

# Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants

Kevin A.T. Silverstein<sup>1\*</sup>, William A. Moskal Jr.<sup>2</sup>, Hank C. Wu<sup>2</sup>, Beverly A. Underwood<sup>2</sup>, Michelle A. Graham<sup>3</sup>, Christopher D. Town<sup>2</sup> and Kathryn A. VandenBosch<sup>1</sup>

<sup>1</sup>Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA,

<sup>2</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, and

<sup>3</sup>USDA-ARS CICGR Unit, Iowa State University, Ames, IA 50011, USA

Received 12 September 2006; revised 26 February 2007; accepted 21 March 2007.

\*For correspondence (fax: + 1 612 625 1738; e-mail silve023@umn.edu).

---

## Summary

Multicellular organisms produce small cysteine-rich antimicrobial peptides as an innate defense against pathogens. While defensins, a well-known class of such peptides, are common among eukaryotes, there are other classes restricted to the plant kingdom. These include thionins, lipid transfer proteins and snakins. In earlier work, we identified several divergent classes of small putatively secreted cysteine-rich peptides (CRPs) in legumes [Graham *et al.* (2004) *Plant Physiol.* 135, 1179–97]. Here, we built sequence motif models for each of these classes of peptides, and iteratively searched for related sequences within the comprehensive UniProt protein dataset, the Institute for Genomic Research's 33 plant gene indices, and the entire genomes of the model dicot, *Arabidopsis thaliana*, and the model monocot and crop species, *Oryza sativa* (rice). Using this search strategy, we identified ~13 000 plant genes encoding peptides with common features: (i) an N-terminal signal peptide, (ii) a small divergent charged or polar mature peptide with conserved cysteines, (iii) a similar intron/exon structure, (iv) spatial clustering in the genomes studied, and (v) overrepresentation in expressed sequences from reproductive structures of specific taxa. The identified genes include classes of defensins, thionins, lipid transfer proteins, and snakins, plus other protease inhibitors, pollen allergens, and uncharacterized gene families. We estimate that these classes of genes account for ~2–3% of the gene repertoire of each model species. Although 24% of the genes identified were not annotated in the latest *Arabidopsis* genome releases (TIGR5, TAIR6), we confirmed expression via RT-PCR for 59% of the sequences attempted. These findings highlight limitations in current annotation procedures for small divergent peptide classes.

**Keywords:** antimicrobial peptides, cysteine-rich peptides, defensins, lipid transfer protein, thionin, RALF.

---

## Introduction

Plants have evolved a complex array of defenses against pathogens and predators (Dangl and Jones, 2001; Mysore and Ryu, 2004; Veronese *et al.*, 2003). The plant cell wall serves as an initial pre-formed barrier to microbial penetration. Antimicrobial compounds such as triterpene and steroidal saponins are constitutively produced to deter would-be invaders (Osbourn, 1996). If the cell is penetrated by an intruder, recognition leads to the production of toxic reactive oxygen intermediates (Alvarez *et al.*, 1998; Jabs *et al.*, 1997; Kawano, 2003), cell-wall strengthening (Brisson *et al.*, 1994) and lignification (Cano-Delgado *et al.*, 2003; Vance *et al.*, 1980), activation of numerous protein kinase pathways (Scheel, 1998; Zhang and Klessig, 2001),

and the production of a range of signaling intermediates (Kunkel and Brooks, 2002). These latter signaling events can lead to localized reactions such as a hypersensitive response (HR) in which the cell undergoes a programmed death (Greenberg and Yao, 2004; Lam *et al.*, 2001), or to the release of a wide range of antimicrobial compounds such as phytoalexins (Darvill and Albersheim, 1984), and pathogenesis-related (PR) proteins (>100 amino acids) and small peptides (<100 amino acids; Van Loon and van Kammen, 1970; Van Loon and van Strien, 1999). Additionally, the plant may be immunized against unrelated pathogens via systemic acquired resistance (Durrant and Dong, 2004; Ross, 1961).

Antimicrobial peptides (AMPs) are one of several types of PR peptides that are abundant among eukaryotic organisms (Boman, 2003; Broekaert *et al.*, 1997; Bulet *et al.*, 2004; Garcia-Olmedo *et al.*, 1998; Theis and Stahl, 2004). Their mode of action typically involves disruption of the pathogen's membrane via both specific and non-specific electrostatic and hydrophobic interactions with cell surface groups (Shai, 2002; Thevissen *et al.*, 2003). However, evidence exists to support additional modes of action, such as ribosomal inactivation (Endo *et al.*, 1987; Vivanco *et al.*, 1999) and related forms of protein inhibition (Chen *et al.*, 2002; Garcia-Olmedo *et al.*, 1983). Although numerous simple, single-helical AMPs have been identified, most are larger peptides (20–100 amino acids) that are rich in one or two amino acids, often proline, glycine, histidine, arginine, tryptophan, or cysteine (Boman, 2003; Bulet *et al.*, 2004).

Cysteine-rich peptides (CRPs) are particularly well represented among plants, and are divided into many classes (Boman, 2003; Broekaert *et al.*, 1997; Bulet *et al.*, 2004; Garcia-Olmedo *et al.*, 1998; Theis and Stahl, 2004). The number and arrangement of cysteine residues in the primary sequence distinguishes each CRP class from the others. Pairs of cysteines form characteristic disulfide bridges in distinct 3D structural folds that also typify the class (Broekaert *et al.*, 1997). Plant defensins, which have a characteristic eight-cysteine motif that forms four cysteine bridges in a knottin fold, are the most ancient of cysteine-rich AMPs. Related defensins with a common 3D structure have been identified in diverse eukaryotic taxa, including fungi (Mygind *et al.*, 2005). But plants possess many classes of antimicrobial CRPs that do not appear to have homologs outside the plant kingdom, such as thionins (Florack and Stiekema, 1994; Garcia-Olmedo *et al.*, 1989), lipid transfer proteins (Molina *et al.*, 1993; Terras *et al.*, 1992), and snakins (Berrocal-Lobo *et al.*, 2002; Segura *et al.*, 1999).

Despite the structural differences among CRP classes, members from each class have striking similarities in their molecular properties, spatial expression and function (Broekaert *et al.*, 1997; Garcia-Olmedo *et al.*, 1998). Cysteine-rich peptides have an N-terminal signal peptide and a positively charged mature peptide with conserved cysteine residues. Members of each class may be constitutively produced as a front-line defense in peripheral cell layers of nutrient-rich structures (e.g. flowers and seeds), or induced by various defense or stress-signaling pathways (Broekaert *et al.*, 1997), particularly the jasmonate/ethylene (JA/ET) pathway (Farmer and Ryan, 1990; Thomma *et al.*, 1998). In addition to active microbicidal roles that define the AMPs, several AMP homologs have been determined to have more passive defense functions that deter predation, such as allergenicity (Himly *et al.*, 2003; van Ree, 2002) and trypsin or  $\alpha$ -amylase inhibition (Bloch and Richardson, 1991; Melo *et al.*, 2002; Odani *et al.*, 1983). In fact, additional classes of CRPs found in plants with a protease inhibitory

function share all of the properties of the AMPs described above except for their microbicidal activity (Laskowski and Qasim, 2000; Rawlings *et al.*, 2004; Ryan, 1990). These include Bowman Birk (Birk, 1985; Mello *et al.*, 2003), Kunitz (Bauw *et al.*, 2006; Talyzina and Ingvarsson, 2006), and Kazal-type (Laskowski and Qasim, 2000) inhibitors.

Some CRPs have non-defense roles. Several types of these proteins are expressed specifically in flowers and seeds, where they play reproductive regulatory roles, such as the stigma-specific STIG1 family (Goldman *et al.*, 1994), the defensin-like S-locus cysteine-rich (SCR) proteins (Schopfer *et al.*, 1999; Takayama *et al.*, 2000b), and related pollen coat proteins (PCPs; Takayama *et al.*, 2000a; Watanabe *et al.*, 2000). Other types of CRPs include proteins that regulate plant growth and development, such as the rapid alkanization factor (RALF) proteins (Pearce *et al.*, 2001), lipid transfer protein (LTP)-like xylogens (Motose *et al.*, 2004) and the gibberellin-stimulated GASA/GAST family of proteins (Aubert *et al.*, 1998).

Although the number and diversity of characterized plant CRPs is large, it is thought that many more members of this superfamily remain to be discovered. With the advent of genome-scale sequencing technologies, new opportunities exist to discover novel CRPs. By analyzing the whole-genome expressed sequence tag (EST) libraries from numerous plants, and the available *Arabidopsis thaliana* and *Medicago truncatula* genome sequence data, we recently determined that plant genomes typically possess several hundred distinct defensin-like sequences (DEFLs; Graham *et al.*, 2004; Silverstein *et al.*, 2005). These findings were in stark contrast to the total of 15 defensins that had previously been described in *Arabidopsis* (Thomma *et al.*, 2002), and indicated that the majority of DEFLs had remained unannotated in the *Arabidopsis* genome (TIGR5, TAIR6). Other classes of CRPs share the same molecular properties that made DEFLs difficult to identify, namely small size and extreme sequence divergence. Therefore, we set out to develop methods that would exhaustively search the genomes of the model dicot, *Arabidopsis*, the model monocot, *Oryza sativa* (rice), and the expressed sequences of higher plants for unidentified CRPs.

## Results

We used multiple sequence alignments of CRP groups identified in previous work (Graham *et al.*, 2004; Silverstein *et al.*, 2005) as a starting point for iterative hidden Markov model (HMM) builds and database searches. Some of these initial groups appeared to have uncharacterized cysteine arrangements, whereas the cysteine patterns of others resembled known CRPs such as defensins, thionins, and RALFs (Silverstein *et al.*, 2006). Seven successive iterations of HMM builds and searches were performed, after which no new significant hits were identified, as described in Experimental procedures.

*Small genes encoding putatively secreted cysteine-rich peptides are abundant in plants*

We identified 12 824 distinct plant CRPs among the expressed sequences of 33 plants (145 721 of 4 801 711 total ESTs scanned). As shown in Tables 1 and 2, these sequences were clustered by the search protocol into 516 sequence-related subgroups. These subgroups are grouped into 24 primary classes based on their characteristic cysteine motifs. Many of these CRPs had cysteine motifs similar to those of known proteins (Table 1), but others, to our knowledge, have not been previously characterized (Table 2). An exhaustive list of the 516 subgroups and their characteristics is provided in Table S1. FASTA files, trimmed alignments and final HMMs for all groups are also given as Supplementary Data.

During the course of the iterations, additional groups of putatively secreted sequences were encountered but not examined for various reasons. Plantacyanins (also known as basic blue copper proteins; Murata *et al.*, 1982) were not analyzed because they often had two or fewer cysteines in a non-conserved arrangement. Several groups were skipped despite a large number of conserved cysteines because they were excessively large for AMPs (>150 amino acids) and/or had a complicated gene structure (more than three exons). These included the family of X8 module glycosyl hydrolase-associated domains (six core cysteines; Henrissat and Davies, 2000), thaumatins and related osmotins (16 cysteines; Van Loon and van Strien, 1999), extracellular dermal glycoproteins (Sato *et al.*, 1992) and related xyloglucan-specific fungal endoglucanase inhibitors (12 cysteines; Qin *et al.*, 2003), and a novel family of 33-kDa secretory proteins with 12 cysteines that are associated with a group of serine/threonine kinases.

*Each cysteine-rich peptide class contains divergent subgroups that have expanded extensively within taxonomic groups*

Within most classes of CRPs, a large subgroup can be found that is composed of sequences that are well conserved across diverse plant taxa. This is illustrated by the top multiple sequence subalignment in Figure 1a–c for LTP, thionin, and RALF classes, respectively. Contrasting with this observation, these same CRP classes also have multiple divergent subgroups that often are specific to a particular taxon (examples are shown in the additional subalignments for Figure 1a–c). Table 3 shows that a large number of subgroups from the most abundant LTP-like, thionin-like and RALF-like classes are nearly entirely composed of sequences from a single taxonomic family or order. The taxonomic specificity in Table 3 is statistically significant in all summarized cases ( $P < 0.05$ , Fisher exact test; see Experimental procedures and Table S1), which is

not surprising given the large size of the subgroups involved.

*Cysteine-rich peptides of most classes have subgroups that are over-represented in expressed sequence tags from reproductive structures*

Cysteine-rich peptide-encoding ESTs were identified from the 33 plant Gene Indices at The Institute for Genomic Research, and annotated according to the morphological structure from which they originated. After excluding those ESTs from 'mixed tissues' or 'whole plants' or other ambiguous annotations, 4 002 137 ESTs were annotated as having originated from reproductive (floral, seed, or fruit structures, as enumerated in the Experimental procedures) or vegetative plant structures. Of these, 37% were from reproductive structures, including 16% from seeds. Despite these large proportions, many of the LTP-like, thionin-like, and RALF-like subgroups were statistically over-represented in transcripts derived from seeds and other reproductive structures (Table 4, Table S1). Over-representation of a particular morphological structure suggests that more transcripts were observed from that structure than would be expected by chance given the sample size ( $P < 0.05$ , Fisher exact test; see Experimental procedures).

*Cysteine-rich peptide fusions occur frequently with other cysteine-rich peptide, glycine-rich protein and proline-rich protein domains*

Within each class of CRP, numerous examples of transcripts encoding CRP–CRP, CRP–glycine-rich protein (GRP), and CRP–proline-rich protein (PRP) fused peptides were observed. Cysteine-rich peptide subgroups with these properties are marked in Table S1, and a schematic depiction of the relative abundance of each type of fusion is given in Table 5. The most frequent type of fusion involved PRPs anchored at the C-terminal end of the predicted protein (affecting 33 subgroups). Tandem duplications of like-class CRP motifs within a single predicted protein were also highly prevalent (affecting 23 subgroups).

*The verification rate and accuracy of our Arabidopsis gene predictions is good*

A large percentage of CRPs from several classes that we identified in the model plant *A. thaliana* were not annotated previously (highlighted in bold in Table 1). In total, we identified 196 novel genes. This represents 24% of all CRPs (39% if we exclude genes in the DEFL class, which were annotated largely because of our own direct submission to TAIR in earlier work; Silverstein *et al.*, 2005). These 196 predicted, unannotated genes were submitted to TAIR, where they were assigned official Arabidopsis Genome

**Table 1.** Characteristics of previously described cysteine-rich peptide classes in plants identified in this work by iterative searches

| Group description                   | Group ID range <sup>a</sup> | Unigene count (12 824 total) | EST count (145 721 total) | Subgrp count (516 total) | At. <sup>b</sup> count (825 total) | Os. <sup>b</sup> count (598 total) | % At. hits unannotated <sup>c</sup> | % Os. hits unannotated <sup>c</sup> | Approx. size <sup>d</sup> (aa) | Cysteine arrangement <sup>e</sup>                                                                                                                                             |
|-------------------------------------|-----------------------------|------------------------------|---------------------------|--------------------------|------------------------------------|------------------------------------|-------------------------------------|-------------------------------------|--------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| LTP/2S ECA 1                        | CRP3600–CRP5080             | 4038                         | 67501                     | 128                      | 276                                | 188                                | 40                                  | 9                                   | 65–90                          | CX(6,15)CX(9,31)CCX(8,21)CX(13,35)CX(5,18)C<br>CX(5,13)CX(14,20)CCX(8,10) CX(10,32)C                                                                                          |
| Defensin/DEFL                       | CRP0000–CRP1530             | 2320                         | 13635                     | 135                      | 323                                | 93                                 | 80 <sup>f</sup>                     | 42                                  | 40–70                          | CX(4,25)CX(12,12)CX(3,4) CX(3,17)CX(4,32)CX(1,6)C<br>CX(3,21)CX(12,12)CX(3,4) CX(3,15)CX(4,23)CCC<br>CX(2,14)CX(3,5)CX(3,16) CX(4,28)CXC<br>CX(3,5)CX(8,17) CX(4,6)C          |
| Glutenin/gliadin/prolamins          | CRP5090–CRP5270             | 1298                         | 19750                     | 26                       | 0                                  | 30                                 | 0                                   | 3                                   | 90–145                         | CX(6,7)CX(15,25)CX(6) CCX(11)CX(35–80) CX(7)C<br>CX(7) CX(19) CX(6) CC X(40–65) C                                                                                             |
| Hevein                              | CRP5600–CRP5820             | 966                          | 8192                      | 24                       | 16                                 | 25                                 | 0                                   | 8                                   | 130–155                        | CX(29)CCX(11)CX(75–100)CX(7)C<br>CX(1,8)CX(4,5)CCX(5)CX(6) CX(3,5)CX(3,4)C                                                                                                    |
| Pollen Ole e I                      | CRP3300–CRP3510             | 647                          | 6059                      | 25                       | 35                                 | 52                                 | 3                                   | 4                                   | 100–130                        | CX(2,3)CX(19,23)CX(31,42) CX(8,16) CX(32,54)C<br>CX(2,3)CX(19,22)CX(9,13) CX(31,32)CX(11,14)C                                                                                 |
| Kunitz type inhibitor               | CRP6010–CRP6180             | 620                          | 7969                      | 20                       | 7                                  | 1                                  | 0                                   | 0                                   | 90–125                         | CX(42,51)CX(39,58)CX(1,10) CX(1,3)CX(1,6)C<br>CX(40,50)CX(34,61)CX(3,11)C                                                                                                     |
| RALF                                | CRP1700–CRP2120             | 568                          | 3566                      | 43                       | 39                                 | 46                                 | 41                                  | 11                                  | 60–90                          | CX(4,14)CX(22,51)CX(6,12) CX(5,14)CX(5,6)C<br>CX(3,12)CX(5,21)CX(5,6)C<br>CX(5,6)C                                                                                            |
| Thionin                             | CRP2200–CRP2610             | 506                          | 5967                      | 43                       | 71                                 | 44                                 | 75                                  | 9                                   | 25–55                          | CX(3)CX(3,4)CX(4,32) CX(2,3)CX(3,4)C = A<br>CCX(10,11)CX(8,10)CX(5) C?X(7,10)C = B<br>BX(8,22)A                                                                               |
| Bowman Birk inhibitor               | CRP3100–CRP3190             | 405                          | 4257                      | 10                       | 0                                  | 13                                 | 0                                   | 8                                   | 60–85                          | AX(4,10)CX(3)CX(16,24)CX(3)C<br>CX(3)CX(17,20)CX(3)CX(7,9)A                                                                                                                   |
| GASA/GAST/Snakin                    | CRP2700                     | 445                          | 3790                      | 1                        | 15                                 | 11                                 | 7                                   | 1                                   | 50–75                          | CCX(2)CX(1,2)CX(7,9)CX(2)CX(8,29) CX(7,11) CX(2,3)C<br>CCX(2) CX(8,9) CX(6) CX(3) CX(2)CX(11,15)CX(7,15) CX(1,3)C<br>CCX(4)                                                   |
| Maternally-expressed gene (MEG)/Ae1 | CRP5300–CRP5520             | 231                          | 1060                      | 17                       | 17                                 | 27                                 | 76                                  | 41                                  | 60–70                          | CX(3)CX(3)CX(7,11)CX(3)CX(2) CCX(2)CX(1,3)CX(11)CX(1,2) CX(11,14)-KCP.<br>CX(4,9)CX(4,12)CX(4,10) CX(6,12) CX(3)C<br>CX(2) CX(2) CX(4,16)CX(3,4) CCX(4)CX(5,10)CX(6)CX(2,10)C |
| Proteinase inhibitor II             | CRP5545–CRP5560             | 119                          | 847                       | 4                        | 1                                  | 5                                  | 0                                   | 20                                  | 30–50                          | CX(3)CX(7,9)CX(10,13)CX(1,2) CX(7)CX(5,11)CX(10,12)C                                                                                                                          |
| Root cap/LEA                        | CRP2820–CRP2850             | 115                          | 469                       | 4                        | 6                                  | 9                                  | 0                                   | 11                                  | 45–60                          | CX(3,6)CX(3)CX(3)CX(3) CX(2) CX(3,4)CX(1,2)CX(5,7)CX(12–24)C<br>CX(5,9)CX(6)CX(3)CX(12,13)CX(3) CX(2,4)CX(3) CX(6) CX(6) C                                                    |
| Antimicrobial peptide MBP-1         | CRP2900–CRP3000             | 97                           | 1339                      | 10                       | 1                                  | 15                                 | 0                                   | 0                                   | 50–60                          | CX(3)CX(6,14)CX(3)C                                                                                                                                                           |
| Legumin/Albumin-1                   | CRP1640–CRP1650             | 80                           | 668                       | 2                        | 0                                  | 0                                  | 0                                   | 0                                   | 15–25                          | CX(3)CX(6,14)CX(3)C                                                                                                                                                           |
|                                     |                             |                              |                           |                          |                                    |                                    |                                     |                                     | 65–90                          | CX(3,6)CX(5,9)CX(3,5)CX(5,14) CX(16,20)CX(5)CX(8)CX(10,12)C<br>CX(6) CX(7) CX(4,5) CX(9,11)C                                                                                  |
|                                     |                             |                              |                           |                          |                                    |                                    |                                     |                                     | 25–30                          |                                                                                                                                                                               |

Table 1. Continued

| Group description    | Group ID | Unigene count (12 824 total) | EST count (145 721 total) | Subgrp count (516 total) | At. <sup>b</sup> count (825 total) | Os. <sup>b</sup> count (598 total) | % At. hits unannotated <sup>c</sup> | % Os. hits unannotated <sup>c</sup> | Approx. size <sup>d</sup> (aa) | Cysteine arrangement <sup>e</sup>                                                   |
|----------------------|----------|------------------------------|---------------------------|--------------------------|------------------------------------|------------------------------------|-------------------------------------|-------------------------------------|--------------------------------|-------------------------------------------------------------------------------------|
| Kazal type inhibitor | CRP5940  | 40                           | 128                       | 1                        | 2                                  | 2                                  | 0                                   | 50                                  | 45–55                          | CX(7,12)CX(3,5)CX(6)CX(8) CX(4)CX(10)C                                              |
| Stig1                | CRP3080  | 39                           | 44                        | 1                        | 6                                  | 4                                  | 0                                   | 25                                  | 75–90                          | CX(2,7)CX(8,14)CCX(3,4) CX(9)CX(2)CX(3)CX(5,7) CCX(3)CX(9)CX(2)C-X(3) CX(2,5)CX(4)C |

<sup>a</sup>Arbitrary subgroup identifiers assigned in this work, enumerated in Table S1.

<sup>b</sup>Total number of family members identified in the *Arabidopsis thaliana* (At.) genome (TAIR version 6), or the *Oryza sativa* (Os.) genome (TIGR version 4).

<sup>c</sup>Percentages are shown in bold for classes of CRPs that were significantly underpredicted in the TAIR version 6 annotation.

<sup>d</sup>Approximate size of the portion of the mature peptide from the first to the last conserved cysteine residue. Regions outside this range were omitted because they rarely were conserved within a subgroup or class.

<sup>e</sup>Regular expression patterns shown match the majority of sequences within a class. For some classes, multiple patterns were typical (shown here aligned relative to each other). Conserved cysteine residues are shown in bold (C); X denotes any amino acid; numbers in curly braces indicate a range of variable residues; ? indicates an optional residue; A and B are arbitrary tags used to abbreviate two common thionin motifs appearing alone in some sequences, or fused to other motifs elsewhere.

<sup>f</sup>Percentages as reported earlier (Silverstein *et al.*, 2005).

Abbreviations: LTP, lipid transfer protein; DEFL, defensin like; GASA/T, gibberellic acid stimulated arabidopsis/tomato; LEA, late embryogenesis abundant.

Initiative (AGI codes), and will appear in the next annotation release (TAIR7). The AGI codes and molecular characteristics for all 825 *Arabidopsis* CRPs including chromosomal position, gene structure, predicted signal peptide cleavage location, predicted mature peptide molecular weight and isoelectric point, nucleotide sequence and external database cross-references are provided in Table S2 with additional detail in Tables S3 and S4. Ninety-four per cent of all *Arabidopsis* CRPs (annotated and unannotated) had either no intron or a single intron, typically between the exons encoding the signal peptide and mature peptide. Some members of the LTP and GASA/GAST-like subgroups had one or two additional introns in the region encoding the mature peptide.

We attempted to verify the expression of a subset of the newly predicted CRPs plus some of the still unverified DEFLs identified earlier (Silverstein *et al.*, 2005) via reverse transcriptase polymerase chain reaction (RT-PCR), rapid amplification of cDNA ends (RACE), and full-length (FL) sequencing. All sequences have been deposited in GenBank. The predicted genes that correspond to each accession are indicated in Table S2. Out of 77 non-DEFL CRPs and 40 new DEFLs attempted, 44/77 (57%) and 25/40 (63%), respectively, were successfully detected and sequenced (Figure S1). This is comparable to the verification rate observed for other *Arabidopsis* hypothetical and intergenic predictions (Xiao *et al.*, 2005). Among the recovered sequences that could confirm or refute our gene structure prediction, 60/69 (87%) had the same intron/exon structure as predicted (including 19 partial length sequences).

In addition to the targeted CRPs, a parallel gene prediction verification effort at the Institute for Genomic Research (TIGR) was under way (Moskal *et al.*, 2007), and yielded expression evidence for 19 DEFLs and 39 other CRPs (Figure S1). Interestingly, four of the genes that this parallel effort confirmed overlap the set of 117 targeted genes in this study, even though we failed to amplify them here. This suggests that some of our 'failed' targets could still be verified given another attempt or a different set of primers. Among these 58 additional verified sequences, 47 confirm our predicted gene structure, and 11 lacked enough sequence information to confirm or refute the prediction. Summarizing across both studies, our total percentage of gene structure predictions consistent with the experimental evidence was 92% (102/111).

As an additional source of expression verification for novel predicted genes, we consulted the massively parallel signature sequencing (MPSS) database (Meyers *et al.*, 2004). We determined that 12% and 22% of our novel CRPs and DEFLs, respectively, had evidence of expression via MPSS tags that match conservatively (i.e. were present in multiple MPSS runs, and exactly matched within the predicted coding sequence). Consistent with Table 4 and our earlier observations (Silverstein *et al.*, 2005), the

**Table 2.** Characteristics of uncharacterized cysteine-rich peptide (CRP) classes in plants identified in this work by iterative searches

| Group description   | Group ID range <sup>a</sup> | Unigene count (12 824 total) | EST count (145 721 total) | Subgrp count (516 total) | At. <sup>b</sup> count (825 total) | Os. <sup>b</sup> count (598 total) | % At.hits unannotated | % Os.hits unannotated | Approx. size <sup>c</sup> (aa) | Cysteine arrangement <sup>d</sup>                    |
|---------------------|-----------------------------|------------------------------|---------------------------|--------------------------|------------------------------------|------------------------------------|-----------------------|-----------------------|--------------------------------|------------------------------------------------------|
| Novel CRP 1 (NCRP1) | CRP2855–CRP2865             | 75                           | 388                       | 5                        | 4                                  | 5                                  | 0                     | 20                    | 30–40                          | CX(4,6)CX(6)CX(3,31)CX(3)CX(2,4)CX(3)                |
| Novel CRP 2 (NCRP2) | CRP3040                     | 42                           | 396                       | 1                        | 2                                  | 1                                  | 0                     | 0                     | 100–110                        | CX(3)CX(11)CX(8)CX(3)CX(12)CX(33,37)CX(23,27)C       |
| Novel CRP 3 (NCRP3) | CRP3210                     | 22                           | 73                        | 1                        | 2                                  | 9                                  | 0                     | 44                    | 100–110                        | CX(6,9)CX(22,33)CX(3)CX(9)CX(26,27)C                 |
| Novel CRP 4 (NCRP4) | CRP5880                     | 4                            | 21                        | 1                        | 0                                  | 0                                  | 0                     | 0                     | 65–70                          | CX(12)CX(21)CX(8)CX(12)C                             |
| Novel CRP 5 (NCRP5) | CRP1620                     | 4                            | 15                        | 1                        | 0                                  | 0                                  | 0                     | 0                     | 30–35                          | CX(6)CX(7,8)CX(3)CX(10)C                             |
| Novel CRP 6 (NCRP6) | CRP5850                     | 4                            | 5                         | 1                        | 0                                  | 0                                  | 0                     | 0                     | 30–35                          | CX(12)CX(5)CX(6)CX(3)C                               |
| Novel CRP 7 (NCRP7) | CRP5830                     | 4                            | 4                         | 1                        | 0                                  | 0                                  | 0                     | 0                     | 80–90                          | CX(6)CX(3)CX(12,14)CX(8)CX(19)CX(16,17)CX(2,3)CX(5)C |

<sup>a</sup>Arbitrary subgroup identifiers assigned in this work, enumerated in Table S1.<sup>b</sup>Total number of family members identified in the *Arabidopsis thaliana* (At.) genome (TAIR version 6), or the *Oryza sativa* (Os.) genome (TIGR version 4)<sup>c</sup>Approximate size of the portion of the mature peptide from the first to the last conserved cysteine residue. Regions outside this range were omitted because they rarely were conserved within a subgroup or class.<sup>d</sup>Regular expression patterns shown match the majority of sequences within a class. Conserved cysteine residues are shown in bold (C); X denotes any amino acid; numbers in curly braces indicate a range of variable residues.

majority of CRPs and DEFLs were observed preferentially in silique and inflorescence libraries. The MPSS identification status and morphological structure specificity for all Arabidopsis CRPs (including DEFLs) is indicated in Table S2 with a complete listing of all matches being provided in Table S5.

#### *Rice has significantly fewer total and novel cysteine-rich peptides than Arabidopsis*

We identified a total of 598 CRPs in the rice genome. Of these, 91 (15%) were novel and 68 (11%) were previously predicted in release 4.0 of TIGR's rice annotation, but had significant structural flaws (e.g. additional or missing exons, start sites omitting the signal peptide or key cysteines). Thirty-eight per cent of the CRPs that were annotated had the simple description 'expressed protein' or 'hypothetical protein'. Full details on the position, subgroup classification, status at TIGR and sequence of each rice CRP are provided in Table S6. A summary of the counts of rice genes in each CRP class is shown in Tables 1 and 2, in a column adjacent to the corresponding information in Arabidopsis.

Several observations are noteworthy in comparing the counts between the two species: (i) Arabidopsis has significantly more CRP genes than rice (38% more). This is most pronounced within the class of DEFLs, where Arabidopsis has more than three times as many genes. (ii) Despite the overall abundance of CRPs in Arabidopsis, a few classes have greatly expanded in rice. These include the prolamins, the Pol Ole e I allergens, the Bowman Birk inhibitors, the maternally expressed gene (MEG) family, and the MBP-1 antimicrobial peptide classes. (iii) Finally, the percentage of CRPs that were not annotated in the Arabidopsis genome is generally higher than the unannotated percentage in rice.

#### *Families of related cysteine-rich peptides are clustered in the Arabidopsis and rice genomes*

As shown in Figures 2 and 3, CRPs from the three non-DEFL classes with the most members (lipid transfer protein, thionin, and RALF-like genes) are clustered in their chromosomal position in both the Arabidopsis and rice genomes. On average, 85% of sequences shown within clusters (100 000 bp) are from the same sequence-related subgroups (Tables S2 and S3). Interestingly, 10 clusters of four or more sequences among these three classes were composed entirely of unannotated Arabidopsis CRPs (open and filled circles in Figure S2). Some of these clusters are remarkably large – most notably an LTP-like cluster on chromosome 5 and a thionin-like cluster on chromosome 1 that have 32 and 14 novel tandem-duplicated genes, respectively. Similar spatial clustering was observed for Arabidopsis DEFLs (Silverstein *et al.*, 2005) and is also clearly apparent among rice DEFLs (Figure S2). In contrast to Arabidopsis, however,

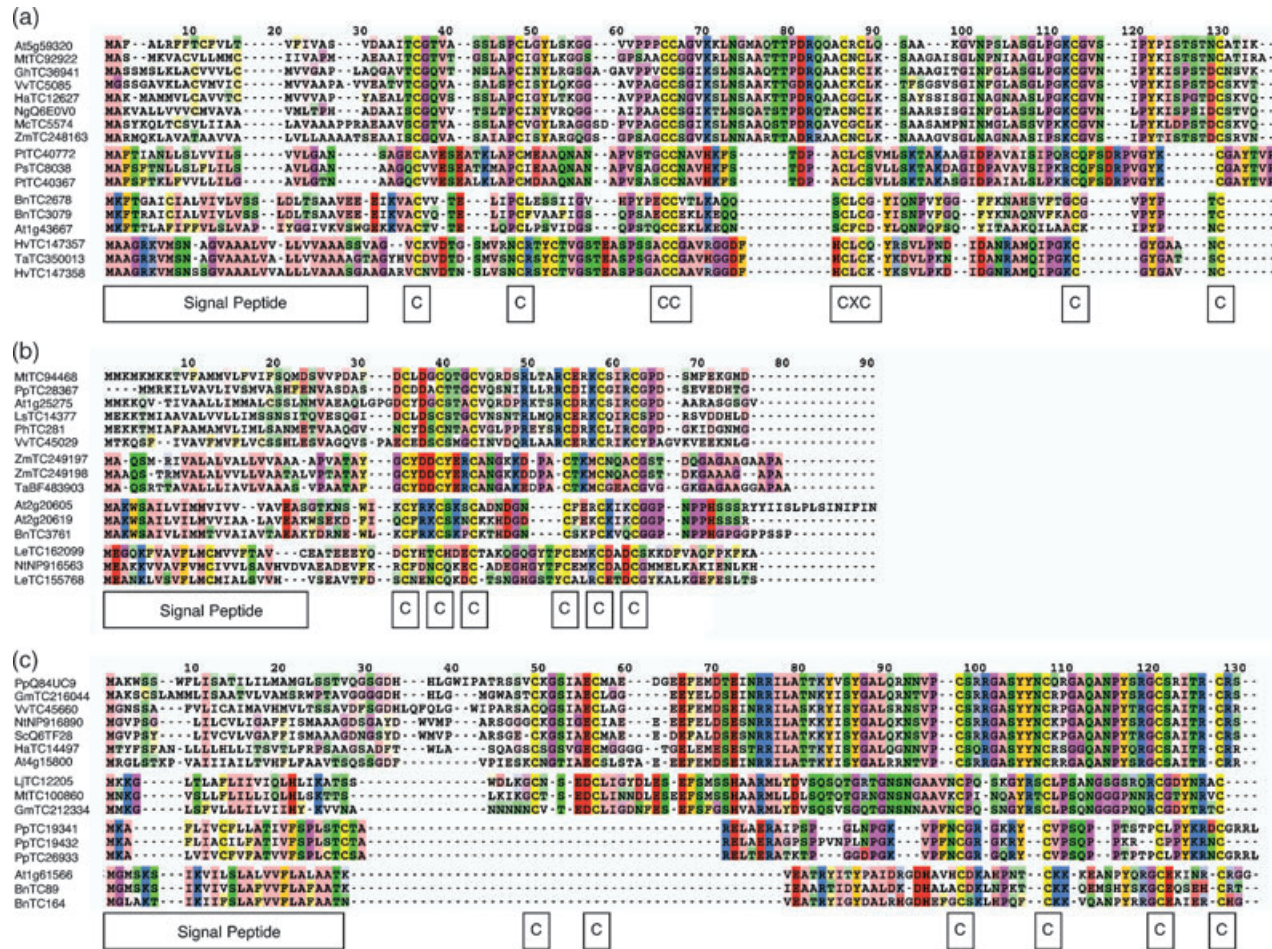


**Figure 1.** Representative alignments of CRP subgroups illustrate the sequence diversity within a class.

- (a) Lipid transfer protein-like sequences.  
 (b) Thionin-like sequences.  
 (c) Rapid alkanization factor-like sequences.

Representative sequences were aligned with CLUSTALW and visualized via Jalview with the shade consensus menu option. For all classes, the primary alignment of six to eight selected sequences comes from a subgroup that has multiple members per species spanning a diverse range of taxonomic families. Three additional alignments for each class are shown with selected members from subgroups that are dominated by the following individual taxonomic families or orders, respectively: (a) Coniferales, Brassicaceae, Poaceae; (b) Poaceae, Brassicaceae, Solanaceae; (c) Fabaceae, Salicaceae, Brassicaceae. The signal peptide region and conserved class-specific cysteine residues are displayed schematically below the alignments.

Sequence identifiers are prefixed with initials to designate the species: At, *Arabidopsis thaliana*; Bn, *Brassica napus*; Gh, *Gossypium hirsutum*; Gm, *Glycine max*/soja; Ha, *Helianthus annuus*; Hv, *Hordeum vulgare*; Le, *Lycopersicon esculentum*; Lj, *Lotus japonicus*; Ls, *Lactuca sativa*; Mc, *Mesembryanthemum crystallinum*; Mt, *Medicago truncatula*; Ng, *Nicotiana glauca*; Nt, *Nicotiana tobacum*; Ph, *Petunia hybrida*; Pp, *Populus* spp.; Ps, *Picea* spp.; Pt, *Pinus* spp.; Sc, *Secale cereale*; Ta, *Triticum aestivum*; Vv, *Vitis vinifera*; Zm, *Zea mays*.

only a single cluster was found in rice among these four CRP classes that was completely composed of novel genes – a DEFL cluster of 11 novel tandem-duplicated genes on rice chromosome 10.

As a general rule, sequences belonging to a cluster of four or more CRPs within 100 000 bp in *Arabidopsis* belong to a subgroup that is not found (or at least not clustered) in rice, and vice versa. The one exception to this rule is the LTP-like group CRP4820 which is highly represented among both monocots and dicots (320 and

414 unigenes, respectively). *Arabidopsis* has 33 genes in this subgroup, which are circumscribed by boxes in Figure 2. Rice has 37 genes from this subgroup similarly marked in Figure 3. On average, each rice gene in CRP4820 encodes a protein that is 51% identical and 64% similar to its closest *Arabidopsis* homolog. In fact, an alignment of the encoded proteins between the monocot and dicot model species shows striking conservation in the mature peptide throughout this subgroup (Figure S3).

**Table 3.** Counts for taxon-specific sequence-related cysteine-rich peptide subgroups resembling plant lipid transfer proteins, thionins, and rapid alkanization factor-like sequences

| Taxonomic group | Subgroup count <sup>a</sup> |              |           | Unigene counts in taxon-specific subgroups <sup>b</sup> |              |           | EST counts in taxon-specific subgroups <sup>c</sup> |              |           |
|-----------------|-----------------------------|--------------|-----------|---------------------------------------------------------|--------------|-----------|-----------------------------------------------------|--------------|-----------|
|                 | LTP-like                    | Thionin-like | RALF-like | LTP-like                                                | Thionin-like | RALF-like | LTP-like                                            | Thionin-like | RALF-like |
| Poaceae         | 45                          | 21           | 16        | 865 (21)                                                | 224 (44)     | 221 (39)  | 16 273 (24)                                         | 4039 (68)    | 1463 (41) |
| Brassicaceae    | 12                          | 4            | 4         | 103 (3)                                                 | 77 (15)      | 48 (8)    | 196 (0.3)                                           | 247 (4)      | 202 (6)   |
| Coniferales     | 7                           | 0            | 1         | 123 (3)                                                 | –            | 9 (2)     | 8826 (13)                                           | –            | 42 (1)    |
| Fabaceae        | 6                           | 0            | 5         | 51 (1)                                                  | –            | 42 (7)    | 784 (1)                                             | –            | 245 (7)   |
| Solanaceae      | 3                           | 1            | 0         | 49 (1)                                                  | 20 (4)       | –         | 200 (0.3)                                           | 125 (2)      | –         |
| Asparagales     | 1                           | 0            | 0         | 7 (0.2)                                                 | –            | –         | 27 (0.04)                                           | –            | –         |
| Asteraceae      | 1                           | 0            | 1         | 5 (0.1)                                                 | –            | 4 (0.7)   | 56 (0.08)                                           | –            | 17 (0.5)  |
| Vitaceae        | 1                           | 0            | 2         | 10 (0.2)                                                | –            | 10 (2)    | 467 (0.7)                                           | –            | 34 (1)    |
| Salicaceae      | 0                           | 0            | 2         | –                                                       | –            | 39 (7)    | –                                                   | –            | 281 (8)   |

<sup>a</sup>Subgroups having four or more unigene sequences with  $\geq 95\%$  coming from the specified taxonomic family or order. Total subgroups in each class are: LTP-like, 128; thionin-like, 43; RALF-like, 43.

<sup>b</sup>Total number of unique sequences coming from the specified taxon-specific CRP subgroups. Unigene sequences were collected among the Institute for Genomic Research's (TIGR's) plant gene indices (GIs), Uniref100 and the Arabidopsis and *M. truncatula* genomes. The numbers in parentheses indicate the percentage of the total unigenes assigned to the specified CRP class shown in Table 1.

<sup>c</sup>Total number of expressed sequence tags (ESTs) in TIGR's plant GIs. The numbers in parentheses indicate the percentage of the total number of ESTs assigned to the specified CRP class shown in Table 1.

**Table 4.** Counts for morphological structure-specific cysteine-rich peptide (CRP) subgroups resembling plant lipid transfer proteins (LTPs), thionins, and rapid alkanization factor-like sequences (RALFs)

| Structural group                     | Subgroup count <sup>a</sup> |              |           | Unigene counts in taxon-specific subgroups <sup>b</sup> |              |                    | EST counts in taxon-specific subgroups <sup>c</sup> |              |                     |
|--------------------------------------|-----------------------------|--------------|-----------|---------------------------------------------------------|--------------|--------------------|-----------------------------------------------------|--------------|---------------------|
|                                      | LTP-like                    | Thionin-like | RALF-like | LTP-like                                                | Thionin-like | RALF-like          | LTP-like                                            | Thionin-like | RALF-like           |
| Reproductive <sup>§</sup>            | 45                          | 18           | 7         | 987 (24)                                                | 191 (38)     | 85 (15)            | 15 434 (23)                                         | 2214 (37)    | 384 (11)            |
| Seeds <sup>§</sup>                   | 9                           | 0            | 1         | 183 (5)                                                 | –            | 3 (0.5)            | 10 059 (15)                                         | –            | 146 (4)             |
| Roots                                | 1                           | 6            | 1         | 8 (0.2)                                                 | 41 (8)       | 9 <sup>*</sup> (2) | 446 (0.7)                                           | 144 (2)      | 39 <sup>*</sup> (1) |
| Above ground vegetative <sup>§</sup> | 0                           | 0            | 1         | –                                                       | –            | 23 (4)             | –                                                   | –            | 92 (3)              |

<sup>a</sup>Subgroups having four or more unigene sequences with  $\geq 95\%$  coming from the specified morphological structure groups. Total subgroups in each class are: LTP-like, 128; thionin-like, 43; RALF-like, 43.

<sup>b</sup>Total number of unique sequences coming from the specified structure-specific CRP subgroups. Unigene sequences were collected among the Institute for Genomic Research's (TIGR's) plant gene indices (GIs), Uniref100 and the Arabidopsis and *M. truncatula* genomes. The numbers in parentheses indicate the percentage of the total unigenes assigned to the specified CRP class shown in Table 1.

<sup>c</sup>Total number of expressed sequence tags (ESTs) in TIGR's plant GIs. The numbers in parentheses indicate the percentage of the total number of ESTs assigned to the specified CRP class shown in Table 1.

<sup>§</sup>These structures are itemized in **Experimental procedures**.

<sup>\*</sup>Root nodules also included.

## Discussion

*Despite differing arrangement and number of cysteines, the various classes of cysteine-rich peptides have many common features*

Regardless of class type, plant CRPs, including the DEFLs identified earlier (Silverstein *et al.*, 2005), share similarities in molecular properties, genome organization, taxonomic distribution, spatial expression and function. Nearly all have an N-terminal signal sequence and a small divergent charged or polar mature peptide with conserved cysteines (Figure 1,

Tables 1 and 2). In the Arabidopsis and rice genomes, genes encoding CRPs are clustered (Figures 2 and 3) and appear as a single exon or two exons with a conserved intron position. Cysteine-rich peptide sequence-related subgroups are over-represented in specific taxonomic families (Table 3) and in ESTs from reproductive structures (Table 4). Further, members from several distinct classes have been shown to have functions related to antimicrobial activity (Broekaert *et al.*, 1997), protease inhibition (Bloch and Richardson, 1991; Melo *et al.*, 2002; Odani *et al.*, 1983), allergenic cross-reactivity (Himly *et al.*, 2003; van Ree, 2002), and reproductive regulation (Goldman *et al.*, 1994; Schopfer *et al.*, 1999).

**Table 5.** Tally of apparent CRP fusions with PRPs, GRPs, and other CRPs.

| Observed pattern <sup>a</sup> | No. subgroups affected | No. unique sequences |
|-------------------------------|------------------------|----------------------|
| <b>SP – CRP – PRP</b>         | 33                     | 1269                 |
| <b>SP – CRP – CRP</b>         | 23                     | 329                  |
| <b>SP – PRP – CRP</b>         | 18                     | 88                   |
| <b>SP – CRP – CRP'</b>        | 16                     | 65                   |
| <b>SP – CRP* GRP CRP*</b>     | 13                     | 361                  |
| <b>SP – GRP – CRP</b>         | 6                      | 72                   |
| <b>SP – CRP* PRP CRP*</b>     | 5                      | 76                   |
| <b>SP – PRP – CRP – PRP</b>   | 4                      | 70                   |
| <b>SP – CRP – GRP</b>         | 3                      | 14                   |

<sup>a</sup>Schematic depiction of fusion type. SP, Signal Peptide; CRP, Cysteine-Rich peptide; PRP, Proline-Rich peptide; GRP, Glycine-Rich peptide; CRP' indicates a distinct cysteine-rich domain of a different class; CRP\* indicates an interrupted domain.

*The maternally expressed gene family is notable for its under-annotation in the Arabidopsis and rice genomes*

In this work, we have identified several poorly characterized CRP classes (Tables 1 and 2). Among these, the MEG family (Gutierrez-Marcos *et al.*, 2004) stands out because 13 of the 17 Arabidopsis and 11 out of the 27 rice family members were previously unannotated. We verified the expression of six novel Arabidopsis MEG-related genes, and the MPSS data suggest that two more of them are specifically expressed in reproductive structures (Table S2). The function of this moderately sized diverse family is unknown, but the class includes a family (CRP5495 in Table S1) of basal layer antifungal peptides (Serna *et al.*, 2001) that is specific to the same maize basal endosperm transfer layer cells as *MEG1* (Gutierrez-Marcos *et al.*, 2004). The latter antifungal peptides are lacking one and sometimes two pairs of cysteines relative to other members in the class, but otherwise have a conserved cysteine spacing pattern and similar morphological structure specificity (maternal endosperm).

*Most cysteine-rich peptide classes have members fused to proline-rich, glycine-rich, or similar cysteine-rich peptides*

Subgroups of nearly every CRP class have an N-terminal (just downstream of the signal peptide), C-terminal, or internal PRP or GRP domain (often alternating with serine, threonine, asparagine, or glutamine) (Table 5, Table S1). Chimeric PRP–CRP and GRP–CRP sequences have been noted before (Broekaert *et al.*, 1997; Cassab, 1998; Silverstein *et al.*, 2006) for a few individual CRPs, but the widespread occurrence of such fusions throughout the CRP classes is noteworthy. Some have speculated that these chimeras may be anchored to the plant cell wall like other characterized PRPs and GRPs, where they may act as sentinels or direct microbicides when proteolytically

cleaved by predators (Broekaert *et al.*, 1997; Cassab, 1998; Silverstein *et al.*, 2006).

In addition to PRP and GRP fusions, internally duplicated and fused CRP motifs occur in almost every class (Table S1). Previously, defensin–defensin fusions have been noted (Thomma *et al.*, 2002). In Table S1 there are even cases of apparent fusions between distinct classes of CRPs. For example, subgroups CRP2230–CRP2290 appear to be defensin–thionin fusions. They have an N-terminal CS $\alpha\beta$  motif (C...CXXXC...C...CXC; X = any amino acid; ... = variable number of amino acids) characteristic of defensins (Lay *et al.*, 2003; Zhu *et al.*, 2005) fused to a C-terminal thionin motif (CXXXCXXXC...CXXXCXXXC). Several 3D structures have been determined for the N-terminal portion of these thionins by NMR and X-ray crystallography, as well as for plant defensins. Purothionin A-1 [SwissProt ID P01543; Protein Data Bank (PDB) code 1bhp] from subgroup CRP2260 is an example of a thionin with the N-terminal CS $\alpha\beta$  motif (Stec *et al.*, 1995) while NaD1 (SwissProt ID Q8GTM0; PDB code 1mr4) is an example of a typical plant defensin (Lay *et al.*, 2003). Both of these structures begin with a beta strand spanning the first Cys residue followed by a helix that spans the CXXXC portion of the motif, and end in a beta strand that spans the CXC region. The only topological difference between the N-terminal thionin domain and the defensin is the substitution of a helix in the thionin for the defensin beta strand that spans the variable-spaced Cys residue between the CXXXC and CXC motifs.

*Plant cysteine-rich peptide classes may be evolving by rearrangements of cysteine motifs*

The diversity of CRPs in plants is immense. Even within some sequence-related subgroups there are clear examples where individual sequences within an alignment have undergone internal duplication (e.g. At5g04045 and BnCD819151 in Figure 4) followed by subsequent loss of a pair of cysteines (e.g. compare CRP0675 with CRP0670 or compare the members within CRP1120). Additionally, there are many examples where some sequences within an alignment have undergone rearrangements of their cysteines: one example involving DEFL-class peptides is shown in Figure 4. Another extreme example is provided in Figure S4, where the complete repertoire of 30 closely-related seed-storage prolamins in rice (found only at four distinct genomic loci) have anything from 0 to 13 cysteines. Such rearrangements led to the creation of a class of scorpion toxins with novel functions from ancestral defensins within the framework of the CS $\alpha\beta$  structural motif (Zhu *et al.*, 2005). A small, compact structure with multiple cysteine bridges provides a thermally stable framework (Creighton, 1993) that is robust to changes in primary sequence (Zhu *et al.*, 2005), and is highly evolvable (Bloom *et al.*, 2006).

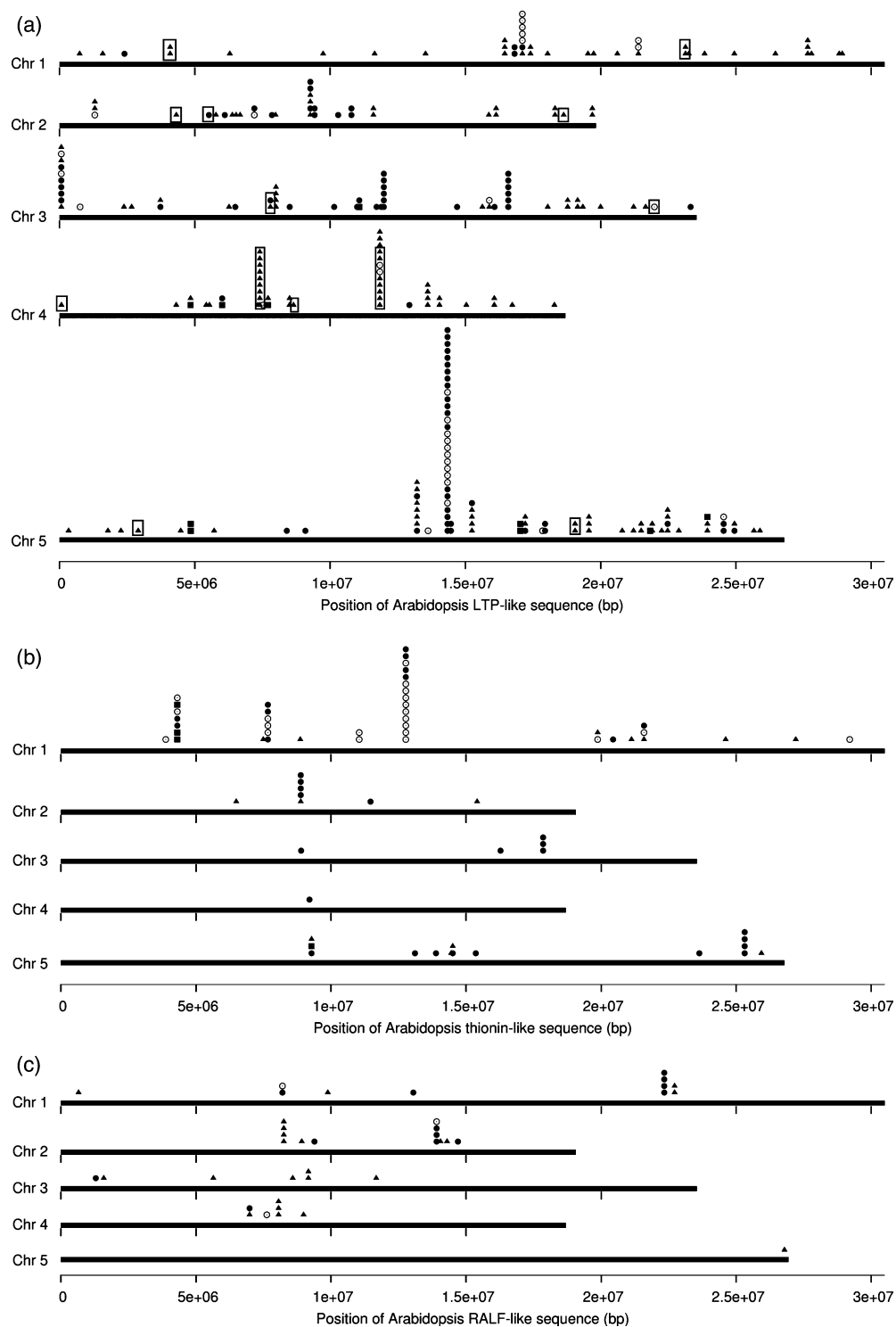


**Figure 2.** Clustering and prediction status of cysteine-rich peptide (CRP) sequences in the Arabidopsis genome.

(a) Lipid transfer protein-like sequences.

(b) Thionin-like sequences.

(c) Rapid alkanization factor-like sequences.

Each symbol represents a single CRP-encoding gene. Symbols falling within a 100 000-bp window are stacked vertically. The type of symbol indicates the status of the gene in the Arabidopsis Information Resource annotation version 6: ▲, expressed; ■, annotated without expression evidence; ●, unannotated; ○, unannotated with expression verified in this work. Symbols circumscribed by a box belong to the prolific proline-rich protein-LTP fusion subgroup CRP4820 discussed in the text.

We have observed that the various classes of CRPs are subject to frequent internal duplications and cysteine rearrangements (Figure 4, Table S1). It is possible that some of the CRP classes have evolved from a common ancestor. Indeed, some sequences score well against both thionin and GASA/GAST HMMs (e.g. GmTC225453, GmTC225456, McBE036009, PvTC1686,  $E$ -value  $< 10^{-5}$ ). However, this is questionable because distinct CRP classes often have quite different 3D folds (Broekaert *et al.*, 1997), and physical constraints could be at play forcing convergently evolved similarities.

*Differences exist in the cysteine-rich peptide repertoire of the model monocot rice and the model dicot Arabidopsis*

Arabidopsis and rice have a very different distribution of gene counts among the 24 primary CRP classes (Table 1). Such differences between the monocot and the dicot might be expected given the large evolutionary distance between the two species, and the observation that highly similar CRPs have expanded in tandem arrays within each genome (Figures 2 and 3). Nevertheless, we did observe one subgroup within the LTP-like class encoding LTP-PRP fusion peptides, CRP4820, that is highly conserved between the two species (average per cent similarity = 64% at the protein level; Figure S3). Further, members of this subgroup are arrayed in multiple clusters throughout both genomes (Figures 2 and 3).

The Arabidopsis genome was found to have 38% more CRP genes than the rice genome. This over abundance of CRPs in Arabidopsis is surprising since we have observed a large number of Poaceae-specific CRP subgroups (200 subgroups in Table S1), each comprise a large number of unigene sequences (Table 3, Table S1). It is possible that this apparent contradiction may be explained by the multiple counting of unigenes within TIGR's gene indices. Among the gene indices, there is often a one-to-many relationship between genomic loci and unigenes since transcripts from alternative splicings, different alleles, and mixed cultivars often do not contig together. For example, Arabidopsis has 62 010 unigenes in AGI 12.1, yet there are only 30 690 genes in the TAIR6 annotation of the genome. Wheat has an astounding 122 282 unigenes in TaGI 10.0, and it is unlikely to have that many genes. The polyploid nature of many of the Poaceae species for which ESTs are available (including hexaploid wheat) probably exaggerates the CRP counts for those species.

Another plausible explanation for the observed under-representation of CRPs in rice may simply be that rice was searched at the end of this study, rather than iteratively like Arabidopsis. Perhaps if we had included the novel rice CRPs in the HMMs, rebuilt the models, and iterated until convergence, we would have found additional genes. Although we cannot discount this possibility, we find it highly unlikely,

especially in accounting for the more than threefold difference observed for the number of DEFLs in rice and Arabidopsis. Rice was highly represented in the iterative phase of the search via the Rice Gene Index, which had 274 018 ESTs collapsed into 89 147 unigenes. Further, Poaceae in general were well represented during the iterations by seven species in the gene indices collectively including 2 084 178 ESTs coalesced into 443 506 unigenes (Table S7).

*The accuracy of the cysteine-rich peptide predictions is likely to be higher than those obtained from generic gene-prediction algorithms*

Several independent lines of evidence suggest that the Arabidopsis and rice CRP predictions reported here are probably more reliable than predictions made by *ab initio* gene-prediction algorithms (i.e. those that do not incorporate knowledge gleaned from related genes): (i) The products of these genes were aligned in families that often included expressed gene products from other species (e.g. At2g20605 and At2g20619 in Figure 1b). In these alignments, the signal peptide, cysteine positions, splice sites, and size of the gene products were all well conserved, giving us confidence in our gene predictions. (ii) Despite the fact that our search algorithm only targeted the mature peptide region, we were able to find a nearby upstream region that encoded a signal peptide, further bolstering our confidence. (iii) In our attempt to verify a subset (117) of our Arabidopsis DEFL and non-DEFL CRP genes via RT-PCR, we recovered 59% of them. This is remarkably high, considering that these genes are each known to have expression patterns restricted to specific morphological structures (typically seeds, pollen, or other reproductive structures), and yet our RNA pool was from a mixture mostly devoid of reproductive structures (less than 10% of the total pool). (4) Many additional predictions have verified expression in the MPSS database (Meyers *et al.*, 2004). (All those marked 'T' in the column 'MPSS Support?' in Table S2.) Our cross-references to this database were very conservative, as we only specified a 'T' in that column if it overlapped the actual coding sequence (CDS) and if the match was verified by multiple MPSS sequencing runs.

*Genome annotation is incomplete: family-specific searches should be performed routinely to find superfamilies of small divergent genes*

In this work we identified 196 unannotated CRPs in Arabidopsis plus 91 novel CRPs in rice. These are in addition to the ~200 unannotated DEFLs reported earlier (Silverstein *et al.*, 2005). Yet the Arabidopsis genome sequence has been nearly complete for several years. Since the original publication of the genome sequence where 25 498 genes were reported (The Arabidopsis Genome Initiative, 2000),

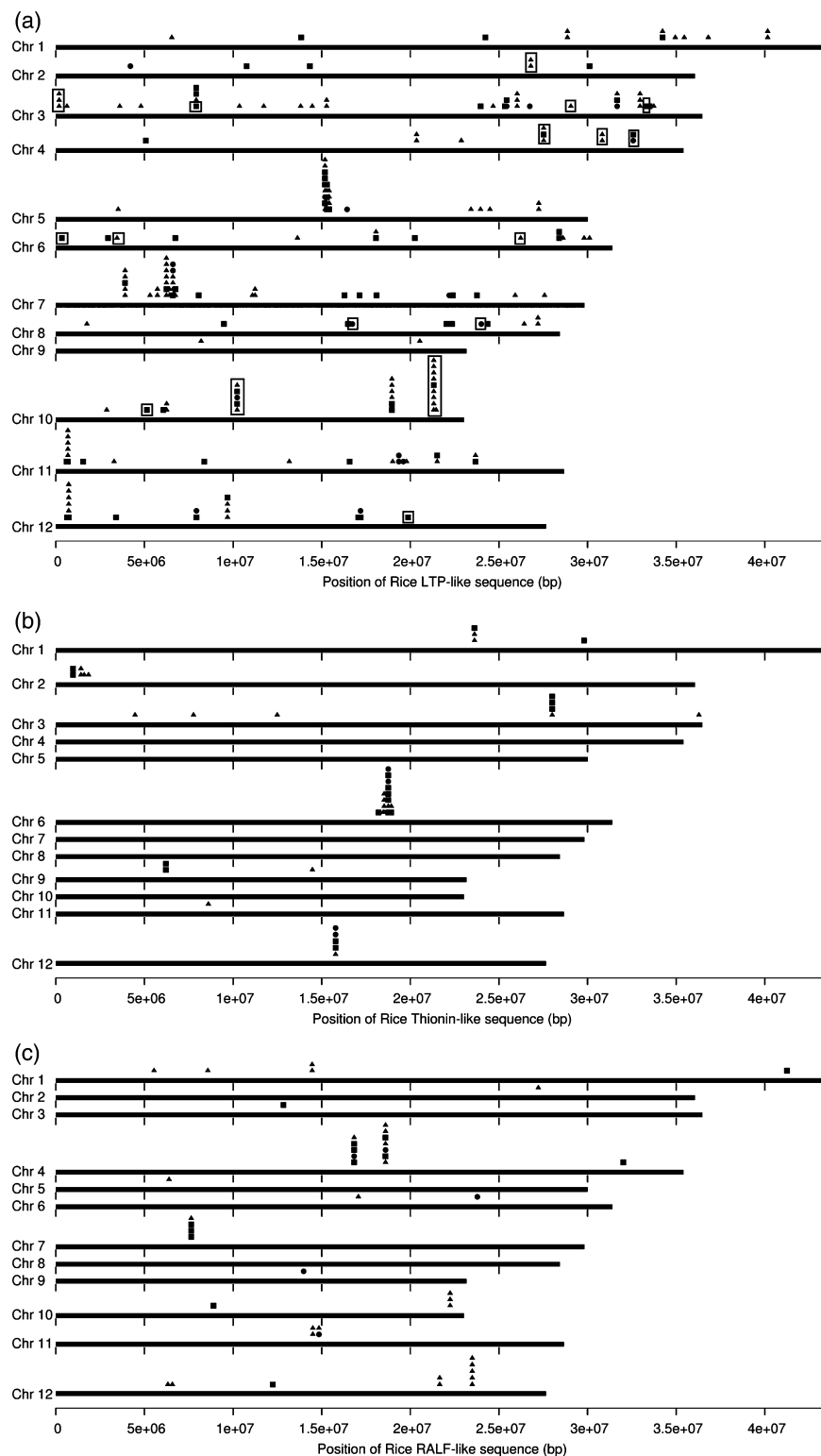


**Figure 3.** Clustering and prediction status of cysteine-rich peptide (CRP) sequences in the rice genome.

(a) Lipid transfer protein-like sequences.

(b) Thionin-like sequences.

(c) Rapid alkanization factor-like sequences.

Each symbol represents a single CRP-encoding gene. Symbols falling within a 100 000-bp window are stacked vertically. The type of symbol indicates the status of the gene in the Institute for Genomic Research rice genome annotation version 4: ▲, expressed; ■, annotated without expression evidence; ●, unannotated. Symbols circumscribed by a box belong to the prolific proline-rich protein-LTP fusion subgroup CRP4820 discussed in the text.

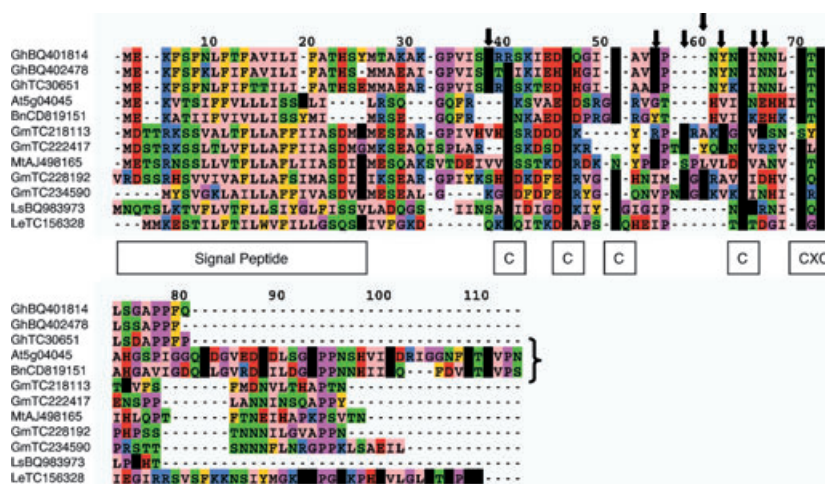


**Figure 4.** Example of a cysteine-rich peptide subgroup that displays evidence for cysteine re-arrangements and internal duplications. Representative sequences from CRP1110 were aligned with CLUSTALW with minor manual adjustments and visualized via Jalview with the Zappo color scheme. Cysteines are shown as black filled boxes for emphasis. The signal peptide region and conserved class-specific cysteine residues are displayed schematically below the alignment. Thick black arrows at the top of the alignment indicate columns where one or more sequence members have an additional cysteine residue relative to the consensus. A curly brace highlights two sequence members that have a clear repetition of the conserved cysteine motif. A single sequence in the alignment has additional cysteines in a novel arrangement (underlined here).

the number of annotated genes has steadily increased as large-scale experimental verification efforts have identified novel genes. In the intervening years, whole-genome tiling arrays (Yamada *et al.*, 2003), full-length cDNA sequencing (Moskal *et al.*, 2007; Wortman *et al.*, 2003; Xiao *et al.*, 2005), MPSS sequencing (Meyers *et al.*, 2004) and comparisons to draft sequences of a closely related species (Ayele *et al.*, 2005) have each provided evidence for hundreds to thousands of previously unannotated genes. The number of annotated, protein-encoding genes now stands at 30 690 (TAIR6) – 20% more than originally reported in 2000. The final Arabidopsis genome annotation at TIGR (release 5) was completed in 2004 (Haas *et al.*, 2005). The first TAIR release (TAIR6) added only expressed sequence-based and community-supplied structural updates. Genome annotation is an ongoing task and there are probably many more small, divergent genes that remain to be discovered in this and other genomes. More effort should be focused on experimentally identifying genes of this class, such as the proteomic effort recently carried out in *Medicago truncatula* (Zhang *et al.*, 2006).

One relevant feature of the Arabidopsis annotation pipeline is that small genes predicted by gene-finding algorithms (encoding proteins <110 amino acids in length) that lacked explicit expression verification were screened out to avoid abundant false-positive small-gene predictions (Haas *et al.*, 2005). No explicit size cut-off was imposed by the annotators of the rice genome (Ouyang *et al.*, 2007; <http://www.tigr.org/tdb/e2k1/osa1/>), and that probably explains why a higher percentage of genes that we identified in rice were already annotated (Figures 2 and 3; Tables S2 and S6). It is import-

ant to note, however, that the inclusiveness of the rice genome annotation comes at a price – it contains a large number of annotated genes (62 827), of which only 50% have expression evidence. In contrast the Arabidopsis annotation has relatively fewer genes (30 690) with higher confidence (76% with expression evidence).

Our procedure for generating and searching iterative sequence motifs is an effective means of identifying whole families of related genes when applied to both expressed and genomic sequence databases. Initially we demonstrated the effectiveness of this procedure on the superfamily of defensin-like genes (Silverstein *et al.*, 2005), but have expanded it here with similar success to a larger set of CRPs. Others have demonstrated the utility of iterative family searches for several families of small divergent genes (Ride *et al.*, 1999; Vanoosthuysen *et al.*, 2001; Cock and McCormick, 2001). Thus, the time appears right for annotation groups to apply similar techniques more systematically to identify many of the missing genes in genomes.

## Experimental procedures

Upon request, all novel materials described in this publication will be made available in a timely manner for non-commercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining any permission will be the responsibility of the requestor. The cloned sequences reported in this paper have been deposited in the GenBank database. Identified CRP sequences from the Arabidopsis genome were provided to TAIR and AGI gene codes were assigned. Supplementary Data are provided online.

### Iterative motif searching

The search procedure was essentially as described previously (Silverstein *et al.*, 2005). Starting with an initial set of 108 subgroup HMMs enumerated below, we scanned TIGR's 33 plant gene indices, the near-complete Arabidopsis genome, and all available *Medicago* genomic BAC sequences using *hmmsearch* from the HMMer 2.3.2 package (Durbin *et al.*, 1998) with default parameters. Genomic sequences were split into 2000-bp fragments such that each genomic fragment overlapped the previous fragment by 1000 bp. All sequences were translated in six frames prior to search.

All identified hits among the expressed sequences with  $E$ -value < 10 were manually examined and candidates for retention were selected if they met the following criteria: (i) they had an N-terminal signal peptide and the C-terminal cysteine-rich portion was 110 amino acids or fewer, or (ii) they clearly matched the cysteine arrangement spanning a subregion of one of the multiple sequence alignments from which the HMMs were based. Genomic hits with  $E$ -value < 10 were manually examined with similar criteria, except that it was necessary to scan up to 1200 bp (400 amino acids in the translation in all three same sense frames) upstream of the hit region for an N-terminal signal peptide.

Candidate hits were assigned to the HMM that had the best score (i.e. lowest  $E$ -value) and added to the sequence list for that subgroup. Once all new sequences were added to each model, they were realigned via CLUSTALW version 1.83 (Thompson *et al.*, 1994), and the resulting alignment was manually inspected and adjusted via Jalview (Clamp *et al.*, 2004). If the alignment was clearly heterogeneous, the subgroup was divided into two or more new subgroups according to the dendrogram obtained by CLUSTALW. Once all new hits were assigned to the appropriate model, all alignments were trimmed to the conserved mature peptide region (typically from first to last cysteine residue). New HMMs corresponding to the trimmed alignments were generated via *hmmbuild* -g and *hmmcalibrate* from the HMMer 2.3.2 package (Durbin *et al.*, 1998). These new HMMs were used to scan the target databases for the next iteration. Iterations were terminated when no new significant hits were found.

After termination of the iterative search procedure, all HMMs were scanned against the UniProt comprehensive protein dataset (Apweiler *et al.*, 2004). UniProt hits from plants only were added to the best-matched model using the criteria described above for expressed sequences. At this point, all collected CRPs were compared against themselves via BLASTN with options -F F -e 1e-20 (Altschul *et al.*, 1997) and redundant sequences (those with 100% identity over their entire length to an equal or larger-sized CRP) were removed while retaining their identifiers in the FASTA definition line of the equivalent retained sequence. Finally, sequences from each model were realigned, the alignments were trimmed, and the HMMs were rebuilt. As a self-consistency check, all CRPs were scanned against the final models via *hmmpfam* to verify that they matched their assigned model better than any other. Sequences that matched another model were reassigned, and the two affected models were rebuilt. Exceptions to this latter procedure included: (i) poor quality sequences with Ns and frameshift errors, (ii) incomplete sequences that visually fit with the assigned model but score better against a related truncated alignment, (iii) some sequences belonging to highly related and overlapping hevein/chitinase, glutenin, or prolamin models that are marked in Table S2.

The entire translated rice genome, split into overlapping 2000-bp fragments, was scanned against the final 516 HMMs via *hmmpfam* (Durbin *et al.*, 1998) using default parameters. Hits to the genome were manually inspected and assigned to the subgroup model with the best  $E$ -value as described above. A parallel search against the collected TIGR Rice 4.0 CDS sequences was made, and hits

corresponding to the same genomic loci were coalesced, with differences in gene structure noted.

### Initial motif models

The starting HMMs included DEFL subgroups 31.01–31.11, thionin-like subgroup 38, RALF-like subgroups 427, 560, and 655, and leginsulin subgroup 666, all initially identified in a search among ESTs for legume-specific genes (Graham *et al.*, 2004), plus 92 DEFL subgroups CRP0000–CRP1520 identified in subsequent work (Silverstein *et al.*, 2005).

### Target sequence databases

Target databases for *hmmsearch* included the unigene sequences from all 33 plant gene indices at TIGR (Quackenbush *et al.*, 2001), the comprehensive uniref100 collection (version 2.7) of known protein sequences (Apweiler *et al.*, 2004), the TAIR version 6 Arabidopsis genome (Haas *et al.*, 2005; Rhee *et al.*, 2003), the TIGR version 4 rice genome (Ouyang *et al.*, 2007) and the emerging genomic sequence data from the *M. truncatula* sequencing project (August, 2005) (Young *et al.*, 2005). The TIGR plant gene indices used are itemized in Table S7. The total unigene counts and ESTs among all included TIGR GIs were 1 083 977 and 4 801 711, respectively.

### Morphological structure category assignment

In determining the structural specificity of identified subgroups the following terms were considered. Reproductive structures (1 494 949 ESTs) included: aleurone, anther, carpel, caryopsis, coleoptile, cone, crown, ear (maize), egg, embryo, endosperm, fiber (cotton), floret, flower, fruit, glume, grain, head (wheat), inflorescence, kernel, maternal, ovary, ovule, panicle, pedicel, pericarp, pistil, pod, pollen, scutellum, seed, silk (maize), silique, sperm cell, spike, strobili, tassel. Seed structures (642 525 ESTs) included: aleurone, caryopsis, coleoptile, egg, embryo, endosperm, fiber (cotton), grain, kernel, pod, seed, and silique. Above-ground vegetative structures (1 468 315 ESTs) included: apex, apical meristem, cotyledon, culm, epidermis, epicotyl, hypocotyls, leaves, needle, petiole, phloem, rosette, sheath, shoots, stalk, stems, trichome, and vegetative.

Morphological structures of origin for all EST libraries among TIGR's 33 plant gene indices were determined by careful examination of the GenBank records, and merging this extracted information with the library descriptions at TIGR. The total number of ESTs from identifiable structures was 4 002 137. The compiled list of structural origins is available upon request.

### Statistical significance of taxon and structure specificity for cysteine-rich peptide subgroups

We used the Fisher exact probability test (Siegel, 1956) to determine whether sequence subgroups were statistically enriched by unigene sequences from a given taxonomic species, family or order. Detailed computations for all subgroups in Table S1 having at least 75% of the unigene sequences derived from a single taxonomic group are provided in Table S8. Although we have calculated chi-squared probabilities alongside the Fisher probabilities in Table S8 for comparison, only the exact Fisher values are reported in the subgroup summary table, Table S1. No statistical assessment was made for subgroups apparently specific to the Brassicaceae or

Fabaceae, since the total number of independently sampled *Arabidopsis* and *Medicago* sequences is undefined due to the inclusion of their entire respective translated genomes. Statistics were computed for Poaceae since the rice genomic sequence was searched after the completion of the study (the results of which are not reflected in the counts in Tables S1 and S8).

The probability calculations to assess structural specificity were similar to those above, replacing unigene counts with EST counts. We made three separate assessments of the Fisher exact probabilities for the structural specificity of each subgroup, and report the most conservative (i.e. highest *P*-value) among the three in Table S1. Detailed probability calculations for each of the three sets of computations are provided as separate sheets in Table S8. The first calculation uses global counts for each structural group in the TIGR GLs, and the total number of ESTs from identifiable structures at TIGR. In the second approach, we break down the counts by taxonomic family (or order in a few select cases). In the third approach, counts are broken down by individual species. Additional details and motivation for these statistical procedures are provided in the Appendix S1.

### Calculation of molecular properties

Several molecular properties were predicted for *Arabidopsis* CRPs and recorded in Table S2. Splice sites for gene prediction were predicted using the NetPlantGene server (Hebsgaard *et al.*, 1996) and via consensus with aligned closely related expressed genes in the same subgroup. Upstream signal peptides were confirmed via the SignalP 3.0 server (Bendtsen *et al.*, 2004). Theoretical isoelectric points (pI) and molecular weights (MW) were computed using ExPASy's pI/MW tool (Gasteiger *et al.*, 2005).

### Primer design

The PCR primers were generated for 40 DEFL and 77 non-DEFL *Arabidopsis* predictions using a primer design pipeline that ensured high throughput and specificity. First, the software exploited the capabilities of MIT primer3 (Rozen and Skaletsky, 2000) to design hundreds of primer set candidates that fit experimental requirements (minimum/optimal/maximum  $T_m = 65/70/71^\circ\text{C}$ ) for each of the 117 target gene predictions. Next, candidate primer sequences were aligned via WU-BLAST (Gish, <http://blast.wustl.edu>) to the *Arabidopsis* CDS library (Haas *et al.*, 2005) in order to discard those with secondary hits reporting 80% or higher identity. Finally, primer sets were sorted by specificity, self-complementarity, and primer3 penalty scores to achieve selection of the best primers for every target. The primer design pipeline was implemented in object-oriented Perl modules and hosted in a relational MySQL database. Designed primer sequences and their predicted  $T_m$  values are given in Table S9.

### Synthesis of complementary DNA

*Arabidopsis* SMART cDNA populations, one each for 5' and 3' RACE reactions, were generated as per the manufacturer's protocol (Clontech, <http://www.clontech.com/>) using pools of RNA from a range of morphological structures and treatments that included heat shock, cold shock, young plant, *Xanthomonas*-infected, *Pseudomonas*-infected, tissue culture, inflorescence, roots, 2,4-dichlorophenoxy acetic acid, salt stress, indole-3-acetic acid, UV and  $\text{H}_2\text{O}_2$  as described earlier (Xiao *et al.*, 2005), with the addition of whole inflorescences lacking developed siliques to the pool. Inflorescences made up approximately 1/14 of the total RNA in the pool.

### Polymerase chain reactions

The RACE conditions (per 25  $\mu\text{l}$  reaction) were: 2.5  $\mu\text{l}$  10 $\times$  PCR buffer, 0.5  $\mu\text{l}$  100 mM dNTP mix, 0.5  $\mu\text{l}$  PCR Advantage 2 Polymerase mix (Clontech), 0.5  $\mu\text{l}$  10  $\mu\text{M}$  adapter/vector primer, 4  $\mu\text{l}$  1.25  $\mu\text{M}$  gene-specific primer, 0.5  $\mu\text{l}$  template (BD SMART 5' or 3' RACE-ready cDNA)

### Rapid amplification of cDNA ends

Program RACE1 on MJ Research Tetrad: 94.0 $^\circ\text{C}$  for 30 sec, followed by five cycles of 94.0 $^\circ\text{C}$  for 5 sec, 72.0 for 4 min, followed by five cycles of 94.0 $^\circ\text{C}$  for 5 sec, 70.0 $^\circ\text{C}$  for 4 min, followed by 25 cycles of 94.0 $^\circ\text{C}$  for 5 sec, 68.0 $^\circ\text{C}$  for 4 min, a final elongation at 68.0 $^\circ\text{C}$  for 5 min before holding at 4.0 $^\circ\text{C}$ .

### Subcloning/sequencing

The RACE products were immediately subcloned into the pCR4.1-TOPO-TA vector and used to transform chemically competent 'OneShot' *Escherichia coli* (Invitrogen, <http://www.invitrogen.com/>). Twelve colonies from each transformation were selected and sequenced.

### The PCR/RACE product analysis

Sequences were trimmed of Poly(A) tails, clustered and mapped to the TIGR5 annotation using PASA (Haas *et al.*, 2003) that provides a graphical user interface to visualize and examine the assembled sequences and their locations within the genome. The best assemblies were forwarded to our Gateway open reading frame (ORF) cloning pipeline.

### Full-length sequencing

The open reading frames (ORFs) of each RACE sequence assembly were targeted for ORF cloning. The PCR primers were designed by software that walked 14–35 nucleotides upstream and downstream from the ATG start codon and stop codon positions, respectively, to yield primers with a  $T_m$  value of 60 $^\circ\text{C}$ . The ORFs were amplified and recombinationally cloned into the Gateway<sup>®</sup> pDONR221 as described previously (Underwood *et al.*, 2006). Eight clones from each transformation were selected and sequenced with sequencing primers JX (5'-TAAAACGACGGCCAGTCTTAAGCTC-3') and JZ (5'-ACTCACTATAGGGGATATCAGCTGGATG-3') at the JCVI Joint Technology Center (Rockville, MD, USA). Sequences were trimmed and assembled with TIGR Assembler (Sutton *et al.*, 1995). Sequence assemblies were verified against gene-specific primers and aligned to the RACE target with SIM4 (Florea *et al.*, 1998) for alignment inspection.

### Acknowledgements

The authors wish to thank the Minnesota Supercomputer Institute (MSI) for computer use, which significantly accelerated the iterative HMM search procedure. We also gratefully acknowledge David Swarbreck at TAIR, plus Matthew Campbell and Haining Lin at TIGR for carefully inspecting each of our *Arabidopsis* and rice predictions and assigning official locus identifiers. The work at the University of Minnesota was supported by National Science Foundation award IOB-0516811 and the University of Minnesota College of Biological

Sciences, and at TIGR by the National Science Foundation (grant no. DBI-0312656).

## Supplementary Material

The following supplementary material is available for this article online:

**Table S1.** Properties of cysteine-rich peptide sequence-related subgroups among plants.

**Table S2.** Properties of cysteine-rich peptides in *Arabidopsis thaliana*.

**Table S3.** Intron characteristics for cysteine-rich peptides in *Arabidopsis thaliana*.

**Table S4.** External database cross-references for cysteine-rich peptides in *Arabidopsis thaliana*.

**Table S5.** Massively parallel signature sequencing matches for cysteine-rich peptides in *Arabidopsis thaliana*.

**Table S6.** Properties of cysteine-rich peptides in *Oryza sativa*.

**Table S7.** The Institute for Genomic Research plant gene indices scanned for cysteine-rich peptide sequences.

**Table S8.** Statistical computations for cysteine-rich peptide sequence-related subgroup taxon and morphological structure specificity.

**Table S9.** Primer design characteristics for cysteine-rich peptides attempted in this study.

**Figure S1.** Venn diagram describing the expression verification status of *Arabidopsis* cysteine-rich peptides.

**Figure S2.** Clustering and prediction status of defensin-like sequences in the rice genome.

**Figure S3.** Protein alignment showing the mature peptide of CRP4820 members, which encode lipid transfer protein-proline-rich protein fusion peptides that are highly conserved between *Arabidopsis* and rice.

**Figure S4.** Alignment of all rice prolamin sequences, providing evidence of vast differences in cysteine content among close paralogs.

## Appendix S1 Supplementary methods

**Data S1** FASTA files for *Arabidopsis* cysteine-rich peptide sequences (genomic, coding sequence, peptide)

**Data S2** FASTA files for rice cysteine-rich peptide sequences (genomic, coding sequence, peptide)

**Data S3** FASTA files for plant cysteine-rich peptide sequences (coding sequence, peptide)

**Data S4** Motif models and alignments for plant cysteine-rich peptide sequence groups (WIN-ZIP archive, MAC/UNIX tar-gzip archive)

This material is available as part of the online article from <http://www.blackwell-synergy.com>

## References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- Alvarez, M.E., Pennell, R.I., Meijer, P.J., Ishikawa, A., Dixon, R.A. and Lamb, C. (1998) Reactive oxygen intermediates mediate a systemic signal network in the establishment of plant immunity. *Cell*, **92**, 773–784.
- Apweiler, R., Bairoch, A., Wu, C.H. et al. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119.
- Aubert, D., Chevillard, M., Dorne, A.M., Arlaud, G. and Herzog, M. (1998) Expression patterns of GASA genes in *Arabidopsis*

*thaliana*: the GASA4 gene is up-regulated by gibberellins in meristematic regions. *Plant Mol. Biol.* **36**, 871–883.

- Ayele, M., Haas, B.J., Kumar, N., Wu, H., Xiao, Y., Van Aken, S., Utterback, T.R., Wortman, J.R., White, O.R. and Town, C.D. (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*. *Genome Res.* **15**, 487–495.
- Bauw, G., Nielsen, H.V., Emmersen, J., Nielsen, K.L., Jorgensen, M. and Welinder, K.G. (2006) Patatins, Kunitz protease inhibitors and other major proteins in tuber of potato cv. Kuras. *FEBS J.* **273**, 3569–3584.
- Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795.
- Berrocal-Lobo, M., Segura, A., Moreno, M., Lopez, G., Garcia-Olmedo, F. and Molina, A. (2002) Snakin-2, an antimicrobial peptide from potato whose gene is locally induced by wounding and responds to pathogen infection. *Plant Physiol.* **128**, 951–961.
- Birk, Y. (1985) The Bowman-Birk inhibitor. Trypsin- and chymo- trypsin-inhibitor from soybeans. *Int. J. Pept. Protein Res.* **25**, 113–131.
- Bloch, C. Jr and Richardson, M. (1991) A new family of small (5 kDa) protein inhibitors of insect alpha-amylases from seeds of sorghum [*Sorghum bicolor* (L.) Moench] have sequence homologies with wheat gamma-purothionins. *FEBS Lett.* **279**, 101–104.
- Bloom, J.D., Labthavikul, S.T., Otey, C.R. and Arnold, F.H. (2006) Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA*, **103**, 5869–5874.
- Boman, H.G. (2003) Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.* **254**, 197–215.
- Brisson, L.F., Tenhaken, R. and Lamb, C. (1994) Function of oxidative cross-linking of cell wall structural proteins in plant disease resistance. *Plant Cell*, **6**, 1703–1712.
- Broekaert, W.F., Cammue, B.P.A., De Bolle, M.F.C., Thevissen, K., De Samblanx, G.W. and Osborn, R.W. (1997) Antimicrobial peptides from plants. *Crit. Rev. Plant Sci.* **16**, 297–323.
- Bulet, P., Stocklin, R. and Menin, L. (2004) Anti-microbial peptides: from invertebrates to vertebrates. *Immunol. Rev.* **198**, 169–184.
- Cano-Delgado, A., Penfield, S., Smith, C., Catley, M. and Bevan, M. (2003) Reduced cellulose synthesis invokes lignification and defense responses in *Arabidopsis thaliana*. *Plant J.* **34**, 351–362.
- Cassab, G.I. (1998) Plant cell wall proteins. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 281–309.
- Chen, K.C., Lin, C.Y., Kuan, C.C., Sung, H.Y. and Chen, C.S. (2002) A novel defensin encoded by a mungbean cDNA exhibits insecticidal activity against bruchid. *J. Agric. Food Chem.* **50**, 7258–7263.
- Clamp, M., Cuff, J., Searle, S. and Barton, G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
- Cock, J.M. and McCormick, S. (2001) A large family of genes that share homology with CLAVATA3. *Plant Physiol.* **126**, 939–942.
- Creighton, T.E. (1993) *Proteins: structures and molecular properties*. 2nd edn. New York, NY: W. H. Freeman and Company.
- Dangl, J.L. and Jones, J.D. (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.
- Darvill, A.G. and Albersheim, P. (1984) Phytoalexins and their elicitors – a defense against microbial infection in plants. *Annu. Rev. Plant Physiol.* **35**, 243–275.
- Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.
- Durrant, W.E. and Dong, X. (2004) Systemic acquired resistance. *Annu. Rev. Phytopathol.* **42**, 185–209.
- Endo, Y., Mitsui, K., Motizuki, M. and Tsurugi, K. (1987) The mechanism of action of ricin and related toxic lectins on eucary-

- otic ribosomes. The site and the characterization of the modification in 28S ribosomal RNA caused by the toxins. *J. Biol. Chem.* **262**, 5908–5912.
- Farmer, E.E. and Ryan, C.A. (1990) Interplant communication: airborne methyl jasmonate induces synthesis of proteinase inhibitors in plant leaves. *Proc. Natl Acad. Sci. USA*, **87**, 7713–7716.
- Florack, D.E. and Stiekema, W.J. (1994) Thionins: properties, possible biological roles and mechanisms of action. *Plant Mol. Biol.* **26**, 25–37.
- Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974.
- Garcia-Olmedo, F., Carbonero, P., Hernandez-Lucas, C., Paz-Ares, J., Ponz, F., Vincente, O. and Sierra, J.M. (1983) Inhibition of eukaryotic cell-free protein synthesis by thionins from wheat endosperm. *Biochim. Biophys. Acta*, **740**, 52–56.
- Garcia-Olmedo, F., Rodriguez-Palenzuela, X., Hernandez-Lucas, C., Ponz, F., Marana, C., Carmona, M.J., Lopez-Fando, J.J., Fernandez, J.A. and Carbonero, P. (1989) The thionins: a protein family that includes purothionins, viscotoxins and crambin. In *Oxford Surveys of Plant Molecular and Cell Biology*. Vol. 6 (Mifflin, B.J. ed.). Oxford, UK: Oxford University Press, pp. 31–60.
- Garcia-Olmedo, F., Molina, A., Alamillo, J.M. and Rodriguez-Palenzuela, P. (1998) Plant defense peptides. *Biopolymers*, **47**, 479–491.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. (2005) Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook* (Walker, J.M., ed.). Totowa, NJ: Humana Press.
- Goldman, M.H., Goldberg, R.B. and Mariani, C. (1994) Female sterile tobacco plants are produced by stigma-specific cell ablation. *EMBO J.* **13**, 2976–2984.
- Graham, M.A., Silverstein, K.A.T., Cannon, S.B. and VandenBosch, K.A. (2004) Computational identification and characterization of novel genes from legumes. *Plant Physiol.* **135**, 1179–1197.
- Greenberg, J.T. and Yao, N. (2004) The role and regulation of programmed cell death in plant-pathogen interactions. *Cell Microbiol.* **6**, 201–211.
- Gutierrez-Marcos, J.F., Costa, L.M., Biderre-Petit, C., Khbaya, B., O'Sullivan, D.M., Wormald, M., Perez, P. and Dickinson, H.G. (2004) Maternally expressed gene1 is a novel maize endosperm transfer cell-specific gene with a maternal parent-of-origin pattern of expression. *Plant Cell*, **16**, 1288–1301.
- Haas, B.J., Delcher, A.L., Mount, S.M. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666.
- Haas, B.J., Wortman, J.R., Ronning, C.M. *et al.* (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol.* **3**, 7.
- Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P. and Brunak, S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24**, 3439–3452.
- Henrissat, B. and Davies, G.J. (2000) Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiol.* **124**, 1515–1519.
- Himly, M., Jahn-Schmid, B., Dedic, A. *et al.* (2003) Art v 1, the major allergen of mugwort pollen, is a modular glycoprotein with a defensin-like and a hydroxyproline-rich domain. *FASEB J.* **17**, 106–108.
- Jabs, T., Tschöpe, M., Colling, C., Hahlbrock, K. and Scheel, D. (1997) Elicitor-stimulated ion fluxes and O<sub>2</sub><sup>-</sup> from the oxidative burst are essential components in triggering defense gene activation and phytoalexin synthesis in parsley. *Proc. Natl Acad. Sci. USA*, **94**, 4800–4805.
- Kawano, T. (2003) Roles of the reactive oxygen species-generating peroxidase reactions in plant defense and growth induction. *Plant Cell Rep.* **21**, 829–837.
- Kunkel, B.N. and Brooks, D.M. (2002) Cross talk between signaling pathways in pathogen defense. *Curr. Opin. Plant Biol.* **5**, 325–331.
- Lam, E., Kato, N. and Lawton, M. (2001) Programmed cell death, mitochondria and the plant hypersensitive response. *Nature*, **411**, 848–853.
- Laskowski, M. and Qasim, M.A. (2000) What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochim. Biophys. Acta*, **1477**, 324–337.
- Lay, F.T., Schirra, H.J., Scanlon, M.J., Anderson, M.A. and Craik, D.J. (2003) The three-dimensional solution structure of NaD1, a new floral defensin from *Nicotiana glauca* and its application to a homology model of the crop defense protein alfA. *J. Mol. Biol.* **325**, 175–188.
- Mello, M.O., Tanaka, A.S. and Silva-Filho, M.C. (2003) Molecular evolution of Bowman-Birk type proteinase inhibitors in flowering plants. *Mol. Phylogenet. Evol.* **27**, 103–112.
- Melo, F.R., Rigden, D.J., Franco, O.L., Mello, L.V., Ary, M.B., Grossi de Sa, M.F. and Bloch, C. Jr (2002) Inhibition of trypsin by cowpea thionin: characterization, molecular modeling, and docking. *Proteins*, **48**, 311–319.
- Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., Ghazal, H. and Decola, S. (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Res.* **14**, 1641–1653.
- Molina, A., Segura, A. and Garcia-Olmedo, F. (1993) Lipid transfer proteins (nsLTPs) from barley and maize leaves are potent inhibitors of bacterial and fungal plant pathogens. *FEBS Lett.* **316**, 119–122.
- Moskal, W.A. Jr, Wu, H.C., Underwood, B.A., Wang, W., Town, C.D. and Xiao, Y. (2007) Experimental validation of novel genes predicted in the un-annotated regions of the Arabidopsis genome. *BMC Genomics*, **8**, 18.
- Motose, H., Sugiyama, M. and Fukuda, H. (2004) A proteoglycan mediates inductive interaction during plant vascular development. *Nature*, **429**, 873–878.
- Murata, M., Begg, G.S., Lambrou, F., Leslie, B., Simpson, R.J., Freeman, H.C. and Morgan, F.J. (1982) Amino acid sequence of a basic blue protein from cucumber seedlings. *Proc. Natl Acad. Sci. USA*, **79**, 6434–6437.
- Mygind, P.H., Fischer, R.L., Schnorr, K.M. *et al.* (2005) Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus. *Nature*, **437**, 975–980.
- Mysore, K.S. and Ryu, C.M. (2004) Nonhost resistance: how much do we know? *Trends Plant Sci.* **9**, 97–104.
- Odani, S., Koide, T. and Ono, T. (1983) The complete amino acid sequence of barley trypsin inhibitor. *J. Biol. Chem.* **258**, 7998–8003.
- Osborn, A.E. (1996) Preformed antimicrobial compounds and plant defense against fungal attack. *Plant Cell*, **8**, 1821–1831.
- Ouyang, S., Zhu, W., Hamilton, J. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887.
- Pearce, G., Moura, D.S., Stratmann, J. and Ryan, C.A. Jr (2001) RALF, a 5-kDa ubiquitous polypeptide in plants, arrests root growth and development. *Proc. Natl Acad. Sci. USA*, **98**, 12843–12847.
- Qin, Q., Bergmann, C.W., Rose, J.K., Saladie, M., Kolli, V.S., Albersheim, P., Darvill, A.G. and York, W.S. (2003) Characterization of a tomato protein that inhibits a xyloglucan-specific endoglucanase. *Plant J.* **34**, 327–338.

- Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Perte, G., Sultana, R. and White, J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159–164.
- Rawlings, N.D., Tolle, D.P. and Barrett, A.J. (2004) Evolutionary families of peptidase inhibitors. *Biochem. J.* **378**, 705–716.
- van Ree, R. (2002) Clinical importance of non-specific lipid transfer proteins as food allergens. *Biochem. Soc. Trans.* **30**, 910–913.
- Rhee, S.Y., Beavis, W., Berardini, T.Z. et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.* **31**, 224–228.
- Ride, J.P., Davies, E.M., Franklin, F.C. and Marshall, D.F. (1999) Analysis of Arabidopsis genome sequence reveals a large new gene family in plants. *Plant Mol. Biol.* **39**, 927–932.
- Ross, A.F. (1961) Systemic acquired resistance induced by localized virus infections in plants. *Virology*, **14**, 340–358.
- Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and biologist programmers. In *Bioinformatics Methods and Protocols* (Krawetz, S. and Misener, S., eds). Totowa, NJ: Humana Press, pp. 365–386.
- Ryan, C.A. (1990) Protease inhibitors in plants: genes for improving defenses against insects and pathogens. *Annu. Rev. Phytopathol.* **28**, 425–449.
- Satoh, S., Sturm, A., Fujii, T. and Chrispeels, M.J. (1992) cDNA cloning of an extracellular dermal glycoprotein of carrot and its expression in response to wounding. *Planta*, **188**, 432–438.
- Scheel, D. (1998) Resistance response physiology and signal transduction. *Curr. Opin. Plant Biol.* **1**, 305–310.
- Schopfer, C.R., Nasrallah, M.E. and Nasrallah, J.B. (1999) The male determinant of self-incompatibility in *Brassica*. *Science*, **286**, 1697–1700.
- Segura, A., Moreno, M., Madueno, F., Molina, A. and Garcia-Olmedo, F. (1999) Snakin-1, a peptide from potato that is active against plant pathogens. *Mol. Plant Microbe Interact.* **12**, 16–23.
- Serna, A., Maitz, M., O'Connell, T. et al. (2001) Maize endosperm secretes a novel antifungal protein into adjacent maternal tissue. *Plant J.* **25**, 687–698.
- Shai, Y. (2002) Mode of action of membrane active antimicrobial peptides. *Biopolymers*, **66**, 236–248.
- Siegel, S. (1956) *Nonparametric statistics: for the behavioral sciences*. New York, NY: McGraw-Hill Book Company.
- Silverstein, K.A.T., Graham, M.A., Paape, T.D. and VandenBosch, K.A. (2005) Genome organization of more than 300 defensin-like genes in Arabidopsis. *Plant Physiol.* **138**, 600–610.
- Silverstein, K.A., Graham, M.A. and VandenBosch, K.A. (2006) Novel paralogous gene families with potential function in legume nodules and seeds. *Curr. Opin. Plant Biol.* **9**, 142–146.
- Stec, B., Rao, U. and Teeter, M.M. (1995) Refinement of purothionins reveals solute particles important for lattice formation and toxicity. Part 2: structure of beta-purothionin at 1.7 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.* **51**, 914–924.
- Sutton, G.S., White, O., Adams, M.D. and Kerlavage, A.R. (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1**, 9–19.
- Takayama, S., Shiba, H., Iwano, M., Asano, K., Hara, M., Che, F.S., Watanabe, M., Hinata, K. and Isogai, A. (2000a) Isolation and characterization of pollen coat proteins of *Brassica campestris* that interact with S locus-related glycoprotein 1 involved in pollen-stigma adhesion. *Proc. Natl Acad. Sci. USA*, **97**, 3765–3770.
- Takayama, S., Shiba, H., Iwano, M., Shimosato, H., Che, F.S., Kai, N., Watanabe, M., Suzuki, G., Hinata, K. and Isogai, A. (2000b) The pollen determinant of self-incompatibility in *Brassica campestris*. *Proc. Natl Acad. Sci. USA*, **97**, 1920–1925.
- Talyzina, N.M. and Ingvarsson, P.K. (2006) Molecular evolution of a small gene family of wound inducible Kunitz trypsin inhibitors in *Populus*. *J. Mol. Evol.* **63**, 108–119.
- Terras, F.R.G., Goderis, I.J., Van Leuven, F., Vanderleyden, J., Cammue, B.P.A. and Broekaert, W.F. (1992) In vitro anti-fungal activity of a radish (*Raogabys satuvys* L.) seed protein homologous to nonspecific lipid transfer proteins. *Plant Physiol.* **100**, 1055–1058.
- The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Theis, T. and Stahl, U. (2004) Antifungal proteins: targets, mechanisms and prospective applications. *Cell. Mol. Life Sci.* **61**, 437–455.
- Thevissen, K., Ferket, K.K., Francois, I.E. and Cammue, B. (2003) Interactions of antifungal plant defensins with fungal membrane components. *Peptides*, **24**, 1705–1712.
- Thomma, B., Eggermont, K., Penninckx, I., Mauch-Mani, B., Vogelsang, R., Cammue, B.P.A. and Broekaert, W.F. (1998) Separate jasmonate-dependent and salicylate-dependent defense-response pathways in Arabidopsis are essential for resistance to distinct microbial pathogens. *Proc. Natl Acad. Sci. USA*, **95**, 15107–15111.
- Thomma, B.P., Cammue, B.P. and Thevissen, K. (2002) Plant defensins. *Planta*, **216**, 193–202.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
- Underwood, B.A., Vanderhaeghen, R., Whitford, R., Town, C.D. and Hilson, P. (2006) Simultaneous high-throughput recombinatorial cloning of open reading frames in closed and open configurations. *Plant Biotechnol. J.* **4**, 317–324.
- Van Loon, L.C. and van Kammen, A. (1970) Polyacrylamide disc electrophoresis of the soluble leaf proteins from *Nicotiana tabacum* var. Samsun and Samsun NN II. Changes in protein constitution after infection with tobacco mosaic virus. *Virology*, **40**, 199–211.
- Van Loon, L.C. and van Strien, E.A. (1999) The families of pathogenesis-related proteins, and their activities, and comparative analysis of PR-1 type proteins. *Physiol. Mol. Plant Pathol.* **55**, 85–97.
- Vance, C.P., Kirk, T.K. and Sherwood, R.T. (1980) Lignification as a mechanism of disease resistance. *Annu. Rev. Phytopathol.* **18**, 259–288.
- Vanoosthuysse, V., Mieg, C., Dumas, C. and Cock, J.M. (2001) Two large *Arabidopsis thaliana* gene families are homologous to the Brassica gene superfamily that encodes pollen coat proteins and the male component of the self-incompatibility response. *Plant Mol. Biol.* **46**, 17–34.
- Veronese, P., Ruiz, M.T., Coca, M.A. et al. (2003) In defense against pathogens. Both plant sentinels and foot soldiers need to know the enemy. *Plant Physiol.* **131**, 1580–1590.
- Vivanco, J.M., Savary, B.J. and Flores, H.E. (1999) Characterization of two novel type I ribosome-inactivating proteins from the storage roots of the andean crop *Mirabilis expansa*. *Plant Physiol.* **119**, 1447–1456.
- Watanabe, M., Ito, A., Takada, Y. et al. (2000) Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of *Brassica campestris* (syn. rapa) L. *FEBS Lett.* **473**, 139–144.

- Wortman, J.R., Haas, B.J., Hannick, L.I. et al.** (2003) Annotation of the Arabidopsis genome. *Plant Physiol.* **132**, 461–468.
- Xiao, Y.L., Smith, S.R., Ishmael, N., Redman, J.C., Kumar, N., Monaghan, E.L., Ayele, M., Haas, B.J., Wu, H.C. and Town, C.D.** (2005) Analysis of the cDNAs of hypothetical genes on Arabidopsis chromosome 2 reveals numerous transcript variants. *Plant Physiol.* **139**, 1323–1337.
- Yamada, K., Lim, J., Dale, J.M. et al.** (2003) Empirical Analysis of transcriptional activity in the Arabidopsis genome. *Science*, **302**, 842–846.
- Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A. and Tabata, S.** (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Physiol.* **137**, 1174–1181.
- Zhang, S. and Klessig, D.F.** (2001) MAPK cascades in plant defense signaling. *Trends Plant Sci.* **6**, 520–527.
- Zhang, K., McKinlay, C., Hocart, C.H. and Djordjevic, M.A.** (2006) The *Medicago truncatula* small protein proteome and peptidome. *J. Proteome Res.* **5**, 3355–3367.
- Zhu, S., Gao, B. and Tytgat, J.** (2005) Phylogenetic distribution, functional epitopes and evolution of the CSalphabeta superfamily. *Cell. Mol. Life Sci.* **62**, 2257–2269.