Microarray-based annotation of the gut transcriptome of the migratory locust,

Title: Microarray-based annotation of the gut transcriptome of the migratory locust, *Locusta migratoria*.

Authors: Jornt Spit[a], Liesbeth Badisco[a], Lucia Vergauwen[b], Dries Knapen[b] , Jozef Vanden Broeck[a]

[a] Department of Animal Physiology and Neurobiology, Zoological Institute K.U.Leuven, Naamsestraat 59, B-3000 Leuven, Belgium

[b] Department of Biology, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerpen, Belgium

Abstract: The African migratory locust, *Locusta migratoria*, is a serious agricultural pest and important insect model to study insect digestion and feeding behavior. The gut is one of the primary interfaces between the insect and its environment. Nevertheless, even though its genome was recently published, knowledge on the gut transcriptome of *L. migratoria* is still very limited. Here, 48802 ESTs were extracted from publicly available databases and their expression in larval gut and/or brain tissue was determined using microarray hybridization. Our data show 2765 transcripts to be predominantly or exclusively expressed in the gut. Out of these sequences, 935 could be functionally annotated. Many transcripts had putative functions closely related to the physiological functions of the gut as a muscular digestive organ and as the first barrier against microorganisms and a wide range of toxins. By means of a ranking procedure based on the relative signal intensity, we estimated 15% of the transcripts to show high expression levels, the highest belonging to diverse digestive enzymes and muscle related proteins. We also found evidence for very high expression of an allergen protein that is closely related to the major allergen from *Blatella germanica*. High expression of an allergen protein could have important implications, since locusts form a traditional food source in different parts of the world, and were also more recently added to the list of edible insects fit for human consumption in Europe. Interestingly, many highly expressed sequences have yet unknown functions. Taken together, the present data provide significant insight into locust larval gut physiology, and will be valuable for future studies on the insect gut.

1. Introduction

The African migratory locust, *Locusta migratoria* is a hemimetabolous insect species belonging to the order of the Orthoptera. These locusts possess the intriguing ability to change between two phenotypes, depending on the density of the population (Verlinden et al., 2009). In their gregarious phase locusts are known to be devastating pest insects that will actively aggregate to form marching hopper bands or flying swarms. Since a single locust can ingest its own body weight in food each day, these swarms can inflict enormous damage to agricultural production in large areas of the world, including Africa, the Middle East and the Indian subcontinent. Today, the control of locust plagues relies mainly on chemical pesticides. These are not only harmful for the environment, but also pose significant risk to human health after occupational exposure. Therefore, there is a constant demand for sustainable pest management strategies, whereby the digestive system forms one of the most attractive targets. The digestive system constitutes the first barrier with the outside world and is hence involved in various physiological and biological processes, including food digestion, immune responses, detoxification, and interactions with hosts and symbionts.

With the development of two EST databases from *L. migratoria* (whole body and CNS) and one EST database from *S. gregaria* (CNS), an abundance of transcript data was made available for locusts (Badisco et al., 2011; Ma et al., 2006; Zhang et al., 2012). In addition, the genome of *Locusta* was also recently published in an effort to create a better understanding of swarm formation and flight behavior (Wang et al., 2014). While the transcript composition of nervous tissue was relatively well studied after the development of the specific CNS derived EST-databases from both *L. migratoria* and *S. gregaria* (Badisco et al., 2011; Zhang et al., 2012), little transcript profiling information is available for the digestive system at the moment. Locusts are however widely used as physiological model organisms regarding the regulation and control of feeding and digestion, and improved knowledge on the gut transcriptome could contribute significantly to a better understanding of their gut physiology.

Therefore, we aimed to use the available sequence data to specifically identify gut-expressed transcripts in 5th larval locusts. By means of two independent self-self microarray hybridizations for two distinct tissues, the gut and the brain, a selection could be made of those ESTs that are present in the gut and/or the brain. Here, sequences that were found to be expressed in gut but not brain were further functionally annotated to shed new light on the complex physiology of the locust digestive system. Since the gut is the single most important organ in digestion, and both tissues are assumed to be involved in the regulation thereof, the resulting subset of sequences can also be valuable for further in depth studies on the regulation of digestion. In addition, the method allowed us to rank the signal intensities, using them as a rough indicator to compare relative transcript

abundance in the gut. The data complements previously published transcript and genomic data, and provide a clear overview of the expressed portion of the genome in the gut.

2. Material and methods

2.1 Rearing of animals and sample collection

Locusts (*L. migratoria*) were reared under crowded conditions with controlled temperature (32 ± 1°C), light (14 h photoperiod) and relative humidity (40–60%) and fed daily with grass. Five day old fifth instar larvae were dissected for use in a microarray hybridization experiment. The main objective of this experiment was to make a selection of transcripts from the total EST-data, that are expressed in the gut and/or brain. Therefore, locusts from different biological conditions (regularly fed, fed diet supplemented with protease inhibitors, and locusts starved for three days) were used to ensure a total representation of all expressed genes for future research purposes. Tissues from 3 times 5 animals were pooled for each condition. Foregut and hindgut combined, midgut, gastric caeca and brain were collected separately. Tissues were dissected in locust saline solution (155 mM NaCl, 6.5 mM KCl, 1.6 mM NaH_2PO_4, 3 mM $NaHCO_3$, 7.7 mM $MgCl_2$, 2.9 mM $CaCl_2$) and immediately transferred to liquid nitrogen. Samples were stored at -80°C until further processing.

2.2 RNA extraction and quality control

Total RNA from dissected gut tissues was extracted using the RNeasy Lipid tissue kit (Qiagen), while total RNA from brain was extracted using the RNeasy Mini Kit (Qiagen). DNaseI treatment was performed to remove traces of genomic DNA contamination. Quality and concentration of the extracted RNA were assessed using the Agilent 2100 Bioanalyzer. For all dissected gut parts (Foregut and hindgut, midgut, caeca), equal quantities of RNA were combined and used as template to produce one pool of 'gut' cRNA.

2.3 Labeling of cRNA

Starting from total RNA, fluorescently labeled cRNA was generated. Labeling of the samples was performed with the Quick Amp Labeling Kit (Agilent Technologies), according to the manufacturer's instructions. In brief, each sample was labeled with a Cy5 (red) and Cy3 (green) fluorescent dye. For both gut and brain sample, 1 µg total RNA from a combination of all pools was used for labeling. mRNA was reverse transcribed to cDNA by using a poly(dT) primer that is coupled to an antisense T7 promotor. The reverse transcriptase also catalyzes synthesis of the second cDNA strand. Prior to cDNA synthesis, a Spike A or Spike B mix was added to samples to be labeled with Cy3 or Cy5, respectively. These Spike mixes contain polyadenylated transcripts from the adenovirus E1A, that are premixed in various quantities and ratios. They are used as a control for the workflow and were used to normalize expression data. In a second step, the second cDNA strand functions as template for the synthesis of cRNA, for which fluorescently labeled dCTPs (Cy3 or Cy5) are provided. The fluorescently

labeled samples were subsequently purified by means of the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. Yield, concentration and quantity of fluorescent labels incorporated were determined by a NanoDrop spectrophotometer. For both brain and gut the yield was higher than 7 µg, and all samples showed an activity higher than 17 pmol Cy per µg cRNA, which meets the minimum requirements of 825 ng total cRNA and 8 pmol Cy per µg cRNA.

2.4 Design of the microarray

A 2 x 105K Custom Gene Expression Microarray slide (Agilent) was spotted with probes that represented 12107 unigenes from LocustDB (Ma et al., 2006), all *L. migratoria* ESTs from the CNS deep-sequencing project (Zhang et al., 2012) that did not appear to be represented in LocustDB, and all *S. gregaria* EST sequences (Badisco et al., 2011) that had no ortholog in the *L. migratoria* (blast-N hits producing E-value < 1e-10 were considered as orthologs). For each of the resulting 48802 unique sequences, two different probes were spotted in addition to standard Agilent control features. Probes were designed by means of the eArray system (Agilent).The best probe methodology was used selecting two probes per target, with a probe length of 60 nucleotides in sense orientation.

2.5 Hybridization, scanning and data analysis

To make a selection of transcripts that are expressed in gut and/or brain we performed self-self hybridizations of both gut and brain tissue separately. Both Cy5- and Cy3-labeled cRNA samples were co-hybridized onto the Agilent Custom Gene Expression Microarray. Once for brain cRNA and once for gut cRNA. Hybridization was performed at 65°C for 17h. Next, the slide was washed: twice in GE Wash Buffer 1 (Agilent) for 1min each, in GE Wash Buffer 2 (Agilent) at 37°C for 1min, in acetonitrile (Sigma) for 1min and in Stabilization and Drying Solution (Agilent) for 30sec. Subsequently, the microarray slide was scanned using a Genepix Personal 4100A confocal scanner (Axon Instruments) at a resolution of 5 µm and excitation wavelengths of 635 nm and 532 nm. The photomultiplier tube voltages for separate wavelengths were adjusted to obtain an overall green/red ratio as close as possible to 1. Theoretically, ratios for all probes are expected to be 1, since we are dealing with self-self hybridizations from a cRNA sample that is labeled in both red (Cy5) and green (Cy3). Images were processed using GenePix Pro 6.0 software (Axon Instruments) for spot identification and quantification of the fluorescent signal intensities. Each transcript was spotted with two different probes. Only one probe per expressed transcript was retained for analysis and future use. For transcripts where both probes emitted a signal, but one probe produced a lower signal for both red and green, the weakest probe was omitted. When neither one of the probes had the lowest intensity for both the red and the green signal, the probe with the smallest deviation from green/red ratio = 1 was selected. Transcripts for which the median foreground (FG) intensity was lower than the average

local background (BG) intensity (calculated over the complete array) plus one standard deviation (SD) for both the brain and gut array were considered unexpressed and were deleted before analysis. Any spot with an intensity of FG > BG + 1SD in either tissue was initially considered to be expressed in that tissue. To be able to roughly compare spot intensities between brain and gut, transcript intensity was calibrated against spiked-in controls, which were added in the same quantity in both samples. This was done by dividing the average signal intensity of the spiked-in control group spots from the brain sample by that of the gut sample and multiplying the original signal intensity of the gut transcripts with the obtained factor. The transcripts were sorted on the average Cy5 and Cy3 intensity in both brain and gut. In addition to transcripts that had a FG > BG + 1SD in only one of both tissues, transcripts with an average intensity ratio of > 20 were also considered tissue specific.

2.6 Annotation of gut specific transcripts

Gut specific transcripts were functionally annotated using the InterProScan (IPS) tool that is integrated in the Blast2GO software (Conesa et al., 2005; Quevillon et al., 2005). When multiple IPS hits were received for the same transcript, the most relevant one was retained. In addition, BLAST searches in the NCBI database were performed and the best hit was used for the sequence description of the transcripts.

3. Results and Discussion

3.1 Microarray hybridization revealed transcripts expressed in gut- but not brain-tissue

In a custom microarray setup, 42861 ESTs from *L. migratoria* and 5941 ESTs from *S. gregaria* were spotted, for a total of 48802 unique sequences. By means of dual-color labeled gut and brain RNA-samples self-self hybridizations, ESTs that were expressed specifically in gut (combination of foregut, midgut, gastric caeca and hindgut), brain or both tissues in 5^{th} instar larvae were identified. 11841 and 2191 transcripts met the initial FG > BG + 1SD criterion for brain or gut, respectively, while 21837 ESTs appeared to be expressed in both tissues. To further account for differences between intensity levels of the same EST that is expressed in both tissues, spot intensities were calibrated against spiked-in controls and compared between brain and gut. Plotting the resulting Cy3- against the Cy5-intensity signal for all transcripts resulted in little deviation from the diagonal axis for both midgut and brain tissue, as expected from self-self hybridizations (fig. 1). Significant variation in mean intensity of the fluorescent signal between transcripts could be observed, and signal intensity was used to estimate and compare expression levels between both tissues. All transcripts that showed intensity levels 20 times higher in one tissue compared to the other were considered to be tissue specific. The cutoff value of 20 was chosen arbitrarily and other sequences with slightly lower ratios may also be predominantly expressed in one of the tissues. Most of the transcripts with an intensity difference over 20 were positioned at the high-end intensity spectrum in one of the tissues, while being very lowly expressed in the other (fig. 1), indicating this was a valuable approach to select additional tissue specific sequences. In total, an extra 574 and 208 transcripts could be designated as respectively gut or brain specific by this method, for a final number of 2765 and 12049, while 21055 transcripts were expressed in both tissues. The reason for higher numbers of 'brain specific' sequences is most likely that the total set of locust EST sequences is biased for sequences from the CNS. Indeed, the ESTs expressed in brain are almost entirely derived from both CNS sequencing projects (88% of all 'brain only' ESTs). Alternatively, it could be possible that the brain expresses more tissue specific sequences. Almost half (43%) of all sequences appear to be expressed in both tissues. This shows the value of microarray-based approaches for tissue specific annotation purposes, since direct annotation of sequenced ESTs from a particular tissue will tend to be biased for high amounts of non tissue specific transcripts. A list of all ESTs that were identified to be expressed in either gut or brain tissue is provided as supplementary data, table S1.

3.2 Functional annotation of gut specific sequences indicates a majority of sequences involved in digestion, transport, defense and peritrophic membrane function

Annotation of CNS derived sequences was already done extensively for both *L. migratoria* and *S. gregaria* (Badisco et al., 2011; Zhang et al., 2012). However, knowledge on the gut transcriptome of locusts is limited and therefore functional annotation of sequences expressed in locust gut but not brain was performed using the InterProScan (IPS) tool integrated in the Blast2GO program. 1331 ESTs produced at least one BLAST hit. However, only 935 out of 2765 ESTs (34%) produced strong IPS results. For each sequence only the most relevant IPS result is retained. A total of 344 unique IPS annotations was obtained. A relatively large proportion (66%) of sequences could not be functionally annotated. High numbers of sequences with no known matches are often observed in the annotation process of insect guts (Coates et al., 2008; Hughes and Vogler, 2006; Pedra et al., 2003), suggesting a significant diversification in insect midgut genes. An overview with all identified transcripts in the gut, IPS annotation, best BLAST hit, mean similarity, and relative ranking based on signal intensity can be found in the supplementary data, table S2.

A summary of some of the most frequent and relevant InterPro families is presented in table 1. Results show large numbers of transcripts primarily associated with food breakdown, transport, detoxification, muscle function, and the peritrophic membrane.

a) *Sequences involved in digestion*

Putative proteases were among the most detected transcripts expressed in the Locust gut, including large numbers of serine (S1) proteases, some smaller numbers of cysteine (C) proteases and different families of metallopeptidases (M), including amino- and carboxypeptidases. These results are in full agreement with our previous findings of high numbers of trypsin- and chymotrypsin-like serine proteases that are active in the midgut of *L. migratoria*, where specific data mining of the LocustDB resulted in 95 transcripts that could be further assembled into 20 unique serine protease encoding sequences (Spit et al., 2014). In addition to serine protease activity, minor cysteine and some carboxypeptidase activity could also be shown in locusts (Spit et al., 2014, 2012). Putative dipeptidases and two sequences with an asparaginase 2 domain, involved in the breakdown of glycoproteins, could also be predicted (table 1).

Many sequences putatively involved in carbohydrate digestion were also identified. Carbohydrates make up an important part of the insect diet, as polysaccharides are major constituents of plant cell walls and starch. The sequences belonging to the glycoside hydrolase superfamily were mostly divided over family 1 (β-glucosidases), family 9 (cellulases), and glycoside hydrolase family 13, which

includes maltases and α-amylases. For several other glycoside hydrolase families such as trehalase, α- and β-galactosidase, α-glucosidase, and β-mannosidase, a single transcript was predicted (table 1).

Several putative lipase transcripts were also detected. In insects, most studies have focused on the roles of lipases in the fat body, where lipids are stored in the form of droplets. In locust fat body, triacylglycerol lipase triggers lipolysis, leading to diacylglycerols, which are released into the hemolymph to provide energy (Auerswald and Gäde, 2006). However, lipases involved in lipid digestion in the midgut were described in a number of insects, including *Rodnius prolixus* (Grillo et al., 2007), *Bombyx mori* (Ponnuvel et al., 2003), and *Helicoverpa armigera* (Sui et al., 2008).

b) Peritrophic membrane function and chitin metabolism

The second most identified InterPro domain in the *L. migratoria* gut was the chitin binding domain type 2 (ChtBD2). It is characterized by six conserved cysteines that form three disulphide bridges. This sequence motif is also often referred to as the peritrophin A domain. It occurs most often in proteins extracted from the peritrophic membrane (PM). Nevertheless, extensive phylogenetic analyses in *Tribolium* have shown other protein families containing ChtBD2s (Jasrapuria et al., 2010).

The number of seemingly unique transcripts that contain (at least one) chitin binding domain is striking. One of the reasons for this could be the highly repetitive domain nature of peritrophins. The number of ChtBD2s in peritrophins can vary, ranging from one to at least 19 (Jasrapuria et al., 2010). Because of ambiguities created in alignments of sequence reads, transcript assembly of sequences containing high numbers of repetitive domains has always proved to be a technical challenge (Treangen and Salzberg, 2012). Here, the maximum number of ChBD2s we could identify in a single transcript was 5 (supp. table S3). Nevertheless, not all transcripts encode a full ORF, suggesting the possible existence of proteins containing higher number of ChtBD2s. In addition, a multiple sequence alignment of all identified ChBD2s also showed that several ChBD2s from different transcripts are perfectly identical (data not shown), suggesting that several of the transcripts could have been misassembled in the original construction of the databases. Further research will be necessary to determine the amount of peritrophin proteins present in the gut, and their number of ChBD2s, in *Locusta migratoria*.

Putative chitin synthase and chitinase genes were also predicted. Blast results indicated the transcripts correspond to *Locusta migratoria* chitin synthase 2 (supp. table S2), which was previously shown to be expressed exclusively in the midgut and gastric caeca (Liu et al., 2012). Midgut chitin synthases and chitinases are known to be involved in the formation and degradation of the PM. Alternatively, they may also be involved in the regulation of molting, as chitin is also a major constituent of the insect cuticle (Merzendorfer, 2003).

c) Sequences involved in detoxification

Different families of detoxification genes have been shown to play essential roles in insect defense against natural and synthetic insecticidal compounds. Several genes associated with detoxification, including carboxylesterases, glutathione S-transferases, UDP-glycosyltransferases, and cytochrome P450 could be identified in the gut specific transcripts of *L. migratoria*.

The results show multiple transcripts encoding a carboxylesterase type B. These enzymes hydrolyze ester bonds from various substrates that posses a carboxylic ester. They are reported to be involved in the detoxification of xenobiotic insecticides such as organophosphates in insects. A recent EST database screen identified 25 different carboxylesterase genes in *L. migratoria*, from which 12 were significantly upregulated in a field-derived malathion resistant colony (Zhang et al., 2011). In addition, 2 carboxylesterases from *L. migratoria*, LmCesA1 and LmCesA2, were characterized into detail and shown to be primarily expressed in gastric caeca, confirming the gut only expression we observed in this study. Both LmCesA1 and LmCesA2 were shown to play a significant role in the detoxification of chlorpyrifos, which is a commonly used organophosphate (Zhang et al., 2013).

Another InterPro family observed in the gut-derived transcripts comprises the glutathione S-transferases (GSTs). GSTs catalyze the conjugation of electrophilic compounds with the thiol group of reduced glutathione, resulting in products that are often more water soluble and excretable. The majority of studies on insect GSTs have focused on their role in detoxifying foreign compounds, in particular insecticides and plant allelochemicals. Additionally, they have been reported to be involved oxidative stress responses (Enayati et al., 2005). These enzymes play an important role in the detoxification of xenobiotic compounds. In *L. migratoria*, the insecticide detoxification potency of 5 GSTs was characterized over the past years. LmGSTs5, LmGSTs3 and LmGSTu1 were shown to be involved in carbaryl, malathion and chlorpyrifos detoxification (Qin et al., 2012; 2013). High expression of these GSTs was detected in midgut, caeca and hindgut. Nevertheless, expression of several GSTs was highest in Malpighian tubules and fat body, two tissues we didn't include in our analysis (Qin et al., 2011).

Another protein family known for several roles in detoxification and stress responses is the cytochrome P450 superfamily. Different members of this superfamily oxidize steroids, fatty acids and xenobiotics, and are important for the destruction and elimination of various plant secondary compounds, such as terpenoids, furanocoumarins, and insecticides, e.g. pyrethroids and organophosphates (Schuler, 2011). Members of cytochrome P450 class E, group I and II were found in our data (supp. table S2). Group I includes the CYP2-family, while group II contains the CYP3- and

CYP4-family. Together with the mitochondrial P450 genes, these are the 4 major insect P450 clades (Feyereisen, 2006).

Five gut specific transcripts were predicted to be UDP-glycosyltransferases. These enzymes catalyze the glucosidation of small hydrophobic molecules, and as such also fulfill major roles in the inactivation of various plant secondary metabolites and xenobiotics (Ahn et al., 2012).

The occurrence of large numbers of all the above detoxification families in *L. migratoria* was also confirmed by the recently published genome, where 68 UDP-glycosultransferase-, 80 carboxyl/ choline esterase-, 94 cytochrome P450- and 28 glutathione S-transferase-genes were identified (Wang et al., 2014). It is suggested that high amounts of these enzymes allow *L. migratoria* to handle a broad range of secondary metabolites present in different host plants.

d) Transporter-like sequences

A relatively large subset of sequences showed homology with transporter-like sequences. Many transcripts belonged to the major facilitator superfamily (MFS), and the ATP-binding cassette (ABC) superfamily. The ABC superfamily forms one of the largest groups of membrane transporters, that can bind and hydrolyze ATP while transporting a large diversity of substrates across membranes. They may have diverse physiological functions, also including roles in xenobiotic resistance (Dermauw and Van Leeuwen, 2013). In locusts, so far, no ABC transporters have been characterized into detail.

The MFS also represents one of the largest families of secondary transporters in the cell and these proteins are capable of transporting small molecules in response to different ion gradients. Members comprise uniporters and symporters, as well as antiporters, and MFS transporters can bind a variety of substrates, including lipids, amino acids, carbohydrates, peptides and ions (Yan, 2013). Among the transcripts we found sequences encoding sugar transporters, an oligopeptide transporter, and an amino acid transporter, all belonging to the MFS superfamily.

e) Immunity

Several transcripts may be associated with locust immunity. C-type lectins, containing a carbohydrate recognition domain, mainly function in innate immunity as microbial pattern recognition molecules (Kanost et al., 2004; Theopold et al., 1999). In addition, peptidoglycan recognition proteins (PGRPs) recognize peptidoglycans from the bacterial cell wall (Aggrawal and Silverman, 2007), while the Toll/interleukin-1 receptor homology (TIR) domain, occurring in Toll-like receptors (TLRs) and the interleukin-1 receptor (IL-1R) superfamily, is also involved in innate antibacterial and antifungal

immunity in insects. Finally, serpins, or serine protease inhibitors are a superfamily of proteins that control protease-mediated processes, and have been associated with different functions in immunity in insects. For example, in *Drosophila*, the Toll-mediated innate immune response and the prophenol oxidase pathway both under control of serpins (Garrett et al., 2009; Reichhart et al., 2011).

Eleven gut transcripts were annotated as members of the pacifastin serine protease inhibitor family. Three precursor genes encoding 9 different inhibitor domains from this family are known to exist in *L. migratoria* (Simonet et al., 2002). All 11 identified transcripts appear to encode the same mature protein, namely LmPP2, containing three inhibitor domains, LmPI-3, LmPI-4 and LmPI-5 (Simonet et al., 2002). LmPP-2 shows a high degree of sequence identity with SGPP-5 from *S. gregaria*. A tissue distribution for this gene showed high expression in foregut and hindgut, with no expression in the midgut (Simonet et al., 2004). The exact functioning of these PIs in locusts remains elusive, nevertheless, roles in insect immunity were suggested (Breugelmans et al., 2009).

f) Insect allergen-related

Interestingly, one other group of highly abundant transcripts comprised those encoding insect allergen-related proteins. Detailed analysis showed that most transcripts in this group could be further assembled into a single sequence that encoded a protein that shows remarkably high sequence similarity with two major allergen sequences from the German cockroach *Blatella germanica* (fig. 2). Bla g 1 was identified as a human allergen from the German cockroach where it is produced by the gut and excreted in the faeces (Arrudai et al., 1995; Gore and Schal, 2005; Randall et al., 2013). Similar excretion in the faeces of locusts could help explain the emergence of locust-related allergies such as asthma (Lopata et al., 2005). The natural functions of these proteins are still unknown, but are suggested to be digestion related, as there is a strong sequence similarity with aspartic proteases, while expression seems to be regulated in response to food intake (Arrudai et al., 1995). Expression of allergens or related proteins could also have important implications for human consumption of locusts. Entomophagy is common to cultures in several parts of the world, and so far, not a lot is known about the potential thread of insect-derived food allergens (Fao, 2013).

3.3 Ranking signal intensities shows that digestive enzymes, allergens and muscle proteins are highly expressed

When we assume that incorporation of fluorescently labeled dCTPs was approximately equal for all cRNAs, average signal intensities can be used to rank the expression of different transcripts, where the brightest spots represent genes that are expressed the most and are ranked the highest. A similar strategy has been used in the characterization of the gut transcriptome of *T. castaneum* (Morris et al., 2009). From the 2765 designated gut specific transcripts, the majority of transcripts were

assigned a relative ranking lower than 10. Only 422 transcripts (15%) had fluorescence intensity levels that were at least 10% of the maximum intensity (fig. 3). The ranking of each transcript is also provided in supplementary table S2. If we select the 150 most intense spots after hybridization of the gut samples, 125 belonged to the subset of sequences determined to have expression only in the gut but not in the brain, while 25 also are expressed in brain tissue (table 2). The latter mostly corresponded to ribosomal protein encoding genes or structural and muscle protein encoding genes. A full list of the transcript IDs together with their observed gut over brain intensity ratio, IPR results, and best BLAST hit is provided in supplementary table S4.

Serine proteases, β-glucosidases and α-amylases are expressed at very high levels, while lipases and other classes of proteases are only moderately or lowly expressed and were not present among the 150 most intense spots from the gut. Other genes that are very highly expressed in the gut, and in much higher levels than in brain, were actin, myosin, troponin and calponin. Actin and myosin are major components of muscle thick and thin filaments, respectively. Troponin is a thin filament-associated protein that is involved in muscle contraction, and calponin is part of a thin-filament based regulatory system (Hooper and Thuma, 2005). Very high expression of such muscle related transcripts likely reflects the muscular nature of the gut, necessary to propel the food bolus through the digestive tract. Interestingly, a relatively large number of transcripts did not share similarity to any known protein. However, the very high expression levels suggest important functions in gut physiology, and it would therefore be interesting to determine the function of the corresponding genes. Finally, the previously identified major allergen-like encoding transcripts were also very highly expressed, further suggesting that this protein could be an important factor in causing locust allergy.

4. Conclusions

An increased knowledge on the gut transcriptome of locusts could significantly improve our understanding of insect gut physiology. Here, we successfully used self-self microarray hybridization with two different tissues to identify and annotate 2765 ESTs that were considered to be expressed specifically in the gut. It has to be noted that only brain and gut tissues were included in our analysis. It is likely that some transcripts that appear gut specific might also be expressed in some other tissues that were not included in our experimental setup. Nevertheless, the observed 'gut-specific' sequences had putative functions closely related with the physiological functions of the gut as a muscular digestive organ and as the first barrier against microorganisms and a wide range of toxins, proving the validity of our approach. A signal intensity ranking allowed us to identify several of the digestive enzymes, a major allergen like protein, and structural proteins as being expressed the most

in the gut. Interestingly, a large number of sequences highly expressed in the gut could not be functionally annotated, suggesting several proteins with yet unknown functions are active in the gut.

The microarray hybridizations were performed using all unique EST transcripts of locusts derived from different databases. However, when studied in detail, in several cases, high transcript sequence redundancy could be observed, suggesting the transcript assembly was of poor quality. This was most clear for highly abundant transcripts like for example serine protease transcripts, major allergen transcripts, chitin binding domain transcripts, and pacifastin related transcripts. Therefore, careful consideration is necessary when extrapolating numbers of transcripts to numbers of genes or proteins. In all cases, further detailed sequence analysis is recommended. Nevertheless, the data provide valuable insight into the expressed part of the genome in different tissues, and thus will proof very useful for future studies on gut physiology in insects.

5. Acknowledgements

The authors explicitly want to thank Roger Jonckers and Evelien Herinckx for carefully managing the locust culture. The authors also gratefully acknowledge the Agency for Science and Technology (IWT) (PhD fellowship obtained by J.S.), the KU Leuven Research Foundation (GOA/11/02) and the Research Foundation of Flanders (FWO) (FWO-G031112N) for financial support.

6. References

Aggrawal, K., Silverman, N., 2007. Peptidoglycan recognition in *Drosophila*. Biochem. Soc. Trans. 35, 1496–1500. doi:10.1042/BST0351496

Ahn, S.-J., Vogel, H., Heckel, D.G., 2012. Comparative analysis of the UDP-glycosyltransferase multigene family in insects. Insect Biochem. Mol. Biol. 42, 133–47. doi:10.1016/j.ibmb.2011.11.006

Arrudai, L., Vailes, L., Mann, J., Shannon, J., Fow, W., Vedvick, T., Hayden, Chapman, M., 1995. Molecular Cloning of a Major Cockroach (*Blattella germanica*) Allergen, Bla g 2. J. Biol. Chem. 270, 19563–19568.

Auerswald, L., Gäde, G., 2006. Endocrine control of TAG lipase in the fat body of the migratory locust, *Locusta migratoria*. Insect Biochem. Mol. Biol. 36, 759–768. doi:10.1016/j.ibmb.2006.07.004

Badisco, L., Huybrechts, J., Simonet, G., Verlinden, H., Marchal, E., Huybrechts, R., Schoofs, L., De Loof, A., Vanden Broeck, J., 2011. Transcriptome analysis of the desert locust central nervous system: production and annotation of a *Schistocerca gregaria* EST database. PLoS One 6, e17274. doi:10.1371/journal.pone.0017274

Breugelmans, B., Simonet, G., van Hoef, V., Van Soest, S., Vanden Broeck, J., 2009. Pacifastin-related peptides: structural and functional characteristics of a family of serine peptidase inhibitors. Peptides 30, 622–32. doi:10.1016/j.peptides.2008.07.026

Coates, B.S., Sumerford, D. V, Hellmich, R.L., Lewis, L.C., 2008. Mining an Ostrinia nubilalis midgut expressed sequence tag (EST) library for candidate genes and single nucleotide polymorphisms (SNPs). Insect Mol. Biol. 17, 607–20. doi:10.1111/j.1365-2583.2008.00833.x

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676. doi:10.1093/bioinformatics/bti610

Dermauw, W., Van Leeuwen, T., 2013. The ABC gene family in arthropods: Comparative genomics and role in insecticide transport and resistance. Insect Biochem. Mol. Biol. 45C, 89–110. doi:10.1016/j.ibmb.2013.11.001

Enayati, a a, Ranson, H., Hemingway, J., 2005. Insect glutathione transferases and insecticide resistance. Insect Mol. Biol. 14, 3–8. doi:10.1111/j.1365-2583.2004.00529.x

Fao, 2013. Edible insects. Future prospects for food and feed security, Food and Agriculture Organization of the United Nations.

Feyereisen, R., 2006. Evolution of insect P450. Biochem. Soc. Trans. 34, 1252–5. doi:10.1042/BST0341252

Garrett, M., Fullaondo, A., Troxler, L., Micklem, G., Gubb, D., 2009. Identification and analysis of serpin-family genes by homology and synteny across the 12 sequenced Drosophilid genomes. BMC Genomics 10, 489. doi:10.1186/1471-2164-10-489

Gore, J.C., Schal, C., 2005. Expression, production and excretion of Bla g 1, a major human allergen, in relation to food intake in the German cockroach, *Blattella germanica*. Med. Vet. Entomol. 19, 127–34. doi:10.1111/j.0269-283X.2005.00550.x

Grillo, L.A.M., Majerowicz, D., Gondim, K.C., 2007. Lipid metabolism in *Rhodnius prolixus* (Hemiptera: Reduviidae): Role of a midgut triacylglycerol-lipase. Insect Biochem. Mol. Biol. 37, 579–588. doi:10.1016/j.ibmb.2007.03.002

Hooper, S., Thuma, J., 2005. Invertebrate muscles: muscle specific genes and proteins. Physiol. Rev. 1001–1060. doi:10.1152/physrev.00019.2004.

Hughes, J., Vogler, A.P., 2006. Gene expression in the gut of keratin-feeding clothes moths (Tineola) and keratin beetles (Trox) revealed by subtracted cDNA libraries. Insect Biochem. Mol. Biol. 36, 584–92. doi:10.1016/j.ibmb.2006.04.007

Jasrapuria, S., Arakane, Y., Osman, G., Kramer, K.J., Beeman, R.W., Muthukrishnan, S., 2010. Genes encoding proteins with peritrophin A-type chitin-binding domains in *Tribolium castaneum* are grouped into three distinct families based on phylogeny, expression and function. Insect Biochem. Mol. Biol. 40, 214–27. doi:10.1016/j.ibmb.2010.01.011

Jeong, K.Y., Lee, H., Shin, K.H., Yi, M., Jeong, K., Hong, C.-S., Yong, T.-S., 2008. Sequence polymorphisms of major German cockroach allergens Bla g 1, Bla g 2, Bla g 4, and Bla g 5. Int. Arch. Allergy Immunol. 145, 1–8. doi:10.1159/000107460

Kanost, M.R., Jiang, H., Yu, X.-Q., 2004. Innate immune responses of a lepidopteran insect, *Manduca sexta*. Immunol. Rev. 198, 97–105.

Liu, X., Zhang, H., Li, S., Zhu, K.Y., Ma, E., Zhang, J., 2012. Characterization of a midgut-specific chitin synthase gene (LmCHS2) responsible for biosynthesis of chitin of peritrophic matrix in *Locusta migratoria*. Insect Biochem. Mol. Biol. 42, 902–10. doi:10.1016/j.ibmb.2012.09.002

Lopata, a L., Fenemore, B., Jeebhay, M.F., Gäde, G., Potter, P.C., 2005. Occupational allergy in laboratory workers caused by the African migratory grasshopper *Locusta migratoria*. Allergy 60, 200–5. doi:10.1111/j.1398-9995.2005.00661.x

Ma, Z., Yu, J., Kang, L., 2006. LocustDB: a relational database for the transcriptome and biology of the migratory locust (*Locusta migratoria*). BMC Genomics 7, 11. doi:10.1186/1471-2164-7-11

Merzendorfer, H., 2003. Chitin metabolism in insects: structure, function and regulation of chitin synthases and chitinases. J. Exp. Biol. 206, 4393–4412. doi:10.1242/jeb.00709

Morris, K., Lorenzen, M.D.M., Hiromasa, Y., Tomich, J.M., Oppert, C., Elpidina, E.N., Vinokurov, K., Jurat-Fuentes, J.L., Fabrick, J., Oppert, B., 2009. *Tribolium castaneum* larval gut transcriptome and proteome: a resource for the study of the coleopteran gut. J. Proteome Res. 8, 3889–3898. doi:10.1021/pr900168z

Pedra, J.H.F., Brandt, A., Westerman, R., Lobo, N., Li, H.-M., Romero-Severson, J., Murdock, L.L., Pittendrigh, B.R., 2003. Transcriptome analysis of the cowpea weevil bruchid: identification of putative proteinases and alpha-amylases associated with food breakdown. Insect Mol. Biol. 12, 405–12.

Ponnuvel, K.M., Nakazawa, H., Furukawa, S., Asaoka, A., Ishibashi, J., Tanaka, H., Yamakawa, M., 2003. A lipase isolated from the silkworm *Bombyx mori* shows antiviral activity against nucleopolyhedrovirus. J. Virol. 77, 10725–10729. doi:10.1128/JVI.77.19.10725-10729.2003

Qin, G., Jia, M., Liu, T., Xuan, T., Yan Zhu, K., Guo, Y., Ma, E., Zhang, J., 2011. Identification and characterisation of ten glutathione S-transferase genes from oriental migratory locust, *Locusta migratoria manilensis* (Meyen). Pest Manag. Sci. 67, 697–704. doi:10.1002/ps.2110

Qin, G., Jia, M., Liu, T., Zhang, X., Guo, Y., Zhu, K.Y., Ma, E., Zhang, J., 2013. Characterization and functional analysis of four glutathione S-transferases from the migratory locust, *Locusta migratoria*. PLoS One 8, e58410. doi:10.1371/journal.pone.0058410

Qin, G., Jia, M., Liu, T., Zhang, X., Guo, Y., Zhu, K.Y., Ma, E., Zhang, J., 2012. Heterologous expression and characterization of a sigma glutathione S-transferase involved in carbaryl detoxification from oriental migratory locust, *Locusta migratoria manilensis* (Meyen). J. Insect Physiol. 58, 220–7. doi:10.1016/j.jinsphys.2011.10.011

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120. doi:10.1093/nar/gki442

Randall, T. a, Perera, L., London, R.E., Mueller, G. a, 2013. Genomic, RNAseq, and molecular modeling evidence suggests that the major allergen domain in insects evolved from a homodimeric origin. Genome Biol. Evol. 5, 2344–58. doi:10.1093/gbe/evt182

Reichhart, J.M., Gubb, D., Leclerc, V., 2011. The *Drosophila* serpins: Multiple functions in immunity

and morphogenesis, Methods in Enzymology. doi:10.1016/B978-0-12-386471-0.00011-0

Schuler, M. a, 2011. P450s in plant-insect interactions. Biochim. Biophys. Acta 1814, 36–45. doi:10.1016/j.bbapap.2010.09.012

Simonet, G., Claeys, I., Van Soest, S., Breugelmans, B., Franssens, V., De Loof, A., Vanden Broeck, J., 2004. Molecular identification of SGPP-5, a novel pacifastin-like peptide precursor in the desert locust. Peptides 25, 941–50. doi:10.1016/j.peptides.2004.03.005

Simonet, G., Claeys, I., Vanderperren, H., November, T., De Loof, a, Vanden Broeck, J., 2002. cDNA cloning of two different serine protease inhibitor precursors in the migratory locust, *Locusta migratoria*. Insect Mol. Biol. 11, 249–56.

Spit, J., Breugelmans, B., van Hoef, V., Simonet, G., Zels, S., Vanden Broeck, J., 2012. Growth-inhibition effects of pacifastin-like peptides on a pest insect: the desert locust, *Schistocerca gregaria*. Peptides 34, 251–7. doi:10.1016/j.peptides.2011.06.019

Spit, J., Zels, S., Dillen, S., Holtof, M., Wynant, N., Vanden Broeck, J., 2014. Effects of different dietary conditions on the expression of trypsin- and chymotrypsin-like protease genes in the digestive system of the migratory locust, *Locusta migratoria*. Insect Biochem. Mol. Biol. 48, 100–9. doi:10.1016/j.ibmb.2014.03.002

Sui, Y.-P., Wang, J.-X., Zhao, X.-F., 2008. Effects of classical insect hormones on the expression profiles of a lipase gene from the cotton bollworm (*Helicoverpa armigera*). Insect Mol. Biol. 17, 523–9. doi:10.1111/j.1365-2583.2008.00820.x

Theopold, U., Rissler, M., Fabbri, M., Schmidt, O., Natori, S., 1999. Insect glycobiology: a lectin multigene family in *Drosophila melanogaster*. Biochem. Biophys. Res. Commun. 261, 923–7. doi:10.1006/bbrc.1999.1121

Treangen, T.J., Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46. doi:10.1038/nrg3117

Verlinden, H., Badisco, L., Marchal, E., Van Wielendaele, P., Vanden Broeck, J., 2009. Endocrinology of reproduction and phase transition in locusts. Gen. Comp. Endocrinol. 162, 79–92.

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., Hao, S., Chen, B., Ma, Z., Yu, D., Xiong, Z., Zhu, Y., Fan, D., Han, L., Wang, B., Chen, Y., Wang, J., Yang, L., Zhao, W., Feng, Y., Chen, G., Lian, J., Li, Q., Huang, Z., Yao, X., Lv, N., Zhang, G., Li, Y., Wang, J., Wang, J., Zhu, B., Kang, L., 2014. The locust genome provides insight into swarm formation and long-distance flight. Nat. Commun. 5, 2957. doi:10.1038/ncomms3957

Yan, N., 2013. Structural advances for the major facilitator superfamily (MFS) transporters. Trends Biochem. Sci. 38, 151–9. doi:10.1016/j.tibs.2013.01.003

Zhang, J., Li, D., Ge, P., Yang, M., Guo, Y., Zhu, K.Y., Ma, E., Zhang, J., 2013. RNA interference revealed the roles of two carboxylesterase genes in insecticide detoxification in *Locusta migratoria*. Chemosphere 93, 1207–1215. doi:10.1016/j.chemosphere.2013.06.081

Zhang, J., Zhang, J., Yang, M., Jia, Q., Guo, Y., Ma, E., Zhu, K.Y., 2011. Genomics-based approaches to screening carboxylesterase-like genes potentially involved in malathion resistance in oriental migratory locust (*Locusta migratoria manilensis*). Pest Manag. Sci. 67, 183–190.

Zhang, Z., Peng, Z.-Y., Yi, K., Cheng, Y., Xia, Y., 2012. Identification of representative genes of the central nervous system of the locust, *Locusta migratoria manilensis* by deep sequencing. J. Insect Sci. 12, Article 86. doi:10.1673/031.012.8601

Fig. 1. Scatter plot of the Cy3- and Cy5-intensity for all unique transcripts after self-self hybridization. RNA-samples were labeled with Cy5 or Cy3 and were subsequently hybridized in a dual color manner for both brain (A) and gut (B). Axes are scaled logarithmically. Blue colored spots are internal spiked-in controls, green colored spots are transcripts that showed an intensity of brain/gut > 20, and red colored spots showed an intensity of gut/brain > 20.

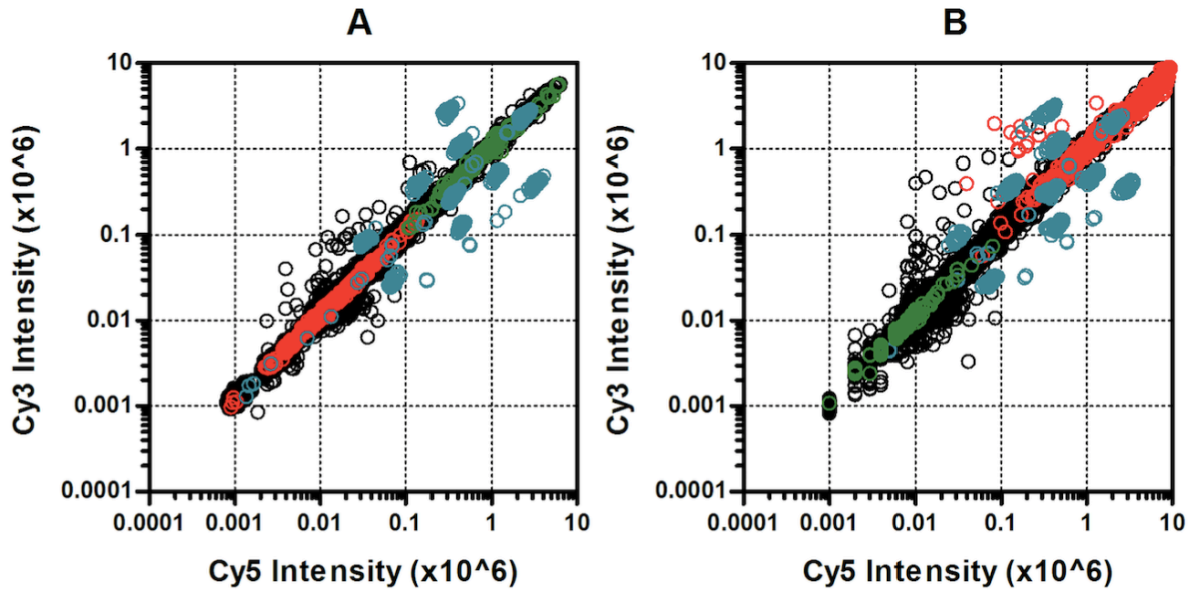

Fig. 2. Multiple sequence alignment of a major allergen-related protein identified in *L. migratoria* (L.m majA.) and two variants of *Blatella germanica* major allergen Bla g 1. Bla g 1.02 and Bla g 1.0101 accesion numbers AAD13531, AAD13530. Alignment was carried out in MUSCLE. Cutoff for shading was set at 100%. Shaded in black are identical amino acid residues, shaded in grey are similar amino acid residues

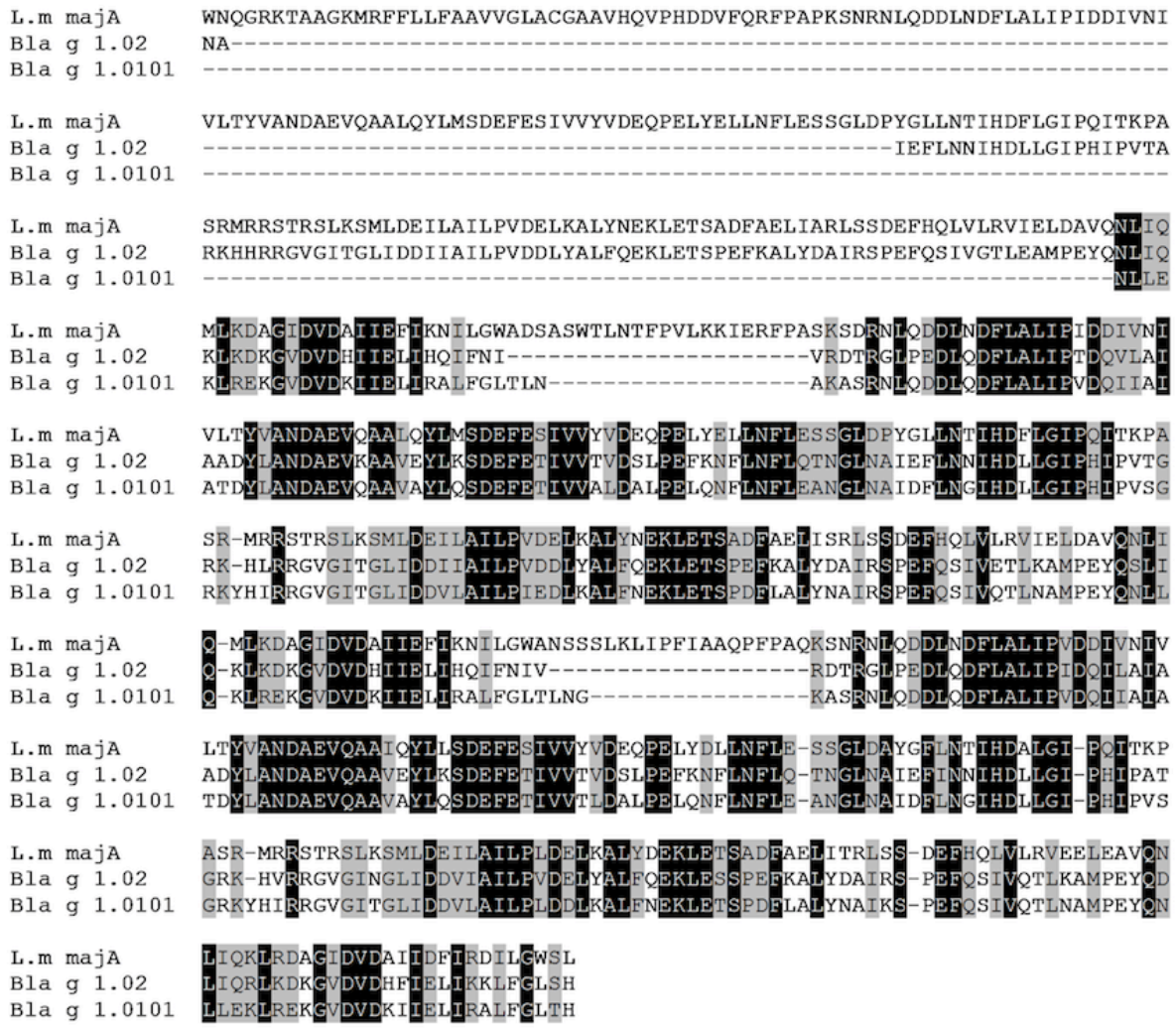

Fig. 3. Histogram based on relative intensity ranking. The relative ranking is calculated relative to the maximum intensity level, set at 100%. The majority of transcripts show low expression levels, <10% of the maximum intensity.