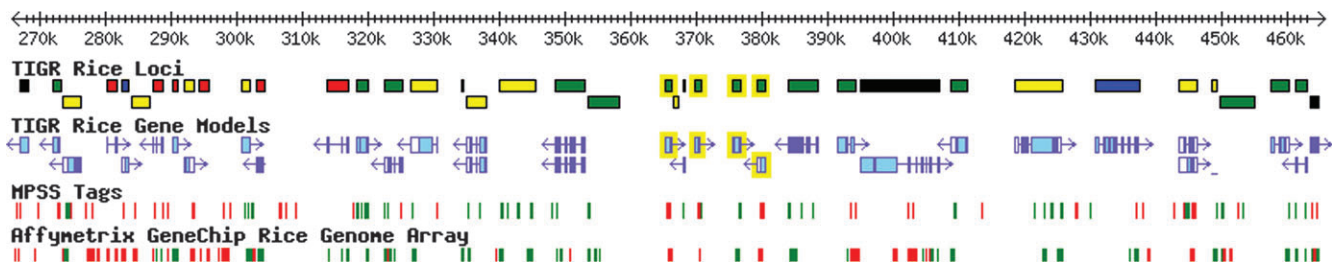


Figure 1. Genomic Organization of Ory s 1 Loci on Chromosome 3.

Ory s 1 loci (green, highlighted in yellow) are, from left: LOC_Os03g01610, LOC_Os03g01630, LOC_Os03g01640, and LOC_Os03g01650. TIGR Rice Gene Loci line shows gene placement, with orientation indicated in line labeled TIGR Rice Gene Models. The high degree of sequence overlap is reflected in overlapping MPSS signatures and Affymetrix GeneChip probe sets, indicated in red, contrasted with unique sequences in green. (Graphics modified from TIGR Rice Genome Browser.)

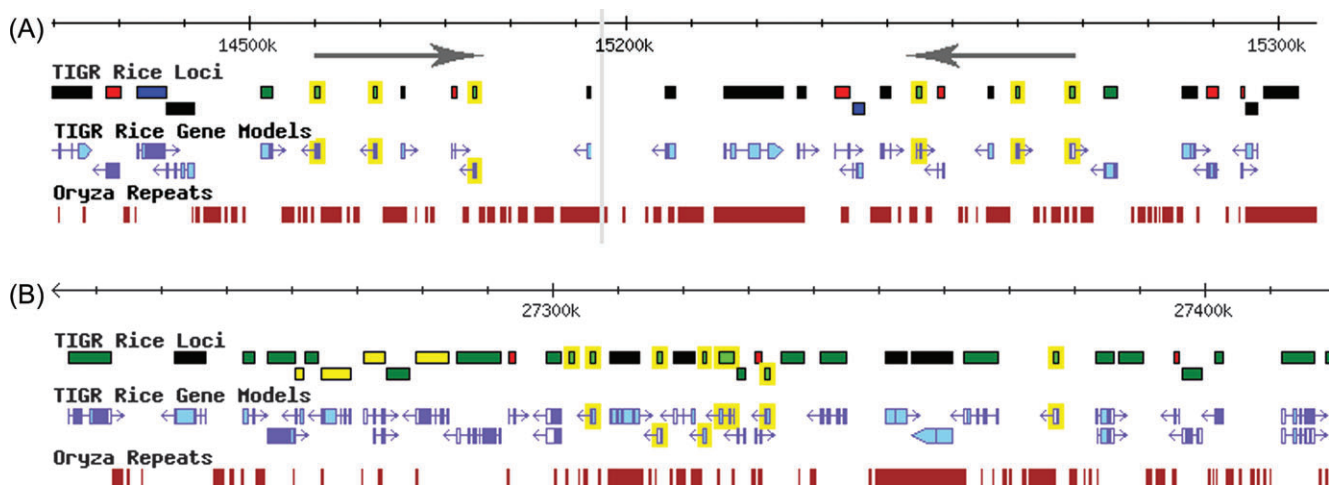


Figure 2. Genomic Organization of Ory s 2 Loci (Green, Highlighted in Yellow) on Chromosome 4.

(A) Reflects mirror images of loci (left: LOC_Os04g25150, LOC_Os04g25160, LOC_Os04g25190, and right: LOC_Os04g26190, LOC_Os04g26220, and LOC_Os04g26230).

(B) Chromosome 6 loci have a tight distribution of related sequences on (-) strand. Oryza repeats illustrate repeat elements, including LTRs, flanking the coding regions of Ory s 2 loci. (Graphics modified from TIGR Rice Genome Browser.)

1a precursor with the third highest abundance in the Affymetrix microarray, 177 ESTs, 34th MPSS abundance and strong proteome support (see supplementary data). Two others (LOC_Os08g44790 and LOC_Os12g36040) are also highly up-regulated and apparently pollen-specific. These are obvious candidates but have not yet been tested for allergenic activity.

Profilin A (Ory s 12)

Ory s 12 is the pollen allergen corresponding to pollen-expressed profilin, which has long been known as a potent allergen (Valenta et al., 1991). Since the known functional role of profilin—principally as a protein involved with controlling actin polymerization—is exclusively intracellular, the release of profilin from pollen as an allergen is likely the result of protein released from damaged, non-intact or dead pollen, possibly forming small aerosol particles that may be inhaled (Swoboda et al., 2004). In rice, pollen-related profilin A transcripts appear to be highly expressed (intensity 8.120 log₂) and up-regulated 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; there was strong proteome support (supplementary data). In rice, the two loci active exclusively in pollen are LOC_Os10g17660 and LOC_Os10g17680. These are inverted on opposite strands and share high sequence similarity, differing by only six nucleotides. In the rice genome, there are only three loci devoted to profilin, of which two seem only to be expressed in pollen (see supplementary data). All other sporophytic expression seems to correspond to a remaining profilin-2 (LOC_Os06g05880). 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. These appear to be strong candidates.

Ory s 23

Ory s 23 was the sixth most abundant transcript according to the 57K microarray, second according to MPSS (mpss.udel.edu/rice/, dataset NPO) with a fold up-regulation of between ×1000 (MPSS) and ×8000 (57K chip). These transcripts are translated and abundant, occurring as water-soluble secreted proteins according to proteome evidence from mature pollen and in-vitro germinated pollen (Dai et al., 2006, 2007). Surely this degree of transcription and translation would suggest an important role in facilitating rice pollen function, but its role is as yet undetermined. Ory s 23 has as its closest homologue Cyn d 23, which was first described in *Cynodon dactylon* as a 112 amino acid pollen allergen peptide (AAP80170.1). The similarity between this gene and the version in *Cynodon dactylon* is 37% amino acid identity with 55% positives and two significant gaps, $p < 5e^{-14}$ and therefore there may be significant differences in function of the protein, its potential substrate target or both. While allergenic response is suggested by its possessing an IUIS designation (www.allergen.org/), the med-

ical possibilities of this potential allergen protein are not currently characterized in rice. If this pollen product is similar to other described secreted pollen allergens, as a secreted product, it might be expected to have enzymatic, possibly cell wall-altering, activity, which may function in facilitating pollen tube passage within the gynoecium (Valdivia et al., 2007) or some other highly up-regulated pollen function. These appear to be strong candidates.

The Orys23 gene family in rice consists of 17 very closely related loci, sharing high sequence synonymy and separation by less than 200 kbp, with all copies of the gene located within the interval from TIGR loci LOC_Os09g23899 to LOC_Os09g24150. All are on the Crick (–) strand and 13 of these bear UTR sequences and strong EST support (<http://rice.tigr.org/>). The coding region of each locus consists of an intronless 405 nucleotide sequence encoding a 134 amino acid protein with signal peptide at position 16 (ASA-VL). These peptide sequences have 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.

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 the fact that this is a non-coding-region, with the exception of three upstream sequences that have apparently been displaced (see supplementary data). The low sequence variability expressed by Ory s 23 suggests either a very short evolutionary 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 relatively little variability except along two motifs, the suggestion is that this has a short history and that a faithful copy mechanism that can capture non-coding regions seems a likely explanation. The grouping of different sequence motifs as well as the distance between coding sequences suggests that multiple loci may have been inserted at the same time, potentially in pairs.

Transposable element (TE) motifs are linked to three additional nearly complete Ory s 23 sequences. One open reading frame has a long leading retrotransposon sequence that constitutes a properly terminated Ory s 23 sequence at LOC_Os09g23980. Two shorter, trailing TE motifs that appear anomalously terminated are connected to reading frames at LOC_Os09g24150, with a 14-bp overlapping transposon motif, and LOC_Os09g23930, with a 17-bp overlapping retrotransposon motif at the end of an otherwise normal Ory s 23 sequence. The anomalous flanking sequences have high homology with TE sequences located on other chromosomes, suggesting multiple independent insertions of the TE sequences, which may have potentially arisen as an inadvertent consequence of TE activity itself. The presence of multiple repetitive elements in broad intervals between tandem Ory s 23 sequences and

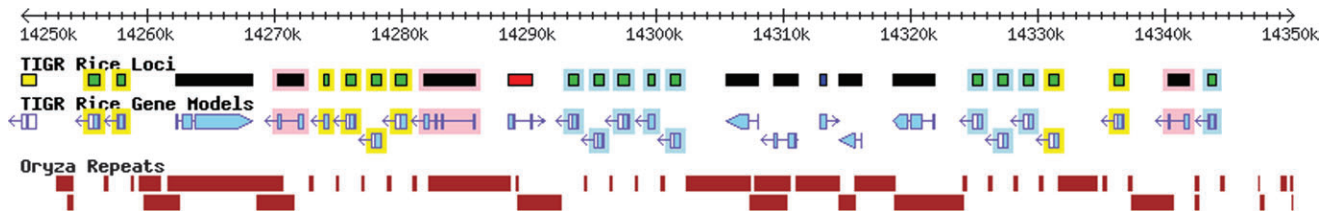


Figure 3. Genomic Organization of *Ory s* 23 Loci.

Rectangles with green interiors display two variants (highlighted in light blue and yellow) that differ by six amino acids on chromosome 9; leftmost *Ory s* 23 locus is LOC_Os09g23899; rightmost is LOC_Os09g24159. *Ory s* 23 loci are intronless, separated by narrow repetitive intervals of typically ≈ 1400 bp—a feature of the *Ory s* 23 gene family. Transposable element (TE) motifs (highlighted in pink) form three nearly complete *Ory s* 23 sequences with leading or trailing TE motifs. *Oryza* repeats indicate repeat elements, including LTRs, flanking the coding regions. (Graphic modified from TIGR Rice Genome Browser.)

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 that 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 (see supplementary data).

Other Putative Allergens

Among the 2615 protein families of angiosperms, Radauer and Breiteneder (2006) found that allergenic pollen proteins occur in 29 of these, eight of which are common in Poaceae; half of these groups are covered above: expansin-related products, ‘Ole e 1-like’ proteins, and profilin. Among these, the remaining groups include EF hand, glycosyl hydrolase family 28, and FAD binding.

The EF-hand protein family is responsible for the rice allergen known as *Ory s* 7 and is highly up-regulated and expressed in rice pollen at two loci: LOC_Os08g44660, with an intensity of $7.894 \log_2$, up-regulated $12.526 \log_2$ -fold ($\sim 4000\times$), making it one of the most highly up-regulated sequences, and LOC_Os12g12730, with an intensity of $7.538 \log_2$, up-regulated $9.671 \log_2$ -fold ($\sim 700\times$). A third product (LOC_Os04g45180) is highly expressed, but only about four times more than seedling, suggesting wide sporophytic expression as well.

Pectate lyase is highly up-regulated ($>10 \log_2$ -fold) and expressed ($>7 \log_2$) in pollen at five loci on the rice genome in the 57K chip, with one locus also highly expressed in MPSS (LOC_Os06g05260), but no proteome support. The five pollen-

expressed loci share sequences that distinguish them from the predominantly sporophytically expressed loci; four are located on chromosome 6, one on chromosome 2. The most closely related are LOC_Os06g05209 and LOC_Os06g05272, which are on (–) and (+) strands, respectively. These appear to be weak candidates (supplementary data).

Glycosyl hydrolase family 28 proteins are most abundantly represented in rice pollen by the exopolygalacturonases, which have well documented representatives in the grasses that elicit an allergic response (Swoboda et al., 2004). Of 68 loci that have been annotated as polygalacturonases in the rice genome (TIGR, v5), there were eight with ESTs that displayed $>6 \log_2$ higher expression than in seedlings, and expression is absent in other tissues (GEOS data). Three are most likely to be pollen allergen candidates: LOC_Os06g35320, LOC_Os06g35370, and LOC_Os02g10300, each of which is highly supported by microarray, MPSS and proteome (supplementary data). Each has a signal peptide and is expected to represent a secreted protein that may express wall-modifying properties that may alter PT walls or their penetration into the gynoecium. All polygalacturonases expressed highly in pollen seem to have introns, but there seems to be an inverse relationship between the numbers of introns and pollen expression (supplementary data).

FAD-binding motif is present in only one representative, annotated as a reticuline oxide precursor, an oxidoreductase gene typically involved in secondary metabolism (similar to tetrahydroprotoberberine synthase, At5g44410, $1.00e^{-137}$); the predicted protein has a signal peptide at position 29 and is likely secreted. With a \log_2 intensity of 8.508 and very high up-regulation over seedling controls ($12.414 \log_2$), the expression profile is similar to that of pollen allergens, and has MPSS and proteome support.

Evaluation of Proteins for Potential Allergenic Activity

Clearly, clinical testing is needed to determine the medical relevance of each of these protein candidates to determine their allergenic potential in humans. Skin testing, radioallergosorbent test (RAST), immunoblotting and allergen-induced lymphoproliferation have been used to assess asthmatic children and revealed three major allergen molecular weight classes of approximately 16, 26, and 32 kDa, as well as other

minor allergens not described in detail (Tsai et al., 1990). These molecular weights correspond to a number of candidates; for example, Ory s 1 is undoubtedly represented by the 32-kDa protein (Ory s 1 has a reported MW of 30 kDa, according to Kerim et al., 2003a, 2003b, and 34 kDa according to Xu et al., 1995). The 26-kDa line likely also represents a form of expansin, presumably one of the top two candidates under 'additional potential Group I, II/III allergens'. The 16-kDa line may represent a composite of Ory s 2, which has been measured at 12–13 kDa (Kerim et al., 2003b), profilin Ory s 12 measured at ~14 kDa and Ory s 23, predicted at 13.99 kDa. The others may represent minor allergens that could become problematic with chronic exposure, as with most pollens. Rice pollen allergy has remained relatively little studied compared to other grasses. Population centers are typically remote to rice paddies and have more conspicuous allergenic threats. Another contributing factor may be that rice pollen is extremely short-lived. *Oryza sativa* pollen viability is reduced below 50% within 6 min after leaving the anther and essentially all are non-viable in a mere 20 min (Song et al., 2001).

Expression and Potential Amplification of Putative Pollen Allergens

Pollen allergens are among the most highly expressed among any cell or tissue type in rice, independently of the method of normalization used (in this study, Affymetrix GCOS, GCRMA, dChip, RMA, and PLM yielded similarly high intensities compared to seedling controls and GEO datasets). That pollen expresses a highly divergent and abundant complement of pollen transcripts is supported by MPSS (Nobuta et al., 2007; www.mpss.udel.edu/rice/, dataset NPO) and has been noted in regard to the *Arabidopsis* transcriptome as well (Becker and Feijo, 2007). Proteome studies also support that rice pollen allergens are translated at high levels of abundance (Kerim et al., 2003a, 2003b; Imin et al., 2004; Dai et al., 2006, 2007).

Most highly expressed pollen allergens are present in multiple copies, often are intronless, and the most abundant are often grouped. Additionally, not only the directly expressed sequences but also some non-coding portions of these gene loci may be highly conserved, including UTRs and upstream regulatory sequences. These multiple genes could possibly have arisen through genomic duplication, as occurred in rice at approximately 57 MYA (Wang et al., 2007), but the similarity of these sequences, as well as their typical restriction to a single chromosome, suggests a more recent origin. Many of the most abundantly transcribed pollen proteins are encoded by loci lacking introns or containing a reduced number of introns. Since introns in some cases may be recruited in post-transcriptional control (such as through the generation of siRNAs, Piriyaongsa and Jordan, 2008), elimination of introns could potentially release translational regulation, leading to further enhanced expression.

Commonly, highly expressed sequences were encoded in pairs on opposite DNA strands grouped within several kbp. Sequence duplication, accompanied by production of nearby pseudogenes, has been suggested as a signature for the

activity of transposable elements, particularly retrotransposons, in creating chimeral genes (Wang et al., 2006). Retrotransposon-mediated gene duplication has been reported in other crop plants and appears to account for much of the relatively rapid origin and high degree of variability in tomato fruit morphology (Xiao et al., 2008). These gene insertions and duplications in tomato appeared to be primed by the occurrence of repeat sequences near critical existing genes. In *Oryza*, many of the most highly transcribed pollen allergen loci also appear to be linked with nearby LTR-associated sequences. Frequently, gene duplication in eukaryotes has been reported to occur through non-homologous DNA end-joining as a common motif (Gu and Lieber, 2008), as well as illegitimate recombination whereby a DNA repair mismatch may insert a duplicated strand of DNA that is incorporated into the genome (Guan et al., 2007). However, reported bursts of LTR-retrotransposon activation (Vitte et al., 2007) may leave LTR repeats flanking duplicated sequences.

The extreme degree of precision that seems to be evident in the insertion of Ory s 23 sequences (including nearly 100% duplication of non-coding, as well as coding sequences) seems highly unusual unless it is a relatively recent occurrence. Although genomic sequences have reportedly been precisely removed by retrotransposable elements (van de Lagemaat et al., 2005), similar precision in insertion has not. These segments of repeated nucleotide sequences may thus contribute to retrotransposon insertion but also favor alternative methods of illegitimate DNA mismatch as well. In general, retrotransposable elements tend to activate in the context of hypomethylated DNA. Hypomethylation may also trigger read-through transcription that inadvertently activates linked genes (Cheng et al., 2006). It may be noteworthy that pollen is reported to display especially low levels of DNA methylation compared to sporophytic cells (Oakeley et al., 1997), which, if true, could potentially favor such high transcription rates.

Amplification of pollen allergen transcription, particularly in grasses, is well known, and evidence suggests a role for retrotransposition. Wang et al. (2006) found numerous examples of chimeric retrogenes in rice representing captured mRNA messages that have been reverse transcribed and inserted, lacking introns, often in reverse direction; current bioinformatic data on putative pollen allergens concur with this possibility, as well as attachment to signature LTR sequences. As younger retrogenes accumulate indels rapidly, LTR flanking regions will decay, and the retrogene sequence will soon lose direct evidence of its LTR origin (Ma et al., 2004). Such a mechanism could provide positive feedback for the production of highly transcribed sequences and account for amplification over time.

METHODS

Microarray

Affymetrix 57K Rice GeneChip™ microarray provides essentially genomic coverage of genes. The microarray consists of 57 381 probe sets with 182 control and reporter probe sets.

Specific putative pollen allergy candidates were selected from expression in protein groups noted for human allergenic responses in grass pollen (Radauer and Breiteneder, 2006) and from genomic predictions (Jiang et al., 2005). Screened pollen allergens in this study were identified as candidates based on high up-regulation and expression in pollen. Transcription and translation were detected using the Affymetrix 57K rice genome microarray, MPSS and proteome data. Overall expression profiles of candidate genes are available on Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/, platform GPL2025).

Pollen Isolation

Mature anthers were collected from field-grown rice (*Oryza sativa* ssp. *japonica*, cultivar Katy) at the Dale Bumpers National Rice Research Center and University of Arkansas Extension Station near Stuttgart, Arkansas, to obtain pollen isolates. Collections were made on three different days using three different field sites, forming biological/technical replicates that were kept separated at all subsequent steps in processing. Field material displayed high vigor and no sign of disease, which was supported by absence calls for major reporter pathogens on the Affymetrix 57K rice microarray controls.

RNA Preparation

Mature viable rice pollen (Figure 4) was isolated by grinding collected anthers in a cold 45% sucrose solution. Pollen cytoplasm was collected by filtering out debris using a 100- μ m nylon mesh membrane and pelleting the resulting fraction at 300 g for 3 min. Pollen was washed once in cold 45% sucrose solution and stored frozen at or below -80°C until use. Seedlings were collected as a sporophytic control from rice seed of Katy germinated in soil and cultured to developmental stage V3 (collar forms on leaf 3 on main stem), according to Counce et al. (2000), and harvested. Seedling tissue was frozen in liquid nitrogen, ground into a fine powder and then stored until needed. All samples were stored at or below -80°C until RNA isolation. Total RNA was purified using the RNeasy plant mini kit according to manufacturer's instructions (Qiagen). RNA

concentration and quality of pollen and seedlings were determined by routine spectrophotometric measurement and agarose gels. 100 μ g total RNA of seedlings and mature pollen were used for probe preparation for each of the three biological/technical replicates performed.

Oligonucleotide Microarray Hybridization and Data Collection

Since the amount of starting total RNA was low, the GeneChip Two-Cycle cDNA Synthesis Kit from Affymetrix (Santa Clara, CA, USA) was used for target preparation with signal amplification. The Affymetrix GeneChip Rice Genome oligonucleotide array was used for the hybridization (45°C for 16 h) with a mixture containing 15 μ g of fragmented cRNA. Subsequent washing and staining steps were performed on a GeneChip Fluidics Station 450 and the chips were scanned on a GeneChip Scanner 3000. All experimental procedures strictly followed instructions specified in the Affymetrix GeneChip Expression Analysis Technical Manual. Instrument control and data collection were carried out with GeneChip Operating Software (GCOS, ver. 1.1.1). In order to minimize experimental variability, standardized microarray operation was performed by an experienced investigator throughout the study. The quality and quantity of the original RNA samples and of the cRNA probes generated were determined with the Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, USA), and by spectrophotometric measurements at 260 and 280 nm on a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA). The microarray data generated from all chips met the quality-control criteria set by Affymetrix. Consistency of the data was supported by Pearson's coefficient of correlation (r), which, for pollen biological/technical replicates, $r = 0.9898, 0.9931, \text{ and } 0.9971$, and, for control seedling replicates, $r = 0.9918, 0.9925, \text{ and } 0.9929$. Calls for transcripts were calculated based on the Affymetrix PM/MM method using GCOS software (version 1.1.1) as being present ($p < 0.05$), absent ($p > 0.10$) or marginal ($0.05 < p < 0.10$). The values for probe set intensity and up-regulation were calculated using GC Robust Multi-array Average (GCRMA)

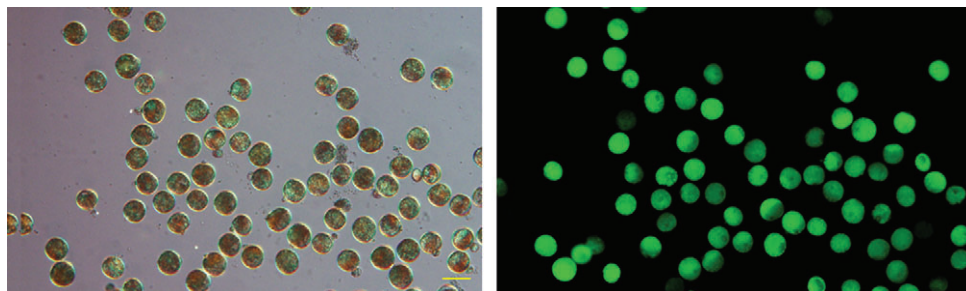


Figure 4. Photomicrograph of Isolated Pollen in Differential Interference Contrast (Left) and Fluorescein Diacetate Viability Screen in Epifluorescence Microscopy (Right).

Bar = 50 μ m.

background adjustment, quantile normalization, and median-polish summarization on Affymetrix microarray probe-level data using afflmmGUI (Irizarry et al., 2003; Wu et al., 2004). Values correspond to the A and M parameters, with B parameter (p values) indicating consistency of the grouped data. To determine probe matches, probe sequences were compared with TIGR rice pseudochromosome sequences (v. 5) using blastn to determine number of probes with exact 25-mer match. These are listed in parentheses in the tables of the charts. dChip software (Li and Wong, 2003; www.biostat.harvard.edu/complab/dchip/) was used for additional independent analysis of data and for GEO comparisons (at a CEL file level) of different sporophytic tissues.

SUPPLEMENTARY DATA

Supplementary Data are available at www.mplant.oxfordjournals.org.

FUNDING

Financial support was provided by Australian Research Council (PLB, MBS), University of Melbourne and University of Oklahoma (SDR).

ACKNOWLEDGMENTS

We thank Dr Xiaoping Gou, Dr Tong Yuan, Xiaoping Wei, and Cal Lemke, University of Oklahoma, for excellent field, technical, and growth assistance, and Dr Xinkun Wang, University of Kansas Genomics Facility for Affymetrix processing. Special thanks are due to Dr Yulin Jia, Dale Bumpers National Rice Research Center, Stuttgart, Arkansas, USA, for providing field material and advice during our collections and to Prof. Karen Moldenhauer, University of Arkansas Extension Station, Stuttgart, for the seed. We thank Professor Terry Speed and the staff of WEHI software development for help with statistical analysis. No conflict of interest declared.

REFERENCES

- Asero, R., Amato, S., Alfieri, B., Folloni, S., and Mistrello, G. (2007). Rice: another potential cause of food allergy in patients sensitized to lipid transfer protein. *Int. Arch. Allergy Immunol.* **143**, 69–74.
- Becker, J.D., and Feijo, J.A. (2007). How many genes are needed to make a pollen tube? Lessons from transcriptomics. *Ann. Bot.* **100**, 1117–1123.
- Bennetzen, J.L. (2007). Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* **10**, 176–181.
- Cheng, C., Daigen, M., and Hirochika, H. (2006). Epigenetic regulation of the rice retrotransposon Tos17. *Mol. Genet. Genomics.* **276**, 378–390.
- Counce, P.A., Keisling, T.C., and Mitchell, A.J. (2000). A uniform, objective, and adaptive system for expressing rice development. *Crop Science.* **40**, 436–443.
- Dai, S.J., Chen, T.T., Chong, K., Xue, Y.B., Liu, S.Q., and Wang, T. (2007). Proteomic identification of differentially expressed proteins associated with pollen germination and tube growth reveals characteristics of germinated *Oryza sativa* pollen. *Mol. Cell Proteomics.* **6**, 207–230.
- Dai, S.J., Li, L., Chen, T.T., Chong, K., Xue, Y.B., and Wang, T. (2006). Proteomic analyses of *Oryza sativa* mature pollen reveal novel proteins associated with pollen germination and tube growth. *Proteomics.* **6**, 2504–2529.
- Gu, J., and Lieber, M.R. (2008). Mechanistic flexibility as a conserved theme across 3 billion years of nonhomologous DNA end-joining. *Genes Dev.* **22**, 411–415.
- Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics.* **175**, 933–943.
- Holmes-Davis, R., Tanaka, C.K., Vensel, W.H., Hurkman, W.J., and McCormick, S. (2005). Proteome mapping of mature pollen of *Arabidopsis thaliana*. *Proteomics.* **5**, 4864–4884.
- Imin, N., Kerim, T., Rolfe, B.G., and Weinman, J.J. (2004). Effect of early cold stress on the maturation of rice anthers. *Proteomics.* **4**, 1873–1882.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* **4**, 249–264.
- Jiang, S.Y., Jasmin, P.X., Ting, Y.Y., and Ramachandran, S. (2005). Genome-wide identification and molecular characterization of Ole_e_1, Allerg_1 and Allerg_2 domain-containing pollen-allergen-like genes in *Oryza sativa*. *DNA Res.* **12**, 167–179.
- Kerim, T., Imin, N., Weinman, J.J., and Rolfe, B.G. (2003a). Proteome analysis of male gametophyte development in rice anthers. *Proteomics.* **3**, 738–51.
- Kerim, T., Imin, N., Weinman, J.J., and Rolfe, B.G. (2003b). Proteomic analysis reveals developmentally expressed rice homologues of grass group II pollen allergens. *Functional Plant Biology.* **30**, 843–852.
- Li, C., and Wong, W.H. (2003). DNA-chip analyzer (dChip). In *The Analysis of Gene Expression Data: Methods and Software*, G. Parmigiani, Garrett, E.S., Irizarry, R. and S.L. Zeger, eds (New York: Springer), pp. 120–141.
- Ma, J.X., Devos, K.M., and Bennetzen, J.L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869.
- Mohapatra, S., Lockey, R., and Shirley, S. (2005). Immunobiology of grass pollen allergens. *Curr. Allergy Asthma Rep.* **5**, 381–387.
- Nobuta, K., et al. (2007). An expression atlas of rice mRNAs and small RNAs. *Nat. Biotechnol.* **25**, 473–477.
- Noir, S., Brautigam, A., Colby, T., Schmidt, J., and Panstruga, R. (2005). A reference map of the *Arabidopsis thaliana* mature pollen proteome. *Biochem. Biophys. Res. Commun.* **337**, 1257–66.
- Oakeley, E.J., Podesta, A., and Jost, J.P. (1997). Developmental changes in DNA methylation of the two tobacco pollen nuclei during maturation. *Proc. Natl Acad. Sci. U S A.* **94**, 11721–11725.
- Piriyapongsa, J., and Jordan, I.K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA.* **14**, 814–821.

- Radauer, C., and Breiteneder, H.** (2006). Pollen allergens are restricted to few protein families and show distinct patterns of species distribution. *J. Allergy Clin. Immunol.* **117**, 141–147.
- Sampedro, J., and Cosgrove, D.** (2005). The expansin superfamily. *Genome Biol.* **6**, 242.1–11.
- Sampedro, J., Lee, Y., Carey, R.E., de Pamphilis, C., and Cosgrove, D.J.** (2005). Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. *Plant J.* **44**, 409–419.
- Sen, M.M., Adhikari, A., Gupta-Bhattacharya, S., and Chanda, S.** (2003). Airborne rice pollen and pollen allergen in an agricultural field: aerobiological and immunochemical evidence. *J. Environ. Monit.* **5**, 959–962.
- Sheoran, I.S., Ross, A.R.S., Olson, D.J.H., and Sawhney, V.K.** (2007). Proteomic analysis of tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*) pollen. *J. Exp. Bot.* **58**, 3525–3535.
- Sheoran, I.S., Sproul, K.A., Olson, D.J.H., Ross, A.R.S., and Sawhney, V.K.** (2006). Proteome profile and functional classification of proteins in *Arabidopsis thaliana* (Landsberg erecta) mature pollen. *Sex. Plant Reprod.* **19**, 185–196.
- Song, Z.P., Lu, B.R., and Chen, J.K.** (2001). A study of pollen viability and longevity in *Oryza rufipogon*, *O. sativa*, and their hybrids. *International Rice Research Notes.* **26**, 31–32.
- Swoboda, I., et al.** (2004). Molecular characterization of polygalacturonases as grass pollen-specific marker allergens: expulsion from pollen via submicronic respirable particles. *J. Immunol.* **172**, 6490–6500.
- Tsai, Y.T., Chen, S.H., Lin, K.L., and Hsieh, K.H.** (1990). Rice pollen allergy in Taiwan. *Ann. Allergy* **65**, 459–462.
- Valdivia, E.R., Sampedro, J., Lamb, J.C., Chopra, S., and Cosgrove, D.J.** (2007). Recent proliferation and translocation of pollen group 1 allergen genes in the maize genome. *Plant Physiol.* **143**, 1269–81.
- Valenta, R., Duchene, M., Pettenburger, K., Sillaber, C., Valent, P., Bettelheim, P., Breitenbach, M., Rumpold, H., Kraft, D., and Scheiner, O.** (1991). Identification of profilin as a novel pollen allergen; IgE autoreactivity in sensitized individuals. *Science.* **253**, 557–560.
- van de Lagemaat, L.N., Gagnier, L., Medstrand, P., and Mager, D.L.** (2005). Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res.* **15**, 1243–1249.
- Vitte, C., Panaud, O., and Quesneville, H.** (2007). LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics.* **8**, 218.
- Wang, W., et al.** (2006). High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell.* **18**, 1791–1802.
- Wang, X., Tang, H., Bowers, J.E., Feltus, F.A., and Paterson, A.H.** (2007). Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics.* **177**, 1753–1763.
- Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., and Spencer, F.** (2004). A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909–917.
- Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J., and van der Knaap, E.** (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science.* **319**, 1527–1530.
- Xu, H., Theerakulpisut, P., Goulding, N., Suphioglu, C., Singh, M.B., and Bhalla, P.L.** (1995). Cloning, expression and immunological characterization of *Ory s 1*, the major allergen of rice pollen. *Gene.* **164**, 255–259.
- Xu, H.L., Goulding, N., Zhang, Y., Swoboda, I., Singh, M.B., and Bhalla, P.L.** (1999). Promoter region of *Ory s 1*, the major rice pollen allergen gene. *Sex. Plant Reprod.* **12**, 125–126.