Insect Molecular Biology

Insect Molecular Biology (2011) **20**(6), 745–762 doi: 10.1111/j.1365-2583.2011.01104.x

Pyrosequencing the midgut transcriptome of the brown planthopper, *Nilaparvata lugens*

X. Peng*, W. Zha*, R. He†, T. Lu‡, L. Zhu*, B. Han‡ and G. He*

**State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan, China; †Arizona Genomics Institute, University of Arizona, Tucson, AZ, USA; and ‡The National Center for Gene Research, Institute of Plant Physiology and Ecology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai, China*

Abstract

The brown planthopper, *Nilaparvata lugens*, is a serious pest threatening rice production across the world. To identify the main features of the gene expression and the key components of the midgut of *N. lugens* responsible for nutrition, xenobiotic metabolism and the immune response, we used pyrosequencing to sample the transcriptome. More than 190 000 clean sequences were generated, which led to about 30 000 unique sequences. Sequence analysis indicated that genes with abundant transcripts in the midgut of *N. lugens* were mainly sugar hydrolyases and transporters, proteases and detoxification-related proteins. Based on the sequence information, we cloned the candidate sucrase gene; this enzyme is likely to interact with the perimicrovillar membrane through its highly hydrophobic C-terminal region. Many proteases were identified, which supported the hypothesis that *N. lugens* uses the proteolysis system for digestion. Scores of detoxification genes were newly identified, including cytochrome P450s, glutathione S-transferases, carboxylesterases. A wealth of new transcripts possibly participating in the immune response were described as well. The gene encoding a peptidoglycan recognition protein was cloned. Unlike in *Acyrthosiphon pisum*, the immunodeficiency pathway may be present in *N. lugens*. This is the first global analysis of midgut transcriptome from *N. lugens*.

Keywords: *Nilaparvata lugens*, midgut, pyrosequencing, sucrase, xenobiotic metabolism, innate immune, peptidoglycan recognition protein.

Introduction

The brown planthopper (*Nilaparvata lugens* Stål; Hemiptera, Delphacidae) is a specialist herbivore of rice (*Oryza sativa* L.). One of the most destructive pests in the world, *N. lugens* causes physiological abnormalities to rice plants, and transmits rice viruses (Wei *et al*., 2009). Traditionally, rice farmers depend on chemical insecticides to control *N. lugens* (Hao *et al*., 2008), but the long-term application of insecticides has caused *N. lugens* to develop resistance, which may result in pest resurgence and environmental pollution (Liu *et al*., 2003; Cohen, 2006). Another strategy for *N. lugens* control is to use host-plant resistance by breeding rice varieties with *N. lugens* resistance genes through molecular marker-assisted selection; however, *N. lugens* can overcome these rice varieties with resistance loci (Jena & Kim, 2010). Identifying the components involved in important physiological processes, e.g. nutrition, xenobiotic metabolism and immune response, especially in the midgut of *N. lugens*, will facilitate an understanding of the adaptation mechanism of *N. lugens* to host plants and the development of more suitable methods to control this agricultural pest.

The midgut is the portion of the insect's body that actively interacts with the physical environment (Wang & Granados, 2001). This tissue plays several critical roles, including nutrition digestion and uptake, detoxification and defence. It has been predicted that sucrases in the midgut of phloem-sucking insects break down sucrose into absorbable glucose and fructose and polymerize glucose into oligosaccharides to reduce the osmotic pressure in the alimentary canals (Price *et al*., 2007a, b). Although scores of genes encoding for sugar transporters have been cloned, the enzymes responsible for sugar

First published online 15 September 2011.

Correspondence: Guangcun He, College of Life Sciences, Wuhan University, Wuhan 430072, China. Tel.: +86 27 68752384; fax: +86 27 68752327; e-mail: gche@whu.edu.cn

hydrolysis have not been systemically identified in *N. lugens* (Price *et al*., 2007b; Kikuta *et al*., 2010). Traditionally, it was thought that proteases were not employed by sap-sucking phytophagous insects for digestion, because there were not enough proteins in phloem sap from plants; however, recent results suggest otherwise (Cristofoletti *et al*., 2003; Rispe *et al*., 2008). To date, two trypsin-like genes and one cathepsin B-like gene have been cloned in *N. lugens* (Foissac *et al*., 2002; Yang *et al*., 2005), but it remains unclear whether or not proteases are used by *N. lugens* for digestion.

Research on insecticide resistance of many insects has highlighted that genes involved in xenobiotic metabolism are essential for environment adaptation. Meanwhile, it has also been shown that the transcripts of the key components for this process, including cytochrome P450s (CYPs), glutathione S-transferases (GSTs) and carboxylesterases (COEs), can be induced in the midgut (Enayati *et al*., 2005; Chung *et al*., 2009; Yu *et al*., 2009). Since resistant rice varieties show higher levels of toxic compounds (Bing *et al*., 2007), it is reasonable to assume that detoxifying genes are related to the adaptation of *N. lugens* to host plants (Yang *et al*., 2005, 2010). Genome-wide annotations of model insects have indicated that xenobiotic metabolism genes are all multifunctional super-families and some members may evolve very rapidly (Ranson *et al*., 2002; Chung *et al*., 2009; The International Aphid Genomics Consortium, 2010). It is therefore worthwhile to discover the diversities and identify new members of those genes in the midgut of *N. lugens* using transcriptome analysis.

The midgut of insects is also an important site for pathogen entry (Hakim *et al*., 2010), but the immune function of insect midguts has been ignored by most studies (Siva-Jothy *et al*., 2005). Pathogens in insects trigger systemic responses mainly through four defence mechanisms: the Toll pathway, the immunodeficiency (Imd) pathway, the c-Jun N-terminal kinase (JNK) pathway and the Janus kinase/signal transducers and activators of transcription (JAK/STAT) pathway (Gerardo *et al*., 2010). *N. lugens* needs yeast-like endosymbionts to acquire essential substances for survival and harbours several kinds of bacterial microbes (Suh *et al*., 2001; Tang *et al*., 2010). Three viruses have been characterized in *N. lugens*, including *Nilaparvata lugens reovirus* (NLRV), Himetobi P virus (HiPV) and *Nilaparvata lugens* commensal X virus (NLCXV) (Nakashima *et al*., 2006). In addition, the natural enemies of *N. lugens* nymphs and adults contain 16 parasites belonging to Hymenoptera, Strepsiptera and Diptera, one species of nematode and seven species of fungus (Chiu, 1979). The basic immune response in *N. lugens* at the molecular level, especially in the midgut, however, has not been studied.

As an economically important pest with currently unavailable genomic data, expressed sequence tag (EST) collection is a feasible method for new gene discovery. Noda *et al*. performed the first extensive EST collection in *N. lugens* from 15 different tissues or developmental stages, including the midgut, through Sanger technique, and generated more than 37 000 EST sequences (Noda *et al*., 2008). They found a trypsin-like protease gene abundantly expressed in the midguts of male or female adults. Xue *et al*. used short-read sequencing technology (Illumina) to compare the gene expressions for developmental stages, sexes and wing forms of *N. lugens* (Xue *et al*., 2010). This research has shown important information and provided valuable data for further research on gene functions in *N. lugens*; however, tissue-specific transcriptome analysis is still insufficient. More detailed gene expression information is needed to identify the features in the midgut at the molecular level.

454, pyrosequencing, a high throughput technique developed by 454 Life Sciences, is popularly used to perform deep sequencing of transcriptomes (Margulies *et al*., 2005). Using sequences generated by pyrosequencing, extensive gene discovery programmes and genetic analyses are feasible for non-model insect species (Droege & Hill, 2008), including *N. lugens*. In the present study we applied pyrosequencing to sample the transcriptome of the *N. lugens* nymph midgut. In combination with full-length cDNA cloning and PCR validation, we detailed the gene expression profile in the midgut of *N. lugens* and detected key components that may play important roles in nutrition digestion, xenobiotic metabolism and the innate immune response. Our results could broaden our insight into the functions of the midgut of *N. lugens* at the molecular level and advance the understanding of the molecular mechanism on the adaptation of this pest to the environment. Furthermore, the data collected in the present study provide a valuable resource for further functional genomics study of the midgut of *N. lugens*.

Results

Pyrosequencing, assembly and similarity analysis of Nilaparvata lugens *midgut transcriptome*

Through pyrosequencing, we obtained a total of 323 957 reads, 193 168 among which were clean, high quality sequences of ESTs (Table 1). After assembly, 10 848 contigs and 20 650 singletons were generated. The contig sizes varied considerably. More than 40% of the contigs consisted of two reads, and 310 contigs consisted of 50 reads or more. For the unique sequences, contigs and singletons, we performed homology searches against the non-redundant protein database and NCBI est_others database, respectively. Among unique sequences, ≈15%

Table 1. Summary statistics for *Nilaparvata lugens* midgut expressed sequence tag assembly and annotation

Assembly	
Total number of reads	323 957
Average read length	189
Number of reads that entered in the assembly	193 168
Total number of contigs	10 848
Total number of singletons	20 650
N50 contig size	289 bp
Annotation	
%unique sequences with a blast hit against nr	15.2%
%unique sequences with a blast hit against est_others	38.1%
%unique sequences with at least one GO term	8.6%*
%unique sequences with an EC number	2.9%*

*The transcripts identified from viruses, NLRV, HiPV, TATV and CcBV, were not included in the analysis.

(4773 unique sequences) showed significant matches with proteins deposited in the public database. More than 38.1% were matched to nucleic acid sequences from the NCBI est_others database. The *e*-value and similarity distributions of these BLASTX hits are shown in Fig. 1A and B. In all, 62 contigs and 28 singletons were matched to sequences from four viruses, NLRV, HiPV, *Taterapox virus* (TATV), and *Cotesia congregata bracovirus* (CcBV): 52, 12, nine and three unique sequences, respectively. Previous research has shown that the midgut is a major infection and propagation site of HiPV (Nakashima *et al*., 2006), which is consistent with our results. CcBV is a mutualism virus with endoparasitic wasps (Friedman & Hughes, 2006), therefore, some *N. lugens* might harbour this virus as a result of parasitic wasps. These virus-derived sequences were not included in the process of further analysis.

BLASTX analysis suggested that the most frequently hit proteins were from *Tribolium castaneum* (Coleoptera; 3500 matches), followed by those from *Nasonia vitripennis* (Hymenoptera), *Acyrthosiphon pisum* (Hemiptera), *Apis mellifera* (Hymenoptera) and *Pediculus humanus* (Psocodea) (Fig. 1C). To analyse the similarities between the sequences from the midgut of *N. lugens* and model insects in detail, we further searched our sequences against *ab initio*-predicted proteins from *A. pisum* and *T. castaneum*, as well as *Drosophila melanogaster*. As expected, the annotation percentage was significantly positively correlated with query length (Fig. 2A). Most identifications by proteins from the three species were overlapped. A total of 4390 sequences (14.0%) were identified when combining the proteins from the three insect models ($e < 10^{-5}$). In addition, 9.23% had homologies to those proteins from at least one gene model at an *e*-value of 10^{-10}, and 4.05% at an *e*-value of 10^{-20} (Fig. 2B). Noda *et al*. sequenced approximately 37 000 ESTs from *N. lugens*, including some sequences from the midgut (Noda *et al*., 2008). Through alignment analysis, we found that 38.0% of our unique midgut sequences overlapped with the ESTs sequenced by Noda *et al*.; 64.0% of their EST sequences were found among our unique sequences.

Identification of simple sequence repeat loci

Microsatellites, or simple sequence repeats (SSRs), are tandem repeated motifs of 1–6 bases that have been widely used in population and conservation genetics, molecular epidemiology and pathology, and gene mapping (Chistiakov *et al*., 2006). By screening all the unique sequences, 478 microsatellite loci with di-, tri-, and tetra-nucleotide repeats were detected in the *N. lugens* midgut transcriptome data. SSRs with tri-nucleotide repeats were the most abundant (344 loci, 72.0%) and the most frequent motifs were (CTT)n (38 loci) and (AGA)n (33 loci). In addition, there were 70 loci with 4-base pair (bp) repeat units and 64 SSR loci with 2-bp repeat units. These SSRs represent a valuable resource for future genetic analysis of *N. lugens*.

Gene ontology and protein functions

Gene ontology (GO) assignments, enzyme classifications (EC), as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) classifications, were applied to the predicted midgut proteins. In total, 10 201 terms were assigned to 2722 sequences, and 1195 EC numbers were assigned to a total of 918 unique sequences (Table 1). Catalytic, binding, cellular process and metabolic activity are dominant, as expected for insect midgut proteins (Fig. 3). 1006 sequences were mapped to currently known KEGG pathways with 101 unique maps. Most sequences with EC annotations were involved in energy metabolism, and other pathways well represented by unique sequences from the *N. lugens* midgut were carbohydrate metabolism and nucleotide metabolism (Table 2). Energy metabolism was found to be the most active pathway in the midgut. Twenty-six sequences were predicted to be involved in xenobiotic biodegradation and metabolism. In addition, 12 sequences were mapped to signal transduction pathways in the environmental information process category, while 18 sequences were mapped to translation pathways in the genetic information process category and 10 sequences were mapped to immune system pathways.

The InterPro database was also employed to determine the likely functions of predicted proteins in the midgut and a summary of the 25 most frequent classifications is listed in Table S1. Conserved domains from peptidases, glycoside hydrolases, and components responsible for xenobiotic metabolism (CYP, Thioredoxin fold) were dominant. Immunoglobulin-like fold was also frequent (Table S1).

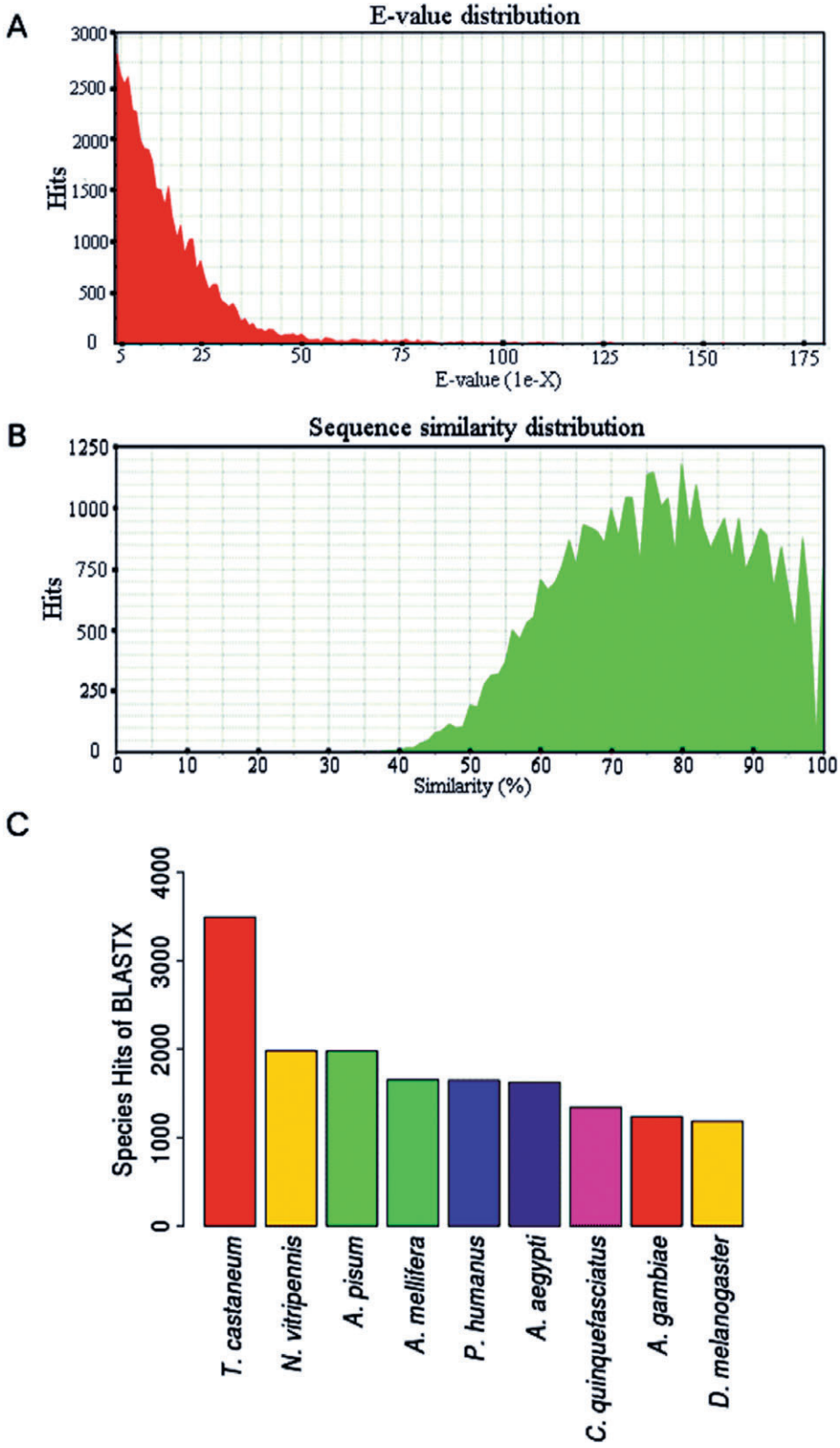

Figure 1. Summary of homology searches (BLASTX) of *Nilaparvata lugens* midgut 454 data against the non-redundant protein database. (A) The *e*-value distribution of the top BLASTX hits for each unique sequence. The cut-off used in this case was 10^{-3} and HSP $\geq$ 33. (B) Similarity distribution of the top BLAST hits for each unique sequence. (C) The most frequent species distribution of the BLASTX hits. Species abbreviations: *Tribolium casteneum*; *Nasonia vitripennis*; *Acyrthosiphon pisum*; *Apis mellifera*; *Pediculus humanus*; *Aedes aegypti*; *Culex quinquefasciatus*; *Anopheles gambiae*; *Drosophila melanogaster*.

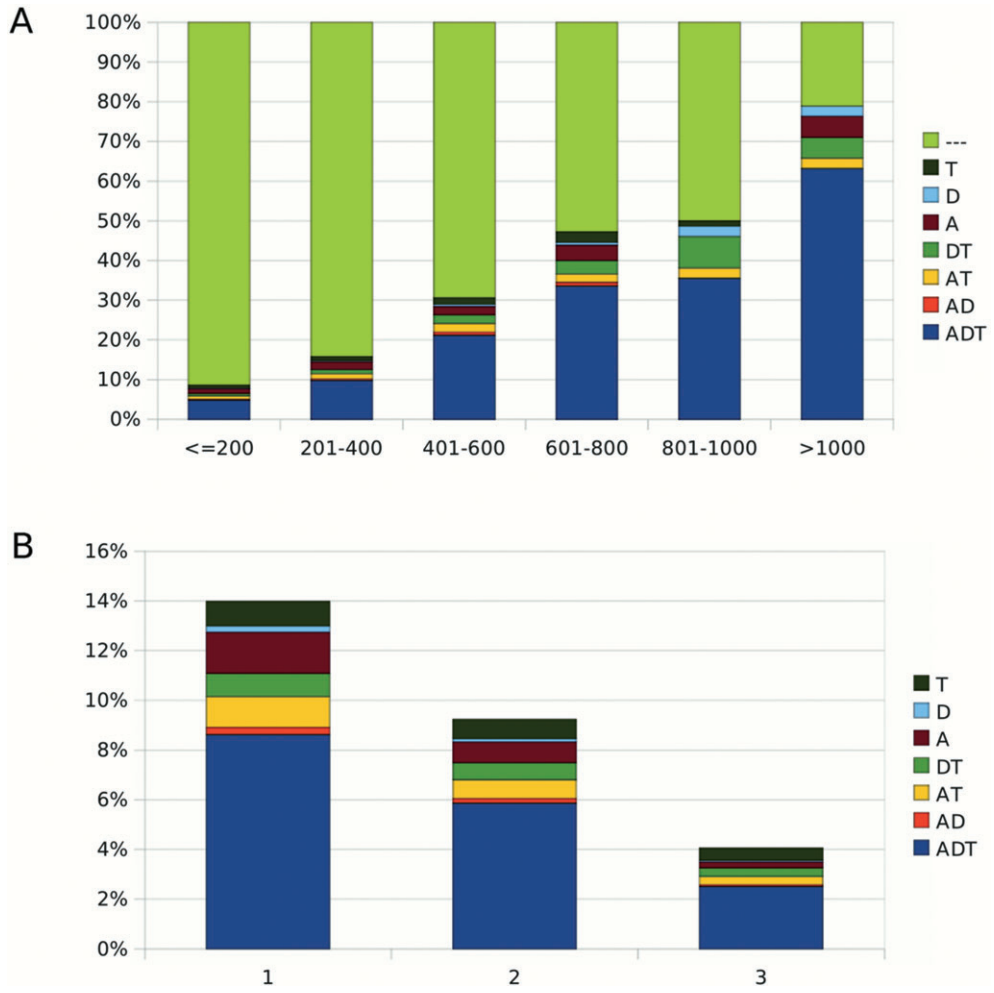

Figure 2. Proportions of *Nilaparvata lugens* unigenes with homology to gene models. Proportions of *N. lugens* unigenes (contigs + singletons) with and without (-) homology to aphid (A, *Acyrthosiphon pisum*,), fruitfly (D, *Drosophila melanogaster*,), and red flour beetle (T, *Tribolium castaneum*) gene models. (A) Effect of the sequence length on the proportion of homology to gene models (*e*-value 10^{-5}). (B) Proportion of *N. lugens* unique sequences with and without homology to gene models at three different *e*-values ($1{:}10^{-5}$, $2{:}10^{-10}$, $3{:}10^{-20}$). We used BLASTX to compare our sequences with *ab initio* protein predictions of *A. pisum*, all translation-r5.26 proteins of fruitfly, and *ab initio* protein predictions of red flour beetle.

Identification of the most abundant transcripts in the midgut

Transcript levels of specific genes can be quantified using their EST abundance in the library (Khajuria *et al.*, 2009). There were 310 contigs with at least 50 reads each, among which 147 sequences can be annotated by BLASTX and 29 were determined as transcripts from mitochondria. The most abundant cluster was Contig8098, which was represented by 16 890 high quality reads. This sequence could not be annotated by BLASTX, but showed high similarity to some sequences in the NCBI est_others database. These sequences had been deposited as ESTs from the *Oryza* plants. One distinct example was EST00447 from a subtracted cDNA library of *Oryza minuta* (accession number: CD026509, *e* = 10–161, score = 579). Meanwhile, when compared with sequences from the NCBI_nr nucleotide database, Contig8089 was best matched with AY166969, an *O. sativa* resistance-related non-coding RNA gene, *nc1* (*e*-value = 0, score = 2187). It was interesting that a *N. lugens* gene that was highly expressed in the midgut was significantly matched to that from a rice plant. Table 3 shows well annotated sequences with at least 50 reads, except components from mitochrondria. The results suggested that transcripts for serine and cysteine peptidases, sugar hydrolyases and transporters, detoxification-related proteins were abundant in the midgut of *N. lugens*. In addition, some unique sequences were also identified as abundant transcripts, sharing no similarity to proteins in current databases.

To validate the expression of transcripts identified through transcriptrome sequencing, 32 unique sequences were selected for reverse transcription-PCR (RT-PCR)

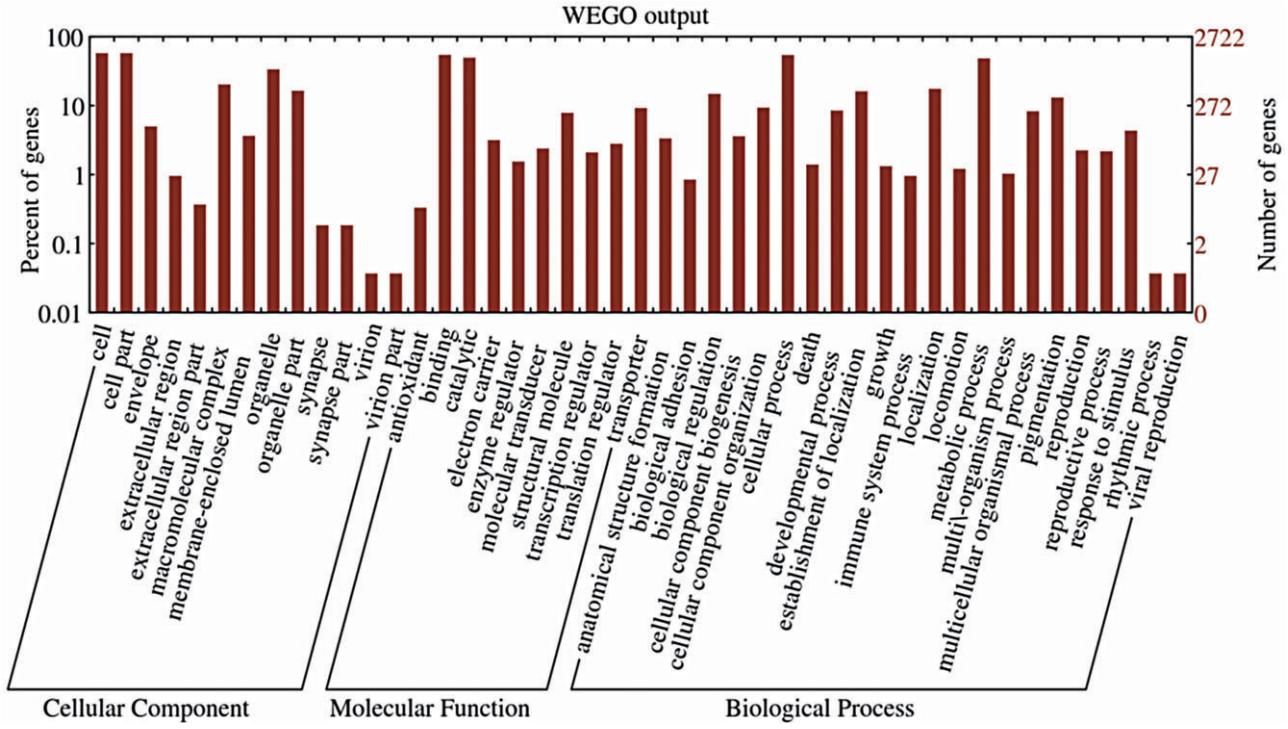

Figure 3. Gene ontology annotation of the predicted proteins in the midgut of *Nilaparvata lugens*.

analysis (Fig. 4A). The results confirmed that the abundant transcripts were expressed highly in the midgut of *N. lugens* and some also in heads, legs and fat body. The RT-PCR analysis on sequences involved in sugar and protein hydrolysis, xenobiotic metabolism and immune response also validated that these genes were expressed in the midgut of *N. lugens* and some may be expressed mainly in the midgut (Fig. 4A).

Table 2. Kyoto Encyclopedia of Genes and Genomes (KEGG) biochemical mappings for unique sequences annotated by gene ontology with enzyme classification number

KEGG categories represented	Unique sequences	Percentage
Metabolism		
Amino acid metabolism	80	7.95
Biosynthesis of other secondary metabolites	19	1.89
Carbohydrate metabolism	130	12.92
Energy metabolism	182	18.09
Lipid metabolism	40	3.98
Glycan biosynthesis and metabolism	24	2.39
Biosynthesis of polyketides and terpenoids	3	0.30
Metabolism of other amino acids	25	2.49
Metabolism of cofactors and vitamins	49	4.87
Nucleotide metabolism	95	9.44
Xenobiotics biodegradation and metabolism	26	2.58
Overview	293	29.13
Environmental information processing		
Signal transduction	12	1.19
Genetic information processing		
Translation	18	1.79
Organismal systems		
Immune system	10	0.99

Transcripts for important proteins identified in Nilaparvata lugens *midgut*

Enzymes involved in sugar hydrolysis. Through GO vocabulary analysis, we noted that 22.6% of terms from molecular function annotation showed hydrolase activity, which included 6.6% for peptidase activity, 2.9% for hydrolase, acting on ester bonds, 0.8% hydrolase activity, acting on glycosyl bond. In all, three kinds of enzymes responsible for sugar hydrolysis were identified, including alpha-glucosidase (maltase), beta-glucosidase and alpha-amylase (Table S2). Interestingly, we identified three sequences encoding enzymes participating in cell-wall hydrolysis or cellulose hydrolysis. Contig2824 could be translated into a polypeptide with 46 amino acids, which could be well matched to the cellulase from the Gram-positive bacteria, *Streptomyces avermitilis* MA-4680 (*e*-value = 4e-20, identities = 96%, at the protein level). Furthermore, a singleton, 225026_3593_0722, was also determined as a transcript for an enzyme with cell wall-associated hydrolyase activity. The deduced polypeptide of this singleton showed high similarity to the cell wall-associated hydrolases from gammaproteobacteria (*e*-value 9e-30, identities = 99%, at the protein level). In

Table 3. The most abundant transcripts identified in the midgut of *Nilaparvata lugens*

Contig ID	Tentative annotation	Conserved domain		Number of reads
Contig1263	Ribosomal protein l8e	IPR002171	Ribosomal protein L2	3705
Contig2681	Ribosomal protein l6e			2990
Contig5009	Chymotrypsin-like protein			2304
Contig1950	Kallikrein related-peptidase 9			861
Contig9932*	Ferritin			731
Contig3650	Ribosomal protein l27e	IPR001141	Ribosomal protein L27e	703
Contig4596	rrna promoter binding protein			692
Contig4010	60 s ribosomal protein l7	IPR016082	Ribosomal protein L30, ferredoxin-like fold domain	340
Contig8460*	Trypsin	IPR001254	Peptidase S1/S6, chymotrypsin/Hap	325
Contig5238*	Facilitative hexose transporter 1	IPR003663	Sugar/inositol transporter	308
Contig913*	Calcineurin-like phosphoesterase domain containing 1			253
Contig6765*	Juvenile hormone esterase	IPR002018	Carboxylesterase, type B	237
Contig9910*	Cathepsin b	IPR000169	Peptidase, cysteine peptidase active site	220
Contig1616	af373879_1peritrophin-like protein 1	IPR002557	Chitin binding domain	196
Contig5315	Legumain	IPR001096	Peptidase C13, legumain	178
Contig8238*	Phosphatidylserine decarboxylase	IPR003817	Phosphatidylserine decarboxylase-related	176
Contig3742	Gamma-interferon-inducible lysosomal thiol reductase	IPR004911	Gamma interferon inducible lysosomal thiol reductase GILT	160
Contig2069*	Ferritin 2 isoform 1	IPR001519	Ferritin	159
Contig554	Cell division	IPR000038	Cell division/GTP binding protein	155
Contig10241	Glutamate dehydrogenase (nad +)			155
Contig7892	Ribosomal protein l34	IPR008195	Ribosomal protein L34e	130
Contig8855	Ribosomal protein s11	IPR000266	Ribosomal protein S17	128
Contig892	Viral a-type inclusion protein			126
Contig2674	Ribosomal protein s2	IPR000851	Ribosomal protein S5	116
Contig868	Cathepsin b			115
Contig7653*	Kallikrein related-peptidase 6 precursor	IPR001254	Peptidase S1/S6, chymotrypsin/Hap	115
Contig10180	Inositol-3-phosphate synthase a	IPR002587	Myo-inositol-1-phosphate synthase	115
Contig5698	Cathepsin b-like cysteine proteinase precursor	IPR000169	Peptidase, cysteine peptidase active site	109
Contig7212	Gamma-interferon inducible lysosomal thiol	IPR004911	Gamma interferon inducible lysosomal thiol reductase GILT	108
Contig7606	Legumain precursor (asparaginyl endopeptidase) (cysteine 1)			103
Contig5693	Cathepsin b precursor	IPR000169	Peptidase, cysteine peptidase active site	96
Contig7435	Apolipoprotein d	IPR000566	Lipocalin/cytosolic fatty-acid binding protein	94
Contig3010	Ribosomal protein s9			93
Contig10774	Elongation factor 1-alpha	IPR004160	Translation elongation factor EFTu/EF1A, C-terminal	93
Contig2525	Ribosomal protein l23	IPR000218	Ribosomal protein L14b/L23e	89
Contig9748	Maltase precursor	IPR013781	Glycoside hydrolase, subgroup, catalytic core	88
Contig9834*	Alpha-glucosidase	IPR006047	Glycosyl hydrolase, family 13, catalytic domain	86
Contig4652	Transcription factor iifalpha	IPR008851	Transcription initiation factor IIF, alpha subunit	85
Contig283	Ribosomal protein l7ae	IPR002415	H/ACA ribonucleoprotein complex, subunit Nhp2, eukaryote	84
Contig4208	Alkaline phosphatase			84
Contig5608	Translationally controlled tumor protein	IPR011057	Mss4-like	83
Contig6404	Charged multivesicular body protein	IPR005024	Snf7	81
Contig8958	Trypsin-like protease			78
Contig3383	Serine protease	IPR001254	Peptidase S1/S6, chymotrypsin/Hap	77
Contig1103	Ribosomal protein l14	IPR002784	Ribosomal protein L14	73
Contig2380	40 s ribosomal protein s20	IPR001848	Ribosomal protein S10	72
Contig9328	Ribosomal protein s24e	IPR001976	Ribosomal protein S24e	69
Contig6269	Trypsin 7	IPR001254	Peptidase S1/S6, chymotrypsin/Hap	68
Contig4945	phd finger protein			64
Contig5591*	Cytochrome p450	IPR001128	Cytochrome P450	62
Contig10418	Ribosomal protein l36			62
Contig186	Lim and sh3 domain protein			61
Contig4961	Ribosomal protein 49	IPR001515	Ribosomal protein L32e	61
Contig5487	40 s ribosomal protein sa			60
Contig9799	Peptidase c13 family protein	IPR001096	Peptidase C13, legumain	60
Contig673	Draper	IPR013032	EGF-likeregion, conserved site	
Contig572	Formin 3 cg33556-pb			59
Contig5265	Cathepsin b-like cysteine proteinase precursor			59
Contig6163	Ferritin	IPR001519	Ferritin	56
Contig9214	Glyceraldehyde-3-phosphate dehydrogenase			56
Contig1130	Septin 7a			55
Contig8651	Cardioacceleratory peptide receptor			55

Table 3. *Continued*

Contig ID	Tentative annotation	Conserved domain		Number of reads
Contig4076	Conserved plasmodium protein			54
Contig5756	Ribosomal protein s15a	IPR000630	Ribosomal protein S8	54
Contig6229	Dipeptidyl peptidase iii	IPR005317	Peptidase M49, dipeptidyl-peptidase III	54
Contig1100	dna-binding protein			53
Contig8566	Peptidyl-prolyl *cis-trans* isomerase cyp6			53
Contig2369	ctlh2_anoga ame: full = ctl-like protein 2			52
Contig6026	Cathepsin l-like proteinase	IPR013128	Peptidase C1A, papain	52
Contig7705	Ribosomal protein l31	IPR000054	Ribosomal protein L31e	52
Contig7156	Mob2	IPR005301	Mob1/phocein	52
Contig2474	Ribosomal protein s11			51
Contig5086	Midgut trypsin	IPR001254	Peptidase S1/S6, chymotrypsin/Hap	51
Contig5375	Transcription initiation factor iif alpha subunit	IPR008851	Transcription initiation factor IIF, alpha subunit	51
Contig158	Sarcoplasmic calcium-binding protein 2 cg14904-pa	IPR002048	Calcium-binding EF-hand	50
Contig2096	Troponin t	IPR001978	Troponin	50
Contig9619	Ubiquitin specific protease 14 (trna-guanine transglycosylase)			50
Contig10030	Nucleoside diphosphate kinase	IPR001564	Nucleoside diphosphate kinase	50

*These sequences were selected for RT-PCR analysis.
Note: All sequences identified as abundant transcripts were first annotated by BLASTX. Each sequence with an open reading frame (ORF) no less than 150 was extracted, then the largest ORF for each sequence was translated into polypeptide, which was used for InterPro analysis. Only unique sequences that can be clearly annotated are shown. Components identified as transcripts from mitochondria are neglected.

addition, another singleton could be matched well to the putative cellulase from *A. pisum* well (*e*-value = 6e-06, at the nucleotide level) (Table S2).

The polypeptides of the three unique sequences, Contig4453, Contig9834 and Contig2569, which were determined as transcripts for alpha-glucosidases, all had the signature from glycosyl hydrolase, family 13, catalytic domain (IPR006047) (Table S2). Contig9834 is 1206 bp long with the deduced polypeptide of 399 amino acids and RT-PCR analysis indicated that this gene was mainly expressed in the midgut (Fig. 4A). This unique sequence was probably the transcript from the candidate sucrase gene in *N. lugens*. Based on the sequence information of Contig9834, the full-length cDNA sequence (*NlSu1*) was cloned through rapid amplification of cDNA ends (RACE). *NlSu1* was 2000 bp long with a 24-bp 5′untranslated region (UTR), a 1914-bp coding region and a 60-bp 3′UTR region (Fig. 5A). Contig9834 was exactly the middle part of the transcript. The protein encoded by the cDNA comprised 637 amino acids. The polypeptide had a predicted signal peptide, with a putative cleavage site at the carboxyl side of Ala^{29} (SignalP 3.0). Removing the signal peptide led to a mature protein of 608 amino acids with a theoretical Mr of 70 527Da and a predicted pI of 5.23. This protein contained an amylase domain and highly conserved sites to family 13 of the glycosyl hydrolases (Fig. 5A). Furthermore, *NlSu1* grouped with a subset of proteins with known alpha-glucosidase activities (EC 3.2.1.20) and had closer evolutionary distance to the sucrase cloned in *A. pisum* (DQ223541) (Fig. 5B). Hydropathy analysis suggested that the catalytic domain was quite hydrophilic, while two regions were predominantly hydrophobic (Fig. 5C). One was in the N-terminal region, which was the part from the signal peptide. The other is in the C-terminal region. Considering no transmembrane helice or anchor predicted for this protein, the highly hydrophobic C-terminal region could function as a membrane association region.

Proteases. Proteases identified in the midgut of *N. lugens* mainly included trypsin-like or chymotrypsin-like proteins, cathepsins, aminopeptidases, dipeptidyl petidases, and carboxypeptidases. The transcripts of trypsin/chymotrypsin-like or cathepsin-like proteases were expressed highly in the midgut of *N. lugens* (Table 3). In total, 29 unique sequences were identified as trypsin-like or chymotrypsin-like serine protease transcripts. Through the InterProScan analysis, 12 of those trypsin-like sequences were found to contain the protein signature from Peptidase S1/S6, chymotrypsin-Hap (IPR001254) (Table S2). In the meantime, we found top-hits of the trypsin/chymotrypsin like sequences represented 16 proteins from 15 species, which indicated the diversity of trypsin/chymotrypsin in the *N. lugens* midgut. We cloned the full-length cDNA encoding a novel trypsin-like protein (*NlTry-theta*) through screening the midgut cDNA library of *N. lugens*. The *NlTry-theta* was highly expressed in the midgut (Fig. 4B). The deduced protein had 274 amino acids containing a domain belonging to the Tryp_SPc family and may be secreted to the gut after synthesis based on a predicted signal peptide. This new trypsin-like protein had no significant similarity to the two trypsin-like proteins cloned in *N. lugens* earlier (Foissac *et al.*, 2002; Yang *et al.*, 2005). Cathepsins were another kind of pro-

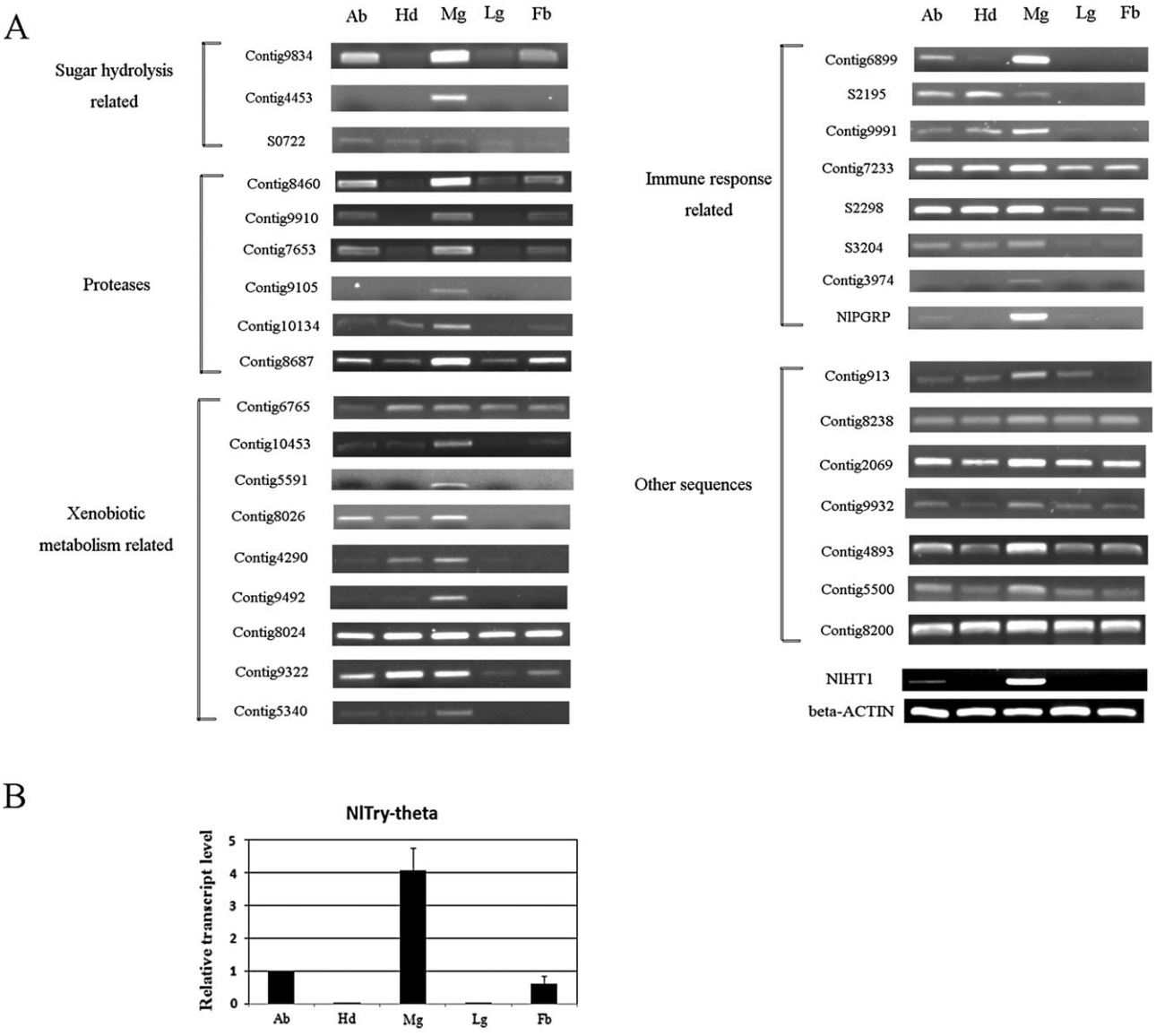

Figure 4. Reverse transcription (RT)-PCR and quantitative real-time PCR analyses to validate the expression of unique sequences in *Nilaparvata lugens*. (A). RT-PCR analysis of the selected genes. *NlHT1* is the sugar transporter gene first cloned by Price *et al.*, which is expressed mainly in the midgut of *N. lugens*. Fourteen unique sequences were from transcripts identified as abundant transcripts, which were listed in Table 3, except Contig8024 (a putative carboxylesterase), Contig8200 (a putative H^+ transporting ATP synthase), Contig4893 (lack of homology to known sequences), Contig5500 (lack of homology to known sequences). S0722, S2195, S2298 and S3204 refer to four singletons: 225026_3593_0722, 060247_2362_2195, 304903_3177_2298, 245955_3925_3204, respectively. (B). qPCR analysis of *NlTry-theta* (GenBank accession number: JF915745). Abbreviations: Ab, abdomen; Hd, head; Mg, midgut; Lg, leg; Fb, fat body. beta-ACTIN was used for normalization and the expression level is shown relative to the abdomen expression level, which is set to one. The mean ± SD was used for analysis of the relative transcript levels for each sample using the $\Delta\Delta C_t$ method.

tease with abundant trancripts. A total of 41 unique sequences were identified as cathepsins, 33 of which were from cathepsin B, in addition to several cathepsin L- and D-like sequences. There are also several sequences for aminopeptidases and carboxypeptidases.

Proteins involved in detoxification. It has been accepted that xenobiotic metabolism mediated by CYPs, GSTs or COEs is a universal mechanism of insecticide resistance in insects (Després *et al.*, 2007; Strode *et al.*, 2008). According to our research, 18 unique sequences coding CYPs were found, which had been first matched by 15 CYP proteins from 13 species (Table S3). Seven CYP-like sequences shared the signature from CYP (IPR001128), while one NADPH-CYP oxidoreductase-like sequence had the protein signature of FAD-binding, type 1 (IPR003097) (Table S3). CYP unique sequences identified here were mostly from the CYP3 clade (12 unique

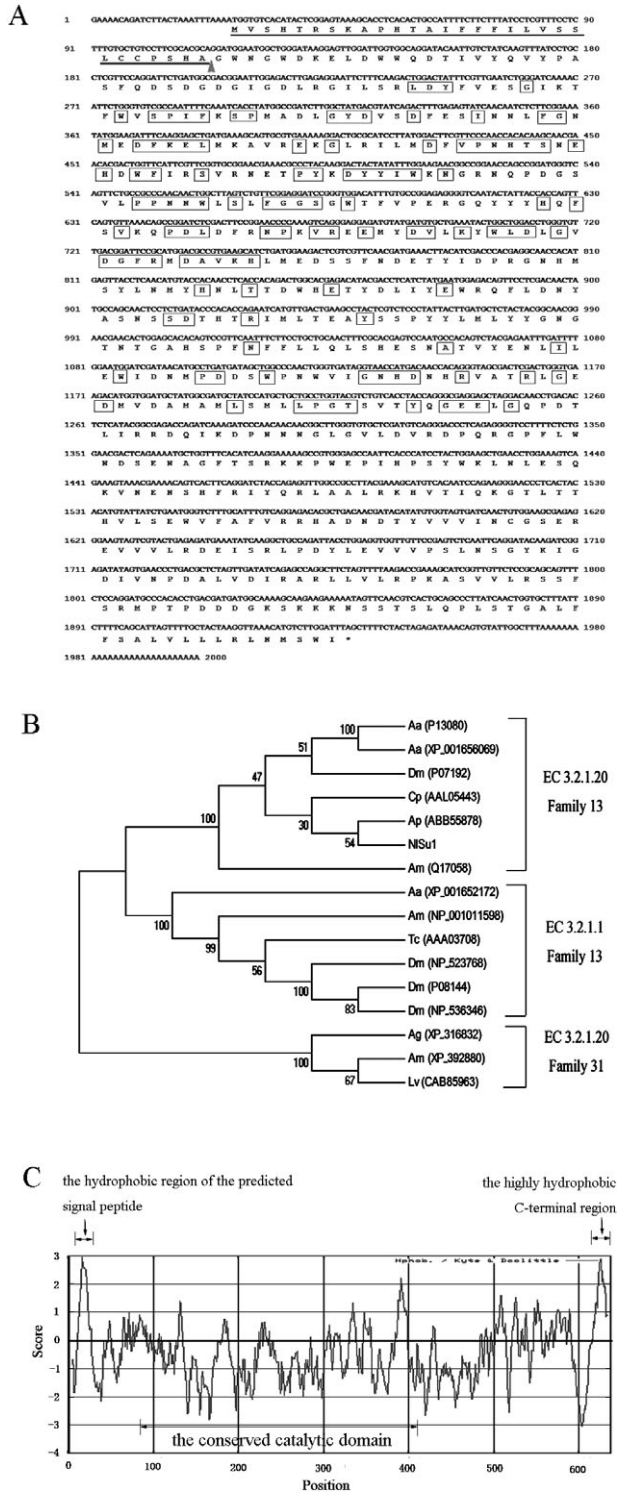

Figure 5. Molecular analysis of the candidate sucrase gene (*NlSu1*) from *Nilaparvata lugens* (GenBank accession number: JN382244). This gene was cloned through rapid amplification of cDNA ends. The unique sequence, Contig9834 identified as the putative sucrase, was used to design specific primers. (A) Nucleotide and deduced amino acid sequence of *NlSu1*. The putative signal peptide, predicted by SignalP 3.0, is underlined. The red arrow shows the cleavage site. Highly conserved regions or sites are boxed, which are identical to Pfam00128. (B) Rooted phylogenetic tree derived from an alignment of selected insect alpha-glucosidase like sequences. Except NlSu1, these sequences are from *Drosophila melanogaster* (Dm), *Acyrthosiphon pisum* (Ap), *Aedes aegypti* (Aa), *Tribolium castaneum* (Tc), *Anopheles gambiae* (Ag), *Litopenaeus vannamei* (Lv), *Culex pipien* (Cp). Numbers on the branches represent level of confidence as determined as by bootstrap analysis (10000 replicates). (C) Hydrophobicity plot (ProtScale) of NlSu1 (hydrophobic +ve, hydrophilic -ve), showed N- and C- ternimal regions.

sequences) and the CYP4 clade (four unique sequences) (Table S3, Fig. 6A), which was consistent with the previous conclusion that genes in the CYP3 or CYP4 clades are numerous among insect P450 genes (Feyereisen, 2006). Phylogenetic analysis suggested that Contig5591 was a member from the CYP4 clade. Compared with other cloned CYP 4 members in *N. lugens*, this gene evolved very rapidly (Fig. 6A).

Eight GST unique sequences were discovered, which distributed in the Delta/Eplison class (five sequences), the Sigma class (two sequences) and the Omega class (one sequence), respectively (Table S3). The polypeptides of Contig4290 and Contig9267 had signatures of GST, C-terminal (IPR010987 and IPR00406). The deduced polypeptide from Contig10257 was a putative GST protein belonging to the Sigma class, while the only one GST cloned in *N. lugens* (AAM21583) was from the Delta/Epsilon class (Fig. 6B). The Sigma GST class distributed in a diverse range of species and was single copy gene in all species studied (Enayati *et al*., 2005); however, the Delta/Epsilon classes of GST are unique to insects and had been linked to specific environment adaption (Ranson *et al*., 2002). In addition, 16 COEs were discovered, including seven sequences for juvenile hormone esterases and one unique sequence for acetycholinesterase.

Immune response-related proteins. The midgut is the most significant portion of the insect's body that interacts with the environment, therefore, its important role in immune response has been emphasized (Zaidman-Rémy *et al*., 2006; Freitak *et al*., 2007; Pauchet *et al*., 2009). The *N. lugens* midgut transcriptome contained a wealth of transcripts corresponding to genes involved in the immune response. PGRPs, Imd protein, Iap2, JNK were present in the midgut of *N. lugens*. In addition, transcripts from GNBP, SR-c and SR class b were identified (Table 4). Other important immune participants, including prophenoloxidase, superoxidase dismutase and peroxidases are listed in Table 3. Some components in the RNA interference (RNAi) pathway, including piwi and argonaute 2-like sequences were also identified in this study, which suggested that the conserved antiviral mechanism was also employed by *N. lugens*.

It has been shown that PGRPs are involved in at least three pathways in the immune response of insects, such

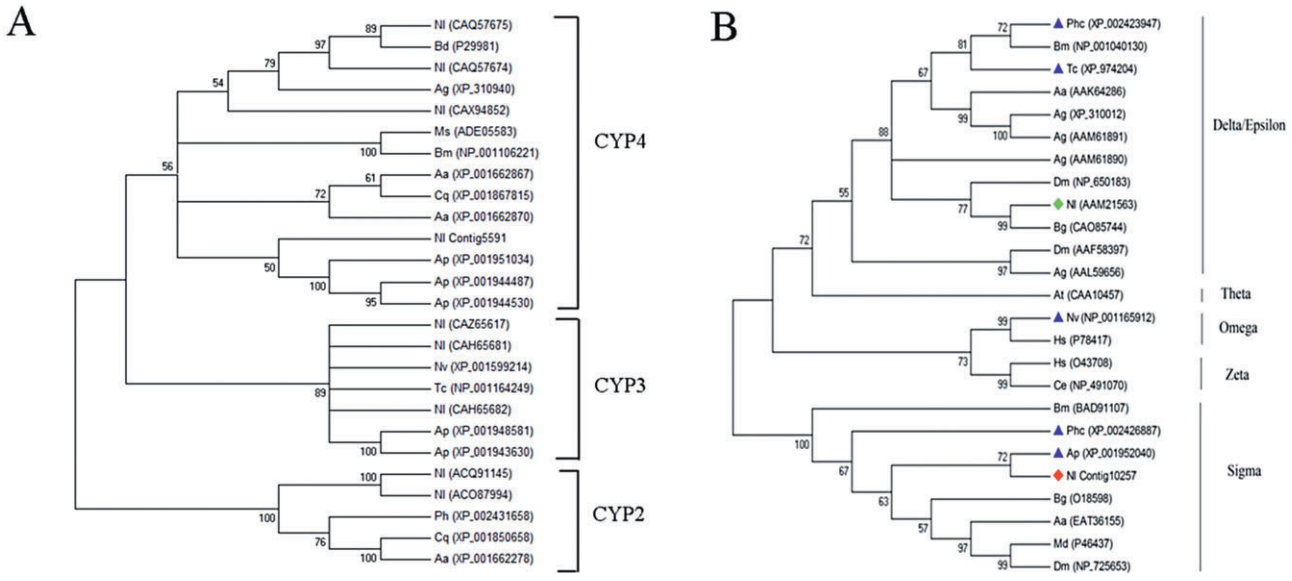

Figure 6. Phylogenetic trees of cytochrome P450s (CYPs) and glutathione S-transferases (GSTs). (A) CYPs. (B) GSTs. The insect CYP genes or GSTs were from *Blaberus discoidalis* (Bd), *Manduca sexta* (Ms), *Mayetiola destructor* (Md), *Tribolium castaneum* (Tc), *Anopheles gambiae* str. PEST (Ag), *Acyrthosiphon pisum* (Ap), *Pediculus humanus corporis* (Phc), *Apis mellifera* (Am), *Aedes aegypti* (Aa), *Culex quinquefasciatus* (Cq), *Nasonia vitripennis* (Nv), *Bombyx mori* (Bm), *Drosophila melanogaster* (Dm), *Blattella germanica* (Bg), *Arabidopsis thaliana* (At), *Homo sapiens* (Hs), *Caenorhabditis elegans* (Ce), and *Nilaparvata lugens* (Nl). The amino acid sequences deduced from Contig5591 and Contig10257 (marked as red triangles) were used for the analysis, respectively. GSTs from non-*N. lugens* species that could be best matched to other GST like unique sequences were marked as blue triangles. The sequences marked as green triangles were GST genes that had cloned from *N. lugens*. The trees were constructed using the neigbour-joining method using MEGA 4.0, besides cut-off value for consensus tree was set to 50%. The numbers on the tree nodes represent the percent bootstap support in 10 000 trials.

as the Imd/JNK pathway, the Toll pathway, and the prophenoloxidase cascade. In the present study, we identified at least four kinds of PGRPs, three of which encoded proteins with signatures from the N-acetylmuramoyl-L-alanine amidase domain (IPR002502) (Table 4). A cDNA clone encoding a putative PGRP protein (NlPGRP) was isolated from the midgut cDNA library, which contained an amidase-activity domain. The deduced polypeptide was 123 amino acid residues. When searched against other PRGPs characterized in model insects, it appeared that the N-terminal region was truncated. Thus, the full-length cDNA of *NlPGRP* was obtained through both 5′ and 3′ RACE. This PGRP gene in *N. lugens* was 968 bp long with a 41-bp 5′UTR, a 651-bp coding region and 276-bp 3′UTR region. NlPGRP had 216 aa residues with a predicted molecular mass of 24.04 kDa. A signal peptide was predicted for NlPGRP and the putative cleavage site was positioned between 33 and 34 (Fig. 7). RT-PCR analysis indicated that *NlPGRP* was mainly expressed in the midgut (Fig. 4A). There was no potential transmembrane helix predicted in this protein.

Discussion

Pyrosequencing is an efficient technique that has been applied for gene expression profile analysis in specific tissues of insects (Pauchet *et al.*, 2009, 2010). In the present study, through deep sequencing, we characterized the transcriptome of the midgut of *N. lugens*, the portion of the insect's body that plays critical roles in nutrition absorption and interaction with the physical environment. In total, we obtained 31 498 unique sequences that expressed in the midgut of *N. lugens*, 38% were overlapped with ESTs released by Noda *et al*. Combined annotations from BLASTX, GO, KEGG and InterPro, we showed that many of the predicted proteins with functions expected for the midgut with likely roles in digestion, xenobiotic metabolism or the innate immune response, whereas many showed no similarity to proteins deposited in current databases. RT-PCR analysis of selected unique sequences validated the genes' expression in the midgut of *N. lugens* (Fig. 4A).

Some sequences showed high similarities to those from microbes or plants. Ninety unique sequences that were putative transcripts from viruses were not included for further annotation. Another seven unique sequences were found to be first matched to proteins from microbes, which include two sequences from gammaproteobacteria, one sequence from alphaproteobacteria, betaproteobacteria, Actinobacteria, Wolbachia endosymbiont and Curvibacter putative symbiont of hydra magnipapillata, respectively. Through homology searching against the est_others database, 124 unique sequences matched ESTs from *Oryza* plants, which included *O. sativa* and *O. minuta*. We also

Table 4. Putative immunity regulatory genes identified

Regulatory protein	Function	Unique sequences
PGRP	Peptidoglycan recognition protein	Contig4334(1.3e-21) IPR002502 Contig6422(3.4e-14) Contig6899(1.3e-22) IPR002502 Contig9586(3.7e-8) IPR002502
GNBP	Gram negative binding protein	Contig1399(1e-5)
scavenger receptor sr-c	Scavenger receptors exhibit broad affinity towards both Gram-positive and Gram-negative bacteria	233582_2700_2048(5.5e-4)
scavenger receptor class b	Scavenger receptors exhibit broad affinity towards both Gram-positive and Gram-negative bacteria	Contig6731(1.1e-9)
Iap2	Inhibitor of apoptosis	Contig9876 (4.1e-12) IPR001370 291634_2779_2433(4.3e-14) Contig5326 (4.4e-9) 001187_1378_3010(7.3e-9) 283435_2726_3949(2.3e-10)
serpin	Serine protease inhibitors	Contig9183 (5.2e-23)
Kazal-type inhibitor	Serine protease inhibitors that belong to Merops inhibitor families	Contig1107(4.2e-4)
Imd	Death domain-containing protein, that similar to receptor interacting protein of TNF-pathway	431404_2150_3349 (0.14)
JNK	c-Jun N-terminal kinase	060247_2362_2195 (3.3e-25)
Toll-like receptor	The key receptor in Toll pathway	Contig6907 (3.5e-6) 164659_2559_0969 (1.1e-4)
MAPK	Mitogen-activated protein kinase	060247_2362_2195(3.3e-25)
JSAP1	jnk sapk-associated protein	Contig5342(3.4e-46) 188705_2880_2945(1.0e-18)
prophenoloxidase	Prophenoloxidase cascade, a part of the antimicrobial defenses in insects and generates melanin in the cuticle	279209_2275_0783 (2e-17)
superoxidase dismutase	Catalyze the dismutation of superoxide into oxygen and hydrogen peroxide	Contig9322 (1.5e-20)
TPX	thioredoxin peroxidase	Contig5220 (8.2e-21) Contig9991 (1.1e-52 165567_3438_0166 (6.0e-19)
GPX	glutathione peroxidase	Contig7233 (6.8e-31) Contig10692 (5.3e-23)
CATs	catalase	Contig5340 (4.1e-19) 338830_2241_1479 (9.1e-28)
nitric oxide synthase	Production nitric oxide, which is toxic to both parasites and pathogens	281537_2794_1691(2.6e-9)
Smt3(SUMO)	Ubiquitin-like proteins	304903_3177_2298 (4.1e-31) 283419_2347_1188(5.5e-17)
transferrin	Ion delivery	Contig9440(2.6e-6)
piwi-like protein	P:gene silencing by RNA	45955_3925_3204(3.9e-18)*
argonaute 2	Catalytic component of the RNA-induced silencing complex	Contig3974(1.0e-18)
Other RNAi-pathway participants	P:RNA interference C:RNA-induced silencing complex	226409_2833_2919(3.4e-14)*
	P:gene silencing	154363_2743_0337(2.4e-23)*
	P:gene silencing by miRNA	Contig5451(1.0e-13)*
	P:chromatin silencing	072382_2168_3148(8.6e-29)*

*The roles of those sequences are shown as gene ontology vocabularies in the function column, and P refers to biological process and C refers to cellular component.

noted that six unique sequences shared high homologies to proteins from rice plants. A total of 71 unique sequences could be significant matched to the EST, CD026509 from *O. minuta*, and 36 unique sequences showed significant homology to the EST, CX114475, from *O. sativa*. At present, for the limited data, it is hard for us to determine if these unique sequences were derived from microbes harboured by *N. lugens*, from the host plant or truly from *N. lugens*.

As for plant phloem sap-sucking insects, sucrase-mediated hydrolysis contributes to two important physiological processes, including carbon acquisition from dietary and osmosregulation (Foissac *et al*., 2002; Price *et al*., 2007a). Research on *A. pisum* has proved that the sucrase, APS1, is located to the posterior midgut and tightly associated with membrane, which is required for sucrase activity (Price *et al*., 2007a). In the midgut of *N. lugens*, three sugar hydrolysis-related genes, including alpha-glucosidase (maltase), beta-glucosidase and alpha-amylase were detected. The beta-glucosidase identified in the midgut may function as a xenobiotic detoxificator, for this enzyme could act on a variety of substances, such as plant glycosides, which are toxic to insects (Pankoke *et al*., 2010). Based on sequence information from transcriptome data, we cloned the candidate sucrase gene (*NlSu1*) in *N. lugens*, which was mainly expressed in the

Figure 7. Nucleotide and the deduced amino acid sequence of NlPGRP (GenBank accession number: JN382245). The putative signal peptide, as predicted by SignalP3.0, is underlined by red line. The red arrow shows the putative cleavage site. The regions or sites highly conserved are boxed.

midgut (Fig. 4A). NlSu1 grouped with a subset of proteins with known alpha-glucosidase activities (EC 3.2.1.20) and had closer evolutionary distance to the sucrase cloned in *A. pisum* (Fig. 5B). Like the sucrase identified in *A. pisum*, the C-terminal region of the protein was predominantly hydrophobic and no transmembrane helix or GPI-anchor site was predicted (Price *et al*., 2007a). Apparently, NlSu1 could use the C-terminal region as a membrane association region, which may ensure the enzyme activity. Employing the hydrophobic region in C-terminus to interact with the perimicrovillar membrane may be a common mechanism used by sucrases in hemiptera.

The transcripts for enzymes involved in proteolysis were highly abundant in the midgut of *N. lugens* (Table 3, Table S2). The diversities of proteases were also noted. For instance, top-hits of trypsin-like unique sequences were 16 proteins from 15 species and at least three kinds of cathepsin B were abundant. The Contig8460, consisted of 325 reads, matched well to the trypsin-like gene cloned in *N. lugens* (CAC87119) (Identities = 283/287, at the protein level). The polypeptide encoded by this sequence had all the three catalytic active sites (H, D, S). The new trypsin in *N. lugens*, NlTry-theta, had a putative signal peptide, and two of the three catalytic active sites were identified. Contig9910 with 220 reads, shared high similarity to the cathepsin B in *N. lugens* (CAC87118) (Identities = 345/347, at the protein level) and the deduced polypeptide had three eukaryotic thiol (cysteine) proteases active sites (cysteine, QGnCGSCWAvSV; histidine, GRHAVKVIGWG; asparagine, YWLvQNSWdydWGdkGLFkI). The RT-PCR analysis suggested all of Contig8460, Contig9910 and Contig9105 (a cathepsin L like transcript) were mainly expressed in the midgut, so did *NlTry-theta*, the transcript level of which was determined by quantitative real-time PCR analysis (Fig. 4). Although transcripts for proteases have been identified in digestive canals of hemipteran insects (Foissac *et al*., 2002; Cristofoletti *et al*., 2003), determining their actual roles is difficult. Recent reports proved the digestive role of some proteases in the midgut of hemiptera (Foissac *et al*., 2002; Cristofoletti *et al*., 2003; Kehr, 2006; Rispe *et al*., 2008). In this experiment, abundant transcripts from different kinds of peptidases, especially serine and cysteine proteases, identified in the midgut strengthen the possibility that proteases may be used by *N. lugens* for digestion. An arguable conclusion is that trypsin appeared early in the digestive systems of invertebrates and during the process of evolution serine proteinase was lost by Auchenorrhyncha-like hemipterans and then lysosomal proteinase was used to digest proteins (Cristofoletti *et al*., 2005; Muhlia-Almazan *et al*., 2008). Thus, in phytophagous hemiptera, cysteine proteases were considered to be the most likely digestive proteases. The digestive roles of cathepsin B, L and D have been shown (Cristofoletti *et al*., 2005; Koo *et al*.,

2008; Padilha *et al*., 2009); however, Foissac *et al*. showed that trypsin-like serine protease from *N. lugens* was responsible for most of hydrolysis activity against synthetic peptide substrates (Foissac *et al*., 2002). The optimal pH environments for trypsin and cathepsin activities are opposing: acidic pH optima for cathepsins and alkaline pH optima for trypsins. The gut pH of *N. lugens* is about 6 (Foissac *et al*., 2002), while rice phloem sap has a pH of ≈8 (Tanaka *et al*., 2009). Cristofoletti *et al*. (2003) found that the pH varies greatly along the midgut of aphid from pH 5.5 to pH 8.5. Thus, enzymes from serine proteases and cysteine proteases may function for digestion in different parts of the midgut. Recent studies on plant phloem suggested that many kinds and significant amounts of proteins are present in phloem saps, including rice phloem sap (Kehr, 2006; Atkins *et al*., 2011). So wasting such a rich nitrogen source may be not economic for *N. lugens*. Proteases in insects have also proved to be participants for development, metamorphosis, even immune response (Cristofoletti *et al*., 2005; Zhao *et al*., 2010). Aminopeptidases and carboxypeptidases have been divided into digestive and regulatory versions in insects (Isoe *et al*., 2009; Lomate & Hivrale, 2010).

In insects, CYPs, COEs, and GSTs are largely responsible for detoxification and insecticide resistance (Ranson *et al*., 2002). Genome-wide annotations have proved that most insects with genome data available have scores of CYP-coding genes, for example, 85 CYPs for *D. melanogaster* (Chung *et al*., 2009). It has been also shown that there are 83 CYPs, 20 GSTs, 29 COEs in the *A. pisum* genome (The International Aphid Genomics Consortium, 2010). Using *D. melanogaster*, Chung *et al*. (2009) showed that 29/85 CYPs were expressed in the midgut. Insect CYPs fall into four clades: the CYP2 clade, the CYP3 clade, the CYP4 clade and the mitochondrial CYP clade. Many genes from the CYP3 clade or the CYP4 clade participate in xenobiotic metabolism, and evolve very rapidly (Feyereisen, 2006). Using transcriptome analysis in the present study, we identified 18 unique sequences encoding CYPs, among which 12 kinds of transcripts were newly identified (Table S3). Most of the identified CYPs were grouped into the CYP3 or CYP4 clade (Table S2). Contig5591, a member from the CYP4 clade, may be restricted to the midgut of *N. lugens* and phylogenetic analysis indicated this gene evolved very rapidly (Figs 4A and 6A). This gene may contribute significantly to the host adaptation of *N. lugens* to rice plants. Among the GSTs, we identified a member from the Sigma class, which is encoded by a single gene in all other characterized insects. This component was more homologous with that from *A. pisum* (XP_001952040). Other members from Delta/Epsilon or Omega classes were also newly identified (Fig. 6B). The Delta/Epsilon class of GSTs is restricted to insects and linked to the adaptation to specific environments (Enayati *et al*., 2005). Meanwhile, 14 COEs were also determined, including one acetylchonesterase and three juvenile hormone esterases (Table S3). Abundant and diverse detoxification-related genes detected here further emphasized the xenobiotic metabolism physiology in the midgut of *N. lugens*.

Insects use an innate immune system for defence. In *D. melanogaster*, microbial agents can activate four signalling pathways, i.e. the Toll pathway, the Imd pathway, the JNK pathway and the JAK/STAT pathway (Boutros *et al*., 2002). However, genomic data have shown that the Imd pathway was completely missing in *A. pisum*, because central participants in the Imd pathway, like PGRPs, IMD, and Iap2, were not identified in the genome (Gerardo *et al*., 2010; The-International-Aphid-Genomics-Consortium, 2010). An Imd-like transcript was identified in *N. lugens* midgut transcriptome data and this insect also has homologies for other Imd/JNK pathway members (Iap2, JNK, PGRP). The results revealed that the missing Imd pathway in *A. pisum* may be present in *N. lugens*. In addition, the scavenger class C, another immune participants missed in *A. pisum*, was also present in *N. lugens*. Components for the Toll pathway were also identified, which included GNBPs, PGRPs, even toll-like receptors. In addition, the prophenoloxidase cascade was also expected to be intact in *N. lugens* based on the prophenoloxidase determined (Table 3). All PGRPs can be divided into two classes: the long form and the short form (Dziarski, 2004). Long PRGPs mediate Imd pathway activation mainly against Gram-negative bacteria in *Drosophila*. Short PGRPs can recognize Gram-positive bacteria and then activate the Toll pathway and the prophenoloxidase cascade (Dziarski, 2004; Gorman *et al*., 2007). Through the transcriptome analysis four kinds of PGRPs were determined, which may include both long and short forms. The PGRP like gene (*NlPGRP*) cloned in *N. lugens* encodes a protein containing an amidase activity domain, like most insect PGRPs with a recognized function (Dziarski, 2004), with a putative signal peptide (Fig. 7). This NlPGRP should be a short form of PGRPs, which may function extracellularly and was similar to the original PGRP discovered in *Bombyx mori* (Dziarski, 2004). RT-PCR analysis indicated that *NlPGRP* mainly expressed in the midgut (Fig. 4A). Several members responsible for the RNAi pathway were noted, as well, which suggested that RNAi, the conserved mechanism, was also employed by *N. lugens* to combat viral invasion (Table 3). Thus, the transcriptome analysis suggested that *N. lugens* used most of the conserved mechanisms identified in well characterized homometabolous insects, e.g. *D. melanogaster*, for its immune response. Among the hemimetabolous group of insects, the genome of the pea aphid, *A. pisum*, was the first to be sequenced and has been used as the model insect for research in other

species from hemiptera. Our results indicated that the difference between *A. pisum* and *N. lugens* at molecular level may be greater than expected.

In addition to the midgut, the salivary gland of hemiptera is the other key tissue interacting with the environment. To fully understand the mechanism of the interaction between rice plants and *N. lugens*, as well as the high adaptation of this insect, gene expression data from the midgut and the salivary glands are both important. Recent research has also shown heterogeneous features of insect midguts. For instance, the midgut of *D. melanogaster* is divided into at least 13 distinct regions based on the difference in gene expression patterns (Chung *et al*., 2009), and the midgut of *A. pisum* consists of four regions labelled V1-V4 (Cristofoletti *et al*., 2003). Thus, further research on the different parts of the midgut in *N. lugens* may also be necessary.

In conclusion, we have shown that pyrosequencing was an effective approach to understanding gene expression patterns in the midgut of *N. lugens*. Combining pyrosequencing with other molecular techniques, we identified the main features of gene expression and key components responsible for nutrition, xenobiotic metabolism and immune responses in the midgut of *N. lugens*. The data obtained through the present study should facilitate an understanding of the mechanism on the host adaptation of *N. lugens*. The sequence information obtained could be valuable for further evolutionary and genomic study. Key components identified may also be targets for developing suitable methods to control this agricultural pest.

Experimental procedures

Insect rearing, midgut RNA preparation

The *N. lugens* population was reared on TN1 rice plants, which is susceptible to *N. lugens*, in the greenhouse at the Genetics Institute, Wuhan University. Three to four-instar nymphs of *N. lugens* were collected and their midguts were dissected using fine needles under an stereomicroscope (OLYMPUS SZX7). Before dissection, the insects were chilled on ice and placed in a Petri dish that had been brushed with chilled insect physiological buffer (0.65% NaCl water solution). The dissected midguts were immediately placed in TRIzol reagent (Invitrogen, Carlsbad, CA, USA). Three hundred midguts were collected for total RNA extraction.

cDNA library construction, pyrosequencing and sequence preprocessing

A SMART cDNA library construction kit (Clontech, Mountain View, CA, USA) was used to construct a midgut cDNA library following the manufacturer's protocol. For pyrosequencing, cDNA of *N. lugens* midgut were also synthesized using this kit, except that the modified CDS III/3′ cDNA synthesis primer, 5′-TAGAGACC GAGGCGGCCGACATGTTTTGTTTTTTTTCTTTTTTTTTVN-3′ was used. Pyrosequencing was performed using the 454 sequencing system at the Arizona Genomics Institute, University of Arizona, AZ, USA. Reads with low quality values were removed and sequences with poly (A) and vector-contaminated zones were trimmed. CAP3 was used to assemble high quality reads using default parameters (Huang & Madan, 1999). We assigned a contig to a cluster when it had at least two reads, while those with only one read were considered to be singletons. The data analysis was mainly performed on a work-station (HP Z800) operated by a Linux (ubuntu 9.10) system. The raw data were submitted to Short Read Archive at NCBI with accession number SRP007536.

To obtain more transcript information on the midgut of *N. lugens*, we also sequenced randomly selected cDNA clones from the library using the Sanger method.

BLAST homology searches and sequence annotations

NCBI_nr protein database and est_others database (March, 2010) were first downloaded from the National Center for Biotechnology Information (NCBI). Then all unique sequences were searched against NR database using BLASTX with *e*-value less than 10^{-3}, v = 10 and b = 10 and searched against est_others database using BLASTN with the *e*-value set to 1e-10. To analyse the similarities between the transcripts from the midgut of *N. lugens* and those from other model insects in detail, proteins of *A. pisum* and *T. castenum* were downloaded from NCBI and those from *D. melanogaster* were downloaded from the Flybase (March, 2010). Sequences that could be matched to proteins from model insects with *e*-values lower than 10^{-5}, 10^{-10} and 10^{-20}, were recorded, respectively. Before our research, Noda *et al*. have released more than 30 000 ESTs, generated by the Sanger method (Noda *et al*., 2008). To analyse the similarities, we also downloaded those sequences from NCBI, and pairwise comparisons were performed using BLASTN with the *e*-value set to $1e^{-5}$.

BLAST2GO (Version 2.4.2) software was employed to deal with the BLASTX results in XML format (high-scoring segment pairs ≥33) and then to perform the functional annotation by GO vocabularies, EC codes, KEGG metabolism pathways (Gotz *et al*., 2008). WEGO from BGI was used for GO vocabulary distribution plotting (Ye *et al*., 2006). To further attempt to link the unique sequences to the functional domain database, each sequence with an open reading frame (ORF) no less than 150 was extracted, then the largest ORF for each sequence was translated using getorf (EMBOSS package). InterProScan analysis was performed on those polypeptides through BLAST2GO. The most likely InterPro annotations were assigned to relevant sequences. SMART or ScanProsite analysis was performed if necessary (SMART: http://smart.embl-heidelberg.de/; ScanProsite: http://prosite.expasy.org/scanprosite/).

cDNA cloning of candidate genes

Rapid amplification of cDNA ends was used to clone the full-length cDNA of the candidate sucrase and PGRP genes. 5′-Full RACE kit and 3′-Full RACE kit were used to perform the 5′ and 3′RACE according to the instructions (TaKaRa, Kyoto, Japan), respectively. The specific primers used in this procedure are listed in Table S4.

Reverse transcription-PCR and quantitative PCR analysis

To validate expression of the transcripts in the midgut of *N. lugens*, RT-PCR analysis was performed. Three instar nymphs

were used to dissect the midgut, head, leg, fat body and whole abdomen. For each sample, about 100 nymphs were used. Total RNA from the tissues was extracted using TRIzol reagent (Invitrogen). RNA was treated for DNase I (MBI Fermentas, Flamborough, ON, Canada) to remove DNA contamination. Four point five micrograms of total RNA was reverse transcribed with Oligo(dT)18 primers using RevertAid First strand cDNA sysnthesis kit (MBI Fermentas, Flamborough, ON, Canada). Each cDNA sample was diluted 20-fold to be used for amplification. The primers used are listed in Table S4. The running parameters were as follows: 94°C for 5 min followed by 40 cycles of 94°C for 10 s, 55°C or 53°C for 10 s, 72°C for 20 s; and 72°C for 5 min. PCR products were analysed on a 1.5% agarose gel electrophoresis followed by ethidium bromide staining. Each target was amplified at least twice.

The qPCR reaction was performed on RotorGene 6000 (Corbett Research, Sydney, Australia). Each amplification reaction was performed using a 10-ul reaction with the following steps: 95°C for 30 s, followed by 45 cycles of 95°C for 5 s, 55°C for 10 s and 72°C for 15 s. Beta-Actin was used as the internal reference gene (Liu *et al*., 2008). For data analysis, triplicate measures for each target were averaged.

Simple sequence repeat locus and molecular analysis of cloned sequences

Perl script SSRIT, was downloaded from GRAMENE (http://www.gramene.org/db/markers/ssrtool) and used to identify SSR loci. Signal peptide and transmembrane helix prediction was performed using the CBS prediction servers (http://www.cbs.dtu.dk/services/). Hydropathy analysis was performed using ProScale with Kyte & Doolittle scale (http://expasy.org/tools/protscale.html). The MEGA4.0 software was used to construct the consensus phylogenetic tree using the Neighbour-joining method (Tamura *et al*., 2007). To evaluate the branch strength of the phylogenetic tree, bootstrap analysis of 10 000 replicates was performed.

Acknowledgements

The authors thank Dr Yeisoo Yu, at the Arizona Genomics Institute, University of Arizona, for valuable comments on this manuscript. The work was supported by the Fundamental Research Funds for the Central Universities (No. 1104002) and the Special Fund for Commonwealth Industry (200803003) from the Ministry of Agriculture, P. R. China.

References

Atkins, C., Smith, P. and Rodriguez-Medina, C. (2011) Macromolecules in phloem exudates – a review. *Protoplasma* **248**(1): 165–172.

Bing, L., Hongxia, D., Maoxin, Z., Di, X. and Jingshu, W. (2007) Potential resistance of tricin in rice against brown planthopper *Nilaparvata lugens* (Stål). *Acta Ecol Sin* **27**: 1300–1306.

Boutros, M., Agaisse, H. and Perrimon, N. (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Dev Cell* **3**: 711–722.

Chistiakov, D.A., Hellemans, B. and Volckaert, F.A.M. (2006) Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. *Aquaculture* **255**: 1–29.

Chiu, S. (1979) Biological control of the brown planthopper, *Nilaparvata lugens*. In *Brown Planthopper: Threat to Rice Production in Asia* (International Rice Research Institute, ed.), pp. 335–355. International Rice Research Institute, Los Baños, Philippines.

Chung, H., Sztal, T., Pasricha, S., Sridhar, M., Batterham, P. and Daborn, P.J. (2009) Characterization of *Drosophila* melanogaster cytochrome P450 genes. *Proc Natl Acad Sci USA* **106**: 5731–5736.

Cohen, E. (2006) Pesticide-mediated homeostatic modulation in arthropods. *Pestic Biochem Physiol* **85**: 21–27.

Cristofoletti, P.T., Ribeiro, A.F., Deraison, C., Rahbe, Y. and Terra, W.R. (2003) Midgut adaptation and digestive enzyme distribution in a phloem feeding insect, the pea aphid *Acyrthosiphon pisum*. *J Insect Physiol* **49**: 11–24.

Cristofoletti, P.T., Ribeiro, A.F. and Terra, W.R. (2005) The cathepsin L-like proteinases from the midgut of *Tenebrio molitor* larvae: sequence, properties, immunocytochemical localization and function. *Insect Biochem Mol Biol* **35**: 883–901.

Després, L., David, J.-P. and Gallet, C. (2007) The evolutionary ecology of insect resistance to plant chemicals. *Trends Ecol Evol* **22**: 298–307.

Droege, M. and Hill, B. (2008) The genome sequencer FLX system – longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol* **136**: 3–10.

Dziarski, R. (2004) Peptidoglycan recognition proteins (PGRPs). *Mol Immunol* **40**: 877–886.

Enayati, A.A., Ranson, H. and Hemingway, J. (2005) Insect glutathione transferases and insecticide resistance. *Insect Mol Biol* **14**: 3–8.

Feyereisen, R. (2006) Evolution of insect P450. *Biochem Soc Trans* **34**: 1252–1255.

Foissac, X., Edwards, M.G., Du, J.P., Gatehouse, A.M. and Gatehouse, J.A. (2002) Putative protein digestion in a sap-sucking homopteran plant pest (rice brown plant hopper; *Nilaparvata lugens*: Delphacidae) – identification of trypsin-like and cathepsin B-like proteases. *Insect Biochem Mol Biol* **32**: 967–978.

Freitak, D., Wheat, C., Heckel, D. and Vogel, H. (2007) Immune system responses and fitness costs associated with consumption of bacteria in larvae of *Trichoplusia ni*. *BMC Biol* **5**: 56.

Friedman, R. and Hughes, A.L. (2006) Pattern of gene duplication in the *Cotesia congregata* Bracovirus. *Infect Genet Evol* **6**: 315–322.

Gerardo, N.M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S.M., de Vos, M. *et al*. (2010) Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. *Genome Biol* **11**: R21.

Gorman, M.J., Wang, Y., Jiang, H. and Kanost, M.R. (2007) *Manduca sexta* hemolymph proteinase 21 activates prophenoloxidase-activating proteinase 3 in an insect innate immune response proteinase cascade. *J Biol Chem* **282**: 11742–11749.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J. *et al*. (2008) High-throughput functional

annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**: 3420–3435.

Hakim, R.S., Baldwin, K. and Smagghe, G. (2010) Regulation of midgut growth, development, and metamorphosis. *Annu Rev Entomol* **55**: 593–608.

Hao, P., Liu, C., Wang, Y., Chen, R., Tang, M., Du, B. *et al*. (2008) Herbivore-induced callose deposition on the sieve plates of rice: an important mechanism for host resistance. *Plant Physiol* **146**: 1810–1820.

Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res* **9**: 868–877.

Isoe, J., Zamora, J. and Miesfeld, R.L. (2009) Molecular analysis of the *Aedes aegypti* carboxypeptidase gene family. *Insect Biochem Mol Biol* **39**: 68–73.

Jena, K. and Kim, S.-M. (2010) Current status of brown planthopper (BPH) resistance and genetics. *Rice* **3**: 161–171.

Kehr, J. (2006) Phloem sap proteins: their identities and potential roles in the interaction between plants and phloem-feeding insects. *J Exp Bot* **57**: 767–774.

Khajuria, C., Zhu, Y.C., Chen, M.S., Buschman, L.L., Higgins, R.A., Yao, J. *et al*. (2009) Expressed sequence tags from larval gut of the European corn borer (*Ostrinia nubilalis*): exploring candidate genes potentially involved in *Bacillus thuringiensis* toxicity and resistance. *BMC Genomics* **10**: 286.

Kikuta, S., Kikawada, T., Hagiwara-Komoda, Y., Nakashima, N. and Noda, H. (2010) Sugar transporter genes of the brown planthopper, *Nilaparvata lugens*: a facilitated glucose/fructose transporter. *Insect Biochem Mol Biol* **40**: 805–813.

Koo, Y.D., Ahn, J.E., Salzman, R.A., Moon, J., Chi, Y.H., Yun, D.J. *et al*. (2008) Functional expression of an insect cathepsin B-like counter-defence protein. *Insect Mol Biol* **17**: 235–245.

Liu, S., Yang, B., Gu, J., Yao, X., Zhang, Y., Song, F. *et al*. (2008) Molecular cloning and characterization of a juvenile hormone esterase gene from brown planthopper, *Nilaparvata lugens*. *J Insect Physiol* **54**: 1495–1502.

Liu, Z., Han, Z., Wang, Y., Zhang, L., Zhang, H. and Liu, C. (2003) Selection for imidacloprid resistance in *Nilaparvata lugens*: cross-resistance patterns and possible mechanisms. *Pest Manag Sci* **59**: 1355–1359.

Lomate, P.R. and Hivrale, V.K. (2010) Partial purification and characterization of *Helicoverpa armigera* (Lepidoptera: Noctuidae) active aminopeptidase secreted in midgut. *Comp Biochem Physiol B Biochem Mol Biol* **155**: 164–170.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Muhlia-Almazan, A., Sanchez-Paz, A. and Garcia-Carreno, F.L. (2008) Invertebrate trypsins: a review. *J Comp Physiol [B]* **178**: 655–672.

Nakashima, N., Kawahara, N., Omura, T. and Noda, H. (2006) Characterization of a novel satellite virus and a strain of Himetobi P virus (*Dicistroviridae*) from the brown planthopper, *Nilaparvata lugens*. *J Invertebr Pathol* **91**: 53–56.

Noda, H., Kawai, S., Koizumi, Y., Matsui, K., Zhang, Q., Furukawa, S. *et al*. (2008) Annotated ESTs from various tissues of the brown planthopper *Nilaparvata lugens*: a genomic resource for studying agricultural pests. *BMC Genomics* **9**: 117.

Padilha, M.H.P., Pimentel, A.C., Ribeiro, A.F. and Terra, W.R. (2009) Sequence and function of lysosomal and digestive cathepsin d-like proteinases of *Musca domestica* midgut. *Insect Biochem Mol Biol* **39**: 782–791.

Pankoke, H., Bowers, M.D. and Dobler, S. (2010) Influence of iridoid glycoside containing host plants on midgut [beta]-glucosidase activity in a polyphagous caterpillar, *Spilosoma virginica* Fabricius (Arctiidae). *J Insect Physiol* **56**: 1907–1912.

Pauchet, Y., Wilkinson, P., van Munster, M., Augustin, S., Pauron, D. and ffrench-Constant, R.H. (2009) Pyrosequencing of the midgut transcriptome of the poplar leaf beetle *Chrysomela tremulae* reveals new gene families in Coleoptera. *Insect Biochem Mol Biol* **39**: 403–413.

Pauchet, Y., Wilkinson, P., Vogel, H., Nelson, D.R., Reynolds, S.E., Heckel, D.G. *et al*. (2010) Pyrosequencing the *Manduca sexta* larval midgut transcriptome: messages for digestion, detoxification and defence. *Insect Mol Biol* **19**: 61–75.

Price, D.R.G., Karley, A.J., Ashford, D.A., Isaacs, H.V., Pownall, M.E., Wilkinson, H.S. *et al*. (2007a) Molecular characterisation of a candidate gut sucrase in the pea aphid, *Acyrthosiphon pisum*. *Insect Biochem Mol Biol* **37**: 307–317.

Price, D.R.G., Wilkinson, H.S. and Gatehouse, J.A. (2007b) Functional expression and characterisation of a gut facilitative glucose transporter, NlHT1, from the phloem-feeding insect *Nilaparvata lugens* (rice brown planthopper). *Insect Biochem Mol Biol* **37**: 1138–1148.

Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M.V. *et al*. (2002) Evolution of supergene families associated with insecticide resistance. *Science* **298**: 179–181.

Rispe, C., Kutsukake, M., Doublet, V., Hudaverdian, S., Legeai, F., Simon, J.-C. *et al*. (2008) Large gene family expansion and variable selective pressures for cathepsin b in aphids. *Mol Biol Evol* **25**: 5–17.

Siva-Jothy, M.T., Moret, Y. and Rolff, J. (2005) Insect immunity: an evolutionary ecology perspective. *Advances in Insect Physiology* Vol. 32: 1–48.

Strode, C., Wondji, C.S., David, J.-P., Hawkes, N.J., Lumjuan, N., Nelson, D.R. *et al*. (2008) Genomic analysis of detoxification genes in the mosquito *Aedes aegypti*. *Insect Biochem Mol Biol* **38**: 113–123.

Suh, S.-O., Noda, H. and Blackwell, M. (2001) Insect symbiosis: derivation of yeast-like endosymbionts within an entomopathogenic filamentous lineage. *Mol Biol Evol* **18**: 995–1000.

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.

Tanaka, T., Suzui, N., Hayashi, H., Yamaya, T. and Yoneyama, T. (2009) Cytosolic glutamine synthetase is present in the phloem sap of rice (*Oryza sativa* L.). *Soil Sci Plant Nutr* **55**: 102–106.

Tang, M., Lv, L., Jing, S., Zhu, L. and He, G. (2010) Bacterial symbionts of the brown planthopper, *Nilaparvata lugens* (Homoptera: Delphacidae). *Appl Environ Microbiol* **76**: 1740–1745.

The International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* **8**: e1000313.

Wang, P. and Granados, R.R. (2001) Molecular structure of the peritrophic membrane (PM): identification of potential PM target sites for insect control. *Arch Insect Biochem Physiol* **47**: 110–118.

Wei, Z., Hu, W., Lin, Q., Cheng, X., Tong, M., Zhu, L. *et al*. (2009) Understanding rice plant resistance to the brown planthopper (*Nilaparvata lugens*): a proteomic approach. *Proteomics* **9**: 2798–2808.

Xue, J., Bao, Y.-Y., Li, B.-L., Cheng, Y.-B., Peng, Z.-Y., Liu, H. *et al*. (2010) Transcriptome analysis of the brown planthopper *Nilaparvata lugens*. *PLoS ONE* **5**: e14233.

Yang, Z., Zhang, F., He, Q. and He, G. (2005) Molecular dynamics of detoxification and toxin-tolerance genes in brown planthopper (*Nilaparvata lugens* Stal., Homoptera: Delphacidae) feeding on resistant rice plants. *Arch Insect Biochem Physiol* **59**: 59–66.

Yang, Z., Zhang, Y., Liu, X. and Wang, X. (2010) Two novel cytochrome P450 genes CYP6CS1 and CYP6CW1 from *Nilaparvata lugens* (Hemiptera: Delphacidae): cDNA cloning and induction by host resistant rice. *Bull Entomol Res* **FirstView**: 1–9.

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z. *et al*. (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* **34**: W293–W297.

Yu, Q.-Y., Lu, C., Li, W.-L., Xiang, Z.-H. and Zhang, Z. (2009) Annotation and expression of carboxylesterases in the silkworm, *Bombyx mori*. *BMC Genomics* **10**: 553.

Zaidman-Rémy, A., Hervé, M., Poidevin, M., Pili-Floury, S., Kim, M.-S., Blanot, D. *et al*. (2006) The *drosophila* amidase PGRP-LB modulates the immune response to bacterial infection. *Immunity* **24**: 463–473.

Zhao, P., Wang, G.-H., Dong, Z.-M., Duan, J., Xu, P.-Z., Cheng, T.-C. *et al*. (2010) Genome-wide identification and expression analysis of serine proteases and homologs in the silkworm *Bombyx mori*. *BMC Genomics* **11**: 405.

Supporting Information

Additional Supporting Information may be found in the online version of this article under the DOI reference: 10.1111/j.1365-2583.2011.01104.x

Table S1. Summary of the 25 most common InterPro entries found in the *Nilaparvata lugens* nymph midgut database.

Table S2. The transcripts for selected proteases and sugar hydrolyases identified in the midgut of *Nilaparvata lugens*.

Table S3. Genes linked to xenobiotic metabolism.

Table S4. Primers used in this research.

Please note: Neither the Editors nor Wiley-Blackwell are responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.