Transcriptomic and proteomic analysis of putative digestive proteases in the salivary gland and gut of *Empoasca* (Matsumurasca) *onukii* Matsuda

Ensi Shao
Fujian Agriculture and Forestry University https://orcid.org/0000-0001-9825-0857
Yujuan Song
Fujian Agriculture and Forestry University
Yaomin Wang
Fujian Agriculture and Forestry University
Yichen Liao
Fujian Agriculture and Forestry University
Yufei Luo
Fujian Agriculture and Forestry University
Sijun Liu (sliu@iastate.edu)
Iowa State University
Xiong Guan
Fujian Agriculture and Forestry University
Zhipeng Huang (zhipengh6841@163.com)
Fujian Agriculture and Forestry University

Research article

Keywords: tea green leafhopper, RNA-Seq, proteomics, enzymatic activity, salivary gland, gut

Posted Date: August 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-25607/v2

Version of Record: A version of this preprint was published at BMC Genomics on April 15th, 2021. See the published version at https://doi.org/10.1186/s12864-021-07578-2.

Abstract

Background: Infestation by tea green leafhoppers, Empoasca (Matsumurasca) onukii , could cause a series of biochemical changes in tea leaves. As a typical cell-rupture feeder, E. onukii secretes proteases while probing with its stylet into the tender shoots of tea plants (Camellia sinensis). This study identified and analyzed proteases specifically expressed in the salivary gland (SG) and gut of E. onukii through enzymatic activity assays, complemented with an integrated analysis of transcriptome and proteome data.

Results: In total, 129 contigs representing seven types of putative proteases were identified. Transcript abundance of digestive proteases and enzymatic activity assays showed that cathepsin B-like protease, cathepsin L-like protease, and serine proteases (trypsin- and chymotrypsin-like protease) were highly abundant in the gut while moderately abundant in the SG. The abundance pattern of digestive proteases in the SG and gut of E. onukii differed from that of other hemipterans including Nilaparvata lugens , Laodelphax striatellus , Acyrthosiphum pisum , Halyomorpha halys and Nephotettix cincticeps . Phylogenetic analysis showed that aminopeptidase N-like proteins and serine proteases abundant in the SG or gut of hemipterans were distributed to two distinct clusters.

Conclusions: Altogether, this study provide insightful information on the digestive system of E. onukii and observed different patterns of proteases abundant in the SG and gut of E. onukii , in comparison with other five hemipteran species. These results will be beneficial in understanding the interaction between tea plants and E. onukii .

Background

The tea green leafhopper, *Empoasca* (Matsumurasca) *onukii* Matsuda (Hemiptera: Cicadellidae) is an important insect-pest of the tea plant, *Camellia sinensis* (L.) O. Kuntze (Theaceae), in tea-producing countries in Asia [1-3]. *E. onukii* was previously identified as two species, *E. onukii* and *E. vitis*. A recent detailed study of morphometric characteristics has suggested that *E. onukii* and *E. vitis* belong to the same species [3]. *E. onukii* damages tea plants by piercing a needle-like stylet into tea buds, leaves and tender shoots to suck the sap, resulting in a typical syndrome called hopper burn, which dramatically reduces the yield and quality of tea leaves [1]. It has been estimated that the annual economic losses caused by this insect could be up to ~15-50% in China alone [4]. Currently, management of *E. onukii* infestation relies largely on applications of chemical pesticides, leading to potential problems of pesticide resistance and detrimental effects on the environment and human health [5-8]. Biological controls of *E. onukii* have also been reported but with very limited success [9-11]. Knowledge of the biochemistry and physiology of the ingestion system of *E. onukii* may support the innovation of novel strategies for the management of this serious insect-pest.

Phytophagous hemipteran insects adapt piercing-sucking mouthparts for sap-feeding. There are three types of plant sap-feeding behaviors, e.g., phytophagous salivary sheath feeding, osmotic-pump feeding,

and cell-rupture feeding [12, 13]. Studies on stylet activity suggested that *E. onukii* is a typical cell-rupture feeder which secretes enzymes while probing its stylet into plant tissues and then feeds on the mesophyll or stem parenchyma cells [1, 14]. The piercing-sucking feeding requires injecting digestive enzymes from the insect's salivary gland (SG) into plants for extraoral digestion. Hence, a detailed investigation into salivary protein compositions can help to understand the physiology of hemipteran insect digestive systems and potentially lead to develop novel strategies to manage these insect-pests.

Salivary protein profiles have been studied at both transcriptomic and proteomic levels for a wide range of hemipteran insects, including stink bugs, aphids, leafhoppers and planthoppers [15-23]. Salivary proteomic profiling has revealed that saliva from hemipteran insects mainly contains various enzyme families, including oxidoreductases, hydrolases, transferases, Ca^{2+}-binding proteins and proteases/peptidases [15, 18, 20, 24, 25]. However, most of those early studies identified only protease groups, rather than protease species, due to the lack of genomic information.

Transcriptomic analysis coupled with proteomic profiling is an effective strategy for identification of functional proteins in the SG and gut of hemipterans [21, 26]. For example, proteases in the SG and gut of the two devastating pentatomid stink bugs, *Halyomorpha halys* and *Nezara viridula*, were investigated to identify proteases highly expressed in each tissue [21, 26]. Furthermore, mapping of SG transcripts of *H. halys* to the protein profiles of watery saliva, which is the saliva secreted into the host plants, revealed 22 abundant digestive proteases. In addition, the majority of these transcripts of the proteases highly accumulated in the principle salivary gland (PSG) of *H. halys* and *N. vididula*. These results indicated that in both stink bugs, PSGs were the major sources to release proteases into the saliva [21].

Although proteases and nucleases participating in extraoral digestive activities have been previously investigated in the SG and guts of some hemipteran insects, knowledge about composition of proteases in digestive tissues of *E. onukii* is still limited. We previously analyzed the transcriptome sequencing data derived from the gut of *E. onukii*, which identified putative transcripts coding proteins of digestive proteases, detoxification enzymes and immune response factors [27]. In the present study, we have extended our investigation to transcriptome and proteomics of the SG and gut of *E. onukii* in order to investigate tissue-specific expressions of proteases. In addition, we have also compared the abundance of proteases in the SG and gut of several hemipterans to assess common proteases across the species. Identification and analysis of digestive proteases in the SG and gut of *E. onukii* not only provide information for further understanding of the interaction between *E. onukii* and tea plants but also assist in the innovation of novel strategies in the management of this important pest.

Results

Protease activity in the salivary gland and gut of *E. onukii*

Enzymatic activities of leucine aminopeptidase, cathepsin B-, cathepsin L-, trypsin- and chymotrypsin-like protease in the SG and gut were examined, and the results are shown in Fig. 1. The overall enzymatic

activities of digestive proteases were much higher in the gut than in the SG, except for leucine aminopeptidase activities, which showed similar activity units in the SG and gut. Activities of cathepsin B-, cathepsin L-, trypsin- and chymotrypsin-like protease were significantly higher in the gut than in the SG (Fig. 1). Cathepsin B- and L-like protease showed the highest activity in the gut with low to moderate activity was observed in the gut by trypsin-like protease, chymotrypsin-like protease and leucine aminopeptidases (Fig. 1).

Transcriptomic and proteomic analysis of proteins in the SG and gut of *E. onukii*

1) RNA sequence assembly and functional annotations

The obtained sequence reads were assembled either by using tissue specific RNA reads or by using the pooled reads derived from RNA of both the SG and gut. BUSCO analysis indicated that the quality of the assembly was high. Details of sequence assembly are summarized in Additional file 1: Table S1. The generated contigs were annotated by BLASTx against the NCBI (National Center for Biotechnology Information) nr database. Around 42%, 57% and 38% of the contigs assembled from the SG, gut and pooled reads, respectively, hit to proteins (Additional file 1: Table S1). The protein hits for the three sets of contigs were similar, which was reflected in species distribution of the hits (Additional file 2: Fig. S1). The majority of hits (41-43%) were from hemipterans, followed by ~22% hit to proteins of Blattaria (Table 1).

2) Mapping of the assembled genes to the proteomic profiles derived from the SG and gut

To investigate expression of the *E. onukii* genes at the protein level, we used the assembled transcripts as a database and mapped the peptide sequences resulted from proteomic sequencing of SG and gut proteins to the protein sequences translated from the assembled contigs. The peptide mapping results are summarized in Additional file 1: Table S1. In total, 4,457 unique transcripts in the SG and 3,784 transcripts in the gut were mapped by the peptides derived from proteomic sequencing. Numbers of the mapped proteins from the SG, gut and from both tissues are shown in Fig. 2a. The majority of proteomic peptides mapped transcripts were identified in both SG and gut, with 18.6% (945) and 13.7% (654) of transcripts predicted to be specifically expressed in the SG and gut respectively. These results indicated that SG and gut of *E. onukii* generally provide common functions in the digestion system. However, unique proteins identified in the SG and gut reflect that the two tissues also play different biological functions in *E. onukii*. In addition, certain numbers of the transcripts identified from the proteomic profiles were associated with the increase of the FPKM (expected number of Fragments Per Kilobase of transcript sequence per Million base pairs sequenced) value of the transcripts (Fig. 2b), which were similar to the previous observations in the investigation of proteases of stinkbugs [21, 26].

3) Identification of protease genes

We identified 930 unique contigs encoding putative proteases from SG and gut tissues by analyzing the BLAST annotation results. These proteases included aminopeptidase, carboxypeptidase, dipeptidase, aspartic protease, cathepsin B-like protease, cathepsin L-like protease and serine protease (trypsin-,

chymotrypsin- and elastase-like protease). However, only 129 (14%) of the protease proteins were mapped by at least 2 unique peptides derived from proteomic sequencing profiles (Table 2), which included 26 aminopeptidases, 12 carboxypeptidase, 11 dipeptidases, 8 aspartic proteases, 14 cathepsin B-like proteases, 18 cathepsin L-like proteases and 40 serine proteases (Additional file 3). Fifty-two of the putative proteases mapped by proteomic profiles contained a signal peptide, which were potential digestive proteases. The other 77 proteases had either no signal peptides or the signal peptides could not be determined due to lacking the N terminal sequences (Additional file 3). The protein sequences of the possible secreted proteases in the gut or SG of *E. onukii* are provided in Additional file 4.

4) Expression of protease transcripts

To assess overall expression of the transcripts in the SG and gut, the FPKM of each contig was estimated. The FPKM data were converted to log scales, and boxplots showing medians and the full range of the variations of FPKMs are presented in Fig. 3a. In addition, boxplots using the FPKM value of the transcripts encoding proteases are also shown in Fig. 3a. The majority of the transcripts had very low FPKM values, and the median FPKM was only 1.47 for the gut and 2.23 for the SG, respectively (Fig. 3a). On the other hand, the proteases had much higher RNA expression levels. The median FPKM of proteases was 84.43 in the gut, which was 54-fold higher than the median FPKM calculated from the FPKM of all transcripts (Fig. 3a). It is notable that proteases were highly transcribed in the gut of *E. onukii*. Seven out of the top ten most abundant transcripts in the gut were proteases (Additional file 5). Transcripts of protease in the SG were relatively lower, compared with those in the gut of *E. onukii,* but still higher than the average FPKM of the total SG contig set. The median FPKM of proteases in the SG was 8.70, which was around 4-fold higher than that of the total transcripts (Fig. 3a), although no transcripts of protease were among the top 100 most expressed proteins in the SG (Additional file 5).

The FPKM of each identified protease in the SG or gut is listed in Additional file 3 and presented by a heatmap (Additional file 2: Fig. S2). Of all the identified proteases, 107 of them had an FPKM $\geq$ 1,000, or 50 $\leq$ FPKM < 1000. The FPKM values of 49 proteases were higher than 1,000 in the gut, while those of only 10 proteases were above 1,000 in the SG (Additional file 3). Differential analysis by DESeq2 showed that 85 out of 129 putative protease genes had over 2 folds of expression level in the gut than in the SG, while only 13 putative protease genes had more than 2 folds of expression level in the SG than in the gut (Additional file 3). Different types of proteases apparently differentially expressed in the SG or gut. Cathepsin-like proteases and serine proteases were the most abundant proteases among all the proteases analyzed. In addition, the highly transcribed cathepsin B- and cathepsin L-like proteins in the gut also showed relatively higher FPKM values in the SG than the other cathepsin-like proteins (Additional file 3). Serine proteases were also abundant in the gut. Seven out of 40 potential serine proteases had FPKM values over 10,000. Contrary to this, only 2 serine proteases (EMoSerineProtease-24 and -26) showed FPKM values over 1,000 in the SG (Additional file 3). In the aminopeptidase group, EMoAminopeptidase-10 was the only aminopeptidase showing an FPKM over 1,000 in the gut. Aminopeptidases were also highly transcribed in the SG of *E. onukii*. FPKMs of EMoAminopeptidase-1 and -3 in the SG were 710.05 and 755.95, respectively, which showed 3.50 and 6.18 fold changes against

their expression level in the gut (Additional file 3). Compared to other protease groups, carboxypeptidases and dipeptidases were relatively lower in expression. EMoCarboxypeptidase-3 was the most abundant carboxypeptidase in the gut, and EMoCarboxypeptidase-5 and -6 were moderately transcribed (FPKM below 100). To the contrary, EMoCarboxypeptidase-7, -8 and -9 were abundant in the SG but were less transcribed in the gut. In addition, the most abundant dipeptidase was EMoDipeptidase-2 (FPKM=402.60) in the gut.

5) Tissue specific distributions of protease proteins

The number of potential secreted proteases identified from the SG and gut or found in both tissues are shown in Fig. 3b. The majority of the proteases (59%) were distributed in both tissues with 8 and 34 proteases uniquely found in the SG and gut, respectively. Among protease groups, the same numbers of aminopeptidases were found in the SG and gut respectively, while more cathepsin-like proteases and serine proteases were found in the gut tissues than in the SG (Additional file 1: Table S2). These results were consistent with the results of enzymatic activity tests, in which higher cathepsin and serine protease activities were observed in the gut, while not many differences in aminopeptidase activities between the SG and gut were detected. Various numbers of proteases identified in the SG and gut provided explanations of the enzymatic activities. Notably, EMoCathepsin L-16 and EMoSerineProtease-21 were highly expressed in the gut with FPKMs of 44,267 and 3,460, respectively (Additional file 3). However, from the proteomic profiles, these protease proteins were only found in the SG, but not observed in the gut.

Comparison of the top expressed proteases in *E. onukii* and in five other hemipterans

The SG and gut of *E. onukii* expressed almost similar proteases (Table 3), which suggested that the SG and gut of *E. onukii* may play similar roles in food digestion. To determine whether other hemipteran insects have similar protease distributions, we analyzed available RNA-Seq data isolated from SGs or guts of five hemipteran insects, including two rice planthoppers (*N. lugens* and *L. striatellus*), one rice leafhopper (*N. cincticeps*), one stink bug (*H. halys*) and one aphid (*A. pisum*). Information of proteases in *N. lugens*, *L. striatellus*, *A. pisum*, *H. halys* and *N. cincticeps* is shown in Additional file 6. Due to the lack of genomic data for *N. cincticeps*, translated protein sequences of predicted proteases in *N. cincticeps* are shown in Additional file 4. The top ten proteases of each insect based on FPKM values were selected and listed in Table 3. The results showed that aminopeptidase, carboxypeptidase and dipeptidase were among the most expressed proteases in the transcripts of the SG of *N. lugens*, *L. striatellus*, *A. pisum* and *N. cincticeps*. Different from these four insects, cathepsin-like proteases and serine proteases were among the most expressed proteases in the SG of *E. onukii* and *H. halys* (including PSG and accessory SG, ASG). Only one carboxypeptidase was included in the top ten most abundant digestive proteases in the SG of *E. onukii* and the PSG of *H. halys*, respectively. The most abundant transcripts of proteases were serine proteases (trypsin- and chymotrypsin-like) or cysteine proteases (cathepsin B- and L-like) in the guts of these hemipterans (RNA-Seq data of *N. cincticeps* gut is unavailable) including *E. onukii*.

Different from other four hemipterans, cathepsin L-like protease is the most abundant protease family in the gut of *E. onukii* and *H. halys*.

Phylogenetic analysis of potential digestive proteases isolated from *E. onukii* and other hemipteran insects

Phylogenetic trees of aminopeptidases, cathepsin B- and L-like proteins or serine proteases isolated from SGs and guts of six hemipteran insects (*E. onukii*, *A. pisum, H. halys, N. cincticeps, N. lugens* and *L. striatellus*) were constructed. In addition, phylogenetic analyses of the proteases from hemipteran insects and from other insect groups were also performed.

1) Aminopeptidase

There are many types of aminopeptidases [28]. Nine different aminopeptidase groups were identified in the transcriptome of the SG and gut of *A. pisum, H. halys, N. cincticeps*, *N. lugens*, *L. striatellus* and *E. onukii*. The phylogenetic analysis grouped the aminopeptidases into 12 clades (Fig. 4a). Aminopeptidase N (APN) was the largest group and was distributed in four clades (A, B, F and G). The rest of the aminopeptidases were divided into eight clades representing varieties of the aminopeptidase group (Fig. 4a). Insect aminopeptidase N normally contains a gluzincin motif (GXMEN) and a zinc-binding motif ($HEXXHX^{18}E$) [29]. We noticed that these two motifs were only observed in the APNs in groups A and B, but not in groups F and G. Groups F and G were mainly constituted by deduced aminopeptidases of *E. onukii* and *N. cincticeps*, respectively (Fig. 4b). In addition, APNs of group A had generally higher transcriptional levels in the gut than in the SG of the five hemipteran insects. In contrast, group B APNs showed higher expression levels in the SG than in the gut (Fig. 4a). The differentially expressed APNs of groups A and B found in the SG and gut were uniquely clustered when the APNs of more insect orders were introduced for phylogenetic tree construction (Additional file 2: Fig. S3).

2) Cathepsin B- and L-like protein

Cathepsin B- and L-like proteins identified from the SG and gut of *E. onukii* and five other hemipterans were clearly grouped into two distinct clusters, e.g., the cathepsin B group and the cathepsin L group, in the phylogenetic tree (Fig. 5). In the group containing cathepsin L-like proteins, the cathepsin L-like proteases of *E. onukii* clustered together in the same clade except for EMoCathepsin L-1, which is closely related to NcCathepsin L-5 (Fig. 5). The cathepsin B-like proteases, which showed higher transcriptional levels in the gut of *E. onukii* (EMoCathepsin B -3, -5, -6, -7, -9, -10, -11, -12 and -13) were clustered together. On the other hand, EMoCathepsin B-8, the only cathepsin B-like protein of *E. onukii* showing a higher expression level in the SG, was grouped with four other SG enriched cathepsin B-like proteases, e.g., HhCathepsin B-1, ApCathepsin B-3, LsCathepsin B-2, and NlCathepsin B-4 (Fig. 5).

3) Serine protease

The serine proteases (trypsin-like, chymotrypsin-like, elastase-like and other serine proteases) of *E. onukii* and five other hemipteran insects were mainly divided into two major clusters in the phylogenetic tree (Fig. 6). Most of the cluster I serine proteases showed low to moderate transcriptional abundance in the gut, while higher expressions in the gut were observed in most of the cluster II serine proteases. Significantly, a branch of serine proteases in cluster I, including NcSerineProtease-6, EMoSerineProtease-4, -36, ApSerineProtease-1, -2, -6, and multiple other serine proteases, which were highly transcribed in the SG of stink bugs [21, 26], were grouped together (the clade colored red in Fig. 6). The putative serine proteases from *E. onukii* were mostly grouped to cluster II with the exception of EMoSerineProteases-4, -35, -36 and -37 (Fig. 6). In addition to the two major clusters, EMoSerineProtease-35 and seven other serine proteases from *A. pisum, H. halys,* and *L. striatellus* were clustered into two small clades, which were independent from other serine proteases (clusters III and IV). To understand whether the cluster II serine protease was distinct from serine proteases abundant in the SG, venom serine proteases from various insect orders were appended to the phylogenetic analysis. Interestingly, when venom serine proteases of multiple insect orders were included in the phylogenetic analysis, the serine proteases in cluster II, which were highly expressed in the gut of hemipterans, were again grouped into a distinct clade and located at the root of the tree (Additional file 2: Fig. S4).

Discussion

The primary goal for this study was to identify digestive proteases in *E. onukii* by analyzing transcriptomic and proteomic profiles derived from the SG and gut of *E. onukii*. We discovered at least 129 of the protease sequences, 52 of them containing signal peptides, which were putative digestive proteases. These proteases include aminopeptidase, carboxypeptidase, dipeptidase, aspartic protease, cysteine protease (cathepsin B- and L- like), and serine protease (trypsin-, chymotrypsin- and elastase-like protease). Furthermore, cysteine proteases and serine proteases showed significantly higher activity in the gut than in the SG, while almost equal enzymatic activity of aminopeptidase were determined in both tissues (Fig. 1). erine proteases (trypsin- and chymotrypsin-like proteases) and cysteine proteases (cathepsin B- and cathepsin L-like) were identified as the major digestive enzymes in the midguts of phloem sap sucking hemipterans although the phloem sap is nutritionally inadequate [21-23, 26, 30, 31]. Our results are consistent to the previous investigations on digestive proteases of hemipteran insects. It is interesting to observe that both cysteine and serine proteases are accumulated in guts of hemipteran insects. These two groups of proteases require different pH for their activities. The optimum pH for cysteine proteases is in lower pH (pH<7.0), while serine proteases are alkaline proteases [31, 32]. One possible explanation for the existing of both cysteine and serine proteases in the gut of *E. onukii* is that pH of gut environment of hemipterans is varied in their digestive tracts. Pondus hydrogenii variations in midgut of *Aphis gossypii* was observed with changes from acidic (stomach) to lower alkaline (central and posterior midgut) [33].

In the digestive system of insects, digestive proteases are produced mainly by the midgut epithelial cells and secreted into the lumen where the food bolus passes through [34, 35]. However, transcription or translation of digestive proteins could be regulated by insect feeding behaviors, developmental stages,

and food sources. For instance, enzymic activities in salivary gland and midgut of the mirid bug *Apolygus luncorum* were regulated by sex, age and food resources (plant or animal sources) [22]. In addition, symbiotic bacteria could also regulate digestive protease expressions [30]. Differential expressions of various proteases in the SG and gut were also observed in pentatomid stink bugs, *N. viridula* [26] and *H. halys* [21]. Comparing to these two species of stink bug, *E. onukii* appears to have a very different regulation of protease expression in its digestive system. In general, more proteases were identified in the gut from both transcriptome and proteome data in *E. onukii*. Of the total of 129 protease proteins mapped by the peptide profiles, only 16 proteases were unique in the SG, and 2 (EMoCathepsin L-16 and EMoSerine protease-21) of them had a much higher transcription level (log_2 fold changes, $log_2FC > 6.5$) in the gut (Additional file 3). These results suggested that some proteases, especially cathepsin-like proteases and serine proteases, might be transferred from the gut to the SG. Cathepsin-like proteases could possibly be delivered from gut to saliva were implied from a study on *H. halys*. Two highly expressed cathepsin-like proteases in the gut of *H. halys* were detected in water saliva [21]. An earlier investigation on *Tuberaphis styraci* aphid suggested that the 1st-instar aphid soldiers could inject midgut expressed cathepsin B-like proteases through their stylets into enemies [36]. Our observation seems to support that hemipteran insects may utilize midgut expressed proteases for extraoral digestion although the actual mechanism remains to be investigated. Studies of protein composition of the tea leaf showed that 15% of the dry weight of tea leaves was proteins [37]. Therefore, a high abundance of digestive enzymes in the gut of *E. onukii* may be functional in the digestion of ingested proteins from the mesophyll and stem parenchyma cells of tea plants.

Recently, transcriptomic and proteomic analysis of *H. halys* identified various proteases form saliva of *H. halys*. The vast majority of the proteases found from WS are serine proteases which are highly expressed in PSG of *H. halys*, but also include cathepsin-like proteases and peptidases [21]. Significantly, almost all the proteases recovered from WS are proteases with the highest transcript level, suggesting the proteases with higher transcriptional level are more likely to be functional in food digestion. Hence, we analyzed transcriptomes of five hemipteran insects including two planthoppers (*N. lugens* and *L. striatellus*), one leafhopper (*N. cincticeps*), one aphid (*A. pisum*) and one stink bug (*H. halys*). The top ten most highly expressed proteases in the SG and gut of *E. onukii* were compared to those in the five hemipterans (Table 3). As expected, the top 10 proteases in PSG of *H. halys* are 9 serine proteases and one carboxypeptidase, which were previously found in WS of *H. halys* [21]. These results further prove that proteases with high transcriptional level in digestion related organs are likely to be involved in food digestion. It is clear that the top ten most highly expressed protease transcripts (mainly were cathepsin-like and serine proteases) in the SG and gut of *E. onukii* are very similar, indicating that both the SG and gut of *E. onukii* play important roles in food digestion. The SGs and guts of the three rice feeders (*N. lungens*, *L. striatellus* and *N. cincticeps*) present similar protease compositions (Table 3). In the gut of three rice feeders and one aphid, cathepsin-like proteases or serine proteases were the most abundant proteases. In addition to cathepsin-like and serine proteases, aminopeptidases were also included in the top expressed proteases in the SG of these four hemipterans. It is not clear whether the aminopeptidases were released into the saliva. An observation was made that no aminopeptidase was identified in the saliva of *H. halys*,

although low to moderate expression of aminopeptidases was found in the SG of *H. halys* and *N. viridula* [21]. In present study, no aminopeptidase was in the top ten lists of PSG and ASG of *H. halys* (Table 3). It is believed that *E. onukii* is a typical cell-rupture feeder [1], and *H. halys* was reported to have both cell-rupture and sheath forming feeding behaviors [38], while other hemipterans investigated in this report are sheath forming feeders [12, 39-42]. Cell-rupture feeders feed intracellularly on the mesophyll and parenchyma cells by rapidly moving their stylets and continually secreting saliva to digest plant tissues *in vitro*, followed by the sucking of processed soup through the stylets [14]. Hence, the ingested plant juices containing rich protein components are required for further processing in the alimentary tract. This predicted feeding behavior might explain why the SG and gut of *E. onukii* contain similar digestive proteases. On the other hand, the sheath forming hemipterans generally suck sap from vascular tissues (phloem and xylem) [14, 43]. Hence, the nutritional components of sap ingested from vascular tissues are significantly different from the juices of the broken mesophyll and parenchymal cells in non-vascular tissues. The SGs of the three rice feeders (*N. lungens*, *L. striatellus* and *N. cincticeps*) present similar protease compositions (Table 3). Consequently, different feeding behaviors and food sources may play important roles in impacting protease compositions in the SG of hemipterans.

Phylogenetic analysis of aminopeptidases from six hemipteran insects showed that the APNs of hemipterans, which contain a gluzincin motif and a zinc-binding motif, were divided into two distinct clades; one clade (A) had a higher transcript level in the gut, while the APNs of the other clade (B) were more abundant in the SG (Fig. 4). In the phylogenetic tree constructed with the APNs from multiple insect orders, hemipteran APNs in clades A and B of the phylogenetic tree in Fig. 4 were again separated respectively into two clades (Additional file 2: Fig. S3). Notably, APNs of other insect families were also divided into two clusters (Additional file 2: Fig. S3). Hence, APNs in insect digestive tissues may be of two types, one mainly expressed in the SG and the other highly expressed in the gut. Further study is needed to clarify the current observations.

Similar to the clusters of hemipteran APNs, the majority of serine proteases from six hemipteran insect species were divided into two distinct clusters from the root of the phylogenetic tree (Fig. 6). Based on the transcriptomic level, some serine proteases in cluster I (the red clade) are likely to be the venom proteases that were highly transcribed in the SG, while the majority of cluster II serine proteases were abundant in the gut. To further verify the isolation of serine proteases in cluster II from other serine proteases, venom serine proteases from hemipterans and other insect orders were supplemented to the phylogenetic analysis. Interestingly, with the addition of venom serine proteases from other insect orders, the cluster II hemipteran serine proteases formed a unique branch, which was distinct from all venom serine proteases from multiple orders of insect include Hemiptera (Additional file 2: Fig. S4). These results indicated that evolutionary cases by dividing serine proteases into a gut abundant group and a venom group happened prior to the emergence of Hemiptera.

Conclusions

In summary, we found that both the SG and gut of *E. onukii* express similar groups of proteases, and most of the proteases were highly transcribed in the gut. We also compared the most expressed proteases in the SG and gut of *E. onukii* with five other hemipteran insects and observed that the five analyzed hemipterans had different protease distributions in the SG and gut compared to those of *E. onukii*. The variation in the proteases expressing in the SG and gut could be associated with their feeding behaviors and food sources. Phylogenetic analysis suggests that the proteases in favor of expression in the SG or gut were separated in evolutionary directions. Tea cultivars resistant to the infestation of *E. onukii* were previously reported [11, 44]. Proteases inhibitors are potential biocontrol agents for insect-pests [45-47]. The information presented in this work enriches our understanding of digestive systems of *E. onukii* and could be targets for development of proteases inhibitors and other biocontrol agents to manage this important tea insect-pest. This work could also provide a knowledge base for exploration of the interaction mechanisms between tea plants and *E. onukii.*

Methods

Insects, tissue collections and sample preparations

E. onukii used in this research was originally collected from a tea field located at the Fujian Agriculture and Forestry University, Fuzhou, Fujian Province, China, and has been maintained in laboratory conditions for about 2 years. The *E. onukii* colony was raised on shoots of tea and was bred in water and kept at 28℃ with a photoperiod of 14:10 (light:dark) h in an insectary. The 3rd instar nymphs of *E. onukii* were collected from water planted fresh tea shoots and used for isolation of guts and salivary glands. To dissect the gut and salivary gland, the insects were immobilized on ice for several minutes before they were dissected. SG isolation was performed under a stereomicroscope (VHX-5000, KEYENCE). The morphology of the SG is illustrated in Fig. 7. Methods used for the gut dissections have been previously described [27].

To determine enzyme activities, six hundred salivary glands or guts were homogenized in 600 μL PBS (137 mM NaCl, 2.68 mM KCl, 8.1 mM Na_2HPO_4, 1.47 mM KH_2PO_4, pH 7.4) in a 1.5 mL microcentrifuge tube. Preparation of the homogenized samples was replicated three times, and the protein concentrations were normalized by results of the Bradford method [48]. For RNA extractions, about 1,500 SGs and 1,000 guts were isolated, and the tissues were quickly washed in a diethylpyrocarbonate (DEPC)-treated PBS solution after dissection. The tissues were kept in RNAhold (TransGen Biotech, China). For proteomic sequencing, 2,500 SGs or 1,500 guts were quickly washed and resuspended in PBS. The tissues were then immediately frozen with liquid nitrogen and kept at -80℃. Tissue collections for RNA sequencing and proteomic analysis were repeated three times.

Activity assays of digestive proteases

For enzyme activity assays, collected SG or gut tissues were homogenized by a mortar homogenizer. Leucine aminopeptidase activity was measured using leucine p-nitroanilide (LpNA) (Sigma, US) as a

chromogenic substrate [49]. Tissue homogenates were diluted to 100 μg/mL by 0.1 M Tris-HCl (pH 8.6). Eighteen micrograms of each sample were mixed with 1 mM prewarmed (30℃) substrate buffer (0.1 M Tris-HCl, 1 mM LpNA, pH 8.6). The enzymatic reaction was monitored for optical absorbance increase at 405 nm at 30℃ in a Microplate Reader (SPARK 10M, Tecan) for 10 min (observation was carried out every 10 s, and 60 cycles were conducted for each reaction). Trypsin activity was determined by mixing 5 μg protein in 3 mL of 1 mM Na-benzoyl-L-arginine p-nitroanilide (BApNA，Sigma, US) in 50 mM Tris-HCl (pH 8.0). The enzymatic reaction at 28℃ was monitored in a UV-VIS spectrophotometer (TU-l950, PERSEE) by recording the optical absorbance at 405 nm for 40 min (observation was carried out every 30 s, and 80 cycles were conducted for each reaction). Activities of cathepsin L- and B-like protease were assayed by using carbobenzyloxy-Phe-Arg-(7-amino-4-methyl-coumarin) (Z-FR-AMC, Santa Cruz Biotechnology, US) as a fluorescent substrate. Briefly, 5 μg of the protein was suspended in the activation buffer (10 mM Tris-HCl, pH 7.5), at a final volume of 60 μL including Z-FR-AMC (0.1 mM). The enzymatic reaction was monitored at 30℃ in a Microplate Reader (SPARK 10M, Tecan) with an optical absorbance increase excitation wavelength at 380 nm and an emission wavelength at 460 nm for 40 min (observation was carried out every 30 s, and 80 cycles were conducted for each reaction). One unit of enzyme (U) is defined as the amount that hydrolyzes 1 μmol of substrate per minute. Mean enzyme activity were calculated from three readings for one replication. The activity assay of chymotrypsin was carried out by use of a chymotrypsin activity assay kit (A080-3-1, Nanjing Jiancheng Bioengineering Institute, China) according to the manufacturer's instructions. All enzymatic reactions were replicated at least three times for the statistical analysis, using Student's *t*-test by Prism (version 8.2.0).

RNA extraction, library preparation, Illumina sequencing and *de novo* assembly

Collected tissues were homogenized moderately with a hand homogenizer. Total RNA was extracted by use of HP Total RNA Kit (Omega, USA) according to the manufacturer's protocol. Quantity of extracted RNA was confirmed with Nanodrop (Bio-Rad, USA) and quality of RNA was verified by Agilent 2100 (Agilent, Germany) and electrophoresis gel analysis. About 1.5 μg of total RNA extracted from each sample was used for the generation of sequencing libraries. Sequencing libraries were constructed by the use of NEBNext® ULtra^{TM} RNA Library Prep Kit for Illumina® (NEB, USA) following the manufacturer's recommendations. The qualified cDNA library was clustered through the Illumina cBot system and sequenced on an Illumina HiSeq 2500 platform to generate 150 nt pair-ended reads (Novogene Bioinformatics Technology Co., Ltd., Beijing, China). The original image data were processed with Illumina GA Pipeline v1.3 to clean reads, followed by the removal of adapter sequences, empty reads and low-quality reads. The clean reads obtained from SG or gut samples were uploaded to NCBI SRA databse (BioProject ID: PRJNA606974) and assembled by Trinity Assembly (version 2.8.5) [50] with the default parameter setting. The assembly was also conducted with all reads pooled together. Tissue-specific RNA sequencing reads from SGs or guts of *Nilaparvata lugens* (SRR5149721 and SRR8189329), *Laodelphax striatellus* (SRR1617628 and SRR1617623), *Nephotettix cincticeps* (SRR018462), *Acyrthosiphum pisum* (SRR7037541 and SRR7037537) and *Halyomorpha halys* (SRX6717290, SRX6717291, SRX6717292) were downloaded from the NCBI Sequence Read Archive (SRA) database and separately assembled by

Trinity as described above. To assess the completeness of the assembled data, the assembled transcripts were analyzed using the arthropod gene sets in the BUSCO database as references by BUSCO program (Benchmarking Universal Single-Copy Orthologs) [51].

Transcriptome analysis

Functional annotation of the *de novo* assembled contigs was conducted by the BLASTx search against the NCBI nr (non-redundant) and the Swiss-Prot databases with an E-value cutoff of 10^{-5}. The protein sequences of the contigs with positive hits were translated for further analysis. Potential coding sequences (CDS) of the contigs without hits by BLAST annotation were predicted by ESTScan (version 3.0.3). Relative expression of the transcripts in pooled assembled *E. onukii* datasets was estimated using the RSEM software (version 1.3.1) [52]. Reads mapping and FPKM value were performed by Bowtie. FPKM of protease genes were divided by that of reference gene in either SG or gut for normalization. Reference genes from each insect specie were selected according to previous studies [53-56]. Information of selected reference genes were shown in Additional file 6. Relative expression of protease genes in the SG and gut of *N. lugens*, *L. striatellus*, *A. pisum*, and *H. halys* were presented by calculation of $\log_2$FC of FPKM value (gut VS SG). Median values of the FPKM were calculated and plotted by boxplot functions using a Python script. Heatmaps of FPKM were generated using TBtools software (version 0.6652) [57].

Extraction of total protein and LC-MS/MS analysis

Tissue samples were thoroughly milled in liquid nitrogen by a mortar homogenizer. The milled powders were mixed with a lysis buffer (50 mM Tris-HCl, pH 8.0, 8 M urea and 0.2% SDS) and incubated with ultrasonication on ice for 5 min followed by centrifugation at 12,000 g for 15 min at 4℃. Samples were mixed with 2 mM dithiothreitol and incubated at 56℃ for 1 h, followed by adding sufficient iodoacetic acid and incubating for 1 h at room temperature in darkness. The samples were then mixed with 4 volumes of cold acetone, vortexed and placed at -20℃ overnight, followed by centrifugation at 12,000 g for 15 min at 4℃. The supernatants were discarded, and the pellets were washed twice with cold acetone. The washed pellets were dissolved in dissolution buffer (0.1 M triethylammonium bicarbonate, pH 8.5, 8 M urea), and the protein concentration was determined by the Bradford assay. Protein samples were sent to Novogene Co. Ltd (China) for further preparation followed by liquid chromatography (LC) - electrospray ionization (ESI) tandem mass spectrometry (MS/MS) analysis by Novogene Co. Ltd (China).

Identification of protease proteins from transcriptomic and proteomic data

Putative protease proteins were identified by searching for the positive hits of proteases from BLAST results. The assembled transcripts annotated as aminopeptidase, aspartic protease (cathepsin D), carboxypeptidase, dipeptidase, cathepsin B, cathepsin L, trypsin, chymotrypsin, elastase, cysteine protease and serine protease were identified as putative proteases. The putative protease fragments (>100 aa) obtained from contig sets of the gut, SG and pooled reads were further examined by BLASTp to sort for duplicate protease sequences. Sequences that hit the same accession numbers were manually checked by alignment of these sequences using MEGAX (version 10.1.7) built-in MUSCLE and ClustalW

programs to determine whether they were derived from the same genes. A similar process was also used for identified protease proteins from transcriptomes of *N. lugens, L. striatellus, N. cincticeps* and *A. pisum* species. Signal peptidases of 5' end completed proteins were predicted by SignalP-5.0 server [58].

To identify protease protein from the proteomic data, putative amino acid sequences of proteins encoded by the contigs assembled from the gut, SG and pooled reads were translated using TransDecoder (v5.5.0). (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875132/). The protein sequences were mapped to the peptide libraries resulted by LC-MS/MS analysis using Discoverer 2.2 software (Thermo Fisher Scientific, Bremen, Germany). A protein sequence mapped by at least two different peptides was selected. Selected proteins predicted to contain a signal peptide at the N terminus were determined as putative digestive proteases.

Phylogenetic analysis

Aminopeptidase, cysteine proteases (cathepsin B- and cathepsin L-like protease) and serine proteases (trypsin-, chymotrypsin- and elastase-like protease) discovered from *A. pisum, N. lugens, L. striatellus, N. cincticeps, H. halys* [21] and *E. onukii* were included in the phylogenetic analysis. Information about proteins used in the phylogenetic analysis is shown in Additional file 3 and 6. For the phylogenetic analysis of aminopeptidase- or cathepsin-like proteins from multiple orders of insects, reference proteins were randomly selected from the BLASTp results using EMoAminopeptidases and EMoCathepsins, respectively, as queries against the NCBI nr database. For the phylogenetic analysis of serine proteases from multiple insect orders, reference proteins were randomly selected from previously reported insect venom serine proteases [21]. Plantae aminopeptidase N-like proteins, hemipteran aspartic proteases and cathepsin-like proteases of *H. halys* were included in the phylogenetic tree of aminopeptidase N, cathepsin-like protease and serine protease to form an outgroup. Protein sequences of each protease group were aligned in batches with MAFFT [59]. Aligned sequences were then used for phylogenetic analysis by the maximum likelihood (ML) method. ModelFinder [60] implemented in IQ-TREE [61] was used to choose the best partitioning scheme and models. ML analysis was performed using IQ-TREE with 10,000 ultrafast bootstraps [62]. Constructed trees were uploaded to Interactive Tree of Life (http://itol.embl.de) for visualization and annotations.

Abbreviations

SG: Salivary gland

PSG: principle salivary gland

FPKM: Reads per kilo base per million mapped reads

SRA: Sequence read archives

PBS: Phosphatic buffer solution

RNA: Ribonucleic acid

cDNA: Complementary DNA

NCBI: National Center for Biotechnology Information

Log_2FC: Log_2 fold changes

CDS: Sequence coding for amino acids in protein

DEPC: Diethylpyrocarbonate

LpNA: leucine p-nitroanilide

BApNA: Na-benzoyl-L-arginine p-nitroanilide

Z-FR-AMC: Carbobenzyloxy-Phe-Arg-(7-amino-4-methyl-coumarin)

ANOVA: One-way analysis of variance

LC-MS/MS: Liquid chromatography - electrospray ionization tandem mass spectrometry

EMo: *Empoasca onukii*

Nl: *Nilaparvata lugens*

Ls: *Laodelphax striatellus*

Nc: *Nephotettix cincticeps*

Ap: *Acyrthosiphum pisum*

Hh: *Halyomorpha halys*

Declarations

Ethics approval and consent to participate

The insects used for the experiments are not considered an endangered or protected species, and their collection and breeding are legal in China.

Consent for publication

Not applicable.

Availability of data and materials

All raw sequences of *E. onukii* were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (BioProject: PRJNA606974). Raw sequencing reads from SGs or guts of *Nilaparvata lugens* (SRR5149721 and SRR8189329), *Laodelphax striatellus* (SRR1617628 and SRR1617623), *Nephotettix cincticeps* (SRR018462), *Acyrthosiphum pisum* (SRR7037541 and SRR7037537) and *Halyomorpha halys* (SRX6717290, SRX6717291, SRX6717292) were downloaded from the NCBI Sequence Read Archive database.

Competing interests

The authors declare that they have no competing interests. The authors of this article are faculties of Fujian Agriculture and Forestry University or Iowa State University. None has any involvement with any organization that could be interpreted as leading to a particular bias with respect to the subject of this study.

Funding

This study was financed, by grants from the National Key R&D Program of China (2017YFE0121700) to XG and National Natural Science Foundation of China (31772539) to ZH. ES was supported by the FAFU Science Fund for Distinguished Young Scholars (XJQ201819), Fujian Agriculture and Forestry University Construction Project for Technological Innovation and Service System of Tea Industry Chain (K1520007A03) and Science and Technology Innovation Fund of FAFU (CXZX2018040).

Authors' contributions

ES conceived and designed research, conducted experiments, analyzed data and wrote the manuscript. YS bred experimental insects, prepared samples, and assayed proteases activity. YW, YL and YL bred experimental insects and prepared samples. SL conceived and designed research, analyzed data, revised the manuscript. XG supported the study and wrote the manuscript. ZH analyzed data, supported the study and wrote the manuscript. All authors read and approved the submitted version.

Acknowledgements

We thank Connie Allison for critical reading and editing of the manuscript.

References

1. Jin S, Chen ZM, Backus EA, Sun XL, Xiao B: **Characterization of EPG waveforms for the tea green leafhopper, Empoasca vitis Göthe (Hemiptera: Cicadellidae), on tea plants and their correlation with stylet activities**. *Journal of insect physiology* 2012, **58**(9):1235-1244.
2. Peng P, Tang M, Hou Y-j, Lin Q, Huang S-j, Deng M, Hu X, Zhang Y: **Study on the Effect and Characters of Yellow Sticky Trap Sticking Aleurocanthus spiniferus and Empoasca vitis Gothe in Tea Garden**. *Southwest China Journal of Agricultural Sciences* 2010, **1**:020.

3. Qin D, Zhang L, Xiao Q, Dietrich C, Matsumura M: **Clarification of the identity of the tea green leafhopper based on morphological comparison between Chinese and Japanese specimens**. *PloS one* 2015, **10**(9):e0139202.
4. Zhang Z, Luo Z, Gao Y, Bian L, Sun X, Chen Z: **Volatiles from non-host aromatic plants repel tea green leafhopper Empoasca vitis**. *Entomologia Experimentalis et Applicata* 2014, **153**(2):156-169.
5. Wei Q, Yu H-Y, Niu C-D, Yao R, Wu S-F, Chen Z, Gao C-F: **Comparison of Insecticide Susceptibilities of Empoasca vitis (Hemiptera: Cicadellidae) from Three Main Tea-Growing Regions in China**. *Journal of economic entomology* 2015, **108**(3):1251-1259.
6. Guo H: **Research progress of tea tree major pest, Empoasca vitis**. *Jiangsu Agric Sci* 2011, **1**:132-134.
7. Wei Q, Mu X-C, Yu H-Y, Niu C-D, Wang L-X, Zheng C, Chen Z, Gao C-F: **Susceptibility of Empoasca vitis (Hemiptera: Cicadellidae) populations from the main tea-growing regions of China to thirteen insecticides**. *Crop Protection* 2017, **96**:204-210.
8. Bian L, Sun XL, Luo ZX, Zhang ZQ, Chen ZM: **Design and selection of trap color for capture of the tea leafhopper, E mpoasca vitis, by orthogonal optimization**. *Entomologia Experimentalis et Applicata* 2014, **151**(3):247-258.
9. Hazarika LK, Bhuyan M, Hazarika BN: **Insect pests of tea and their management**. *Annual review of entomology* 2009, **54**:267-284.
10. Ye G-Y, Xiao Q, Chen M, Chen X-x, Yuan Z-j, Stanley DW, Hu C: **Tea: biological control of insect and mite pests in China**. *Biological Control* 2014, **68**:73-91.
11. Yorozuya H: **Analysis of tea plant resistance to tea green leafhopper, E mpoasca onukii, by detecting stylet-probing behavior with DC electropenetrography**. *Entomologia Experimentalis et Applicata* 2017, **165**(1):62-69.
12. Sharma A, Khan A, Subrahmanyam S, Raman A, Taylor G, Fletcher M: **Salivary proteins of plant-feeding hemipteroids–implication in phytophagy**. *Bulletin of Entomological Research* 2014, **104**(2):117-136.
13. Miles PW: **The saliva of Hemiptera**. In: *Advances in insect physiology.* vol. 9: Elsevier; 1972: 183-255.
14. Backus EA: **Sensory systems and behaviours which mediate hemipteran plant-feeding: a taxonomic overview**. *Journal of Insect Physiology* 1988, **34**(3):151-165.
15. Hattori M, Komatsu S, Noda H, Matsumoto Y: **Proteome analysis of watery saliva secreted by green rice leafhopper, Nephotettix cincticeps**. *PloS one* 2015, **10**(4):e0123671.
16. Matsumoto Y, Suetsugu Y, Nakamura M, Hattori M: **Transcriptome analysis of the salivary glands of Nephotettix cincticeps (Uhler)**. *Journal of insect physiology* 2014, **71**:170-176.
17. Huang H-J, Liu C-W, Huang X-H, Zhou X, Zhuo J-C, Zhang C-X, Bao Y-Y: **Screening and functional analyses of Nilaparvata lugens salivary proteome**. *Journal of proteome research* 2016, **15**(6):1883-1896.
18. Harmel N, Letocart E, Cherqui A, Giordanengo P, Mazzucchelli G, Guillonneau F, De Pauw E, Haubruge E, Francis F: **Identification of aphid salivary proteins: a proteomic investigation of Myzus persicae**.

Insect molecular biology 2008, **17**(2):165-174.

19. Ji R, Yu H, Fu Q, Chen H, Ye W, Li S, Lou Y: **Comparative transcriptome analysis of salivary glands of two populations of rice brown planthopper, Nilaparvata lugens, that differ in virulence**. *PloS one* 2013, **8**(11):e79612.
20. Liu X, Zhou H, Zhao J, Hua H, He Y: **Identification of the secreted watery saliva proteins of the rice brown planthopper, Nilaparvata lugens (Stål) by transcriptome and Shotgun LC–MS/MS approach**. *Journal of insect physiology* 2016, **89**:60-69.
21. Liu S, Bonning BC: **The Principal Salivary Gland is the Primary Source of Digestive Enzymes in the Saliva of the Brown Marmorated Stink Bug, Halyomorpha halys**. *Frontiers in physiology* 2019, **10**:1255.
22. Li W, Zhao X, Yuan W, Wu K: **Activities of digestive enzymes in the omnivorous pest Apolygus lucorum (Hemiptera: Miridae)**. *Journal of economic entomology* 2017, **110**(1):101-110.
23. Lomate PR, Bonning BC: **Distinct properties of proteases and nucleases in the gut, salivary gland and saliva of southern green stink bug, Nezara viridula**. *Scientific reports* 2016, **6**:27587.
24. Carolan JC, Fitzroy CI, Ashton PD, Douglas AE, Wilkinson TL: **The secreted salivary proteome of the pea aphid Acyrthosiphon pisum characterised by mass spectrometry**. *Proteomics* 2009, **9**(9):2457-2467.
25. Cooper WR, Dillwith JW, Puterka GJ: **Comparisons of salivary proteins from five aphid (Hemiptera: Aphididae) species**. *Environmental Entomology* 2011, **40**(1):151-156.
26. Liu S, Lomate PR, Bonning BC: **Tissue-specific transcription of proteases and nucleases across the accessory salivary gland, principal salivary gland and gut of Nezara viridula**. *Insect biochemistry and molecular biology* 2018, **103**:36-45.
27. Shao E-s, Lin G-f, Liu S, Ma X-l, Chen M-f, Lin L, Wu S-q, Sha L, Liu Z-x, Hu X-h: **Identification of transcripts involved in digestion, detoxification and immune response from transcriptome of Empoasca vitis (Hemiptera: Cicadellidae) nymphs**. *Genomics* 2017, **109**(1):58-66.
28. Taylor A: **Aminopeptidases: structure and function**. *The FASEB Journal* 1993, **7**(2):290-298.
29. Adang MJ: **Insect aminopeptidase N**. In: *Handbook of proteolytic enzymes.* Elsevier; 2004: 296-299.
30. Akman Gündüz E, Douglas A: **Symbiotic bacteria enable insect to use a nutritionally inadequate diet**. *Proceedings of the Royal Society B: Biological Sciences* 2008, **276**(1658):987-991.
31. Foissac X, Edwards M, Du J, Gatehouse A, Gatehouse J: **Putative protein digestion in a sap-sucking homopteran plant pest (rice brown plant hopper; Nilaparvata lugens: Delphacidae)–identification of trypsin-like and cathepsin B-like proteases**. *Insect Biochemistry and Molecular Biology* 2002, **32**(9):967-978.
32. Wang P, Zhao J-Z, Rodrigo-Simón A, Kain W, Janmaat AF, Shelton AM, Ferré J, Myers J: **Mechanism of resistance to Bacillus thuringiensis toxin Cry1Ac in a greenhouse population of the cabbage looper, *Trichoplusia ni***. *Applied and environmental microbiology* 2007, **73**(4):1199-1207.

33. Deraison C, Darboux I, Duportets L, Gorojankina T, Rahbé Y, Jouanin L: **Cloning and characterization of a gut-specific cathepsin L from the aphid *Aphis gossypii***. *Insect molecular biology* 2004, **13**(2):165-177.
34. Law JH, Dunn PE, Kramer KJ: **Insect proteases and peptidases**. *Adv Enzymol Relat Areas Mol Biol* 1977, **45**:389-425.
35. Macedo MR, Freire MdGM: **Insect digestive enzymes as a target for pest control**. *Invertebrate Survival Journal* 2011, **8**(2):190-198.
36. Kutsukake M, Shibao H, Nikoh N, Morioka M, Tamura T, Hoshino T, Ohgiya S, Fukatsu T: **Venomous protease of aphid soldier for colony defense**. *Proceedings of the National Academy of Sciences* 2004, **101**(31):11338-11343.
37. Chacko SM, Thambi PT, Kuttan R, Nishigaki I: **Beneficial effects of green tea: a literature review**. *Chin med* 2010, **5**(13):1-9.
38. Serteyn L, Ponnet L, Backus EA, Francis F: **Characterization of electropenetrography waveforms for the invasive heteropteran pest, *Halyomorpha halys*, on *Vicia faba* leaves**. *Arthropod-Plant Interactions* 2020, **14**(1):113-126.
39. Will T, Vilcinskas A: **The structural sheath protein of aphids is required for phloem feeding**. *Insect biochemistry and molecular biology* 2015, **57**:34-40.
40. Youn YN: **Electrically recorded feeding behavior of *Nephotettix cincticeps***. *Journal of Asia-Pacific Entomology* 1998, **1**(2):147-161.
41. SoGAWA K: **Studies on the salivary glands of rice plant leafhoppers: II. Origins of the structural precursors of the sheath material**. *Applied Entomology and Zoology* 1967, **2**(4):195-202.
42. Jing P, Bai S, Liu F: **Prelimitary research on electrical penetration graph (EPG) waveforms in relation to feeding behavior of *Laodelphax striatellus***. *Chinese Journal of Applied Entomology* 2013, **50**(3):758-763.
43. Ahmad A, Kaushik S, Ramamurthy V, Lakhanpaul S, Ramani R, Sharma K, Vidyarthi A: **Mouthparts and stylet penetration of the lac insect *Kerria lacca* (Kerr)(Hemiptera: Tachardiidae)**. *Arthropod structure & development* 2012, **41**(5):435-441.
44. Takeda Y: **Studies on variations in genetic resources of tea in Japan and application to tea [*Camellia sinensis*] breeding**. *Bulletin of the National Institute of Vegetable and Tea Science (Japan)* 2002.
45. Lawrence PK, Koundal KR: **Plant protease inhibitors in control of phytophagous insects**. *Electronic Journal of Biotechnology* 2002, **5**(1):5-6.
46. Singh S, Singh A, Kumar S, Mittal P, Singh IK: **Protease inhibitors: recent advancement in its usage as a potential biocontrol agent for insect pest management**. *Insect Science* 2020, **27**(2):186-201.
47. Schlüter U, Benchabane M, Munger A, Kiggundu A, Vorster J, Goulet M-C, Cloutier C, Michaud D: **Recombinant protease inhibitors for herbivore pest control: a multitrophic perspective**. *Journal of Experimental Botany* 2010, **61**(15):4169-4183.

48. Bradford MM: **A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding**. *Analytical biochemistry* 1976, **72**(1-2):248-254.
49. Valaitis AP, Augustin S, Clancy KM: **Purification and characterization of the western spruce budworm larval midgut proteinases and comparison of gut activities of laboratory-reared and field-collected insects**. *Insect biochemistry and molecular biology* 1999, **29**(5):405-415.
50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q: **Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data**. *Nature biotechnology* 2011, **29**(7):644.
51. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015:btv351.
52. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC bioinformatics* 2011, **12**(1):1.
53. Wu W, Liu H, Dong Y, Zhang Y, Wong S-M, Wang C, Zhou Y, Xu Q: **Determination of Suitable RT-qPCR Reference Genes for Studies of Gene Functions in *Laodelphax striatellus* (Fallén)**. *Genes* 2019, **10**(11):887.
54. Yang C, Pan H, Liu Y, Zhou X: **Selection of reference genes for expression analysis using quantitative real-time PCR in the pea aphid, *Acyrthosiphon pisum* (Harris)(Hemiptera, Aphidiae)**. *PloS one* 2014, **9**(11):e110454.
55. Maroniche GA, Sagadín M, Mongelli VC, Truol GA, del Vas M: **Reference gene selection for gene expression studies using RT-qPCR in virus-infected planthoppers**. *Virology Journal* 2011, **8**(1):308.
56. Mogilicherla K, Howell JL, Palli SR: **Improving RNAi in the Brown Marmorated Stink Bug: Identification of target genes and reference genes for RT-qPCR**. *Scientific Reports* 2018, **8**(1):1-9.
57. Chen C, Xia R, Chen H, He Y: **TBtools, a Toolkit for Biologists integrating various biological data handling tools with a user-friendly interface**. *BioRxiv* 2018:289660.
58. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H: **SignalP 5.0 improves signal peptide predictions using deep neural networks**. *Nature biotechnology* 2019, **37**(4):420.
59. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Molecular biology and evolution* 2013, **30**(4):772-780.
60. Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS: **ModelFinder: fast model selection for accurate phylogenetic estimates**. *Nature methods* 2017, **14**(6):587.
61. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Molecular biology and evolution* 2014, **32**(1):268-274.
62. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS: **UFBoot2: improving the ultrafast bootstrap approximation**. *Molecular Biology and Evolution* 2017, **35**(2):518-522.

Tables

Table 1. Distribution of the top hit annotations against the NCBI nr database in different insect orders

Insect order	salivary gland specific assembly	gut specific assembly	combined assembly
Hemiptera	43.16%	40.88%	40.63%
Blattaria	21.88%	21.35%	22.15%
Coleoptera	4.69%	4.13%	4.97%
Lepidoptera	5.69%	3.82%	3.01%
Hymenoptera	2.41%	2.41%	2.42%
Diptera	0.51%	0.62%	0.57%
Others	21.66%	26.78%	26.26%

Table 2. Summary of identified digestive proteases

Category of protease	Nr annotated contig	Proteome identified	Percentage
Aminopeptidase	148	26	18%
Carboxypeptidase	89	12	13%
Dipeptidase	46	11	24%
Aspartic protease [a]	39	8	21%
Cathepsin B-like	68	14	21%
Cathepsin L-like	197	18	10%
Serine protease [b]	343	40	12%
Total	930	129	14%

[a] Aspartic protease includes cathepsin D and aspartic protease.

[b] Serine protease includes trypsin, chymotrypsin and elastase.

Table 3. Top 10 digestive proteases abundant in the salivary gland and gut

N. lugens		*L. striatellus*		*A. pisum*		*H. halys*			*N. cincticeps*	*E. onukii*	
Salivary gland	Gut	Salivary gland	Gut	Salivary gland	Gut	Primary salivary gland	Accessory salivary gland	Gut	Salivary gland	Salivary gland	Gut
NlCathepsinB_4	NlSerineProtease_1	LsCarboxypeptidase_10	LsSerineProtease_1	ApAminopeptidase_14	ApCathepsinB_8	HhSerineProtease_25	HhSerineProtease_5	HhCathepsinL_8	NcAminopeptidase_9	EMoSerineProtease-24	EMoCathepsin L-15
NlCathepsinL_1	NlCathepsinB_3	LsCarboxypeptidase_7	LsCathepsinB_1	ApCathepsinL_1	ApAminopeptidase_3	HhSerineProtease_9	HhSerineProtease_29	HhCathepsinL_5	NcAminopeptidase_16	EMoCathepsin L-15	EMoCathepsin L-4
NlAminopeptidase_10	NlSerineProtease_2	LsCarboxypeptidase_3	LsSerineProtease_17	ApAminopeptidase_26	ApDipeptidase_2	HhSerineProtease_27	HhSerineProtease_23	HhCathepsinL_7	NcAminopeptidase_10	EMoSerineProtease-26	EMoSerineProtease-24
NlAminopeptidase_4	NlCathepsinB_5	LsCathepsinB_1	LsAsparticProtease_1	ApAminopeptidase_19	ApAminopeptidase_5	HhCarboxypeptidase_27	HhCathepsinL_1	HhCathepsinB_3	NcAsparticProtease_1	EMoCathepsin L-5	EMoCathepsin L-5
NlCathepsinL_3	NlSerineProtease_4	LsDipeptidase_3	LsSerineProtease_16	ApDipeptidase_2	ApCathepsinB_9	HhSerineProtease_26	HhCarboxypeptidase_16	HhCathepsinL_10	NcCathepsinB_1	EMoCathepsin L-4	SerineProtease-26
NlAsparticProtease_1	NlCathepsinB_1	LsAminopeptidase_14	LsSerineProtease_19	ApAminopeptidase_3	ApCathepsinB_1	HhSerineProtease_18	HhCathepsinL_21	HhCathepsinL_14	NcCathepsinL_5	EMoCathepsin L-3	EMoCathepsin L-3
NlCarboxypeptidase_1	NlSerineProtease_7	LsCarboxypeptidase_11	LsSerineProtease_18	ApCathepsinF_1	ApCathepsinB_6	HhSerineProtease_27	HhCathepsinB_1	HhCathepsinL_13	NcAsparticProtease_2	EMoCathepsin L-14	EMoCathepsin L-16
NlDipeptidase_1	NlAminopeptidase_3	LsAsparticProtease_2	LsCathepsinB_3	ApAminopeptidase_10	ApCathepsinB_11	HhSerineProtease_24	HhCarboxypeptidase_1	HhAminopeptidase_11	NcAminopeptidase_8	EMoCathepsin L-16	EMoSerineProtease-12
NlSerineProtease_1	NlSerineProtease_6	LsAminopeptidase_11	LsSerineProtease_15	ApCarboxypeptidase_1	ApCathepsinB_5	HhSerineProtease_4	HhAsparticProtease_2	HhCathepsinL_9	NcCathepsinL_3	EMoSerineProtease-12	EMoCathepsin B-6
NlAminopeptidase_11	NlAminopeptidase_2	LsAminopeptidase_5	LsSerineProtease_12	ApCathepsinK_1	ApSerineProtease_3	HhSerineProtease_28	HhSerineProtease_19	HhCathepsinL_4	NcAminopeptidase_14	EMoCarboxypeptidase-7	EMoCathepsin L-14

Detailed information of proteases in the table are shown in additional file 3 and 6.

Figures

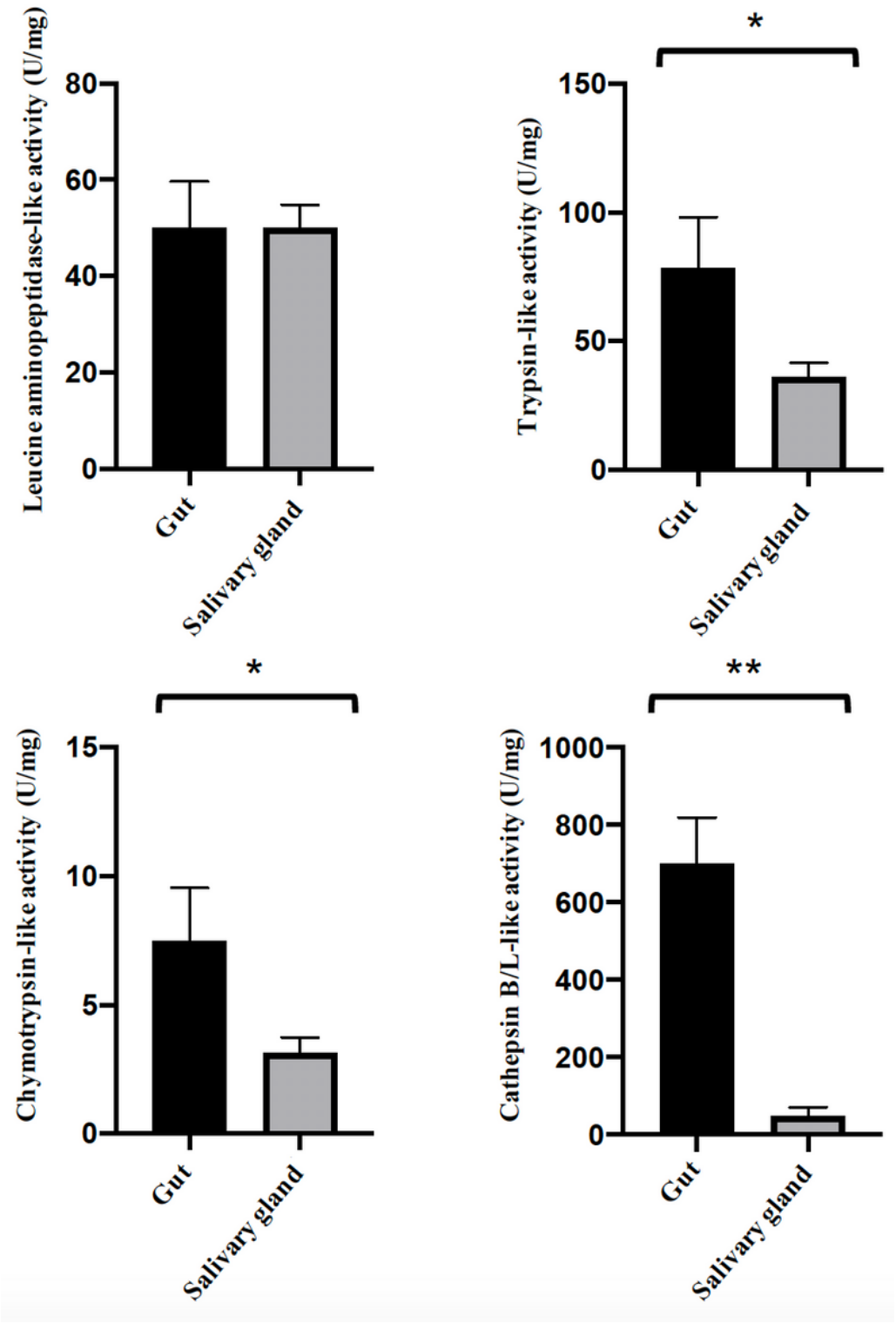

Figure 1

Enzymatic activity of selected proteases in homogenates of salivary glands and guts of E. onukii Error bars indicate the standard deviation from the mean for three replications. Statistical comparisons were conducted between the enzymatic activity of salivary glands and guts (*0.01<P <0.05, **P<0.01, Student's t-test).

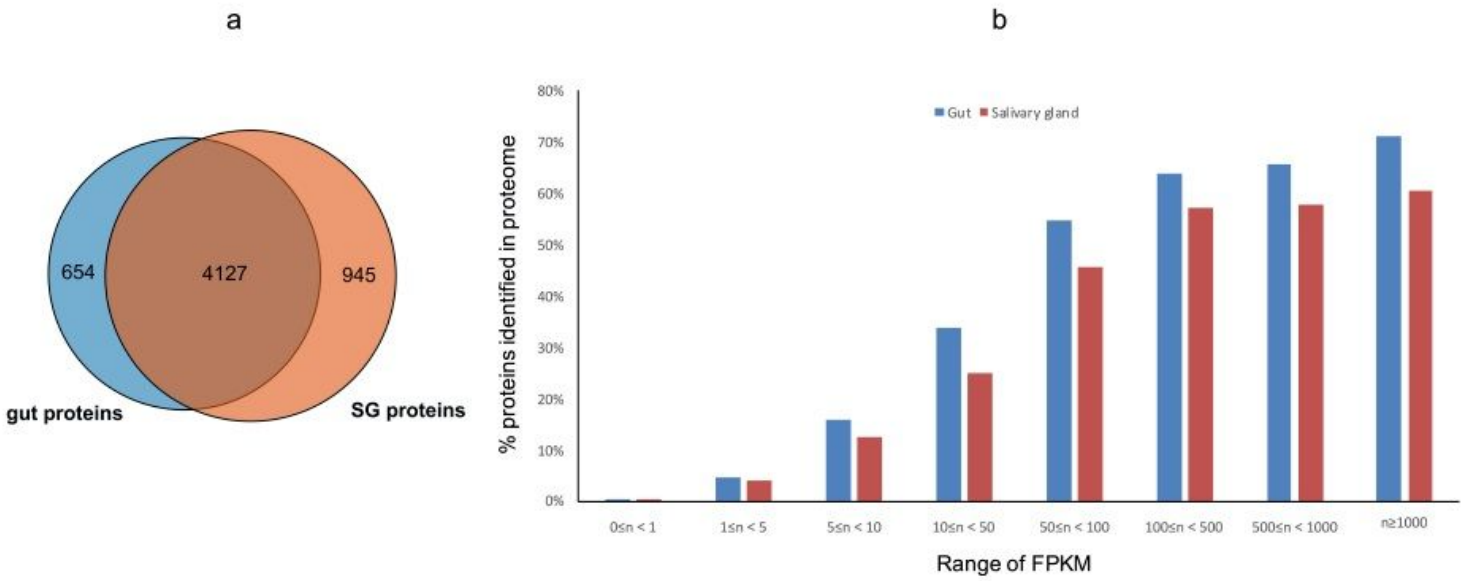

Figure 2

Summary of proteins mapped to proteomic peptides Panel a: Number of proteins identified in the SG, gut or both tissues. Panel b: Correlations between transcript abundance and the probability of proteins mapped to proteomic peptides. Translated protein sequence sets derived from the salivary glands and gut were respectively used as targets for mapping of peptides that resulted from the proteomic analysis of the corresponding tissue. The proportion of proteins detected from the proteome increased with increasing FPKM values.

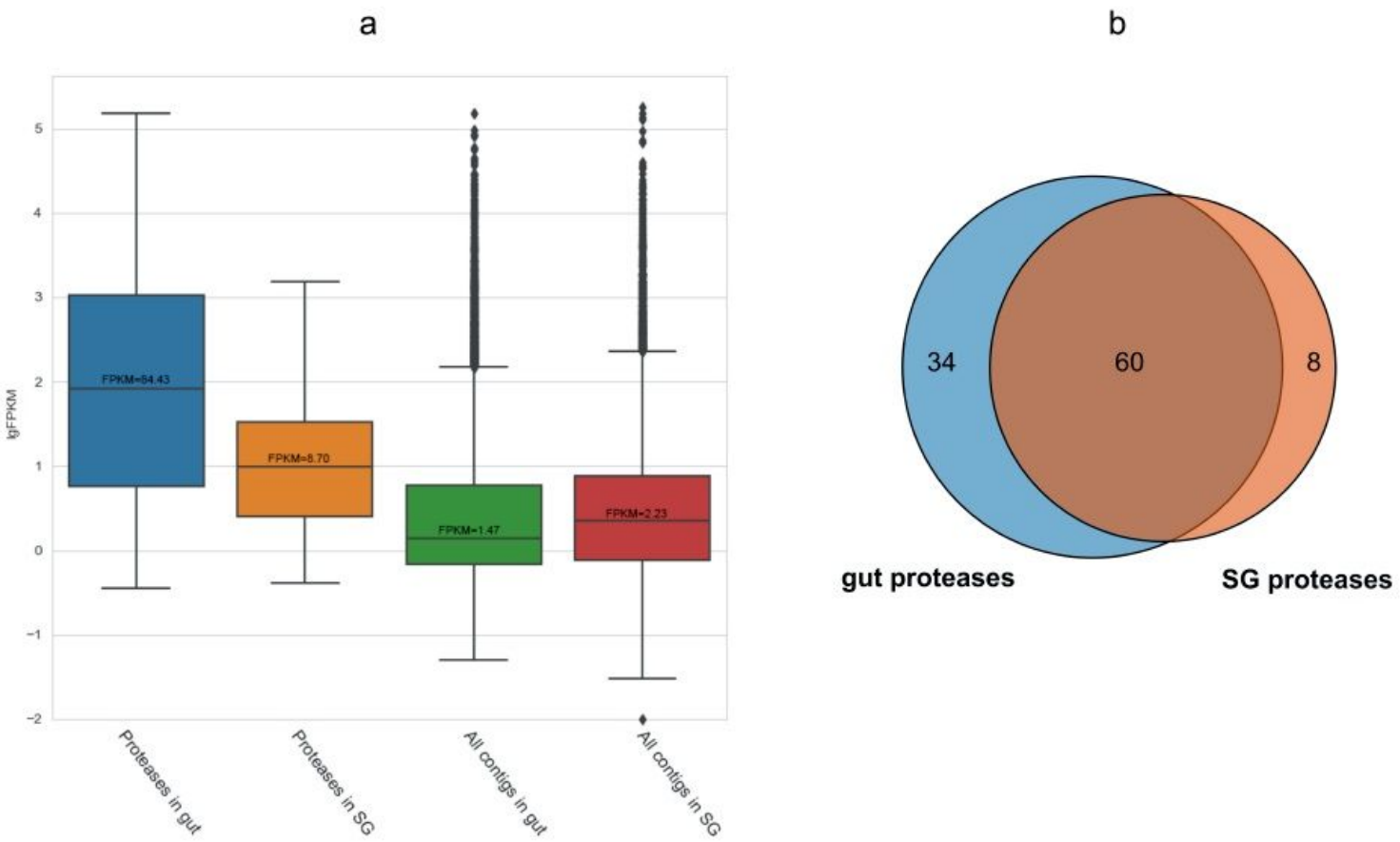

Figure 3

Transcriptomic abundance of proteases mapped to proteomic peptides Panel a: Median FPKM of predicted proteins in the SG and gut of E. onukii. Error bars indicate the standard deviation of the median FPKM of proteins in three biological replications. Panel b: Number of proteases identified in the SG, gut or both tissues.

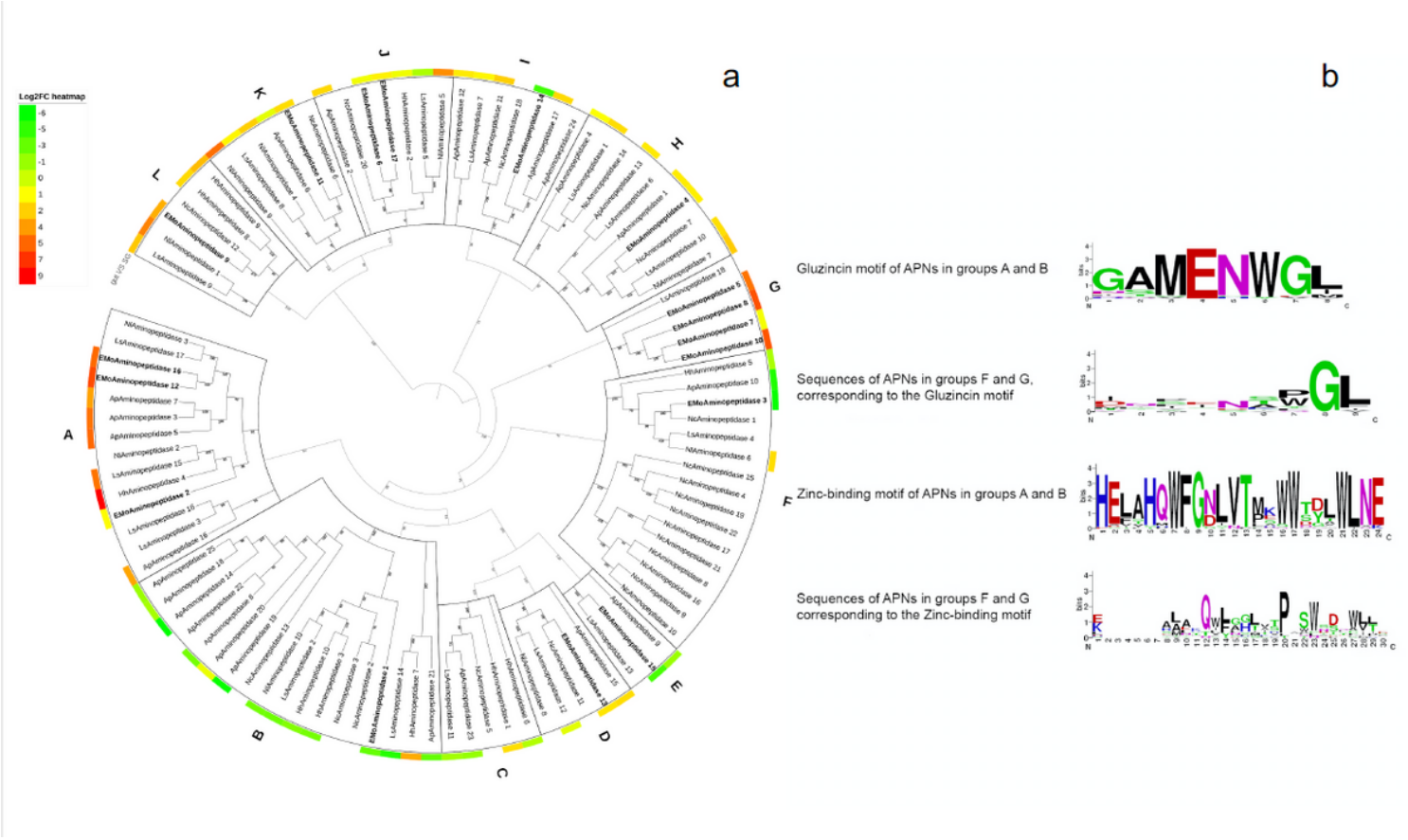

Figure 4

Phylogenetic tree containing aminopeptidases from six hemipteran species Protein sequences of aminopeptidase-like proteins from Nilaparvata lugens (Nl), Laodelphax striatellus (Ls), Nephotettix cincticeps (Nc), Acyrthosiphum pisim (Ap), Halyomorpha halys (Hh) and Empoasca (Matsumurasca) onukii (EMo) were aligned and used for the phylogenetic analysis through the maximum likelihood strategy (Panel a). Information regarding the proteins is shown in Files S3 and S6. Proteins from E. onukii are indicated in bold. Proteins in the phylogenic tree mentioned in the paper were arrowed. Groups A, B, F and G: proteins annotated by aminopeptidase N; Group C: glutamyl aminopeptidase; Group D: puromycin sensitive aminopeptidase; Group E: endoplasmic reticulum aminopeptidase; Group H: methionine aminopeptidase; Group I: xaa-Pro aminopeptidase; Group J: aminopeptidase NPEPL1; Group K: cytosol aminopeptidase; Group L: aminopeptidase W07G4.4e. Relative expression level was presented by log2FC (gut VS SG) and shown by a colored circle outside the tree, from green (log2FC = -6) to red (log2FC = 9). Proteins with no log2FC of FPKM were left blank. The sequence logos of the gluzincin motif and zinc-binding motif from aminopeptidase N-like proteins in groups A, B, F and G are shown on panel b.

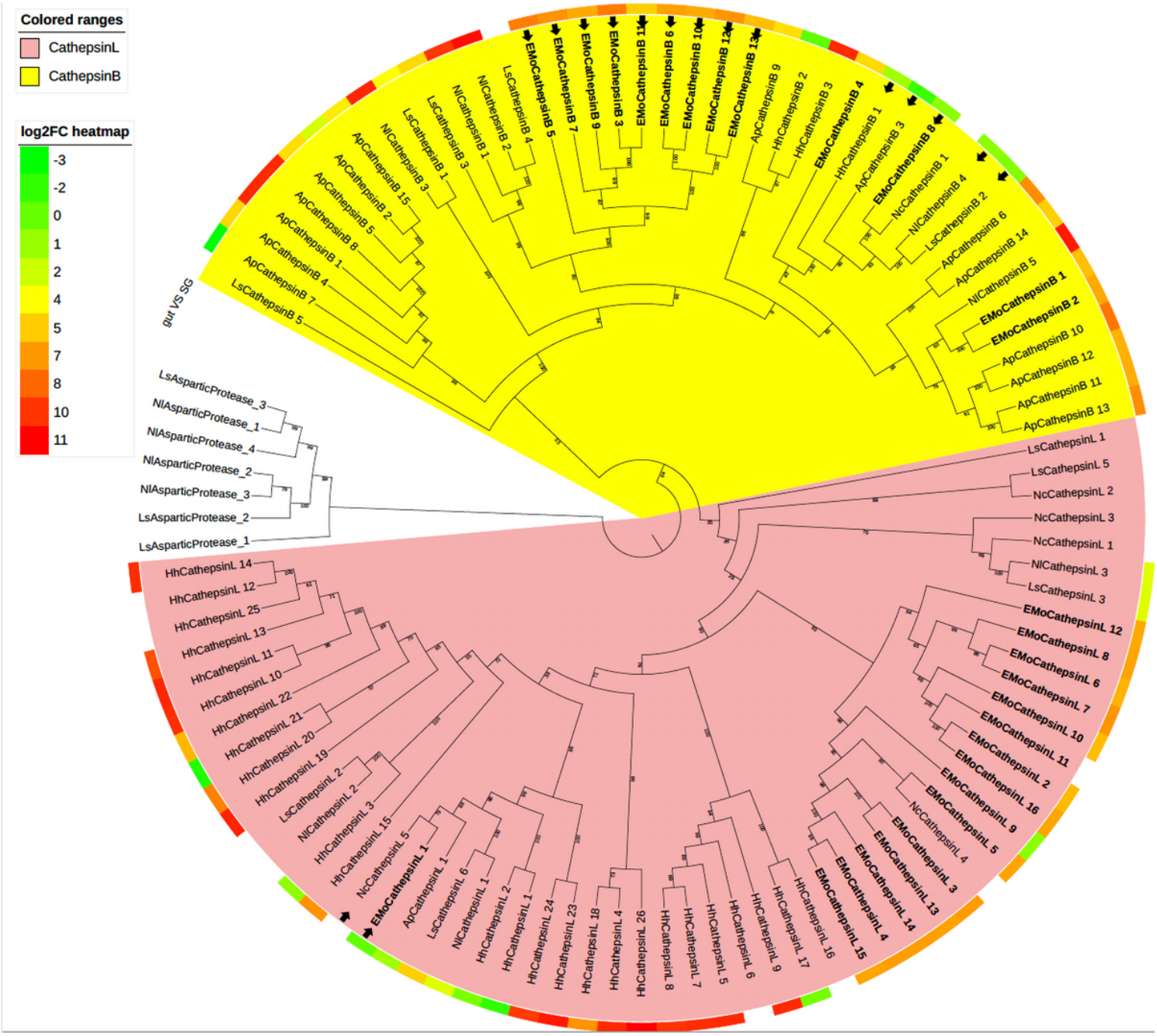

Figure 5

Phylogenetic tree containing cathepsin B- and cathepsin L-like proteins from six hemipteran species Protein sequences of cathepsin B- and cathepsin L-like proteins from N. lugens (Nl), L. striatellus (Ls), N. cincticeps (Nc), A. pisim (Ap), H. halys (Hh) and E. onukii (EMo) were aligned and used for the phylogenetic analysis through the maximum likelihood strategy. Cathepsin D-like proteins of N. lugens and L. striatellus were clustered as the outgroup. Information regarding the proteins is shown in Files S3 and S6. Proteins from E. onukii are indicated in bold. Proteins in the phylogenic tree mentioned in the paper were arrowed. The clade highlighted by pink and yellow contains cathepsin B- and cathepsin L-like proteins, respectively. Relative expression level was presented by log2FC (gut VS SG) and shown by a

colored circle outside the tree, from green (log2FC = -3) to red (log2FC = 11). Proteins with no log2FC of FPKM were left blank.

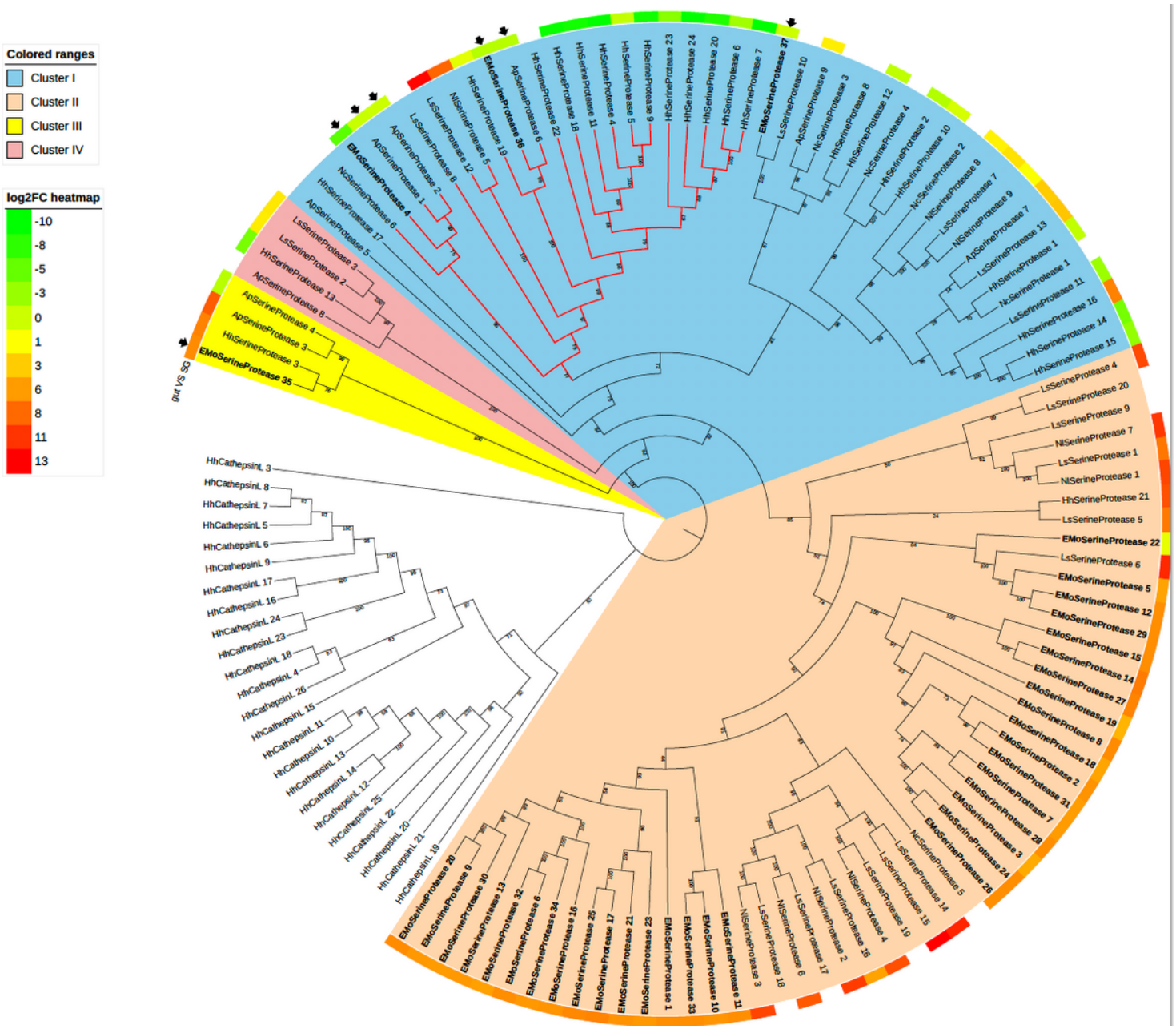

Figure 6

Phylogenetic tree containing serine protease-like proteins from six hemipteran species Protein sequences annotated by trypsin, chymotrypsin and elastase from N. lugens (Nl), L. striatellus (Ls), N. cincticeps (Nc), A. pisim (Ap), H. halys (Hh) and E. onukii (EMo) were aligned and used for the phylogenetic analysis through the maximum likelihood strategy. Information regarding the proteins is shown in Files S3 and S6. Proteins from E. onukii are indicated in bold. Proteins in the phylogenic tree mentioned in the paper were arrowed. Clades highlighted by blue and orange indicate two clusters of serine protease. The branch highlighted by red indicates a cluster of serine protease showing relatively higher FPKMs in the salivary gland than in the gut. Relative expression level was presented by log2FC (gut VS SG) and shown by a

colored circle outside the tree, from green (log2FC = -10) to red (log2FC = 13). Proteins with no log2FC of FPKM were left blank.

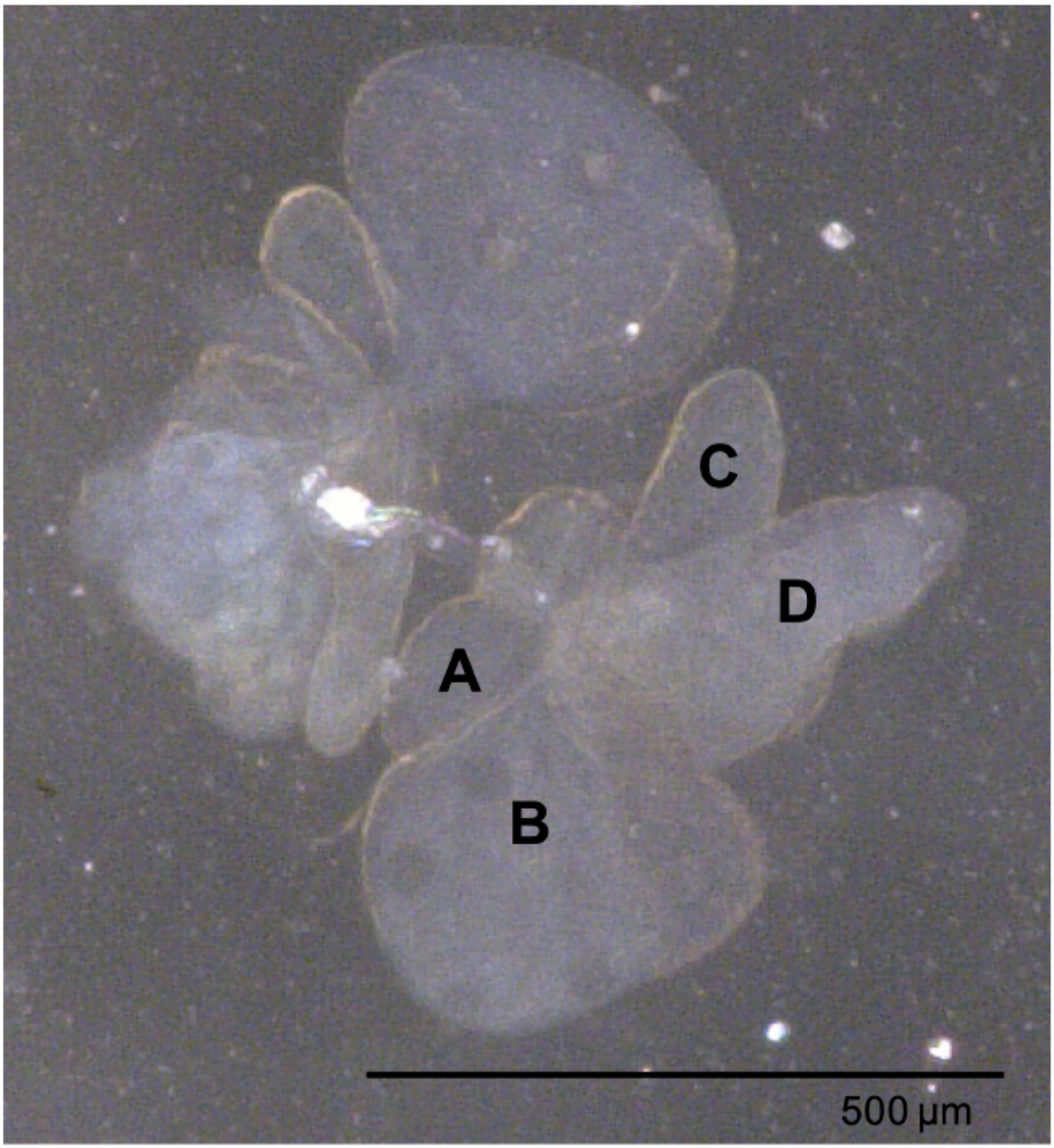

Figure 7

Excised salivary glands of Empoasca (Matsumurasca) onukii nymphs A-D indicate four lobes of one side of the salivary gland.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2.pdf
- Additionalfile3.xlsx
- Additionalfile4.txt
- Additionalfile5.xlsx
- Additionalfile6.xlsx