Transcriptomic and Proteomic Analysis of *Arion vulgaris*—Proteins for Probably Successful Survival Strategies?

Tanja Bulat[1], Roman Smidak[1], Fernando J. Sialana[1], Gangsoo Jung[1], Thomas Rattei[2], Martin Bilban[3], Helmut Sattmann[4], Gert Lubec[5]*, Jana Aradska[1]*

1 Department of Pediatrics, Medical University of Vienna, Vienna, Austria, 2 Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria, 3 Department of Laboratory Medicine and Core Facility Genomics, Medical University of Vienna, Vienna, Austria, 4 Third Zoological Department, Museum of Natural History Vienna, Vienna, Austria, 5 Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria

* gert.lubec@univie.ac.at (GL); jana.aradska@gmail.com (JA)

Abstract

The Spanish slug, *Arion vulgaris*, is considered one of the hundred most invasive species in Central Europe. The immense and very successful adaptation and spreading of *A. vulgaris* suggest that it developed highly effective mechanisms to deal with infections and natural predators. Current transcriptomic and proteomic studies on gastropods have been restricted mainly to marine and freshwater gastropods. No transcriptomic or proteomic study on *A. vulgaris* has been carried out so far, and in the current study, the first transcriptomic database from adult specimen of *A. vulgaris* is reported. To facilitate and enable proteomics in this non-model organism, a mRNA-derived protein database was constructed for protein identification. A gel-based proteomic approach was used to obtain the first generation of a comprehensive slug mantle proteome. A total of 2128 proteins were unambiguously identified; 48 proteins represent novel proteins with no significant homology in NCBI non-redundant database. Combined transcriptomic and proteomic analysis revealed an extensive repertoire of novel proteins with a role in innate immunity including many associated pattern recognition, effector proteins and cytokine-like proteins. The number and diversity in gene families encoding lectins point to a complex defense system, probably as a result of adaptation to a pathogen-rich environment. These results are providing a fundamental and important resource for subsequent studies on molluscs as well as for putative antimicrobial compounds for drug discovery and biomedical applications.

Introduction

Arion vulgaris as the only terrestrial gastropod, is considered among the 100 worst invasive species in Europe. The spread of invasive species, including slugs, is an increasing problem worldwide with an important economical, ecological, health and social impact. The slug *A. vulgaris* is considered a serious pest, both in agriculture and private gardens. Moreover *A. vulgaris*

may act as a vector for some pathogenic bacteria, like *Listeria monocytogenes* [1], *Clostridium botulinum* [2],[3] and host species for some parasites, as *Angiostrongylus vasorum* that can cause serious and potentially fatal disease in dogs and other canids [4]. *A. vulgaris* belongs to a species complex *Arion ater* that can only be distinguished by dissecting their reproductive organs.

Currently, transcriptomic studies on molluscs have been restricted to an ecological framework in the marine and fresh water gastropods. However, no transcriptomic or proteomic study has been carried out so far in *A. vulgaris*. Next-generation sequencing has offered a powerful and cost-efficient technique for the generation of transcriptomic datasets in non-model species using diverse platforms such as the Illumina HiSeq, Roche 454, Pacific Biosystems, and Applied Biosystems SOLiD. Several non-model organisms have been characterized by transcriptome sequencing [5–7], which has provided a better understanding of these species.

Since mRNA expression does not necessarily reflect changes at the protein level, complementary proteomic studies can yield more comprehensive biological insight. Mass spectrometry analysis allows large proteome surveys and comparative analysis, where quantities of hundreds or thousands of proteins can be compared between various conditions. However, the lack of a genomic resource for these animals is the major reasons for restricting proteomic applications.

In the current study, the Illumina HiSeq 2000 platform has been used to generate an *A. vulgaris* transcriptome-based protein database. The transcriptome data generated in this study provide a first comprehensive and valuable genomic resource for future research on this slug. With the objective to obtain insight into the slug proteome gel-based proteomic, analysis from the *A. vulgaris* mantle was performed.

Methods

Collection of Arion vulgaris

A. vulgaris specimens were collected in May 2014 in a location (48° 19′ 40″ N, 16° 12′ 34″ E) in Woerdern, Austria. A total of 15 specimens ranging from 7 to 8 g were collected. Species identification was carried out with the assistance of the Zoological Department of Museum of Natural History in Vienna (H.S.). According to national law no permission is required to collect and use invertebrates.

Transcriptome sequencing and assembly

Isolation of RNA was performed with the RNeasy kit (Qiagen, Hilden, Germany). Following removal of the gut, two specimens of *Arion vulgaris* were immersed in liquid nitrogen and pulverized with the help of a mortar and pestle. Pooled total RNA was subjected to RNA-Seq following the mRNA sequencing protocol provided by New England Biolabs (NEBNext kit, NEB, Frankfurt, Germany). The first step in the workflow involved purifying the poly(A)-containing mRNA molecules using two rounds of poly(T)-oligo–attached magnetic beads. After purification, mRNAs were fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments were copied into first-strand complementary DNA (cDNA) using reverse transcriptase and random primers. This was followed by second-strand cDNA synthesis using DNA polymerase I and RNase H (NEB, Frankfurt, Germany). These cDNA fragments subsequently went through an end-repair process, the addition of a single A base and ligation of the adaptors. The products were then purified and enriched with 15 cycles of PCR (with a size distribution of ~200–250 bp) to create the final cDNA library. Finally, the adaptor-ligated DNA was sequenced for 100 cycles on a HiSeq2000 sequencing system (Illumna, San Diego, USA) in a paired-end manner according to the manufacturer's

instructions. The generated reads were filtered and trimmed by prinseq-lite using parameters trim_qual_right 30, min_qual_mean 30, min_len 70, trim_tail_right 6, trim_tail_left 6, and rRNA sequences were removed using SortMeRNA 1.99 [8]. Digital normalization (maximal coverage 30) and transcript assembly were performed using Trinity pipeline (release 2013-02-25, [9]).

Transcriptome annotation

Protein-coding sequences were identified in the assembly using the TransDecoder tool and PFAM version 27.0. Sequence similarities were obtained by blastx searches (cut-off E-value 1E-6) of the contigs and predicted CDS, respectively, against the NCBI non-redundant protein database (version July 2014). The assembled transcripts were further annotated with GO terms, PFAM protein domains and Enzyme Commision (EC) numbers using web platform FastAnnotator [10] with the default search parameters. Duplicates and substrings were removed from the translated sequence dataset and resulting protein sequences were post-assembled using an in-house script to reduce redundancy. Final assembly containing 47,451 non-redundant protein sequences were used as an expression dataset for MS protein identification.

Protein extraction

Individuals of *A. vulgaris* were frozen immediately after harvest in liquid nitrogen and stored at -80°. Mantles were dissected directly before protein extraction procedure and homogenized in urea buffer (7 M urea, 2 M thiourea, 4% CHAPS, 2 M Tris (pH 8.8), 0.5% Carrier (3–11 pH), 1% DTT, 1 mM EDTA, 1 mM PMSF, Protease Inhibitor C). Protein samples were cleaned using a 2-DE clean up kit (BioRad, Hercules, CA, USA) and protein concentration was esti-mated by the Pierce 660 kit (ThermoScientific, Rockford, USA).

Two dimensional gel electrophoresis (2-DE) and nano-LC-ESI-MS/MS analysis

2-DE was performed as reported previously with minor modifications [11–13]. 750 µg of pro-teins were loaded on 18 cm 3–11 linear IPG strips (GE Healthcare). IEF was performed in a IPGphore electrophoretic system (GE Healthcare, Uppsala, Sweden) using a protocol with gradually increasing voltage from 200 to 8,000 V at 4 V/min. Prior to second dimension elec-trophoresis, IPG strips were incubated in 10 ml of equilibration buffer (50 mM Tris-HCl, pH 8.8, 6 M urea, 30% glycerol, 2% SDS and 0.002% bromophenol blue) containing 1% DTT for 15 min with gentle shaking followed by incubation in 10 mL of equilibration buffer with 4% iodoacetamide. The second-dimensional separation was performed on 10–16% gradient SDS-PAGE gels. After fixation for 4 h in 50% methanol and 10% acetic acid, gels were stained overnight with the colloidal coomassie blue (Novex life technologies, Invitrogen, CA). Molecular masses were determined by comparison with precision protein standard markers (#1610373, BioRad, Hercules, CA, USA) spanning the 10 to 250 kDa molecular weight range. All spots were excised from 2-DE gels and digested with trypsin and/or chymotrypsin. The gel pieces were cut into small pieces and washed with 50% ACN in 10mM ammonium bicarbonate for 20 min with vortexing and then 20 min with 10mM ammonium bicarbonate. These two steps were repeated until the gel was completely destained. 100% ACN was added for 10min and gel pieces were dried completely using a SpeedVac concentrator for 30 min at 30°C. Cyste-ine residues were reduced with 10 mM DTT at 56°C for 30 min and alkylated with 55 mM iodoacetamide for 45 min at 22°C. After washing with 10 mM ammonium bicarbonate and dehydratation with 100% ACN, proteins were digested with 12.5 ng/µL trypsin (Promega,

Mannheim, Germany) solution buffered in 10 mM ammonium bicarbonate for 16 h (overnight) at 37°C. The supernatant was transferred to new LoBind 0.5 mL tubes and peptides were extracted with 1% formic acid and then with 15% ACN/0.1% formic acid [14]. 20 uL of extracted peptides were analyzed by nano-LC-ESI-(CID/ETD)-MS/MS essentially as described previously [14]. The HPLC used was an Ultimate 3000 system (Dionex Corporation, Sunnyvale, CA) equipped with a PepMap100 C-18 trap column (300 µm× 5 mm) and PepMap100 C-18 analytical column (75 µm× 250 mm). The gradient was (A = 0.1% formic acid in water, B = 0.08% formic acid in ACN) 4–30% B from 0 to 105 min, 80% B from 105 to 110 min, 4% B from 110 to 125 min. The Amazon speed ETD (Bruker Daltonics, Bremen, Germany) was used to record peptide spectra over the mass range of m/z 400–1,400, and MS/MS spectra in information-dependent data acquisition over the mass range of m/z 100–2,800. Repeatedly, MS spectra were recorded followed by three data-dependent CID MS/MS spectra. An active exclusion of 0.4 min after two spectra was used to detect low abundant peptides. The voltage between ion spray tip and spray shield was set to 1,400 V. Drying nitrogen gas was heated to 150°C and the flow rate was 3 L/min. The collision energy was set automatically according to the mass and charge state of the peptides chosen for fragmentation. Multiple charged peptides were chosen for MS/MS experiments due to their good fragmentation characteristics. MS/MS spectra were interpreted and peak lists were generated by Data Analysis 4.1 (Bruker Daltonics, Bremen, Germany).

One dimensional (1-D) SDS-PAGE and LC-MS/MS

Although 2-DE followed by LC-MS/MS is a proven method for protein separation and identification, it suffers from poor sensitivity, poor representation of very acidic or basic proteins and low solubility of hydrophobic proteins. Therefore, we also separated proteins by 1-D SDS-PAGE coupled with high resolution LC−MS/MS. 50 µg of extracted proteins were mixed with Laemmli buffer (150 mM Tris HCl pH 6.8, 300 mM DTT, 6% SDS, 0.3% bromophenol blue, 30% glycerol) at a 1:3 volume ratio. Proteins in each sample were separated by 10% SDS-PAGE and stained by Blue silver [15]. After destaining with Milli-Q water, gels were cut into 12 slices with proteins of different mass. Proteins underwent in-gel trypsin digestion as above and were analyzed using a LTQ-Orbitrap Velos (ThermoFisher Scientific, Waltham, MA, USA) coupled with nano-LC (Dionex Ultimate 3000) as described previously [16].

Database search

Data generated from Amazon speed ETD for each of the 2-DE protein spots were analyzed by searching the mRNA-derived *Arion vulgaris* database (47,451 sequences) including 115 commonly observed contaminants with Mascot Search enigma (version 2.4) using the Mascot Daemon interface (Matrix Science, London, UK). Detailed search criteria were used as follows; enzyme: trypsin or chymotrypsin with a maximum of two missing cleavage sites; search mode: MS/MS ion search with decoy database search included; fixed modification: carbamidomethylation (C); variable modification: oxidation (M); search mode: MS/MS ion search with decoy database search included; peptide mass tolerance 0.35 Da; MS/MS mass tolerance ± 0.35 Da.

For the mass spectrometry data generated from the LTQ Orbitrap Velos, the MS raw files were first processed with Proteome Discoverer 1.4 (ThermoFisher Scientific, IL, USA) to generate separate Mascot generic files. The database search was performed using Mascot version 2.4 (MatrixScience, London, UK) against the combined database as mentioned above. The search criteria were 10 ppm for precursor and 0.5 Da for fragments; search mode: MS/MS ion search with decoy database search included; fixed modification: carbamidomethyl (cysteine); variable modification: oxidation (methionine).
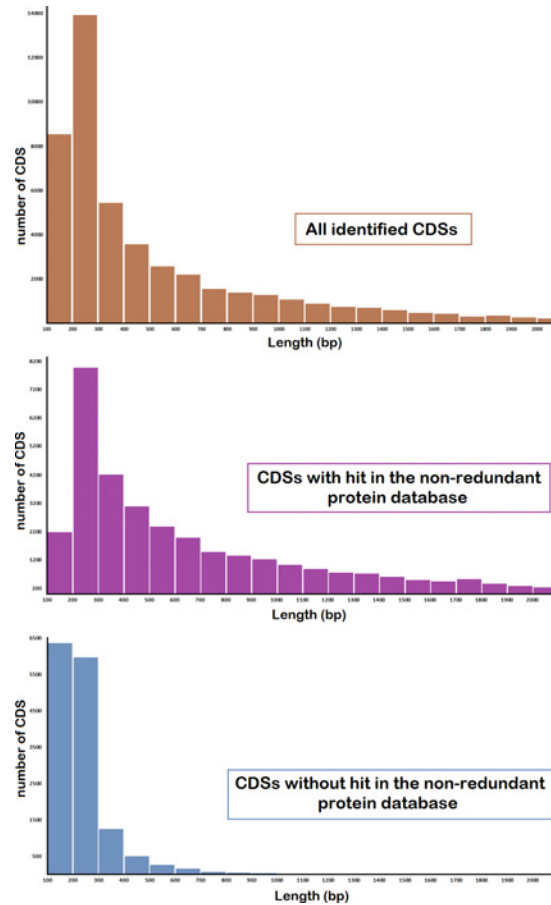
Fig 1. Protein-coding sequences (CDS) distribution showing the majority of sequences in the range from 200 to 300 bp.

Peptide identifications were filtered to a < 1% false discovery rate (FDR) using the target-decoy strategy [17]. Acceptance parameter for protein identification was a minimum of one unique peptide and two distinct peptides.

Results

Transcriptome of whole body

The slug-specific database was constructed using a RNAseq and *de novo* assembly strategy. A cDNA library from the adult slug was generated and using Ilumina HiSeq 2000 technology an approximate number of 339 millions of paired end reads was obtained. *De novo* transcript assembly performed by Trinity pipeline [9] resulted in 136,406 contigs with average lengths of 671.04 bp and N50 of 971 bp, meaning that 50% of the assembled sequences are 971 bp long or longer. 53,523 protein-coding sequences (CDS) were identified using the TransDecoder tool from the Trinity package with PFAM version 27.0. Fig 1 shows the distribution of CDS annotated contig lengths. All CDS annotated contigs of the final assembly were subjected to blastx analysis against the NCBI non-redundant protein database (nr) using the web platform Fastannotator [10] with a cut-off value of 1E-6. Duplicates and substrings were removed from the translated sequence dataset and resulting protein sequences were post-assembled using an in-house script to reduce redundancy. The final assembly contains 47,451 non-redundant protein
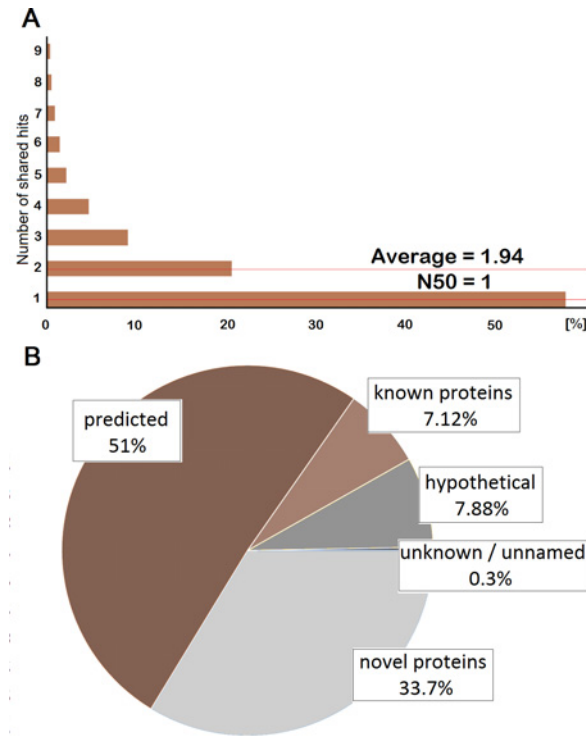
Fig 2. **Characteristics of sequences from *A. vulgaris* transcriptome.** (a) Distribution of shared top-hits. (b) Classification of top-hits in the current transcriptome.

sequences. On average 1.94 sequences of the final assembly share the same hit against the NCBI nr database (Fig 2A), due to sequencing errors and/or potential isoforms. As a result, 31,463 sequences (66.3%) were assigned to at least one protein and 15,988 (33.7%) transcripts did not have a significant hit in the NCBI nr database probably because of the lack of molecular data of mollusc species. Among 31,463 matches, only 3,380 (10.7%) were known proteins, 24,217 are predicted (77%), 3,739 are hypothetical (11.9%), and 127 are either uncharacterized, unknown or unnamed proteins (0.4%) (Fig 2B). Sequences without significant similarity to sequences present in the nr database are being referred as putative novel genes. These include taxonomically constricted genes derived from ancestral genes or appeared *de novo* from non-coding sequences [18]. The high number of transcripts lacking detectable homology with protein sequences described in the database might be also due to the high number of short transcripts (147–300 bp) in the final assembly. Of the 47,451 transcripts in the final assembly, 46.5% (22,055 transcripts) are shorter than 300 bp (Fig 1). It has been shown that non-coding RNAs [19] as well as small open reading frames are potential sources of sequences lacking detectable homology with protein sequences in nr databases. A high number of putative novel genes was identified but additional genetic as well as proteomic studies are needed to annotate them correctly and this was one of the aims of the current proteomic study described below.

39% (18,561) of all sequences were functionally annotated with Gene Onthology (GO) terms using BLAST2GO tool of Fastannotator [20]. The results were summarized to the categories "biological process" (33.9%), "molecular function" (28.2%), and "cellular component" (34.8%). Within the biological process classification, metabolic processes (66%), biological regulation (43%) and response to stimulus (29%) were the most representative (Fig 3). In the molecular function category, the highest percentage of GO terms corresponded to "binding"
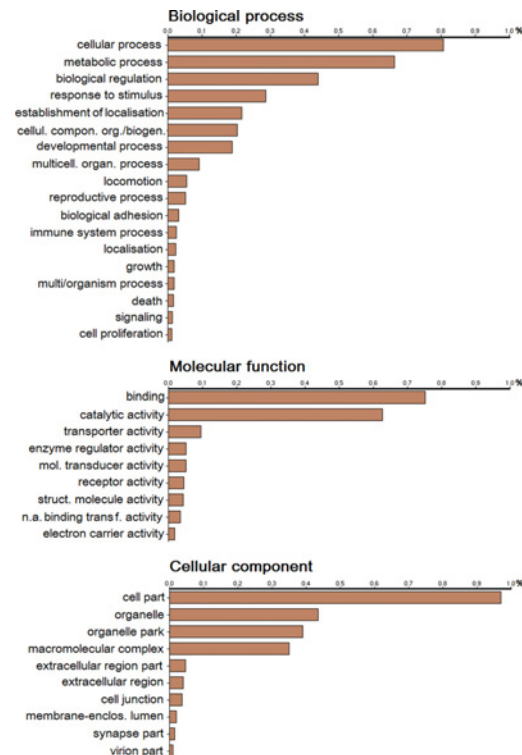
Fig 3. Gene Ontology (GO) analyses of the *A. vulgaris* transcriptome.

(75%), "catalytic activity" (62%) and "transporter activity" (9.8%) (Fig 3). For the cellular component classification, the cluster sizes of "organelle/organelle part", "macromolecular complex" and "extracellular region" were relatively large (Fig 3). Domain analysis against the Pfam database identified 16,994 of entries to have at least one domain of which 471 sequences without blastx hit were further annotated. The zinc-finger double domain zf-H2C2_2 (PF13465) was the most abundant domain followed by domain of unknown function DUF4200 (PF13863) and ankyrin repeat Ank_5 (PF13857) (S1 Table). Ankyrin domains are among the most common structural motifs in known proteins. 1,796 of transcripts were assigned with EC (Enzyme Commission) numbers.

We also analyzed taxonomic distribution of the best-match species of the annotable slug sequences. The highest proportion of the best blastx hits was found for *Aplysia californica* (65%), *Crassostea gigas* (8.7%) and *Capitella teleta* (2.5%) (Fig 4). Despite the limited number of molluscan sequences in the public databases, more than 80% of the sequences best matched to molluscan species. Assignment of transcripts also to chordate, plants, fungi, bacteria and viruses was probably due to more functional data on these species as well as contamination from environmental organisms during sample preparation has to be also considered.

Proteome of slug mantle

With the objective to gain insight into the slug proteome gel-based proteomic analysis from the *A. vulgaris* mantle was performed. In total, 356 spots corresponding to 833 unique proteins were clearly resolved in 2-DE reproducible gels (S2 Table). Fig 5 shows the representative 2-DE gel covering pH 3–10. Most of the spots were located between pH 5 to 9 over a broad range of MW (10–250 kDa). Among these proteins, 814 were assigned to protein sequence with significant hit in NCBI nr database and 19 represent a proteins with no significant homology in nr
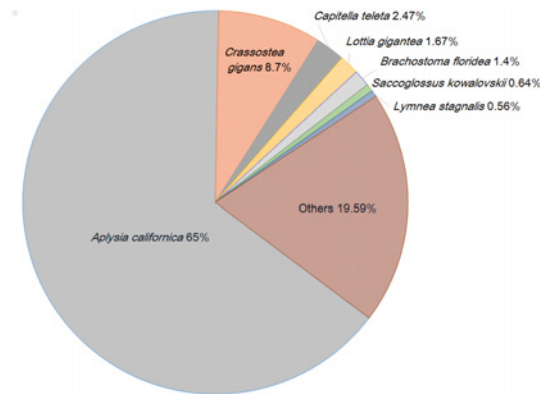
Fig 4. Species distribution of best blastx hits against the nr database.

doi:10.1371/journal.pone.0150614.g004

database. Several putative novel proteins have been identified in several spots representing isoforms with different posttranslational modifications deduced from different position in the gel. Because of the limitation of 2-DE for very small and very large proteins, alkaline proteins and hydrophobic proteins, we also analyzed the protein samples by 1-D SDS-PAGE coupled with LC–MS/MS. 2011 proteins were identified in two biological replicates from 1-D SDS-PAGE (S3 Table). Combining the results from 1-D and 2-D gel separation methods, a total of 2129 proteins were identified with minimum of one unique peptide and two distinct peptides (S4 Table). The overall false-positive assignment was estimated around 1% by a target-decoy search strategy suggesting high quality of our dataset. 1770 identified proteins (83.1%) had matches to Gene Ontology (GO) term annotations and were categorized to molecular function (1647 proteins), cellular component (1337 proteins), and biological process (1596 proteins). Of these, 23.7% have functions that are associated with cytoskeletal protein binding. The mantle, made up of muscle and skin, hence many of the cytoskeletal and muscle-related proteins were observed. The highest number of identified peptides in the combined proteomic approach were observed for paramyosine, myosine, actin and arginine kinase. Arginine kinase belongs to a class of kinases that play a crucial role in invertebrates in the maintenance of ATP levels by the phosphorylation of phosphagens which then serve as a high energy source for rapid ATP replenishment [21]. The GO terms with the largest number of assigned sequences in the biological process category were small molecule metabolic process, catabolic process, biosynthetic process and response to chemical stimulus. 1882 (88.3%) had at least one PFAM domain match and 398 (18.7%) an EC number. The RNA binding motifs RRM_6 and RRM_1 (PF14259 and PF00076) were the most abundant domains followed by Tropomyosin_1 domain (PF12718) and EF hand domains (PF13833, PF13405). (S4 Table).

Combined transcriptomic and proteomic analysis revealed the presence of a high number of sequences sharing carbohydrate-recognition domains, like lectin domains, proteins with C1q domain and toxin-like proteins were present as well (S5 Table).

Discussion

The combined transcriptomic and proteomic approach allowed the identification of a great number of new transcripts and proteins related to metabolic, functional and cellular components. These transcriptomics and proteomics data are by far the most comprehensive among terrestrial gastropods. Moreover, one may speculate that innate immune response-related and defense elements identified in this study may contribute to uncover the success of *A. vulgaris* and indeed, *A.vulgaris* is spreading Europe-wide and is well protected against tentative
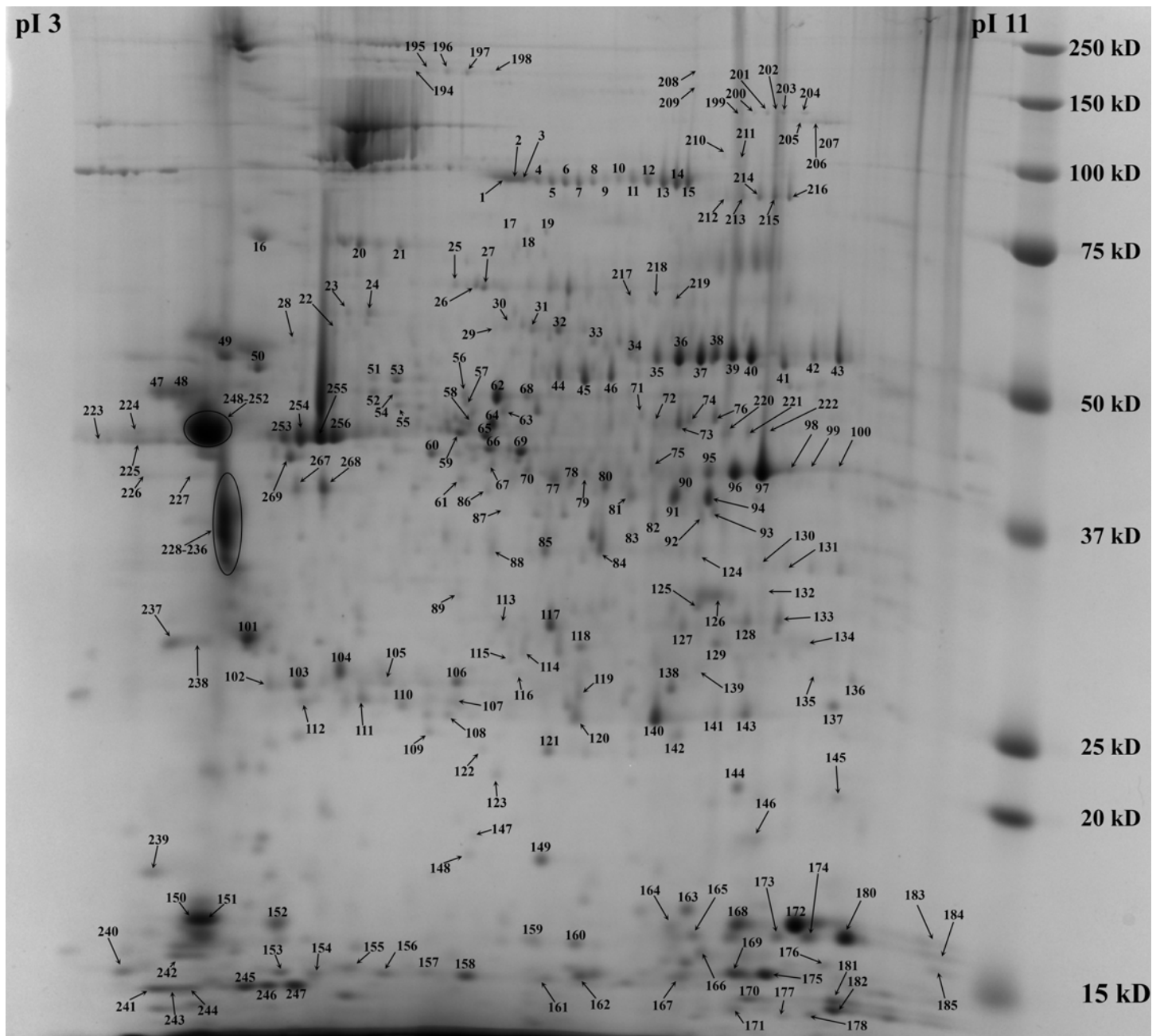
Fig 5. Representative 2-DE gel image from slug mantle showing protein spots selected for mass spectrometry analysis. The identified proteins are listed in S2 Table.

predators and pathogens [22–24]. 2129 *A. vulgaris* mantle proteins were identified through this study, markedly expanding the list of known proteins in terrestrial gastropods, demonstrating the powerfulness of gel-based approach (2-DE and 1-D SDS-PAGE) combined with mass spectrometry identification (LC-ESI-MS/MS). Of the two protein identification approaches used in this study, the 1-D approach has shown markedly higher sensitivity as demonstrated by the significantly higher number of detected proteins. The 2-DE method is useful in separating protein isoforms with different isoelectric point and protein mobility [25].
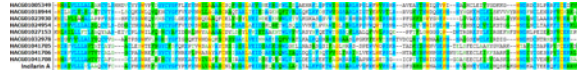
Fig 6. Domain architecture of C-lectin-like molecules found in the transcriptome of *A. vulgaris*.

doi:10.1371/journal.pone.0150614.g006

Slugs belonging to the one of the most successful phyla, Mollusca, lack clear evidence of adaptive immunity. They contend with many pathogens, including bacteria, fungi, viruses and several lineages of specialized eukaryotic parasites, and how they without adaptive immunity effectively and sufficiently defend themselves, is not well understood. The innate immune system in gastropoda is provided by physical barriers (e.g., shell, skin and epithelium), an evolutionary archaic mechanism of molecular "self-non-self" recognition, as well as by a variety of denfense-related factors. The soft, moist slug body protected by a ciliated, mucus-producing epithelium that provides an initial physical barrier to colonization by the pathogens, plays a crucial role in host defense. The diversification of immune systems during evolution involves the expansion of particular gene families in given phyla. Analysis of transcriptomic data from *A. vulgaris* shows a comprehensive repertoire of genes related to innate immunity including many associated pattern recognition, effector proteins and cytokine-like proteins, such as lectins, complement-like proteins, peptidoglycan-recognition proteins (PGRPs), lipopolysaccharide and β1,3-glucan-binding proteins, fibrinogen-related proteins (FREPs), pore-forming membrane attack/perforin (MACPF) domain proteins and toxin-like proteins that indicates an extraordinary complexity of immune system with a high degree of pathogen specificity and immune-priming.

Lectin-like proteins

Lectins play an important role in „self-non-self"recognition and clearance of invaders in gastropods. The slug repertoire of lectin-like proteins bearing conserved carbohydrate-recognition domains (CRDs) is highly diversified, including C-type lectin family, galectin, malectin, H-type lectins, L-type lectins, Ricin-B lectins and calnexin. The diversity of CRDs in lectins may reflect the different functions that the proteins perform. In total, lectin-associated functions may be attributed to 149 transcripts, 37 were unambigously identified at the protein level (S5 Table). These lectin-like proteins show low overlap between species of invertebrates, suggesting relatively rapid evolution of pattern recognition proteins involved in innate immunity. Most of transcripts/proteins described belong to tree lectin families, C-type, R-type and H-type.

C-type lectins

C-type lectins are a superfamily of diverse proteins able to bind specific carbohydrates in a Ca^{2+}-dependent manner. The CRDs contain characteristic double-loop structures stabilized by two highly conserved disulfide bridges located at the bases of the loops [26]. The second loop is structurally and evolutionarily flexible and is involved in Ca^{2+}-dependent carbohydrate binding. In total, 47 transcripts encoding proteins with C-type lectin CRDs have been found and 14 were identified also at the protein level. 17 of 21 complete transcripts have a predicted signal peptide indicating that they are secreted by cells to exert their functions (S5 Table) (Fig 6). Fig 7 shows remarkable sequence variability within incilarin-like proteins identified in our study. Incilarins are C-type lectins originaly isolated from the water-soluble fraction of the body surface mucus of the land slug, *Incilaria fruhstorfer*, possessing hemagglutination activity [27].

R-type lectins

The R-type lectins are members of a superfamily of proteins which contain a carbohydrate-recognition domain (CRD) structurally similar to the CRD in ricin. All identified R-type lectins,
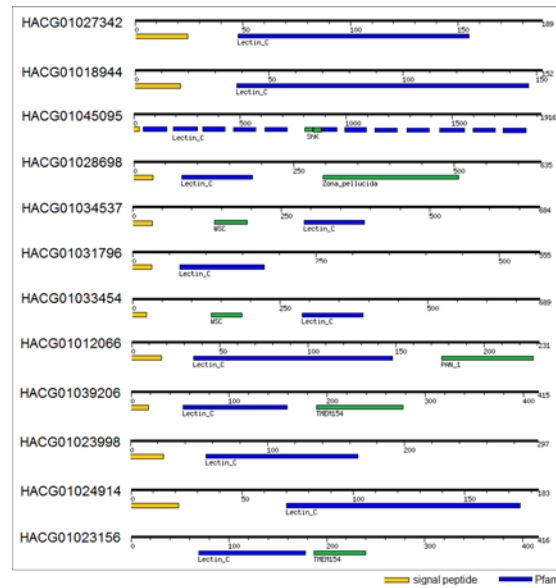
Fig 7. Alignment of amino acid sequences of incilarin-like proteins from *A. vulgaris*.

are chimeric proteins consisting of an A chain with enzymatic activity (galactosyltransferase) linked through a disulfide bridge with a B chain with lectin activity (Fig 8). An analysis of the transcriptome predicts 32 unigenes coding proteins of R-type lectins. R-type lectine-like proteins were not identified at the protein level probably due to tissue-specific expression.

H-type lectins

The H-type lectins are proteins which contain a carbohydrate-recognition domain (CRD) structurally similar to *Helix pomatia* agglutinin (HPA) [28]. It has been shown that HPA is part of the innate immunity system of *H. pomatia* and a component of perivitelline fluid protecting fertilized eggs from bacteria. We identified 12 transcripts, 7 at the protein level. H-type lectins from *A. vulgaris* have a similar size and are predicted to share the same hexameric arrangement achieved by conserved cysteine residues involved in the intermolecular disulfide bridge formation (Figs 8 and 9).

Galectins

Galectins represent a family of structurally-related diverse lectins with carbohydrate binding specificity primarily to β-galactoside residues. Screening of the *A. vulgaris* transcriptome
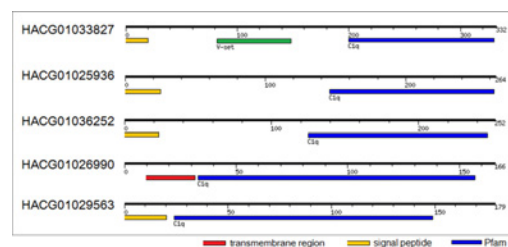

Fig 8. Domain architecture of galectins, H-type lectins, R-type lectins, L-type lectins and calnexin-like proteins found in the transcriptome of *A. vulgaris*.
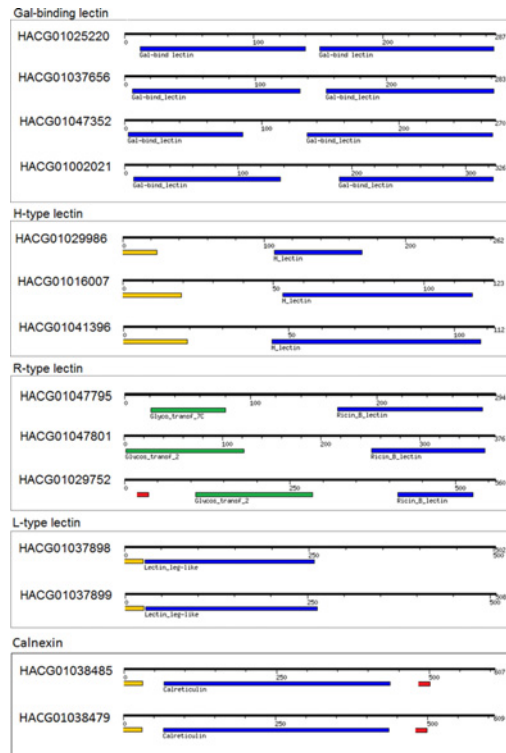
Fig 9. Alignment of amino acid sequences of H-type lectins from *A. vulgaris* with *Helix pomacea* and *Cepaea hortensis*.

doi:10.1371/journal.pone.0150614.g009

revealed that ten unigenes coding proteins share a galectin domain. Galectin domain-containing proteins expressed in *A. vulgaris* are not polymorphic, only a tandem-repeat type galectin has been identified ([Fig 8](#)). The absence of a signal sequence is consistent with other galectins, which are known to be actively secreted from diverse tissue via a "nonclassical" secretory pathway [29]. All of these lectins are most similar to gastropod galectins (*Aplysia californica*, *Biomphalaria glabrata*, etc.;blastp E value ranges from 0 to 2E–32).

L-type lectins

A member of L-type (legume-like) lectin family, ERGIC-53 is known as a pattern recognition receptor involved in the immune system of *E. sinensis* [30]. Transcriptomic analysis revealed the presence of two L-type lectin sequences with highest similarity to homolog of ERGIC-53 protein from *Littorina littorea* (blastp: E = 0.0; Identity = 57–58%). ERGIC-53 homologs share the characteristic architecture with an extracellular signal peptide and a single N-terminal L-type carbohydrate binding site ([Fig 8](#)).

Calnexin

Members of the evolutionarily conserved calnexin/calreticulin-superfamily bind to oligosaccharides containing terminal glucose residues. Calnexin acts as one of the pattern recognition receptors and has a crucial role in shrimp antibacterial immunity [31]. The search of *A. vulgaris* transcriptomic data revealed four calreticulin-like unigenes coding calnexin proteins with a highest similarity to *Aplysia californica* homologs (blastp: E = 0.0; Identity = 70–74%).
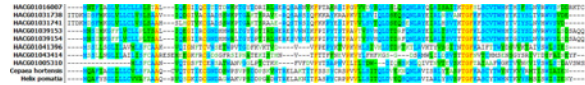
Fig 10. Domain architecture of C1q domain-containing proteins found in the transcriptome of *A. vulgaris*.

Calnexin homologs contain a signal sequence and an additional C-terminal transmembrane helix (Fig 8).

C1q domain-containing proteins

The C1q domain-containing proteins (C1qDC) possessing lectin-like features are a family of proteins characterized by a globular C1q domain [32] regarded as an important player in innate immunity of bivalvia molluscs [33–37]. C1qDC proteins participate in several immune responses, such as pathogen recognition [38], microorganism agglutination [39] and mediating cell migration [40]. In addition, the C1q domain is involved in other immunological processes, such as phagocytosis [41], neutralization of viruses, cell adhesion and clearance of apoptotic cells. It is proposed that C1qDC proteins can activate an ancient complement system by the lectin pathway prior to the evolution of immunoglobulins [41]. The C1q domain has been considered as an extremely efficient pattern recognition domain with highly adaptive binding properties. The extreme versatility of C1q is due to the capability of the C1q domain to bind a variety of self and non-self ligands, including lipopolysaccharides [42], virus envelope proteins, outer membrane proteins from Gram-negative bacteria, phospholipids and some acute-phase proteins.

Based on sequence homology, molecular architecture and domain similarity, 34 C1q domain-containing sequences from the *A. vulgaris* transcriptome may be classified as members of the C1q family (S4 Table). Ten transcripts have been identified also at the protein level. Most of the complete C1qDC sequences displayed a signal peptide or a transmembrane domain on the N-terminus (Fig 10).

Toxin-like proteins

Escaping predation is essential to survival. To reduce predation, organisms have developed a diverse defence mechanisms. A few candidate effector genes were found in the slug transcriptome, including a family of potential pore-forming membrane attack/perforin (MACPF) domain proteins. Pore-forming membrane attack/perforin (MACPF) domain proteins have been described to be involved in the biochemical defence of apple snail eggs against predators [43].

Transcriptomic analysis of *A. vulgaris* revealed a novel toxin-like protein showing significant similarity to agatoxin, expressed in spider glands with a characteristic cysteine motive in the mature peptide. The agatoxin-homolog from *A. vulgaris* shares characteristic architecture with signal peptide, prepeptide sequence and characteristic cysteine motive in the mature peptide (Fig 11).

Fig 11. Alignment of amino acid sequences of toxin-like proteins from *A. vulgaris* with U8-agatoxine–Ao1a from *Agelena orientalis*.

Taken together, one may speculate that the diversity and multitude of proteins with defensive potential along with already reported defensins (diterpene) of this slug [44] may contribute to success and survival of this terrestrial mollusc.

Data Accessibility

The sequence data has been submitted to the sequence read archive (SRA) database of GenBank (http://www.ncbi.nlm.nih.gov/sra) with the BioProject accession number PRJEB7891.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [45] via the PRIDE partner repository with the dataset identifier PXD002078 and 10.6019/PXD002078.

Supporting Information

S1 Table.
(XLSX)

S2 Table. List of 814 proteins identified from the mantel of *A. vulgaris* using 2-D electrophoretic separation.
(XLSX)

S3 Table. List of 2011 proteins identified from the mantel of *A. vulgaris* using 1-D electrophoretic separation.
(XLSX)

S4 Table. List of 2129 proteins identified from the mantel of *A. vulgaris*.
(XLSX)

S5 Table. List of lectin-like sequences and toxin-like sequences identified in *A. vulgaris*.
(XLSX)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JA GL. Performed the experiments: TB FJS GJ. Analyzed the data: RS TR JA. Contributed reagents/materials/analysis tools: MB. Wrote the paper: JA GL. Species identification: HS.

References

1. Gismervik K, Aspholm M, Rorvik LM, Bruheim T, Andersen A, Skaar I (2015) Invading slugs (*Arion vulgaris*) can be vectors for *Listeria monocytogenes*. J Appl Microbiol 118: 809–816. doi: 10.1111/jam.12750 PMID: 25580873

2. Gismervik K, Bruheim T, Rorvik LM, Haukeland S, Skaar I (2014) Invasive slug populations (*Arion vulgaris*) as potential vectors for *Clostridium botulinum*. Acta Vet Scand 56: 65. doi: 10.1186/s13028-014-0065-z PMID: 25277214

3. Stalder GL, Loncaric I, Walzer C (2014) Diversity of enterobacteria including beta-lactamase producing isolates associated with the Spanish slug (*Arion vulgaris*). Sci Total Environ 479–480: 11–16. doi: 10.1016/j.scitotenv.2014.01.103 PMID: 24534696

4. Patel Z, Gill AC, Fox MT, Hermosilla C, Backeljau T, Breugelmans K, et al. (2014) Molecular identification of novel intermediate host species of *Angiostrongylus vasorum* in Greater London. Parasitol Res 113: 4363–4369. doi: 10.1007/s00436-014-4111-6 PMID: 25195057

5. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, *Basommatophora*, *Pulmonata*), and a comparison of assembler performance. BMC Genomics 12: 317. doi: 10.1186/1471-2164-12-317 PMID: 21679424

6. Amin S, Prentis PJ, Gilding EK, Pavasovic A (2014) Assembly and annotation of a non-model gastropod (*Nerita melanotragus*) transcriptome: a comparison of de novo assemblers. BMC Res Notes 7: 488. doi: 10.1186/1756-0500-7-488 PMID: 25084827

7. Sun J, Wang M, Wang H, Zhang H, Zhang X, Thiyagarajan V, et al. (2012) De novo assembly of the transcriptome of an invasive snail and its multiple ecological applications. Mol Ecol Resour 12: 1133–1144. doi: 10.1111/1755-0998.12014 PMID: 22994926

8. Kopylova E, Noe L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in meta-transcriptomic data. Bioinformatics 28: 3211–3217. doi: 10.1093/bioinformatics/bts611 PMID: 23071270

9. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652. doi: 10.1038/nbt.1883 PMID: 21572440

10. Chen TW, Gan RC, Wu TH, Huang PJ, Lee CY, Chen YY, et al. (2012) FastAnnotator—an efficient transcript annotation web tool. BMC Genomics 13 Suppl 7: S9. doi: 10.1186/1471-2164-13-S7-S9 PMID: 23281853

11. Bae N, Lodl M, Pollak A, Lubec G (2012) Mass spectrometrical analysis of bilin-binding protein from the wing of *Hebomoia glaucippe* (Linnaeus, 1758) (Lepidoptera: Pieridae). Electrophoresis 33: 1787–1794. doi: 10.1002/elps.201100569 PMID: 22740467

12. Ahmed KE, Chen WQ, John JP, Kang SU, Lubec G (2010) Complete sequencing of the recombinant granulocyte-colony stimulating factor (filgrastim) and detection of biotinylation by mass spectrometry. Amino Acids 38: 1043–1049. doi: 10.1007/s00726-009-0312-1 PMID: 19526310

13. Chen WQ, Li L, Lubec G (2014) Proteomic differences between white and brown adipocytes. Amino Acids 46: 995–1008. doi: 10.1007/s00726-013-1657-z PMID: 24390458

14. Kang SU, Heo S, Lubec G (2011) Mass spectrometric analysis of GABAA receptor subtypes and phosphorylations from mouse hippocampus. Proteomics 11: 2171–2181. doi: 10.1002/pmic.201000374 PMID: 21538884

15. Candiano G, Bruschi M, Musante L, Santucci L, Ghiggeri GM, Carnemolla B, et al. (2004) Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. Electrophoresis 25: 1327–1333. PMID: 15174055

16. Bennett KL, Funk M, Tschernutter M, Breitwieser FP, Planyavsky M, Ubaida Mohien C, et al. (2011) Proteomic analysis of human cataract aqueous humour: Comparison of one-dimensional gel LCMS with two-dimensional LCMS of unlabelled and iTRAQ(R)-labelled specimens. J Proteomics 74: 151–166. doi: 10.1016/j.jprot.2010.10.002 PMID: 20940065

17. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4: 207–214. PMID: 17327847

18. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20: 1313–1326. doi: 10.1101/gr.101386.109 PMID: 20651121

19. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157: 77–94. doi: 10.1016/j.cell.2014.03.008 PMID: 24679528

20. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676. PMID: 16081474

21. Voncken F, Gao F, Wadforth C, Harley M, Colasante C (2013) The phosphoarginine energy-buffering system of trypanosoma brucei involves multiple arginine kinase isoforms with different subcellular locations. PLoS One 8: e65908. doi: 10.1371/journal.pone.0065908 PMID: 23776565

22. Pfenninger M, Weigand A, Balint M, Klussmann-Kolb A (2014) Misperceived invasion: the Lusitanian slug (*Arion lusitanicus* auct. non-Mabille or *Arion vulgaris* Moquin-Tandon 1855) is native to Central Europe. Evol Appl 7: 702–713. doi: 10.1111/eva.12177 PMID: 25067951

23. Hatteland BA, Roth S, Andersen A, Kaasa K, Stoa B, T S (2013) Distribution and spread of the invasive slug *Arion vulgaris* MoquinTandon in Norway. Fauna norvegica 32: 13–26.

24. Kozlowski J (2007) The distribution, biology, population dynamics and harmfulness of *Arion lusitanicus* Mabillee, 1868 (Gastropoda: *Pulmonata*: *Arionidae*) in Poland. Journal of Plant Protection Research 47: 219–230.

25. Sun J, Zhang H, Wang H, Heras H, Dreon MS, Ituarte S, et al. (2012) First proteome of the egg perivitelline fluid of a freshwater gastropod with aerial oviposition. J Proteome Res 11: 4240–4248. doi: 10.1021/pr3003613 PMID: 22738194

26. Zelensky AN, Gready JE (2005) The C-type lectin-like domain superfamily. FEBS J 272: 6179–6217. PMID: 16336259

27. Yuasa HJ, Furuta E, Nakamura A, Takagi T (1998) Cloning and sequencing of three C-type lectins from body surface mucus of the land slug, *Incilaria fruhstorferi*. Comp Biochem Physiol B Biochem Mol Biol 119: 479–484. PMID: 9734332

28. Sanchez JF, Lescar J, Chazalet V, Audfray A, Gagnon J, Alvarez R, et al. (2006) Biochemical and structural analysis of *Helix pomatia* agglutinin. A hexameric lectin with a novel fold. J Biol Chem 281: 20171–20180. PMID: 16704980

29. Nickel W (2003) The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. Eur J Biochem 270: 2109–2119. PMID: 12752430

30. Huang Y, Tan JM, Wang Z, Yin SW, Huang X, Wang W, et al. (2014) Cloning and characterization of two different L-type lectin genes from the Chinese mitten crab *Eriocheir sinensis*. Dev Comp Immunol 46: 255–266. doi: 10.1016/j.dci.2014.04.015 PMID: 24796868

31. Zhang Q, Wang XQ, Jiang HS, Jia WM, Zhao XF, Wang JX (2014) Calnexin functions in antibacterial immunity of *Marsupenaeus japonicus*. Dev Comp Immunol 46: 356–363. doi: 10.1016/j.dci.2014.05.009 PMID: 24858031

32. Carland TM, Gerwick L (2010) The C1q domain containing proteins: Where do they come from and what do they do? Dev Comp Immunol 34: 785–790. doi: 10.1016/j.dci.2010.02.014 PMID: 20214925

33. Gerlach D, Schlott B, Schmidt KH (2004) Cloning and expression of a sialic acid-binding lectin from the snail *Cepaea hortensis*. FEMS Immunol Med Microbiol 40: 215–221. PMID: 15039097

34. Allam B, Pales Espinosa E, Tanguy A, Jeffroy F, Le Bris C, Paillard C (2014) Transcriptional changes in Manila clam (*Ruditapes philippinarum*) in response to Brown Ring Disease. Fish Shellfish Immunol 41: 2–11. doi: 10.1016/j.fsi.2014.05.022 PMID: 24882017

35. Gestal C, Pallavicini A, Venier P, Novoa B, Figueras A (2010) MgC1q, a novel C1q-domain-containing protein involved in the immune response of *Mytilus galloprovincialis*. Dev Comp Immunol 34: 926–934. doi: 10.1016/j.dci.2010.02.012 PMID: 20219531

36. He X, Zhang Y, Yu F, Yu Z (2011) A novel sialic acid binding lectin with anti-bacterial activity from the Hong Kong oyster (*Crassostrea hongkongensis*). Fish Shellfish Immunol 31: 1247–1250. doi: 10.1016/j.fsi.2011.08.021 PMID: 21906682

37. Xu T, Xie J, Li J, Luo M, Ye S, Wu X (2012) Identification of expressed genes in cDNA library of hemocytes from the RLO-challenged oyster, *Crassostrea ariakensis* Gould with special functional implication of three complement-related fragments (CaC1q1, CaC1q2 and CaC3). Fish Shellfish Immunol 32: 1106–1116. doi: 10.1016/j.fsi.2012.03.012 PMID: 22484279

38. Yu Y, Huang H, Wang Y, Yu Y, Yuan S, Huang S, et al. (2008) A novel C1q family member of amphioxus was revealed to have a partial function of vertebrate C1q molecule. J Immunol 181: 7024–7032. PMID: 18981122

39. Kong P, Zhang H, Wang L, Zhou Z, Yang J, Zhang Y, et al. (2010) AiC1qDC-1, a novel gC1q-domain-containing protein from bay scallop *Argopecten irradians* with fungi agglutinating activity. Dev Comp Immunol 34: 837–846. doi: 10.1016/j.dci.2010.03.006 PMID: 20346969

40. Tahtouh M, Croq F, Vizioli J, Sautiere PE, Van Camp C, Salzet M, et al. (2009) Evidence for a novel chemotactic C1q domain-containing factor in the leech nerve cord. Mol Immunol 46: 523–531. doi: 10.1016/j.molimm.2008.07.026 PMID: 18952286

41. Wang L, Wang L, Zhang H, Zhou Z, Siva VS, Song L (2012) A C1q domain containing protein from scallop *Chlamys farreri* serving as pattern recognition receptor with heat-aggregated IgG binding activity. PLoS One 7: e43289. doi: 10.1371/journal.pone.0043289 PMID: 22905248

42. Zhang H, Song L, Li C, Zhao J, Wang H, Qiu L, et al. (2008) A novel C1q-domain-containing protein from Zhikong scallop *Chlamys farreri* with lipopolysaccharide binding activity. Fish Shellfish Immunol 25: 281–289. doi: 10.1016/j.fsi.2008.06.003 PMID: 18603000

43. Dreon MS, Frassa MV, Ceolin M, Ituarte S, Qiu JW, Sun J, et al. (2013) Novel animal defenses against predation: a snail egg neurotoxin combining lectin and pore-forming chains that resembles plant defense and bacteria attack toxins. PLoS One 8: e63782. doi: 10.1371/journal.pone.0063782 PMID: 23737950

44. Schroeder FC, Gonzalez A, Eisner T, Meinwald J (1999) Miriamin, a defensive diterpene from the eggs of a land slug (*Arion sp.*). Proc Natl Acad Sci U S A 96: 13620–13625. PMID: 10570122

45. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 32: 223–226. doi: 10.1038/nbt.2839 PMID: 24727771