Tissue-specific Proteogenomic Analysis of *Plutella xylostella* Larval Midgut Using a Multialgorithm Pipeline

Xun Zhu‡**, Shangbo Xie§**, Jean Armengaud¶, Wen Xie‡, Zhaojiang Guo‡, Shi Kang‡, Qingjun Wu‡, Shaoli Wang‡, Jixing Xia‡, Rongjun He‡, and Youjun Zhang‡

The diamondback moth, *Plutella xylostella* (L.), is the major cosmopolitan pest of brassica and other cruciferous crops. Its larval midgut is a dynamic tissue that interfaces with a wide variety of toxicological and physiological processes. The draft sequence of the *P. xylostella* genome was recently released, but its annotation remains challenging because of the low sequence coverage of this branch of life and the poor description of exon/intron splicing rules for these insects. Peptide sequencing by computational assignment of tandem mass spectra to genome sequence information provides an experimental independent approach for confirming or refuting protein predictions, a concept that has been termed proteogenomics. In this study, we carried out an in-depth proteogenomic analysis to complement genome annotation of *P. xylostella* larval midgut based on shotgun HPLC-ESI-MS/MS data by means of a multialgorithm pipeline. A total of 876,341 tandem mass spectra were searched against the predicted *P. xylostella* protein sequences and a whole-genome six-frame translation database. Based on a data set comprising 2694 novel genome search specific peptides, we discovered 439 novel protein-coding genes and corrected 128 existing gene models. To get the most accurate data to seed further insect genome annotation, more than half of the novel protein-coding genes, i.e. 235 over 439, were further validated after RT-PCR amplification and sequencing of the corresponding transcripts. Furthermore, we validated 53 novel alternative splicings. Finally, a total of 6744 proteins were identified, resulting in one of the most comprehensive proteogenomic study of a nonmodel animal. As the first tissue-specific proteogenomics analysis of *P. xylostella*, this study provides the fundamental basis for high-throughput proteomics and functional genomics approaches aimed at deciphering the molecular mechanisms of resistance and controlling this pest. *Molecular & Cellular Proteomics* 15: 10.1074/mcp.M115.050989, 1791–1807, 2016.

Thousands of eukaryotic genomes have been sequenced with the rapid development of massive parallel sequencing technologies. The annotation of their genomic DNA sequences is a prerequisite to delineate conserved elements and understand the specific encoded functions. Today, conventional approaches used to identify protein-coding genes are highly dependent on predictions based on computational algorithms and homology searches against known proteins. Nonmodel organisms, such as most arthropods, are by nature distantly related to well-studied organisms (1). As a result, their genomes encode a large majority of proteins with little similarity to known proteins, challenging the annotation process. Much recent focus on computational gene finding is using transcript evidence to complement this prediction-based approach. Indeed, the use of information from deep sequencing of mRNA-derived cDNA libraries (RNA-seq) or expressed sequence tag (EST) libraries can dramatically improve the genome annotation confidence (2–4). However, this analysis remains at the transcript level and cannot make a crystal-clear distinction between coding and noncoding sequences in many cases. A growing body of evidence suggests that a discrepancy exists between protein isoforms that are transcribed *versus* those that are actually translated (5). Also, there is some evidence for genes, when sampling proteins by mass spectrometry, that are not always visible at the transcript level (6). Moreover, transcript evidence is commonly confounded by prespliced messages, nontargeted expression noise, ncRNA, and lack of strand and frame information. As a result, genome annotations are imperfect (7). Many genes still remain to be confirmed, the proteins being labeled as hypothetical or conserved hypothetical whether their sequences are unique, *i.e.* orphan, or found in other closely...
related organisms only, or found in a wider range of organisms. The real existence and function of their products should be established (8). These cases question the final outcome of the gene finding procedure. To improve genome annotation, it has been proposed to use high-throughput proteomics data that allows to definitively prove the real existence of the gene products, a concept called “proteogenomics” (9, 10).

Tandem mass spectrometry (MS/MS), as a key technology for examining the proteome, has emerged as a sensitive high-throughput method in which thousands of proteins can be identified in a single experiment (11). MS/MS-certified peptides can be used as experimental evidence to identify protein-coding genes and correct the possible annotation errors (11). Proteogenomics, sensu stricto (12), always involves four steps when applied to eukaryotes: (1) harvesting peptide sequences by means of MS/MS, (2) comparing these sequences to the previously published annotation of protein sequences of the organism, (3) analyzing the six-frame translation of the entire genome to establish novel genome search specific peptides (GSSPs), and (4) deciphering the possible splice variants. The peptide sequences are used, once mapped back to the entire genome, for confirming, refuting, or refining the existing gene annotations. Other sources of evidence can be also incorporated in the proteogenomics procedure.

Proteogenomics has been successfully applied to re-annotate several bacterial and archaeal genomes as well as fungi: Roseobacter denitrificans (13), Tistila consotensis (14), Mycoplasma pneumoniae (15), Rhodopseudomonas palustris (16), Shewanella oneidensis (zref17), Thermococcus gammatolerans (18), Deinococcus deserti (19, 20), Salmonella thyphimu- rium (21), Mycobacterium tuberculosis (22), Shigella flexner I (23), Leptospira interrogans (24), Ruegeria pomeroyi (25), Mycobacterium genus (26), Candida glabrata (27), and Bradyrhizobium japonicum (28). Because of the difficulty to carry out, few proteogenomic studies have been till now reported for higher organisms, such as for Human (29), the amphibiod Gammarus fossarum (30), the insect Anopheles gambiae (31), and the model plant Arabidopsis thaliana (32, 33).

Arthropoda is the largest animal phylum on Earth, in terms of both population and diversity. It includes insects, millipedes, mites, spiders, and crustaceans such as shrimp and crabs. More than 1 million species have been described just for insects. As herbivores and pollinators of plants, arthropods (especially insects) are essential to numerous terrestrial, aquatic, and marine ecosystems. Although some are competing with humans for crop plants and transmitting diseases to humans and livestock, their molecular study is just in its infancy. Over fifty insect species have been genome sequenced and this knowledge will rapidly expand in the near future by means of the 5000 arthropod genomes initiative (i5K) (34). Proteogenomics data are of utmost interest for seeding the best genome annotation of this key animal phylogeny.

The diamondback moth (DBM), Plutella xylostella (Lepidoptera: Plutellidae), is the major cosmopolitan pest of cruciferous crops (35). The cost of global damage because of the insect pest and its control has been estimated as US$4–5 billion annually worldwide (36). As the troublesome insect pest has been especially problematic, the use of insecticides has been the only successful way to control it. However, extensive insecticide application has led to a rapid increase in its resistance to not only chemical pesticides but also biological pesticides, including pyrethroids, agricultural antibiotics, and Bacillus thuringiensis (Bt) toxins, one of the most successful microbial insecticides used for pest population control (37). Molecular studies have been extensively conducted previously in P. xylostella, especially in terms of insect physiology and insecticide resistance, including data on chemosensory proteins (38), immunity and defense in insect (39), hormonal regulation (40), cuticle function (41), and the mechanisms of resistance to insecticide (42), especially to Bt toxins (43).

The larval midgut lumen of lepidopteran, as a hostile environment for proteins, is a main tissue for not only food digestion and nutrient absorption for the insect body, but also protection from pathogen invasion for the insect, which is also an important target for chemical and biological insecticides or controlling strategies. As the digestive enzymes and defensive molecules in the insect midgut persist in this challenging environment, it is of utmost interest to study how this complex bioreactor functions at the protein level (44). Some studies in Lepidoptera midgut reported involvement of lipases, pro- teases, and carbohydrates in digestion, as well as cytochrome P450, glutathione-s-transferases (GST), and carboxylesterases in xenobiotic metabolism (45). For explaining Bt-resistance, receptors in the midgut of the target pest were investigated. Omics approaches started to be applied to study the physiological and toxicological changes at a global scale in the midgut. ESTs and cDNA microarray technology were used to study immune-inducible genes in P. xylostella (39). More recently, Xie et al. (46) provided more insights into P. xylostella EST database to understand the molecular mechanism of insecticide resistance, using an Illumina-based transcriptome profiling technique. cDNA libraries were used to develop genomic approaches to enrich mRNAs and identify key players from the surrounding cells, but they failed to identify proteins secreted into the lumen and their temporal dynamics. Liu et al. (47) identified the midgut protein profile of sixth instar larvae of Spodoptera litura by HPLC-ESI-MS/MS. In the larval midgut of Bombyx mori 1100 proteins were detected using a 2-DE/MS approach (48). The 2-DE/MS/MS de novo peptide sequencing technology was used to detect major proteins in the midgut lumen of the cotton bollworm Helicoverpa armigera (44).

The draft sequence of the P. xylostella genome was released recently (49) and provided the conditions to uncover the molecular function of the larval midgut in P. xylostella. In the present study, we have carried out an in-depth proteomic analysis of midgut proteins of P. xylostella using shotgun nanoLC-MS/MS. The mass spectrometry-derived data were
analyzed by a multialgorithm pipeline using a six-frame translation of genome sequences in addition to searches of a protein database of *P. xylostella*. We used the MS/MS-certified GSSPs to obtain new gene models that have been further validated by a comparative approach based on the BLAT and geneewise tools. The resulting mass spectrometry-certified genome annotation comprises a total of 6764 proteins resulting in the most comprehensive proteogenomic study of arthropods and allows a strong genome annotation seed for this huge branch of animal life.

**MATERIALS AND METHODS**

**Insect Strains and Toxins**—The two *P. xylostella* strains used in this study have been described elsewhere (50–52). A highly inbred laboratory strain of *P. xylostella*, DBM1Ac-S (also known as Geneva 88), was used as the susceptible strain. It was collected in 1988 in the fields at the New York State Agricultural Experiment Station, Robbins Farm, Geneva, NY, and has been maintained on JingFeng No. 1 cabbage (*Brassica oleracea* var. *capitata*) Beijing, China in our laboratory without insecticide exposure for more than ten years since 2003. A Cry1Ac-resistant strain of *P. xylostella* near-isogenic to the susceptible DBM1Ac-S strain, NIL-R (also known as BC6F4), was derived from a Cry1Ac-resistant population collected from a cruciferous field in Loxahatchee (FL) that had been extensively treated with Bt var. *kurstaki* (Btk) formulations. The NIL-R strain had been backcrossed with the DBM1Ac-S strain seven times. Larvae from the NIL-R strain displayed 50% fold resistance to Cry1Ac protoxin. All the strains were reared on JingFeng No. 1 cabbage without exposure to any Bt toxins or chemical pesticides. The raising conditions of the two biological replicates were described elsewhere (54). About 5000 4th instar larvae for each biological replicate were dissected to obtain 0.4–0.6 g midgut homogenates (wet weight), thus a total of 25,000 larvae were used for the five biological replicate were dissected to obtain 0.4–0.6 g midgut homogenates (wet weight), thus a total of 25,000 larvae were used for the five biological replicates. About 300–400 μg BBMV proteins were obtained from each midgut sample. Purified BBMV proteins (100 μl) were dissolved in 100 μl protein loading buffer and denatured at 95 °C for 5 min prior to SDS-PAGE in 12.5% polyacrylamide gels with a migration of 20 min at 10 mA and then 1.5 h at 20 mA and Tris-glycine-SDS buffer (10 mM Tris, 50 mM glycine, 0.1% SDS, pH 8.0) as running buffer. Gels were stained with Simply Blue Safe Stain, Comassie Brilliant Blue G250 from Invitrogen, Beijing, China. Purified BBMV proteins were quantified using the method of Bradford with bovine serum albumin as standard, and then flash frozen and kept in aliquots at −80 °C until used.

Each gel lane was cut into 20 pieces in equal size cut from top to bottom. Those resulting polyacrylamide bands were digested by trypsin according to Shchvchenko et al. (55). Those gel pieces were treated with 0.2 ml of 100 mM NH4HCO3 in 50% acetonitrile for 40 min at 37 °C, and then with acetonitrile for 5 min. Those gel pieces were dried for 20 min under a vacuum centrifuge. For in-gel digestion, those dry gel pieces were treated with 25 μl NH4HCO3 containing 10 μl dithiothreitol at 56 °C for 45 min. Gel pieces were cooled to room temperature and then treated for 45 min with 25 μl NH4HCO3 containing 55 μl iodoacetamide. The gel pieces were then rinsed with 25 μl NH4HCO3 (100 μl) for 10 min. After dehydration in 100 μl of acetonitrile, the gel pieces were swollen by rehydration in 100 μl of 25 mM NH4HCO3. And then the same volume of acetonitrile was added to shrink again. The gel pieces were dried for 20 min under a vacuum centrifuge after removing the liquid phase, swollen for 45 min in 10 μl of digestion buffer (12.5 ng/μl of trypsin, 5 mM CaCl2, 25 mM NH4HCO3) on ice. And then 10 μl of the same buffer without trypsin were used to replace the supernatant, incubated at 37 °C for overnight. Peptides were extracted by 25 μl NH4HCO3 and then, once by 200 μl of 50% acetonitrile and 0.1% Formic acid, eluted twice by 200 μl of 100% acetonitrile and 0.1% Formic acid (20 min for each elution) at room temperature. All the supernatants were collected and merged.

**nanoLC-MS/MS Shotgun Analysis—Chromatography**—was performed using a surveyor LC system (Shimadzu LC-20AD, Shanghai, China). The samples were loaded by the autosampler onto a 2 cm C18 trap column (inner diameter 200 μm). The pump flow rate was set at 15 μl/min for 4 min, then the 44 min gradient was run at 400 nL/min starting from 2 to 35% B (98%ACN, 0.1%FA), followed by 2 min linear gradient to 80%, and maintained at 80% B for 4 min, and finally returned to 2% in 1 min. The peptides were subjected to nanoelectrospray ionization followed by tandem mass spectrometry (MS/MS) in an LTO Orbitrap Velos (Thermo, Shanghai, China) coupled online to the HPLC. Intact peptides were detected in the Orbitrap at a resolution of 60,000. Peptides were selected for MS/MS using high energy collision dissociation (HCD) operating mode with a normalized collision energy setting of 45% and ion fragments were detected in the LTQ. A data-dependent procedure that alternated between one MS scan followed by eight MS/MS scans was applied for the eight most abundant precursor ions above a threshold ion count of 5000 in the MS survey scan with the following Dynamic Exclusion settings: repeat counts, 2; repeat duration, 30 s; and exclusion duration, 120 s. The electrospray voltage applied was 1.5 kV. Automatic gain control (AGC) was used to prevent overfilling of the ion trap; 1 × 106 ions were accumulated in the ion trap for generation of HCD spectra. For MS scans, the m/z scan range was 350 to 2000 Da.

**Database Searches for Peptide and Protein Identification**—Raw data files were processed to generate peak list files by Xcalibur2.1.0 (ReAdW.exe). The filtering parameters used were: (1) allowed precursor mass range was 500–5000 Da; (2) precursor charge state was allowed from 1 to 5; (3) minimum number of peaks in a spectrum was chosen to be 5; (4) signal to noise ratio was set as 3; and (5) for precursors with unrecognized charge state, default charge states of 2 and 3 were allowed. The protein sequences databases generated from several sources were used for spectral searching. Protein sequences from the diamondback moth genome annotations (DBM-DB v1.0) were downloaded from the genome database (http://iae.fafu.edu.cn/DBM/index.php), which were used to construct a protein database (PDB). A six-frame translation of diamondback moth genome, named Genome Database (GDB), was generated based on the DBM-DB v1.0 genome sequence by EMBoss:6.2.0. A threshold of no less than 30 putative amino acids (a.a.) was chosen to discard small open reading frames (from stop codon to stop codon) to minimize the database size. Based on the genome database, an exon skipping database (ASDB) was constructed. The variant peptides generated were appended to the genome translation databases as separate entries. Three different search engines, Mascot (version 2.3.02), omsssa (version 2.1.1), and xtandem (version 2009.10.01.1) were used to analyze the data. The parameters for database search were as follow: peptide mass tolerance 20 ppm; fragment mass tolerance 0.6 Da; fixed modifications “Carbamidomethyl (C)”; variable modifications “Gln->pyro-Glu (N-term Q), Oxidation (M), Deamidated (NG)” and one miss-cleavage allowed. The results were filtered at 1% FDR at peptide level and each protein identified with at least one unique peptide. Because six different types of searches were performed with the same data set, the identification percentage was calculated based on the following formula: Identification percentage (%) = (the number of protein sequences statistically significant at 1% FDR)/(the total number of protein sequences that were searched).
The goal of this study was to carry out in-depth proteomic profiling to map the proteome of larval midgut of *P. xylostella* at a global scale using a proteogenomic approach. Fig. 1 shows the specific developed workflow. First, protein extracts from the midgut of *P. xylostella* larvae were resolved by SDS-PAGE. The gel was cut into 10 (preliminary experiment) or 20 (four biological replicates) pieces, equal in size each. The five biological replicates were processed leading to a total of 90 peptide fractions, which were analyzed by nanoLC-MS/MS. The resulting MS/MS data were searched using iPeak search algorithm against three databases, i.e. an annotation protein database (PDB), a six-frame translated protein database (GDB), and an alternative splicing database (ASDB), as well as a contaminant database from http://maxquant.org/contaminants.zip with 247 sequences of common contaminants. PDB was used to confirm the gene models annotated in the DBM genome. It comprises all protein sequences from the published version of the genome annotations (DBM-DB.v1.0). GDB was used to find new gene models. It was generated from a six-frame translation of the entire genome (DBM-DB.v1.0). ASDB is an exon skipping database from DBM genome exons that was used to find alternative splicings. For the DBM genome database (DBM-DB), many redundant genomic regions were characterized as “n” to further assemble the DBM genome with less scaffolds. These genomic sequences were removed in the DBM genome version 2. Because the gene sets were not updated in the new genome version (59), the DBM genome database version 1 was used to establish the protein sequence databases described here above. The sizes of the three databases are indicated in Table I.

The iPeak program was developed to automate spectral searches, statistical integration of PSMs and downstream analysis. It is configured with two open source peptide identification algorithms namely OMSSA, XI-Tandem and the commercial software Mascot. These algorithms differ in their basic scoring. For instance, OMSSA uses a Poisson distribution for separating significant matches from random hits (60) whereas XI-Tandem uses a hyper-geometric model (61). Mascot calculates protein ratios using only ratios from the spectra that are distinct for each protein and excluding the shared peptides of protein isoforms. Although there are rescoring algorithms to improve database search results (62, 63), a diverse set of algorithms provide an increased sensitivity for peptide identifications than any single algorithm. iPeak incorporates a reversed protein decoy database strategy to calculate FDR. A total of 876,341 MS/MS spectra were acquired in this study, out of which 300,078 were assigned to peptide spectrum matches (PSMs) with a 1% false discovery rate (FDR) threshold. The complete list of PSM identified in our study along with scores assessed by the three identification algorithms, charge, m/z value, post-translational modifications and genomic coordinates in the DBM genome database is provided in supplemental Data file S1. Fig. 2A depicts all the peptides (16, 151) identified in this study mapped to their genome and protein database searches using Mascot, omssa, and xtandem), peptide sequences assigned to a single spectrum in each search were compared. Spectra that were assigned different sequences in different searches were omitted from further analysis. Protein identification list was generated by grouping proteins based on shared peptides. Proteins that have at least one unique peptide were selected from the group. From a group where no protein could be distinctly selected above others, that is, all proteins had equal evidence, it was represented by only one in the final list marked with an asterisk and remaining equivalent protein IDs were reported in parentheses.

**Peptide Mapping and Analysis with Genome Annotation**—The full list of spectral matched peptides was established as mentioned here above with a FDR cut-off of 1%. The combination of database/peptide and database/template coordinate tables were used to map these peptides back to its genomic location(s), considering the split peptides because of spanning splice junctions. Genome coordinates of all the peptides were found out using blast program. Genome search specific peptides (GSSPs) were identified by excluding those peptides that mapped to known proteins from translated genome database search results. GSSPs were categorized into three parts, including mapping to intergenic region, partially overlapping annotated genes, and completely mapping to annotated genes. Novel genes and gene model modifications thus obtained based on gene prediction software and those peptide evidences were checked based on the conservation across *P. xylostella* using protein blast.

MS/MS spectra used for proposing novel ORFs or changes in gene structure were manually validated. RNA-Seq data downloaded from Gene Ontology databases (http://www.geneontology.org) were used to provide insupplemental Data file S1. Fig. 2 and genomic coordinates in the DBM genome database is provided. Proteogenomics of *P. xylostella* Larval Midgut

**RESULTS**

**Overview of MS/MS Data for Multialgorithmic Proteogenomic Analysis**—The goal of this study was to carry out
Fig. 1. **Overview of the proteogenomic strategy used in this study.** Proteins from the brush border membrane vesicles (BBMV) were extracted from the larval midgut of a susceptible *Plutella xylostella* strain DBM1Ac-S (two biological replicates) and from a Cry1Ac-resistant strain of *P. xylostella* near-isogenic to the susceptible DBM1Ac-S strain NIL-R (two biological replicates). In a preliminary experiment, a biological replicate was obtained after mixed the midguts from the two strains. The resulting five samples of purified BBMV proteins were resolved by SDS-PAGE. A representative SDS-PAGE is shown with lane 1 and lane 2 corresponding to BBMV proteins from DBM1Ac-S strain and NIL-R strain, respectively, and lane M showing the molecular weight markers indicated in kDa. The gel was cut into 10 (preliminary experiment) or 20 (four biological replicates) pieces, equal in size each, as shown. After proteolysis, the resulting peptide fractions were analyzed by nanoLC-MS/MS. Then, MS/MS data were searched using iPeak search algorithm against three databases, *i.e.* an annotation protein database (PDB), a six-frame translated protein database (GDB), and an alternative splicing database (ASDB). The decision tree for validating genes, identifying novel genes, or refining gene structure is shown.
Proteogenomics of *P. xylostella* Larval Midgut

| Table I | Overview of the three databases used in *P. xylostella* proteogenomic analysis |
|---------|-----------------------------------------------------------------------------|

| Database                                      | Number of protein sequences |
|-----------------------------------------------|------------------------------|
| Annotation protein database (PDB)             | 18073                        |
| Translated protein database (GBD)             | 5972969                      |
| Alternative splicing database (ASDB)          | 855515                       |

Genomic loci. The distribution of peptides (Fig. 2B) and proteins (Fig. 2C) identified from the various MS/MS search algorithms applied is also shown. Among this peptide pool, we identified a total of 2694 genome search specific peptides (GSSPs) that could not be mapped to PDB but only to the six-frame translated GDP. These peptides mapped to 58,301 loci on the genome, which were considered for further proteogenomic analysis.

Proteins Identified in *P. xylostella* Larval Midgut—A total of 6764 proteins were identified in the larval midgut of *P. xylostella* by the shotgun ESI MS and proteogenomic analysis against DBM-DB.v1.0 genome database. The list of identified proteins with the corresponding peptide coverage is available in supplemental Data file S2. Distribution of length, molecular mass, isoelectric points (pI), and hydrophobicity of the identified proteins was analyzed (Fig. 2D). As expected, there is a clear preference for detection of larger proteins by bottom-up mass spectrometry, which is in accordance with the notion that larger proteins yield more peptides, hence are more likely to be detected. Molecular mass ranged between 2.85 kDa (scaffold_599_4 9677–9706, an uncharacterized protein) and 1140.9 kDa (Px014167.1 a homolog of Titin protein) with 3263 proteins (48.2%). The pI of the proteins ranged between 3.3 (scaffold_129_3 137870–137941, a predicted protein belonging to a family of proteins involved in signal transduction mechanisms) and 13.1 (scaffold_77_1 111208_111309, an uncharacterized protein) with most pIs between 4 and 12.

Distributions of the preference for hydrophilic proteins conform to the law of normal distribution, in agreement with some previous proteomics reports (64).

Confirmation of Annotated Protein-Coding Genes in Genome—The DBM-DB.v1.0 release of the *P. xylostella* automatic genome annotation contains 18,071 predicted protein-coding models. Based on the proteome of the larval midgut studied here, a total of 2156 peptides were mapped to 1896 gene models annotated in DBM-DB.v1.0. These supporting peptides provide evidence for translation of the existing gene models.

Many protein records in the DBM-DB database are derived from high throughput cDNA experiments or computational gene predictions. Identification of peptides from these proteins serve as confirmation that the locus in question is, in fact, a protein-coding gene. Many of the genes in *P. xylostella* are functionally “hypothetical protein” or “putative protein” as these are conserved across some lepidoptera (*B. mori, Danaus plexippus, Papilio polytes* etc.) and other insects (*Tribolium castaneum, Anopheles darlingi, Culex quinquefasciatus, Solenopsis invicta,Apis mellifera*, etc.) (49). We examined all search results that correspond to proteins with annotations related to “hypothetical protein” or “putative protein” that could be considered doubtful because of the lack of experimental evidence. The search results confirmed the existence of many of these proteins. A total of 207 of these proteins are confirmed by a minimum of two spectra from distinct peptides, many of which have never been shown to be translated by any earlier proteomic study. Supplemental Table S1 summarizes the results.

Refining of Current Gene Models by GSSPs—To revise the annotated gene models, we used the Blastn program to find ORFs in the genomic regions where the set of 2694 GSSPs were mapped. Some intragenic loci were considered to be evidence for gene extension beyond the borders of the existing annotated frame, which suggested that the annotated gene model is incorrect. A total of 355 GSSPs clustered into existing genes that extended the existing exon.

It is difficult to determine whether a novel exon is a refinement of an existing gene model or an additional instance of a novel gene model. Choosing an appropriate distance for clustering is the key to better distinguish between these two cases, which is complicated in eukaryotic organisms as exons from a single gene can be separated by large distances. An increase in distance results in raising the likelihood of GSSPs from adjacent genes being clustered together, whereas a decrease in the distance cutoff results in an increase likelihood of single-gene GSSPs being clustered separately. This distance was reported to be quite different in model organisms (65). We choose a distance cutoff of 500 bp, which represents the most percentile of the intron length distribution and the most percentile of the intergenic distance distribution in DBM-DB.v1.0. Using the distance above, a total of 83 GSSPs clustering with existing models were considered as likely pieces of evidence of gene model extensions, which were also proved to be the instances of novel exon.

A total of 128 gene models were refined, which were supported by at least one reliable GSSP. Supplemental Table S2 compiles these results. Of those, 39 instances of refined genes were supported by at least two GSSPs. To provide additional support to our gene refinements, we checked the conservation of these genes based on EST and RNA-Seq experimental data available from *P. xylostella* using the blat algorithm. For 14 gene fixed events, EST/cDNA sequence were found to support the refined cases. We also used blast and genewise to identify homologs of the predicted amino acid sequence in the animal subset of the National Center for Biotechnology Information nonredundant database (version 2013, August 26th). For 103 proteins, homology protein evidence was found and supported the identification for reliable refined gene models. For 33 genes, homology protein or
EST/cDNA evidence were found to support the refined cases with single GSSP, i.e. the one-hit wonder cases when only one significant PSMs was used for identification. Notably, the exon boundary changes that located at 5′ extension of the first exon supported evidences to correct identification of the translational start codon. We found 27 instances of 5′ extension of the first exon supported by GSSP evidence, suggesting novel translational start codons for these genes. For example, we identified four GSSPs and one shared peptide in the gene Px012027.3 (a muscle myosin heavy chain protein) suggesting the presence of an extension of the protein-coding region as shown in Fig. 3A. Furthermore, the refined case was supported by the detection of a similar protein (gi_156538148_ref_XP_001600210.1) and three EST sequences (Unigene17414_All; Unigene23463_All; Unigene11233_All). Also, we identified three GSSPs corresponding to a

![Proteogenomics of P. xylostella Larval Midgut](image)

**Fig. 2. Proteogenomic identification of peptides and proteins.**  
(A) Genomic view of peptides identified. The outside circle represents the genome of DBM that mapped to silkworm genome LGs; the blue circle shows all peptides identified in the study; the red circle highlights peptides that were used for fixing gene annotation; the green circle shows the peptides used for annotating new genes; the orange circle indicates transcripts mapped onto the DBM genome; the inner circle shows the genes validated by RT-PCR; a 100 Kbp bin was set as a window for counting the number of peptides or transcripts in DBM genome. 
(B) Identification of peptides by MS/MS search algorithms.  
(C) Identification of proteins by MS/MS search algorithms.  
(D) Bias analysis of all identified proteins of the Midgut of P. xylostella.
missing or mispredicted exon upstream of the annotated start codon at locus Px004337.1 (Mitochondrial 2-oxodicarboxylate carrier) as indicated in Fig. 3B. Analysis with blastp recovered a homolog in the bumblebee Bombus terrestris, XP_003401887.1. A final example is the gene Px008471.7 specifying the Ran GTPase-activating protein. We identified three GSSPs upstream of the annotated start codon at the same locus (supplemental Fig. S1), the re-annotation being confirmed with the existence of a homologous protein already noted in Harpegnathos saltator, EFN81060.1.

Alternative Splicing—It is difficult to estimate the true extent of alternative splicing, given that the alternative splice forms are often not as highly expressed, and might not be sampled. Normally, evidence for alternative splicing comes from mRNA sequencing projects, which may include prespliced or contaminating sequences. Mass spectrometry data can confirm the presence of specific isoforms in a sample at the protein level. In this study, MS/MS data were searched against the database of exon–exon junction spanning tryptic peptides of hypothetical splice variants generated using all possible forward combinations of exons in a gene. Using our deep proteogenomic sampling we were able to identify a total of 53 events of novel alternative splicing. Their complete list is reported in supplemental Table S3. It is worth to mention that the splice variant described in Fig. 4 is also seen in other Lepidoptera species, which contained two new alternative protein isoforms. The MS/MS spectrum of the peptide QRIIESYVEK, resulting from splicing together of exons 2 and 54 of the Px013984 gene, has been previously described for the Danaus plexippus dynein heavy chain. At the same time, the peptide sequence MGFLTAMR was identified as a result of splicing of exon 4 to exon 94 of this gene. The identification of an exon–exon junction using the peptide LLILGAISLADIQYETMK in the Px001131 gene is shown in supplemental Fig. S2. The two peptide sequences GREQEAVLQR and GSSILLDNSLQR led to the identification of splicing together of exons 8 and 22 and splicing of exon 9 to exon 14 of the Px012768 gene, as shown in the supplemental Fig. S3. Taken
together, these data show that high-resolution mass spectrometry is a unique tool in annotating novel splice variants even in the absence of transcript evidence. 

**Discovery of Novel Genes—** Clustering of experimental peptide sequences based on MS/MS spectra also allows confirming annotated genes that are specific to one or some genomes, thereby reducing annotation error discrepancies. The 1594 GSSPs were further used to explore the novel protein-coding genes in *P. xylostella* genome. Based on these, we identified 1365 novel protein-coding genes, which had not been reported in DBM-DB.v1.0 annotation. All the identified peptides were found matching with blat queries to the cDNA/EST database of *P. xylotella*. To analyze the function and establish whether homologs exist, we took into consideration the surrounding nucleotide sequence and searched against the animal protein sequences in NCBInr. The refined cases that were identified by single GSSP with only one significant PSM without homology protein or EST/cDNA evidence were considered to be unreliable cases, thus, they were removed from the results. Of the remaining novel genes, 75 have EST and similarities support, in addition to experimental peptides; and 356 genes are supported by the peptides and similarities only. A subset of 200 genes have at least two GSSPs supported. After filtering, the full list of 439 novel protein-coding genes is available in supplemental Table S4 with details of the supporting peptide evidence and genome coordinates. Furthermore, the total of 348 instances for novel protein-coding genes were validated by RNA-Seq data downloaded from the diamondback moth genome database (http://iae.fafu.edu.cn/DBM/index.php). As we mapped these novel genes to DBM-DB V 2.0, there were 326 cases that failed to map to the scaffold of the new DBM genome database, which supported our worthy choice that the DBM-DB V 1.0 should be used for the proteogenomic analysis.

For example, we identified pieces of evidence with 16 GSSPs to support a novel protein-coding gene (scaffold_8_5_683226_684035). Although scaffold_8_5_683226_684035 was not previously functionally annotated in *P. xylostella* genome, we believe this protein and the novel on scaffold 8 of DBM-DB, are NADH dehydrogenase [ubiquinone] iron–sulfur protein based on their similarity to the protein encoded by Gene ID: 512937168 in *B. mori* (e value 3E-155). Fig. 5 shows the characteristics of this novel case including the GSSPs used for its identification, the closest homolog in NCBI nr, XP_004933701.1 (e value 3E-155), and the EST/cDNA evidence.

**Validation of Novel Genes Using RT-PCR and cDNA Sequencing—** As described above, our MS/MS data identified 439 novel protein-coding genes in the midgut of *P. xylostella*. We performed an additional level of validation by performing RT-PCR on selected examples of novel genes. Of all the novel protein-coding genes, 246 (56% of 439) have been confirmed by RT-PCR. A table listing the novel protein-coding genes supported by RT-PCR data including primers and production size is available in supplemental Table S4. Fig. 6 shows an example of a novel gene supported by its RT-PCR sequence. In this case, five GSSPs were found to support this gene (scaffold_451_2_172373_173194), which has high similarity to MRPS35 protein of *Brugia malayi*, XP_001896553.1 and is also supported by two EST sequences (Unigene30158_All; Unigene123403_All). The sequence analysis of RT-PCR products revealed that they corresponded to the novel gene models thus confirming the validity of the proteogenomic approach to uncover novel gene models.

**Most Abundant Proteins in Larval Midgut—** The abundance of all the proteins identified were analyzed. The 100 most abundant proteins are listed in supplemental Table S6, including housekeeping proteins, such as ATP synthase subunit, V-type proton ATPase proteins, ribosomal proteins, tubulin, actin, eukaryotic translation initiation factor, elongation factor, and heat shock proteins. Some proteins involved in cytoskeleton can be found, such as myosin heavy chain proteins and...
paramyosin. Besides housekeeping proteins, proteins involved in the transport and metabolisms of fatty acid, proteins and carbohydrates also are abundant. Notably, among these, the lipid- and fatty acid-related enzymes are in remarkable ratio in the midgut of *P. xylostella*. In the crucifer plant family, the fatty acid content is generally high, as many plants are the main oil crops all over the world (66), such as rapes, *Brassica campestris*. As a specific pest of crucifer plants, *P. xylostella* developed their metabolic system for lipid and fatty acid to digest the food crops and absorb energy from crucifer plants. At the same time, some of these enzymes may play some roles to help DBM to avoid the toxicity from glucosinolates produced by its *Brassicaceae* hosts, as there are many kinds of aliphatic, aromatic, and indole glucosinolates (67). The detoxification-related proteins, including cytochrome P450, glutathione S-transferase, and ATP-binding cassette transporter, contribute also for the highly abundant proteins. These proteins play important roles in xenobiotic detoxification (68) and defense against secondary plant compounds (49). Two proteins, namely thymus-specific serine protease (Px002304.1) and serine protease (Px012830.1), were found to be highly abundant in the midgut. Serine proteases are one of the most important groups of the digestive enzymes in the larval gut and account for about 95% of digestive activity (69). These results are in favor of the idea that *P. xylostella* has coevolved with the crucifer plant family in terms of digestion and absorption of nutrition.

**Functional Classification and Pathway Analysis**—GO analysis was carried out to classify the 2298 total identified proteins based on molecular function, biological process, and cellular localization, which was performed according to the Gene Ontology Annotation (http://www.ebi.ac.uk/goa/) (supplemental Data file S3). Fourteen groups of molecular function were delineated (Fig. 7A). The catalytic activity (1352 proteins; 44% of 2039 proteins) and binding activity (1136; 37%) groups were the top two categories. This indicated that catalytic and binding were extremely active in brush border membrane vesicle (BBMV) of *P. xylostella*, which agrees with reports on other Lepidoptera midguts (44, 47).

Proteins identified after proteogenomic analysis were also classified based on their cellular localization (Fig. 7B). A total of 1353 proteins (21%) were localized in the cytoplasm and associated parts, whereas 1030 (16%), 708 (11%), and 747 (11%) proteins were localized in organelle, organelle part, and macromolecular complex, respectively. Particularly, 599 proteins (9%), 405 (6%), and 204 (3%) proteins were localized in membrane, membrane part and membrane-enclosed lumen, respectively. Membrane-associated proteins are very impor-
tant in organism including cellular signaling and recognition. Most of membrane-associated proteins were identified in midgut proteins, indicating that the midgut plays very important role in the life of *P. xylostella*. Although membrane proteins are usually in low abundance and difficult to extract, the extensive nanoLC-MS/MS analysis done in this study on five biological replicates and proteome fractionation (90 fractions) contribute to detect many low-abundant proteins.

Proteins identified were also classified based on biological process (Fig. 7C). The cellular processes (1484 proteins, 18%), the metabolism processes (1320 proteins, 16%) and the single-organism processes (948 proteins, 11%) groups were the most abundant. The developmental process, multi-cellular organismal process, localization, response to stimulus, establishment of localization cellular component organization or biogenesis, and regulation of biological process groups were almost 5% of all the proteins identified. And then few proteins fell into other terms. Less than 0.01% (only one protein) were directly involved in cell killing group. Almost all the physiological and biochemical events occurred in the midgut. These results proved the predominance of the midgut in the life of *P. xylostella*, with metabolic processes associated with the feeding functions such as food digestion and nutrient absorption but also associated with the tissue formation or degeneration. This is different from the report in the midgut proteome of *S. litura* (47), but is in line with the report of the
midgut transcriptome of *P. xylostella* (46). One of the main reasons may be that the more complex biological characters of *P. xylostella* makes the midgut more multifunctional to adapt the environment to evolve to become the most destructive pest all over the world. This result simply proves that the proteogenomic analysis is an unbiased approach, which
make the possibility to detect an extensive catalog of proteins included in all possible biological processes.

To identify the biological pathways that are active in *P. xylostella* midgut, all identified proteins were mapped to the reference canonical pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) (70). In total, 2744 proteins were assigned to 253 KEGG pathways (Fig. 8). Carbohydrate metabolism (592 members), Lipid metabolism (463 members), and Digestive system (427 members) were the top three pathway groups. This proved that the midgut is the primary site of food digestion, nutrient absorption and energy production. In *P. xylostella*, lipid and carbohydrate are the main energy sources. This provides more insights into the coevolutions between the crucifer plant family and *P. xylostella*.

**DISCUSSION**

*P. xylostella* has developed resistance to various synthetic and biological pesticides, and has been an important model system to study the molecular mechanisms underlying the development of insecticide resistance. The major groups of proteins potentially involved in xenobiotic detoxification and the targets of the major classes of synthetic insecticides were identified in the midgut. The proteins potentially involved in insecticide metabolic resistance (summarized in supplemental Table S7) include conventional detoxification enzymes such as cytochrome P450 monoxygenase (CYP), glutathione S-transferase, carboxylesterase, ATP-binding cassette (ABC) transporter, UDP-glycosyltransferases (UGTs), NADH dehydrogenase, catalase, peroxidase, superoxide dismutase, chymotrypsin, proteinase/protease, and trypsin. CYP is a family of oxidase enzymes involved in metabolism of an extremely large number of endogenous and exogenous compounds (71). A total of 62 distinct CYPs were identified in the midgut; it would be interesting to know the functions of these CYPs in the midgut and assessed their dynamics when the animal is subjected to exogenous compounds. Such targeted proteomic approach could be worth study to define the function of each of these numerous isoforms, taking into account as prioritizing criterion their abundances found here. The glutathione S-transferase GSTs are involved in detoxification of xenobiotics. Again, numerous isoforms have been detected. Eighteen proteins of GSTs were found in larval midgut of *P.*
Proteogenomics of *P. xylostella* Larval Midgut

*xylostella*, which included almost all six classes of GSTs in insects (72, 73). The insect carboxylesterase and ATP-binding cassette (ABC) transporter gene families are the main members involved in metabolism and regulated absorption of both insecticides and plant secondary metabolites, respectively (68, 74). Along with glutathione S-transferases and cytochrome P450 monooxygenases (P450s), they represent the so-called performance genes affecting growth and survival of insect larvae on host plants. In this study, carboxylesterase proteins (21 members) and ATP-binding cassette (ABC) transporters (38 members), respectively, were identified. UDP-glycosyltransferases (UGTs), as do other important enzymes, play an important role in the detoxification of xenobiotics and transport of secondary metabolites in insects, as they catalyze the conjugation of a range of small lipophilic compounds with sugars, developing their solubility in water (75). Sixteen UGTs proteins were identified. The enzymes including chymotrypsin, proteinase/protease, and trypsin were associated with the Bt toxicity activity in larval midgut. Although there are few reports on chymotrypsin of *P. xylostella* (46), we pointed at the existence of four distinct chymotrypsins.

Further, receptors potentially associated with the Bt toxicity and resistance in the *P. xylostella* larval midgut were highlighted, including alkaline phosphatase (ALP), cadherin, aminopeptidase N (APN), ABC transporter, and glycosphingolipid. Recently, an ABCC2 gene in *Heliothis virescens* was genetically linked to the Cry1Ac resistance. A loss-of-function mutation in ABCC2 led to the loss of Cry1Ac binding to membrane vesicles, suggesting ABC transporters may play a key role in the mode of action of Bt toxins (76). Moreover, Baxter et al. (42) cloned an ABCC2 gene in *P. xylostella*, and genetically mapped it at a locus controlling the Bt Cry1Ac resistance. In our laboratory, although we found that cadherin is not involved in *P. xylostella* Bt Cry1Ac resistance (52), we revealed that down-regulation of ALP, ABCC2, ABCC3, and a novel ABCC1 (white) genes are all associated with Cry1Ac resistance in *P. xylostella* (77, 78). Acetylcholinesterase (AChE), as a hydrolase, which hydrolyzes the neurotransmitter acetylcholine was the target of organophosphorus insecticide; chloride channel was studied as insecticidal target (79). Ryanodine receptors (RyRs) are the major cellular mediator of calcium-induced calcium release in animal cells. Diamide insecticides control insects by the activation of RyRs that leads to uncontrolled calcium release in muscle (80). In this study, proteins including acetylcholinesterase (one member), chloride channel (three members), and ryanodine receptor (one member), respectively, were identified.

Our extensive proteogenomic study for a complex organism exemplifies the application of high resolution mass spectrometry data for validating known and novel protein-coding genes and correcting wrong structural annotations. Proteogenomics is becoming a popular method for confirming predicted gene structures and accelerating genome curation, and numerous recent applications have been reported for improving human genome annotation (81–83). Regarding nonmodel animals for which no closely related genome is known, its application remains a great challenge (1). In this study, we have endeavored to apply proteogenomics to the midgut of *P. xylostella* using a multialgorithmic pipeline to gain molecular understanding of this animal. Using identified GSSPs, we discovered novel protein-coding genes and corrected wrongly annotated gene models. Furthermore, novel alternative splicings were identified and validated. The DBM draft sequence was only recently published and annotation efforts are still in the early stages. The MS/MS data analyzed in this study largely support the DBM annotations. In-depth analysis identified proteins putatively involved in insecticide resistance. The presence, at high levels, of numerous enzymes of the glycolysis pathway, as well as absorption and transport of fatty acids and lipids, indicate that active metabolic processes of carbohydrates and lipids occurred in the larval midgut of *P. xylostella*. However, as the first tissue-specific proteogenomics analysis of *P. xylostella*, we have provided examples showing the unique role proteogenomics can play in building the most accurate and descriptive structural annotation. Clearly, our results show that there is a need for continued improvements to existing gene models, annotation of missing genes, and corrections to the genome sequence itself to provide *P. xylostella* researchers with the most accurate representation of the DBM genome, transcriptome, and proteome upon which their research relies, and to a wider extent to all those interested in Arthropods as these proteogenomics data can be used to better annotate new genome sequences of this yet molecularly poorly characterized group of animals, especially within the i5k framework. This study provides the fundamental basis for future elucidation of the molecular mechanism involved in insecticide resistance. With the increasing number of genome sequenced organisms, we anticipate that proteogenomics should be a valuable tool and an integral part of future genome sequencing efforts to support novel information to the biomedical community.

This research was supported by the Agricultural Science and Technology Innovation Program (ASTIP), the National Natural Science Foundation of China (30871659), the 863 Program (2012AA101502), the Special Fund for Agro-scientific Research in the Public Interest (201203038) and Beijing Key Laboratory for Pest Control and Sustainable Cultivation of Vegetables.

© This article contains supplemental material.

To whom correspondence should be addressed: Department of Plant Protection, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, No. 12, South Street of Zhong-Guan-Cun, Haidian District, Beijing 100081, China. Tel.: +86-10-82109518; Fax: +86-10-82109518; E-mail: zhangyoujun@caas.cn.

** These authors contributed equally to this article. Authors declare no conflicts of interests exist.

** REFERENCES

1. Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., and Hartmann, E. M. (2014) Nonmodel organisms, a species endangered by
13. Bland, C., Hartmann, E. M., Christie-Oleza, J. A., Fernandez, B., and Jaffe, J., Berg, H., and Church, G. (2004) Proteogenomic mapping as a
14. Rubiano-Labrador, C., Bland, C., Miotello, G., Guerin, P., Pible, O., Baena, S., and Armengaud, J. (2014) Proteogenomic insights into salt tolerance
11. Yates, J. R., Eng, J. K., and McCormack, A. L. (1995) Mining genomes: the proteomics-based confirmation of protein expression and correction of annotation errors in the Brucella abortus genome. BMC Genomics 11, 300
5. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary based confirmation of protein expression and correction of annotation
13, 361–3625
21, Ansong, C., Tolic, N., Purvine, S., Porwollik, S., Jones, M., Yoon, H., Payne, S., Martin, J., Burnet, M., Monroe, M., Venepally, P., Smith, R., Peterson, S., Heffron, F., McClelland, M., and Adkins, J. (2011) Experimental annotation of post-translational features and translated coding regions in the pathogen Salmonella typhimurium. BMC Genomics 12, 433
20. Baudet, M., Ortel, P., Gaillard, J.-C., Fernandez, B., Guerin, P., Enjalbal, C., Subra, G., de Groot, A., Barakat, M., Dedieu, A., and Armengaud, J. (2010) Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of noncanonical translation initiation codons. Mol. Cell. Proteomics 9, 415–426
21. Ansong, C., Tolic, N., Purvine, S., Porwollik, S., Jones, M., Yoon, H., Payne, S., Martin, J., Burnet, M., Monroe, M., Venepally, P., Smith, R., Peterson, S., Heffron, F., McClelland, M., and Adkins, J. (2011) Experimental annotation of post-translational features and translated coding regions in the pathogen Salmonella typhimurium. BMC Genomics 12, 433
20. Baudet, M., Ortel, P., Gaillard, J.-C., Fernandez, B., Guerin, P., Enjalbal, C., Subra, G., de Groot, A., Barakat, M., Dedieu, A., and Armengaud, J. (2010) Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of noncanonical translation initiation codons. Mol. Cell. Proteomics 9, 415–426
21. Ansong, C., Tolic, N., Purvine, S., Porwollik, S., Jones, M., Yoon, H., Payne, S., Martin, J., Burnet, M., Monroe, M., Venepally, P., Smith, R., Peterson, S., Heffron, F., McClelland, M., and Adkins, J. (2011) Experimental annotation of post-translational features and translated coding regions in the pathogen Salmonella typhimurium. BMC Genomics 12, 433
Proteogenomics of \textit{P. xylostella} Larval Midgut

Koh, Y. H. (2011) RNA interference of pheromone biosynthesis-activating neuropeptide receptor suppresses mating behavior by inhibiting sex pheromone production in \textit{Plutella xylostella} (L.). \textit{Insect Biochem. Mol. Biol.} 41, 236–243

40. Atre, J. A., Vandenberg, J. D., and Cantone, F. A. (1999) Pathogenicity of \textit{Paecilomyces fumosoroseus} isolates to diamondback moth, \textit{Plutella xylostella}: correlation with spore size, germination speed, and attachment to cuticle. \textit{J. Invertebr. Pathol.} 73, 322–333

41. Bautista, M. A., Miyata, T., Miura, K., and Tanaka, T. (2009) RNA interference-mediated knockdown of a cytochrome P450, CYP6BG1, from the diamondback moth, \textit{Plutella xylostella}, reduces larval resistance to permethrin. \textit{Insect Biochem. Mol. Biol.} 39, 38–46

42. Baxter, S. W., Badenes-Pérez, F. R., Morrison, A., Vogel, H., Crickmore, N., Kain, W., Wang, S., Heidig, D. G., and Jiggins, C. D. (2011) Parallel evolution of \textit{Bacillus thuringiensis} delta-endotoxins. \textit{Insect Biochem. Mol. Biol.} 41, 2117–2147

43. Pauchet, Y., Munier, I., Neefjes, J. J., and de Bruijn, H. J. (2011) A heterozygous moth genome provides insights into herbivory and defense. \textit{PLoS Genet.} 7, e1002257

44. Pauchet, Y., Wilkinson, P., Vogel, H., Nelson, D. R., Reynolds, S. E., Heckel, D. G., and Fiehn, O. (2010) Protein profiles of the midgut of \textit{Spodoptera litura} larvae at the sixth instar feeding stage by shotgun ESI-MS approach. \textit{Int. J. Biol. Sci.} 6, 236–243

45. Zangerl, A. R., Liao, L. H., Jogesh, T., and Berenbaum, M. R. (2012) Diversity in lepidopteran serine proteases. \textit{Cell. Mol. Biol. Lett.} 17, 61–75

46. Liu, J., Zheng, S., Liu, L., Li, L., and Feng, Q. (2010) Protein profiles of the midgut of \textit{Spodoptera litura} larvae at the sixth instar feeding stage by shotgun ESI-MS approach. \textit{Int. J. Biol. Sci.} 6, 236–243

47. Zangerl, A. R., Liao, L. H., Jogesh, T., and Berenbaum, M. R. (2012) Diversity in lepidopteran serine proteases. \textit{Cell. Mol. Biol. Lett.} 17, 61–75

48. You, M., Yue, Z., He, W., Yang, X., Gie, M., Zhan, D., Baxter, S. W., Badenes-Perez, F. R., Morrison, A., Vogel, H., Crickmore, N., Kain, W., Wang, S., Heidig, D. G., and Jiggins, C. D. (2011) Parallel evolution of \textit{Bacillus thuringiensis} delta-endotoxins. \textit{Insect Biochem. Mol. Biol.} 41, 2117–2147

49. Pauchet, Y., Muller, I., Neefjes, J. J., and de Bruijn, H. J. (2011) A heterozygous moth genome provides insights into herbivory and defense. \textit{PLoS Genet.} 7, e1002257

50. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. \textit{Anal. Chem.} 68, 850–855

51. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., C. J., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. \textit{The Gene Ontology Consortium. Nat. Genet.} 25, 25–29

52. Dorus, S., Busby, S. A., Gerike, U., Shabanowitz, J., Hunt, D. F., and Karr, T. L. (2006) Genomic and functional evolution of the \textit{Drosophila melanogaster} sperm proteome. \textit{Nat. Genet.} 38, 1440–1445

53. Vezzai, A. J., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Oommen, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. \textit{Nat. Biotechnol.} 32, 223–226

54. Tang, W., Yu, L., He, W., Yang, G., Ke, F., Baxter, S. W., You, S., Douglas, C. J., and You, M. (2014) \textit{DBM}-DB: the diamondback moth genome database. \textit{Database bat087}

55. Geer, L. Y., Markey, S. P., Kovalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrom-
naling pathway alters expression of midgut ALP and ABCC genes and causes resistance to Bacillus thuringiensis Cry1Ac toxin in diamondback moth. PLoS Genet. 11, e1005124

78. Guo, Z., Kang, S., Zhu, X., Xia, J., Wu, Q., Wang, S., Xie, W., and Zhang, Y. (2015) Down-regulation of a novel ABC transporter gene (Pxwhite) is associated with Cry1Ac resistance in the diamondback moth, Plutella xylostella (L.). Insect Biochem. Mol. Biol. 59, 30–40

79. Bloomquist, J. R. (2003) Chloride channels as tools for developing selective insecticides. Arch. Insect Biochem. Physiol. 54, 145–156

80. Lahm, G. P., Cordova, D., and Barry, J. D. (2009) New and selective ryanodine receptor activators for insect control. Bioorg. Med. Chem. 17, 4127–4133

81. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J. C., Liebler, D. C., and the, N. C. (2014) Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–387

82. Horvatovich, P., Lundberg, E. K., Chen, Y.-J., Sung, T.-Y., He, F., Nice, E. C., Goode, R. J., Yu, S., Ranganathan, S., and Baker, M. S. (2015) Quest for missing proteins: update 2015 on chromosome-centric human proteome project. J. Proteome Res. 14, 3415–3431

83. Carapito, C., Lane, L., Benama, M., Opsomer, A., Mouton-Barbosa, E., Garrigues, L., Gonzalez de Peredo, A., Burel, A., Bruley, C., Gateau, A., Bouyssié, D., Jaquinod, M., Cianferani, S., Burlet-Schiltz, O., Van Dorselaer, A., Garin, J., and Vandenbrouck, Y. (2015) Computational and mass-spectrometry-based workflow for the discovery and validation of missing human proteins: application to chromosomes 2 and 14. J. Proteome Res. 14, 3621–3634