IMMUNITY OR DIGESTION: Glucanase activity in a glucan-binding protein family from Lepidoptera

Yannick Pauchet, Dalial Freitak, Hanna M. Heidel-Fischer, David G. Heckel, and Heiko Vogel

From Department of Entomology, Max Planck Institute for Chemical Ecology, Beutenberg Campus, Hans-Knoell-Str. 8, 07745 Jena, Germany.

Running Head: Lepidopteran glucan-binding proteins

Address correspondence to: Yannick Pauchet, Present address: School of Biosciences, University of Exeter in Cornwall, Penryn, Cornwall, TR10 9EZ, United Kingdom. Phone: ++441326371852, Fax: ++441326253638, E-mail: Y.Pauchet@exeter.ac.uk

The cell surfaces of microorganisms display distinct molecular patterns formed from lipopolysaccharides, peptidoglycans or β-1,3-glucans. Binding of these surfaces by pattern recognition proteins such as β-1,3-glucan recognition proteins (βGRPs) activates the immune response in arthropods. We identified a 40 kDa β-1,3-glucan-binding protein with sequence similarity to previously characterized lepidopteran βGRPs from hemolymph, but unlike these it is secreted into the larval gut lumen and is an active β-1,3-glucanase. This glucanase was not detected in hemolymph. Its mRNA is constitutively and predominantly expressed in the midgut and is induced there when larvae feed on diet containing bacteria. Homologs of this predominantly midgut-expressed gene from many Lepidoptera possess key residues shown to be part of the active site of other glucanases, and form a cluster that is distinct from previously-described βGRPs. In addition, this group includes proteins from insects such as the Anopheles gambiae GNPB subgroup B for which a catalytic role has not been previously suspected. The current domain classification does not distinguish between the catalytic and noncatalytic clades, and should be revised. The noncatalytic βGRPs may be evolutionarily derived from this newly described enzyme family which continues to function catalytically in digestion and/or pathogen defence.

Recognition of invading organisms as non-self is a crucial pre-requisite for an immune response. In arthropods, immune response is triggered in the hemolymph by so-called Pattern Recognition Proteins or Receptors (PRPs or PRRs) which bind to classes of bacteria- and fungi-specific polysaccharides such as peptidoglycans, β-1,3-glucans, lipopolysaccharides (LPS) and lipotechoic acid (LTA) (1,2). Little detail is known for Lepidoptera but in Drosophila these binding interactions lead to activation of the innate immune response, largely regulated by two main pathways, the Toll and Imd pathways. The innate immune system contains three main effector mechanisms: the cellular response, the humoral response and melanization. The activation of the signaling pathways leads to a cascade of events that result in the induction of defensive genes in hemocytes and the fat body, including antimicrobial peptide genes (3), and also to the activation of the prophenoloxidase (proPO) cascade. Phenoloxidase (PO) is an essential enzyme for the cellular immune responses (4) but is also involved in other developmental and defensive processes such as wound healing and sclerotization (1).

Gram negative bacteria binding proteins (GNBPs) and β-1,3-glucan recognition proteins (βGRPs) have been extensively studied as pattern recognition proteins in Lepidoptera (5-10). These proteins are produced in the fat body and secreted into the caterpillar’s hemolymph. Some are constitutively present (5,8) whereas others are induced upon microbial infection (6). Recognition of β-1,3-glucans by these proteins appears to be due to two distinct β-1,3-glucan-binding domains:
a strongly binding N-terminal glucan recognition domain and a weakly-binding C-terminal glucanase-like domain. Neither domain possesses β-1,3-glucanase activity (11).

The immune response has been extensively studied in Lepidoptera by introducing microbes into the hemocoel of laboratory-reared insects by mechanical wounding. However, since herbivorous lepidopteran larvae consume large amount of plant material during their development, infection via food is very likely. The potential role of the midgut in triggering the immune response of herbivorous larvae has been poorly studied, but there is increasing evidence that it participates in this process. Freitak et al. (12) showed that in larvae of Trichoplusia ni that were fed artificial diet containing bacteria, several immune-related genes were up-regulated and immune-related proteins increased in concentration in the hemolymph.

If the midgut were to play a role in sensing potentially invasive bacteria, one would expect to find pattern recognition proteins expressed in that tissue. Bacteria present in the Drosophila gut are thought to release peptidoglycan fragments that bind to specific PGRP (peptidoglycan recognition protein) receptors and induce the imd pathway (13-15). Simpson et al. (16) described a sequence (EpGRP1) with similarity to β-1,3-glucan binding proteins from Anopheles gambiae and Bombyx mori in ESTs derived from the larval midgut of Epiphyas postvittana. In a proteomic study of the lumen contents of larval midgut of Helicoverpa armigera, we (17) discovered a similar protein that we named GH16BetaGRP-1 (Glycosyl Hydrolase family 16, β-glucan recognition protein 1) secreted into the lumen. The full-length cDNA showed similarities to β-1,3-glucan recognition proteins from the hemolymph of B. mori and Manduca sexta. Since β-1,3-glucanase activity had not been previously reported from lepidopteran midguts, we suggested the possibility of an immune-related function.

Here we describe the purification and characterization of GH16betaGRP-1 from the midgut lumen of H. armigera. The protein does indeed possess β-1,3-glucanase activity and can account for the majority of such activity in the lumen. Thus we have renamed this protein as H. armigera β-1,3-glucanase-1. The mRNA is predominantly expressed in midgut, and the protein product is secreted into the lumen and persists stably there. Expression is increased when the insect feeds on diet containing bacteria. We found similar sequences from a number of lepidopteran species derived from cDNAs generated from midgut tissue samples. These form a novel protein family that is characterized by a highly conserved signature sequence including two glutamate residues previously shown to be necessary for catalytic activity in other β-1,3-glucanases (18). This family is related to but distinct from a family of previously-described βGRP/GNBP proteins primarily found in lepidopteran hemolymph which do not possess this conserved site. Certain proteins from insects for which a catalytic role has not been suspected, including the Anopheles gambiae GNBP subgroup B, belong to the catalytic clade and we suggest that the domain nomenclature be revised accordingly. These features suggest that lepidopteran larvae secrete an active β-1,3-glucanase into the midgut lumen which may function in digestion of β-1,3-glucans released by commensal or invading bacteria.

EXPERIMENTAL PROCEDURES

Insects. The TWB strain of H. armigera Hübner (Lepidoptera: Noctuidae) was collected from the vicinity of Toowoomba, Queensland, Australia. Neonates were reared on a chemically defined diet containing only casein as the protein source and no plant-derived material, as described by Vanderzant (1968) at 26 °C with a 16:8 (L:D) photoperiod. Sample preparation. Gut lumen sample was prepared essentially as described by Pauchet et al. (17). Briefly, midguts were dissected from actively feeding second day fifth-instar larvae in ice-cold phosphate-buffered saline (PBS). Peritrophic matrix containing the food bolus was pulled out of the midgut with forceps and gently homogenized by 10 strokes in a Potter-Elvehjem homogenizer in PBS pH 7.5 containing a cocktail of protease inhibitors (Complete EDTA-free, Roche Applied
Science) in order to release soluble proteins. After centrifugation (30,000xg, 30 min, 4 °C), the supernatant containing the gut lumen soluble proteins was kept and protein concentration was determined using the Protein Dye reagent (BioRad) and bovine serum albumin (BSA) as standard. Hemolymph was collected via lateral punctuation of abdominal part of larvae with micropipette. Collected hemolymph was immediately diluted with ice-cold N-Phenyldiurea-saturated PBS containing protease inhibitors and hemocytes were removed by centrifugation (3,000xg, 15 min, 4 °C).

Curdlan pull-down assay. Gut lumen and hemolymph samples (1 ml each) were incubated 30 min at 22 °C with 2.5 mg curdlan (an insoluble β-1,3-glucan isolated from Alcaligenes faecalis, Sigma). After centrifugation of the sample/curdlan mixture (16,000xg, 5 min, 4 °C), the supernatant corresponding to the unbound fraction was saved. The curdlan pellet was washed 5 times with PBS followed by 2 washes with PBS containing 1 M NaCl and finally 3 more times with PBS. Bound proteins were eluted from curdlan by boiling 10 min in SDS-PAGE sample buffer. After centrifugation (16,000xg, 5 min, 4 °C), the supernatant was analyzed by SDS-PAGE.

Protein analysis by mass spectrometry. This was done essentially as described by Pauchet et al. (17). Protein bands were excised manually from the SDS-PAGE gels and were destained, trypsinized, and extracted using an Etten TA Digester running the Digester Version 1.10 software (GE Healthcare Bio-Sciences AB). Trypsin digestion was carried out overnight with 50 ng of porcine trypsin (Promega) at 37 °C. First step analysis of the tryptic peptides was done using a MALDImicro MX mass spectrometer (Waters) used in reflectron mode and was calibrated using a tryptic digest of bovine serum albumin (MPrep, Waters). The MALDI-TOF spectra searches were performed with the Protein Lynx Global Server software, version 2.2 (PLGS 2.2, Waters) against the NCBI insecta database (downloaded on 15 March 2008 from http://www.ncbi.nlm.nih.gov/database, 372,057 entries). The search parameters were as follows: peptide tolerance of 80 ppm, one missed cleavage, carbamidomethyl modification of cysteines, and possible Met oxidation. An estimated calibration error of 0.05 D and a minimum of four peptide matches were the criteria for obtaining positive database hits.

The MALDI-TOF peptide signal intensities were used to estimate the volume of the remaining sample to be used for the subsequent nanoLC-MS/MS de novo sequence analysis. Liquid chromatography-tandem mass spectrometry was performed to acquire fragmentation data from selected peptides. Aliquots of tryptic peptides were injected on a CapLC XE 2D nanoLC system (Waters). After concentration and desalting, eluted peptides were transferred to the NanoElectroSpray source of a Q-TOF Ultima tandem mass spectrometer (Waters). MS/MS spectra were collected by MassLynx v4.0 software (Waters). ProteinLynx Global Server Browser v.2.2 software (PLGS 2.2, Waters) was used for baseline subtraction and smoothing, deisotoping, and de novo peptide sequence identification.

Using PLGS 2.2, CID spectra were interpreted de novo to yield peptide sequences. Sequences with a ladder score (percentage of expected y- and b-ions) exceeding 30 were then used in a homology-based search strategy using the MS BLAST program (19). MS BLAST was developed to utilize redundant, degenerate and partly inaccurate peptide sequences in similarity searches of protein databases that may be derived from organisms phylogenetically distant from the study species. The WU-BLAST2 BLASTP search engine (W. Gish 1996, http://blast.wustl.edu) is employed with parameter values that disallow gaps within a peptide, and that score only the most significant match in the case of several peptide candidates covering the same region in the target sequence. In addition, the PAM30MS matrix which accounts for the inability to distinguish I and L residues, and allows for unknown residues X, is used in the blastp similarity search (19). Scoring of the significance of peptide matches is not based on E- or p-values of the individual HSPs (high-scoring segment pairs) but instead on pre-computed threshold scores conditional on the number of query peptides and HSP hits. The color-
coded output produced by the MS BLAST script identifies in red those target sequences with scores exceeding 99% of queries utilizing randomized peptide sequences by chance. Computational studies have estimated a false positive rate of <3% (20). We installed an in-house MS BLAST server for searching NCBI_insecta and a locally generated EST database from H. armigera midgut and fat body cDNA libraries (5,685 protein sequences). Additional information on MS BLAST statistics and scoring can be found in Shevchenko et al. (19).

Beta-glucanase activity assays. Lichenan, Barley β-glucan and carboxymethyl (CM)-cellulose were from Sigma. CM-curdlan, endo-β-1,3-glucanase from Trichoderma sp. And cellulase from Trichoderma sp. were both from Megazyme.

Plate assays were performed on 1.5% agar Petri dishes containing 0.25 % of each substrate in 50 mM Citrate-Phosphate buffer pH 6.0. Wells within plates were made by puncturing the agar with a plastic pipette and removing the agar plug by suction. Two and a half micrograms gut luminal proteins in 2.5 µl PBS were tested together with 2.5 µl PBS as negative control, and 1 µl of a 100 times diluted solution of each endo-β-1,3-glucanase and cellulase as positive controls. After 16 hours incubation at 37 °C, activity was revealed by staining with Congo Red (0.1 % (w/v) in water) for 30 min and destaining with 1M NaCl also for 30 min. Plates were scanned using a GS800 densitometer (BioRad).

For zymograms, samples were resolved by native PAGE using 4-15 % polyacrylamide gels (Ready Gels, BioRad) and Tris-Glycine pH 8.3 as running buffer. Electrophoresis was performed at 150 V for 60 min at 4 °C. After the run, gels were washed 3 times 15 min in 50 mM citrate-phosphate buffer pH 6.0 before being overlayed on a CM-curdlan plate prepared as described above. Incubation and Congo Red staining/destaining were performed as described above.

Cell culture and transfection. The Ha_GH16betaGRP-1 cDNA was amplified from a H. armigera larval midgut cDNA library clone using primers GRP1-F (5'-CAAGCCCCAATGCGAACGTAGTC-3') and GRP1-R (5'-GCAAGCACCAGTGGTCGTTAAGCAGGCGT-3') and was inserted in pIB/V5-His TOPO (Invitrogen) by TA cloning according to the supplier’s instructions. Positive clones were selected and correct insertion was confirmed by sequencing.

Trichoplusia ni High Five cells (Invitrogen) were cultured at 27 °C in Express Five serum-free medium (Gibco) supplemented with 2 mM L-Glutamine and 10 μg/ml Gentamycin. Cells were transfected in 90 mm-diameter Petri dishes with 12 μg plasmid DNA using Insect Gene Juice (Novagen) as transfection reagent. Forty-eight hours post-transfection, culture medium was saved and the cells were detached from the plate and pelleted by centrifugation. After two washes with PBS, cell pellets were resuspended in cold PBS/Complete. Expression was analyzed by western blot using the anti-V5-HRP antibody (Invitrogen).

Preparation of cDNA libraries. TRIzol Reagent (Invitrogen) was used to isolate total RNA from whole larvae or dissected tissues according to the manufacturer’s protocol. After DNase treatment, total RNA were further purified by using the RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer’s protocol. Poly(A)+ mRNA were purified by binding to an oligo d(T) column (RNA Purist, Ambion). The generation of T. ni whole-larvae cDNA library has been described in Freitak et al. (12). Generation of Plutella xylostella, Pieris rapae, Colias eurytheme, Anthoracharis cardamines and Delias nigrina has been described in Fischer et al. (21). For H. armigera midgut and fat body and Spodoptera littoralis whole-larvae, normalized, full-length enriched cDNA libraries were generated using a combination of the SMART cDNA library construction kit (Clontech) and the Trimmer-Direct cDNA normalization kit (Evrogen) generally following the manufacturer’s protocol but with several important modifications. In brief, 2 μg of poly(A)+ mRNA was used for
Reverse transcription was performed with a mixture of several reverse transcription enzymes for 1h at 42 °C and 90 minutes at 50 °C. cDNA size fractionation was performed with SizeSep 400 spun columns (GE Healthcare) that resulted in a cutoff at ~200 bp. The full-length-enriched cDNAs were cut with SfiI and ligated to pDNR-Lib plasmid (Clontech). Ligations were transformed into *E. coli* ELECTROMAX DH5α-E electro-competent cells (Invitrogen).

**Generation of EST Databases.** Plasmid minipreparation from bacterial colonies grown in 96 deep-well plates was performed using the 96 robot plasmid isolation kit (Eppendorf) on a Tecan Evo Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5' termini of cDNA libraries was carried out on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems). Vector clipping, quality trimming and sequence assembly was done with the Lasergene software package (DNAStar Inc.). Blast searches were conducted on a local server using the National Center for Biotechnology Information (NCBI) blastall program. Protein domains were determined by searching the NCBI Conserved Domain Database (CDD) at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. Sequences were aligned using ClustalW software (22).

**BAC library screening and sequencing.** *H. armigera* BAC library nylon filters were washed, blocked and hybridized with horseradish peroxidase (HRP)-labelled DNA fragments containing part of the *GH16betaGRP-1* gene. Labelling, hybridization and probe detection were done according to specifications in the ECL DNA labeling and detection kit (GE Healthcare). Positive clones were isolated from glycerol stocks, grown in Terrific Broth and BAC DNA was isolated with the Nucleobond Xtra Midi Kit according to the manufacturers’ instructions (Macherey-Nagel). BAC Genomic DNA quantity was measured photospectrometrically on a Nanodrop ND1000 and all of the positive clones were digested with EcoRI and HindIII to identify clone diversity and redundant inserts, blotted and re-hybridized to verify positive inserts. The BAC DNA was sheared into two different size ranges (1-1.5 kb and 4-5 kb) with a Hydroshear device (Molecular Devices), blunted with the Quick Blunting Kit (NEB), isolated from an agarose gel, column purified and ligated into the pUC19-Smal vector (Fermentas). Ligations were transformed into *E. coli* ELECTROMAX DH5α-E electro-competent cells (Invitrogen). Plasmid preparation, sequencing and assembly were performed as mentioned above.

**Feeding assays.** Fifteen early fifth-instar *H. armigera* larvae were fed for 24h on artificial pinto-bean based diet soaked with overnight liquid culture of either *S. cerevisiae* or *M. luteus* or *E. coli*. Alternatively, larvae were fed on artificial diet soaked in solutions containing either β-glucans (0.25 % w/v) or bacterial cell wall polysaccharides (LTA or LPS, 0.05 % w/v). Larvae fed on sterile artificial diet were used as a control group. Three times 5 larvae per diet were used for detecting diet-induced changes in *GH16betaGRP-1* expression in midgut tissue. Experiments were repeated 3 times with different patches of eggs.

**Quantitative real-time PCR.** Total RNA was prepared from dissected insect midguts, salivary glands, fat bodies and rest of bodies (except head capsules) according to the method described above. Five hundred nanograms of DNA-free total RNA was converted into single-stranded cDNA using a mix of random and oligo-dT20 primers according to the ABgene protocol (ABgene). As an endogenous control gene, large ribosomal protein 1 (*Lrp1*) was used (forward primer: 5’-CACATCAGCAAAACACATCAACC and Reverse primer: 5’-AGAAGTGAAGCGGCTGAAAG). Gene-specific primers were designed on the basis of sequence obtained for *GH16betaGRP-1* gene (forward primer: 5’-CTGGGATCAGGAAACACTGCT and reverse primer: 5’-TACACGATCCGCCAGTGTAGGAA). Q-RT-PCR was done in optical 96-well plates on a MX3000P Real-Time PCR Detection System (Stratagene) using the Absolute QPCR SYBR green Mix (ABgene) to monitor double-stranded DNA synthesis in combination with ROX as a passive reference dye included in the PCR master mix.
Phylogenetic analysis. Genes were aligned by their amino acid sequences using MAFFT multiple sequence alignment program (23). The amino acid alignment was then used for phylogenetic analysis.

The phylogenetic reconstruction was done by Bayesian inference using Mr. Bayes 3.1 (24). The prior was set for the amino acid models to mix, thereby allowing model jumping between fixed-rate amino acid models. Markov Chain Monte Carlo runs were carried out for 1,000,000 generations after which log likelihood values showed that equilibrium had been reached after the first 400 generations in all cases, and those data were discarded from each run and considered as ‘burnin’. Two runs were conducted for the dataset showing agreement in topology and likelihood scores.

Phylogenetic analysis was furthermore performed using the Neighbour-Joining (NJ) method (TREECON) based on the MAFFT alignments. Distance calculations were performed after Tajima & Nei and bootstrap analysis, running 1000 bootstrap samples. The Neighbour-joining and the Bayesian tree topologies including their general subfamily relationships and node supports were in agreement.

RESULTS

A β-1,3-Glucan Binding Protein from H. armigera larval midgut. Midgut lumen contents of unchallenged H. armigera larvae were screened for glucan binding proteins using curdlan, an insoluble β-1,3-glucan isolated from the Gram negative bacterium A. faecalis, as an affinity matrix in pull-down assays. The bound fraction was freed from the matrix by boiling and after SDS-PAGE, a single main protein band (apparent MW: 40 kDa) was detected along with trace amounts of other proteins (Fig.1A). The peptide mass fingerprint of this main protein band by MALDI-TOF/MS analysis exactly matched the cDNA sequence of H. armigera GH16betaGRP-1 (ABU98621.1, Supplementary Information) showing that this previously identified protein, which had been named on the basis of sequence similarity to other Lepidopteran proteins, was the main β-1,3-glucan binding protein in the larval midgut lumen.

Expression patterns of this gene in different tissues were further investigated by quantitative real time PCR (RT-qPCR) (Fig. 1B). The expression level was very high in larval midgut; much lower but detectable in salivary glands, hemocytes, and the fat body; and undetectable in the rest of the larval body tissue.

Glucan-binding protein has β-1,3-glucanase activity. We screened for glucanase activity in H. armigera larval midgut lumen contents by plate assays on carbohydrate polymers followed by staining of any remaining polymer with Congo Red (Fig. 2A). The only tested substrate that was degraded by the lumen sample was CM-curdlan which contains exclusively β-1,3-glucan linkages. Neither cellulose (exclusively β-1,4 linkages) nor lichenan and Barley β-glucan (with mixed β-1,3 and β-1,4 linkages) were degraded by the lumen sample. The latter two substrates could be partially degraded by fungal enzymes specific for β-1,3-glucans or β-1,4-glucans. However, the lumen sample was incapable of degrading the β-1,3-glucan linkages in the lichenan or barley β-glucan (Fig 2A).

To investigate the correlation between β-1,3-glucan binding and glucanase activity, an untreated gut lumen sample was compared with the flow-through from the curdlan pull-down assay by native PAGE followed by an overlay with CM-curdlan and staining with Congo Red (Fig. 2B). Although the resolution of the zymogram was poor, a broad band indicating activity could be detected only in the untreated sample whereas no activity was detected in the flow-through. The broad activity band is likely due to the major band seen in Fig. 1A after the curdlan pull-down, which is poorly resolved on the native gel; although a contribution of some of the fainter bands in Fig. 1A cannot be completely ruled out.

To further test whether the glucan-binding protein itself could be responsible for this β-1,3-glucanase activity, its coding sequence fused to a C-terminal V5/(His)$_6$ tag was transiently expressed in High Five insect cells (Fig. 2C). Using the anti-V5 antibody in a western blot, a protein with an apparent molecular weight of 44 kDa was detected
in the culture medium of transfected High Five cells, and a lesser amount was associated with the crude membrane fraction. The increase in apparent molecular weight of about 4 kDa between the gut lumen form of the protein and its recombinant form expressed in insect cells is due to the V5/(His)_6 tag. A zymogram using CM-curdlan as substrate revealed a single activity band in the culture medium of cells transfected with the protein construct but not with empty vector (Fig. 2D). These results show that the protein that we had named GH16betaGRP-1 is an active β-1,3-glucanase and suggest that the activity detected in the gut lumen is mainly due to this protein. We thus rename this protein as β-1,3-glucanase-1 from *Helicoverpa armigera*.

**Induction of glucanase-1 after ingestion of bacteria.** *H. armigera* larvae were fed for 24 h on artificial diet containing one of three types of non-pathogenic microbes: Gram negative (*Escherichia coli*) and Gram positive (*Micrococcus luteus*) bacteria and yeast (*Saccharomyces cerevisiae*). Expression of glucanase-1 was investigated by RT-qPCR (Fig. 3). A ~2.5-fold increase in mRNA levels was observed for larvae fed on Gram negative bacteria as well as a ~4.5-fold increase for larvae fed on Gram positive bacteria. No significant induction was observed for larvae fed on yeast.

Alternatively, larvae were fed on diet containing bacterial cell wall polysaccharides: LTA for Gram positive or LPS for Gram negative bacteria. A ~2-fold increase in mRNA levels was observed for larvae fed on LTA as well as a ~1.8-fold increase for larvae fed on LPS. In addition, no significant induction was observed for larvae fed on artificial diet containing one of three types of β-glucans (Curdlan, Lichenan and Barley β-glucan).

**Glucanase-1 is not detected in the hemolymph.** To further confirm the expression specificity of glucanase-1 in the midgut, affinity to curdlan was used to pull down proteins from *H. armigera* hemolymph sample of unchallenged larvae (Fig. 4). Seven protein bands were recovered and tryptic peptides were sequenced de novo and identified. None of the proteins matched the glucanase-1 sequence although two (bands b and c) having an apparent MW of around 56 kDa were identified as βGRPs (Table 1) by their matching ESTs from *H. armigera* fat body. Predicted protein sequences of βGRP-1 and -2a are very similar to previously characterized proteins purified from larval hemolymph of the lepidopteran species *M. sexta* and *B. mori* (8,9). In addition to βGRPs, prophenoloxidase (proPO) subunit 2 (band a), a proPO-activating factor-like protein (serine-proteinase like protein 1, band e), a C-type lectin (band f) and a serine-proteinase homolog (band g) were identified in the hemolymph curdlan pull-down fraction (Table 1). All hits obtained using MS-BLAST were of high confidence as given by the search engine except for the C-type lectin (Supplementary Information). This protein was identified with only a single 9 amino acid peptide (SVIPGNFDK) that was returned as a “borderline” hit by MS BLAST. We took this hit into consideration because the predicted MW of the full-length protein (36.8 kDa) was similar to the apparent MW observed on the gel (Fig. 4).

**Glucanase-1 is conserved among Lepidoptera and other insect orders.** The *H. armigera* glucanase-1 is 375 amino acids long and it contains a predicted signal peptide with a cleavage site between amino acid 17 and 18 suggesting secretion. Residues 40 through 375 correspond to a conserved domain identified as the glycosyl hydrolase family 16 (pfam00722 Glyco_hydro-16, E = 4e-10) and the subfamily beta-1,3-glucan recognition proteins (cd02179 GH16_beta_GRP, E = 3e-103) as indicated by a search against the Conserved Domain Database. Furthermore, a conserved GH16 active site was found between amino acids 188 and 199 including 2 glutamate residues in positions 188 and 193 that have been shown to be crucial for catalytic activity in the glucanase from *Bacillus licheniformis* (18) (Fig. 5A).
hydrolase 16 family, and has not yet been assigned to a conserved domain family, but it occurs in a variety of beta-1,3-glucan binding proteins isolated from lepidopteran hemolymph. Fabrick et al. (11) denoted this domain as CRD (carbohydrate recognition domain) and showed that truncated proteins from Plodia interpunctella expressing only CRD could bind to curdlan and could activate the prophenoloxidase cascade. Putative orthologs to H. armigera Glucanase-1 were found in several species including B. mori, Spodoptera frugiperda, E. postvittana, Ostrinia nubilalis and P. interpunctella (Supplementary information). EST coverage was sufficient to obtain 3 full-length sequences. The O. nubilalis sequence was truncated missing 30 nucleotides, and the full-length sequence was kindly provided by Dr. Siu Fai Lee (University of Melbourne, CESAR centre). All EST accession numbers used to obtain the data presented here are given in Supplementary Information. An orthologous protein annotated as beta-1,3-glucanase from the sugarcane borer Diatraea saccharalis was found in Genbank (ABR28479); this protein is truncated missing a few amino acids at the amino-terminus. Finally, a S. frugiperda ortholog was also found in Genbank (ABR28478). This protein differs by 4 amino acid substitutions from the protein we obtained via ESTs. The Genbank protein is two amino acids longer than our protein due to a repetition of amino acids 165-166: DW (GATTGG) in our sequence, DWDW (GATTGGGATTGG) in the ABR28478 sequence. Similar to the H. armigera sequence, the 4 orthologs are 375 amino acids long and harbor a conserved GH16 active site (Figure 5A).

Orthologous glucanase genes in H. armigera and B. mori. To obtain the intron/exon organization of the Glucanase-1 gene, a probe from the H. armigera cDNA sequence was used to screen a BAC library, yielding a 120 kb clone. When sequenced, it was found to contain the entire Glucanase-1 genomic region (3407 bp) consisting of 9 exons, with the initiator methionine codon at the start of exon 2 (Fig 6A). For an interspecific comparison, the B. mori cDNA sequence was used in a blastn search of the whole genome shotgun (wgs) sequences at NCBI, identifying a total of 5 hits (AADK01004415, BAAB01063034, BAAB01133393, PREDICTED: similar to β-1,3-glucanase, XP_970010.1). All of these sequences have around 60-65% amino acid identity compared to lepidopteran Glucanase-1 and all of them possess a conserved GH16 active site (Fig. 5A). Orthologous sequences were not found in Drosophila sp. or in the honeybee genome.

Identification of 4 distinct betaGRP clades in Lepidoptera. To examine the relationships among βGRP proteins in Lepidoptera, a total of 32 sequences from 17 species (for details refer to Supplementary information S3) were collected and used to construct a Bayesian phylogeny (Fig. 5C). The analysis revealed that these sequences clustered in two distinct clades. One of them falls into three subclades (clades 1, 2 and 3) containing many proteins that had previously been found in insect hemolymph: Mse_BGRP1 and 2 (6,8), Bmo_BGRP and p50_GNBP (7,9), Pin_BGRP (5), and Ha_BGRP1 and 2a (the present study). The second cluster (clade 4) contains the H. armigera Glucanase-1 protein isolated from midgut, and sequences from cDNA libraries made from midgut tissue of different Lepidoptera species. This cluster is clearly separated from the other clades by a posterior probability of one and a large branch length. The topology of the tree is unchanged when the comparison is restricted to the glycosyl hydrolase 16 domains of the proteins, and is similar when the neighbor-joining method is used (data not shown). This phylogeny suggests an ancient duplication event leading to paralogues that have different tissue specificity as well as function.
BAAB01043260, BAAB01083370) that clustered together in one unique contig of 18,022 bp which covers the entire Glucanase-1 gene between base pair numbers 9395 and 15120 for a total of 5725 bp. Similar to the H. armigera one, the B. mori gene consists of 9 exons with the transcription initiation methionine codon also at the beginning of exon 2 (Fig. 6A).

We further identified the flanking genes of Glucanase-1 in both species (Fig. 6B). To do so, we recovered a scaffold of the B. mori genome assembly from SilkDB (scaffold000559, http://silkworm.genomics.org.cn/) that covered the entire B. mori Glucanase-1 gene as well as flanking genes. For H. armigera, a part of the 120 kb BAC clone described above was used for the comparative analysis. All of the sequences were blasted via the NCBI webpage against the non-redundant database using BLASTX. A total of 10 and 7 genes were found on the H. armigera and B. mori sequences respectively (Fig. 6B). The genomic regions of the 2 species contain four genes in common: Transcription factor IIE, Tyrosine receptor kinase, GTPase and Glucanase-1, with similar but not identical orientations. No additional or duplicated BGRP gene was identified within the respective genomic regions. These data strongly suggest that the B. mori and H. armigera Glucanase-1 genes are orthologous and that a larger genomic region is highly syntenic between these two lepidopteran species.

**DISCUSSION**

Here we present the molecular characterization of a member of a new class of β-1,3-glucan recognition protein in Lepidoptera that binds strongly to curdlan, is predominantly expressed in the midgut, is located in the gut lumen thus in direct contact with the food bolus, and exhibits β-1,3-glucanase activity.

Glucanase-1 orthologs are present in many insect orders and thus this BGRP/GNBP gene family is not mosquito-specific as suggested by Warr et al. (26). Within the Lepidoptera, we can find a clear distinction between two classes of βGRPs, one possessing the signature of a glucanase active site and expressed primarily in midgut, and the other lacking this signature but possessing an additional C-terminal domain and expressed in other tissues including hemolymph. The difference in function however is not clear yet. We found that the expression of the gene coding for this protein is induced in the midgut of larvae fed on Gram negative or Gram positive bacteria in the diet, suggesting a role in immune response.

The result of our curdlan pull-down assay with the hemolymph sample was complex and several unexpected proteins were identified apart from the βGRPs 1 and 2a. Finding a C-type lectin associated with curdlan was not surprising because lectins are well known polysaccharide-binding proteins and a previously characterized protein from H. armigera has been shown to have some affinity for curdlan (27). The presence of proPO subunit 2, a proPO-activating factor protein (serine proteinase-like protein) and a serine-proteinase homolog was surprising and is unlikely due to direct binding to curdlan. ProPO is responsible for melanization and has no known affinity for polysaccharides (28) similar to proPO-activating factors (1). PRRs (in our case βGRPs and C-type lectins), proPO and proPO-activating factors all participate in the proPO activation cascade (1) and consequently might at some point interact together. Hence, our data suggest that we have precipitated a complex of these proteins. This is consistent with the suggestion of Yu et al. (29) that, in M. sexta, Immunolectin-2 (a C-type lectin), after binding to surface polysaccharides of microbes, forms a complex together with serine-proteinase homologs and proPO-activating factors.

Our results on the dietary induction of Glucanase-1 by bacteria as well as by LTA and LPS correlate well with observations of one of its ortholog in the mosquito A. gambiae, GNBP-B1, for which the expression is induced upon bacterial challenge with both Gram negative and Gram positive bacteria with a stronger effect observed with the latter (30). No induction of GNBP-B1 is observed upon challenge with S. cerevisiae (30). Furthermore, Warr et al. (26) showed that knocking down the expression of GNBP subgroup B genes in adult female mosquitoes by RNAi increased their susceptibility to immune challenge...
and subsequently their death rate. Our findings suggest that these results should now be considered in light of a possible enzymatic role of GNBP-B proteins. A similar knock-down approach with lepidopteran Glucanase-1 genes combined with bioassays testing larvae behavior feeding on highly microbial-contaminated diet would be a valuable tool in understanding the potential role in immune defense of Glucanase-1 in caterpillars. Development of gene manipulation techniques for Lepidoptera is a growing research field, but their availability is still limited although some progress has been made recently especially with *H. armigera* (31).

In Lepidoptera and other insect orders, pattern recognition proteins including βGRPs/GNBPs trigger an innate immune response after recognition of PAMPs by activating the prophenoloxidase cascade (1,3) and the Toll receptor pathway in hemocytes (32). These events take place in the hemolymph and thus it is unlikely that Glucanase-1 could activate this cascade from its site of activity in the midgut lumen. Furthermore, the presence of two distinct groups of βGRPs in Lepidoptera suggests an ancient duplication event leading to paralogous genes. Subfunctionalization and neofunctionalization could have led to different functions of both these paralogs. Hence, hemolymph specific βGRPs might trigger innate immune response, whereas midgut specific βGRPs (Glucanase-1) possibly fulfill a different function within the organism.

The recognition that glucanases and non-catalytic glucan-binding proteins are similar but distinct clades is hampered by current nomenclature, including the misleading definition of a conserved domain in the NCBI CDD database. Version 2.15 lists a Conserved Domain, cd02179: G16 beta_GRP, described as follows: "Beta-GRP (beta-1,3-glucan recognition protein) is one of several pattern recognition receptors (PRRs), also referred to as biosensor proteins, that complexes with pathogen-associated beta-1,3-glucans and then transduces signals necessary for activation of an appropriate immune response ... ". This description is supported by citations (8, 11) describing studies on the non-catalytic hemolymph bGRP proteins from Lepidoptera. However, none of these lepidopteran proteins lacking the GH16 active site with two glutamates are among the 14 sequences used to construct the consensus sequence for this cd02179 domain. Instead, all 14 sequences possess the active site, including 7 from fungi, 2 from earthworms, 4 from crustaceans, and 2 mosquito sequences (one is a GNBP-B from *A. gambiae*, which as far as we know has not been tested for glucanase activity). Thus biological information from those sequences has been associated with the definition of a domain possessing the catalytic site, yet sequences from Lepidoptera have not been used to represent a separate domain lacking the catalytic site. Recognition of an additional subfamily containing bona-fide GRPs lacking the active site would better reflect the existence of similar but distinct catalytic and non-catalytic clades that we have found.

Our data show that glucanase-1 hydrolyzes specifically β-1,3-glucans. The major source for this type of polysaccharides is the cell wall of bacteria, yeast and fungi. One of its ortholog in *A. gambiae*, GNBP-B4, binds to the surface of both Gram negative and Gram positive bacteria (26), but its potential beta-1,3-glucanase activity has not been tested yet. Genta et al. (33) purified a 46 kDa endo-β-1,3-glucanase from salivary glands of the cockroach *Periplaneta americana*. Their purified protein was able to lyse *S. cerevisiae* cells in hypotonic medium lacking nutrients, making their content available as a nutritive source, but no lysis was observed in isotonic nutritive medium. The authors also pointed out that although it is conceivable that this enzyme may protect midgut cells from microbial invasion, its presence in saliva and the large amounts of fungi usually found in detritus, the major food source of *P. americana*, points more strongly to a digestive role. We could not observe any lytic activity by plate assay on any microbe tested using partially purified Glucanase-1 from culture medium of High Five transfected cells (data not shown) although the recombinant protein exhibits glucanase activity. We cannot exclude that this lack of activity is due to a too low concentration of recombinant Glucanase-1. A more thorough investigation of the potential anti-
microbial activity of Glucanase-1 should be undertaken.

Similarly, it has been proposed that β-1,3-glucanase from plants may act synergistically with other hydrolases, such as chitinases and proteinases, to disrupt the structural integrity of fungal cell wall (34). Alternatively they may release oligosaccharides from β-1,3-glucan substrates that may serve as chemical signals leading to the activation of other defense responses (35).

In addition, the midgut of herbivorous caterpillars is composed of a highly diverse microbial community that might contribute to its physiological function (36-38) and it is unclear how Glucanase-1 could differentiate between them, as the molecular motifs activating the innate immune response are present in both beneficial and intruding microorganisms. One hypothesis is that Glucanase-1 has a similar function as PGRP-LB in Drosophila in preventing local immune activation by commensal gut bacteria (15). PGRP-LB, in contrast to PGRP-LC, is an active amidase able to digest peptidoglycan molecules released by dividing gram-negative bacteria. PGRP-LB regulates the level of immune response by scavenging peptidoglycan present in the Drosophila larval gut. If this scavenging effect is overwhelmed, by bacterial infection for example, the excess peptidoglycan is recognized by the nonenzymatic PGRP-LC which activates the immune response through the Imd pathway. In the case of peptidoglycan released by commensal gut flora, the scavenging effect of PGRP-LB is sufficient to prevent activation of the Imd pathway and thus to inhibit any immune response.

Our results raise the following questions: (i) Can Glucanase-1 be considered as a true pattern recognition protein? (ii) What is the role of Glucanase-1 in the midgut: Immune defence protein or digestive enzyme? The present study is a first step in understanding the physiological role of this new lepidopteran midgut-specific βGRP protein family. More investigations have to be conducted in order to fully address these questions.

REFERENCES

1. Kanost, M. R., Jiang, H., and Yu, X. Q. (2004) Immunol Rev 198, 97-105
2. Medzhitov, R., and Janeway, C. A., Jr. (1997) Cell 91(3), 295-298
3. Kim, Y. S., Ryu, J. H., Han, S. J., Choi, K. H., Nam, K. B., Jang, I. H., Lemaitre, B., Brey, P. T., and Lee, W. J. (2000) J Biol Chem 275(42), 32721-32727
4. Cerenius, L., and Soderhall, K. (2004) Immunol Rev 198, 116-126
5. Fabrick, J. A., Baker, J. E., and Kanost, M. R. (2003) Insect Biochem Mol Biol 33(6), 579-594
6. Jiang, H., Ma, C., Lu, Z. Q., and Kanost, M. R. (2004) Insect Biochem Mol Biol 34(1), 89-100
7. Lee, W. J., Lee, J. D., Kravchenko, V. V., Ulevitch, R. J., and Brey, P. T. (1996) Proc Natl Acad Sci U S A 93(15), 7888-7893
8. Ma, C., and Kanost, M. R. (2000) J Biol Chem 275(11), 7505-7514
9. Ochiai, M., and Ashida, M. (2000) J Biol Chem 275(7), 4995-5002
10. Shin, S. W., Park, S. S., Park, D. S., Kim, M. G., Kim, S. C., Brey, P. T., and Park, H. Y. (1998) Insect Biochem Mol Biol 28(11), 827-837
11. Fabrick, J. A., Baker, J. E., and Kanost, M. R. (2004) J Biol Chem 279(25), 26605-26611
12. Freitak, D., Wheat, C. W., Heckel, D. G., and Vogel, H. (2007) BMC Biol 5, 56
13. Liehl, P., Blight, M., Vodovar, N., Boccard, F., and Lemaitre, B. (2006) PLoS Pathog 2(6), e56
14. Nehme, N. T., Liegeois, S., Kele, B., Giammarinaro, P., Pradel, E., Hoffmann, J. A., Ewbank, J. J., and Ferrandon, D. (2007) *PLoS Pathog* 3(11), e173
15. Zaidman-Remy, A., Herve, M., Poidevin, M., Pili-Floury, S., Kim, M. S., Blanot, D., Oh, B. H., Ueda, R., Mengin-Lecreulx, D., and Lemaître, B. (2006) *Immunity* 24(4), 463-473
16. Simpson, R. M., Newcomb, R. D., Gatehouse, H. S., Crowhurst, R. N., Chagne, D., Gatehouse, L. N., Markwick, N. P., Beuning, L. L., Murray, C., Marshall, S. D., Yauk, Y. K., Nain, B., Wang, Y. Y., Gleave, A. P., and Christeller, J. T. (2007) *Insect Mol Biol* 16(6), 675-690
17. Pauchet, Y., Muck, A., Svatos, A., Heckel, D. G., and Preiss, S. (2008) *J Proteome Res* 7(4), 1629-1639
18. Juncosa, M., Pons, J., Dot, T., Querol, E., and Planas, A. (1994) *J Biol Chem* 269(20), 14530-14535
19. Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P., Ens, W., and Standing, K. G. (2001) *Anal Chem* 73(9), 1917-1926
20. Habermann, B., Oegema, J., Sunyaev, S., and Shevchenko, A. (2004) *Mol Cell Proteomics* 3(3), 238-249
21. Fischer, H. M., Wheat, C. W., Heckel, D. G., and Vogel, H. (2008) *Mol Biol Evol* 25(5), 809-820
22. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res* 25(24), 4876-4882
23. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002) *Nucleic Acids Res* 30(14), 3059-3066
24. Ronquist, F., and Huelsenbeck, J. P. (2003) *Bioinformatics* 19(12), 1572-1574
25. Bulmer, M. S., and Crozier, R. H. (2006) *Mol Biol Evol* 23(2), 317-326
26. Warr, E., Das, S., Dong, Y., and Dimopoulos, G. (2008) *Insect Mol Biol* 17(1), 39-51
27. Chai, L. Q., Tian, Y. Y., Yang, D. T., Wang, J. X., and Zhao, X. F. (2008) *Dev Comp Immunol* 32(1), 71-83
28. Hall, M., Scott, T., Sugumaran, M., Soderhall, K., and Law, J. H. (1995) *Proc Natl Acad Sci USA* 92(17), 7764-7768
29. Yu, X. Q., Jiang, H., Wang, Y., and Kanost, M. R. (2003) *Insect Biochem Mol Biol* 33(2), 197-208
30. Dimopoulos, G., Richman, A., Muller, H. M., and Kafatos, F. C. (1997) *Proc Natl Acad Sci USA* 94(21), 11508-11513
31. Mao, Y. B., Cai, W. J., Wang, J. W., Hong, G. J., Tao, X. Y., Wang, L. J., Huang, Y. P., and Chen, X. Y. (2007) *Nat Biotechnol* 25(11), 1307-1313
32. Ao, J. Q., Ling, E., and Yu, X. Q. (2008) *Mol Immunol* 45(2), 543-552
33. Genta, F. A., Terra, W. R., and Ferreira, C. (2003) *Insect Biochem Mol Biol* 33(11), 1085-1097
34. Mauch, F., Mauch-Mani, B., and Boller, T. (1988) *Plant Physiol* 88(3), 936-942
35. Lamb, C. J., Lawton, M. A., Dron, M., and Dixon, R. A. (1989) *Cell* 56(2), 215-224
36. Ping, L., Buchler, R., Mithofer, A., Svatos, A., Spiteller, D., Dettner, K., Gmeiner, S., Piel, J., Schlott, B., and Boland, W. (2007) *Environ Microbiol* 9(6), 1572-1583
37. Spiteller, D., Dettner, K., and Boland, W. (2000) *Biol Chem* 381(8), 755-762
FOOTNOTES

We are particularly grateful to A Muck and A Svatoš for the Mass Spec support and to D Schnabelrauch and H Ringys-Beckstein for providing technical assistance for sequencing. We are grateful to Dr Siu Fai Lee for providing us with his *O. nubilalis Glucanase-1* sequence. This work was supported by the Max Planck Society. DF is supported by an International Max Planck Research School (IMPRS) PhD fellowship.

FIGURE LEGENDS

**Figure 1:** Purification and tissue expression of *H. armigera* glucanase-1. (A) Curdlan pull-down assay on gut larval lumen sample. Proteins were separated by SDS-PAGE using a 12% polyacrylamide resolving gel and stained with coomassie. Lane 1: 10 µg gut lumen proteins, lane 2: 10 µg gut lumen flow-through, lane 3: 15 µl elution fraction. Tryptic peptides obtained from band 1 were analyzed by MALDI-TOF/MS (see Supplementary information S2). (B) Tissue expression analysis of glucanase-1 by real-time quantitative RT-PCR, the different tissue investigated were midgut, salivary glands, hemocytes, fat body and rest body. Relative fold differences were calculated in comparison to glucanase-1 expression in a pooled sample of adults and pupae for which relative fold expression was set to 1 (whole animals).

**Figure 2:** Glucanase activity assays on *H. armigera* larval gut lumen. (A) Assays were performed on 1.5% agar plates containing 0.25% of several types of beta-glucans: CM-curdlan (beta-1,3-glucan), lichenan and barley beta-glucan (mixed beta-1,3:1,4-glucans) and CM-cellulose (beta-1,4-glucan). Activity was revealed by Congo Red staining. Lane 1: larval gut lumen, lane 2: PBS (negative control), lane 3: endo beta-1,3-glucanase from *Trichoderma* sp. and lane 4: cellulase from *Trichoderma* sp. (both positive controls). (B) Beta-1,3-glucan zymogram of *H. armigera* larval gut lumen. Ten micrograms proteins were separated by native PAGE and the gel overlaid on a CM-curdlan-containing agar plate. Activity was revealed by Congo red staining. Lane 1: gut lumen proteins, lane 2: gut lumen proteins after curdlan-pull down assay (flow-through). (C) Analysis of the expression of *H. armigera* glucanase-1 in transfected *T. ni* High Five cells. Ten micrograms of proteins from culture medium and crude membrane fraction of both glucanase-1- and empty vector-transfected cells were resolved by SDS-PAGE using a 12% polyacrylamide resolving gel, then transferred to PVDF and revealed by western blot using the anti-V5-HRP antibody. Line 1: culture medium from control cells, Line 2: culture medium from glucanase-1-transfected cells, line 3: crude-membrane fraction from control cells, line 4: crude-membrane fraction from glucanase-1-transfected cells. (D) Beta-1,3-glucanase activity in culture medium from control and glucanase-1-transfected cells was investigated by zymogram. Ten micrograms of proteins were resolved by native PAGE and the gel overlaid on a CM-curdlan-containing agar plate. Activity was revealed by Congo red staining. Line 1: culture medium from glucanase-1-transfected cells, line 2: culture medium from control cells.

**Figure 3:** Variation of expression of glucanase-1 in the midgut of *H. armigera* larvae upon feeding on microbial- or polysaccharides-contaminated diet. Larvae were fed for 24h on artificial diet contaminated with either microbes (*E. coli*, *M. luteus* and *S. cerevisiae*) or bacterial cell wall polysaccharides (LTA and LPS) or various beta-glucans (Lichenan, Curdlan and Barley beta-glucan). Relative fold differences were
calculated in comparison to the data obtained for larvae fed on sterile artificial diet for which relative fold expression was set to 1.

**Figure 4:** Curdlan pull-down assay on larval hemolymph sample collected from unchallenged larvae. Proteins were separated by SDS-PAGE using a 10% polyacrylamide resolving gel and stained with coomassie. Lane 1: 10 µg cell-free hemolymph proteins, lane 2: 10 µg cell-free hemolymph flow-through, lane 3: 15 µl elution fraction. Tryptic peptides obtained from bands a-g were analyzed by nanoLC-MS/MS and interpreted de novo, search results are given in Table 1.

**Figure 5:** (A) Alignment of the predicted GH16 active site of 2 lepidopteran Glucanase-1 and GNBPs from termite and mosquitoes compared with two lepidopteran hemolymph-specific proteins. Consensus active site for GH16 family was obtained from PROSITE (http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00794). Bmo: *B. mori*, Har: *H. armigera*, Nfu: *N. fumigatus*, Aae: *A. aegypti*, Aga: *A. gambiae* and Msc: *M. sexta*. (B) Schematic structure of lepidopteran Glucanase-1 and hemolymph-specific betaGRPs. CRD: Carbohydrate Recognition Domain. (C) Bayesian tree (rooted) of putative betaGRPs protein sequences expressed in several lepidopteran species. Sequences were obtained from Genbank, in-house EST databases and SilkDB. Posterior probability for each branch point is given. Relevant information concerning the sequences used for this Phylogenetic analysis is summarized in Supplementary Information S3.

**Figure 6:** (A) Schematic representation of the intron/exon structure of *H. armigera* and *B. mori* Glucanase-1 genes. Boxes represent exons: white boxes correspond to the gene CDS and gray boxes represent the 5’- and 3’-UTR found on the transcripts. Size (in nucleotides) is shown below each exon. Lines represent intron insertions (drawn to scale), and the size of each intron is indicated above the line. Initiator methionine codon (ATG) and stop codon (TAA) are also indicated for both species. (B) Schematic representation of the genomic region around the Glucanase-1 gene deduced from a *B. mori* scaffold obtained from NCBI and sequencing of a *H. armigera* BAC clone. Black arrows: putative genes common between the 2 species, grey arrows: other putative genes.
Table I: Protein identification from the curdlan pull-down assay on *H. armigera* larval hemolymph

| band | database searched | hit | species            | accession | de novo sequenced peptides<sup>3</sup> | peptides matched by MS BLAST | MS BLAST score<sup>4</sup> |
|------|-------------------|-----|--------------------|-----------|-------------------------------------|-----------------------------|---------------------------|
| a    | ncbi_insecta      | proPO subunit 2 | *H. armigera*     | AAZ52554  | 5                                   | 5                           | 336                       |
| b    | Ha_EST_db         | beta-1,3-glucan recognition protein 2a | *H. armigera* | EU770382* | 4                                   | 1                           | 78                        |
| c    | ncbi_insecta      | beta-1,3-glucan recognition protein 1  | *B. mori*       | NP_001036840 | 5                                   | 3                           | 182                       |
| c    | Ha_EST_db         | beta-1,3-glucan recognition protein 1  | *H. armigera*   | EU770381*  | 5                                   | 1                           | 101                       |
| d    | not identified    |     |                    |           | 6                                   |                             |                           |
| e    | Ha_EST_db         | Serine-proteinase like protein 1      | *H. armigera*   | EU770389*  | 3                                   | 2                           | 145                       |
| f    | Ha_EST_db         | C-type lectin                         | *H. armigera*   | EU770390*  | 3                                   | 1                           | 64                        |
| g    | ncbi_insecta      | masquerade-like serine protease       | *B. mori*       | NP_001037053 | 6                                   | 2                           | 113                       |
| g    | Ha_EST_db         | Serine-proteinase homolog-1           | *H. armigera*   | EU770391*  | 6                                   | 4                           | 192                       |

<sup>1</sup>For hits obtained by searching ncbi_insecta, the annotation is the one given by the database. For hits obtained by searching our *H. armigera* EST database, the annotation has been given according to the *M. sexta* homologs (Kanost et al., 2004).<sup>2</sup>Genbank accession numbers (*from this study).<sup>3</sup>Details on the *de novo* sequenced peptides are given in Supplementary Information S4.<sup>4</sup>MS BLAST scoring is detailed in the Material and Methods section.
Figure 1
Figure 2
Figure 3

Figure 4
**Figure 5**

**A**

GH16 consensus

|    | L | D | L | M | L | K | P |
|----|---|---|---|---|---|---|---|
| GH16 | I | I | Q | S | V | V | N |
| GH16 | F | F | A |

Har-Glucl 186-8 G E I D L V E S R G N A-199
Bmo-Glucl 188-8 G E I D L V E S R G N K N-199
Aar-GNBP 187-7 R T R N R T R N R N T-198
Ase-GNBP 204-8 G E I D L M E S R G N L D-215
Nfu-GNBP 188-8 G E I D L T E S R G N K Q-199

Mse-BGRF2 317-8 G L M R V A F V R G N D V-330
Har-BGRF2a 324-8 G M N R V A F A R G N P S-337

**B**

**Glucanase**

- Signal peptide
- Hemolymph beta-GRP
- CRD

**C**

- **Not_gluclase**
- **Tmo_BGRP**
- **Dme_GNBP3**
- **Mse_BGRF2**
- **Pra_BGRP**
- **yYHar_BGRF2a**
- **Tc_BGRF2**
- **yGHar_BGRF2b**
- **Sk_BGRF2**
- **coDn_BGRP**
- **Yp_BGRP**
- **Mse_BGRP1**
- **Dme_BGRP**
- **Bmo_BGRP**
- **gsoHar_BGRP1**
- **Aar_BGRP**
- **Bmo_BGRP2**
- **Ase_BGRP**
- **Pra_BGRP2**
- **Mso_BGRF2**
- **Hcu_BGRP**
- **Hcu_BGRP2**
- **coDoD_BGRP**
- **Pra_BGRP3**
- **gsoAme_GNBP1**
- **Dme_GNBP1**
- **Nl_GNBP1**
- **Pyr_glucl**
- **Pin_glucl**
- **Eps_glucl**
- **Dsa_glucl**
- **Ceu_glucl**
- **coPra_glucl**
- **Aar_glucl**
- **Bmo_glucl**
- **Onu_glucl**
- **Hiu_glucl**
- **HGr_glucl**
- **Spi_glucl**
- **coAqA_GNBP83**
- **Ase_GNBP**

**Clade 2**

Lepidopteran homologous-specific betaGRPs

**Clade 1**

**Clade 3**

Lepidopteran midgut-specific β-1,3-gluclase-1
Figure 6
