RESEARCH ARTICLE

Evolution of the defensin-like gene family in grass genomes

JIANDONG WU, XIAOLEI JIN, YANG ZHAO, QING DONG, HAIYANG JIANG and QING MA*

Key Laboratory of Crop Biology of Anhui Province, School of Life Sciences, Anhui Agricultural University, Hefei 230036, People's Republic of China

Abstract

Plant defensins are small, diverse, cysteine-rich peptides, belonging to a group of pathogenesis-related defense mechanism proteins, which can provide a barrier against a broad range of pathogens. In this study, 51 defensin-like (*DEFL*) genes in Gramineae, including brachypodium, rice, maize and sorghum were identified based on bioinformatics methods. Using the synteny analysis method, we found that 21 *DEFL* genes formed 30 pairs of duplicated blocks that have undergone large-scale duplication events, mostly occurring between species. In particular, some chromosomal regions are highly conserved in the four grasses. Using mean K_S values, we estimated the approximate time of divergence for each pair of duplicated regions and found that these regions generally diverged more than 40 million years ago (Mya). Selection pressure analysis showed that the *DEFL* gene family is subjected to purifying selection. However, sliding window analysis detected partial regions of duplicated genes under positive selection. The evolutionary patterns within *DEFL* gene families among grasses can be used to explore the subsequent functional divergence of duplicated genes and to further analyse the antimicrobial effects of defensins during plant development.

[Wu J., Jin X., Zhao Y., Dong Q., Jiang H. and Ma Q. 2016 Evolution of the defensin-like gene family in grass genomes. *J. Genet*. **95**, 53–62]

Introduction

Since the isolation of HNP123 from human neutrophils in 1984 (Yang *et al.* 2004), many defensins, which are extensively distributed in animals, plants and insects, have been identified and isolated. Depending on the structure of the precursor proteins, plant defensins can be grouped into two major classes. The first and largest class consists of an endoplasmic reticulum signal sequence and a mature defensin domain, and the second class contains larger precursors of ~33 amino acids in the C-terminus (Prema and Pruthvi 2012). Initially, defensins were considered to comprise small multigene families; however, this view is controversial due to the discovery of more than 90 *DEFL* genes in rice and over 300 members in *Arabidopsis* and legumes (Graham *et al*. 2004; Silverstein *et al*. 2005, 2007).

Plant defensins comprise a class of small and diverse cysteine-rich antimicrobial peptides that share a complex three-dimensional folding structure stabilized by eight conserved disulphide-linked cysteines (Thomma *et al.* 2002; Giacomelli *et al.* 2012). A comparison of primary amino acid sequences of plant defensins indicates a rich diversity of variants, despite their highly conserved structural features, which is responsible for the functional diversity observed in defensins (Lay and Anderson 2005; Carvalho Ade and Gomes 2009; Stotz *et al.* 2009). Previous studies have demonstrated that plant defensins can inhibit fungal growth from the extracellular or intracellular sides of fungal cells through interactions with fungal-specific cell wall and plasma membrane components (Aerts *et al.* 2008). Defensins are usually expressed as precursor proteins with signal peptides, and they execute antimicrobial functions after a series of processes (Meyer *et al.* 1996). Since the discovery of their potent antifungal activity, defensins have been widely used in agrobiotechnology to generate disease-resistant crops.

Gramineae, which evolved ~70 million years ago (Mya) from a common ancestor, includes a number of important agronomic crops, such as rice, maize and sorghum. Although the origin of these crops can be dated to ~50 to 65 Mya, the family has now expanded to over 10,000 species (Kellogg 2001). Three of the domesticated grasses, i.e. rice, wheat and maize, account for approximately half of total world food production. With the advent of comparative mapping, comparative genome analyses has demonstrated that gene orders among related plant species have been largely conserved during the course of evolution (Gale and Devos 1998; Keller and Feuillet 2000; Paterson *et al.* 2000), and the extensive differences in genome size observed among related species

*For correspondence. E-mail: qingmad@163.com.

Keywords. defensin-like genes; Gramineae; gene evolution; duplicated genes; selection pressure.

are determined by polyploid events and transposable element duplication (SanMiguel *et al.* 1998; Bennetzen 2000). The chromosomal organization of grass has remained largely conserved for 60 million years (Myr), but small local rearrangements and duplications are clearly a common feature of grass genome evolution.

In the present study, we analysed interspecific and intraspecific collinear relationships and duplication events of *DEFL* genes among four Gramineae plants, including *Oryza sativa*, *Zea mays*, *Brachypodium distachyon* and *Sorghum bicolor*. The major objectives of this study were as follows: (i) to identify *DEFL* genes in these four plant species; (ii) to determine which genes exhibit synteny; (iii) to estimate the approximate time that duplication events occurred; and (iv) to analyse the environmental selection pressure on this family. Comparison and analysis of intraspecific and interspecific gene rearrangement, gene loss and gene frequency are important for understanding the relationship between genome structure and gene function, and for implementing strategies for crop improvement.

Materials and methods

Identification of DEFL genes

Recent versions of genome sequences, coding sequences (CDS) and protein sequences for rice, maize, sorghum and brachypodium were downloaded from the following sources: rice from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu//), maize from B73 Maize Genome Project (http://www.maizesequence.org/index.html), sorghum from DOE-JGI Community Sequencing Program (CSP) (http://www.phytozome.net/sorghum.php) and brachypodium from *Brachypodium distachyon* information resource (http://www.brachypodium.org/).

Firstly, successive iterations of the hidden Markov model (HMM) were adopted to identify all possible *DEFL* genes in four genome sequences with BLASTP searches ($E \leq 0.001$). To find potent *DEFL* genes not predicted by the first method, a second method was employed. The identified set of *DEFL* genes from *Arabidopsis*, soybean and capsicum was used as a query to search for additional genes through BLAST searches of the genome sequences (Graham *et al*. 2004; Silverstein *et al*. 2005, 2007). All candidate sequences that met the standards were confirmed to be real by Pfam (http://pfam.sanger.ac.uk/) and SMART (http://smart.embl-heidelberg.de/) analysis. Then, all confirmed DEFL proteins were aligned using ClustalW, and all identical sequences were checked manually to remove redundant sequences. Finally, this restricted set of genes was further filtered to include only those with a clear signal peptide by scanning with the SignalP program ($P \geq 0.9$).

Phylogenetic analysis

The complete amino acid sequences of all *DEFL* genes from brachypodium, rice, maize and sorghum were clustered and aligned with ClustalW (Thompson *et al.* 2002). The phylogenetic trees were plotted using MEGA 4.0 (Tamura *et al.* 2007) with the neighbour-joining (NJ) method with default parameters; bootstrap analysis was performed using 1000 replicates with the pairwise deletion option.

Categorization of DEFL expansion

To categorize the apparent expansion of *DEFL* gene families, the process of segmental and tandem duplication was analysed based on the similarity among sets of *DEFL* genes as markers for regions involved in such duplications (Vision *et al.* 2000). Toward this goal, the chromosome locations of all members were investigated. Tandem duplications are usually defined as multiple genes located within the same intergenic region or within a neighbouring intergenic region.

To determine whether two *DEFL* genes reside within a duplicated block, high similarity must be detected between their flanking genes at the amino acid level (Maher *et al.* 2006). Primarily, all *DEFL* genes in four species were respectively selected as anchors, and all flanking genes 100-kb upstream and downstream of each *DEFL* were then compared by BLASTP to identify duplicated genes between two independent regions (Li *et al.* 2014). Then, the total number of flanking genes ($E \leq 10^{-10}$) between two anchor points was calculated. Additionally, microsynteny analysis between species was performed, employing the same method as above, except the best nonself hit ($E \leq 10^{-20}$) (Sato *et al.* 2008).

Estimation of synonymous substitution and duplication event dates

To calculate the synonymous substitution (K_a) and nonsynonymous substitution (K_s), codons were extracted for each amino acid that was aligned between genes using protein alignment as a guide, excluding regions containing gaps. DnaSP 5.0 program (Librado and Rozas 2009) was used to calculate the substitutions for each pair of conserved flanking genes within each *DEFL* duplicated region. $K_s > 2.0$ would be excluded due to the risk of saturation (Blanc and Wolfe 2004). For each pair of duplicated regions, the mean K_s value was calculated for individual homologues within flanking conserved genes and used to determine the approximate time of divergence (T) with the equation $T = K_s/2E$; E represents a constant rate of synonymous substitutions for monocots of 6.5×10^{-9} substitutions per synonymous site per year (Gaut *et al.* 1996).

Detecting the selection pressure on DEFL genes

To detect the selection pressure on *DEFL* genes, K_a/K_s ratios were calculated for each duplicated region using DnaSP 5.0. Moreover, a sliding window of 30 bp under a step size of 6 bp was used to calculate K_a/K_s ratios. Both K_a and K_s were calculated according to Nei and Gojobori (1986).

Results

Genomewide identification of DEFL genes

By performing HMM and BLAST searches, we identified a total of 41 putative defensin sequences in the four plant species examined. An additional 17 sequences were predicted using the second process. Finally, we identified a total of 51 *DEFL* genes (nine in brachypodium, 12 in rice, 20 in maize and 10 in sorghum) after excluding redundant sequences or sequences without signal peptides. We gathered information about these *DEFL* genes, such as open reading frame (ORF) length, pI (isoelectric point) and location on chromosomes (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/; figure 1). A previous study has shown that plant DEFL proteins are usually small peptides that are basic in the mature form and contain a signal peptide (Carvalho Ade and Gomes 2009). The current study shows that DEFL proteins range in size from roughly 71–129 amino acids (aa) long, with an average length of ∼88 aa, which corresponds to the results of a previous study. Further, the smallest average length is ∼84 aa (in brachypodium), while the largest DEFL protein is ∼91 aa (in maize). The pI values of these proteins range from 5.17 to 10.05, with an average value of ∼8.

Phylogenetic and sequence structure analysis

Based on the protein sequences, we constructed a phylogenetic tree with default parameters. Although many gene clusters were present in each clade, it was still difficult to classify some genes into classes due to their low bootstrap values (<50%), whereas we also identified some genes with high similarity, such as SbDf-02/ZmDf-13 and BdDf-03/OsDf-01 (figure 1 in electronic supplementary material). On the other hand, the *DEFL* genes formed sister pairs both within and between species. In particular, sister pairs between species were primarily formed by sorghum and maize orthologues with very strong bootstrap support, which may be related to the fact that sorghum and maize have a relatively close relationship. The different degree in orthologous relationships among sorghum, maize, rice and brachypodium is due to genetic relatedness of these genomes. Similar results were also observed in the previous study (Muthamilarasan *et al*. 2014). To further understand defensin sequence structure, we carried out exon–intron structure analysis (http://gsds.cbi.pku.edu.cn/). The results showed that most *DEFL* genes generally contain two exons and one intron (figure 2 in electronic supplementary material), which is consistent with previous results (Giacomelli *et al.* 2012). Except for OsDf-03/06/09/10/11, SbDf-07/08/09, ZmDf-06 and

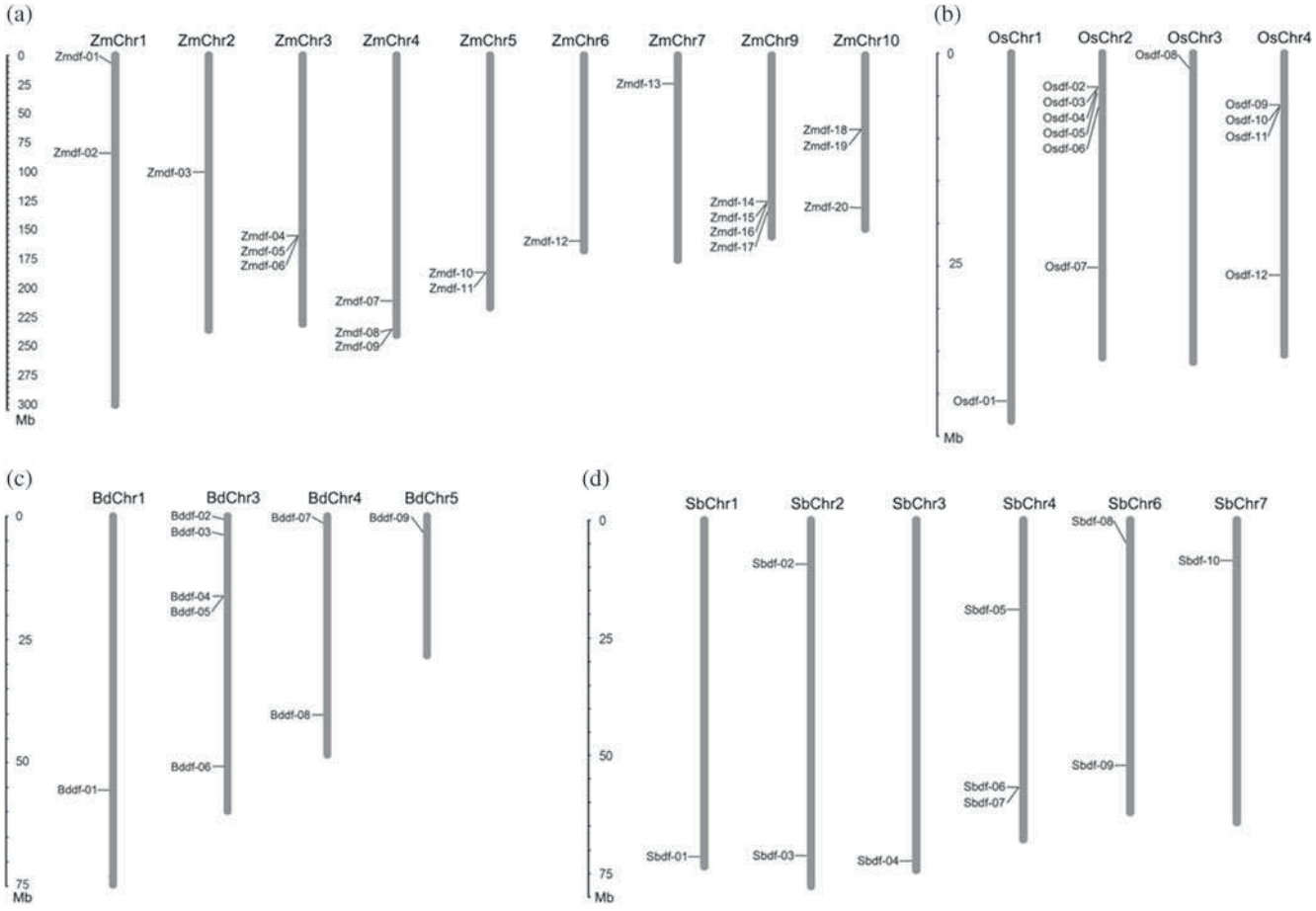

Figure 1. Chromosomal locations of *DEFL* genes in four species. (a) *Z. mays*, (b) *O. sativa*; (c) *B. distachyon*; (d) *S. bicolor*.

BdDf-05/08/09, which contain a longer intron, the remaining *DEFL* genes typically contain a small intron of ~200 bp.

Considering the cysteine signatures of defensin genes, we next performed analysis of protein sequence structure (figure 2). We found that in these proteins, four cysteines are completely conserved, with one, three, four and seven sites among eight conserved cysteines, respectively. We also found some other relatively conserved amino acids, such as glycine and serine. The lengths of signal peptides differ among sequences; most contain ~25 to 30 amino acids, and several conserved leucines exist in all of the signal peptides.

Microsynteny analysis

We analysed the chromosomal locations of *DEFL* genes among the four grass species, which revealed several gene clusters (figure 3 in electronic supplementary material). By examining the chromosomal information for each gene and comparing the evolutionary relationship between these genes, we found that 16 genes experienced apparent tandem duplications (table 1 and figure 3 in electronic supplementary material). Nevertheless, the tandem duplication events only occurred in maize and rice, in the gene pairs OsDf-02/03/04/05, OsDf-09/10/11, ZmDf-04/05/06, ZmDf-08/09, ZmDf-10/11 and ZmDf-14/15, respectively.

To determine whether the flanking genes of *DEFL* have undergone large-scale duplication events during the evolution of *DEFL* gene families, we compared the flanking genes of any two *DEFL* genes. If three or more flanking members had a best nonself match according to BLASTP ($E \leq 10^{-10}$ within species and $E \leq 10^{-20}$ between species) (Sato *et al.* 2008), we considered that these members belonged to a duplicated block. Ultimately, we detect a total of 21 (41%) genes involved in large-scale duplication events, with a maximum number of seven in rice and a minimum number of four in brachypodium. Additionally, the 21 *DEFL* genes form 30 pairs of duplicated blocks, including 28 pairs occurring between species. Although maize has the most *DEFL* genes (20), only five genes have gone through the duplication procedure. These results demonstrate that the occurrence of large-scale duplication events is not directly related to the number of genes in a family.

In particular, colinearity at the map level was plotted using the obtained duplicated blocks (figure 3). According to the figure, the greatest gene density was observed in the rice genome, and maize has the smallest gene density among the four grass plants examined. We also found that seven

Figure 2. Main domain alignment of DEFL proteins in four grass species.

DEFL genes were primarily involved in a mutual large-scale duplication event, which constituted 20 pairs of duplicated blocks and only two pairs within species (figure 3c). Figure 3 also shows that the gene order was generally conserved among species. Further, several duplicated regions appear to have transposed to nearby colinear positions. Seven *DEFL* gene homologues from rice and maize, however, were not detected in brachypodium genomic DNA, suggesting

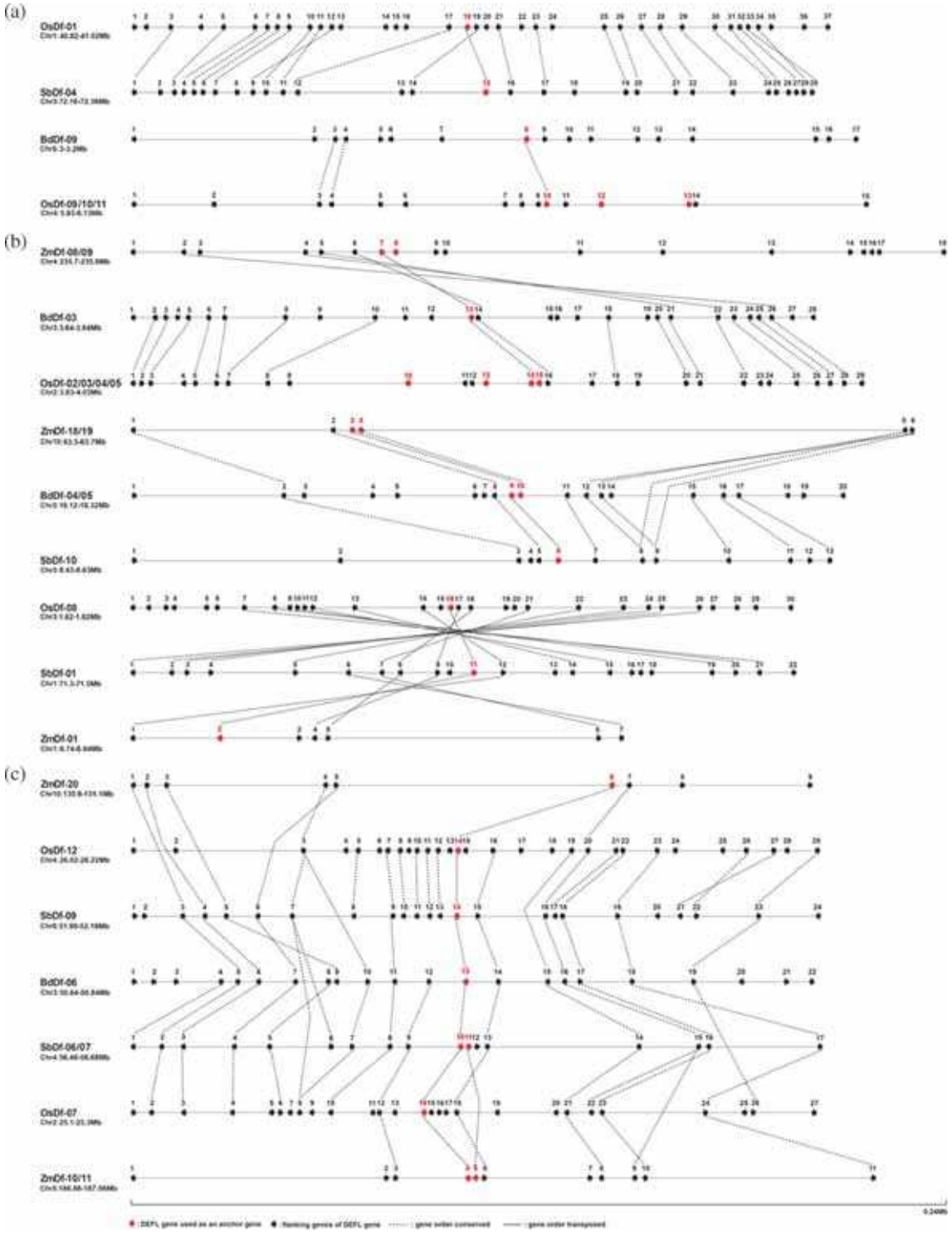

Figure 3. Comparative maps of *DEFL* genes and their flanking genes within syntenic chromosomal intervals between four grass species. The red arrows represent anchor *DEFL* genes, and the flanking genes are indicated by black arrows. The gene order is indicated from left to right for each segment. Conserved gene pairs between the segments are connected with lines (genes in order) and dotted lines (genes in inversion).

either extensive sequence divergence or the absence of these genes in brachypodium. Further, the large-scale duplication event of brachypodium primarily occurred on chromosome 3.

Estimation of duplication event date

An attempt was made to estimate the dates of the large-scale duplication events, assuming that synonymous silent substitutions per site occur at a constant rate over time. For this analysis, we calculated the K_s between the coding sequences of each paralogous pair and used the mean K_s to estimate the approximate date of duplication event, with an estimated rate of silent-site substitutions of 6.5×10^{-9} substitutions per synonymous site per year. Information about the mean K_s values for each duplication event and the estimated dates are shown in table 1. The results show that the timing of large-scale duplication events was divergent, with values ranging from 12 to 67 Myr, primarily in the past 40–55 Myr. Additionally, we plotted the number of duplication blocks against the average K_s value of the block pairs to find a distribution that reflected the approximate age of the duplication

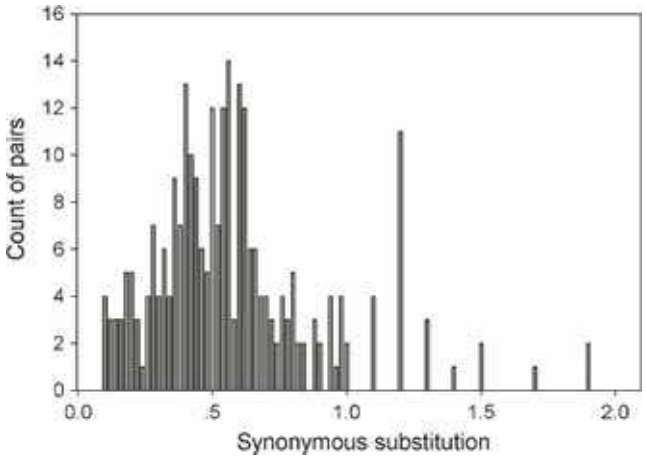

Figure 4. K_a/K_s ratios of the duplicated *DEFL* genes and their flanking paralogous in four grass species are shown above the dots. The *x* and *y* axes denote the synonymous distance and K_a/K_s ratios for each pair, respectively.

(figure 4). This figure shows that the K_s values were mostly distributed around 0.5 synonymous substitutions per site, which can be referred to as 'recent duplications'. However,

Table 1. Estimation of the dates of large-scale duplication events in four grasses.

Duplicated pair	*n*	Minimum K_s	Maximum K_a	K_s (mean ± SD)	Date (Myr)
ZmDf-01 & OsDf-08	5	0.37	0.62	0.49±0.11	37.75
ZmDf-10 & OsDf-07	6	0.17	0.87	0.54±0.26	41.82
ZmDf-10 & OsDf-12	5	0.35	1.63	0.88±0.48	67.51
ZmDf-20 & OsDf-07	6	0.31	1.90	0.82±0.61	63.06
ZmDf-20 & OsDf-12	3	0.35	0.98	0.57±0.35	44.15
ZmDf-08 & BdDf-03	4	0.40	0.66	0.56±0.11	43.24
ZmDf-20 & BdDf-06	7	0.28	0.99	0.73±0.28	55.82
ZmDf-10 & BdDf-06	7	0.27	0.64	0.54±0.14	41.62
ZmDf-18 & BdDf-04	6	0.45	0.87	0.62±0.14	47.74
ZmDf-01 & SbDf-01	6	0.10	0.26	0.17±0.06	13.12
ZmDf-10 & SbDf-06	8	0.12	0.42	0.21±0.09	15.98
ZmDf-10 & SbDf-09	3	0.45	0.97	0.65±0.28	49.87
ZmDf-18 & SbDf-10	5	0.09	0.25	0.15±0.07	11.83
ZmDf-20 & SbDf-06	6	0.40	1.35	0.86±0.37	66.15
ZmDf-20 & SbDf-09	7	0.14	0.34	0.21±0.07	16.46
OsDf-01 & SbDf-04	23	0.26	1.27	064±0.28	49.23
OsDf-07 & SbDf-06	13	0.12	0.62	0.44±0.15	33.85
OsDf-07 & SbDf-09	11	0.29	1.18	0.67±0.30	51.43
OsDf-08 & SbDf-01	15	0.21	0.80	0.50±0.16	38.36
OsDf-12 & SbDf-06	8	0.31	1.46	0.74±0.37	56.92
OsDf-12 & SbDf-09	15	0.30	1.15	0.55±0.20	42.45
OsDf-04 & BdDf-03	17	0.32	1.15	0.51±0.19	39.23
OsDf-07 & BdDf-06	14	0.11	0.73	0.42±0.17	32.47
OsDf-09 & BdDf-09	3	0.54	1.17	0.85±0.32	65.58
OsDf-12 & BdDf-06	9	0.35	1.47	0.79±0.37	60.70
SbDf-06 & BdDf-06	14	0.09	0.70	0.46±0.15	35.64
SbDf-09 & BdDf-06	12	0.30	1.89	0.85±0.43	65.13
SbDf-10 & BdDf-04	9	0.37	0.79	0.57±0.15	43.85
OsDf-12 & OsDf-07	8	0.28	1.19	0.74±0.36	56.88
SbDf-09 & SbDf-06	9	0.32	1.14	0.66±0.27	50.77

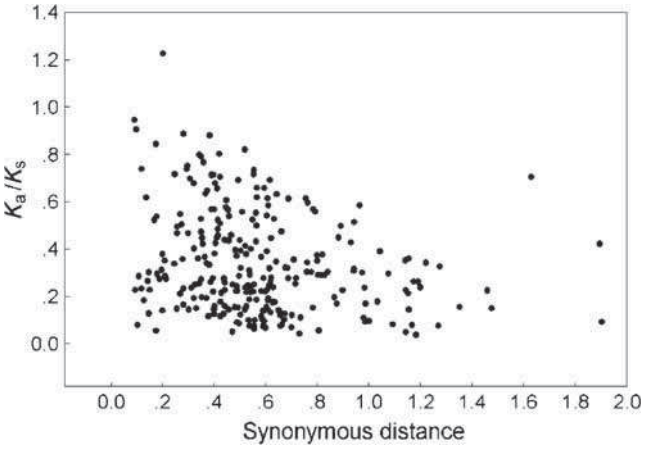

Figure 5. Distribution of synonymous substitution (K_s) of duplicated gene pairs in four grass plants. The histograms indicate synonymous substitutions between pairs (*x*-axis) and the number of duplicated gene pairs (*y*-axis).

the partial K_s values were much higher than 1.0 synonymous substitution per site, and almost all occurred between flanking genes, suggesting that those genes may have diverged from more distant duplication events.

Selection pressure on DEFL genes

The K_a/K_s ratios for 30 pairs of *DEFL* paralogues were calculated to elucidate the evolutionary constraints acting on the *DEFL* gene family. The pairwise comparison results showed that most K_a/K_s ratios were smaller than 1, suggesting that the duplicated regions were subjected to purifying selection (figure 5). Considering the possibility that overall strong purifying selection can cover up positive selection on some individual codon sites, we performed further analysis using sliding window analysis of K_a/K_s for each syntenic gene pair. When the proportion of differences is equal or higher than 0.75, the K_a/K_s ratios of certain coding region cannot be calculated. Figure 6 shows that positive selection actually occurred on some regions, despite the fact that whole genes were subjected to purifying selection, particularly for ZmDf-20/BdDf-07 and SbDf-09/BdDf-07.

Discussion

The *DEFL* gene family was first reported in a woody plant species, *Vitis vinifera*. Subsequently, the number of reports on the plant *DEFL* gene family has increased rapidly and even more defensins have been isolated from plants. Here, we identified a total of 51 *DEFL* genes in the four grass species examined using a bioinformatic approach. Each *DEFL* gene family appeared to be smaller than the number previously identified in grasses and legumes using an analogous method (Graham *et al*. 2004; Silverstein *et al*. 2005, 2007). Phylogenetic analysis of DEFL proteins performed in the current study showed that some genes could barely be classified into groups due to their low bootstrap values (<50%). On the other hand, we noted that orthologue pairs of maize and sorghum DEFL proteins were more prevalent in the phylogenetic tree, illustrating that some ancestor *DEFL* genes have existed before the divergence of maize and sorghum. Exon–intron structure analysis was also conducted to support the phylogenetic relationships of DEFL proteins in the four grasses. Corresponding to a previous study, the majority of *DEFL* genes are characterized by a typical gene structure of two exons and one intron, and the length of the intron is more variable than the exon size (Giacomelli *et al.* 2012).

Comparative genome analyses have demonstrated that gene orders among related plant species have been largely conserved over the course of million years of evolution, despite the large differences between species in terms of genome size and chromosome number (Devos and Gale 2000). The highly similar arrangement of duplicated block for seven *DEFL* genes among the four species examined (figure 3c) suggests that these chromosomal segments have undergone a few rearrangements since their divergence. In other words, these chromosomal regions are highly conserved, suggesting the crucial roles of these seven *DEFL* genes in the plant defense system. Actually, a few chromosomal rearrangements were recognized that are likely to have occurred before Poaceae divergence. In particular, we detected the existence of several-to-one microsynteny between rice and sorghum, which indicates that several flanking genes of one anchor are likely to have been derived from one flanking gene of another, indicating that duplication events, random translocations and even insertion events may have occurred during evolutionary history (data not shown). Even if it is limited to small regions, the existence of colinearity among the four species would allow direct exploitation of the genomic sequence of one species for the identification of candidate genes in the other species. Although this study employed bioinformatic methods to assess gene duplication and its colinearity with the genome sequence dataset, the features of genome conservation and divergence that we described are meaningful for comparative analysis of plant genomes.

Protein-coding genes often evolve from several duplication mechanisms such as local rearrangements, large-scale chromosomal duplication and genomewide duplication events (Bowers *et al.* 2003; Lawton-Rauh 2003; Blanc and Wolfe 2004). Various genomic analyses in rice, sorghum and maize have suggested an ancestral whole-genome duplication predating the divergence of the different cereal genomes (Paterson *et al.* 2004; Wang *et al.* 2005; Wei *et al.* 2007). Whole duplication of the maize genome through allotetraploidization was founded and characterized further through the evolutionary analysis of duplicated genes (Gaut and Doebley 1997) and by interspecific comparisons between orthologous loci in rice, sorghum and maize (Swigonova *et al.* 2004). By investigating the genomic positions of *DEFL* genes, we demonstrated that *DEFL* genes have evolved through both segmental duplications and tandem duplications in the genomes of rice,

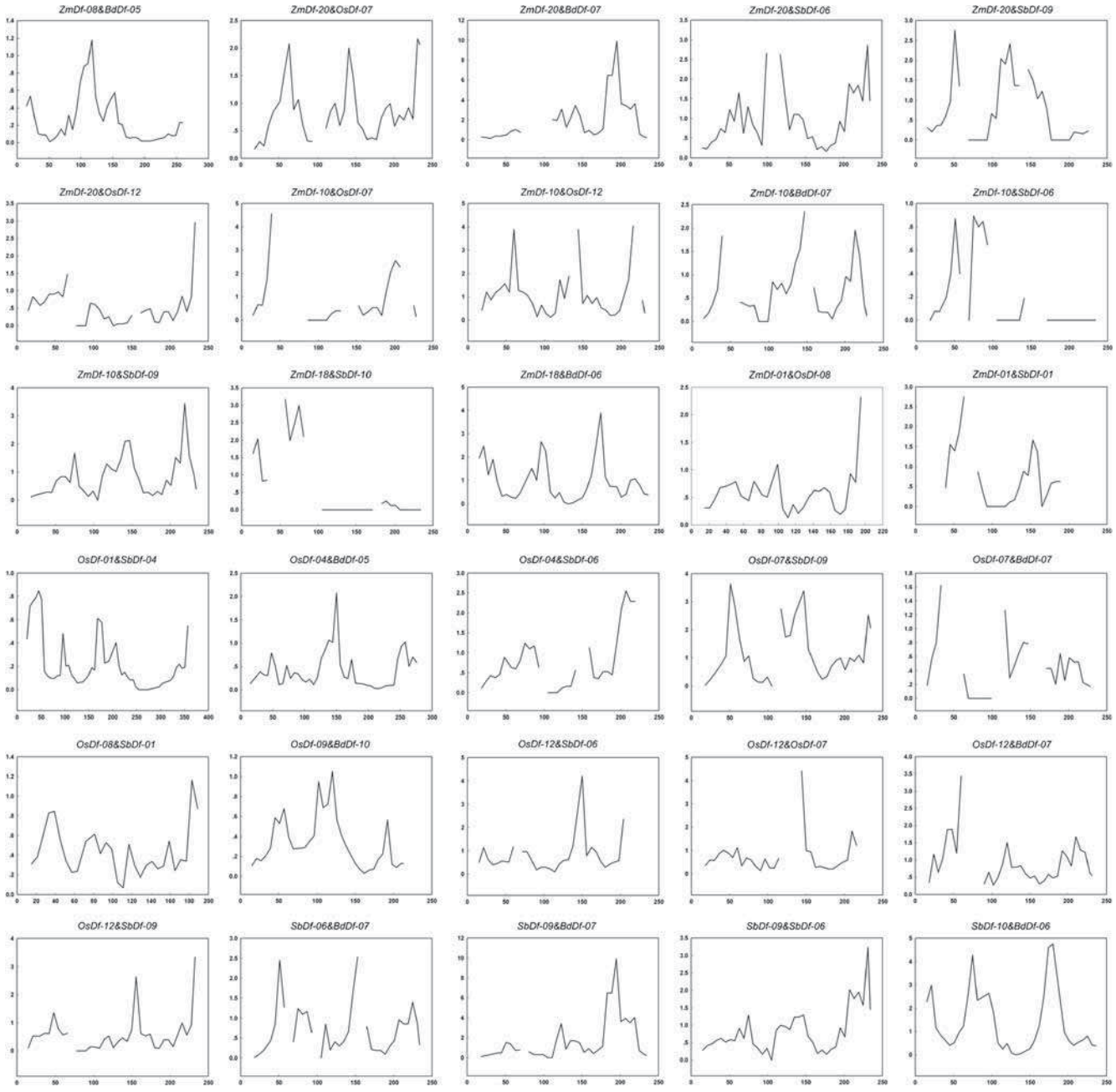

Figure 6. Sliding window analysis of duplicated *DEFL* genes in four grass species. The window size is 30 bp, with a step size of 6 bp. The *x*-axis denotes the nucleotide position. The *y*-axis denotes K_a/K_s ratio. The gaps represent K_a/K_s ratios that could not be computed.

maize,sorghum and brachypodium. In particular, tandem duplications in maize and rice may lead to increased transcript and protein accumulation or functional diversification. Analysing the occurrence of duplication events, we found that the number of *DEFL* genes in duplicated blocks between species (28 pairs) was higher than that within species (only two pairs), illustrating that ancient large-scale duplication may have been followed by gene loss and rearrangement. A total of 10 (33%) duplicated pairs have 10 or more conserved flanking protein-coding genes, showing that duplication events have resulted in the expansion of the *DEFL* gene family and their flanking genes. Previous studies have demonstrated that some *DEFL* genes are densely rich in transposable elements, and a high density of retrotransposons corresponds to a rapidly evolving portion of the genome and often includes resistance genes, which may explain the high duplication rate of *DEFL* genes in our study (Bertioli *et al.* 2009; Giacomelli *et al.* 2012) In particular, we determined that rice shares a remarkably high level of conserved synteny with sorghum and brachypodium, indicating that the duplication events caused much greater expansion between these species than between the others. These duplications

and comparisons with grass plants indicated common and lineage-specific patterns of conservation between the different genomes.

The distribution of K_s value was examined to clarify the differences among duplicated regions. However, when we calculated the K_s value for each duplicated region, the values for several pairs could not be calculated due to the higher proportion of differences among the genes, which resulted in the failure to estimate the duplication date and selection pressure of these duplicated pairs. The results demonstrate that the K_s value of each duplicated pair differed from each other, illustrating that duplicated pairs may have different ages of divergence. Our analysis of selection pressure on *DEFL* genes within and between species demonstrates that the *DEFL* gene families and their flanking protein-coding genes are subjected to purifying selection, with one exception. Subsequently, sliding window analysis was conducted to detect gene sequences with unusual selection pressure, which provided more detailed insights to help elucidate the environmental effects on this family. The results confirm that some codon sites suffered strong positive selection even if the entire regions were under purifying selection.

We demonstrated that in general, plant *DEFL* gene families evolved through duplication events, which is similar to the events that drive the evolution of protein-coding genes. The evolutionary patterns within *DEFL* gene families among the four grass plants can be used for further experimental analysis of their antimicrobial effects in plant development and can enable us to explore the subsequent functional divergence of the duplicated genes Our methods can also be exploited to analyse other species, including a family similar to the four grass plants examined.

Acknowledgements

This work was supported by the National Key Technologies Research and Development Programme of China (2012BAD20B00) and Anhui Provincial Natural Science Foundation (1408085MKL35). We would like to thank members of Key Laboratory of Crop Biology of Anhui province for their assistance in this study.

References

Aerts A. M., François I. E. J. A., Cammue B. P. A. and Thevissen K. 2008 The mode of antifungal action of plant, insect and human defensins. *Cell Mol. Life Sci.* **65**, 2069–2079.

Bennetzen J. L. 2000 Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**, 1021–1029.

Bertioli D. J., Moretzsohn M. C., Madsen L. H., Sandal N., Leal-Bertioli S. C., Guimarães P. M. *et al.* 2009 An analysis of synteny of Arachis with Lotus and Medicago sheds new light on the structure, stability and evolution of legume genomes. *BMC Genomics* **10**, 45.

Blanc G. and Wolfe K. H. 2004 Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678.

Bowers J. E., Chapman B. A, Rong J. and Paterson A. H. 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438.

Carvalho Ade O. and Gomes V. M. 2009 Plant defensins—prospects for the biological functions and biotechnological properties. *Peptides* **30**, 1007–1020.

Devos K. M. and Gale M. D. 2000 Genome relationships: the grass model in current research. *Plant Cell* **12**, 637–646.

Gale M. D. and Devos K. M. 1998 Plant comparative genetics after 10 years. *Science* **282**, 656–659.

Gaut B. S., Morton B. R., McCaig B. C. and Clegg M. T. 1996 Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. USA* **93**, 10274–10279.

Gaut B. S. and Doebley J. F. 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA* **94**, 6809–6814.

Giacomelli L., Nanni V., Lenzi L., Zhuang J., Serra M. D., Banfield M. J. *et al.* 2012 Identification and characterization of the defensin-like gene family of grapevine. *Mol. Plant Microbe Interact.* **25**, 1118–1131.

Graham M. A., Silverstein K. A., Cannon S. B. and VandenBosch K. A. 2004 Computational identification and characterization of novel genes from legumes. *Plant Physiol.* **135**, 1179–1197.

Keller B. and Feuillet C. 2000 Colinearity and gene density in grass genomes. *Trends Plant Sci.* **5**, 246–251.

Kellogg E. A. 2001 Evolutionary history of the grasses. *Plant Physiol.* **125**, 1198–1205.

Lawton-Rauh A. 2003 Evolutionary dynamics of duplicated genes in plants. *Mol. Phylogenet. Evol.* **29**, 396–409.

Lay F. T. and Anderson M. A. 2005 Defensins—components of the innate immune system in plants. *Curr. Protein Pept. Sci.* **6**, 85–101.

Li Z., Jiang H., Zhou L., Deng L., Lin Y., Peng X. *et al.* 2014 Molecular evolution of the HD-ZIP I gene family in legume genomes. *Gene* **533**, 218–228.

Librado P. and Rozas J. 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452.

Maher C, Stein L. and Ware D. 2006 Evolution of *Arabidopsis* microRNA families through duplication events. *Genome Res.* **16**, 510–519.

Meyer B., Houlne G., Pozueta-Romero J., Schantz M. L. and Schantz R. 1996 Fruit-specific expression of a defensin-type gene family in bell pepper. Upregulation during ripening and upon wounding. *Plant Physiol.* **112**, 615–622.

Muthamilarasan M. R., Khandelwal C. B., Yadav V. S., Bonthala Y. Khan and Prasad M. 2014 Identification and molecular characterization of MYB transcription factor superfamily in C4 model plant foxtail millet (*Setaria italica L.*) *PLoS One* **9**, e109920.

Nei M. and Gojobori T 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426.

Paterson A. H., Bowers J. E., Burow M. D., Draye X., Elsik C. G., Jiang C. X. *et al.* 2000 Comparative genomics of plant chromosomes. *Plant Cell* **12**, 1523–1540.

Paterson A. H., Bowers J. E. and Chapman B.A. 2004 Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. USA* **101**, 9903–9908.

Prema G. and Pruthvi T. 2012 Antifungal plant defensins. *Curr. Biotica* **6**, 254–270.

SanMiguel P., Gaut B. S., Tikhonov A., Nakajima Y. and Bennetzen J. L. 1998 The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–50.

Sato S., Nakamura Y., Kaneko T., Asamizu E., Kato T., Nakao M. *et al.* 2008 Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227–239.

Silverstein K. A., 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., Moskal W. A. Jr., Wu H. C., Underwood B. A., Graham M. A. *et al.* 2007 Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J.* **51**, 262–280.

Stotz H. U, Thomson J. G. and Wang Y 2009 Plant defensins: defense, development and application. *Plant Signal Behav.* **4**, 1010–1012.

Swigonova Z., Lai J. S., Ma J. X., Ramakrishna W., Llaca V., Bennetzen J. L. and Messing J. 2004 Close split of sorghum and maize genome progenitors. *Genome Res.* **14** 1916–1923.

Tamura K., Dudley J., Nei M. and Kumar S. 2007 MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.

Thomma B. P, Cammue B. P. and Thevissen K. 2002 Plant defensins. *Planta* **216**, 193–202.

Thompson J. D, Gibson T. J. and Higgins D. G. 2002 Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* 2–3.

Vision T. J, Brown D. G. and Tanksley S. D. 2000 The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117.

Wang X., Shi X., Hao B., Ge S. and Luo J. 2005 Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* **165** 937–946.

Wei F., Coe E. D., Nelson W., Bharti A. K., Engler F., Beetler E. and Fuks G. 2007 Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet.* **3**, e123.

Yang D., Biragyn A., Hoover D. M., Lubkowski J. and Oppenheim J. J. 2004 Multiple roles of antimicrobial defensins, cathelicidins, and eosinophil-derived neurotoxin in host defense. *Annu. Rev. Immunol.* **22**, 181–215.

Received 15 April 2015, in revised form 8 June 2015; accepted 6 July 2015
Unedited version published online: 9 July 2015
Final version published online: 20 January 2016