A combined bioinformatics and LC-MS-based approach for the development and benchmarking of a comprehensive database of Lymnaea CNS proteins

Article (Published Version)

© 2022. Published by The Company of Biologists Ltd | Journal of Experimental Biology (2022) 225, jeb243753. doi:10.1242/jeb.243753

RESEARCH ARTICLE

A combined bioinformatics and LC-MS-based approach for the development and benchmarking of a comprehensive database of *Lymnaea* CNS proteins

Sarah Wooller[1,*], Aikaterini Anagnostopoulou[2,*], Benno Kuropka[3], Michael Crossley[2], Paul R. Benjamin[2], Frances Pearl[1], Ildikó Kemenes[2], György Kemenes[2,‡] and Murat Eravci[2,‡]

ABSTRACT

Applications of key technologies in biomedical research, such as qRT-PCR or LC-MS-based proteomics, are generating large biological (-omics) datasets which are useful for the identification and quantification of biomarkers in any research area of interest. Genome, transcriptome and proteome databases are already available for a number of model organisms including vertebrates and invertebrates. However, there is insufficient information available for protein sequences of certain invertebrates, such as the great pond snail *Lymnaea stagnalis*, a model organism that has been used highly successfully in elucidating evolutionarily conserved mechanisms of memory function and dysfunction. Here, we used a bioinformatics approach to designing and benchmarking a comprehensive central nervous system (CNS) proteomics database (LymCNS-PDB) for the identification of proteins from the CNS of *Lymnaea* by LC-MS-based proteomics. LymCNS-PDB was created by using the Trinity TransDecoder bioinformatics tool to translate amino acid sequences from mRNA transcript assemblies obtained from a published *Lymnaea* transcriptomics database. The blast-style MMSeq2 software was used to match all translated sequences to UniProtKB sequences for molluscan proteins, including those from *Lymnaea* and other molluscs. LymCNS-PDB contains 9628 identified matched proteins that were benchmarked by performing LC-MS-based proteomics analysis with proteins isolated from the *Lymnaea* CNS. MS/MS analysis using the LymCNS-PDB database led to the identification of 3810 proteins. Only 982 proteins were identified by using a non-specific molluscan database. LymCNS-PDB provides a valuable tool that will enable us to perform quantitative proteomics analysis of protein interactomes involved in several CNS functions in *Lymnaea*, including learning and memory and age-related memory decline.

KEY WORDS: *Lymnaea*, Central nervous system, Liquid chromatography–mass spectrometry, Bioinformatics, Proteomics database

[1]Bioinformatics Group, School of Life Sciences, University of Sussex, Brighton, BN1 9QG, UK. [2]Sussex Neuroscience, School of Life Sciences, University of Sussex, Brighton, BN1 9QG, UK. [3]Institute for Chemistry and Biochemistry, Freie Universität Berlin, 14195 Berlin, Germany.
*These authors contributed equally to this work

‡Authors for correspondence (m.eravci@sussex.ac.uk, g.kemenes@sussex.ac.uk)

A.A., 0000-0002-2212-1714; B.K., 0000-0001-5088-6346; M.C., 0000-0002-3120-4124; P.R.B., 0000-0002-1021-3558; F.P., 0000-0002-8210-4393; I.K., 0000-0002-8722-8766; G.K., 0000-0003-2004-8725; M.E., 0000-0003-4786-9179

Received 5 November 2021; Accepted 17 March 2022

INTRODUCTION

Protein networks perform key functions in all living organisms. The study of such complex biological functions has led to the development of liquid chromatography–mass spectrometry (LC-MS)-based platforms that enable researchers to perform quantitative system-wide analysis of proteomes, including protein–protein interactions, post-translational modifications and spatial localization of proteins, even at the single cell level. Furthermore, several labelling and pre-fractionation techniques allow researchers to improve the detection of low abundance and possibly functionally relevant proteins.

The great pond snail, *Lymnaea stagnalis*, is used as a model organism in a wide range of biological research fields, such as the study of host–parasite interactions, ecotoxicology, evolution, developmental biology, learning and memory, ageing and age-related memory decline, genome editing, '-omics' and human disease modelling (Benjamin et al., 2021; Benjamin and Kemenes, 2020; Fodor et al., 2020a; Rivi et al., 2020).

In our laboratory, we are interested in elucidating evolutionarily conserved molecular mechanisms involved in the formation of long-term memory after classical conditioning and memory impairment associated with ageing as well as amyloid-β-induced memory decline. In memory consolidation and underlying synaptic plasticity, we and others have already established important roles for a variety of key enzyme proteins, such as nitric oxide synthase (NOS) (Kemenes et al., 2002), mitogen-activated protein kinase (MAPK) (Ribeiro et al., 2005), protein kinase A (PKA) (Kemenes et al., 2006), Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) (Naskar et al., 2014) and the transcription factors cAMP-response element binding protein 1 (CREB1) (Ribeiro et al., 2003; Sadamoto et al., 2004) and CCAAT enhancer binding protein (C/EBP) (Hatakeyama et al., 2006). In age-related memory decline in *L. stagnalis*, we have found several impaired signalling pathways, including pituitary adenylate cyclase activating polypeptide (PACAP) and insulin-like growth factor-1 (IGF-1); when these are restored by administering exogenous PACAP or IGF-1, the age-related learning deficiency is reversed (Pirger et al., 2014). We have already established *L. stagnalis* as a useful invertebrate model for amyloid-β-associated memory impairment. Specifically, we found that amyloid-β peptides 1–42 and 25–35 were able to inhibit long-term memory recall in *L. stagnalis* (Ford et al., 2017; Ford et al., 2015). Recently, the sequencing of the whole transcriptome of the central nervous system (CNS) of *L. stagnalis* identified several evolutionarily conserved sequences of genes involved in human ageing, and age-related and amyloid-β-induced memory loss, including gelsolin, presenilin, huntingin, Parkinson disease protein 7 (PARK-7/DJ-1) and amyloid precursor protein (Fodor et al., 2021; Fodor et al., 2020b). However, the lack of

Journal of Experimental Biology (2022) 225, jeb243753. doi:10.1242/jeb.243753

comprehensive proteomics information in *L. stagnalis* has hindered further progress with research aimed at understanding the roles of these proteins in the context of large-scale protein networks in the CNS.

Although quantitative reverse transcription–polymerase chain reaction (qRT-PCR) is a well-established method to quantify specific gene transcripts of interest, it does not allow the quantification and detection of the function of a specific protein in its protein network because of the possibility of post-translational modifications, e.g. phosphorylation. To enable us to provide these types of important information, in the present study we performed large-scale proteomics experiments using Nanoscale liquid chromatography–mass spectrometry (nanoLC-MS) to analyse protein expression and post-translational modification in the CNS of *L. stagnalis*.

One important prerequisite for a successful proteomic workflow and the accurate identification of proteins is the existence of a protein sequence database of the organism of interest that can be used for comparison of the peptide sequences acquired from tandem mass spectrometry (MS/MS) fragmentation spectra with protein sequences in the database of the same organism.

In the Universal Protein Knowledgebase (UniProtKB) (UniProt, 2019), the only available and useful protein database for *L. stagnalis* consists of 519 proteins, of which 48 have been reviewed but the remaining 471 have not (Uniprot.org – last modified on 3 August 2020). Because of this lack of specific protein sequence information, previous proteomics analysis of *L. stagnalis* was performed by utilizing the entire UniProtKB/Swiss-Prot database with protein sequences from all available organisms (Rosenegger et al., 2010; Silverman-Gavrila et al., 2011), or using a Metazoa-specific database (Giusti et al., 2013) to identify proteins.

To prepare a more comprehensive and representative database for CNS proteins of *L. stagnalis*, which will enable us to identify proteins accurately by liquid chromatography with tandem mass spectrometry (LC-MS/MS) analysis, we selected the transcriptome dataset of Sadamoto et al. (2012) from the available transcriptome datasets (Bouetard et al., 2012; Davison and Blaxter, 2005; Dong et al., 2021; Feng et al., 2009; Sadamoto et al., 2004; Sadamoto et al., 2012) published in the NCBI database because it contains the transcriptome of the whole *L. stagnalis* CNS including the buccal 'learning' ganglia and the central cerebral ring.

The transcripts from the Sadamoto et al. (2012) dataset (NCBI accession number PRJDB98) were filtered for coding regions and the remaining transcripts were then searched for homology against all UniProtKB molluscan entries. The resulting *Lymnaea* CNS protein database, LymCNS-PDB, was then utilized in a proof of principle experiment by performing LC-MS-based proteomics analysis with proteins isolated from the CNS of *L. stagnalis*. Furthermore, we prepared a database (DB) with all proteins from the LymCNS-PDB using matching amino acid sequences of the other molluscan species from UniProtKB. As the two databases were the same size, we were able to compare the number of identifications from a *L. stagnalis*-specific database with those in a non-specific molluscan database.

MATERIALS AND METHODS

Experimental animals

Specimens of *Lymnaea stagnalis* (Linnaeus 1758) were raised in the breeding facility of the University of Sussex, where they were kept in 20–22°C copper-free water under a 12 h light and dark cycle. They were fed on lettuce 3 times and a vegetable-based fish food twice a week.

Preparation of CNS samples

Whole CNS samples from 90 snails aged 3 months and weighing approximately 1.5 g were prepared as follows. The shell of the snail was cut, and the body carefully removed and pinned to a Sylgard-coated dish containing Hepes-buffered saline, then dissected under a stereomicroscope (E-Zoom6, Edmund Optics, Barrington, NJ, USA). The CNS was accessed by an incision in the dorsal body region isolated from the buccal mass by the severing of all the peripheral nerves, and then immediately placed in Eppendorf tubes on dry ice. Three tubes, each containing 30 CNS, were stored at −80°C.

Protein extraction, digestion and prefractionation

Frozen CNS samples were thawed for 30 min at room temperature (RT; 20–22°C) before adding lysis buffer (6 mol l^{-1} urea, 2 mol l^{-1} thiourea in 10 mmol l^{-1} Hepes pH 8.0) and 30 ceramic beads (1.4 mm zirconium oxide beads) for homogenization in a Precellys 24 Homogeniser (Bertin Instruments) using two cycles for 30 s at 6800 rpm with a 60 s break in between, followed by a centrifugation for 1 h at 14,000 ***g*** to remove the debris. The supernatant was transferred to a fresh tube and the extracted proteins were reduced for 30 min in 10 mmol l^{-1} dithiothreitol and alkylated for 30 min in 55 mmol l^{-1} iodoacetamide (in the dark). Proteins were first pre-digested with LysC for at least 3 h at RT and after dilution with 3 volumes of 50 mmol l^{-1} ammonium bicarbonate buffer, the main digestion was performed with trypsin overnight at RT. Peptide samples (in triplicate) were desalted with C18 Hypersep cartridges (Thermo Fisher) and eluates were concentrated in a SpeedVac concentrator (Savant) and prefractionated into 6 fractions using the immobilized pH gradient (IPG) strip-based peptide fractionation method as previously described (Eravci et al., 2014).

LC-MS analysis

The 6 desalted peptide fractions were analysed in triplicate (18 samples in total) by a reversed-phase capillary nano liquid chromatography system (Ultimate 3000, Thermo Scientific) connected to a Q Exactive HF mass spectrometer (Thermo Scientific). Samples were injected and concentrated on a trap column (PepMap100 C18, 3 µm, 100 Å, 75 µm i.d.×2 cm, Thermo Scientific) equilibrated with 0.05% trifluoroacetic acid in water. After switching the trap column inline, LC separations were performed on a capillary column (Acclaim PepMap100 C18, 2 µm, 100 Å, 75 µm i.d.×25 cm, Thermo Scientific) at an eluent flow rate of 300 nl min^{-1}. Mobile phase A contained 0.1% formic acid in water, and mobile phase B contained 0.1% formic acid in 80% acetonitrile, 20% water. The column was pre-equilibrated with 5% mobile phase B and peptides were separated using a gradient of 5–44% mobile phase B within 100 min. Mass spectra were acquired in a data-dependent mode utilizing a single MS survey scan (*m*/*z* 350–1650) with a resolution of 60,000 in the Orbitrap, and MS/MS scans of the 15 most intense precursor ions with a resolution of 15,000. HCD fragmentation was performed for all ions with charge states of 2+ to 5+, normalized collision energy of 27 and an isolation window of 1.4 *m*/*z*. The dynamic exclusion time was set to 20 s. Automatic gain control (AGC) was set to 3×106 for MS scans using a maximum injection time of 20 ms. For MS2 scans, the AGC target was set to 1×105 with a maximum injection time of 25 ms.

MS and MS/MS raw data were analysed using the MaxQuant software package (version 1.6.12.0) with an implemented Andromeda peptide search engine (Tyanova et al., 2016). Data were searched against the FASTA formatted protein database of *L. stagnalis* described here.

Journal of Experimental Biology (2022) 225, jeb243753. doi:10.1242/jeb.243753

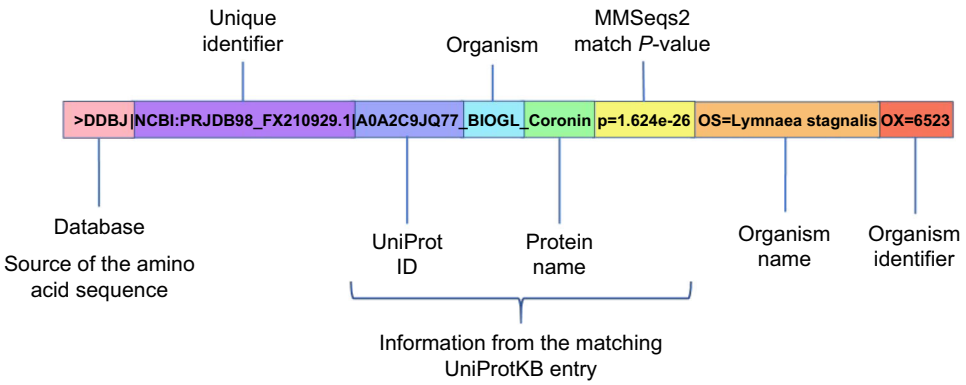

Fig. 1. Design of the header composition used for the *Lymnaea stagnalis* central nervous system proteomics database (LymCNS-PDB) following the UniProtKB parsing rules for FASTA headers.

Database construction

To predict coding regions in the NCBI PRJDB98 transcriptome dataset, we analysed all available transcript assemblies using the Trinity TransDecoder (version 5.5.0) bioinformatics tool. Prior to this, we had downloaded all available Swiss-Prot and TrEMBL amino acid sequences for molluscan proteins from UniProtKB and characterized them by whether they were sequences for *L. stagnalis*, other euthyneura species or other molluscan species.

All sequences were compared with the transdecoded database from the NCBI PRJDB98 transcriptome dataset mentioned above, in order to find potential matches using the blast-style program Many-against-Many sequence searching (MMSeqs2) (Steinegger and Soding, 2017). In each case, proteins were matched by preference, first against *L. stagnalis*, then other euthyneura and, finally, against other molluscan proteins. Duplicate proteins as well as matches to proteins termed 'uncharacterized' or 'hypothetical' were removed from the final dataset.

For all matching entries, the header of the TrEMBL/Swiss-Prot protein was used to update the headings of the respective uncharacterized *L. stagnalis* amino acid sequences from the transdecoded PRJDB98 transcriptome dataset, indicating the original GenBank accession number, the UniProt ID and the protein name of the matching molluscan protein, and the *P*-value for matching accuracy provided by the MMSeqs2. The constructed database LymCNS-PDB is saved in a 'FASTA' format following the UniProtKB parsing rules for FASTA headers (see Fig. 1).

To compare our database with the non-specific database, we prepared a molluscan database with all proteins identified from our LymCNS-PDB database using the amino acid sequences of the matching molluscan species from UniProtKB. To allow optimal comparison, the two databases are the same size and contain the same number of proteins.

RESULTS

Construction of the LymCNS-PDB

The NCBI PRJDB98 transcriptome dataset (Sadamoto et al., 2012), which contains 116,265 transcripts, was analysed with a Trinity TransDecoder (version 5.5.0) to predict coding regions in the present assemblies; 22,180 'transdecoded' transcripts were then used for homology searches with MMSeqs2 against 211,200 protein entries from the UniProtKB database with the preference for *Lymnaea stagnalis*>Euthyneura>Mollusca. We were able to match 16,142 transcripts to 9628 proteins from the UniProtKB database (Fig. 2).

Three-quarters of all matching entries were from organisms with a large number of available protein sequences in UniProtKB: *Mizuhopecten yessoensis* (MIZYE; Yesso scallop with 22,614 protein entries matching to 4781 PRJDB98 transcripts of 3469 unique proteins; *Crassostrea gigas* (CRAGI; Pacific oyster) with 27,077 protein entries matching to 2672 PRJDB98 transcripts of 1997 unique proteins and *Biomphalaria glabrata* (BIOGL), a species of pulmonate freshwater snail, with 31,775 protein entries matching to 3183 PRJDB98 transcripts of 1729 unique proteins.

A rather low number of matching entries were from the following molluscan organisms: a mixture of several *Eupulmonata* (9EUPU), with the majority of protein entries (75%) from *Arion vulgaris* within this taxonomic clade of air-breathing snails, with 65,368 protein entries matching to 1072 PRJDB98 transcripts of 614 unique proteins; *Elysia chlorotica* (ELYCH; eastern emerald elysia) with 23,887 protein entries matching to 928 PRJDB98 transcripts of 525 unique proteins; *Lymnaea stagnalis* (LYMST; great pond snail) with 442 protein entries matching to 1245 PRJDB98 transcripts of 391 unique proteins; *Aplysia californica* (APLCA; California sea hare) with 443 protein entries matching to 755 PRJDB98 transcripts of 136 unique proteins; *Pomacea canaliculata* (POMCA; golden

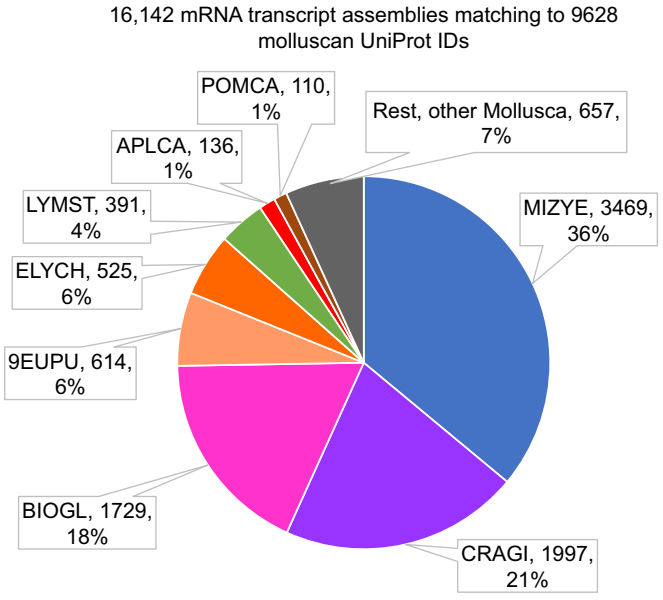

Fig. 2. Distribution of matching transcript assemblies of the NCBI PRJDB98 transcriptome dataset to proteins of different organisms within the UniProtKB Mollusca database. MIZYE, *Mizuhopecten yessoensis*; CRAGI, *Crassostrea gigas*; BIOGL, *Biomphalaria glabrata*; 9EUPU, *Eupulmonata*/majority *Arion vulgaris* at 75%; ELYCH, *Elysia chlorotica*; LYMST, *Lymnaea stagnalis*; APLCA, *Aplysia californica*; POMCA, *Pomacea canaliculata*.

Journal of Experimental Biology (2022) 225, jeb243753. doi:10.1242/jeb.243753

Table 1. Overview of the number of protein entries for the respective organisms in UniProtKB database

Organism	Abbreviation	No. of UniProt protein entries	Matching proteins in LymCNS-PDB	Unique proteins identified by MS analysis with LymCNS-PDB
Mizuhopecten yessoensis	MIZYE	22,614	3469	1478
Crassostrea gigas	CRAGI	27,077	1997	775
Biomphalaria glabrata	BIOGL	31,775	1729	590
Eupulmonata	9EUPU	65,368	614	275
Elysia chlorotica	ELYCH	23,887	525	177
Lymnaea stagnalis	LYMST	442	391	121
Aplysia californica	APLCA	443	136	82
Pomacea canaliculate	POMCA	21,514	110	32
Rest of Mollusca		237,428	657	280
Sum				3810

The table shows the number of matching proteins in the LymCNS-PDB database and the number of unique proteins identified by MS analysis using the LymCNS-PDB database in comparison to the number of unique proteins identified by MS analysis using a molluscan database from UniProtKB (December 2020).

apple snail) with 21,514 protein entries matching to 144 PRJDB98 transcripts of 110 unique proteins; and the remaining 237,428 molluscan protein entries matching to 1029 PRJDB98 transcripts of 657 unique proteins (Fig. 2 and Table 1).

LC-MS-based proteomics analysis

To demonstrate the suitability of our newly created protein database (LymCNS-PDB), we performed a MS-based proteomics experiment of the CNS from *L. stagnalis*, which resulted in the identification of 3810 unique proteins. The identified proteins were derived from the following molluscan organisms with significant matching homology to the corresponding amino acid sequences of the *L. stagnalis* transcripts from the PRJDB98 dataset: *Mizuhopecten yessoensis* (MIZYE) with 1478 unique proteins; *Crassostrea gigas* (CRAGI) with 775; *Biomphalaria glabrata* (BIOGL) with 590; *Eupulmonata* (9EUPU) with 275; *Elysia chlorotica* (ELYCH) with 177; *Lymnaea stagnalis* (LYMST) with 121; *Aplysia californica* (APLCA) with 82; *Pomacea canaliculata* (POMCA) with 32; and other molluscan organisms with 280 (Fig. 3 and Table 1). In contrast, when the same experimental data were analysed against the non-specific molluscan database, only 920 unique proteins were identified with 44 proteins from *Mizuhopecten yessoensis*, 121 proteins from *Crassostrea gigas*, 341 proteins from *Biomphalaria glabrata*, 104 proteins from *Eupulmonata*, 52 proteins from *Elysia chlorotica*, 121 proteins from *Lymnaea stagnalis*, 41 proteins from *Aplysia californica*, 3 proteins from *Pomacea canaliculate* and 34 proteins from other molluscan species.

To provide a functional categorization of all proteins included in our LymCNS-PDB, protein sequences were further annotated and classified based on EuKaryotic Orthologous Groups (KOG) categories by using RPSBLAST 2.2.15 on NCBI KOG 2/2/2011 database (http://weizhong-lab.ucsd.edu/webMGA/server/kog/). Fig. 4 shows the distribution of KOG annotations evaluated on the proteins in the LymCNS-PDB and Table S1 provides more detailed information on this distribution. The similarity of this KOG annotation pattern to the distribution in the comparison of KOG annotations of protein-coding transcripts expressed in the CNS of key vertebrate and invertebrate neuroscience model organisms presented by Dong et al. (2021) indicates that all those functional categories are also present in our LymCNS-PDB.

DISCUSSION

The pond snail *L. stagnalis* is an invertebrate model organism used highly successfully in both basic and translational neuroscience research aimed at understanding the neural and circuit mechanisms of a variety of behaviours because of its numerically simple and well-characterized CNS (Benjamin, 2008; Fodor et al., 2020a; Rivi et al., 2020). However, the molecular characterization of the different functions of the CNS has been limited by the lack of a comprehensive proteomics database. In this investigation, we set out to develop an extensive proteomics database based on the mRNA transcript assemblies from the NCBI Bioproject PRJDB98 obtained by *de novo* sequencing and transcriptome analysis (Sadamoto et al., 2012).

Although a recently published paper by Dong et al. (2021) presented a more comprehensive *L. stagnalis* transcriptomic database, it only included RNA transcripts from the ring without the buccal ganglia, and the identification of protein sequences was focused mainly on different ion channels (Dong et al., 2021). In our approach, we developed a more general proteomics database that includes proteins involved in several CNS functions such

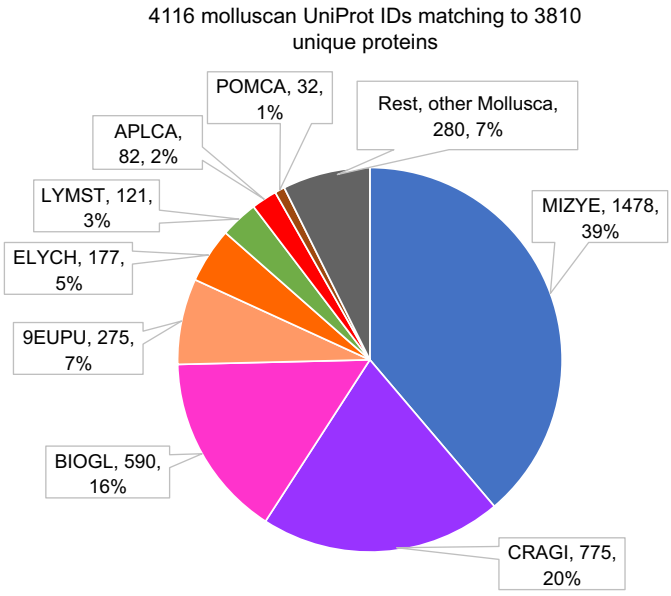

Fig. 3. Distribution of matching organisms of the identified proteins by LC-MS analysis using the LymCNS-PDB. MIZYE, *Mizuhopecten yessoensis*; CRAGI, *Crassostrea gigas*; BIOGL, *Biomphalaria glabrata*; 9EUPU, *Eupulmonata*/majority *Arion vulgaris* at 75%; ELYCH, *Elysia chlorotica*; LYMST, *Lymnaea stagnalis*; APLCA, *Aplysia californica*; POMCA, *Pomacea canaliculata*.

Journal of Experimental Biology (2022) 225, jeb243753. doi:10.1242/jeb.243753

as learning, fundamental decision making and feeding-related motivational states, among others. As previous published studies have identified the buccal ganglia as the location of circuitry involved in the expression of both appetitive and aversive memories (Ito et al., 2012; Marra et al., 2010), encoding hunger states (Crossley et al., 2016; Staras et al., 2003) as well as fundamental decision making (Crossley et al., 2018), we developed a proteomics database created from a transcriptomic database that included RNA transcripts from the buccal ganglia as well as the ring ganglia. For this reason, we used the *L. stagnalis* transcriptomics database published by Sadamoto et al. (2012) as this database contains the transcriptome from the whole CNS, not just the ring ganglia (Sadamoto et al., 2012).

By matching all mRNA transcript assemblies from the NCBI Bioproject PRJDB98 to all available molluscan proteins on the UniProtKB database, we succeeded in creating the proteomics database LymCNS-PDB with 9628 proteins, containing the translated amino acid sequences of their respective mRNAs from the *L. stagnalis* CNS, as well as obtaining all the other information (e.g. protein name) from their matching molluscan counterparts from UniProtKB. Most of the matches to certain molluscan species were due to the large number of available protein sequences of this organism in UniProtKB. Species that have a smaller number of identified proteins have a greater match compared with those with a larger number of identified proteins. For example, *Mizuhopecten yessoensis* with 22,614, *Crassostrea gigas* with 27,077 and *Biomphalaria glabrata* with 31,775 protein entries in UniProtKB were matched to 3469, 1997 and 1729 PRJDB98 sequences, respectively. These matches represent 75% of all the matching proteins (Fig. 2), even though two of these three organisms, *M. yessoensis* and *C. gigas*, have the most distant phylogenetic relationship to *L. stagnalis* amongst all of the matching organisms (Fig. 5). In contrast, the group with a very high number of UniProtKB entries (65,368), *Eupulmonata* (with approximately 49,000 UniProtKB entries from *Arion vulgaris*) only matched to 614 PRJDB98 sequences, even though it has a much closer phylogenetic relationship to *L. stagnalis* (Fig. 5).

The number of protein identifications was increased by performing a pre-fractionation of the tryptic peptides from the CNS of *L. stagnalis* using IPG gels before analysis of the fractions by nanoLC-MS (Eravci et al., 2014).

Analysis of the MS/MS fragmentation spectra using the LymCNS-PDB led to the identification of 3810 unique proteins, representing almost 40% of the entire LymCNS-PDB database. To our knowledge, this is the highest number of protein identifications in a proteomics experiment using *L. stagnalis*.

The proportion of different organisms identified by MS analysis using the LymCNS-PDB (Fig. 3) shows the same distribution as that

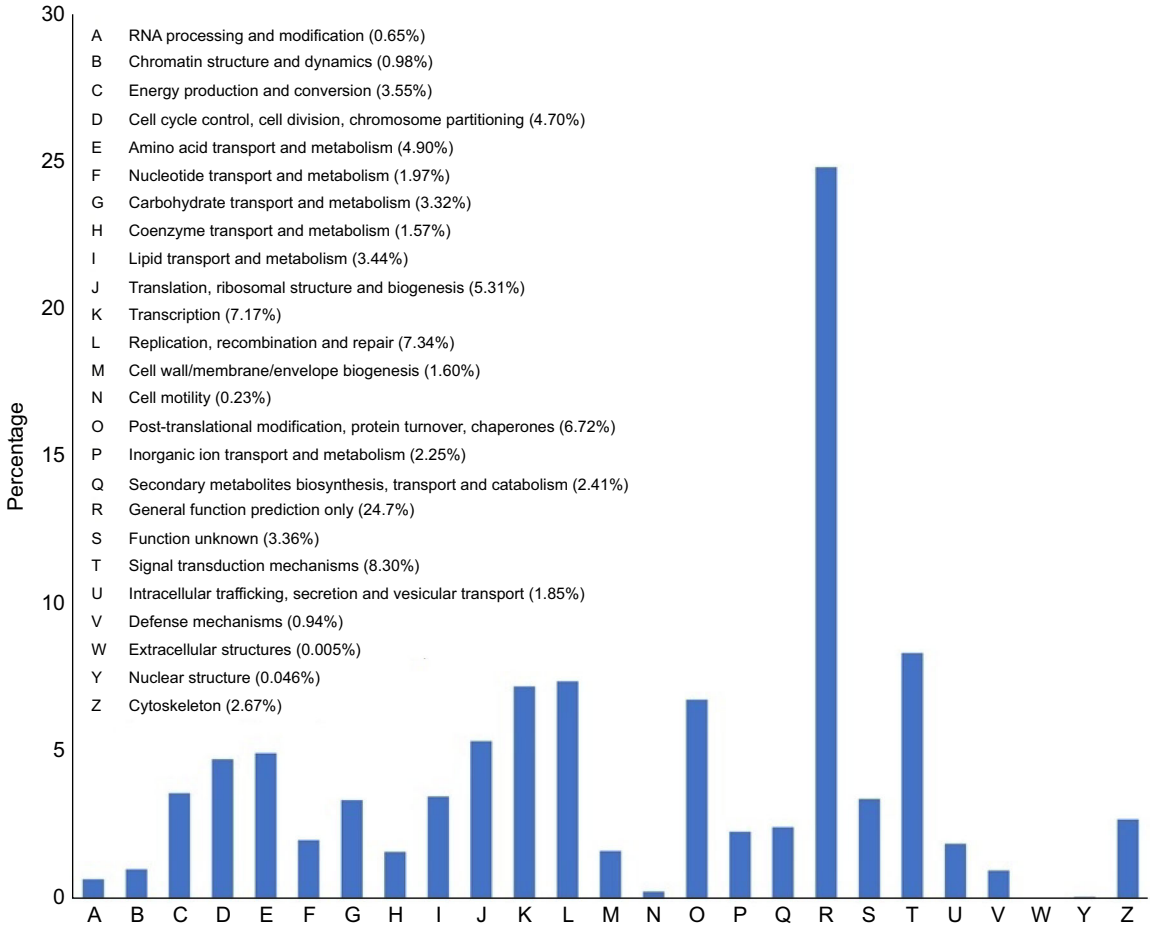

Fig. 4. Distribution of KOG (EuKaryotic Orthologous Groups) annotations of the protein sequences in the LymCNS-PDB. Proteins were annotated and grouped into 25 functional categories (A–Z) using the RPSBLAST 2.2.15 program on NCBI KOG 2/2/2011 database using an e-value <1E−5 cutoff for prediction (http://weizhong-lab.ucsd.edu/webMGA/server/kog). For more details of the distribution, see Table S1.

Journal of Experimental Biology (2022) 225, jeb243753. doi:10.1242/jeb.243753

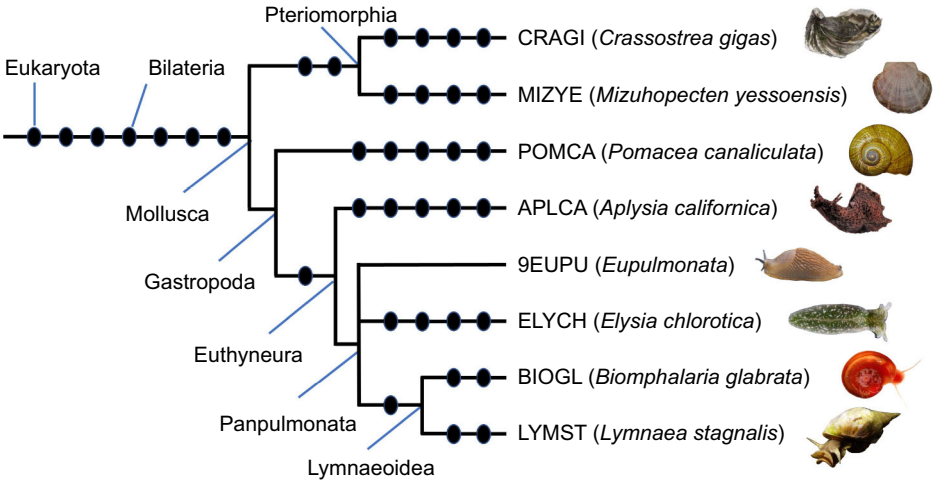

Fig. 5. Phylogenetic tree of organisms with matching entries to the *L. stagnalis* transcripts of the NCBI PRJDB98 dataset. The phyloT V2 (https://phylot.biobyte.de/) branch-based tree was generated based on the NCBI taxonomy database, using MIZYE, CRAGI, BIOGL, 9EUPU, ELYCH, LYMST, APLCA and POMCA as NCBI tree elements. Branching points (nodes) which represent the ancestral organisms and all of the descendants of the terminal taxa (leaf nodes) have been labelled and ovoid shapes on the branches of the tree are unlabelled inner nodes with one child branch. Taxons that share a branch are more closely related to each other compared with other taxa (e.g. *L. stagnalis* is more closely related to *B. glabrata* than to *M. yessoensis*). All pictures from https://commons.wikimedia.org.

for the organisms identified from all molluscan entries in the NCBI Bioproject PRJDB98 assemblies (Fig. 2). The similarity of the two distributions indicates that the LymCNS-PDB can be successfully used for the identification of proteins with extracted samples from *L. stagnalis* CNS without any bias towards one of the matched molluscan organisms.

To compare our database with amino acid sequences specific for *L. stagnalis* with other studies that were using a non-specific database for the identification of proteins from this species (Giusti et al., 2013; Rosenegger et al., 2010; Silverman-Gavrila et al., 2011), we prepared a molluscan database with all proteins from our LymCNS-PDB, using the amino acid sequences of their matching counterparts from other molluscs instead of the *L. stagnalis* sequences derived from the NCBI PRJDB98 transcriptome dataset. In comparison to the 3810 unique proteins identified using the LymCNS-PDB the unspecific molluscan database led to the identification of only 920 unique proteins, which presumably contain orthologous sequences in evolutionarily conserved regions, homologues to the sequences of certain tryptic peptides from the *L. stagnalis* CNS.

This comparison clearly shows the benefits of using a proteomics database with specific amino acid sequences for the organism under investigation, as using the LymCNS-PDB for the identification of proteins from the *L. stagnalis* CNS resulted in a 4 times higher number of identified proteins compared with using the non-specific molluscan database.

We have generated the most extensive proteomics database to date that is specifically for proteins from the CNS of *L. stagnalis*, the LymCNS-PDB. We have successfully used this database, as a proof of principle, to identify proteins in the isolated *L. stagnalis* CNS. Recently, we also successfully used it to reveal quantitative protein expression differences between CNS preparations made from classically conditioned and control animals (Anagnostopoulou et al., 2021b), food-deprived and satiated animals (Eravci et al., 2021), and young and aged animals (Anagnostopoulou et al., 2021a).

The LymCNS-PDB database provides a valuable tool to open new avenues for future research on proteomics to identify and quantify a plethora of proteins, which are involved in the molecular mechanisms of different neurobiological functions in the CNS of *L. stagnalis*, including learning and memory formation, ageing, age-related memory impairment as well amyloid-β-induced memory decline, feeding patterns, defensive responses and neuro-hormonal behavioural circuits involved in reproduction.

Acknowledgements
The authors would like to thank Dr Paul Johnston from the Freie Universität Berlin for analysis of the NCBI PRJDB98 transcriptome dataset with Trinity TransDecoder and providing the transcoded dataset. For mass spectrometry, we acknowledge the assistance of the Core Facility BioSupraMol supported by the Deutsche Forschungsgemeinschaft (DFG).

Competing interests
The authors declare no competing or financial interests.

Author contributions
Conceptualization: S.W., F.P., M.E.; Methodology: S.W., B.K., F.P., M.E.; Software: S.W., F.P., M.E.; Validation: M.E.; Formal analysis: S.W., F.P., M.E.; Investigation: A.A., B.K., M.C., M.E.; Resources: M.E.; Data curation: S.W., M.E.; Writing - original draft: M.E.; Writing - review & editing: A.A., B.K., M.C., P.R.B., I.K., G.K., M.E.; Visualization: M.E.; Supervision: F.P., I.K., G.K., M.E.; Project administration: I.K., G.K., M.E.; Funding acquisition: I.K., G.K.

Funding
This work was funded by Biotechnology and Biology Research Council grant BBSRC/BB/P00766X/1 (to I.K., G.K. and P.R.B.). Open Access funding provided by University of Sussex. Deposited in PMC for immediate release.

Data availability
PRoteomics IDEntifications Database (PRIDE): ProteomeXchange accession no. PXD025591. Project Webpage: http://www.ebi.ac.uk/pride/archive/projects/PXD025591

References

Anagnostopoulou, A., Eravci, M., Crossley, M., Kleine, P., Wayne, S., Benjamin, P. R., Kemenes, G. and Kemenes, I. (2021a). The effects of ageing upon cAMP response element-binding (CREB) proteins and differential protein expression in the CNS of Lymnaea stagnalis. BNA 2021 Festival of Neuroscience Poster abstracts. *Brain and Neuroscience Advances* 5, 26.

Anagnostopoulou, A., Eravci, M., Felletar, I., Benjamin, P. R., Crossley, M., Kemenes, G. and Kemenes, I. (2021b). The involvement of Cyclic AMP response element-binding protein (CREB) transcription factors and global protein expression in memory consolidation in Lymnaea stagnalis. BNA 2021 Festival of Neuroscience Poster abstracts. *Brain and Neuroscience Advances* 5, 71.

Benjamin, P. R. (2008). *Lymnaea*. *Scholarpedia* **3**, 4124. doi: 10.4249/scholarpedia.4124

Benjamin, P. R. and Kemenes, I. (2020). Peptidergic systems in the pond snail *Lymnaea*: From genes to hormones and behaviour. In *Advances in Invertebrate (Neuro)Endocrinology* (ed. S. A.B.L. and I. Orchard), pp. 213-254. Apple Academic Press.

Benjamin, P. R., Kemenes, G. and Staras, K. (2021). Molluscan nervous systems. *eLS Neuroscience* **2**, 1-15.

Bouetard, A., Noirot, C., Besnard, A. L., Bouchez, O., Choisne, D., Robe, E., Klopp, C., Lagadic, L. and Coutellec, M. A. (2012). Pyrosequencing-based transcriptomic resources in the pond snail Lymnaea stagnalis, with a focus on genes involved in molecular response to diquat-induced stress. *Ecotoxicology* **21**, 2222-2234. doi:10.1007/s10646-012-0977-1

Crossley, M., Staras, K. and Kemenes, G. (2016). A two-neuron system for adaptive goal-directed decision-making in *Lymnaea*. *Nat. Commun.* **7**, 11793-11805. doi:10.1038/ncomms11793

Crossley, M., Staras, K. and Kemenes, G. (2018). A central control circuit for encoding perceived food value. *Sci. Adv.* **4**, eaau9180-eaau9190. doi:10.1126/sciadv.aau9180

Davison, A. and Blaxter, M. L. (2005). An expressed sequence tag survey of gene expression in the pond snail *Lymnaea stagnalis*, an intermediate vector of trematodes [corrected]. *Parasitology* **130**, 539-552. doi:10.1017/S0031182004006791

Dong, N., Bandura, J., Zhang, Z., Wang, Y., Labadie, K., Noel, B., Davison, A., Koene, J. M., Sun, H. S., Coutellec, M. A. et al. (2021). Ion channel profiling of the *Lymnaea stagnalis* ganglia via transcriptome analysis. *BMC Genomics* **22**, 18-42. doi:10.1186/s12864-020-07287-2

Eravci, M., Anagnostopoulou, A., Crossley, M., Franklin, J., Singh, G., Lalji, N., Benjamin, P. R., Kemenes, I., Kemenes, G. (2021). Effects of hunger state on protein expression and CREB phosphorylation in the nervous system of *Lymnaea stagnalis*. In BNA 2021 Festival of Neuroscience Poster abstracts. Brain and Neuroscience Advances. 5:125.

Eravci, M., Sommer, C. and Selbach, M. (2014). IPG strip-based peptide fractionation for shotgun proteomics. *Methods Mol. Biol.* **1156**, 67-77. doi:10.1007/978-1-4939-0685-7_5

Feng, Z.-P., Zhang, Z., van Kesteren, R. E., Straub, V. A., van Nierop, P., Jin, K., Nejatbakhsh, N., Goldberg, J. I., Spencer, G. E., Yeoman, M. S. et al. (2009). Transcriptome analysis of the central nervous system of the mollusc *Lymnaea stagnalis*. *BMC Genomics* **10**, 451-465. doi:10.1186/1471-2164-10-451

Fodor, I., Hussein, A. A., Benjamin, P. R., Koene, J. M. and Pirger, Z. (2020a). The unlimited potential of the great pond snail, *Lymnaea stagnalis*. *Elife* **9**, e56962-e56979. doi:10.7554/eLife.56962

Fodor, I., Urban, P., Kemenes, G., Koene, J. M. and Pirger, Z. (2020b). Aging and disease-relevant gene products in the neuronal transcriptome of the great pond snail (*Lymnaea stagnalis*): a potential model of aging, age-related memory loss, and neurodegenerative diseases. *Invert. Neurosci.* **20**, 9-13. doi:10.1007/s10158-020-00242-6

Fodor, I., Svigruha, R., Kemenes, G., Kemenes, I. and Pirger, Z. (2021). The great pond snail (*Lymnaea stagnalis*) as a model of aging and age-related memory impairment: an overview. *J. Gerontol. A Biol. Sci. Med. Sci.* **76**, 975-982. doi:10.1093/gerona/glab014

Ford, L., Crossley, M., Williams, T., Thorpe, J. R., Serpell, L. C. and Kemenes, G. (2015). Effects of Abeta exposure on long-term associative memory and its neuronal mechanisms in a defined neuronal network. *Sci. Rep.* **5**, 10614-10628. doi:10.1038/srep10614

Ford, L., Crossley, M., Vadukul, D. M., Kemenes, G. and Serpell, L. C. (2017). Structure-dependent effects of amyloid-beta on long-term memory in *Lymnaea stagnalis*. *FEBS Lett.* **591**, 1236-1246. doi:10.1002/1873-3468.12633

Giusti, A., Leprince, P., Mazzucchelli, G., Thomé, J.-P., Lagadic, L., Ducrot, V. and Joaquim-Justo, C. (2013). Proteomic analysis of the reproductive organs of the hermaphroditic gastropod *Lymnaea stagnalis* exposed to different endocrine disrupting chemicals. *PLoS One* **8**, e81086-e81099. doi:10.1371/journal.pone.0081086

Hatakeyama, D., Sadamoto, H., Watanabe, T., Wagatsuma, A., Kobayashi, S., Fujito, Y., Yamashita, M., Sakakibara, M., Kemenes, G. and Ito, E. (2006). Requirement of new protein synthesis of a transcription factor for memory consolidation: paradoxical changes in mRNA and protein levels of C/EBP. *J. Mol. Biol.* **356**, 569-577. doi:10.1016/j.jmb.2005.12.009

Ito, E., Otsuka, E., Hama, N., Aonuma, H., Okada, R., Hatakeyama, D., Fujito, Y. and Kobayashi, S. (2012). Memory trace in feeding neural circuitry underlying conditioned taste aversion in Lymnaea. *PLoS One* **7**, e43151-e43156. doi:10.1371/journal.pone.0043151

Kemenes, I., Kemenes, G., Andrew, R. J., Benjamin, P. R. and O'Shea, M. (2002). Critical time-window for NO-cGMP-dependent long-term memory formation after one-trial appetitive conditioning. *J. Neurosci.* **22**, 1414-1425. doi:10.1523/JNEUROSCI.22-04-01414.2002

Kemenes, G., Kemenes, I., Michel, M., Papp, A. and Muller, U. (2006). Phase-dependent molecular requirements for memory reconsolidation: differential roles for protein synthesis and protein kinase A activity. *J. Neurosci.* **26**, 6298-6302. doi:10.1523/JNEUROSCI.0890-06.2006

Marra, V., Kemenes, I., Vavoulis, D., Feng, J., O'Shea, M. and Benjamin, P. R. (2010). Role of tonic inhibition in associative reward conditioning in *Lymnaea*. *Front. Behav. Neurosci.* **4**, 161-167. doi:10.3389/fnbeh.2010.00161

Naskar, S., Wan, H. and Kemenes, G. (2014). pT305-CaMKII stabilizes a learning-induced increase in AMPA receptors for ongoing memory consolidation after classical conditioning. *Nat. Commun.* **5**, 3967-3997. doi:10.1038/ncomms4967

Pirger, Z., Naskar, S., Laszlo, Z., Kemenes, G., Reglodi, D. and Kemenes, I. (2014). Reversal of age-related learning deficiency by the vertebrate PACAP and IGF-1 in a novel invertebrate model of aging: the pond snail (*Lymnaea stagnalis*). *J. Gerontol. A Biol. Sci. Med. Sci.* **69**, 1331-1338. doi:10.1093/gerona/glu068

Ribeiro, M. J., Serfozo, Z., Papp, A., Kemenes, I., O'Shea, M., Yin, J. C., Benjamin, P. R. and Kemenes, G. (2003). Cyclic AMP response element-binding (CREB)-like proteins in a molluscan brain: cellular localization and learning-induced phosphorylation. *Eur. J. Neurosci.* **18**, 1223-1234. doi:10.1046/j.1460-9568.2003.02856.x

Ribeiro, M. J., Schofield, M. G., Kemenes, I., O'Shea, M., Kemenes, G. and Benjamin, P. R. (2005). Activation of MAPK is necessary for long-term memory consolidation following food-reward conditioning. *Learn. Mem.* **12**, 538-545. doi:10.1101/lm.8305

Rivi, V., Benatti, C., Colliva, C., Radighieri, G., Brunello, N., Tascedda, F. and Blom, J. M. C. (2020). *Lymnaea stagnalis* as model for translational neuroscience research: From pond to bench. *Neurosci. Biobehav. Rev.* **108**, 602-616. doi:10.1016/j.neubiorev.2019.11.020

Rosenegger, D., Wright, C. and Lukowiak, K. (2010). A quantitative proteomic analysis of long-term memory. *Mol. Brain* **3**, 9-18. doi:10.1186/1756-6606-3-9

Sadamoto, H., Sato, H., Kobayashi, S., Murakami, J., Aonuma, H., Ando, H., Fujito, Y., Hamano, K., Awaji, M., Lukowiak, K. et al. (2004). CREB in the pond snail *Lymnaea stagnalis*: cloning, gene expression, and function in identifiable neurons of the central nervous system. *J. Neurobiol.* **58**, 455-466. doi:10.1002/neu.10296

Sadamoto, H., Takahashi, H., Okada, T., Kenmoku, H., Toyota, M. and Asakawa, Y. (2012). De novo sequencing and transcriptome analysis of the central nervous system of mollusc *Lymnaea stagnalis* by deep RNA sequencing. *PLoS One* **7**, e42546-e42558. doi:10.1371/journal.pone.0042546

Silverman-Gavrila, L. B., Senzel, A. G., Charlton, M. P. and Feng, Z. P. (2011). Expression, phosphorylation, and glycosylation of CNS proteins in aversive operant conditioning associated memory in *Lymnaea stagnalis*. *Neuroscience* **186**, 94-109. doi:10.1016/j.neuroscience.2011.04.027

Staras, K., Kemenes, I., Benjamin, P. R. and Kemenes, G. (2003). Loss of self-inhibition is a cellular mechanism for episodic rhythmic behavior. *Curr. Biol.* **13**, 116-124. doi:10.1016/S0960-9822(02)01435-5

Steinegger, M. and Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026-1028. doi:10.1038/nbt.3988

Tyanova, S., Temu, T. and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301-2319. doi:10.1038/nprot.2016.136

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506-D515. doi:10.1093/nar/gky1049