Shotgun proteomic analysis of saliva and salivary
gland tissue from the common octopus *Octopus
vulgaris*

AUTHOR NAMES

Legana C.H.W. Fingerhut[†], Jan M. Strugnell[‡], Pierre Faou[§], Álvaro Roura Labiaga[‡], Jia
Zhang[†], Ira R. Cooke^{†,§}

AUTHOR ADDRESSES

[†]Department of Molecular and Cell Biology, James Cook University, Townsville, Queensland
4811, Australia

[‡]Centre for Sustainable Tropical Fisheries and Aquaculture, College of Science and
Engineering, James Cook University, Townsville, Queensland 4811, Australia

[§]Department of Ecology, Environment and Evolution, School of Life Sciences, La Trobe
University, Melbourne, Vic 3086, Australia

[§]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La
Trobe University, Melbourne, Victoria 3086, Australia

[‡]Department of Ecology and Marine Biodiversity, Instituto de Investigaciones Marinas de
Vigo (IIM-CSIC), Vigo 36208, Spain

ABSTRACT

The saliva of the common octopus (*Octopus vulgaris*) has been the subject of biochemical study for over a century. A combination of bioassays, behavioural studies and molecular analysis on *O. vulgaris* and related species suggests that it should contain a mixture of highly potent neurotoxins and degradative proteins. However, a lack of genomic and transcriptomic data has meant that the amino acid sequences of these proteins remain almost entirely unknown. To address this, we assembled the salivary gland transcriptome of *O. vulgaris* and combined it with high resolution mass spectrometry data from the posterior and anterior salivary glands of two adults, the posterior salivary glands of six paralarvae and the saliva from a single adult. We identified a total of 2810 protein groups from across this range of salivary tissues and age classes, including 84 with homology to known venom protein families. Additionally, we found 21 short secreted cysteine rich protein groups of which 12 were specific to cephalopods. By combining protein expression data with phylogenetic analysis we demonstrate that serine proteases expanded dramatically within the cephalopod lineage and that cephalopod specific proteins are strongly associated with the salivary apparatus.

Keywords

cephalopod, saliva, proteogenomics, proteomics, transcriptomics, *Octopus vulgaris*, venom, venomics, serine protease, tandem mass spectrometry

INTRODUCTION

Animal venoms are valuable for biodiscovery. Many animal venoms contain peptides with a high degree of stability, selectivity, and potency that make them excellent candidates for biomimetic drugs ¹. A variety of venom derived drugs have already been developed ² with numerous others undergoing clinical trials ¹. Historically, venom research has been biased towards a limited range of taxa including snakes, scorpions, spiders, and cone snails from which venom is relatively easy to obtain and/or is available in large quantities ¹. The venoms of a phylogenetically diverse range of taxa such as annelids (e.g. leeches), arthropods (e.g. spiders, wasps) cnidarians (e.g. jellyfish, coral), echinoderms (e.g. sea urchins, starfish), and molluscs, notably the coleoid cephalopods (i.e. squid, cuttlefish and octopus) remain substantially unexamined ^{1,3}.

Coleoid cephalopods comprise of over 800 species that span many orders of magnitude in body size and occupy marine habitats worldwide ⁴. The loss of the ancestral molluscan protective shell in the evolution of coleoid cephalopods, in conjunction with their predatory nature, has likely driven venom evolution in the group - enabling efficient handling and immobilisation of prey ⁵.

Venom systems within the subclass are best known from benthic predatory species, in particular octopods and cuttlefishes, which consume a wide range of prey including crustaceans, other molluscs, and fish ^{6,7}. Venom is typically injected into prey via a bite through soft parts such as eyes ⁸ or by a hole drilled into the shell ⁷.

Much of the early research investigating the composition of octopod venom employed bioassays and bioassay guided fractionation techniques ⁹⁻¹². These techniques provided the

ability to associate a fraction with its specific activity. Subsequently, these studies observed a wide range of biochemical activities from cephalopod venom including flaccid paralysis^{10,13}, blood coagulation inhibition¹¹, external digestion¹⁴, haemolysis^{15,16}, and detachment of muscle from the exoskeleton^{12,14} or shell⁷. Taken together this demonstrated that coleoid cephalopod venom includes components to fulfil a diversity of roles in prey capture, handling and digestion. This suggested that octopod venom contains a wide variety of highly bioactive molecules, of which many were proteinaceous. However, molecular characterisation in early studies was often limited to reporting a molecular mass, the isoelectric point of a fraction^{9,12} or simply the bulk amino acid composition¹⁷. Indeed, within octopod venom, only one such protein, eledoisin¹⁸, has been sequenced. More recent studies have used transcriptomic and proteomic surveys of the venom producing posterior salivary glands (PSGs) of cephalopods to detect proteins from families known to be associated with toxic activity in other taxa^{4,19,20}. Key protein families identified include CAPs (Cysteine rich secretory proteins, Antigen 5, Pathogenesis related), carboxypeptidases, chitinases, hyaluronidases, phospholipase A₂ proteins, DNase, tachykinin like peptides, SE-cephalotoxin and serine proteases.

Serine proteases have been implicated in multiple venom related activities including proteolysis in digestion and envenomation as well as anticoagulation which consequently amplifies other toxins^{4,19}. Highly diverse and abundant serine proteases have been observed in transcriptomic studies of the venom of platypuses²¹, the venom glands of remipedes²² and the PSGs of a wide range of cephalopods¹⁹. Phylogenetic analyses of cephalopod serine proteases indicate that these expanded in the ancestor of coleoid cephalopods^{19,20} but since serine proteases are ubiquitously present and diverse in all animals, further work is required to specifically implicate these proteins in the cephalopod venom arsenal. In this paper we will present evidence to show that serine proteases not only underwent a dramatic expansion in the coleoid

lineage, but that this expansion is specifically linked to expression within the salivary glands and saliva itself.

The diversity of venom components in octopods is likely to have been underestimated because previous studies have focussed primarily on adults^{4,19} despite evidence that juveniles also use venom for predation²³. Molecular analyses of the liquefied gut content of octopod paralarvae (planktonic juveniles) detected predominately prawn and crab zoeae, despite the fact that these were a relatively sparse component of the plankton, indicating that these taxa were targeted by paralarvae²⁴. Studies of snakes²⁵ and jellyfish²⁶ have shown that diet changes the biochemical composition of venom which makes it likely that variation in venom composition between life stages of octopus would exist due to the different dietary preferences these exhibit.

In this study a combined venomomics approach will be used to investigate the salivary proteome of the common octopus (*Octopus vulgaris*) in detail. Unlike previous work which focussed purely on PSGs of adults^{4,19} this study includes samples from adult and juvenile PSG, anterior salivary glands (ASGs) of adults as well as the first proteomic analysis of pure octopod saliva. We generate a reference protein database by *de novo* assembly of short read transcriptome data from adult PSG and then combine this with high resolution tandem mass spectrometry data from PSGs of different paralarval life stages as well as ASGs, PSGs, and milked saliva from adults. We combine this comprehensive salivary proteome information with phylogenetic analysis to demonstrate that evolution of cephalopod serine proteases is strongly linked with their expression in saliva.

EXPERIMENTAL SECTION

Sample collection and preparation

Three female and one male adult *O. vulgaris* individuals were caught with artisanal fishing equipment from Ría de Vigo, Spain, in January and February 2013. Octopuses were transported and maintained at the aquaculture facilities of Instituto Español de Oceanografía de Vigo (IEO) and used as broodstock. An additional two male adult octopuses were collected using the same methods at the same location in July 2013. The females spawned naturally, (i.e. they were not induced), between April-May in individual tanks containing an artificial den. Paralarvae hatched approximately 55 days later and all batches were translocated to a 100 litre tank two days post hatching. Two paralarvae were randomly selected at days 1, 20 and 30. The following harvest procedures complied with the Directive 2010/63EU²⁷. The animals were anaesthetised in a solution of seawater and 1.5% magnesium chloride hexahydrate for 10 minutes which was followed by euthanasia by immersion in 3.5% magnesium chloride hexahydrate for 30 minutes. A total of six octopus paralarvae were stored in 80% ethanol at -20°C for two years.

Both PSGs from each paralarvae were dissected under a binocular microscope with 40x magnification in July 2015. One adult octopus was milked using the plastic bag technique as described by Ballering et al.²⁸ Although the animal did bite through the bag, it did not squirt saliva (as depicted in Grisley and Boyle¹²) and the substance collected was scraped from around the bite hole. This was followed by manual dissection of both anterior and posterior salivary glands. Three samples of the salivary glands and saliva from different adult octopuses were extracted and stored at -80°C. PSGs, ASGs and saliva samples were lyophilised and stored at -20°C and PSGs were kept in RNA later in -80°C for the transcriptomic analysis.

RNA extraction, sequencing and transcriptome assembly

RNA was extracted from the posterior salivary gland of a single individual using an RNeasy® Mini Kit. Samples were homogenised using an IKA® desktop homogeniser. Transcriptomes were sequenced by Australian Genome Research Facility (AGRF) using an Illumina HiSeq2000 with 3 µg of RNA to produce a total of 19.1 million 100bp paired end reads.

Trinity (v2.2.0)²⁹ was used to assemble RNA sequencing data from PSG (above). *In silico* read normalisation and read trimming was performed automatically in Trinity and all other settings were left as the default. The resulting assembly had 46,519 contigs and was deposited in the National Center for Biotechnology Information (NCBI)'s transcriptome sequence archive under accession PRJNA464423. Transcripts were then annotated using Trinotate (v3.0.1)(<https://trinotate.github.io>). Trinotate uses TransDecoder to predict coding sequences and performs homology-based functional annotation for both coding and non-coding sequences. Trinotate annotations are based on Basic Local Alignment Search Tool (BLAST)³⁰ search matches to the SwissProt database, HMMER³¹ based searches for Pfam³² conserved domains, signal peptide prediction with SignalP (v4.1)³³ and transmembrane prediction with tmHMM³⁴.

Mass spectrometry

All samples were first dried using a SpeedVac Concentrator for 15 minutes before solubilisation into 50 µL of 7M Urea, 3M Tris pH=8.3. Proteins were then reduced overnight using 0.5 µL of TCEP (tris [2-carboxyethyl] phosphine hydrochloride, 200 mM solution in

water). The alkylation step was as followed: addition of 2 μ L IAA (iodoacetamide) at 1M and incubation for one hour in the dark. The samples were then diluted with 500 μ L of 50 mM Tris (pH 8.3) before addition of one μ g trypsin. Samples were digested overnight at 37°C. The digests were acidified with 1% (v/v) trifluoroacetic acid (TFA) and the peptides desalted on SDB-XC (Empore) StageTips as previously described³⁵.

Two different mass spectrometers were utilised for this study. An Orbitrap Elite™ hybrid ion-trap orbitrap mass spectrometer was used for the paralarvae tissue samples in August 2015. A Q Exactive™ high field orbitrap tandem mass spectrometer was used to analyse the adult tissue samples and saliva in March 2016 (see below for details).

Paralarvae tissue peptides reconstituted in 0.1% TFA and 2% acetonitrile (ACN) were loaded using a Thermo Scientific™ UltiMate™ 3000 RSLCnano system onto a trap column (C18 PepMap 300 μ m ID $\times$ 2 cm trapping column, Thermo-Fisher Scientific) at 15 μ l/min for six minutes. The valve was then switched to allow the precolumn to be in line with the analytical column (Vydac MS C18, 3 μ m, 300 Å and 75 μ m ID $\times$ 25 cm, Grace Pty. Ltd.). The separation of peptides was performed at 300 nl/min at 45 °C using a non-linear ACN gradient of buffer A (0.1% formic acid, 2% ACN) and buffer B (0.1% formic acid, 80% ACN), starting at 5% buffer B to 45% over 105 minutes, then 95% B for five minutes followed by an equilibration step of 15 minutes (0.1% formic acid, 2% ACN). Data were collected on an Orbitrap Elite (Thermo-Fisher Scientific) in Data Dependent Acquisition mode using m/z 300–1500 as MS scan range, CID MS/MS spectra were collected for the 10 most intense ions. Dynamic exclusion parameters were set as follows; repeat count 1, duration 90 seconds, the exclusion list size was set at 500 with early expiration disabled. Other instrument parameters for the Orbitrap were the following: MS scan at 120 000 resolution, maximum injection time 150 ms, AGC target 1

× 10⁶, CID at 35% energy for a maximum injection time of 75 ms with AGT target of 5000. The Orbitrap Elite was operated in dual analyser mode with the Orbitrap analyser being used for MS and the linear trap being used for MS/MS.

For the analysis of the adult tissue samples and saliva on the Q Exactive mass spectrometer, the peptides were reconstituted in 0.1% TFA and 2% acetonitrile (ACN) and loaded at 45 °C onto a C18 PepMap 300 μm ID × 2 cm trapping column (Thermo-Fisher Scientific) at 10 μl/min for six minutes, using a Thermo Scientific™ UltiMate™ 3000 RSLCnano system and washed for six minutes before switching the pre-column in line with the analytical column (BioSphere C18, 1.9 μm, 120 Å and 75 μm ID × 40 cm, NanoSeparation). The separation of peptides was performed at 45 °C, 250 nl/min using a non-linear ACN gradient of buffer A (0.1% formic acid, 2% ACN) and buffer B (0.1% formic acid, 80% ACN), starting from 2% buffer B to 12% B in six min and then to 33% B over 60 minutes followed by 50% B at 70 min. The gradient in then increased from 50% B to 95% B for one min and stayed at 95% B from five min. The column is then equilibrated for 15 minutes (0.1% formic acid, 2% ACN). Data were collected on a Q Exactive HF (Thermo-Fisher Scientific) in Data Dependent Acquisition mode using m/z 350–1500 as MS scan range at 60 000 resolution, HCD MS/MS spectra were collected for the 15 most intense ions per MS scan at 15 000 resolution. Dynamic exclusion parameters were set as follows: exclude isotope on, duration 30 seconds and peptide match preferred. Other instrument parameters for the Orbitrap were: MS maximum injection time 30 ms with AGC target 3 × 10⁶, collision at 28% energy for a maximum injection time of 25 ms with AGT target of 1 × 10⁵ and an isolation window of 1.4 m/z.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE³⁶ partner repository with the dataset identifier PXD010298

Proteogenomics and reference protein database construction

A compact and relatively comprehensive protein database for *O. vulgaris* saliva and salivary glands was constructed by combining predicted coding sequences from the *de novo* assembled *O. vulgaris* transcriptome along with sequences that would otherwise have been rejected as non-coding but whose presence in saliva or PSG was supported by mass spectrometry. Our workflow for constructing this database was similar to that described previously in Whitelaw et al.²⁰ and Caruana et al.³⁷. We used TransDecoder, embedded in the Trinity software version 3.0.1 to predict coding sequences from the transcriptome. We termed this predicted set of proteins “known”. In addition to the “known” proteins we generated a “novel” set consisting of all six-frame translations of transcript sequences longer than 15 amino acids. We combined “known” and “novel” sets into a single database and then searched this against our complete mass spectrometry dataset which consisted of both Q-Exactive and Orbitrap Elite spectra across all samples of PSG and saliva from all stages, 15 samples, (see mass spectrometry section) using the search engines X!Tandem³⁸ and MS-GF+³⁹. Search engine parameters were: Methionine oxidation and N-terminal acetylation for variable modifications, Carbamidomethyl cysteines for fixed modification, precursor mass error 20ppm for both instruments and fragment ion error 0.4 Da for Orbitrap Elite Spectra and 0.1 Da for Q-Exactive spectra. Allowed charges were set to 2+, 3+ and 4+, the cleavage enzyme used was Trypsin with up to two missed cleavages and semicleaved peptides allowed.

Since “known” and “novel” sets will have very different expected error distributions we separated peptide spectrum matches (PSM’s) returned from our search engines into “known” and “novel” sets for downstream statistical analysis with PeptideProphet⁴⁰ and iProphet⁴¹. This resulted in a single iProphet generated pepXML file from each set which we then combined

using ProteinProphet⁴² to generate a single protXML file. Our final protein database was then constructed using all “novel” proteins included in this file along with the full set of “known” proteins from TransDecoder.

All analyses were performed using wrapper scripts for search engines and downstream peptide and protein inference tools implemented as part of the protk (<https://github.com/iracooke/protk>) toolkit, with the workflow implemented as a Rake script.

Quantitative proteomic analysis

In order to perform a quantitative analysis across tissues and life stages, the reference protein database was combined with the mass spectrometry data and analysed in MaxQuant (version 1.6.1.0)⁴³. MaxQuant settings included Carbamidomethyl C as a fixed modification and Oxidation of Methionine and Acetylation of protein N-terminus as variable modifications, decoys and contaminants were included automatically, MS/MS tolerance was set according to instrument type (20ppm for Q-Exactive, 0.5Da for Orbitrap Elite), two missed tryptic cleavages were allowed and PSMs were accepted at a 1% false discovery rate (FDR). To allow comparisons based on approximate absolute protein abundance an intensity Based Absolute Quantification (iBAQ) score was calculated by MaxQuant. The standard MaxQuant Label Free Quantification (LFQ) intensity was used for comparisons of the same protein group across tissues.

Protein sequence analysis and homology

All proteins identified by mass spectrometry were annotated using a suite of methods as follows. Sequence length and overall cysteine content relative to the sequence length (cysteine richness) were quantified. SignalP was used to identify signal peptides. Molecular weight and isoelectric point of each protein was calculated using the ProtParam tool within the Biopython⁴⁴ software package. Functional annotation was inferred by homology against curated databases of venom proteins (ToxProt, ArachnoServer) as well as general proteins (SwissProt). *O. vulgaris* proteins with homology to proteins in these databases were identified using protein-protein BLAST (BLASTP) based on an E-value threshold of $<1 \times 10^{-5}$. Where multiple hits existed (across different databases) the closest homolog (lowest E-value) was retained as the primary annotation.

To identify proteins with regions of high cysteine density we scanned each protein amino acid sequence using a sliding window and counted the number of cysteines within each window; (see https://github.com/Legana/Cysteine_Rich_Region_Finder for implementation). For each protein we recorded the maximum cysteine count resulting from this scan which we refer to as the Cysteine Richness Index (CRI). To determine the threshold CRI and optimal window size for putative toxin discovery, we first calibrated the CRI by analysing all the protein sequences in two venom protein databases: ArachnoServer (spider toxin database) and ToxProt and compared these to a general protein database, SwissProt (see Supplementary Figure S1). On this basis, a window width of 30 amino acids and a threshold CRI of 5 or larger was chosen for selecting proteins most likely to contain cysteine knots with a moderate number of cysteines (~6). While it is certainly true that toxin related cysteine arrangements exist with various numbers of cysteines⁴⁵, these are likely to be captured using an overall combination of cysteine density (measured on the whole protein) and short protein length. Our measure (the CRI) is

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

specifically designed to capture cysteine rich regions indicative of small cysteine knots embedded within a larger (cysteine poor) precursor protein.

Identification of cephalopod specific proteins

In order to identify proteins that were unique to the cephalopod lineage we first searched for homologs in the NCBI non-redundant (nr) database using BLASTP with a relaxed E-value threshold of 1×10^{-3} . All hits up to a maximum of 500 matching each protein were retrieved from this search along with the NCBI taxon identifier of the corresponding species. Taxon identifiers of BLAST search results were then compared against a comprehensive list of taxon identifiers corresponding to the cephalopod lineage. Proteins that either had no BLAST hit, or belonged exclusively to cephalopods, were classified as cephalopod specific.

Phylogenetic analysis of serine proteases

We investigated relationships between the evolutionary history of serine proteases and their relative expression across saliva and salivary glands as follows. Firstly, we collected all cephalopod protein sequences available in the UniProt database (accessed April 2018). These were combined with all sequences identified in the salivary proteome of *O. vulgaris* and with predicted proteins from transcriptome assemblies for the southern blue-ringed octopus (*Hapalochlaena maculosa*)²⁰, the southern sand octopus (*Octopus kaurana*)²⁰ and the common cuttlefish (*Sepia officinalis*) (assembly from sequences available on SRA, see Whitelaw et al.²⁰ for details). We also included predicted proteins available on UniProt for representatives of two other major molluscan groups, a bivalve - the Pacific oyster (*Crassostrea gigas*) and a model gastropod - the owl limpet (*Lottia gigantea*)(see Supplementary Table S1 for full

sequence list). All protein sequences from all taxa were scanned for conserved domains using InterproScan (v5) ⁴⁶ and serine proteases were identified by the presence of Pfam domain (PF00089) from these results. From these proteins all full length serine protease domains (domain length >200AA) were extracted and aligned using MAFFT v7.309 plugin ^{47, 48} implemented in Geneious (version 11.1.3) ⁴⁹. The resulting alignment was trimmed to remove sites with >50% gaps.

The final serine protease alignment contained a total of 325 sequences of which the majority (238) were from a relatively small number of species (*H. maculosa*, *O. kaurna*, *S. officinalis*, *L. gigantea*, *C. gigas*) for which comprehensive whole genome or multi-tissue transcriptome sequencing data are available. The remaining sequences comprised 49 from two studies that performed low depth transcriptome sequencing of posterior salivary glands across a wide range of cephalopod species ^{4, 19}, 33 from the present study, three from the liver of *Heterololigo bleekerii*, one from whole viscera of *Doryteuthis opalescens* and one from *Euprymna scolopes* (tissue not specified).

Phylogenetic analysis based on the aligned serine protease sequences was performed using IQ-TREE ⁵⁰. The optimal evolutionary model (WAG + I + G4) was selected using ModelFinder ⁵¹ followed by a maximum likelihood consensus tree inference using 1000 ultrafast bootstraps ⁵². The consensus tree was imported into RStudio using APE ⁵³ and visualised (unrooted) using ggtree ⁵⁴. This large tree (325 sequences) was used to identify higher level evolutionary relationships, from which a subset of sequences comprising a cephalopod specific clade and its closest outgroup sequence were identified (Figure 3). Sequences belonging to the cephalopod specific clade (along with three outgroup *L. gigantea* sequences) were then plotted alongside

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

proteomic expression data for *O. vulgaris* using the gheatmap function from ggtree (Supplementary Figure S2).

RESULTS AND DISCUSSION

Global proteomic profile of salivary tissues and life stages

In this study we analysed the octopod salivary proteome in unprecedented breadth, including samples from anterior and posterior salivary glands, saliva itself and the posterior salivary glands of paralarvae. The overall proteomic profiles of these four broad classes of tissue are summarised in Figure 1 in terms of the relationship between samples (Figure 1B), and the relative abundance of families of key proteins (Figure 1A, Figure 1C). Multidimensional scaling analysis (Figure 1B) showed that despite being measured on different instruments, the proteomic profiles of adult and juvenile PSGs were much more similar to one another than to those of other tissues such as the ASGs and saliva itself. This suggests that paralarvae even as young as one day old have functional PSGs, being more similar to those of adults than to other tissues.

Figure 1. Global proteome profiles of saliva and related tissues for *Octopus vulgaris*. Part A shows the top 40 most abundant protein groups by sample type. Proteins are grouped into broad families and the abundance of each is shown using a single stacked bar. Individual segments within each bar represent abundance of individual proteins. Note that the horizontal scales of bars differ markedly between tissues and are highest for Adult PSG. Part B is a

multidimensional scaling (MDS) plot based on log normalised label-free quantification (LFQ) intensities of proteins present in all samples. Dots represent individual samples and samples are coloured according to tissue/life-stage. Part C shows the relative abundance of venom related protein families across different tissues. Bars are stacked with individual proteins represented as segments (as in Part A). Abbreviations are ASG: Anterior Salivary Gland, CAP: (Cysteine rich secretory proteins, Antigen 5, Pathogenesis related), PLA2: secreted Phospholipase A₂, PSG: Posterior Salivary Gland.

The most abundant proteins in all proteomic profiles included ubiquitous cellular “housekeeping” proteins such as those involved in DNA binding (Histones), cytoskeletal structure (Actin and Tubulin) and respiration (ATP synthase) (Figure 1A). Proteins that could not be characterised by homology to model species made up a surprisingly large fraction of the most abundant proteins in all tissues (Figure 1A). The fact that these proteins were not only diverse, but also abundant, highlights the relative lack of gene function information for cephalopods or their close relatives.

The proteomic profile of the saliva was somewhat enigmatic. It contained several abundant proteins not found in salivary glands that have previously been identified as key components of squid mucus³⁷ including several abundant intermediate filament proteins and a 70 kDa neurofilament-like protein (Figure 1A). Protein families thought to be key components of the octopod venom system,¹⁹ such as serine proteases, were not abundant in the saliva. Our conclusion is that while we were able to collect some form of salivary secretion it is unlikely to represent the same mixture known to be injected into prey. The saliva we collected was not ejected in a stream as observed in Ballering et al.²⁸, but rather trickled from the octopus’ mouth. Similar exuded fluid has also been analysed by Ballering et al.²⁸ and was not found to contain

1
2
3 proteolytic activity which is consistent with our finding that serine proteases were not abundant
4
5 in this fraction.
6
7
8
9

10 Venom related protein families

11
12
13
14

15 Proteome analysis of the PSG and saliva of *O. vulgaris* revealed 84 protein groups
16
17 homologous to protein families previously identified from venom constituents including: CAP,
18
19 carboxylesterase, chitinase, hyaluronidase, metalloprotease, pacifastin, phospholipase A2, and
20
21 serine protease peptidase s1^{4,55}. Of these, serine proteases and, to a lesser extent, CAP proteins
22
23 were dominant in both diversity and abundance across all salivary sample types. Serine
24
25 proteases were particularly dominant as a proportion of total venom protein iBAQ signal in
26
27 PSG from both adult and paralarval samples (Figure 1A, Figure 1C). CAP proteins were most
28
29 abundant in saliva and ASG.
30
31
32
33
34

35 We found a total of eight CAP proteins of which five were distinct and three were sequence
36
37 variants possibly arising from alternative splicing or fragmentation of our transcriptome
38
39 assembly. This is significantly more than the two found in *H. maculosa* and one in *O. kaurana*
40
41 by Whitelaw et al.²⁰ and reflects the inclusion of ASGs in this study which were a rich source
42
43 of CAPs (Figure 1C; Supplementary Table S2). The abundance of CAPs in *O. vulgaris* suggests
44
45 functional importance of this protein family in cephalopods. Two CAP proteins have also been
46
47 identified in slime secretions from the southern bottletail squid (*Sepiadarium austrinum*)³⁷.
48
49 CAP proteins have been identified from a variety of venomous taxa including spiders, cone
50
51 snails, and lizards and span a range of functions. Multiple experiments on the CRVP_HELHO
52
53 CAP protein within the Mexican beaded lizard (*Heloderma horridum horridum*) show partial
54
55 paralysis, lethargy and a decrease of body temperature in mice⁵⁶ as well as interference with
56
57
58
59
60

various calcium^{57, 58} and potassium channels⁵⁹ in rats. Several CAP proteins have been characterised from the venom of another molluscan lineage, the cone snails^{60,61}, however their biochemical role is uncertain⁶¹. Fry, Roelants and Norman⁴ previously identified a single CAP from *H. maculosa* but found that it was phylogenetically distinct from the CAP sequences identified from *Conus*. Our study confirms that CAP proteins are consistently present in octopus saliva and salivary glands but their role remains unclear.

Chitinases are a family of proteins expected to be present in cephalopod venom based on prior observations from salivary glands^{4,20}, and assays of the activity of saliva extracts^{12,16,62}. Two distinct chitinase protein groups were discovered in the *O. vulgaris* salivary proteome, both of which were secreted and cysteine rich (Supplementary Table S2). Although chitinases were present in ASG, PSG, adults and paralarvae (Figure 1C) only one of these (TRINITY_DN12896_c1_g1_i1) was present in all sample types whereas the other, (TRINITY_DN12584_c0_g1_i1) was restricted to saliva and paralarval PSGs.

Secreted phospholipase A2 (sPLA₂) hydrolyse phospholipids into lipid molecules. sPLA₂s are found in a variety of non-venomous and venomous taxa with various biological functions⁶³. sPLA₂s have convergently evolved for venom related purposes in cnidarians, insects, arthropods, cephalopods and reptiles⁵⁵. sPLA₂s have been categorised into multiple functional groups with varying toxic effects such as neurotoxicity and myotoxicity⁶⁴. The sPLA₂ group is diverse in venomous marine invertebrates⁶⁵. Previous studies on sPLA₂ in octopods have found phospholipase activity from PSG extracts across a range of Antarctic species¹⁵, and have identified a molecule homologous to the *O. vulgaris* sequence identified here in the: octopods *H. maculosa*, *O. kaurana*²⁰; cuttlefishes *Sepia latimanus*, *Sepia pharaonic*; and squid *Sepioteuthis australis*, *Loliolus noctiluca*¹⁹. It would be interesting to examine whether these

sPLA_ss belong to different functional groups between cephalopod species. Unlike serine proteases which appear to have massively diversified in octopods (see below;^{19,20}) we identified a single sPLA_s sequence in *O. vulgaris*.

Homologs to previously identified tachykinins from *O. vulgaris*⁶⁶ were present in the *O. vulgaris* transcriptome but absent in the proteome. The lack of discovered tachykinins in the proteome may be due to low abundance. Alternatively, the *O. vulgaris* samples from this study were from Spain, whereas the *O. vulgaris* from Kanda et al.⁶⁶ originated from Japan. It has been previously noted that the octopus population in Japan is genetically distinct from *O. vulgaris*⁶⁷ which could explain the absence of tachykinins in this study. Further studies have shown distinct morphological differences between the Japanese and Spanish populations of *O. vulgaris* and it has been suggested that the Japanese octopus could in fact be a different species (*Octopus sinensis*)^{68,69}. Whitelaw et al.²⁰ detected a tachykinin at the proteome level in the posterior salivary glands of *O. kaurana* but tachykinins were completely absent in *H. maculosa*^{4,20}. Given that the tachykinin-like peptide eledoisin from the octopus *Eledone cirrhosa* has well established neurotoxic activity¹⁸, it is interesting that the detection of tachykinins in octopod salivary glands is inconsistent. This may be due to genuine differences in venom composition between species or it may reflect difficulty in detecting fully processed tachykinin-like peptides via standard bottom up mass spectrometry approaches.

Short secreted cysteine rich proteins

Short secreted cysteine rich proteins (SSCRs) include a diverse range of potent toxins^{70,71} but they are particularly challenging to study for two reasons; firstly because their small size makes them difficult to identify based purely on genomic or transcriptomic sequences⁷² and secondly

1
2
3 because they frequently do not have homologs in model taxa and lack conserved domains that
4
5 would provide an indication of function. Perhaps as a consequence of this, these proteins form
6
7 a large component of what has been termed the dark proteome ⁷³, proteins with little structural
8
9 or functional information.
10
11
12
13
14

15 By combining proteomic and transcriptomic data in this study we were able to identify a total
16
17 of 21 short (<200 AA), secreted (SignalP), cysteine rich (CRI ≥5 or cysteine density > 5%)
18
19 proteins in the octopus salivary proteome (Supplementary Table S3). A significant fraction of
20
21 these (4/21) were not predicted from transcript sequences by TransDecoder and were only
22
23 identified as a consequence of the inclusion of six-frame translations in our protein database.
24
25 The majority (12/21) of *O. vulgaris* SSCRs were cephalopod specific (no detectable homologs
26
27 outside the Cephalopoda) but all had close homologs in other octopods. Expression of the
28
29 majority of these proteins was highly inconsistent with many being observed in a single sample
30
31 (Figure 2). Nevertheless at least four SSCRs were consistently observed in adult salivary
32
33 glands, of which two were homologous to proteins (NP1, NP2) previously identified from the
34
35 PSG of *H. maculosa* ⁴.
36
37
38
39
40
41

42 Although the majority of *O. vulgaris* SSCRs were unique to cephalopods we identified
43
44 homologs of sPLA₂, GM2 ganglioside activator (GM2A), a Neuropeptide prohormone and a
45
46 pacifastin inhibitor domain containing protein. Of these, sPLA₂ was the only protein strongly
47
48 expressed in PSG of both adult and juvenile octopus (Figure 2). sPLA₂ proteins are associated
49
50 with a wide variety of animal venoms and the same protein was also identified via our search
51
52 for venom protein homologs (see previous section).
53
54
55
56
57
58
59
60

Figure 2. Heatmap showing relative abundance of short secreted cysteine rich proteins. Cell colours show the normalised log₂ LFQ intensity value for each protein (rows) and each sample (columns). Grey cells indicate that the protein was not observed in that sample. Dendrograms show hierarchical clustering of rows and proteins are split into two distinct groups (k means clustering). Sample types are Anterior Salivary Gland (ASG), and Posterior Salivary Gland (PSG). For full details of included proteins see Supplementary Table S3.

Serine proteases

Serine proteases (peptidase s1 family) are a large and widespread family of proteins that occurs across many taxa. In some taxa (e.g. reptiles and insects⁵⁵) serine proteases have been recruited to a role as toxins. In cephalopods, evidence from bioassays⁷⁴ suggests that venom is capable of highly targeted external digestive functions (e.g. targeting muscle attachment points⁷⁵) and serine proteases are a likely candidate for this role¹⁹. Serine proteases were particularly diverse and abundant within the salivary gland and saliva proteome of *O. vulgaris* (Figure 1), which is consistent with previous proteomic work on the posterior salivary glands of other octopods¹⁹. Broad evolutionary relationships between serine proteases from cephalopods and

two outgroup molluscan taxa (Figure 3) revealed a large (175/325 sequences), well-supported clade that was entirely comprised of cephalopod sequences. The remaining sequences in the tree comprised several clades that included a dispersed mix of cephalopod, bivalve, and gastropod sequences. Previous phylogenetic work on cephalopod serine proteases has noted that this family seems to have undergone an expansion prior to the diversification of the cephalopod lineage^{19,20}. Here we provide further context to this hypothesis, with the arrangement of Figure 3 indicating that a cephalopod specific expansion of serine proteases occurred.

Another interesting aspect of cephalopod serine protease evolution becomes apparent when one considers the tissue of origin of the sequences in the tree. Without exception, all of the serine protease sequences identified from salivary gland and saliva in this study fell into the cephalopod specific clade. This clade also contained all of the serine protease sequences previously identified by Ruder et al.¹⁹ in transcriptomic surveys of posterior salivary glands of ten coleoid cephalopod species. In contrast, all of the sequences outside this clade came from species with completely sequenced genomes (*C. gigas*, *L. gigantea*, *O. bimaculoides*) or transcriptome assemblies from a wide range of body tissues (*H. maculosa*, *S. officinalis*). Taken together this data strongly suggests that coleoid cephalopods have not only undergone a dramatic expansion in serine proteases, but that this expansion is tightly associated with their saliva and salivary gland system.

Figure 3: Unrooted phylogenetic tree showing high level evolutionary relationships between serine proteases across coleoid cephalopods (octopus, cuttlefish, squid) and non-cephalopod molluscs (gastropods, bivalves). Node labels represent the percentage of 1000 ultrafast bootstraps generated by iqtree and are shown only for well supported high level nodes. Species included seven octopods: *Abdopus aculeatus*, *Octopus bimaculoides*, *Octopus cyanea*, *Octopus karna*, *Octopus vulgaris*, *Hapalochlaena maculosa*, *Pareledone turqueti*, three cuttlefishes: *Sepia latimanus*, *Sepia officinalis*, *Sepia pharaonis*, five squids: *Doryteuthis opalescens*, *Euprymna scolopes*, *Heterololigo bleekeri*, *Loliolus noctiluca*, *Sepioteuthis australis*, a bivalve, *Crassostrea gigas*, and a gastropod *Lottia gigantea*. Tip points are coloured according to broad taxonomic groupings with all cephalopod taxa in shades of red. Filled points represent sequences where tissue expression could not specifically be attributed to salivary glands or saliva. Open points represent sequences identified from transcriptomic or proteomic studies of

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

salivary glands. The marked clade is exclusive to cephalopods and contains all sequences known to be expressed in posterior salivary glands.

The evolutionary relationships and tissue specific expression of serine proteases within the cephalopod specific clade are shown in detail in Supplementary Figure S2. Although high level relationships within the tree were not well resolved, five well supported clades were identified that contained the majority of *O. vulgaris* sequences. All three species of octopus included in the tree were represented in all clades. Cuttlefish sequences were absent from clades A and E, but this may simply reflect the incompleteness of currently available cuttlefish data. Most sequences were expressed across a wide range of posterior salivary gland samples from adults and juveniles. Expression in the anterior salivary glands and saliva itself was more commonly observed in clades A, B and C rather than D and E.

Proteogenomics

Proteogenomic searches identified a total 353 “novel” proteins (i.e. identified purely through six-frame translation). Combining these “novel” proteins with those predicted by TransDecoder (“known”) resulted in a reference protein database with 18,889 total sequences. Almost all (317/353) “novel” sequences were shorter than the TransDecoder minimum length threshold of 100AA. The remaining 36 long (>100AA) “novel” proteins were all associated with mis-predictions by TransDecoder.

These results add to a growing body of work demonstrating the importance of proteogenomic methods when analysing venom samples by mass spectrometry ²⁰. Although the total number of additional proteins identified due to the inclusion of six-frame translations in the protein

1
2
3 database was relatively small (~1.9% of total), this included a large proportion of short,
4
5 secreted, cysteine rich proteins (19%). Given the importance of these classes of proteins in
6
7 venom research our results show that searching six-frame translations of available nucleic acid
8
9 data provide a significant boost to the sensitivity of mass spectrometry based venomomics studies.
10
11
12
13

14 ASSOCIATED CONTENT

15
16
17
18 **Supporting Information.** The following files are available free of charge. Figure of a
19
20 cysteine density analysis in different protein databases; figure of a phylogenetic tree examining
21
22 serine proteases in cephalopods; spreadsheet of the serine proteases used in the phylogenetic
23
24 trees; spreadsheet of the protein groups with homology to known venom family proteins;
25
26 spreadsheet of proteins which were short, secreted, and cysteine rich; Assembled transcript is
27
28 available on NCBI under accession PRJNA464423; Mass spectrometry proteomics data are
29
30 available via ProteomeXchange with identifier PXD010298: Username:
31
32 reviewer88731@ebi.ac.uk Password: dkeybunR
33
34
35
36

37 Figure S1: Cysteine density of three protein databases (file type, PDF)

38 Figure S2: Phylogenetic tree of serine proteases in cephalopods (file type, PDF)

39 Table S1: Protein groups with homology to known venom protein families (file type, XLSX)

40 Table S2: Short secreted cysteine rich protein groups (file type, XLSX)

41 Table S3: Serine protease sequences used in phylogenetic trees (file type, XLSX)
42
43
44
45
46
47
48

49 AUTHOR INFORMATION

50 51 52 Corresponding author

53
54 Email: legana.fingerhut@my.jcu.edu.au
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ACKNOWLEDGEMENTS

Drs. Manuel Nande and Jorge Hernández for collecting the octopus samples at the Instituto Español de Oceanografía that made this work possible. We would also like to acknowledge the *La Trobe University-Comprehensive Proteomics Platform* for providing key infrastructure and expertise for this study.

ABBREVIATIONS

ASG, anterior salivary gland; BLAST, basic local alignment search tool; CAP, cysteine rich secretory proteins, antigen 5, pathenogenesis related proteins; CRI, cysteine richness index; CRRF, cysteine rich region finder; DNase, deoxyribonuclease; iBAQ, intensity based absolute quantitation; MS/MS, tandem mass spectrometry; ORF, open reading frame; PSG, posterior salivary gland; SSCRs, short secreted cysteine rich proteins; sPLA_s, secreted phospholipases A₂;

REFERENCES

- (1) King, G. F., Venoms as a platform for human drugs: translating toxins into therapeutics. *Expert Opin. Biol. Ther.* **2011**, *11*, (11), 1469-1484.
- (2) Zhang, Y., Why do we study animal toxins? *Zool. Res.* **2015**, *36*, (4), 183-222.
- (3) Fry, B. G.; Koludarov, I.; Jackson, T. N. W.; Holford, M.; Terrat, Y.; Casewell, N. R.; Undheim, E. A. B.; Vetter, I.; Ali, S. A.; Low, D. H. W.; Sunagar, K., Seeing the woods for the trees: understanding venom evolution as a guide for biodiscovery. In *Venoms to Drugs: Venom as a Source for the Development of Human Therapeutics*, The Royal Society of Chemistry: 2015; pp 1-36.
- (4) Fry, B. G.; Roelants, K.; Norman, J. A., Tentacles of venom: toxic protein convergence in the kingdom animalia. *J Mol Evol* **2009**, *68*, (4), 311-321.
- (5) Cooke, I. R.; Whitelaw, B.; Norman, M.; Caruana, N.; Strugnell, J. M., Toxicity in cephalopods. In *Evolution of Venomous Animals and Their Toxins*, Gopalakrishnakone, P.; Malhotra, A., Eds. Springer Netherlands: Dordrecht, 2015; pp 1-15.
- (6) Anderson, R. C.; Mather, J. A., The packaging problem: Bivalve prey selection and prey entry techniques of the *Octopus enteroctopus dofleini*. *J Comp Psychol* **2007**, *121*, (3), 300.
- (7) Blustein, D. H.; Anderson, R. C., Localization of octopus drill holes on cowries. *Am Malacol Bull* **2016**, *34*, (1), 61-64.
- (8) Grisley, M. S.; Boyle, P. R.; Key, L. N., Eye puncture as a route of entry for saliva during predation on crabs by the octopus *Eledone cirrhosa* (Lamarck). *J Exp Mar Biol Ecol* **1996**, *202*, (2), 225-237.

- (9) Cariello, L.; Zanetti, L., α - and β -cephalotoxin: two paralysing proteins from posterior salivary glands of *Octopus vulgaris*. *Comp. Biochem. Physiol., C: Comp. Pharmacol.* **1977**, *57*, (2), 169-173.
- (10) Ghiretti, F., Cephalotoxin: the crab-paralysing agent of the posterior salivary glands of cephalopods. *Nature* **1959**, *183*, (4669), 1192-1193.
- (11) Ghiretti, F., Toxicity of octopus saliva against crustacea. *Ann N Y Acad Sci* **1960**, *90*, (3), 726-741.
- (12) Grisley, M. S.; Boyle, P. R., Bioassay and proteolytic activity of digestive enzymes from octopus saliva. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **1987**, *88*, (4), 1117-1123.
- (13) Pilson, M. E. Q.; Taylor, P. B., Hole drilling by octopus. *Science* **1961**, *134*, (3487), 1366-1368.
- (14) Nixon, M., Is there external digestion by octopus? *J. Zool.* **1984**, *202*, (3), 441-447.
- (15) Undheim, E. A. B.; Georgieva, D. N.; Thoen, H. H.; Norman, J. A.; Mork, J.; Betzel, C.; Fry, B. G., Venom on ice: First insights into Antarctic octopus venoms. *Toxicon* **2010**, *56*, (6), 897-913.
- (16) Key, L. N.; Boyle, P. R.; Jaspars, M., Novel activities of saliva from the octopus *Eledone cirrhosa* (Mollusca; Cephalopoda). *Toxicon* **2002**, *40*, (6), 677-683.
- (17) Songdahl, J. H.; Shapiro, B. I., Purification and composition of a toxin from the posterior salivary gland of *Octopus dofleini*. *Toxicon* **1974**, *12*, (2), 109-112.
- (18) Anastasi, A.; Erspamer, V., The isolation and amino acid sequence of eledoisin, the active endecapeptide of the posterior salivary glands of *Eledone*. *Arch Biochem Biophys* **1963**, *101*, (1), 56-65.

- (19) Ruder, T.; Sunagar, K.; Undheim, E. A. B.; Ali, S. A.; Wai, T.-C.; Low, D. H. W.; Jackson, T. N. W.; King, G. F.; Antunes, A.; Fry, B. G., Molecular phylogeny and evolution of the proteins encoded by coleoid (cuttlefish, octopus, and squid) posterior venom glands. *J Mol Evol* **2013**, 76, (4), 192-204.
- (20) Whitelaw, B. L.; Strugnell, J. M.; Faou, P.; da Fonseca, R. R., Combined transcriptomic and proteomic analysis of the posterior salivary gland from the southern blue-ringed octopus and the southern sand octopus. *J. Proteome Res.* **2016**, 15, (9), 3284-3297.
- (21) Whittington, C. M.; Papenfuss, A. T.; Locke, D. P.; Mardis, E. R.; Wilson, R. K.; Abubucker, S.; Mitreva, M.; Wong, E. S. W.; Hsu, A. L.; Kuchel, P. W.; Belov, K.; Warren, W. C., Novel venom gene discovery in the platypus. *Genome Biol* **2010**, 11, (9), R95-R95.
- (22) von Reumont, B. M.; Blanke, A.; Richter, S.; Alvarez, F.; Bleidorn, C.; Jenner, R. A., The first venomous crustacean revealed by transcriptomics and functional morphology: remipede venom glands express a unique toxin cocktail dominated by enzymes and a neurotoxin. *Mol Biol Evol* **2014**, 31, (1), 48-58.
- (23) Hernández-García, V.; Martín, A. Y.; Castro, J. J., Evidence of external digestion of crustaceans in *Octopus vulgaris* paralarvae. *J Mar Biol Assoc U K* **2000**, 80, (3), 559-560.
- (24) Roura, Á.; González, Á. F.; Redd, K.; Guerra, Á., Molecular prey identification in wild *Octopus vulgaris* paralarvae. *Mar. Biol.* **2012**, 159, (6), 1335-1345.
- (25) Daltry, J. C.; Wuster, W.; Thorpe, R. S., Diet and snake venom evolution. *Nature* **1996**, 379, (6565), 537-40.
- (26) Underwood, A. H.; Seymour, J. E., Venom ontogeny, diet and morphology in *Carukia barnesi*, a species of Australian box jellyfish that causes Irukandji syndrome. *Toxicon* **2007**, 49, (8), 1073-1082.

(27) European Parliament and Council of the European Union, Directive 2010/63/EU on the protection of animals used for scientific purposes. *Off. J. Eur. Communities: Legis.* **2010**, *53*, (L 276), 33-89.

(28) Ballering, R. B.; Jalving, M. A.; VenTresca, D. A.; Hallacher, L. E.; Tomlinson, J. T.; Wobber, D. R., Octopus evenomation through a plastic bag via a salivary proboscis. *Toxicon* **1972**, *10*, (3), 245-248.

(29) Grabherr, M. G.; Haas, B. J.; Yassour, M.; Levin, J. Z.; Thompson, D. A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; Chen, Z.; Mauceli, E.; Hacohen, N.; Gnirke, A.; Rhind, N.; di Palma, F.; Birren, B. W.; Nusbaum, C.; Lindblad-Toh, K.; Friedman, N.; Regev, A., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **2011**, *29*, (7), 644-652.

(30) Altschul, S. F.; Gish, W.; Miller, W.; Meyers, E. W.; Lipman, D. J., Basic Local Alignment Search Tool. *J Mol Biol* **1990**, *215*, (3), 403-410.

(31) Finn, R. D.; Clements, J.; Eddy, S. R., HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **2011**, *39*, (suppl_2), W29-W37.

(32) Finn, R. D.; Bateman, A.; Clements, J.; Coghill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E. L. L.; Tate, J.; Punta, M.; Institutionen för biokemi och, b.; Naturvetenskapliga, f.; Stockholms, u.; Science for Life, L., Pfam: The protein families database. *Nucleic Acids Res* **2014**, *42*, (1), D222-D230.

(33) Petersen, T. N.; Brunak, S.; Von Heijne, G.; Nielsen, H.; Institutionen för biokemi och, b.; Stockholms, u.; Naturvetenskapliga, f., SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, (10), 785-786.

(34) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. L., Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol* **2001**, *305*, (3), 567-580.

(35) Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2*, (8), 1896-1906.

(36) Vizcaíno, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H., 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **2016**, *44*, (1), D447-D456.

(37) Caruana, N. J.; Cooke, I. R.; Faou, P.; Finn, J.; Hall, N. E.; Norman, M.; Pineda, S. S.; Strugnell, J. M., A combined proteomic and transcriptomic analysis of slime secreted by the southern bottletail squid, *Sepiadarium austrinum* (Cephalopoda). *J. Proteomics* **2016**, *148*, 170-82.

(38) Craig, R.; Beavis, R. C., TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, (9), 1466-1467.

(39) Kim, S.; Pevzner, P. A., MS-GF plus makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*.

(40) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, (20), 5383-5392.

(41) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I., iProphet: Multi-level integrative analysis

of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10*, (12), M111-M111.007690.

(42) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, (17), 4646-4658.

(43) Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, (12), 1367-1372.

(44) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, (11), 1422-1423.

(45) Lavergne, V.; Harliwong, I.; Jones, A.; Miller, D.; Taft, R. J.; Alewood, P. F., Optimized deep-targeted proteotranscriptomic profiling reveals unexplored *Conus* toxin diversity and novel cysteine frameworks. *Proc. Natl. Acad. Sci.* **2015**, *112*, (29), E3782-E3791.

(46) Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; Pesseat, S.; Quinn, A. F.; Sangrador-Vegas, A.; Scheremetjew, M.; Yong, S.-Y.; Lopez, R.; Hunter, S., InterProScan 5: genome-scale protein function classification. *Bioinformatics* **2014**, *30*, (9), 1236-1240.

(47) Katoh, K.; Standley, D. M., MAFFT Multiple Sequence Alignment Software version 7: Improvements in performance and usability. *Mol Biol Evol* **2013**, *30*, (4), 772-780.

(48) Katoh, K.; Misawa, K.; Kuma, K. i.; Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **2002**, *30*, (14), 3059-3066.

(49) Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; Thierer, T.; Ashton, B.; Meintjes, P.; Drummond, A., Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)* **2012**, *28*, (12), 1647-1649.

(50) Nguyen, L.-T.; Schmidt, H. A.; von Haeseler, A.; Minh, B. Q., IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **2015**, *32*, (1), 268-274.

(51) Kalyaanamoorthy, S.; Minh, B. Q.; Wong, T. K. F.; von Haeseler, A.; Jermin, L. S., ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, (6), 587-589.

(52) Minh, B. Q.; Nguyen, M. A. T.; von Haeseler, A., Ultrafast Approximation for Phylogenetic Bootstrap. *Mol Biol Evol* **2013**, *30*, (5), 1188-1195.

(53) Paradis, E.; Claude, J.; Strimmer, K., APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **2004**, *20*, (2), 289-290.

(54) Yu, G.; Smith, D. K.; Zhu, H.; Guan, Y.; Lam, T. T.-Y., ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **2017**, *8*, (1), 28-36.

(55) Fry, B. G.; Roelants, K.; Champagne, D. E.; Scheib, H.; Tyndall, J. D. A.; King, G. F.; Nevalainen, T. J.; Norman, J. A.; Lewis, R. J.; Norton, R. S.; Renjifo, C.; de la Vega, R. C. R., The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annu. Rev. Genomics Hum. Genet.* **2009**, *10*, 483-511.

(56) Mochca-Morales, J.; Martin, B. M.; Possani, L. D., Isolation and characterization of Helothermine, a novel toxin from *Heloderma horridum horridum* (Mexican beaded lizard) venom. *Toxicon* **1990**, 28, (3), 299-309.

(57) Morrisette, J.; Krätzschar, J.; Haendler, B.; el-Hayek, R.; Mochca-Morales, J.; Martin, B. M.; Patel, J. R.; Moss, R. L.; Schleuning, W. D.; Coronado, R., Primary structure and properties of helothermine, a peptide toxin that blocks ryanodine receptors. *Biophys J* **1995**, 68, (6), 2280-2288.

(58) Nobile, M.; Noceti, F.; Prestipino, G.; Possani, L. D., Helothermine, a lizard venom toxin, inhibits calcium current in cerebellar granules. *Exp Brain Res* **1996**, 110, (1), 15-20.

(59) Nobile, M.; Magnelli, V.; Lagostena, L.; Mochca-Morales, J.; Possani, L. D.; Prestipino, G., The toxin helothermine affects potassium currents in newborn rat cerebellar granule cells. *J Membr Biol* **1994**, 139, (1), 49-55.

(60) Milne, T. J.; Abbenante, G.; Joel, D. A. T.; Halliday, J.; Lewis, R. J., Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily. *J Biol Chem* **2003**, 278, (33), 31105-31110.

(61) Qian, J.; Guo, Z.-y.; Chi, C.-w., Cloning and isolation of a *Conus* cysteine-rich protein homologous to Tex31 but without proteolytic activity. *Acta Biochim Biophys Sin* **2008**, 40, (2), 174-181.

(62) Grisley, M. S.; Boyle, P. R., Chitinase, a new enzyme in octopus saliva. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **1990**, 95, (2), 311-316.

(63) Schaloske, R. H.; Dennis, E. A., The phospholipase A2 superfamily and its group numbering system. *Biochim. Biophys. Acta* **2006**, 1761, (11), 1246-1259.

(64) Valentin, E., What can venom phospholipases A2 tell us about the functional diversity of mammalian secreted phospholipases A2? *Biochimie* **2000**, 82, (9-10), 815-831.

(65) Razpotnik, A.; Križaj, I.; Šribar, J.; Kordiš, D.; Maček, P.; Frangež, R.; Kem, W. R.; Turk, T., A new phospholipase A2 isolated from the sea anemone *Urticina crassicornis* - its primary structure and phylogenetic classification: New PLA2 from *Urticina crassicornis*. *FEBS J.* **2010**, 277, (12), 2641-2653.

(66) Kanda, A.; Iwakoshi-Ukena, E.; Takuwa-Kuroda, K.; Minakata, H., Isolation and characterization of novel tachykinins from the posterior salivary gland of the common octopus *Octopus vulgaris*. *Peptides* **2003**, 24, (1), 35-43.

(67) Roura, Á.; Antón Álvarez-Salgado, X.; González, Á. F.; Gregori, M.; Rosón, G.; Otero, J.; Guerra, Á., Life strategies of cephalopod paralarvae in a coastal upwelling system (NW Iberian Peninsula): insights from zooplankton community and spatio-temporal analyses. *Fish Oceanogr* **2016**, 25, (3), 241-258.

(68) Amor, M. D.; Norman, M. D.; Roura, A.; Leite, T. S.; Gleadall, I. G.; Reid, A.; Perales-Raya, C.; Lu, C. C.; Silvey, C. J.; Vidal, E. A. G.; Hochberg, F. G.; Zheng, X.; Strugnell, J. M., Morphological assessment of the *Octopus vulgaris* species complex evaluated in light of molecular-based phylogenetic inferences. *Zool. Scr.* **2017**, 46, (3), 275-288.

(69) Gleadall, I. G., *Octopus sinensis* d'Orbigny, 1841 (Cephalopoda: Octopodidae): Valid Species Name for the Commercially Valuable East Asian Common Octopus. *Species Divers* **2016**, 21, (1), 31-42.

(70) Robinson, S. D.; Li, Q.; Bandyopadhyay, P. K.; Gajewiak, J.; Yandell, M.; Papenfuss, A. T.; Purcell, A. W.; Norton, R. S.; Safavi-Hemami, H., Hormone-like peptides in the venoms of marine cone snails. *Gen Comp Endocrinol* **2015**, 244, 11-18.

(71) Linial, M.; Rappoport, N.; Ofer, D., Overlooked short toxin-like proteins: a shortcut to drug design. *Toxins (Basel)* **2017**, *9*, (11), 350.

(72) Dinger, M. E.; Pang, K. C.; Mercer, T. R.; Mattick, J. S., Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comp. Biol.* **2008**, *4*, (11), e1000176.

(73) Perdigão, N.; Heinrich, J.; Stolte, C.; Sabir, K. S.; Buckley, M. J.; Tabor, B.; Signal, B.; Gloss, B. S.; Hammang, C. J.; Rost, B.; Schafferhans, A.; O'Donoghue, S. I., Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* **2015**, *112*, (52), 15898-15903.

(74) Grisley, M. S., Separation and partial characterization of salivary enzymes expressed during prey handling in the octopus *Eledone cirrhosa*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **1993**, *105*, (1), 183-192.

(75) Mather, J. A.; Nixon, M., *Octopus vulgaris* (Cephalopoda) drills the chelae of crabs in Bermuda. *J Molluscan Stud* **1995**, *61*, (3), 405-406.

Graphical abstract

82x44mm (300 x 300 DPI)