NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF LIFE SCIENCES

Coleoid venoms: predicting cephalotoxin function and biotechnological applications from ecological and evolutionary traits

CÁTIA VANESSA CAETANO GONÇALVES

Master in Environmental Engineering

DOCTORATE IN BIOTECHNOLOGY

NOVA University Lisbon

April, 2025

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF LIFE SCIENCES

Coleoid venoms: predicting cephalotoxin function and biotechnological applications from ecological and evolutionary traits

CÁTIA VANESSA CAETANO GONÇALVES

Master in Environmental Engineering

Adviser: Pedro Manuel Brôa Costa
Assistant Professor, NOVA University of Lisbon

Co-advisers: Ana Rita Fialho Grosso
Assistant Professor, NOVA University of Lisbon

Odd André Karlsen
Associate Professor, University of Bergen

Examination Committee:

Chair: Pedro Miguel Ribeiro Viana Baptista
Full Professor, NOVA University of Lisbon

Rapporteurs: Bernardo Afonso De Aranha Alhandra Duarte
Assistant Professor, Faculdade de Ciências da Universidade de Lisboa

Marco Filipe Loureiro Lemos
Associate Professor with Habilitation, Instituto Politécnico de Leiria

Adviser: Pedro Manuel Brôa Costa
Assistant Professor, NOVA University of Lisbon

Members: Francisco Rente de Pina Martins
Assistant Professor, Instituto Politécnico de Setúbal

Pedro Miguel Ribeiro Viana Baptista
Full Professor, NOVA University of Lisbon

Susana Maria Pereira Gaudêncio de Matos
Assistant Researcher, NOVA University of Lisbon

DOCTORATE IN BIOTECHNOLOGY

NOVA University Lisbon
April, 2025

Coleoid venoms: predicting cephalotoxin function and biotechnological applications from ecological and evolutionary traits

To my favourite humans, little Clara and Ricardo.

ACKNOWLEDGMENTS

As science is not made by just one person, a student's PhD thesis, despite representing the hard and often individual work of several years of research, cannot exist without a true team behind it, and I had the best.

For all the support and advice I received during these years, I want to especially thank Pedro Costa and Ana Rita for the scientific debates, true guidance, lunches that helped me to decompress, but above all for being human and friends.

To the entire SeaTox lab team, especially those with whom I shared the bench from the beginning, Ana Rodrigo and Carla Martins, for always being there, for the advice, the friendly shoulder and all the funny moments. To Inês Cabral, for sharing knowledge, kindness, infinite availability, and all the 'you can do it!'. To Mariaelena, Diana Rodrigues, Carolina Madeira, and Inês Padrão, for the scientific motivation, experiences, and good moments. To all our students who passed through the SeaTox lab team, with whom I could also learn and develop interpersonal skills.

To all the members of the Computational Multi-Omics Lab for the scientific inputs, 'palette' discussions, lunches, and overall good moments. To Mafalda Soares, in particular, for her skills in digital drawing, which helped me to improve my own figures.

A big thank you to Fred, who was responsible for catching all the octopuses and cuttlefishes, without which I wouldn't have been able to conduct all the experiments of my PhD.

To professor António Pedro Alves de Matos and Pedro Henriques from the Egas Moniz Centre for Interdisciplinary Research, for their help and assistance with transmission electron microscopy, one of the most amazing microscopy techniques.

A thank you to my co-supervisor Odd André from the University of Bergen and to all the Environmental Toxicology group, especially Rhîan, for all the supervision, warm welcoming, and help in one of the most challenging parts of my PhD.

To my husband, for his infinite patience, for always believing in me even when I didn't, and for his hug that cures any 'storm'.

To my daughter, for teaching me to breathe, relax, and appreciate the little things in life.

To my parents, for helping me become the person I am today and for always being there for me, at any moment.

This work would not have been possible without financial support. For this reason, I would like to acknowledge the Portuguese Foundation for Science and Technology (FCT), I.P. for my PhD grant SFRH/BD/144914/2019 (https://doi.org/10.54499/SFRH/BD/144914/2019), and also Fundo Azul for co-financing the MARVEN project (FA_05_2017_007), within which this thesis was developed.

This work was also financed by national funds from FCT, I.P., in the scope of the project UIDP/04378/2020 and UIDB/04378/2020 of the Research Unit on Applied Molecular Biosciences - UCIBIO and the project LA/P/0140/2020 of the Associate Laboratory Institute for Health and Bioeconomy - i4HB.

"Always do your best in whatever you do."

ABSTRACT

The ocean's immense biodiversity offers almost limitless but little explored bioactives for safer and more cost-effective alternatives to synthetic drugs. However, addressing marine biodiversity, especially invertebrates, while producing validated data for industry is challenging. Albeit often neglected, commercial venomous species such as cuttlefishes and octopuses can provide varied bioactives for the purpose of bioprospecting. Cephalotoxins, secreted by the salivary glands of coleoids, are little-known substances with at least analgesics properties. Benefitting from their economic and ecological relevance, our case studies are the *Octopus vulgaris* and the *Sepia officinalis*. The integration of the toxin-delivery apparatus (i.e., anterior and posterior salivary glands) ecophysiology, combined with toxicology and molecular biology provided comprehensive information on the function and targets of these cephalopods bioactives. Altogether, the microanatomy revealed that the two species secrete different substances. Despite generally similar anatomy between the two salivary glands, characterised by branched glandular tubules bearing mucous and secretory epithelia, the octopus' posterior salivary gland (PSG) is more complex structurally than its anterior counterpart (ASG), which also extends to the simpler cuttlefish PSG. Transcriptomics confirmed the high diversity of venom-peptides secreted by both cephalopods, of which bioactives with potential neurotoxic and endocrine-disrupting effects (e.g., pro-insulins) are highlighted, together with proteins with inhibitory functions. However, the cuttlefish PSG presented more putative cephalotoxin variants than octopus. Seven full coding sequences of interest, such as three cysteine-rich venom proteins, a venom insulin, a cephalotoxin, a chitinase and a hyaluronidase, were isolated and validated. These bioactives exhibited different patterns of sequence variation within venomous taxa, suggesting potential specificity among octopus and cuttlefishes. The crude protein extracts of both PSG were able to activate human G protein-coupled receptors of interest, indicating its great potential to aid in

bioprospection, albeit yielding low combined cytotoxicity, which shows an interesting path towards safer drugs targeting the human druggable proteome.

Keywords: Marine biotechnology; Marine toxins; Cephalopoda; Salivary glands; Transcriptomics; Microscopy

RESUMO

A imensa biodiversidade do oceano é uma fonte quase ilimitada de bioativos, embora pouco explorada, que se podem tornar alternativas seguras e rentáveis para a indústria biofarmacêutica. No entanto, abordar esta biodiversidade, especialmente de invertebrados, enquanto se produz dados válidos para a indústria é desafiante. Embora frequentemente negligenciadas, espécies comerciais venenosas como chocos e polvos podem ser importantes alvos de bioprospecção para novos bioactivos. As cefalotoxinas secretadas pelas glândulas salivares dos coleóides apresentam potencial analgésico, mas são substâncias pouco conhecidas. Beneficiando da sua relevância económica e ecológica, os nossos casos de estudo são o *Octopus vulgaris* e o *Sepia officinalis*. A integração da ecofisiologia do órgão produtor de toxinas (i.e., glândulas salivares anteriores e posteriores), combinada com toxicologia e biologia molecular, forneceu informações importantes sobre a função e os alvos dos bioativos destes cefalópodes. Apesar da aparência anatómica semelhante das suas glândulas salivares, caracterizadas por túbulos glandulares ramificados com epitélios mucosos e secretórios, a glândula salivar posterior (PSG) do polvo parece mais complexa do que anterior (ASG), o que se estende à PSG do choco, indicando secreções diferenciadas. A transcritómica confirmou grande diversidade de péptidos venenosos secretados pelos dois cefalópodes, dos quais se destacam bioativos com potenciais efeitos neurotóxicos e disruptores endócrinos (por exemplo, pró-insulinas), juntamente com proteínas com funções inibitórias. No entanto, a PSG do choco apresentou mais variantes de cefalotoxinas. Foram isoladas e validadas sete sequências codificantes de interesse, nomeadamente três proteínas ricas em cisteína, uma pró-insulina, uma cefalotoxina, uma quitinase e uma hialuronidase. Estes bioativos exibiram diferentes padrões de variação entre diversos taxa venenosos, sugerindo potencial especificidade entre polvos e chocos. Os extratos proteicos brutos de ambas as PSG foram capazes de ativar, com baixa toxicidade, recetores humanos acoplados à proteína G, indicando

grande potencial de bioprospecção e revelando um caminho interessante na busca de novos biofármacos.

Palavras-chave: Biotecnologia marinha; Toxinas marinhas; Cephalopoda; Glândulas salivares; Transcritómica; Microscopia.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACE	Angiotensin-converting enzyme
AGTR2	Angiotensin II receptor
ANOVA	Analysis of Variance
ASG	Anterior salivary gland
AVPR2	Arginine vasopressin receptor
BLASTP	Basic local Alignment Search Tool for Proteins
cAMP	Cyclic adenosine 3′,5′-monophosphate
CAP	CRISP [Cysteine-rich secretory proteins], Antigen 5 [Ag5], and Pathogenesis-related [PR-1])
cDNA	Complementary deoxyribonucleic acid
CFDA-AM	5-carboxyfluorescein diacetate acetoxymethyl ester
CHI	Chitinase
CNS	Central nerve system
CRISP	Cysteine-rich secretory protein
CRVP	Cysteine-rich venom protein

Cryo-EM	Cryo-electron microscopy
CTX	SE-cephalotoxin
DEGs	Differentially expressed genes
DMEM	Dulbecco's modified Eagle's medium
DNA	Deoxyribonucleic acid
EGF	Epidermal growth factor
ESTs	Expressed sequence tags
EU	European Union
FDR	False discovery rate
GLP1R	Glucagon-like peptide receptor
GO	Gene Ontology
GPCRs	G protein-coupled receptors
HYAL	Hyaluronidase
HPLC	High-performance liquid chromatography
HTLA	HEK293 cell line stably expressing a tetracycline transactivator (tTA)-dependent luciferase reporter and β-arrestin2 fused to a TEV protease
LPAR1	Lysophosphatidic acid receptor
MAP	Mitogen-activated protein
MC1R	Melanocortin receptor
mRNA	Messenger ribonucleic acid
NCBI	National Center for Biotechnology Information

NMR	Nuclear magnetic resonance
NMUR2	Neuromedin U receptor
NTSR2	Neurotensin receptor
ORFs	Open reading frames
OsO_4	Osmium tetroxide
PAS	Periodic Acid – Schiff's
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PSG	Posterior salivary gland
PTMs	*Post*-translational modifications
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
RT-qPCR	Quantitative reverse transcription polymerase chain reaction
TEM	Transmission electron microscopy
TPMs	Transcript per millions
TTX	Tetrodotoxin
USA	United States of America
VIns	Venom insulin
VIPR1	Vasoactive intestinal peptide

1

INTRODUCTION

The present section was adapted from the following articles produced and published during this thesis:

MOUTINHO CABRAL I**, GONÇALVES C**, GROSSO AR & COSTA PM (2024) Bioprospecting and marine 'omics': surfing the deep blue sea for novel bioactive proteins and peptides. Frontiers in Marine Science 11:1362697. (DOI: 10.3389/fmars.2024.1362697)

Conceptualisation and writing the original manuscript were shared by **C. Gonçalves** and I. Moutinho Cabral.

GONÇALVES C & COSTA PM (2021) Cephalotoxins: A hotspot for marine bioprospecting? Frontiers in Marine Science 8, 1 – 7. (DOI: 10.3389/fmars.2021.647344)

Formal analysis, investigation and writing the original manuscript were performed by **C. Gonçalves**.

The global ocean is arguably the widest and most diversified source of bioactives. Motivated by its vast coastline and varied biogeography from the Mediterranean to the North Sea, the European Union (EU) set high prospects for turning marine bioprospecting into one of the means to lead the global 'Bioeconomy Revolution' in the near future, as new marine-derived bioproducts join ranks in novel potential human and veterinary pharmaceuticals and biomaterials; feeds and nutraceuticals; eco-friendly pesticides; anti-foulants, and other applications (see for instance Rodrigo and Costa, 2019; Vieira et al., 2020; Gonçalves and Costa, 2021). However, virtually all industrialised nations partake in this race, with potentially high benefits for scientists and their networks, industrialists and the end-point consumer. In addition, it is a major contribution to exploit novel and more sustainable marine resources. Evidently, this last, but not least, gain implies the industrial production of marine bioactives rather than its harvesting from the seas.

Bioactive compounds can be simply defined as substances that can interfere with a biological system, triggering a physiological consequence that can be either a response or a direct effect. Even though the concept became popularised by nutraceutical research (refer to Kussmann et al., 2023, for a recent review), 'natural bioactive' is far from being exhausted by food additive research and applications. In fact, compounds that interfere with specific molecular and downstream physiological processes hold great value as pharmaceuticals, for instance, depending on their specificity, potency and safety (Lindequist, 2016; Bordon et al., 2020). Many marine organisms secrete natural antibiotics, anti-foulants and defensive repellents, without neglecting venomous and poisonous animals that inject (in the first case) cocktails of neurotoxins with potential value as painkillers into their target recipients, just to mention few examples (Shen et al., 2000; Rodrigo and Costa, 2019; Turner et al., 2020; Guryanova et al., 2023). However, despite the promises of the seas as an almost inexhaustible source of novel bioactives, the real number of marine-derived pharmaceuticals, for instance, may seem staggering low (Martins et al., 2014; Cappello and Nieri, 2021). In fact, forty years ago, Colwell (1983) in a notorious paper in *Science*, had already pointed out the advantages and challenges of marine biotechnology and its close association to modern genetic engineering to scale-up production. Still, decades after, only about a dozen drugs derived from marine bioactives were tested and approved for commercial use in the United States of America (USA) and the EU since the late 1960s. This figure includes painkillers like ziconotide (developed from a conotoxin), antivirals like vidarabine (derived from a sponge) and anti-cancer drugs like

trabectedin (from a tropical tunicate), which is an alkaloid currently under application mostly for treating sarcomas (recently reviewed by Cappello and Nieri, 2021). Nonetheless, if we consider the challenges of bioprospecting through the oceans' immense biodiversity and how to translate this endeavour into consumer products we may, at least partly, understand these seemingly meagre results compared to bioproducts of terrestrial origin, which directly or indirectly account for roughly 60% of pharmaceutical bioactives (Montaser and Luesch, 2011; Lindequist, 2016). We can also add the many gaps in the knowledge on basic aspects of the biology of marine organisms, from systematics to ecology and physiology, plus difficulties in accessing open water and deep-sea habitats for prospecting (Molinski et al., 2009; Lindequist, 2016; Rodrigo and Costa, 2019). In addition, producing a natural compound in the laboratory can be cumbersome and may require detailed knowledge of unknown metabolic pathways.

From the current state-of-the-art, we may isolate two main approaches for marine bioprospecting: i) extensively cataloguing new metabolites and ii) directing searches towards specific needs. Whereas the second could systematise bioprospecting in favour of the needs of industry but seems still a mirage in the near future, the first is currently the mainstream approach that, albeit relying on a good deal of fortuitousness, is generating important databases especially (but not exclusively) for metabolites isolated from marine invertebrates and microorganisms (e.g., Hosseini et al., 2022, for a review) yet to be translated into consumer products. Regardless of approach, type of molecule and type of prospective application, it is clear that systematising marine bioprospecting is paramount. With this respect, advances in molecular and computational tools, including public databases, permit retrieving, analysing and storing huge amounts of data from marine biological samples and revolutionising bioprospecting (Ambrosino et al., 2019; Maghembe et al., 2020). These advances include high throughput methods, designed to analyse one or few endpoints (such as proteins or mRNAs) in multiple biological samples in a single run, and 'omics' approaches, conceived to analyse multiple endpoints in single runs. While goals like the identification of multiple species (e.g., microbes) in single samples can bear enormous advantages, it is 'omics' that enable screening for multiple metabolites, proteins and mRNAs, with which databases of bioactives and marine genomic resources are enriched, with the added value of contributing to understand how compounds of interest are produced *in vivo* to as a means to achieve industrial scale-up via genetic engineering using convenient models such as bacteria and yeasts (Pennington et al., 2018; Maghembe et al., 2020).

1.1 Marine toxins in the sights of biotechnology

Marine toxins are likely as diversified as the oceans' immense biodiversity and result from co-evolutionary processes between the organisms that secrete them and their target recipients, whether prey, predators or parasites. Along with other types of natural products, these substances are considered to hold high-value bioproducts due to their potential specificity against molecular receptors and metabolic pathways. Many toxins from animals are peptides or high molecular weight proteins that interact with specific enzymes and ion channels (e.g., sodium and/or potassium ion channels), thus affecting, for example, neuromuscular, cardiovascular and immune systems (Zhang, 2015; Ponte and Modica, 2017). For these reasons, toxinology has been rising as a promising field with direct implications for drug discovery, novel therapeutic alternatives and even ecologically friendly antifoulants and pesticides. Although still lagging behind research on their terrestrial counterparts, such as snakes, bees and scorpions, marine toxinology is on the rise as researchers and bio-based industrialists become increasingly aware of the uncanny value of the oceans as an almost limitless source of novel bioproducts (see for instance Molinski et al., 2009; Leal et al., 2020).

Venoms and poisons are complex mixtures of toxins, enzymes, peptides and salts, which renders molecule isolation, identification and determination of substance-specific bioreactivity challenging. Their efficacy results from the interaction of these compounds, not just from toxins as noxious agents but also from permeabilisers like proteolytic enzymes. Still, several venomous or poisonous species of snakes, lizards, and leeches have already been used to produce toxin-based drugs for clinical applications directed against disease-induced pain (e.g., from osteoarthritis and multiple sclerosis), as anticoagulants for specific coronary interventions and also for hypertension or specific types of diabetes (see Bordon et al., 2020 for a review). Additionally, studies on wasp venoms are revealing promising alternatives for traditional antibiotics and even as antiviral and anti-seizure therapeutics (Vila-Farrés et al., 2012; Sample et al., 2013; Castro e Silva et al., 2020). In turn, pushed by the recent premises of the "Blue Growth" revolution, the last decade witnessed a growing interest in marine bioproducts, toxins included. Indeed, marine ecosystems are nowadays acknowledged to be a highly valuable source of pharmaceutically active toxins (see Greener, 2020). Notwithstanding, marine invertebrates are considered the most promising group of animals possessing toxins in face of their richness. Marine animals such cnidarians (jellyfishes, sea

anemones, hydrozoans), echinoderms (starfishes, sea urchins), annelids (polychaetes), nemertines, bryozoans (moss animals), sponges, tunicates and molluscs (gastropods, cephalopods) stand out as toxin-secreting eumetazoans (see Kem, 2005; Zhang, 2015; Rodrigo and Costa, 2019). As examples, recent investigations highlighted bryostatins, which are heterocyclic molecules produced by bryozoans, and other cyclic peptides produced by endosymbiotic micro-organisms from some tunicates, as high-potential compounds for cancer therapeutics (Kem, 2005; Watters, 2018). As yet another example, clinical trials on toxic peptides from sun anemone have been revealed favourable results for the treatment of autoimmune diseases, due to their effects as strong potassium channel blockers (Prentis et al., 2018). However, the most prominent example among biotechnological applications of marine bioproducts is, again, the approved painkiller ziconotide (commercialised as Prialt®), developed from a ω-conotoxin found in the venom of the cone snail *Conus magus* (Williams et al., 2008). This case has recently been followed by trabectedin (Yondelis®), an anti-cancer quinoline first isolated from a tropical tunicate (Cuevas and Francesch, 2009), which, albeit not primarily a toxin, yields leverage to drug development from marine bioproducts.

As more species of marine invertebrates join the ranks of toxin-secreting marine animals, attention is turning to one of the oldest-known venomous group of molluscs, the cephalopods. Perhaps surprisingly, only a few toxins have yet been described for these predators but the broad range of bioactive compounds in their venomous saliva, which includes neurotoxic proteins and peptides, amines and permeabilising enzymes, yields high promises for marine bioprospecting (see Cooke et al., 2017). Recent advances in marine "venomics" (see von Reumont et al., 2014a) can shed new lights on these obscure toxins. Cephalotoxins, in particular, appear to be a unique class of toxins whose exact nature remains elusive even though their discovery dates from the mid-XX century. Here, is summarise the scant knowledge on these toxins that are produced by the salivary glands of cephalopods, endeavouring a fresh biotechnological perspective for the bioactives secreted by these important and ubiquitous animals.

1.2 Cephalotoxins and other coleoid toxins

The interest on cephalotoxins, albeit when they were not termed as such, rose in the late XIX century when Lo Bianco (1888) studied the toxicity of octopus saliva to crabs. These

observations showed that after injecting secretions from octopus' posterior salivary glands directly into crab gills, the target experienced hindered locomotion, progressing rapidly towards complete immobilisation and eventually death (see also Ghiretti, 1960). The active substance that causes this paralysing action was latter purified by Ghiretti (1959) from the posterior salivary glands of *Sepia officinalis*, identified as a proteinaceous compound and called cephalotoxin. From that time onwards, the composition of salivary secretions of cephalopods gained attention and several constituents have since been isolated and identified, with emphasis on neurotoxins and neuropeptides (Table 1). It is now clear that the complexity of these secretions is reflected in a broad range of biologically active compounds, proteinaceous or not. This includes biological amines, namely tyramine, histamine, *p*-hydroxyphenyl-ethanol-amine (octopamine), 5-hydroxytryptamine (5-HT or serotonin), noradrenaline or dopamine (Henze, 1913; Bottazzi and Valentini, 1924; Erspamer, 1948; von Euler, 1952; Erspamer and Asero, 1953; Hartman et al., 1960). Several enzymes have also been identified in cephalopod saliva, such as hyaluronidase, chitinase, phospholipase A2, peptidase S1, metalloprotease, carboxypeptidase (Romanini, 1952; Fry et al., 2009; Cornet et al., 2014; Whitelaw et al., 2016), as well as tachykinins (Eledoisin, OctTK-I, and OctTK-II) and CAP domain- (CRISP [Cysteine-rich secretory proteins], Antigen 5 [Ag5], and Pathogenesis-related [PR-1]) -bearing proteins (Erspamer and Anastasi, 1962; Kanda et al., 2003; Fry et al., 2009). However, the number of cephalopod species covered in terms of its toxin identification is still acknowledged to be very low (Cooke et al., 2017). From both Superorders of Class Cephalopoda, Octopodiformes accounts for the highest number of investigated species (about 14 from a total of 300). As for Decapodiformes, i.e., cuttlefishes and squids, not even 2% of all known species have been studied for the characterisation of salivary gland secretions. The number of species for which there is detailed information on cephalotoxins is even smaller. The octopuses *Octopus vulgaris*, *Hapalochlaena maculosa*, and *Eledone cirrhosa*, were the first to have α- and β-cephalotoxin, hapalotoxin, and Eledone toxin identified, respectively (McDonald and Cottrell, 1972; Cariello and Zanetti, 1977; Savage and Howden, 1977). Indeed, *Hapalochlaena* spp. (the famous blue-ringed octopus) is particularly known for the potency of its neurotoxic agents, which are lethal and devoid of known antidote. Later, Sheumack et al. (1978) was able to identify tetrodotoxin (TTX) in the posterior salivary glands of the octopus *Hapalochlaena maculosa* as well. This non-peptidic neurotoxin is not, however, exclusive to octopus' salivary glands since it was found in multiple body parts of the animal, including

arms, cephalothorax and abdomen (Yotsu-Yamashita et al., 2007). It must be noted, though, that TTX is a secondary metabolite best-known from pufferfishes (Tetraodontidae) as well as in other marine animals and amphibians, with its synthesis being associated to endosymbiotic bacteria (Lago et al., 2015). Despite the aforementioned studies, information on the bioreactivity and structural properties of cephalotoxins is still very limited. However, cephalotoxins from octopuses and *Sepia esculenta* have been conclusively identified as functional proteins. The latter species, in particular, yielded, to date, the only complete cephalotoxin amino acid sequence available from a curated record at UniProt, which illustrates how little is known about these substances whose existence is known for more than a century. Toxins from *Sepia pharaonis* and *Sepia prashadi*, and from the squid *Doryteuthis (Amerigo) pealeii* have also been identified as unknown cytotoxic proteins secreted by posterior salivary glands (Kem and Scott, 1980; and purified by Karthik et al. (2015, 2019), which may indicate potential cephalotoxin candidates.

Table 1- A comparative overview of cephalotoxins and other major proteinaceous or peptidic neurotoxins considered specific of cephalopods.

Species	Toxin name	Molecular weight (kDa)	Isoelectric point	Known mode-of-action/Bioreactivity	*Post*-translational modifications	Reference
Doryteuthis (Amerigo) pealeii[a]	Unknown cytolytic protein	-	-	-	-	Kem and Scott, 1980
Eledone cirrhosa	Eledone toxin	30 – 70	> 7	Neurotoxin	-	McDonald and Cottrell, 1972
Eledone moschata *Eledone cirrhosa*[b]	Eledoisin[c]	-	-	Hypotensive agent	-	Erspamer and Anastasi, 1962
Enteroctopus dofleini[d]	-	22.15	5.2 – 5.3	Potential neurotoxin	-	Songdahl and Shapiro, 1974
Octopus vulgaris	α-cephalotoxin	91.20	4.5 – 5.1	Neurotoxin (inhibitor of glutamate-mediated synaptic transmission)	-	Cariello and Zanetti, 1977
Octopus vulgaris	β-cephalotoxin	33.90	1.8 – 2.5	Neurotoxin (inhibitor of glutamate-mediated synaptic transmission)	-	Cariello and Zanetti, 1977
Octopus vulgaris	OctTK-I and OctTK-II	-	-	Hypotensive agent	-	Kanda et al., 2003
Sepia esculenta	SE-Cephalotoxin[e]	100	9.08	Neurotoxin	Glycosylation	Ueda et al., 2008
Sepia officinalis	Cephalotoxin	-	-	Neurotoxin	-	Ghiretti, 1959
Sepia pharaonis	-	~50	-	Antibacterial agent	-	Karthik et al., 2015
Sepia prashadi	-	1.96	-	Potential antibacterial and antiviral agent	-	Karthik et al., 2019

[a] Name updated from *Loligo pealei*.
[b] Name updated from *Eledone aldrovandi*.
[c] 12-amino acid neuropeptide.
[d] Name updated from *Octopus dofleini*.
[e] UniProt accession B2DCR8.
[-] Data not available/unknown

1.3 Secretion, delivery, and mode-of-action

As carnivores, cephalopods use envenomation as a means for hunting, however, it may also confer defence against predators (Chichery and Chichery, 1988; Norman and Reid, 2000). Naturally, the secretion of cephalotoxins by salivary glands and their delivery via a wound created by the typical parrot beak of coleoids indicates a major role in feeding. The salivary glands (anterior and posterior) are the constituents of the digestive system responsible for secreting substances that intervene directly in the capture of prey, as well as in lubricating and protecting the digestive tract, besides contributing to the pre-digestive process (see Fernández-Gago et al., 2018 and references therein for details). The anterior salivary glands of cephalopods are located behind the buccal mass and are compound tubuloacinar glands mostly composed by granular and mucous secretory cells (Fernández-Gago et al., 2018). However, the disposition of the posterior salivary glands, where it is believed the active toxins are secreted, is more variable. While octopuses and cuttlefishes possess a pair of posterior salivary glands, squids have only a single gland (see for instance Boucaud-Camou and Boucher-Rodoni, 1983). The posterior salivary glands are adjacent to the digestive gland and are tubular as well. In octopodids, granular and mucous cells are strongly present in glandular epithelia, the latter being responsible for the production of acidic glycoconjugate mucins (Fernández-Gago et al., 2018). In sepioids (cuttlefishes), the mucous cells of salivary glands secrete tryptophan-rich proteinaceous substances (Boucaud-Camou, 1968). Interestingly, enterochromaffin cells (a type of neuroendocrine cells) have been detected in the posterior salivary glands of octopods as well, albeit absent in sepioids (Boucaud-Camou and Boucher-Rodoni, 1983). In turn, salivary gland morphoanatomy for the Teuthoidea (squids) is scarce. In any case, the relationship between salivary gland structure, biochemistry and toxin secretion remains somewhat elusive.

The buccal mass plays a paramount role in toxin delivery. Here is located the posterior salivary gland duct, running within the salivary papilla, and the anterior salivary gland duct, both opening into the buccal cavity. During an attack, the toxic secretions produced by the posterior salivary glands are delivered to the prey through a wound created by quick bite rather than by injection through the two-piece beak (also called rostrum). Toxins with immobilising properties, with the aid of permeabilising enzymes plus cardioexcitatory and vasodilatory substances, are promptly disseminated through the prey's circulatory system, rapidly

affecting mobility and motor coordination (Zhang, 2015; Lobo-da-Cunha, 2019). Likely, the copious amount of mucus comprising neutral glycoproteins, sialic acid, dipeptidase and hyaluronidase secreted by the anterior salivary glands, is the delivery vehicle of the bioactive compounds secreted by the posterior salivary glands (Budelmann et al., 1997; Ponte and Modica, 2017).

The number of detailed studies on the toxicological effects of cephalopod salivary secretions is still considerably low. Most bioassays were performed on decapod crustaceans and in general report the paralysing properties of crude secretions from a few species of octopuses, squids and cuttlefishes (Ghiretti, 1959; McDonald and Cottrell, 1972; Songdahl and Shapiro, 1974; Cariello and Zanetti, 1977; Savage and Howden, 1977; Ueda et al., 2008; Cornet et al., 2014). However, there are some studies performed on mice that suggest toxicity to vertebrates as well (Ueda et al., 2008). Whenever the chemical nature of these secretions was investigated, cephalotoxins were consistently present and considered the compounds with highest toxic activity. Cephalotoxins seem to have a powerful paralysing action against models as distinct as crustaceans and mice by blocking the electrical and mechanical responses of muscles (McDonald and Cottrell, 1972), even though the exact mechanisms are not known. Songdahl and Shapiro (1974) also stated that the speed of the toxicological reaction may be dose-dependent. Further, Ueda et al. (2008) demonstrated that distinct cephalopod species secrete toxins with different potencies against the same target species, i.e., whereas salivary extracts from cuttlefish were lethal to crabs and non-toxic to mice, extracts from squid revealed higher toxicity against mice. Cariello and Zanetti (1977) verified a similar toxicological reaction of the cephalotoxins upon octopus itself, causing significant lethality as well. Additionally, toxicological studies on strains of Gram-positive and Gram-negative bacteria confirmed that peptidic extracts of the *Sepia officinalis* posterior salivary glands revealed bactericidal and antimicrobial activities (Cornet et al., 2014). Disclosing allo- or orthosteric ligands for cephalotoxins in target organisms would provide a major leap in understanding mode-of-action of these toxins and eventually explore their biotechnological potential.

1.4 Molecular characterisation of cephalotoxins

Knowledge on the molecular structure of cephalotoxins is scant or incomplete. The most complete characterisation of these toxins, so far, was obtained by Ueda et al. (2008), who

disclosed a full 1052-amino acid sequence for a cephalotoxin variant from the posterior salivary glands of the golden cuttlefish, *Sepia esculenta*. This form, termed SE-cephalotoxin by the same authors appears to be a ≈ 100 kDa monomeric protein whose amino acid sequence was determined from isolated cDNAs. The same authors also purified the protein by high-performance liquid chromatography (HPLC) and confirmed the findings by Western Blotting. Some of the most distinctive features of the protein is glycosylation at five distinct sites and the presence of four conserved domains, namely EGF-like, Sushi, TSP type-1 and LDL-receptor class A (Figure 1). Some of these, particularly the EFG- (epidermal growth factor) like domain have already been described in venom proteins such as in the gigantoxins from the cnidarian *Stichodactyla gigantea*, which interfere with sodium channels and have paralytic activities against crustaceans (Shiomi et al., 2003). Cephalotoxins may thus be considered as belonging to the class commonly referred to as EGF (neuro)toxins.

Despite the absence of further cephalotoxin mRNA or amino acid sequences from curated databases, cephalotoxin-like proteins (possibly non-toxic) have been found in a number of aquatic organisms, from fish to the exoskeleton of some corals (e.g., Ramos-Silva et al., 2013). However, the role and evolution of these proteins remain obscure. Still, important insights have been provided by Ruder et al. (2013), who used an RNA-Seq transcriptomic approach to study multiple toxins secreted by octopus, cuttlefish and squid, revealing for the first time the existence of cephalotoxin-like proteins in all these three major groups by homology matching.

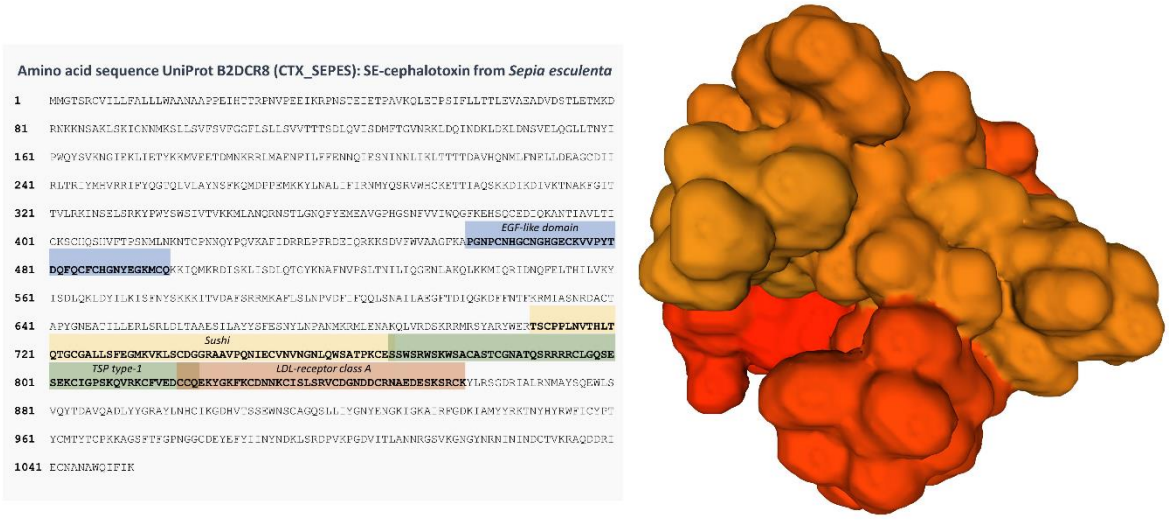


Figure 1 - Sequence and a predicted 3D model of SE-cephalotoxin. Conserved domains are highlighted in the sequence, namely EGF-like, Sushi, TSP type-1, and LDL-receptor class A. Molecule was produced with Swiss-

Model from the UniProt record B2DCR8 (CTX_SEPES). Glycosylation is not represented. See Ueda et al. (2008) for further details.

1.5 Prospects for potential biotechnological applications

As mentioned earlier, it is believed that the immobilising or paralysing activity of cephalotoxins against invertebrate prey, especially arthropods, can result from the interference with specific ion channels or receptors. At the very least, cephalotoxins could be explored for the development of natural and safer pesticides against arthropods (namely insecticides). Nonetheless, verified effects on murines open excellent prospects for biomedical applications, similarly to conotoxins and TTX, especially for the development of safer and non-addictive painkillers. In fact, Cariello and Zanetti (1977) reported cephalotoxins from *O. vulgaris* to be able to block glutamate-mediated synaptic transmission, a signalling process involved precisely in pain sensation (see for instance Wozniak et al., 2012). Nonetheless, only few bioactives from the cephalopod salivary glands are indeed under scientific scrutiny for potential applications. Among such studies, research on extracts from *Sepia pharaonis* and *Sepia prashadi* posterior salivary glands revealed antimicrobial activity against avian bacterial pathogens and also a potential function as an anti-metastatic agent (Karthik et al., 2015, 2017, 2019). These findings suggest that, adding to neurotoxic effects, biocidal and anti-proliferative properties of cephalotoxins and other components of cephalopod venoms can point to novel antibiotics and anti-cancer drugs, further confirming the rising interest of cephalopod toxins for biotechnology. It must be noted that even TTX, albeit being ubiquitous among many animal taxa (cephalopods included) and one of the best-known and most lethal non-peptidic neural ion channel blockers, has also been considered for the development of analgesic and anaesthetic drugs in advanced cancer patients (Hagen et al., 2007). Indeed, under the trade name Tectin, TTX underwent phase III trials for the treatment of pain resulting from chemotherapy in 2014 (Newman and Cragg, 2014). To these applications we may add the potential deployment of TTX in the treatment for opioid dependence and management of withdrawal symptoms, with has already started phase IIa clinical trials as Tetrodin (Butler, 2005; Shi et al., 2009; Song et al., 2011). Similarly, the TTX-based drug Tocudin, has also started preclinical studies as a formulation intended for local anaesthesia (Butler, 2005). These findings highlight the high prospects set upon potential applications of marine animal toxins. Proteinaceous toxins such as cephalotoxins and conotoxins may offer advantages over those

derived from primary or secondary metabolites, as they can be safer, more easily eliminated (by proteolytic activity) and, very importantly, directly cloned into adequate vectors for heterologous expression. In the latter case, constraints posed by protein secretion and *post*-translational modifications, such as glycosylation, can be circumvented by choosing eukaryote models, like yeast.

1.6 Lessons learned and future perspectives

Even though the discovery and description of coleoid venomous secretions dates back to the XIX century, their toxins remain obscure. Cephalotoxins, albeit being consensually found to be the most bioreactive substances in the venom secreted by the posterior salivary gland of these animals (whereas the anterior gland provides the mucin-rich vehicle of delivery), are no exception. Their potent neurotoxic and potential cytotoxic effects, together with the proteinaceous nature, make them particularly appealing for drug discovery. Such endeavour requires, nonetheless, more solid knowledge of their molecular structure and their interaction with target receptors. With this respect, "omics" approaches, particularly proteomics and transcriptomics, can greatly assist in the filling-in of proteins secreted by the salivary glands of cephalopods by screening multiple peptides and proteins in single runs. These approaches may therefore provide a scaffold to support downstream finer structural details of the substances. The "big picture" must also be complemented with finely tuned toxicity and bioreactivity testing to provide the comprehensive bottom-up perspective needed to devise effective applications.

2

OBJECTIVES

The general goals of this thesis consisted in exploring toxins as ecological and evolutionary features as means to identify their function and predict molecular targets and yield its validation for *in silico* approaches designed to predict biotechnological applications for novel bioactives from the target species. Therefore, taking advantage of the importance and abundance of cephalopods from Portuguese coastline, the cephalotoxins from *Octopus vulgaris* and *Sepia officinalis* were explored integrating ecophysiology, toxicology, molecular biology and bioinformatics. Taking this into consideration, the project aimed, specifically, at:

- Isolate and identify the major cephalotoxin isoforms extracted from the salivary glands of *Sepia* and *Octopus* from the Portuguese coast and understand how they are secreted, transported and delivered.
- Characterise native toxins at the molecular and toxicological levels in order to provide a validated and comprehensive overview of toxin function and their sub-cellular receptors.
- Contribute to understand how cephalotoxins evolved using molecular- and phenotype-based phylogenetic models.
- Through a top-down approach, from ecology and evolution to molecular biology, enable translational science by predicting cephalotoxin biotechnological applications and contributing to find novel toxin-secreting cephalopods.
- Make an important contribution to the need to preserve marine biodiversity by demonstrating that it is a major societal benefit, thus assist steering the exploitation of marine resources towards a sustainable 'Biotechnology for Blue Growth' strategy in the EU.

To address these main goals, a set of tasks summarised in Figure 2 were performed. These included acknowledged methods for microanatomical assessment, combined with RNA-Seq based transcriptomic and bioinformatics.

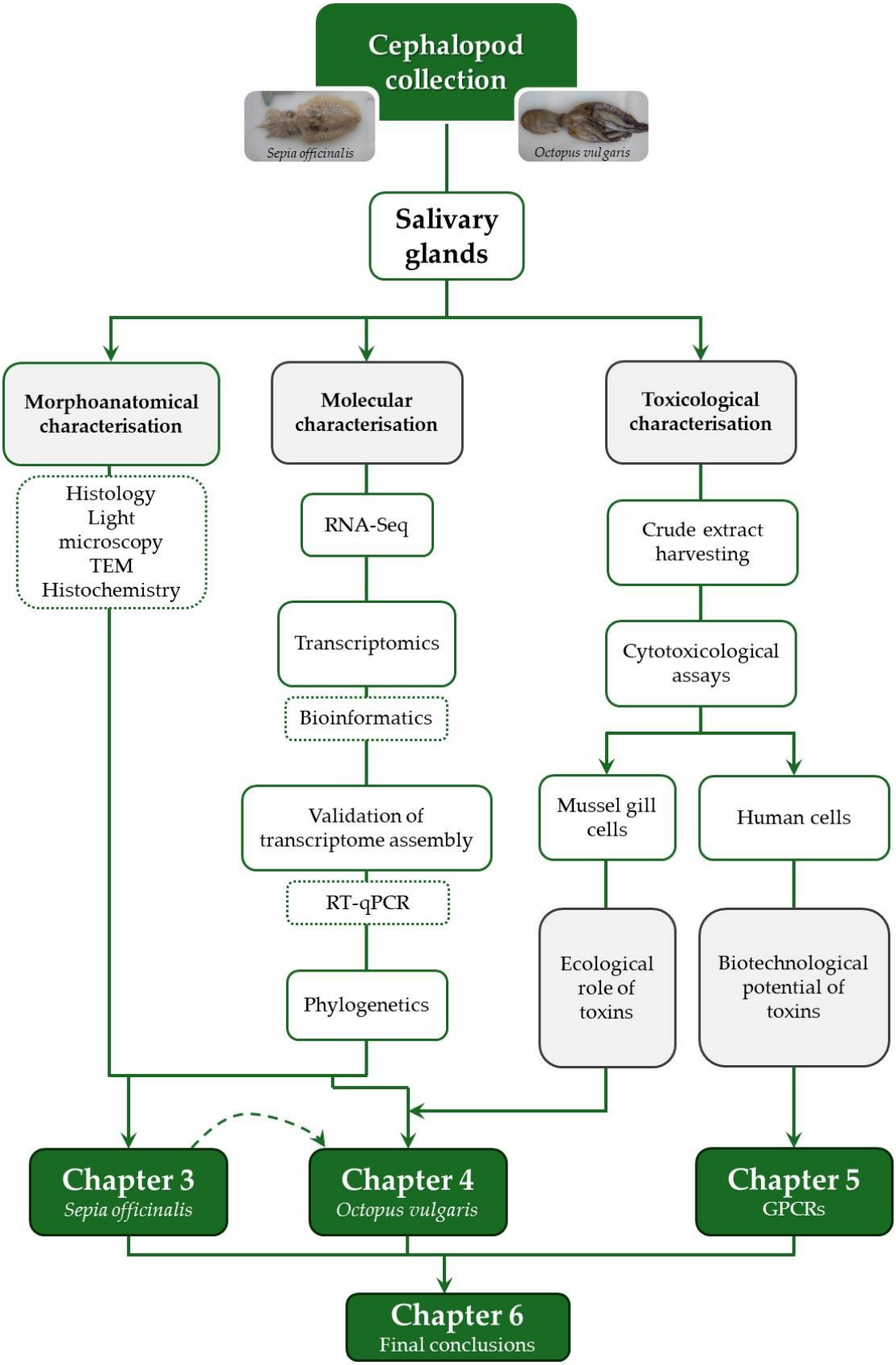


Figure 2 - Flowchart representing the thesis structural organisation.

3

TRANSCRIPTOME PROFILING OF THE POSTERIOR SALIVARY GLANDS OF THE CUTTLEFISH *SEPIA OFFICINALIS* FROM THE PORTUGUESE WEST COAST

This chapter has been published in the following paper:

GONÇALVES C, MOUTINHO CABRAL I, ALVES DE MATOS AP, GROSSO AR & COSTA PM (2024) Transcriptome profiling of the posterior salivary glands of the cuttlefish *Sepia officinalis* from the Portuguese West coast. Frontiers in Marine Science 11:1362824. (DOI: 10.3389/fmars.2024.1362824)

Investigation, methodology and writing the original manuscript were performed by **C. Gonçalves**. Formal analysis and data curation were shared by **C. Gonçalves** and I. Moutinho Cabral.

3.1 Abstract

Cephalopods like octopuses and cuttlefishes are known to secrete a 'toxic saliva' to inject into their prey, especially crustaceans since the XIX century. However, only in the mid-XX century were the first coleoid-specific toxins successfully isolated. Motivated by the growing interest on the global ocean as an almost inexhaustible source of novel bioactive compounds, we used RNA-Seq – based transcriptomics and *de novo* assembly of transcriptomes to screen the posterior salivary glands of *Sepia officinalis* (the common cuttlefish) from the Portuguese West coast for toxins and other bioactive proteins and peptides. Supported by microanatomical analyses, the posterior salivary glands constitute indeed the 'venom gland' whereas the more elusive anterior salivary glands (embedded in the buccal mass) are responsible for the production of mucin-rich saliva that is effectively the vehicle that transports the toxins as the venom is injected into the prey. Indeed, the transcriptomic profiling suggests that the cuttlefish venom is complex mixture of bioactive proteins, among which neurotoxins are major players, together with enzymes whose function is to digest the extracellular matrix to facilitate diffusion of the toxins. Nonetheless, by comparing with previous RNA-Seq data obtained from *S. officinalis* collected from other biogeographical areas, it may be suggested that significant inter-populational variation in venom composition can occur, which may potentially increase the span of bioactives secreted by these animals. We isolated and validated the full coding sequences for three important toxins, a cysteine-rich venom protein (CRVP), a venom insulin (VIns) and a cephalotoxin (CTX). The toxins seem to be relatively conserved among coleoids but diverging from other venomous molluscs such as cone snails. Their properties as potent modulators of glucose (in the case of VIns) and as potential neurotoxins (like CRVP and CTX) can render them primer targets for drug development.

3.2 Introduction

The constant demand for novel effective bioproducts has been attracting growing enthusiasm on oceans as they harbour an immense biodiversity which is linked with a unique chemical diversity. Marine organisms that bear a chemical warfare are particularly interesting for ocean's bioprospection, since they secrete complex cocktails, as a strategy for predation or defence. These bioactive substances, especially toxins, are noteworthy for biotechnological applications since they interfere with specific molecular targets, such as ion channels,

receptors and enzymes, and may consequently impair particular physiological pathways (Calvete et al., 2009). Altogether, the use of these natural bioactives can be a safer, efficient and sustainable therapeutic alternative to the risky and costly process of designing synthetic molecules (see for instance Fusetani, 2009; Rodrigo and Costa, 2019).

The reduced number of substances based on marine compounds translated into the market can be particularly explained by the tremendous marine biodiversity and the downsides associated with it, such as the scarcity of genomic resources. Here, not only can omics shed some light on these less-studied marine species, but they can also be a helpful tool for bioprospecting for novel bioactive compounds with biotechnological applications. Untargeted approaches, such as RNA-Seq, allow to identify multiple molecules without having a reference genome and transcriptome, which is the case of non-conventional model marine species (Martins et al., 2019; Rodrigo and Costa, 2019). Apart from unravelling and characterising novel proteinaceous bioactives, transcriptomics also facilitates the prediction of their mode-of-action and possible targets. Peptides and proteins, especially toxins, are particular noteworthy as candidates for novel drugs as they can interact with specific molecular targets and can potentially be synthetised by heterologous expression.

Studies on marine invertebrates are showing their promising potential for producing molecules, including proteinaceous toxins, with putative biotechnological applications (Leal et al., 2012). The bioactive substances of these marine organisms are presenting interesting biomedical properties (e.g., painkiller, antitumoral, anticoagulant, anti-inflammatory, antimicrobial and antiviral) as well as cosmetical and nutritional applications or even as pesticides (Molinski et al., 2009; Martins et al., 2014; Rodrigo and Costa, 2019, and Cappello and Nieri, 2021). A group of organisms that emerged as a source of novel marine compounds for drug development are the Mollusca. Within this Phylum, cone snails are highlighted as pioneer organisms for marine biotechnologists due to the pharmaceutical value of peptidic bioactives such as conopeptides (Shen et al., 2000; Turner et al., 2020). Indeed, ziconotide (commercialised as Prialt®), a recombinant peptide designed from an ω-conotoxin secreted by the cone-snail *Conus magus*, became one of the first formally approved marine-inspired pharmaceuticals in Europe and the United States (since the early 2000s) for the management of severe chronic pain in patients insensitive or unable to take other treatments, being delivered intrathecally (Olivera et al., 1987; Bowersox et al., 1996; Williams et al., 2008). Adding to these gastropods, coleoid cephalopods are molluscs that have also been suggested to secrete molecules with promising biotechnological proprieties. Indeed, their saliva can be venomous

and contains multiple interesting bioactives, from toxins to permeabilising agents, such as neurotoxins termed cephalotoxins (for a review, refer to Gonçalves and Costa, 2021). The secretion of these substances by the salivary glands can be fundamental for cephalopods to capture their preys (e.g., crabs) as they seemingly have paralysing action. Cephalotoxins are secreted by the posterior salivary glands (PSG), as previously reported by Ghiretti (1959) and Ueda et al. (2008).

As cephalopods are becoming an auspicious group of marine invertebrates for biotechnologists, behavioural biologists and even neurobiologists, as examples, researchers are turning to state-of-the-art omics approaches for bioprospecting these animals for novel bioactives due to the ability to screen for multiple untargeted molecules in single runs (see Albertin and Simakov, 2020, as well as Baden et al., 2023). Despite the relatively scant number of omics studies, comparatively to more conventional model organisms, in marine invertebrates in general, including cephalopods, altogether leading to reduced molecular resources (such as reliable genomic annotation) compared to more conventional model organisms, there are already available literature suggesting molecular approaches as roadmaps towards bioprospecting cephalopods for proteinaceous bioproducts (Ruder et al, 2013; Caruana et al., 2016; Whitelaw et al., 2016). These works are accompanied by studies using proteomics and transcriptomics to describe the venom systems of marine invertebrates, from cone snails (Fassio et al., 2019; Fedosov et al., 2021) to annelids (von Reumont et al., 2014b; Rodrigo et al., 2021; Moutinho Cabral et al., 2022) and even the first known poisonous crustacean, the remipede *Xibalbanus tulumensis* (von Reumont et al., 2017). These works show the potential of these methodologies and novel pipelines of research, from evolution to bioprospecting. In addition, previous analyses of the transcriptome and proteome of the PSG of the cuttlefish *Sepia officinalis* (Linnaeus, 1758) from the coast of France associated with bioassays suggested that these glands could have two functions linked with predation and immune defence as their substances and extracts have paralysing and antimicrobial activities (Cornet et al., 2014).

The present work aims at exploring the PSG of the common cuttlefish, *Sepia officinalis* (Linnaeus, 1758) from the Portuguese West coast, at several levels: (i) comparatively characterise the microanatomy of the PSG and the anterior salivary glands (ASG) and associate their ultrastructure with specific functions, such as the secretion of toxins; (ii) identify and annotate transcripts putatively coding for toxins produced by this organ, in the case using a

RNA-Seq – based transcriptomic approach, and (iii) isolate and analyse the full coding sequence of relevant toxins.

3.3 Materials and methods

3.3.1 Animal collection

Following the implementation of The Three Rs principle (first described by Russel and Burch and presently inserted in the Directive 2010/63/EU on the protection of animals used for scientific purposes) the animals used in this study were provided by local fishermen. Animals were euthanized by sectioning the central brain and the tissues of interest were harvested immediately. Five *Sepia officinalis* (mantle length 14.5 ± 2.0 cm; weight 227.2 ± 68.9 g) were collected from two beaches of the Portuguese West coast: Praia do Norte (38°39′0.153″ N, 9°14′43.876″W) and Praia do Outão (38°29′19.422″ N, 8°56′0.455″ W). The PSG were extracted and divided for both histological analyses and RNA extraction. The ASG were also excised and prepared for histology to add a structural comparison between the two salivary glands.

3.3.2 Sample preparation for microscopy analyses

Samples were immediately fixed after dissection. For optical microscopy, samples were fixed using Davidson's (9–10% v/v formalin, 10% v/v glacial acetic acid and 30% ethanol) and Zenker's (2.5% m/v potassium dichromate, 3% m/v mercury chloride, 1% sodium sulphate and 5% v/v glacial acetic acid) fixatives, both prepared in MilliQ–grade water (> 16 MΩ.cm). Fixation was done at room temperature for 24 h in case of Davidson's and Zenker's fixatives, followed by washing in MilliQ–grade water (4 × 15 min), dehydrated through a progressive series of (aqueous) ethanol: 70% (1 × 30 min), 95% (2 × 15 min) and 100% (3 × 30 min), following by intermediate infiltration and embedding with xylenes (3 × 15 min) and molten paraffin (overnight), respectively. Samples were then sectioned (5 μm thickness) using a Leica RM2245 semi-automated rotary microtome. Histological sections were stained through a tetrachrome staining procedure that combines Alcian Blue for acid sugars; Periodic Acid–Schiff's for neutral polysaccharides; Weigert's Iron Hematoxylin for chromatin, and Picric Acid as counterstain for muscle fibres and cytoplasm, following Costa and Costa (2012). Previous to staining, slides were deparaffinated with xylenes (1 min) and rehydrated through a regressive ethanol series: 100% (1 min), 95% (1 min), 70% (30 s), finalised with MilliQ-grade water (6 min).

Optical microscopy analyses were done with a DM 2500 LED model microscope with a MC 190 HD camera (all from Leica Microsystems). Image acquisition was made using the software Leica Application Suite Version 4.13.0 (Leica Microsystems). Image processing was done with ImageJ (Schneider et al., 2012) and GIMP (Montesanto, 2015).

Glutaraldehyde 2.5% v/v (in 0.1 M sodium cacodylate buffer, pH 7.4) or paraformaldehyde 4% m/v in phosphate-buffered saline (PBS), were used to fix samples for transmission electron microscopy (TEM). The fixation was performed over 2 h at room temperature, followed by washing in cacodylate buffer or PBS (3 × 15 min), *post*-fixation in 1% m/v osmium tetroxide (OsO_4) in 0.1 M sodium cacodylate buffer of PBS and washing in MilliQ–grade water (3 × 15 min). These samples were then embedded in Epoxy resin (Sigma-Aldrich, St. Louis, MO, USA), following Luft's (1961) methodology. In brief, samples were dehydrated through a progressive acetone series: 30%, 60%, 70%, 90% and 100% (15 min each), followed by intermediate infiltration with polypropylene oxide:Epoxy of 2:1, 1:1 and 1:2 (30 min each). A final infiltration was performed with Epoxy resin, in vacuum for 30 min. Resin blocks were left to polymerise overnight at 60°C. Thin sections (50–60 nm) were obtained using a Leica Reichert-Jung Ultramicrotome Ultracut E. Sections were collected onto copper mesh grids and contrasted with 2% (w/v) aqueous Uranyl Acetate for 1 h 20 min and Reynold's lead citrate for 8 min (Venable and Coggeshall, 1965). Grids were analysed using a 100-SX model TEM (JEOL, Tokyo, Japan) operated at 80 keV.

3.3.3 Total RNA extraction and RNA-Seq

After excision, samples were immediately stabilised in RNA*later* reagent (Sigma-Aldrich, St. Louis, MO, USA), where remained during 10 days at 4°C. An exchange to fresh RNA*later* was made during that period. Afterwards, samples were archived at -80°C, until total RNA extraction. Total RNA was extracted from ≈20 mg of tissue per sample (n = 3 PSG from three different individuals) using the RNeasy Protect Mini Kit combined with the RNase-Free DNase Set (all from Qiagen, Hilden, Germany) for an efficient on-column digestion of DNA, following manufacturer instructions. Preliminary quantification of total RNA and initial quality assessment was performed using a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The RNA integrity number (RIN) and quantification of total RNA were obtained using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and were found to range between 4.8. and 7.2, which is deemed adequate for marine invertebrates, with an input above 1 mg uncontaminated RNA (refer to the guidelines

suggested by Gayral et al., 2011; Gallego Romero et al., 2014; Puchta et al., 2020). The library construction of cDNA was carried out using a Stranded mRNA Library Preparation Kit. The generated DNA fragments were then sequenced on the Illumina NovaSeq platform (Illumina, San Diego, CA, USA), with a mean length of 150 bp paired-end reads within a coverage of 40 M reads.

3.3.4 Transcriptome *de novo* assembly and annotation

The quality control check of raw data obtained from RNA-Seq was pre-evaluated with FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The TrimGalore v0.6.6 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to discard low quality reads, i.e., reads under 20 bp of length. After applying the default parameters, less than 0.9% of reads were removed per sample. Filtered data was *de novo* assembled using Trinity v2.6.6 with default parameters (Grabherr et al., 2011; Haas et al., 2013). The quality of transcriptome assembly was evaluated through contigs N50, Ex90N50 and read content statistics, using TrinityStats, Bowtie2 v2.4.1 and Samtools v1.11 (Langmead and Salzberg, 2012; Haas et al., 2013; Langmead et al., 2019; Danecek et al., 2021). Transcript abundance was estimated for the novel assembled transcripts using Kallisto v0.43.0 (Bray et al., 2016). Subsequent analyses were performed using R 4.*x* (Ihaka and Gentleman, 1996), where the packages Tximport v1.28.0 and SeqinR v4.2–30 were used to import data (Charif and Lobry, 2007; Soneson et al., 2015). Normalised gene expression values were assessed through transcript per millions (TPMs). Coding regions were then predicted in the assembled transcriptome through TransDecoder v5.5.0, using the default parameters (Haas et al., 2013). The predicted open reading frames (ORFs) were functionally annotated by scanning for homology against Swiss-Prot (version 18–11-2020 13:41, downloaded from https:// ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz) and five customised UniProtKB databases (UniProt release 2023_03) using BLASTP from NCBI blast+ v2.13.0, with a cut off *e*-value of 1×10^{-5} (Altschul et al., 1990; Camacho et al., 2008; UniProt Consortium, 2021). The five customised databases were built from a subset of either UniProtKB or Swiss-Prot. The customised databases consisted of: (1) The 'SeaTox' database is mostly a customised subset of UniProtKB containing manually annotated toxin- or venom-related proteins from Eukaryota (excluding human), plus unreviewed proteins from Annelida, Cnidaria and Cephalopoda (this database can be freely accessed at https://dx.doi.org/10.5281/zenodo.10418296); (2) The 'Toxin' database is a subset of Swiss-Prot including toxin-related proteins and cephalotoxins;

(3) 'Enzyme' database is a subset of Swiss-Prot including enzyme-related proteins; (4) 'PTM' database is a subset of Swiss-Prot including proteins associated with *post*-translational modifications (PTMs), according to the Gene Ontology annotation; (5) 'Secreted' database is a subset of Swiss-Prot including secreted proteins (UniProtKB, location_sl_0243). In addition, ORFs were also functionally annotated by searching for reference protein domains (*e*-value < 0.05) against the Pfam database (Pfam-A.hmm version 35.0, Nov 2021) using HMMER v3.3.1 (Eddy, 2009; Mistry et al., 2021). The R script of transcriptomics analyses can be found in A.1 Appendix to Chapter 3.

3.3.5 Validation by RT-qPCR

The RNA-Seq results were validated by reverse transcription quantitative polymerase chain reaction (RT-qPCR) of selected representative transcripts that encode for proteins with toxin or hormone activities or interfere with ion channels. The selected transcripts from RNA-Seq data encompassed a cysteine-rich venom protein (CRVP) as cysteine-rich secretory proteins (CRISPs) are often referred to as means to distinguish from their non-toxin counterparts, a cephalotoxin (CTX) and a venom insulin (VIns). The full coding sequence of these transcripts were firstly isolated and amplified by PCR. Primers (Table 2) were evaluated with in silico using PCR Primer Stats (Stothard, 2000) and OligoAnalyzer (https://eu.idtdna.com/calc/analyzer). Specific primers were designed to quantify expressed sequence tags (ESTs) in the coding region by RT-qPCR, avoiding conserved domains. The 18S gene was chosen as housekeeping. The cDNA was synthesised from total RNA samples using the NZY First-Strand cDNA Synthesis Kit (NZYTech, Lisbon, Portugal), following manufacturer instructions. Regions of interest were isolated by PCR in a Biometra TOne 96 gradient thermocycler (Analytik Jena, Jena, Germany) using a PCR kit from Invitrogen (Thermo Fisher Scientific, Waltham, MA, USA). After resolving in a 1.2% agarose gel, PCR products were Sanger-sequenced for sequence confirmation. Gene expression analysis was performed in a Rotor-Gene Q (Qiagen, Hilden, Germany) using the NZY qPCR Green Master Mix (NZYTech, Lisbon, Portugal). The program comprised an initial denaturation procedure (95 °C, 10 min), followed by 40 cycles of denaturation (94 °C, 45 s), annealing (52 °C, 25 s) and extension (72 °C, 30 s). Expression was determined based on a simple ΔCt approach (Silver et al., 2006). Homoscedasticity and normality of data were assessed through Levene's and Shapiro-Wilk's tests, respectively. Differences in gene expression between the three genes

were evaluated through the parametric Tukey's HSD test. A significance level $\alpha = 0.05$ was set for all analyses. Statistics were computed using R 4.*x* (Ihaka and Gentleman, 1996).

Table 2 - Primer sequences used for sequence isolation by PCR and expression analysis by RT-qPCR in the posterior salivary glands of *Sepia officinalis*.

Gene	Primer	PCR/RT-qPCR	Primer sequence (5′- 3′)	Amplicon size
18S	Forward	PCR/RT-qPCR	GTTCCGACCGTAAACG	140 bp
	Reverse	PCR/RT-qPCR	CCTTCCGTCAATTCCT	
CRVP	Forward	PCR	AAGATGAAAGCTGTAGTATC	984 bp
	Reverse	PCR/RT-qPCR	AGGACACTTTTCTTTGC	
CTX	Forward_1	PCR	ATGGTTCTGTGGCAGAG	809 bp
	Reverse_1	PCR	GCCTTCGGTATCTGCAT	
	Forward_2	PCR	CGTCACTTCCGGTTCT	815 bp
	Reverse_2	PCR	CGCTTGAGATCGAAGGA	
	Forward_3	PCR	GCTGTAGACACAAACATTGA	794 bp
	Reverse_3	PCR/RT-qPCR	GCAGTAGCATTCCATTCAA	
	Forward_4	PCR	TGAGTGCAGGTTATCGG	947 bp
	Reverse_4	PCR	AAGCAATATGAGGTCGGTA	
VIns	Forward	PCR/RT-qPCR	GGGACTATGGTCTTAAGTTTC	518 bp
	Reverse	PCR	GTATGCTTATGATGACGAGTC	
CRVP	Forward	RT-qPCR	GAAATCAGGTACTGCTG	151 bp
CTX	Forward	RT-qPCR	CCGCCCTTGAATATCAC	161 bp
VIns	Reverse	RT-qPCR	CTCATTGCAGGTGTGTT	156 bp

The housekeeping gene encoding for the 18S ribosomal subunit was used for normalization (18S), whereas three full coding sequences of venom-related toxins were selected for transcriptome validation: Cysteine-rich venom protein (CRVP), SE-cephalotoxin (CTX) and Venom Insulin (VIns).

3.3.6 Phylogenetics of representative toxin transcripts

Three full coding sequences of the shortlisted transcripts, i.e., cysteine-rich venom protein latisemin (CRVP), a cephalotoxin (CTX) and a venom insulin (VIns) were analysed. Each sequence was scanned for homology against NCBI RefSeq_protein database using BLASTP

(Altschul et al., 1990; Camacho et al., 2008). The search was restricted to related taxonomic groups as 'Lophotrochozoa', phyla 'Arthropoda', 'Mollusca' and 'Cnidaria' as well as the specific classes: 'Cephalopoda', 'Arachnida', 'Gastropoda'. The most significant hits were chosen based on the best matches against venomous or toxin-bearing animals, *e*-value and percentage of identity retrieved from running BLASTP. Sequence alignment and dendrograms were made with MEGA X (Kumar et al., 2018). Phylogenetic consensus trees were produced for each selected sequence using the Maximum Likelihood method and Jones-Taylor-Thornton model, with 800 bootstrap replications (Jones et al., 1992). Tree branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (800 replicates) are shown next to the branches (Felsenstein, 1985). The trees were rooted on related sequences from Arthropoda. Sequence homology-based models of protein structure for each sequence of interest were constructed using Swiss-Model (Waterhouse et al., 2018), which takes AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) as reference, together with data from NMR, Cryo-EM, crystallography and other methods.

3.4 Results

3.4.1 The diverticular nature of the posterior salivary glands

The PSG or venom glands of *Sepia officinalis* are an exocrine paired organ located adjacently to the visceral mass, close to the digestive gland. This organ is connected to the buccal cavity by an excretory duct that runs through the salivary papilla. Macroscopically, the PSG is homogeneous whitish in appearance, but its internal microanatomy reveals that it is formed by branched glandular tubules, seemingly blind-end diverticula (Figure 3A). The tubules are composed by columnar glandular epithelium whose cells are basally basophilic, easily recognised by its bluish or brown-black colours after staining with Harris's hematoxylin or Weigert's iron hematoxylin, respectively (Figures 3A, B). The organ also bears blood vessels and nervous tissue in the intertubular space, the latter of which is invariably formed chiefly by fibrocytes (see Figure 3B insets for examples). The basophilic nature of the basal cytoplasm of cells is mainly due to dense rough endoplasmic reticulum contiguous to nuclei (Figures 3C, D), which is conspicuous under TEM (Figure 4A). Apically, the columnar cells hold picric acid-positive secretory granules, which is indicative of proteinaceous content. Depending on their

maturation stages, these dense protein-containing vesicles (secretory granules) can differ by their hue, which can range from yellowish to orange (visible in Figure 3C). This may be related to the cascade of events involving protein folding and other *post*-translational modifications as vesicles mature *a priori* to secretion. Protein biosynthesis and vesicle maturation is visible under TEM as well, resulting in abundant and often densely packed secretory granules (Figures 4B, C). No evidence was found for goblet (mucus-secreting) cells structurally as well as no evidence for the common histochemical signature of acidic or neutral mucins through Alcian Blue and Periodic Acid – Schiff's (PAS) staining, respectively.

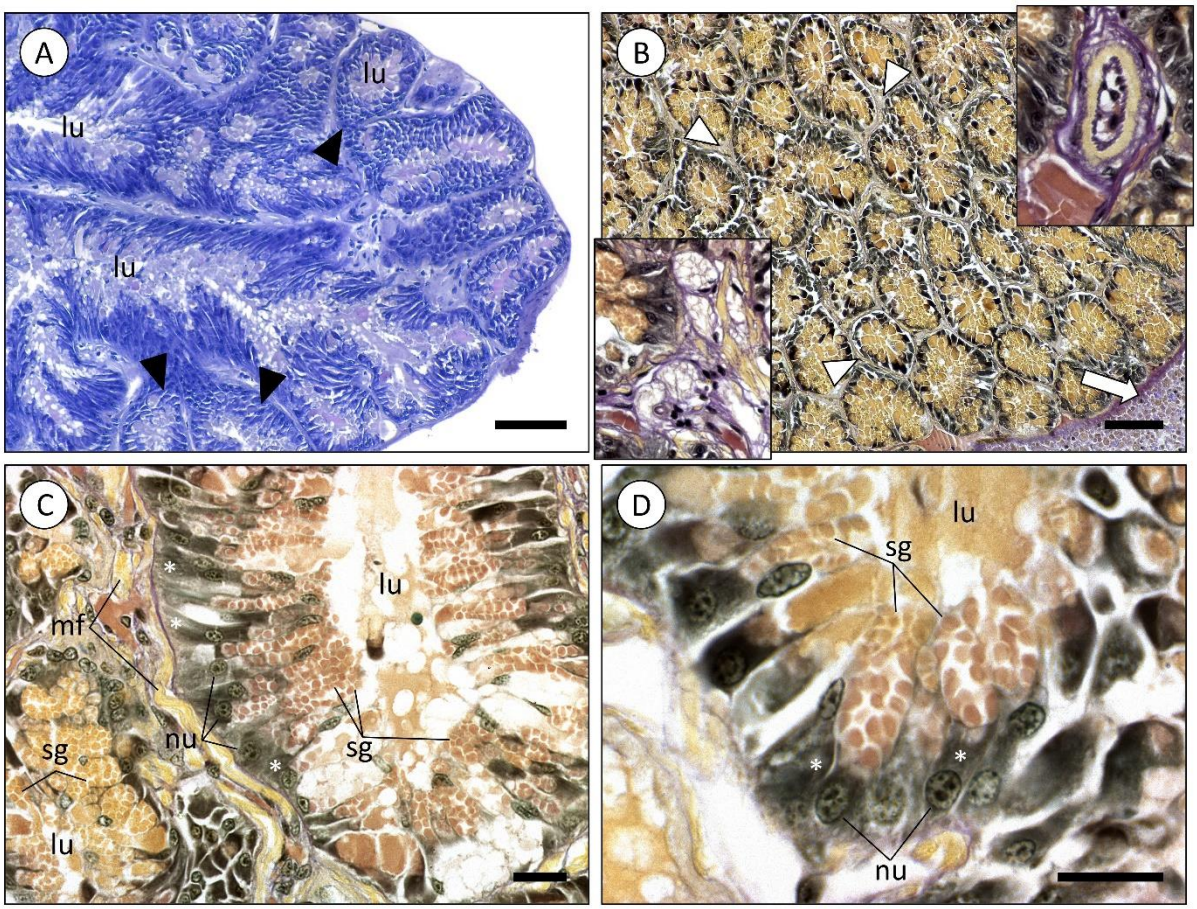


Figure 3 - Histological photomicrographs of the posterior salivary glands of *Sepia officinalis*. **(A)** Overview of the gland structure stained with Hematoxylin & Eosin. Note the presence of the several different-shaped and sized blind-end diverticula, or 'tubules' (arrowheads), that are strongly basophilic basally, indicating dense rough endoplasmic reticulum. **(B)** General appearance of the posterior salivary glands revealed using a histochemical tetrachrome stain. The white arrow denotes the edge of the digestive gland. The secretory granules within glandular cells are stained by picric acid (yellowish), revealing the presence of proteinaceous material. Bottom inset: Detail of nervous tissue inserted within the glandular tubules. Upper inset: Detail of a blood vessel (arteriole-like) within the network of diverticula. **(C)** Diverticula formed by columnar glandular cells. The staining gradient among secretory granules most likely corresponds to different maturation stages. **(D)** Detail of glandular cells with strong

basophilic cytoplasm (basal portion of cells), which indicates abundant rough endoplasmic reticulum (white asterisks). The apical portion of these cells are filled with secretory granules. Abbreviations: lu, lumen; mf, muscular fibres; nu, nucleus. Scale bars: **(A, B)** 100 μm; **(C)** 25 μm; **(D)** 20 μm.

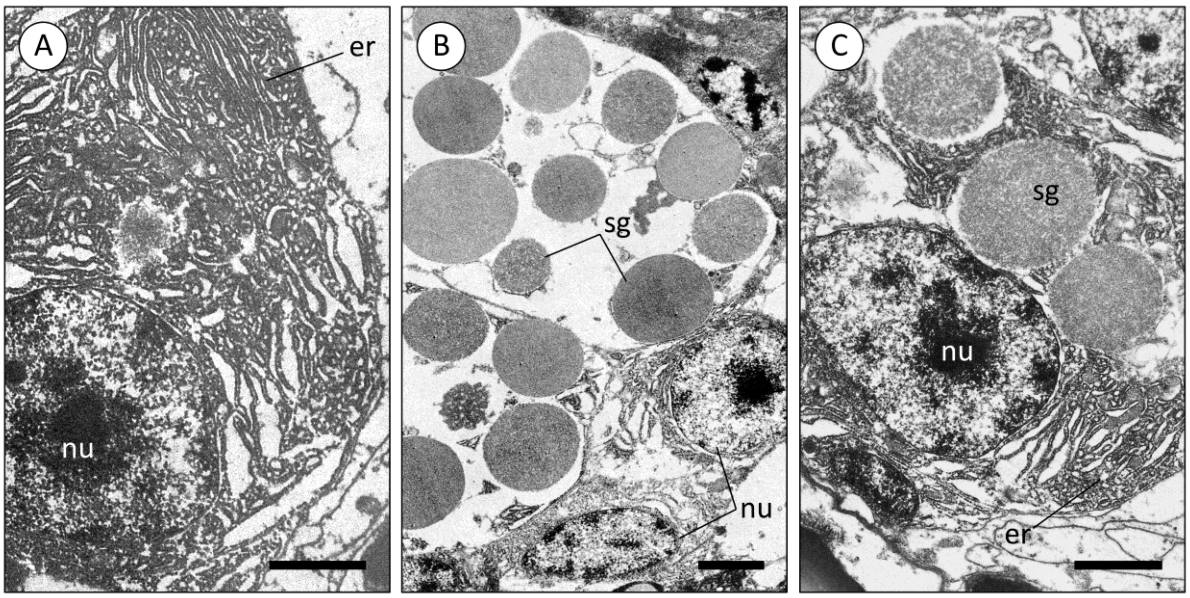


Figure 4 - Transmission electron microscopy images of the posterior salivary gland cells of *Sepia officinalis*. **(A)** Abundant rough endoplasmic reticulum of a glandular cell. **(B)** Secretory granules (protein-dense secretory vesicles) of a glandular cell in different maturation stages. **(C)** Secretory granules located near the endoplasmic reticulum. Abbreviations: er, endoplasmic reticulum; nu, nucleus; sg, secretory granules. Scale bars: **(A, C)** 2 μm; **(B)** 3 μm.

In comparison, the ASG of *Sepia officinalis* are also whitish structures formed by a pair of lobes like the PSG. However, ASG are embedded in the buccal mass, which is mostly comprised of the beak musculature, therefore being harder to localise and isolate. These glands are also microstructurally constant being formed chiefly by a network of branching glandular diverticula (blind-end 'tubules') interlinked by fibrous intertubular tissue and in this case, surrounded by muscular fibres (Figures 5A, B). The diverticula are also comprised of simple columnar secretory epithelium but histochemically distinct from PSG cells. The tubules of the ASG, which is often termed 'mucous gland' in coleoids, are mostly composed by serous and goblet (mucocytes) cells altogether expectedly responsible for the secretion of permeabilising enzymes and copious amounts of mucins, respectively (Figures 5C, D). Serous cells bare also an abundant endoplasmic reticulum and should be responsible for the biogenesis of secretory

granules, now dyed blueish-pink under Alcian Blue and PAS histochemical staining, respectively, revealing acidic and neutral sugars, respectively, presumably in glycoproteins like mucins (Figure 5C). In turn, goblet cells presented less abundance of endoplasmic reticulum and yielded a blue cytoplasm, following staining of mucus sacculi by Alcian Blue for acidic mucins (Figures 5D, E).

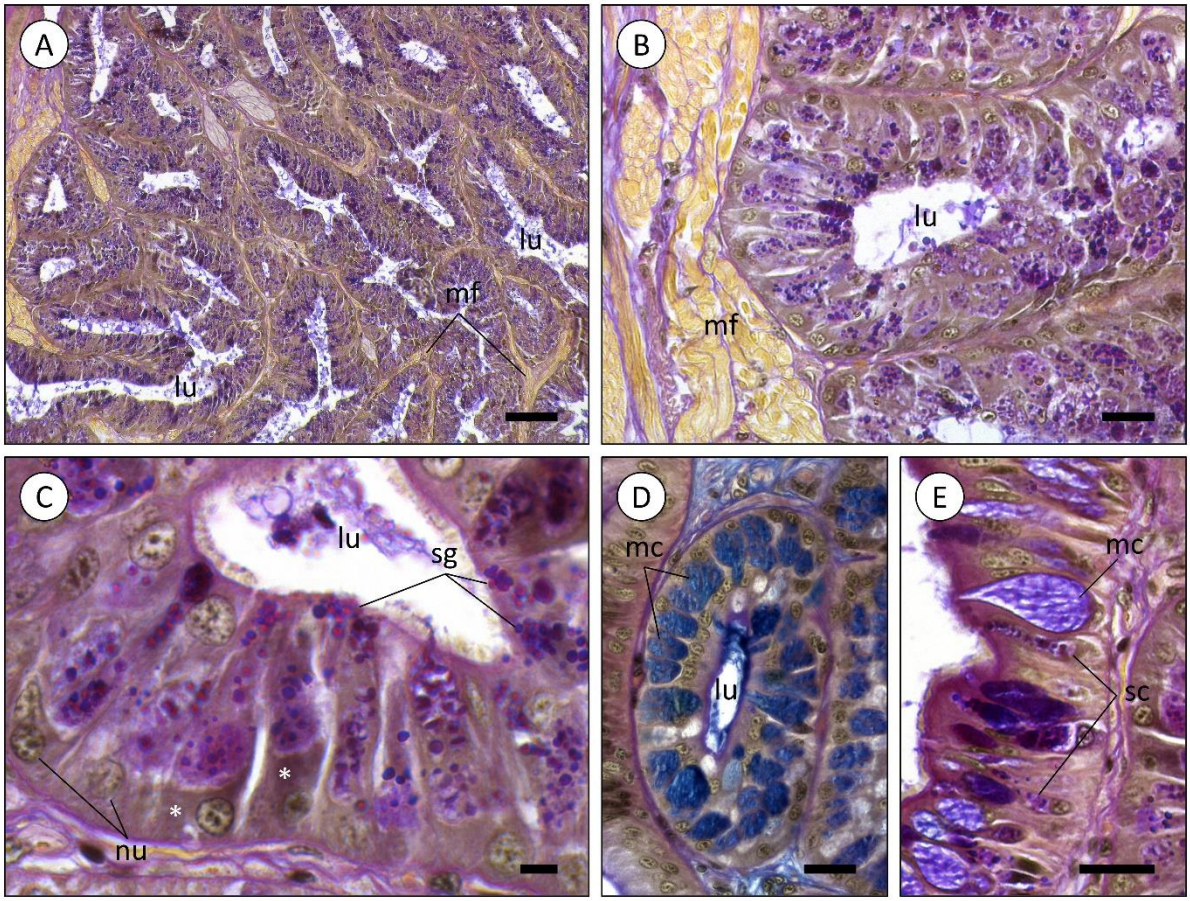


Figure 5 - Histochemistry of the anterior salivary glands of *Sepia officinalis* (tetrachrome stain combining Alcian Blue, PAS, Weigert's iron Hematoxylin and Picric Acid). **(A)** General and homogeneous appearance of anterior salivary gland diverticula ('tubules'). **(B)** A diverticulum of an anterior salivary gland. Comparatively to the posterior salivary glands, diverticula are chiefly composed of mucin-secreting secretory cells. Note the presence of muscular fibres (mf) near the tubules resulting from the glands being embedded in the buccal mass. **(C)** Detail of the secretory granules being produced by tubule cells. **(D)** Diverticulum composed by mucus-secreting cells (mc), stained positive for Alcian blue, revealing acidic mucins. Note mucus, in blue, concentrated at the lumen (lu) of the diverticulum. **(E)** Glandular epithelium bearing serous (sc) and mucocytes (mc), also named goblet cells. Abbreviations: nu, nucleus; asterisk, (rough) endoplasmic reticulum; sg, secretory granules. Scale bars: **(A)** 100 μm; **(B, D, E)** 25 μm; **(C)** 10 μm.

3.4.2 Transcriptome profiles of posterior salivary glands unveil several toxin-encoding transcripts

The assembled transcriptome (Figure 6) of the PSG of *Sepia officinalis* yielded a total of 171,318 transcripts, corresponding to predicted 124,209 genes. Of these, 27,753 resulted in potential open reading frames (ORFs), the shortest of which with 85 aa in length, from which 43.26% are full coding sequences. The initial homology-based search blast against Swiss-Prot (an expertly curated database) revealed about 19,157 annotated ORFs (Figure 6A). These were contrasted against subsets of either UniProtKB or Swiss-Prot databases to survey specifically for toxin-related proteins, enzyme-related proteins, proteins associated with *post*-translational modifications, secreted proteins or overall matches against eukaryote toxin or venom-related proteins (excluding humans), or other unreviewed proteins from Annelida, Cnidaria and Cephalopoda. This analysis resulted in a total of 16 158 annotated ORFs, which is equivalent to ~80% of ORFs annotated through Swiss-Prot. The results detailed in Figure 6B revealed that from those annotated ORFs, more than 1,200 ORFs had hits in all customised databases ('toxin', 'enzyme', 'secreted' and '*post*-translational modifications'). 'Enzyme' detained the most hits, with about 93% of annotated ORFs (~15,000), with ~5,900 exclusive matches. About 45% of all annotated ORFs (~ 7,300) were significantly matched to proteins with known *post*-translational modifications. The annotated ORFs matched against 'Secreted' and 'Toxin', however with relatively lower percentages (about 27% and 21%, respectively, of total annotated ORFs). Since the PSG of *Sepia officinalis* is hypothesized to hold specific functions in secreting toxins and related bioactives, we contrasted the findings against a toxin-oriented customised database termed 'SeaTox' and assessed how many annotated ORFs were in common with the preceding. The results (Figure 6C) showed that from ~1,500 annotated ORFs pertaining to the 'SeaTox' database, about 1,421 ORFs were detected in all databases, whereas 62 hits were exclusive to the 'SeaTox' database. The intersection of all databases revealed that cephalotoxins, cysteine-rich venom proteins and proteins with toxin activity or neurotoxins with metal ion binding activity yielded strongest homology matching, i.e., lowest *e*-value and highest percentage of identity (see chart in Figure 6C). The direct contrast with 'SeaTox' database also revealed high diversity of serine proteases (with 331 hits), followed by toxins, metalloproteases, chitinases, hyaluronidase and hormones (Figure 6D). Still, regardless of overall expression 'toxins' (inferred from annotation keywords) offered the highest number of hits (~750). *Sepia officinalis* transcripts encoded proteins that yielded a total of 203,292 domains, from which 19,873 were present in non-annotated predicted coding regions (5,275 ORFs). On

the other hand, a total of 1,459 putative proteins that were annotated against the 'SeaTox' database also yielded 24,745 conserved protein domains. Within this group of ORFs, domains mostly linked associated with toxins, such as, CAP, Insulin, Kazal, Kunitz_BPTI and ShK were highlighted (Figure 6E), as well as Ldl_recept_a and TSP_1, which are known to be present in cephalotoxins. Other domains (e.g., astacin, reprolysin, trypsin) connected with venoms components, like permeabilising and diffusing agents, were also detected. The expression of the annotated ORFs with match in all five databases is provided in Figure 6F.

Figure 6 - Global transcriptomics profiling of the posterior salivary glands of *Sepia officinalis*. **(A)** Refinement of the whole-transcriptome towards the selection of transcripts of interest. **(B)** Venn diagram showing the number of annotated open reading frames (ORFs) per comparison across four different customised databases ('toxin',

'enzyme', 'secreted' and '*post*-translational modification'). Overlapping regions reveal the number of ORFs with homology matching against two or more databases. **(C)** Venn diagram representing the number of annotated ORFs with homology matching between all the previously-mentioned customised databases and the 'SeaTox' database. The top 10 hits of the common putative proteins conjoining contrasting against all customised databases are listed in the appended chart. **(D)** Categories (inferred from annotation) of proteins of interest with significant match against the 'SeaTox' database. The numbers within figures represent indicate the number of ORFs matching each category, while the radius of circles is representative of expression levels (TPMs). Note that ORFs may be classified in more than one category. **(E)** Conserved protein domains of interest in ORFs annotated against 'SeaTox' database. The coloured circles represent domains most likely associated with toxins. **(F)** Expression levels and *e*-values retrieved from the BLAST analyses results for the annotated ORFs with match in all five databases, highlighting potential genes of interest within the venom secreted by all the annotated ORFs produced by *Sepia*: CRVP (Cysteine-rich venom protein), CTX (SE-cephalotoxin) and VIns (Venom Insulin) through expression levels and the quality of the annotation. The lowest *e*-values for the annotated ORF were represented within the homology-searches against all five customised databases.

From the mostly significantly matched transcripts, we selected three sequences of interest to perform validatory expression analyses by RT-qPCR, based on several factors such as relevance in animal venom, reviewed annotation, expression, availability of full coding sequence and even biotechnological potential. This short list thus includes transcripts that encode toxin-related proteins or venom components namely homologs for a cysteine-rich venom protein (CRVP) with 303 aa, a cephalotoxin (CTX) with 1,042 aa and a venom insulin (VIns) with 161 aa. Overall, comparison of the expression levels and BLAST *e*-values revealed distinct expression profiles these transcripts of interest, with CTX showing higher levels and similarity scores in comparison to the CRVP and VIns. Notably, the transcriptome profiles of the three independent biological replicates revealed high correlation between the annotated transcripts (Spearman's R ~0.74, $p < 0.05$), including for the relevant CRVP, CTX and VIns (Figures 7 and 8). Such findings reinforce the biological relevance and consistency of the toxin-related proteins and venom components that were tentatively identified. Nonetheless, homology-matching analysis against the 'SeaTox' database yielded only eight ORFs that were homologous to proteins and peptides produced by *Conus*. Specifically, seven ORFs had homology-matching against a venom insulin, Con-Ins Im1 (*e*-value ranging from 4.4×10^{-19} to 7.51×10^{-19}) from *Conus imperialis*, whereas the remaining ORF was homologous to a conopressin/conophysin, isoform 2 fragment (*e*-value = 3.82×10^{-28}) from *Conus monile*.

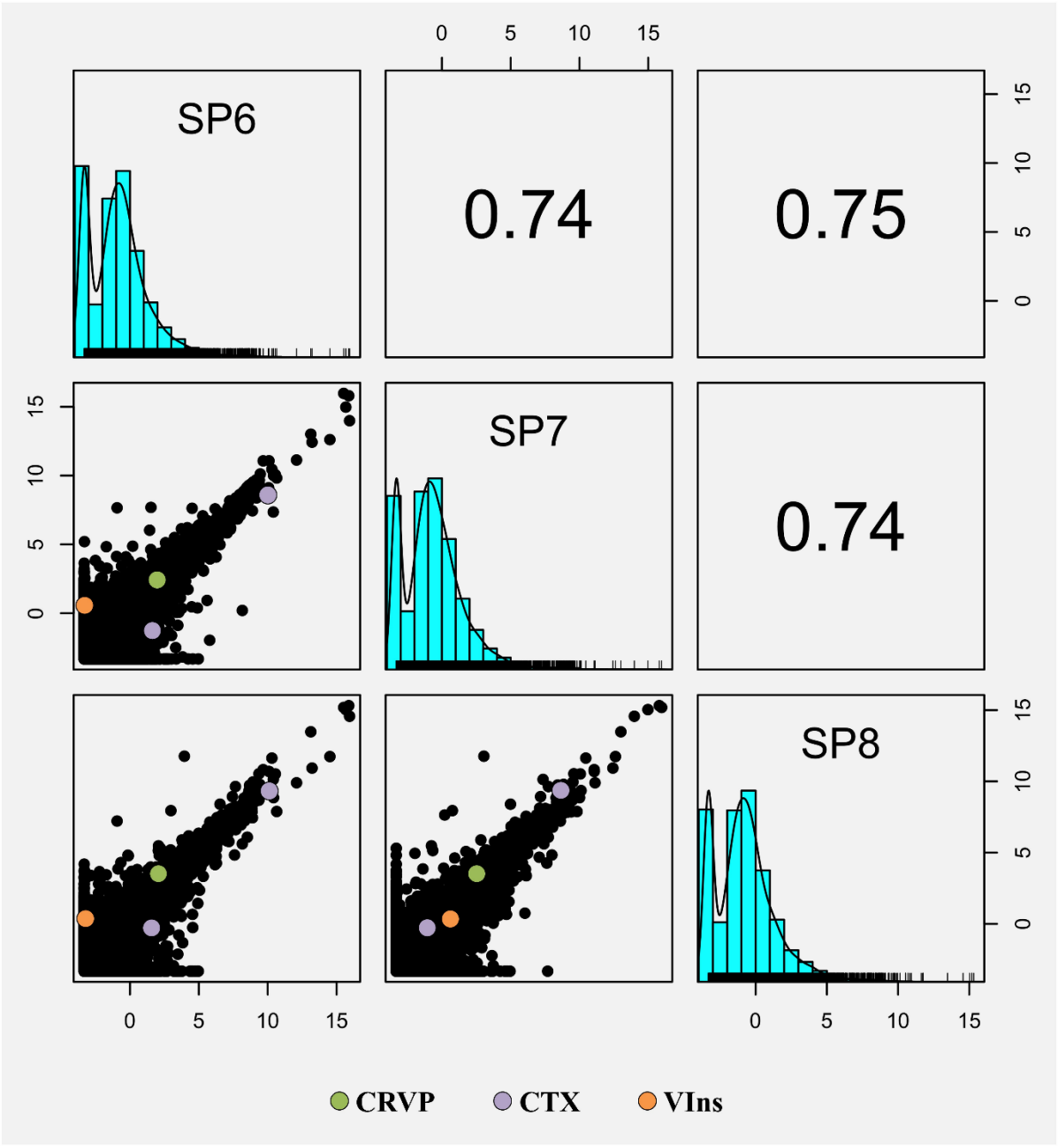


Figure 7 - Generalised pair plot showing all the correlations given by the pairwise comparisons between the three RNA-Seq samples (SP6, SP7 and SP8). Below the diagonal are presented all the pairwise scatter plots comparing all the RNA-Seq samples, whereas Spearman correlation statistic values are displayed above the diagonal. Each point of the scatter plot represents the expression of a given gene (as $\log_2$ TPM). The genes of interest were highlighted in all scatter plots: CRVP (Cysteine-rich venom protein), CTX (SE-cephalotoxin) and VIns (Venom Insulin).

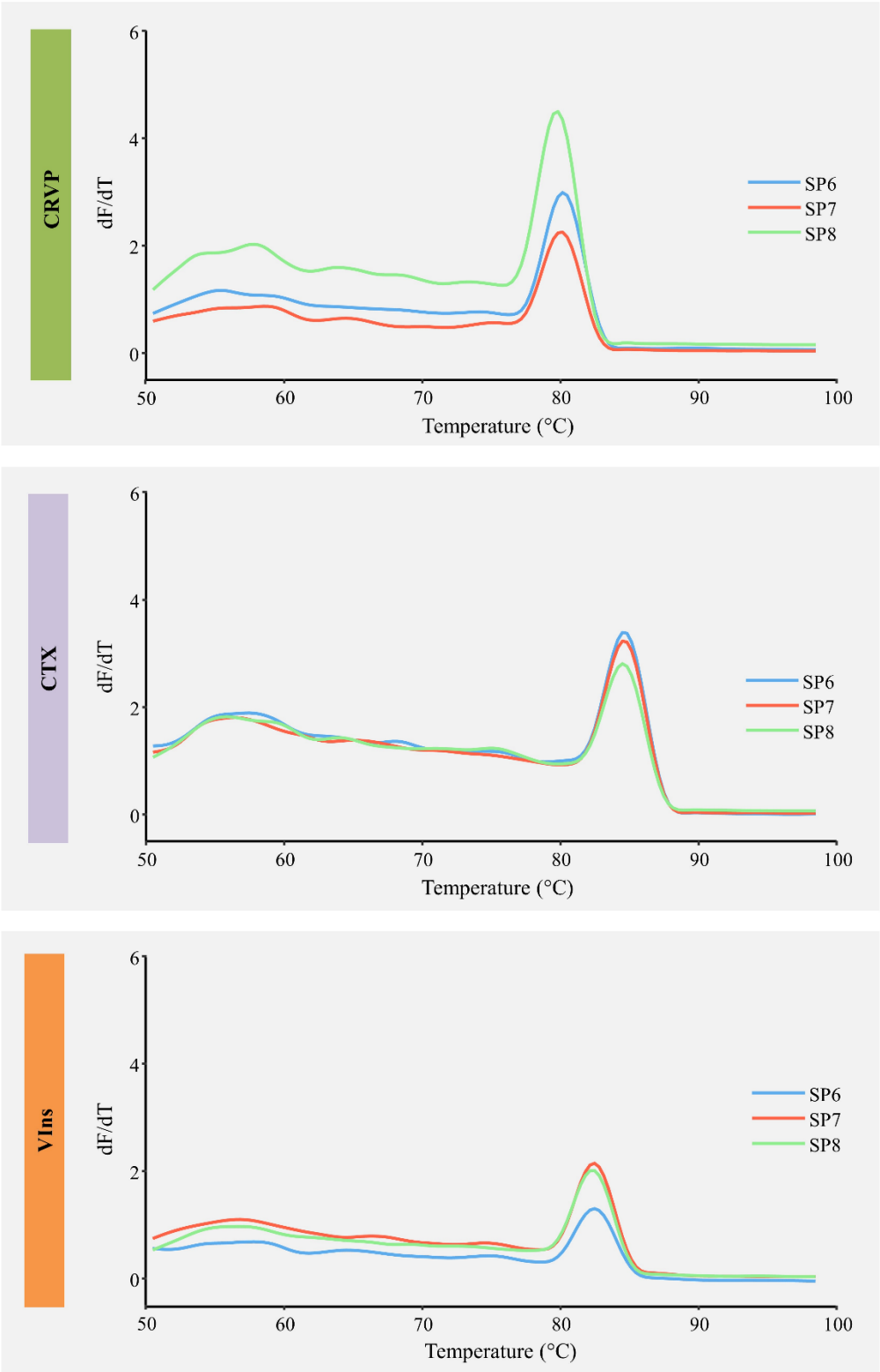


Figure 8 - Melting curve analyses of the three primer pairs used in RT-qPCR for three shortlisted toxin full coding mRNAs isolated from the posterior salivary glands of *Sepia officinalis*. CRVP, Cysteine-rich venom protein; CTX, SE-Cephalotoxin; VIns, venom insulin. Individual replicates are identified as SP6– SP8.

3.4.3 *Sepia officinalis* express toxin-encoding transcripts conserved across venomous taxa

Given the potential biological relevance of the toxin related transcripts CRVP, CTX and VIns (for details on these proteins see Ueda et al., 2008; De Meyts, 2016; Jiráček and Zďáková, 2019; Zhang et al., 2022), these were shortlisted for detailed analysis and validation. Predicting ORFs revealed that all bear different well-conserved domains (Figure 9A). The cysteine-rich venom protein is a calcium channel regulator involved in impairing smooth muscle contraction and possesses a CAP (catabolite activator protein) domain. The toxin CTX has two specific well-conserved domains, namely the Thrombospondin type 1 domain and the Low-density lipoprotein receptor domain class A, the later containing several disulphide-bonding cysteines. The venom insulin acts on prey by inducing a hypoglycaemic shock. This hormone presents the Insulin/IGF/Relaxin family domain that includes secreted regulatory hormones such as insulins, relaxins, insulin-like growth factor and bombyxin. Expression analyses by RT-qPCR involved consensus primer design for the two CTX transcript variants that were found in the assembled transcriptome. Assessing their expression by RT-qPCR confirmed their expression in the PSG (Figure 9B) and revealed comparative higher levels of expression of CTX (Tukey's HSD test, $p < 0.05$). The complete transcripts sequences were further corroborated by Sanger sequencing and then matched against their closest homologs amongst venomous or toxin-bearing animals from eumetazoan taxa namely Arthropoda, Mollusca and Cnidaria (Figure 9C). The phylogenetic trees showed closer similarity of target sequences to different species of cephalopods (both cuttlefishes and octopuses). The findings allowed assuming that CRVP is similar to a cysteine-rich secretory protein (CRISP), probably a highly similar homolog to *S. pharaonis*. In turn, the CTX sequence had also displayed higher similarity with SE-cephalotoxins from other cuttlefishes but unclear positioning with other molluscs. The VIns sequence retrieved from *S. officinalis* was found to pertain to the cephalopod clade, as expected. In the case of VIns, there is a clear distinction between the cephalopod and the gastropod clades, the latter represented by *Conus* and *Aplysia*. However, this distinction is not evident for the two other proteins. The CRVP sequence holds greater similarities to CRISPs from some cnidarians (such as the bryozoan *Bugula*) than to *Aplysia*. In turn, the CTX from *S. officinalis* presented high similarities with other *Sepia* (namely *S. esculenta* and *S. pharaonis*) but higher similarities to some cnidarians than to *Octopus bimaculoides*, which yielded the closest match not only within octopodids but also the remaining molluscs.

Homology-based structural inference on the three proteins CRVP, CTX, VIns yielded global model quality estimate (GMQE, which ranges between 0 and 1) of 0.83, 0.34, and 0.67, respectively (Figure 9D). The most robust model can thus be considered for CRVP (75.83% sequence identity coverage A0A812CAC2.1.A from *Sepia pharaonis*); followed by VIns (97.52% against A0A6B9RMK7.1.A from *Sepia latimanus*) and finally CTX (73.66% against R4G2C8.1.A from *Sepioteuthis australis*).

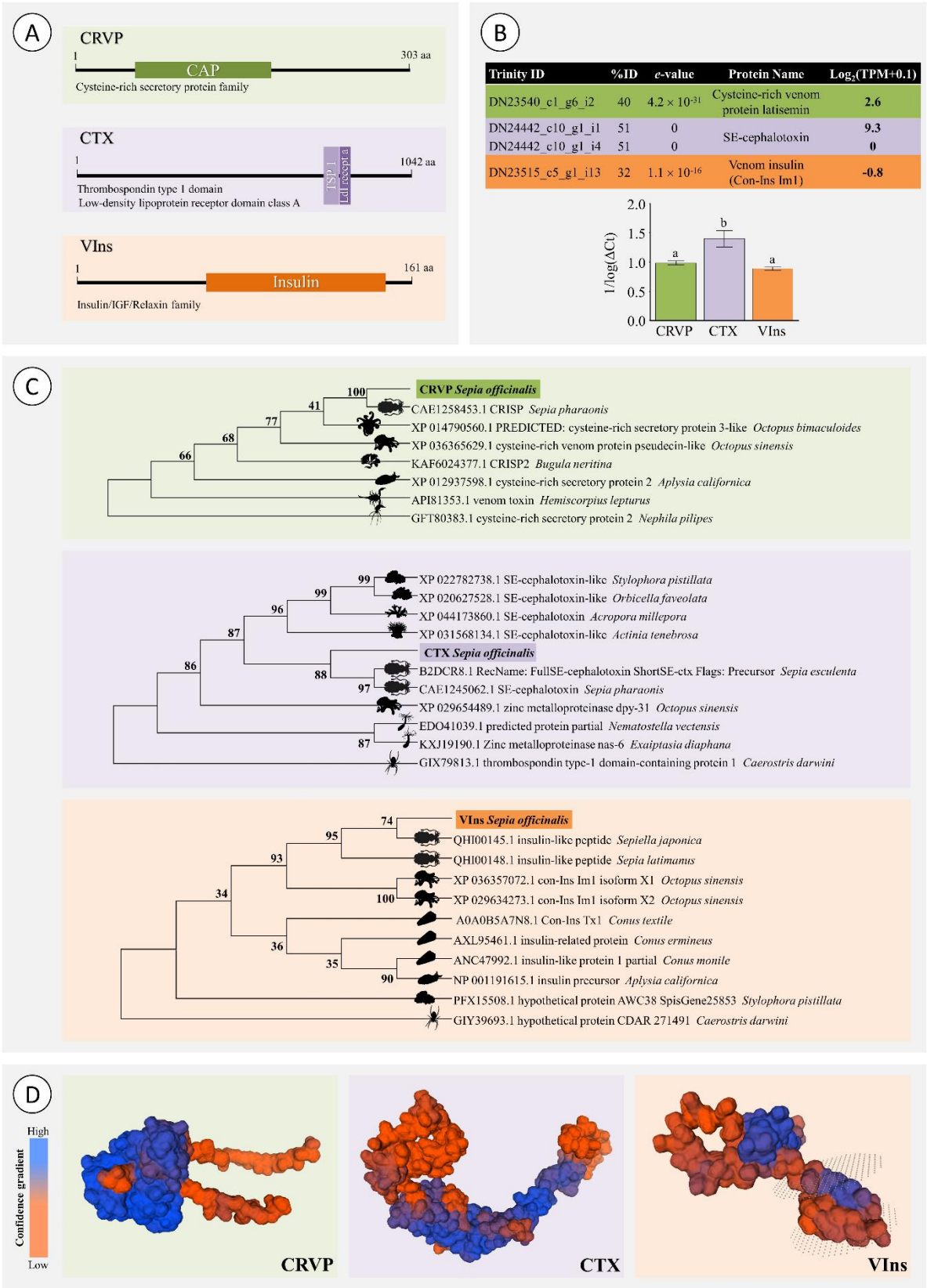


Figure 9 - Validation of the transcriptome of the posterior salivary glands of *Sepia officinalis* by RT-qPCR of representative transcripts. **(A)** Representation of conserved domain locations for each putative protein selected for

transcriptome validation: CRVP (Cysteine-rich venom protein), CTX (SE-cephalotoxin) and VIns (Venom Insulin). **(B)** Description of Expression analysis of selected transcripts (chart) by RT-qPCR (bar plot). Different letters indicate significant differences (Tukey's HSD test, $p < 0.05$). **(C)** Phylogenetic trees of CRVP, CTX and VIns. Each model includes the best matches from venomous or toxin-bearing animals for comparison. The phylogenetic models were produced using Maximum Likelihood and the Jones-Taylor-Thornton model, with 800 bootstrap replications. Bootstrap support values are given for all nodes. **(D)** Predicted homology (reference) -based models of the protein structure for each putative protein of interest: CRVP (Cysteine-rich venom protein), CTX (SE-cephalotoxin) and VIns (Venom Insulin). The latter was predicted to have a transmembrane domain (between dotted surfaces).

3.5 Discussion and Final considerations

The common cuttlefish has two salivary glands, each bearing two lobes. Whereas the ASG is embedded into the buccal mass, and is therefore of difficult access and isolation, the PSG is located near the digestive gland and can be easily isolated and excised. Here we confirmed the very distinct microanatomy between the anterior and posterior salivary glands The dissimilarities have significant implications for their function, even though they share the general diverticula-based, well-irrigated structure that characterises the annex glands of the cephalopod digestive tract (including the digestive gland), thus sharing the same endodermal origin (see Costa et al., 2014 and references therein). The ASG is chiefly involved in the secretion of saliva that constitute a mucin-rich vehicle for the bioactive compounds (mostly proteinaceous) secreted *en masse* by the PSG, which can be inferred from the dense packing of secretory cells and secretory granules within. These are expelled apically into the glandular tubules' lumen through a merocrine (via exocytosis) or most likely, apocrine secretion (involving loss of parts of the cytoplasm) processes (see Rodrigo et al., 2018), even though the exact process could not be determined in the present work. The transcriptomic analyses disclosed the secretion of a cocktail with toxins and matrix-directed proteinases such as metalloproteinases, chitinases and serine proteases that most likely act as permeabilising agents to assist diffusion of toxins. Chitinases, in particular, bare particular relevance for *Sepia*, which is a known predator of decapod crustaceans. Most importantly, we isolated the full coding sequence of a cysteine-rich venom protein, venom insulin and a cephalotoxin, breaking way for future endeavours in recombinant expression and bioactivity testing for potential biotechnological purposes.

From animals collected in 2012 at Normandy, France, Cornet et al. (2014) disclosed that the PSG of *Sepia officinalis* holds a dual role in both predation and immune defence based on

bioactivity assays with crabs and bacteria, respectively. Our findings are generally accordant with the identification of putative venom components inferred from the combination of proteomics and transcriptomics performed by these authors. However, the role of the PSG as an immune barrier, even though plausible, needs further investigation, as the vast majority of proteins that we found pertain to permeabilising enzymes and toxins and not anti-microbial agents like bacterial lipopolysaccharide-biding peptides described by Cornet et al. (2014). In addition, besides its role in lubrication of the gut, the secretion of mucus by goblet cells (which comprise most of the tubule epithelia of the ASG) in the gut and annex glands is *per se* a critically important part of the innate immune system of eumetazoans. In addition, mucus is often the vehicle of antimicrobial and antiviral agents and other defensive molecules against pathogens (recently reviewed by Sheng and Hasnain, 2022). Altogether, our results, i.e., the combination between microanatomy and transcriptomics, suggest a separation of roles between the two salivary glands. Whereas the ASG is mostly involved in defence, the PSG is involved in predation. Interestingly, the microanatomy of gastropods is not as revealing of differences between ASG and PSG, which may be due to far less organ individualisation in the Gastropoda than in the Cephalopoda. Still, our recent investigations on a predatory gastropod of the temperate intertidal that secretes cysteine-rich bioactives, likely toxins, *Nucella lapillus*, showed that its salivary gland blends goblet cell-rich tubules and tubules bearing acinar-like cells almost exclusively (D'Ambrosio et al., 2021), therefore resembling the cuttlefish's ASG and PSG tubules, respectively.

The differences between the present work and Cornet et al. (2014) are likely related to two factors: i) technical, as omics methods and associated databases and algorithms for annotation are constantly evolving, and ii) differences between the two populations of *S. officinalis*. Nonetheless, we highlight the presence of a CRVP and of CTX in common with the work by Cornet et al. (2014), which emphasizes the importance of these potential neurotoxins for the cuttlefish. In fact, high inter-populational variability of venom composition has already been recorded in land animals, with emphasis on snakes, which is considered a serious problem for the treatment of envenomation (e.g., Casewell et al., 2020; Rashmi et al., 2021). Interestingly, Avella et al. (2022) used proteomics to address this issue in a species of viper from Portugal, which further highlights the pertinence of omics methods to study various aspects of venom composition. Even though intra-specific variation of venom composition is little investigated in marine animals besides cone snails (e.g., Fassio et al., 2019; Pardos-Blas et al., 2019), Smith et al. (2023) investigated venom variation in sea anemones combining genomics,

transcriptomics and proteomics and revealed the importance of gene duplication events and strong adaptive pressure to be major factors. Altogether, whether interpopulational venom variability in *S. officinalis* results from different selective pressures or from phenotypic variation (such as that caused by feeding), it is a matter that deserves further investigation. In addition, it also adds to increasing the span of bioactives of interest in this species and marine invertebrates in general.

The existence of a cephalopod 'toxic saliva' that is capable of paralysing crabs is known since the work by Lo Bianco (1888) with octopus, but it was Ghiretti (1959) who first isolated a protein, hitherto named cephalotoxin, as the main causative agent, precisely from the PSG of *S. officinalis*. Since then, our understanding on the complex biochemistry of venoms has greatly refined. The term cysteine-rich venom proteins (CRVPs), which is occasionally used to identify cysteine-rich secretory proteins (CRISPs) in venoms and poisons, have been invariably found in the venoms of both land and marine animals. They are characterised by possessing CAP domains and high bioactivity. The CAP superfamily of proteins, among which CRISPs are one of multiple subfamilies (nine subfamilies are described in mammals alone), holds a wide span of known or predicted functions, from reproduction to immunity (Gibbs et al., 2008). Non-toxin CRISPs are commonly found in male testes of eumetazoans, mammals included, where they are believed to hold protective functions (see for instance the review by Gonzalez et al., 2021). In venoms, they may act as anti-microbials but also as neurotoxins, permeabilising agents and regulators of inflammation (see Tadokoro et al., 2020, for a review). The latter aspect can hold particular importance for venom delivery as proteins that can promote blood flow and prevent immune function and perhaps even clotting. Interestingly, we recently found CRISP signatures in poisonous (non-injected) secretions from marine annelids as well, even in cases where low toxin-specificity is expected (Rodrigo et al., 2021; Moutinho Cabral et al., 2022), which further strengthens the ubiquitous roles of these proteinaceous bioactives in animal venoms.

Venom insulin, in turn, has not been found by Cornet et al. (2014) in *S. officinalis*. However, this hormone-like toxin has been described in cone snails, which release it to the water presumably to disorient schools of small prey fish by eliciting fast hypoglycaemic shock (Safavi-Hemami et al., 2015). This recent discovery is leading to a renewed interest in hormone toxins by biotechnologists, especially because some insulin-like peptides from cone snails (Con-Ins) have been found to act much faster than human insulin (due to rapid activation without multiple cleaving, unlike native human insulin) and have high affinity to human

insulin receptors, which may have very important implications for the treatment of diabetes (De Meyts, 2016; Jiráček and Záková, 2019). The isolation of the full coding sequence for this protein can thus be a major achievement for the bioprospecting for novel peptidic bioactives in molluscs. However Swiss-Model retrieved no sufficient homology between any Con-Ins and the validated insulin-like peptide from *S. officinalis*, which indicates highly dissimilar structures. On the other hand, at this stage the structure of VIns is merely a prediction obtained against a different species of *Sepia*. Accordant with the data from structural predictions, phylogenetics (recall Figure 9) indicate that there can be two distinct clades of venom insulins based on sequence homology, one allocating cephalopods and other the gastropods, including *Conus* and *Aplysia*. The tree thus suggests that the full sequences of these three main toxins present a significant degree of conservation with orthologs from other Cephalopoda, albeit noticeable variation within taxa and that the properties of VIns from these animals, which lacks experimental verification, may be distinct from *Conus*. Interestingly, the trend to separate between the Cephalopoda and Gastropoda clades does not extend to the CRVP/CRISPs or CTXs. The diversity and evolution of CRISPs in venomous animals is better described in snakes and terrestrial arthropods (like scorpions) than in molluscs. Even though they are believed to be relatively well conserved, only a few of these CAP domain-bearing proteins have been fully characterised (see Tadokoro et al., 2020; Zhang et al., 2022). It has been hypothesised that the evolution of CAP proteins (which started in bacteria) has been a multi-step process that potentially involved positive selection through non-toxin CRISPs and potentially from non-toxin to toxin (Vicens and Treviño, 2018; Tadokoro et al., 2020; Zhang et al., 2022). This may explain considerable variability in sequence, structure and function of these proteins even with the same taxon. In turn, cephalotoxin-like toxins (and potentially non-toxin proteins) have recently been found in very distinct animals using omics methods, including cnidarians and even fish, even though not all forms are validated (Moya et al., 2012; Domínguez-Pérez et al., 2018; Silva et al., 2018). Still, the evolution and function of these relatively large proteins (c.a. 500–1,000 aa) is not yet understood, including ancestry and potential convergence. Interestingly, the validated cnidarian form (SwissProt accession B2DCR8) described by Moya et al. (2012) from the coral *Acropora millepora*, is not glycosylated and is devoid of the EGF (epidermal growth factor)-like domain present in *Sepia*, which is a disulphide bond-forming conserved domain common in venom proteins; but either form bear a peptide signal, which indicates they are actively secreted (see also Ueda et al., 2008).

To summarise, the relatively high number of overlapping hits against multiple databases (> 1,000) further sustains high biochemical diversity in the posterior salivary gland (PSG) of *Sepia officinalis*, including aspects that have been little surveyed in cephalopods by marine toxinologists such as toxins and other secreted proteins with *post*-translation modifications that might be needed to assure function and specificity. Since protein folding and *post*-translation modifications of animal proteins can be cumbersome when promoting heterologous expression using prokaryote (e.g., *E. coli*) and even yeast models, this information can be paramount to develop adequate scale-up methods for the production, harvesting and purification for recombinant toxins and other bioactives. The present work confirmed the suspected high diversity of bioactives in the PSG of *S. officinalis*, an organ which should perhaps more correctly be named 'venom gland', as the ASG is chiefly responsible for the secretion of saliva vehicle of bioactives. Even though their specific bioactivity needs investigating, the toxins and accompanying peptides and proteins, of which permeabilising agents such as metalloproteases and serine proteases are well-represented components that are likely able to affect a wide range of prey, especially crustaceans. From the annotated proteins or classes of proteins, we may highlight venom components with potential neurotoxic and endocrine-disrupting effects. The results also suggest, even though this issue needs further enhancement, that can be considerable variability in venom composition between various populations of the common cuttlefish, as it is now known to occur in other venomous animals. However, *Sepia* venom proteins differ considerably from potential homologs in other molluscan taxa, especially the well-known *Conus* gastropods. Finally, it must be emphasised that, even in face of reduced genomic resources, modern omics can successfully be employed to screen for novel bioactives and produced validated results, here at least in the form for three full coding sequences for protein that bear biotechnological potential, a cysteine-rich venom protein (CRVP), a venom insulin (VIns) and a cephalotoxin (CTX).

4

SALIVARY GLAND FUNCTION, COMPLEXITY AND VENOM SECRETION IN CEPHALOPODS: THE *OCTOPUS VULGARIS* CASE STUDY AND A COMPARISON WITH *SEPIA OFFICINALIS*

4.1 Abstract

Marine organisms, especially those venomous species, are gaining momentum for biotechnology and drug discovery due to the ocean's vast span of novel and undiscovered bioactives. Among these, cephalopods such as octopuses and cuttlefishes have long been identified for producing a toxic saliva that remains, however, poorly characterised despite the isolation of a cephalotoxin, a coleoid-specific neurotoxin from the cuttlefish *Sepia esculenta*. The current work combined microscopy techniques with *de novo* transcriptome assembly to characterise the salivary glands of the common octopus (*Octopus vulgaris*) from the Portuguese West coast, searching for toxins and other bioactive proteins, with potential value for biotechnology and biopharma. Microanatomy and transcriptomics profiling suggest that the anterior salivary gland produces mainly mucus and bioactives involved in the upkeeping of protein function and activity whereas the posterior salivary gland produces proteinaceous toxins, in either species. Nonetheless, the cuttlefish, yielded a higher number of toxins even though its posterior salivary gland is anatomically and histochemically simpler than its octopodid counterpart. Altogether, the octopus venom comprises a sizable span of toxins, from neurotoxins to pro-insulins, immunomodulators and permeabilising agents, especially matrix-directed enzymes. We isolated and validated the full coding sequences of four important toxins: two cysteine-rich venom proteins, a chitinase and a hyaluronidase. Phylogenetic analyses unravelled higher similarity of both chitinase and hyaluronidase between octopodid homologs, whereas the toxins are closer to other cysteine-bearing proteins from cuttlefishes. Despite the presence of multiple bioactives in the salivary gland extracts of both *O. vulgaris* and *S. officinalis*, the toxicological experiments with crude venom revealed no significant impairment in viability of mollusc gill cells, which indicates the need for future testing of isolated compounds in non-molluscan, therefore, potentially less sensitive, models.

4.2 Introduction

Marine invertebrates are a most promising group of animals for the bioprospecting for novel bioactives, including those from venoms and poisons. Among these organisms we find the Cephalopoda, i.e., octopuses, cuttlefishes and squids, that albeit little explored by biotechnologists, are known to secrete a 'venomous' saliva for more than a century, even though the first true attempts to characterise their venoms only began in 1950s and only recently the first coding sequence of a cephalotoxin was unravelled (Ueda et al., 2008). Although the blue ringed octopus (*Hapalochlaena* sp.) is probably one of the most important targets of marine toxinologists due to their potent (and lethal) neurotoxic venom (Savage and Howden, 1977; Sheumack et al., 1978), temperate species such as the common octopus (*Octopus vulgaris*) and the common cuttlefish (*Sepia officinalis*) can offer potentially milder venoms yet bearing a broad range of novel and safer bioactives for biopharma and other potential applications. Through distinct omics studies in *Octopus vulgaris*, several candidate proteins and peptides for biotechnological applications were already identified, namely molecules that might have antimicrobial, antitumoral and antiviral activities (among others), even though these have been found chiefly in the mucus produced by the skin and also in the ink, which is a melanin-pigmented mucus-based mixture used to deceive and escape attackers (Imran et al., 2023; Pérez-Polo et al., 2023). In addition, bioactives from different species of cuttlefishes have already been studied for novel antitumor therapeutics. The ink extract of the *Sepia officinalis*, for instance, showed promising analgesic and anti-nociceptive properties in mice and also cytotoxic activities against cancer cell lines (Fahmy and Soliman, 2013). Furthermore, rats fed with a diet supplemented with the 'liver' (digestive gland) oil of *Sepia pharaonis* showed the ability to enhance immune functions and inhibit inflammatory responses (Joseph et al., 2005).

Similarly to the ink, which has dedicated organs for its production and storage (funnel and sac, respectively), the toxins within the venoms of cephalopods are associated with the salivary glands. Indeed, cephalopods hold two twin-lobed salivary glands, anterior and posterior, which are interconnected by a duct. The anterior salivary gland is also connected to the buccal mass whereas the posterior salivary gland is connected to the digestive tract (Guerra, 2019). Although there is evidence that toxins are mainly produced by the posterior salivary gland, which in commonly referred to as the 'venom gland', studies on the identification and description of key anatomical structures that intervene on the production of toxins, permeabilising agents and their mucin-based vehicle (the 'saliva') are scarce.

During the past decade there has been a growing interest in the biochemistry of toxin mixtures from cephalopod salivary glands due to their biotechnological potential and recent studies using high-throughput transcriptomics and proteomics approaches are enhancing their biochemical characterisation (Ménez et a., 2006; Whitelaw et al., 2016). Nevertheless, the biochemical diversity of toxins secreted by these glands remains largely unexplored. With this, we hypothesise that the two salivary glands bear different microanatomical aspects that reflect different functions, secretions and therefore biochemical diversity. We further hypothesise that the two species can hold significant differences that are related to their venoms and ecology and therefore which traits can be indicative of different types of bioactives. Bear this in mind, the current work aims at contributing to the study of the cephalopod 'venom gland' and its potential bioactives by combining the biotechnological perspective with microanatomy and 'omics' as a means to compare the two species and infer on clues to their putative toxins, their diversity and their potential targets and effects.

4.3 Materials and Methods

4.3.1 Animal collection

Eight *Octopus vulgaris* (mantle length 8.6 ± 2.9 cm; weight 334.9 ± 204.9 g) were collected from two beaches of the Portuguese West coast: Praia do Norte (38°39′0.153″ N, 9°14′43.876″W) and Praia da Foz (38°27′56.562″ N, 9°12′2.308″ W). After euthanasia by sectioning the central brain, the tissues of interest were harvested and processed immediately. Both PSG and ASG were extracted and divided for histological analyses and RNA extraction. The PSG of *Sepia officinalis* used for between-species comparative transcriptomics was retrieved from previous work and followed the same sample preparation (see Gonçalves et al. (2024), for details).

4.3.2 Histology

Samples preparation and all procedures and equipment for histological analyses followed the previous work (see Gonçalves et al. (2024), for further details and complete procedures). In brief, the ASG and PSG of *Octopus vulgaris* were used for both optical and transmission electron microscopies. For optical microscopy, samples were fixed using Davidson's and Zenker's fixatives. Paraffin histological sections were stained with Haematoxylin and Eosin and with a tetrachrome staining technique (for further details Costa and Costa (2012)).

Samples for transmission electron microscopy (TEM) were fixed in Glutaraldehyde 2.5% v/v (in 0.1 M sodium cacodylate buffer, pH 7.4), post-fixed in 1% m/v osmium tetroxide in 0.1 M sodium cacodylate buffer and embedded in Epoxy resin. Before the TEM analysis, resin sections were obtained and stained with toluidine blue.

4.3.3 Total RNA extraction and RNA-Seq

Preparation of samples, RNA extraction and RNA-Seq procedures, all reagents, kits and equipment were described in previous work (see Gonçalves et al. (2024) for further details). In brief, total RNA was extracted from ≈16 mg of tissue per sample (n = 3 PSG and n = 3 ASG, from three different individuals). The RNA integrity number (RIN) was found to range between 3.4 and 5.8 and was considered adequate for marine invertebrates. After library construction, the generated DNA fragments were sequenced with a mean length of 150 bp paired-end reads with 40 M reads of depth (Table B.1 in Appendix to Chapter 4).

4.3.4 Transcriptomic analyses

The transcriptome assembly, quality control checking and functional annotation procedures followed the previous work described in Gonçalves et al. (2024). Briefly, the raw data obtained from RNA-Seq was also pre-evaluated through FastQC v0.11.9 and TrimGalore v0.6.6., whereas less than 0.6% of reads were removed per sample, as they were considered as low-quality reads (i.e., reads under 20 bp of length). Transcriptome assembly was performed using Trinity v2.6.6, where a good quality of transcriptome was obtained with contigs N50, Ex90N50 and read content statistics, using TrinityStats, Bowtie2 v2.4.1 and Samtools v1.11 (Langmead and Salzberg, 2012; Haas et al., 2013; Langmead et al., 2019; Danecek et al., 2021). Downstream analyses were performed using R 4.*x* (Ihaka and Gentleman, 1996). Differentially expressed transcripts between the two salivary glands (posterior and anterior) were identified using Kallisto v0.43.0, with $|\log_2 FC| > 1.5$ and False Discovery Rate (FDR)-adjusted $p < 0.05$ as thresholds (Bray et al., 2016). The predicted open reading frames (ORFs) were annotated by scanning for homology against Swiss-Prot (version 18–11–2020 13:41, downloaded from https://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz) and SeaTox database (https://doi.org/10.5281/zenodo.14355541). The SeaTox database is mostly a customised subset of UniProtKB (UniProt release 2024_05), containing manually annotated toxin- or venom-related proteins from Eukaryota (excluding human), plus unreviewed proteins from Annelida, Cnidaria and Cephalopoda. The homology search was performed using BLASTP from NCBI

blast+ v2.10.0+, with a *e*-value cut off of 1×10^{-5} (Altschul et al., 1990; Camacho et al., 2008; UniProt Consortium, 2021). Homology-matching transcripts were functionally annotated through the R package UniproR v2.4.0, including the terms for the Gene Ontology (GO) analyses. The GO enrichment analyses were performed using R and the significance criterion for the Fisher's test was set at FDR-adjusted $p < 0.05$. The R script of transcriptomics analyses can be found in B.2 Appendix to Chapter 4.

4.3.5 Transcriptome validation by RT-qPCR

The RNA-Seq results were validated by reverse transcription quantitative polymerase chain reaction (RT-qPCR) of selected representative transcripts that encode for proteins with toxin or hormone activities or interfere with ion channels. The selected transcripts from RNA-Seq data includes two cysteine-rich venom proteins (CRVP and TX31), a chitinase (CHI) and a hyaluronidase (HYAL). The transcriptome validation followed the same procedures as described in previous work (Gonçalves et al., 2024). In brief, the full coding sequence of the selected transcripts were firstly isolated and amplified by PCR. Primers (Table 3) were designed and evaluated *in silico* using PCR Primer Stats (Stothard, 2000) and OligoAnalyzer (https://eu.idtdna.com/calc/analyzer). Specific primers were designed to quantify expressed sequence tags (ESTs) in the coding region by RT-qPCR, avoiding conserved domains. The 18S and β-actin genes were chosen as housekeeping genes. The primers for 18S gene were designed based on a sequence from *O. vulgaris* deposit in GenBank with the accession FJ617439.1. The β-actin gene primers were retrieved from Sirakov et al. (2009). PCR products were Sanger-sequenced for sequence confirmation. The RT-qPCR program comprised an initial denaturation procedure (95°C, 10 min), followed by 40 to 55 cycles of denaturation (94°C, 45 s), annealing (temperature ranged from 50°C to 54°C, 25 s) and extension (72°C, 30 s). Melting curve analyses for the expression tag of the four toxin full coding sequences can be found in Figure B.3 in Appendix to Chapter 4. Expression analysis was determined based on the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). Homoscedasticity and normality of data were assessed through Levene's and Shapiro-Wilk's tests, respectively. Differences in gene expression between the two organs were evaluated through the parametric Student's *t*-test. A significance level of $\alpha = 0.05$ was set for all analyses. Statistics were computed using R 4.*x* (Ihaka and Gentleman, 1996).

Table 3 - Primers' sequences used for PCR and RT-qPCR. Gene column indicates the housekeeping genes used for normalisation (18S and β-actin) and the genes encoding for the proteins selected for transcriptome validation:

chitinase (CHI), cysteine-rich venom protein latisemin (CRVP), hyaluronidase (HYAL) and cysteine-rich venom protein (TX31). The β-actin primers were retrieved from Sirakov et al. (2009).

Gene	Primer	PCR/RT-qPCR	Primer sequence (5´-3´)	Amplicon size
18S	Forward	RT-qPCR	GTTCCGACCGTAAACG	140 bp
	Reverse	RT-qPCR	CCTTCCGTCAATTCCT	
β-actin	Forward	RT-qPCR	TCCAGGCTGTGTTGTCTCTG	148 bp
	Reverse	RT-qPCR	AGATCACGACCAGCCAAGTC	
CHI	Forward	PCR	CACTTCGCAATGACAATG	1362 bp
	Reverse	PCR	GCCTTACCCACACAACT	
CRVP	Forward	PCR	CCAATGTTCCATTCGTATG	876 bp
	Reverse	PCR	CAAATTACTCACTTGGACAC	
HYAL	Forward_1	PCR	GCAAGGACAGGATTGGGAA	706 bp
	Reverse_1	PCR	TAACCAGCCAAGTCTATCGT	
	Forward_2	PCR	TGATGTCGTCTACTTTAGAGC	817 bp
	Reverse_2	PCR/RT-qPCR	CTCTCAATGATAAGGAAATCCC	
TX31	Forward_1	PCR/RT-qPCR	GTAAATGTGTGCTGGGAATGT	696 bp
	Reverse_1	PCR	CCCATTTCTGAGCTACATA	
	Forward_2	PCR	TGTCAATCCAGCAGCCT	840 bp
	Reverse_2	PCR	CGTGTAGTCTGTGATTCTGTT	
	Forward_3	PCR	GCAGATACTACATTGCTAAGTG	899 bp
	Reverse_3	PCR	GTTGACAATGTGGTCCCT	
CHI	Forward	RT-qPCR	TGTCAAGGCGGAAAATATCC	149 bp
	Reverse	RT-qPCR	GGTTACGACACCCAGTCT	
CRVP	Forward	RT-qPCR	CGTATGTCCTGAGATGTG	174 bp
	Reverse	RT-qPCR	TCTTTGTCTGTCATACCAG	
HYAL	Forward	RT-qPCR	GTGAACAACGCAACTGAAT	155 bp
TX31	Reverse	RT-qPCR	GCTGATTCTTCCAATTCTGCTT	204 bp

4.3.6 Phylogenetic analyses

The phylogenetic analyses were performed in four full coding sequences of the shortlisted transcripts, i.e., two cysteine-rich venom proteins (CRVP and TX31), a chitinase (CHI) and a hyaluronidase (HYAL) and followed previous work procedure (see Gonçalves et al. (2024) for details). Each sequence was scanned for homology against NCBI RefSeq_protein database, however with restrictions to related taxonomic groups as 'Lophotrochozoa', phyla 'Chordata',

'Arthropoda', 'Mollusca' and 'Cnidaria' as well as the specific classes: 'Cephalopoda', 'Arachnida', 'Gastropoda'. The most significant hits were chosen based on the best matches against venomous or toxin-bearing animals, *e*-value and percentage of identity retrieved from running BLASTP. Sequence alignment and dendrograms were made with MEGA X (Kumar et al., 2018). Phylogenetic consensus trees were produced for each selected sequence using the Maximum Likelihood method and Jones-Taylor-Thornton model, with 800 bootstrap replications (Jones et al., 1992). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (800 replicates) are shown next to the branches (Felsenstein, 1985). The trees were rooted on related sequences from Arthropoda or Chordata.

4.3.7 Toxicological analyses

Toxicological assays were performed by exposure mussel gill cells to different concentrations of the salivary glands' crude protein extracts (ASG and PSG of *O. vulgaris* and PSG of *S. officinalis*), following the neutral red uptake assay. In brief, *Mytilus* sp. (length: 4.62 ± 0.31 cm; width: 2.23 ± 0.19 cm) were randomly collected from the Portuguese west coast (38°38'44.943" N, 9°14'36.993" W), in February 2023. All individuals were transported to the laboratory and maintained in controlled conditions (17°C, 30 ‰ salinity, 10-hour photoperiod) until experimental procedure. Gills dissection and subsequent experimental procedures were performed in a sterilised environment. A cell viability assessment was conducted prior to the experiments using 0.4% trypan blue in Dulbecco's phosphate-buffered saline (PBS), pH 7.4. The toxicological experiments were run in a total volume of reaction of 100 µL per well, whereas gill cells at density of 1×10^6 cells/mL were exposed to different concentrations of crude protein extracts for 10 minutes at 18°C, in dark. Control wells (wells with cell suspension in PBS), background wells (wells with only PBS) and positive control wells (wells with cell suspension in PBS treated 1:1 with 30% hydrogen peroxide) were also tested. Neutral red staining protocol was adapted from the Assay Genie's Neutral Red Cell Cytotoxicity Assay Kit (Assay Genie, Dublin, Ireland) for non-adherent cells and from Repetto et al. (2008). Absorbance was read at 540 nm on the Multiskan SkyHigh Microplate Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Neutral red uptake assay was measured as percentage of cell viability. Homoscedasticity and normality of data were assessed through Levene's and Shapiro-Wilk's tests, respectively. Differences between concentrations for each individual crude protein extract were evaluated through the parametric One-Way ANOVA F

test and non-parametric Kruskal-Wallis H test. A significance level of $\alpha = 0.05$ was set for all analyses. Statistics were computed using R 4.*x* (Ihaka and Gentleman, 1996).

4.4 Results

4.4.1 Anterior and posterior salivary glands possess different types of epithelial cells

The salivary glands of the *Octopus vulgaris* are divided into anterior and posterior according to their position along the digestive tract and are both bilobed (Figure 10). Behind the buccal mass, which lodges the mandibular muscles and the beak (a chitinous and articulated structure divided in dorsal and ventral mandibles) are located the two lobes of the anterior salivary glands (ASG). These paired glands are connected by a long secretory duct to the posterior salivary glands (PSG), also composed of two lobes interlinked by short ducts. Both anterior and posterior salivary glands are formed blind-ended branched tubules (diverticula) that bear a simple (single-cell thick) secretory epithelium, however with significant differences between them (Figure 10). The anterior salivary gland, mostly known as the cephalopod mucus gland, is composed of densely-packed tubules connected by a thin layer of vascular-connective tissue (Figures 10 and 11, particularly Figure 11B). The general appearance of this gland is quite homogeneous when stained with Haematoxylin and Eosin (Figure 10). However, two distinct tubules could be distinguished when applying a tetrachrome stain (Figure 11A): (i) tubules rich in mucopolysaccharides (which acquired a dark-purple colour provided by PAS) and (ii) tubules that stained lightly yellowish (colour given by picric acid binding to basic-neutral proteins). Both tubules are formed by goblet (mucus-secreting) and serous cells with identical histochemical signature (PAS-positive), both arranged in simple columnar epithelia. Albeit less common, blueish mucocytes (Alcian Blue positive) were easily detectable in the lightly yellowish tubules (Figure 11C). Serous cells (responsible for the secretion of proteinaceous substances) are characterised by having a brownish colour basally given by the dense rough endoplasmic reticulum (from Weigert's Haematoxylin (Figures 11D and 11F)). Here, the proteins are produced and stored in vesicles (secretory granules), easily observed in TEM analysis (Figures 11E and 11F).

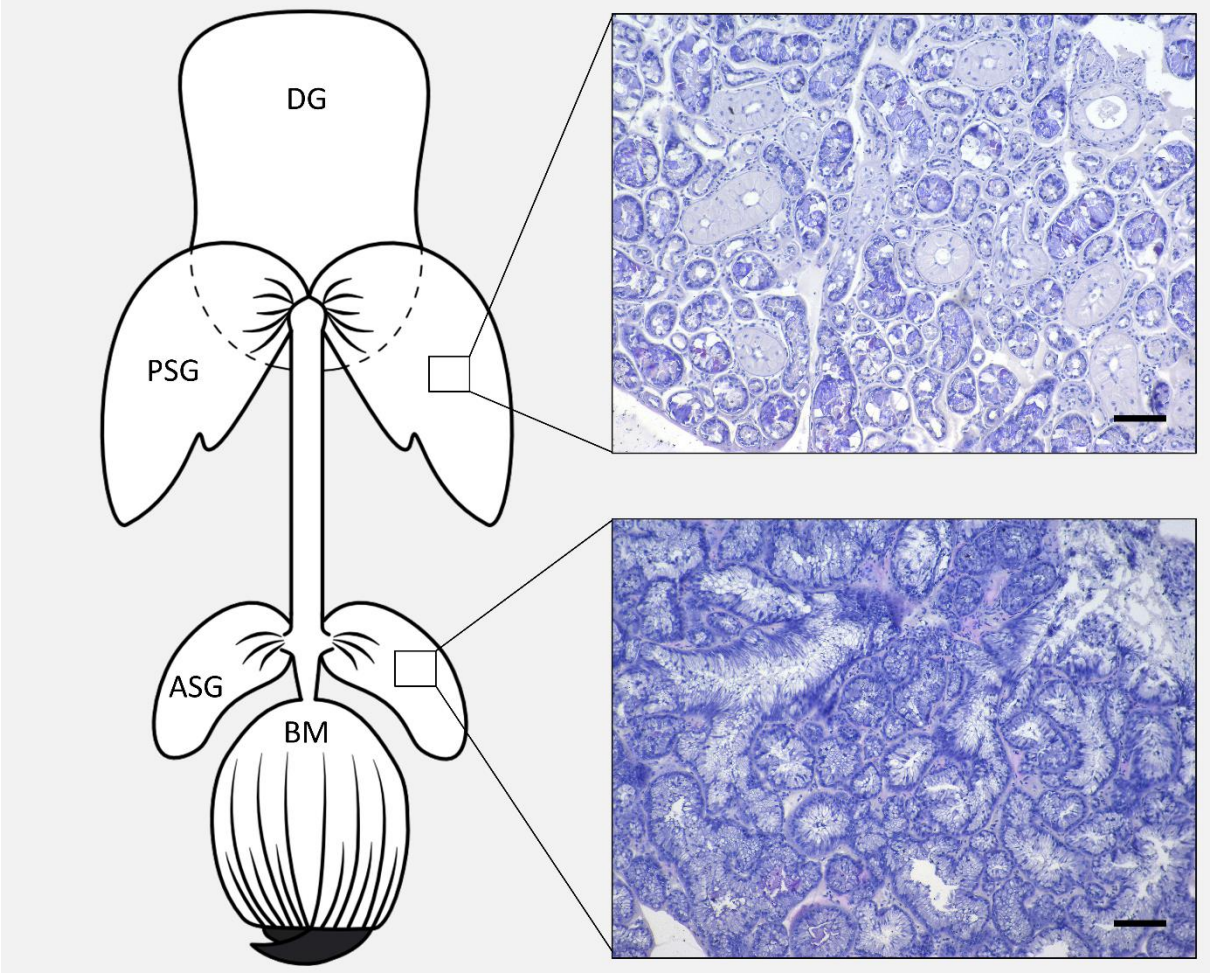


Figure 10 – Schematic representation of the bilobed salivary glands of *Octopus vulgaris*, which area joined by the main salivary duct. Photomicrographs of the anterior and posterior salivary glands show the tubular appearance of both glands stained with Hematoxylin and Eosin. Note the presence of two different epithelia in the posterior salivary gland in comparison of a single epithelium of the anterior salivary gland's tubules. Abbreviations: DG, digestive gland; PSG, posterior salivary gland, ASG, anterior salivary gland; BM, buccal mass. Scale bars: 100 μm.

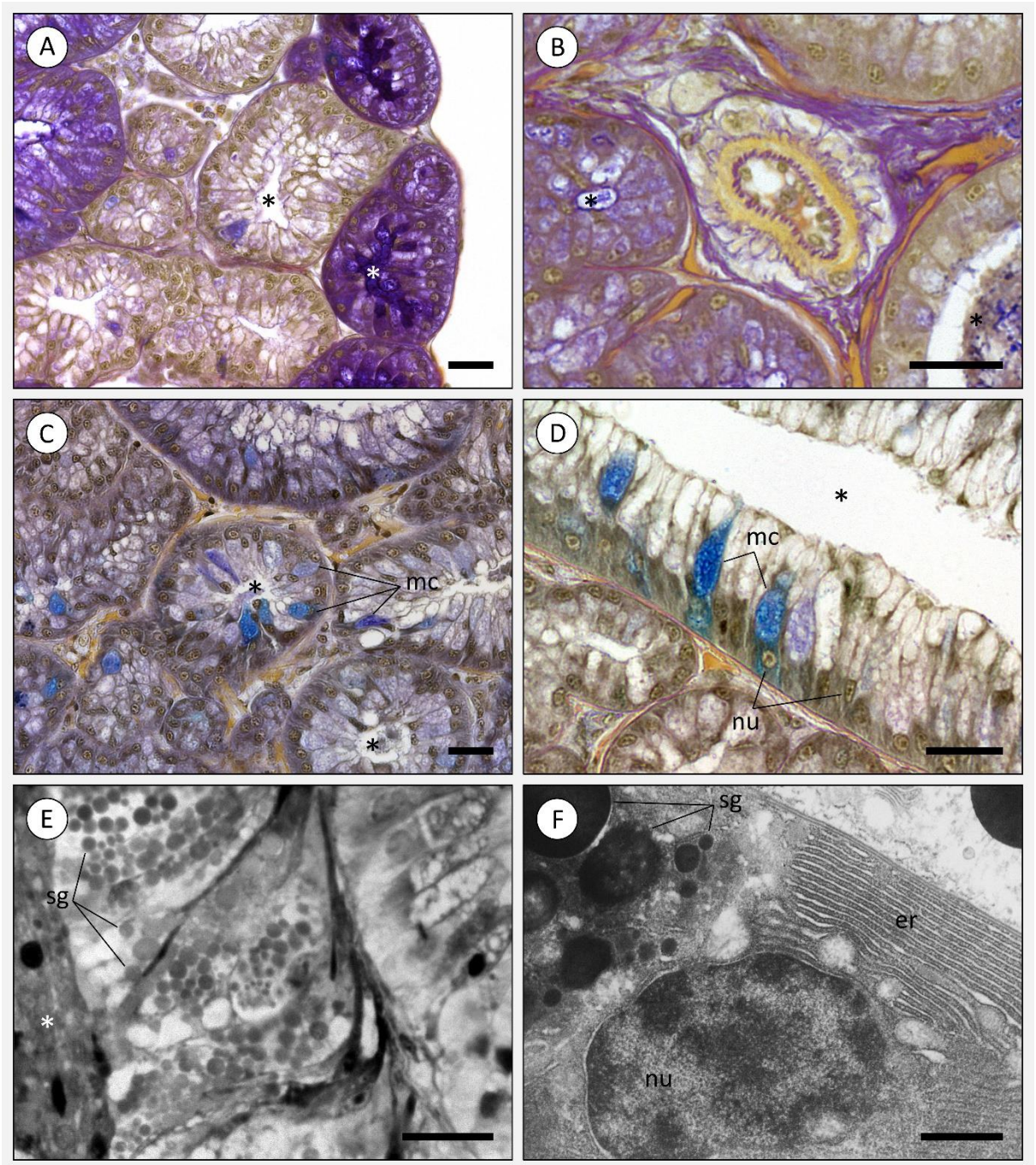


Figure 11 - Photomicrographs of the anterior salivary gland of *Octopus vulgaris* (optical microscopy and transmission electron microscopy). **(A)** Tubular appearance of the anterior salivary gland using a tetrachrome stain (TC). Presence of tubules with two different histochemical signatures: (i) PAS-positive tubules that acquire dark-purple colour and (ii) lightly yellowish tubules. **(B)** Detail of a blood vessel (arteriole-like) compressed between the tubules. **(C)** Homogeneous network of tubules composed of a single-layer columnar epithelium, formed by blueish mucus-secreting cells (mc) and serous cells. **(D)** Detail of a few Alcian Blue-positive mucus-secreting cells (mc) between the serous cells. **(E)** Resin section of an anterior salivary gland tubule representing the columnar cells filled apically with secretory granules (sg), which will be expelled to the lumen (white asterisk). **(F)** Abundant rough endoplasmic reticulum (er) of a glandular cell (transmission electron microscopy). Note the formation and

maturation of the secretory granules close to the nucleus (nu). The black asterisk indicates the lumen of tubules diverticula. Scale bars: **(A – E)** 25 μm; **(F)** 1 μm.

On the other hand, the posterior salivary glands are formed by highly vascular connective tissue within an extensive network of two types of densely stowed tubules whose epithelia differ more considerably comparing to the anterior gland (Figures 12A and 12B). The most abundant tubules present a secretory epithelium with strongly acidophilic mucus-secreting cells (with Alcian Blue-positive mucus sacculi) and serous cells with vesicle content that vary from purple-red to yellowish after tetrachrome staining (Figure 12C). Striated tubules (Figure 12D) are less common and are located between batches of secretory tubules. These tubules are lined by a pseudostratified epithelium that comprises striated cells, cistern cells (firstly named by Matus, 1971) and lumen-lining cells. Located basally, striated cells are characterised by their loops of cellular membrane infoldings, which are arranged radially to the tubule's lumen, and have numerous mitochondria (Figure 12E). Cistern cells are found between the tubule's outer layer (the striated cell layer) and the lumen-lining cells and are composed by multiple infoldings of cellular membrane that connects to the tubule's lumen by crossing the lumen-lining cells (Figure 12F). Both striated and cistern cells are PAS-positive as they acquire a dark-purple colour after tetrachrome staining, despite the lighter tone in cistern cells. The lumen-lining cells have yellowish cytoplasm and are located all around the lumen of the striated tubules. Striated and secretory tubules were commonly found to intersect in proximal area of the gland, thus joining to form the main salivary duct (Figure 12G).

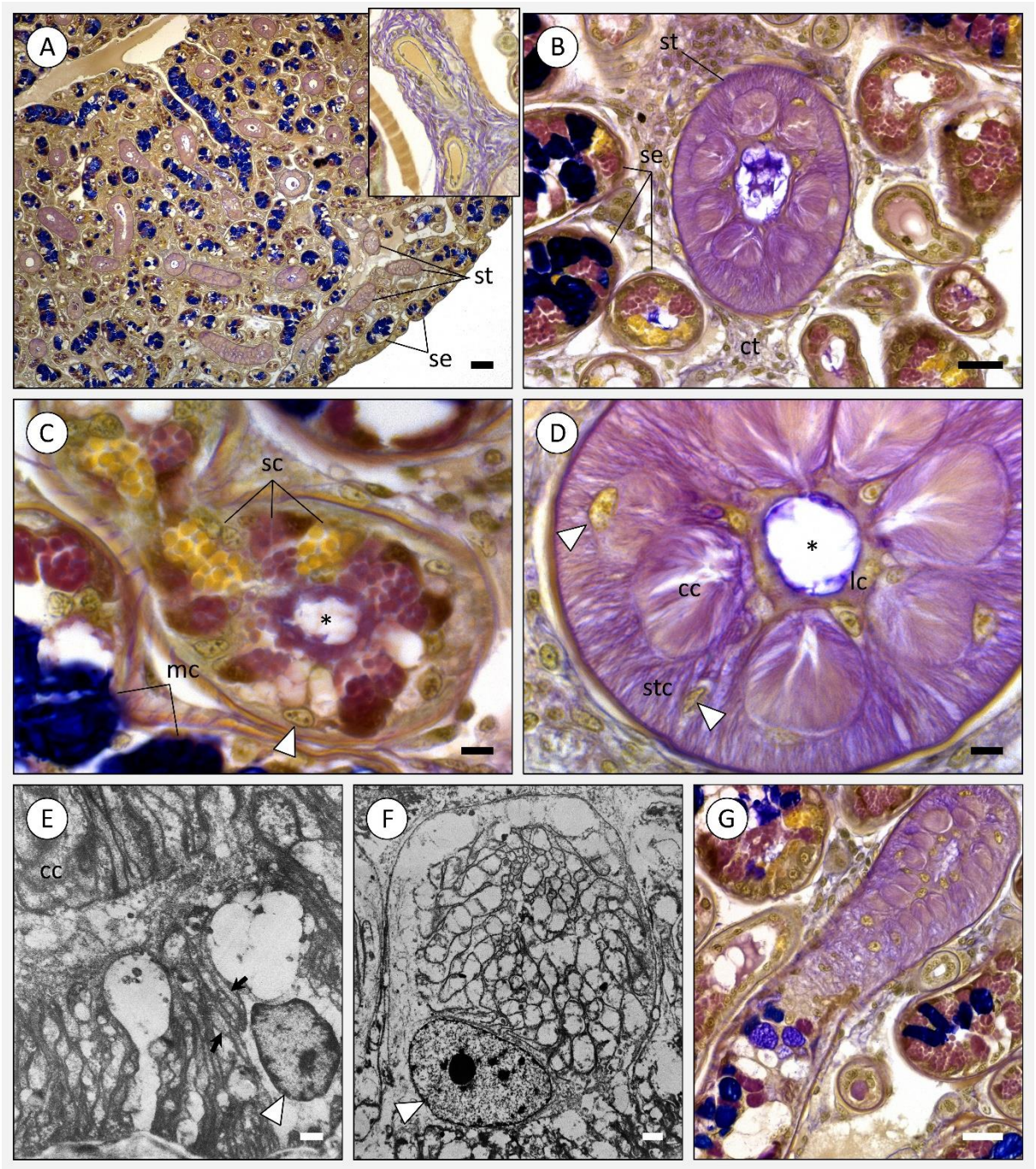


Figure 12 - Optical and transmission electron microscopy (TEM) images from the posterior salivary gland of *Octopus vulgaris*. **(A)** General appearance of the posterior salivary gland through a tetrachrome stain. Presence of two different tubules: the striated tubules (st) and the secretory tubules (se). Inset: Detail of two blood vessels present between the glandular tubules. **(B)** Detail of both striated (st) and secretory (se) tubules organised between the connective tissue (ct). **(C)** Secretory tubule composed of serous cells with purple-red and yellowish secretory granules (tetrachrome stain). These tubules are also formed by acidophilic mucus-secreting cells (mc). **(D)** Epithelium of a striated tubule formed by striated cells (stc), cistern cells (cc) and lumen-lining cells (lc). **(E)** TEM image of a striated cell formed by many loops of cellular membrane infoldings containing numerous mitochondria (black arrows). **(F)** TEM image of a cistern cell showing the multiple cellular membrane infoldings. **(G)** Intersection

of a striated tubule with a secretory tubule commonly found in the proximal area of the gland. Abbreviations: white arrowheads, nucleus; asterisk, lumen. Scale bars: **(A)** 100 μm; **(B, G)** 25 μm; **(C, D)** 10 μm; **(E, F)** 1 μm.

4.4.2 Transcriptomic profiles unveiled different functions between salivary glands

To assess the molecular differences between the two salivary glands, a large-scale transcriptome analysis (RNA-Seq) was performed. The *de novo* assembled transcriptome of the salivary glands of *Octopus vulgaris* yielded a total of 302,475 transcripts, corresponding to 207,109 predicted genes (Figure 13A). Of these, 51,090 resulted in potential open reading frames (ORFs), 53.35% of which were predicted to be full coding sequences, containing the start and stop codons. Comparing the transcript expression levels between both glands revealed that 2.32% of the total transcripts (~7,000 transcripts) are overexpressed in the posterior salivary glands (PSG), compared to 1.29% (~3,900 transcripts) of the total transcripts overexpressed in the anterior salivary glands (Figures 13B and 13C). From the differentially expressed ORFs was possible to annotate ~38% in posterior salivary glands and ~22% in the anterior salivary glands, through homology-based matching against Swiss-Prot (a curated database) (Figure 13D). Among the top ten differentially expressed transcripts, it was possible to identify a putative chitinase (potentially *Ch10*) variant overexpressed in ASG and five transcripts overexpressed in PSG, in this case identified as serine proteases (*Tmprss9*, *Acr*), proteases with inflammatory and anti-inflammatory roles (*Mct1A*), proteins linked with cell growth stimulation and differentiation (*Egf1*) and regulation of neurotransmitters (*DBH*) (Figure 13C).

Gene Ontology (GO) analysis was able to find 141 GO terms (Table B.4 in Appendix to Chapter 4) enclosing the differential expressed genes of both salivary glands, being the top 10 of each category (Molecular Function, Cellular Component and Biological Process) presented in Figure 13E. The anterior salivary glands yielded a significant enrichment of pathways (Fisher's Test, adjusted $p = 0.0088$) associated with protein glycosylation and vesicle-mediated transport between endoplasmic reticulum and Golgi apparatus. These pathways enclose, among others, cargo receptors, proteins with potential roles in transcription regulation and in biosynthesis of *post*-translational modifications and also proteins that promote the formation of transport vesicles. On the other hand, the posterior salivary glands yielded a significant enrichment of pathways associated with serine-type endopeptidase activity, cell adhesion and proteolysis

(Fisher's Test adjusted $p < 0.05$). Albeit with reduced significance (Fisher's Test adjusted $p > 0.05$), the ASG also shown pathways associated to the intracellular protein transport and the Golgi apparatus, in comparison to the PSG, which yielded a high number of GO terms linked with toxins and other venom components, as for instance the calcium and zinc ion bindings, protein kinase activity, signalling receptor binding and actin binding.

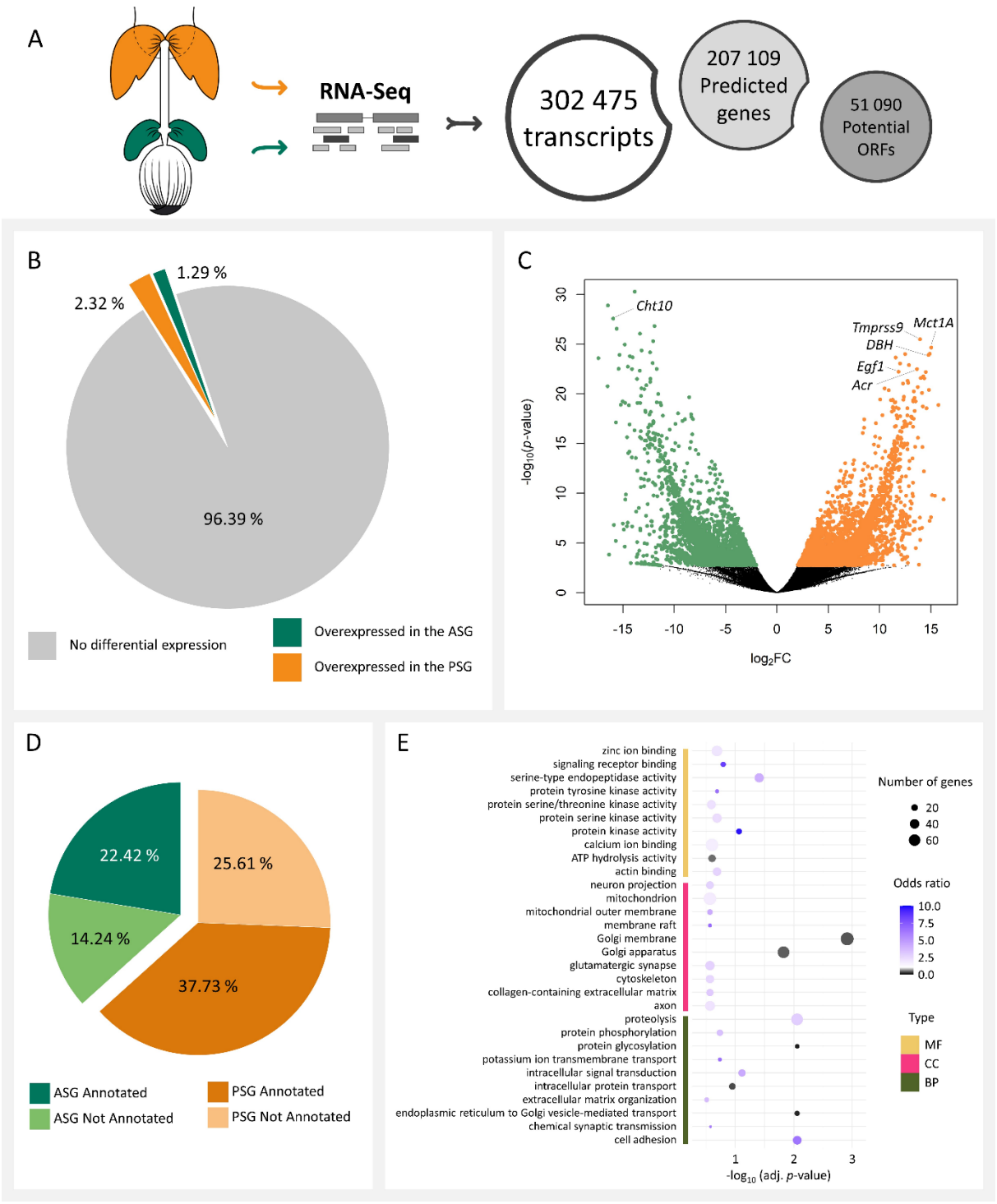


Figure 13 - General transcriptomic profile of the anterior and posterior salivary glands of the *Octopus vulgaris*. **(A)** Whole-transcriptome refinement for the selection of transcripts of interest. **(B)** Pie chart representing the percentages of the differentially expressed transcripts per gland. **(C)** Volcano plot of differentially expressed transcripts under- (green) and overexpressed (orange) in the posterior salivary gland relatively to the anterior salivary gland ($|\log_2 FC| > 1.5$ and FDR-adjusted $p < 0.05$). Top 10 under- and overexpressed transcripts were annotated by scanning for homology against Swiss-Prot database. **(D)** Pie chart showing the percentages per gland of differentially expressed open reading frames (ORFs) annotated through homology against the Swiss-Prot

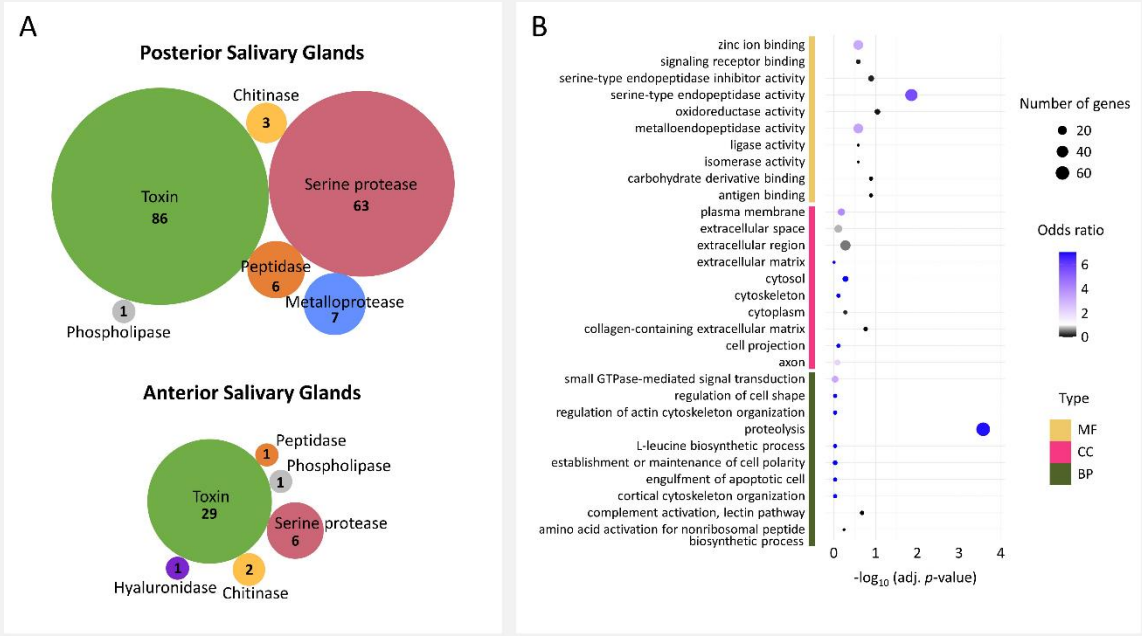
database. **(E)** Bubble plot of gene ontology (GO)/gene enrichment analyses of the differentially expressed genes (DEGs) with hit in Swiss-Prot database, of both salivary glands. Statistics were based on Fisher analysis of Top10 GO terms and GO terms of interest (FDR-adjusted $p < 0.05$). Odds-ratios higher than 1 correspond to a GO term enrichment in the posterior salivary gland and lower than 1 correspond to a GO term enrichment in the anterior salivary gland. Note that the same DEG can bear multiple GO terms.

4.4.3 Salivary glands enclose toxin-like proteins and bioactives

Since the posterior salivary gland returned cues about its function in producing toxins and related bioactives, all transcriptomic data were contrasted against the SeaTox database, a toxin-oriented customised database (see 'Materials and Methods' section for details). Thus, the findings detailed in Figure 14A revealed higher abundance of 'Toxins' and 'Serine proteases' in PSG compared to ASG, as well as 'Chitinases' and 'Peptidases', those in a lower scale of abundance. In addition, the PSG detained exclusively all hits for 'Metalloproteases' and the ASG had the only hit for 'Hyaluronidases'. The Gene Ontology (GO) enrichment analysis using only the SeaTox database selected genes was able to detect more specific GO terms for what is hypothesised to be the major functions of each salivary glands (Figure 14B). Thus, the PSG yielded a significant enrichment (Fisher's Test, adjusted $p < 0.05$) for pathways associated with "serine-type endopeptidase activity" and "proteolysis", where venom proteins, metalloendopeptidases and serine proteases are the most involved proteins. Albeit not significantly enriched, the pathways as "zinc ion binding", "metalloendopeptidase activity", "extracellular matrix", "cell projection" and "axon" also disclosed the involvement of several other proteins in PSG such, proteins containing ShKT domains, Ras-related C3 botulinum toxin substrate 1, Hemolin and proteins already described in snake and stonefish venoms, such as the snake venom metalloproteinase atrolysin-B-like and the Stonustoxin subunit alpha. The GO enrichment analysis in ASG, on the other hand, revealed several pathways associated with "oxidoreductase activity", "serine-type endopeptidase inhibitor activity", "antigen binding", "carbohydrate derivative binding" and "signalling receptor binding", where were detected proteins that can have neurotoxic and immunosuppressive effects, proteins that inhibit the activation of essential enzymes for blood clotting and others that additionally may interfere with platelet function. Gene enrichment analysis also unveiled that the products of expression of several genes are associated to the extracellular region, extracellular space or both in the ASG. Nonetheless, the majority of overexpressed genes associated to extracellular compartments pertains to the PSG. Thus, the genes that could be identified in the ASG encoded, for instance, chitin-binding proteins, protease inhibitors,

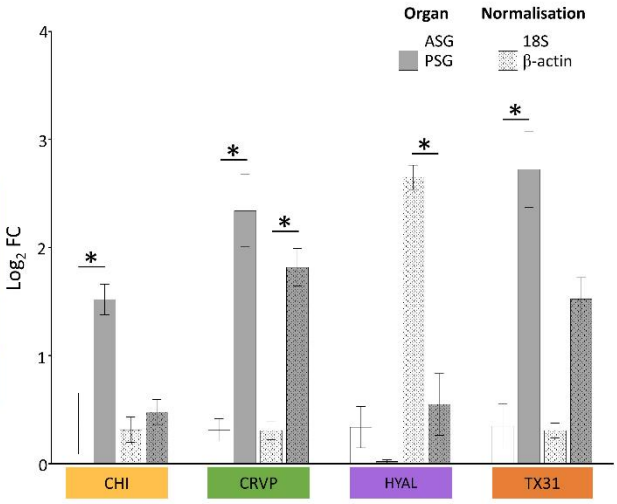
hyaluronidase, phospholipase, peptidase, cysteine-rich venom protein and other proteins with procoagulant activities or that act in blood triglyceride degradation.

Four transcripts were selected for validation of gene expression analyses by RT-qPCR. These were selected based on their function as toxin/venom or permeabilising enzymes, availability of full coding sequence, expression and reviewed annotation. The set precludes a chitinase (CHI) with 418 aa, two cysteine-rich venom proteins (CRVP and TX31) with 281 aa and 500 aa and a hyaluronidase (HYAL) with 436 aa (all accessions are provided in 'Data availability statement' section). The results were found generally accordant with the relative quantification of transcripts by RNA-Seq. Both cysteine-rich venom proteins (CRVP and TX31) and the chitinase revealed significant higher expression levels in PSG comparatively to the ASG (Figure 14C). On the other hand, the expression levels of the hyaluronidase were significantly higher in the ASG in comparison to the PSG, as expected by the gene expression values (Figure 14C). The full coding sequences of the transcripts were also validated by Sanger sequencing and matched against homologs with the most significant hits amongst venomous or toxin-bearing Arthropoda, Mollusca, Cnidaria and Chordata. The phylogenetic trees showed that the target sequences present homology to different species of cephalopods (Figure 14D). Interestingly, both CHI and the HYAL sequences were found to pertain to the octopod clade, revealing higher similarity with other chitinases and hyaluronidases from both *Octopus sinensis* and *Octopus vulgaris*. In turn, both cysteine-rich venom proteins (CRVP and TX31) sequences had displayed higher similarity with other CRVP and cysteine-rich secretory protein (CRISP) from *Sepia officinalis* and *Sepia pharaonis*, than from octopodids.

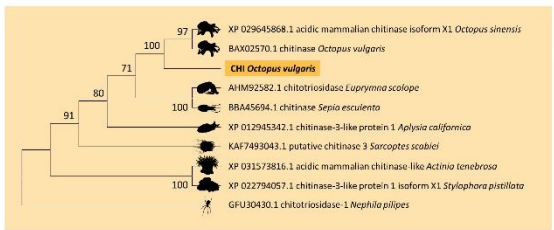


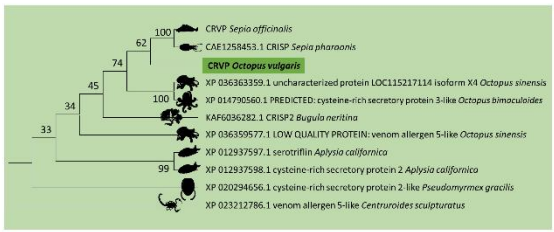
C

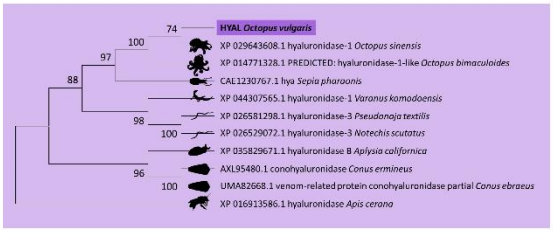
Name	Log_2 FC (PSG)	*p*-value	%ID	*e*-value
Acidic mammalian chitinase	13.4	4.00×10^{-7}	47.9	7.64×10^{-127}
Cysteine-rich venom protein latisemin (CRVP)	10.9	4.41×10^{-21}	37.2	2.08×10^{-29}
Hyaluronidase-1	-5.1	6.99×10^{-10}	35.9	3.25×10^{-87}
Cysteine-rich venom protein (TX31)	10.3	1.07×10^{-11}	34.9	5.45×10^{-32}



D







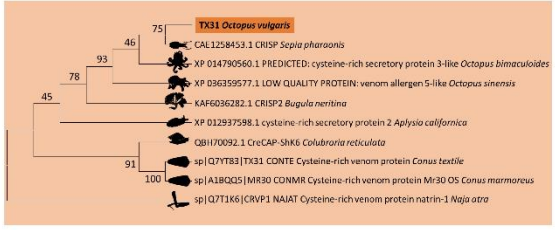


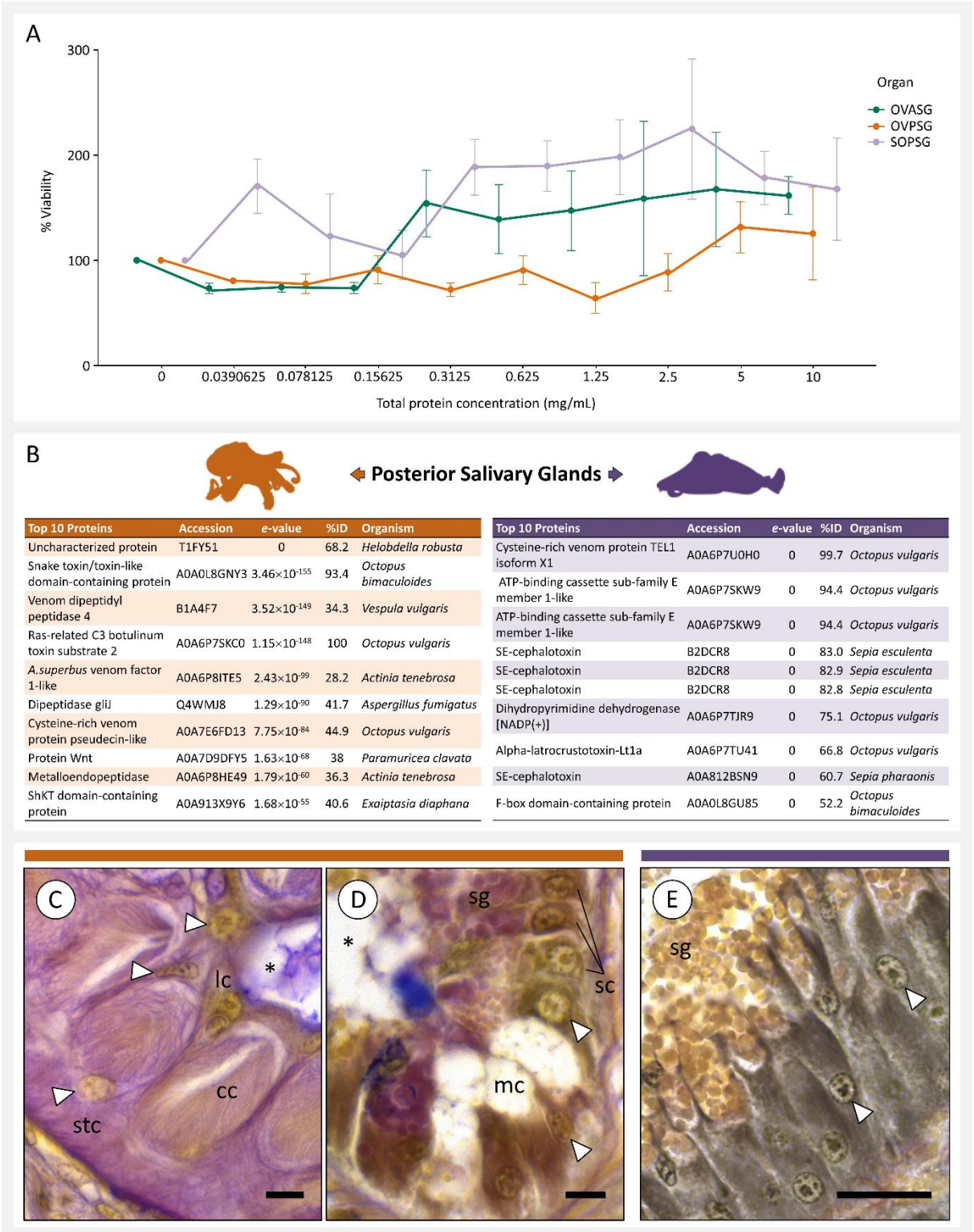
Figure 14 - Transcriptomic profile of the anterior and posterior salivary glands of the *Octopus vulgaris* based on homology matching with the SeaTox database. Transcriptome validation and phylogenetic analyses. **(A)** Categories of proteins of interest with match in SeaTox database. Figures represent the number of genes with hit in each category. Note that genes may be classified in more than one category. **(B)** Bubble plot of gene ontology (GO)/gene enrichment analyses of the differentially expressed genes (DEGs) of both salivary glands with homology matching against SeaTox database. Statistics were based on Fisher analysis of Top10 GO terms and GO terms of interest (FDR-adjusted $p < 0.05$). Odds-ratios higher than 1 correspond to a GO terms enrichment in the posterior salivary gland and lower than 1 correspond to a GO term enrichment in the anterior salivary gland. Note that the same DEG can bear multiple GO terms. **(C)** Description of expression analysis of the selected transcripts (table) used for validatory expression analysis by RT-qPCR (bar plot). Both 18S and β-actin genes were chosen as housekeeping genes for normalisation. Asterisks indicate significant differences between both organs (Student's *t*-test $p < 0.05$). **(D)** Phylogenetic trees of transcripts of interest: chitinase (CHI), cysteine-rich venom proteins (CRVP and TX31) and hyaluronidase (HYAL). Each model includes the best matches from venomous or toxin-bearing animals for comparison. The phylogenetic models were produced using Maximum Likelihood and the Jones-Taylor-Thornton model, with 800 bootstrap replications. Bootstrap support values are given for all nodes.

4.4.4 Crude protein extracts yielded undetermined toxicity to molluscan primary cells

To compare the toxicological effects of the salivary glands extracts at cellular level, primary cells isolated from the gills of the model organism *Mytilus* sp. were exposed to crude protein extracts from PSG and ASG. In addition, venom (PSG) gland extract from *Sepia officinalis* was tested as comparison, as this predator is also known to produce a cocktail with toxins and permeabilising agents in their salivary glands. Exposure to any of the extracted resulted in no significant impairment in cell viability (Kruskal–Wallis *H* or *F*-test $p > 0.05$). Despite the slight decrease of cellular viability in experiments with lower doses of extracts from *Octopus* glands, the results showed an increase of cellular viability percentage at higher concentrations of all extracts, however with no statistical significance, likely reflecting a delayed viability boost of primary cells after harvesting (Figure 15A).

The top ten proteins of the *O. vulgaris* PSG with hits in the SeaTox database ordered by *e*-value and identity percentage are shown in Figure 15B. For inter-species comparison, former transcriptomic data of previous work of the *S. officinalis* PSG was also contrasted against the SeaTox database and the top 10 proteins were retrieved (Figure 15B). The results revealed that the PSG from either species holds toxins and proteins that can produce neurotoxic effects. However, it is noted that PSG of *S. officinalis* is more dedicated to produce toxins in abundance since half of the top ten proteins are cephalotoxin or neurotoxin homologs. In turn, the PSG of *O. vulgaris* presented more ion binding proteins, proteins that modulate cell functions and

venom factors, which should be important assets to facilitate the entry and diffusion of toxins into prey. At histological levels, important differences between both PSG were evident (Figures 15C, 15D and 15E). On one hand, the PSG of *O. vulgaris* possessed different types of tubules (secretory and striated tubules), as described above and detailed again in Figures 15C and 15D. The PSG of *S. officinalis*, in turn, bears a single type of tubule formed by serous cells (Figure 15E).



Top 10 Proteins	Accession	*e*-value	%ID	Organism
Uncharacterized protein	T1FY51	0	68.2	*Helobdella robusta*
Snake toxin/toxin-like domain-containing protein	A0A0L8GNY3	3.46×10^{-155}	93.4	*Octopus bimaculoides*
Venom dipeptidyl peptidase 4	B1A4F7	3.52×10^{-149}	34.3	*Vespula vulgaris*
Ras-related C3 botulinum toxin substrate 2	A0A6P7SKC0	1.15×10^{-148}	100	*Octopus vulgaris*
A.superbus venom factor 1-like	A0A6P8ITE5	2.43×10^{-99}	28.2	*Actinia tenebrosa*
Dipeptidase gliJ	Q4WMJ8	1.29×10^{-90}	41.7	*Aspergillus fumigatus*
Cysteine-rich venom protein pseudecin-like	A0A7E6FD13	7.75×10^{-84}	44.9	*Octopus vulgaris*
Protein Wnt	A0A7D9DFY5	1.63×10^{-68}	38	*Paramuricea clavata*
Metalloendopeptidase	A0A6P8HE49	1.79×10^{-60}	36.3	*Actinia tenebrosa*
ShKT domain-containing protein	A0A913X9Y6	1.68×10^{-55}	40.6	*Exaiptasia diaphana*

Top 10 Proteins	Accession	*e*-value	%ID	Organism
Cysteine-rich venom protein TEL1 isoform X1	A0A6P7U0H0	0	99.7	*Octopus vulgaris*
ATP-binding cassette sub-family E member 1-like	A0A6P7SKW9	0	94.4	*Octopus vulgaris*
ATP-binding cassette sub-family E member 1-like	A0A6P7SKW9	0	94.4	*Octopus vulgaris*
SE-cephalotoxin	B2DCR8	0	83.0	*Sepia esculenta*
SE-cephalotoxin	B2DCR8	0	82.9	*Sepia esculenta*
SE-cephalotoxin	B2DCR8	0	82.8	*Sepia esculenta*
Dihydropyrimidine dehydrogenase [NADP(+)]	A0A6P7TJR9	0	75.1	*Octopus vulgaris*
Alpha-latrocrustotoxin-Lt1a	A0A6P7TU41	0	66.8	*Octopus vulgaris*
SE-cephalotoxin	A0A812BSN9	0	60.7	*Sepia pharaonis*
F-box domain-containing protein	A0A0L8GU85	0	52.2	*Octopus bimaculoides*

Figure 15 - Toxicological comparison and characterisation of salivary glands from *Octopus vulgaris* and *Sepia officinalis*. **(A)** Bar plot representing the percentage of mollusc cellular viability after exposure to different concentrations of the salivary glands' crude protein extracts (ASG and PSG of *O. vulgaris* and PSG of *S. officinalis*). **(B)** Top 10 hits of putative proteins with matching with the SeaTox database of both posterior salivary glands of *O. vulgaris* and *S. officinalis*. **(C)** Histological section of a striated tubule of the PSG of *O. vulgaris*, presenting three

different cells: striated cells (stc), cistern cells (cc) and lumen-lining cells (lc). **(D)** Histological section of a secretory tubule of the PSG of *O. vulgaris* showing the serous cells (sc), bearing several secretory granules (sg), and empty mucocytes (mc). The acidophilic content of mucocytes secreted to the tubule's lumen (asterisk) was stained blue (Alcian Blue). **(E)** Histological section of the PSG of *S. officinalis* detailing the glandular cells filled apically with secretory granules (sg). Abbreviations: white arrowheads, nucleus; asterisk, lumen. Scale bars: **(C, D)** 10 μm; **(E)** 25μm.

4.5 Discussion

The digestive system of cephalopods follows the general molluscan body plan. However, it holds special adaptations toward feeding strategies that combine their skills as problem solvers plus the use of camouflage and visual communication through special modifications of the skin, with the ability to secrete a venomous saliva. The anterior and posterior salivary glands of the common octopus have specific microanatomy and transcriptomic profiles, which reveals distinct functions. Their positioning in the digestive system, including sharing a common duct through which their secretory products are combined and released through the buccal mass, indicate that their distinct secretory products are intended to produce effects that include pre-digestion of food and permeabilising prey to enhance action of stunning toxins and other venom components. Our work highlighted the major differences between the pair of duals-lobed glands, showing that the (somewhat misleadingly) biochemically simpler anterior salivary gland (ASG) chiefly provides the vehicle for the bioactives secreted by the posterior salivary gland (PSG) that with that respect, can effectively be termed 'venom gland'. Altogether, the different substances are secreted by the distinct secretory epithelia of both ASG and PSG with multiple purposes, ranging from predator strategies to the digestive processes, which has been hypothesised by previous authors dealing on histochemistry and biochemistry of the cephalopod digestive tract and annex glands (see Ponte and Modica, 2017, for a review).

The combined outcomes of microanatomy and transcriptomics in the present work showed clear differences between the two salivary glands of *O. vulgaris*, as reported in other cephalopods, with emphasis on the cuttlefish *S. officinalis* (Gonçalves et al., 2024). The first hint for the complementary functions of the two salivary glands was revealed from microanatomy. Comparatively, the ASG is seemingly simpler, being formed by blind-end tubules (branched diverticula) and a single type of secretory epithelium. In its turn, the PSG is a more complex arrangement of tubules with two distinct types of epithelia and higher diversity of cell types (such as goblet, striated and cistern cells). In this gland, the two 'forms' of tubules are not, however, separate structures but instead they result from the differential specialisation within

the same tubule. The clearly distinct biochemical signature and cellular structure of the two types of epithelia and their sequential positioning along the tubules suggest higher specialisation of the posterior gland and concomitantly, an important degree of specificity of the bioactives such as cephalotoxins and matrix-directed proteinases (as permeabilisers) here produced can be expected when compared with the anterior counterpart. This hypothesis is supported by comparative transcriptomics, not only, by the high number (~7,000) of overexpressed transcripts in the PSG relatively to the ASG but also by their function, especially those that encode for toxins and permeabilising enzymes. It should be noted, though, that the two glands share ≈ 96% of the transcriptome (annotated or not). Similarly, our previous work on the PSG of the common cuttlefish revealed that this gland is also formed by branched glandular tubules, yet bearing a single columnar epithelium. Despite presenting a simpler microanatomy compared to the octopus PSG, this organ is highly specialised in producing proteinaceous substances as observed by the copiousness of protein-containing vesicles (secretory granules) in the glandular cells. This is also confirmed by transcriptome profiling, which unveiled several putative toxin-encoding transcripts such as cysteine-rich venom proteins, cephalotoxins, venom insulins, and also permeabilising agents (especially intercellular matrix-directed enzymes).

Regardless of differences in the secretion of bioactive substances, either gland produces its own mucin-rich saliva that may act as the specific vehicle of the substances that ensures its efficient transport and bioactivity (e.g., by maintaining the right pH during transport until they are released through the buccal mass). Indeed, even though either gland produces acidic mucopolysaccharides (long chains of sugars typically found in mucins), the secretion of neutral or more basic mucopolysaccharides is far more evident in the ASG, resulting in its PAS-positive purple tint of tubule epithelial secretory cells. Altogether, the broader span of mucin-like substances produced by the ASG supports the role of this gland as the main provider of vehicle, whereas PSG, which its span of different secretory cells is likely the main secretor of bioactive substances, especially proteins and peptides. It must be noted, though, that the present work somewhat contradicts that of Anadón (2019), who reported two different forms of tubules of the ASG, yet composed of the same cells. Here we suggest that the two 'types' actually correspond to different phases of the cycle of epithelia, i.e., before and after apocrine secretion, resulting in epithelia with vesicle-laden secretory cells and epithelia composed of 'empty' cells. In the PSG, on the other hand, the presence of two types of tubules is far more obvious. One of the types holds both vesicle-secreting (calyx-like) cells and

mucocytes (goblet-like cells), altogether suggesting that proteins are secreted with its mucinous vehicle. This shows that this paired gland, which is known as the venomous gland of coleoid cephalopods, is highly specialised in producing proteinaceous substances. In addition, the second type, i.e., the striated epithelium, grants to the PSG unique abilities in comparison to the ASG. As originally described by Matus (1971) in the posterior salivary gland of *Eledone cirrosa* and *Octopus vulgaris*, striated cells hold numerous mitochondria between the multiple infoldings of the cellular membranes, which indicates that these cells require a heft energy budget to cope with their functions, mostly related with maintenance of the osmotic balance. Additionally, this same author revealed that adjacent cistern cells are involved in active ion transport, which involves high energy demands as well.

Altogether, the findings reported above show that the PSG is more complex structurally and more energy-demanding than the ASG, however, one of its main assets is arguably its direct association with the octopus' central nerve system (CNS). Indeed, both striated and cistern cells of the PSG are associated with bundles of axons containing neurofilaments and nerve terminals (Matus, 1971). The work by House (1980) on the salivary glands of several invertebrates (e.g. snails, octopuses, ticks, cockroaches, moths and blowflies), reviewed the abundant and dual innervation of the *Octopus vulgaris* PSG following the preceding work by Young (1965). The latter author reported that the PSG is provided with two groups of nerves ('dual innervation') embedded in the muscular duct wall, one from the superior buccal lobe of the brain and the other from the peripheral subradular ganglion, which should ultimately control the secretion of the toxin-containing saliva. The same authors stated then, in the mid-1960s, that there is little information on how this ganglion is operated by the brain, an issue that remain to be clarified to the present day. Once within the PSG, the first group of nerves branches to the tubules' epithelia (being called secretory nerves), while the second group is responsible for the control of the muscular contraction of the glands (termed motor nerves) (Ducros, 1972). In accordance, transcriptomics enabled us to identify the transcripts encoding proteins associated with regulation of neurotransmitters among the top ten annotated transcripts that were found overexpressed in the PSG. These outcomes illustrate that the PSG is more enervated and have its function finely tuned by the CNS, which is coherent with the behaviour of such an active predator as *O. vulgaris*.

The *O. vulgaris* venom gland secretes a complex cocktail of matrix directed enzymes plus toxins. In particular, we detected here higher abundance of transcripts encoding for venom proteases, metalloendopeptidases and serine proteases, all significantly enriching the 'serine-

type endopeptidase activity' and 'proteolysis' pathways. These protein homologs contain the peptidase S1 and ShKT domains that are often described as components in venoms of reptiles, caterpillars, wasps, vampire bats, several insects, sea anemones and cephalopods, such the *O. vulgaris* (Fry, 2009; Castaneda, 1995, Fingerhut et al., 2018). The peptidase S1 toxins have been associated to external digestion, immune response and blood coagulation, and were also hypothesised that in the vampire bat (*Desmodus rotundus*) could be involved in activation of other proteins with anticoagulant activity during the capture of prey (Fry, 2009). In turn, ShK domain toxins (ShKT), firstly characterised and isolated as a toxin from the Caribbean Sea anemone *Stichodactyla helianthus*, are potent potassium channel blockers and have been suggested as promising toxin-based drug candidate directed against autoimmune diseases (Castaneda, 1995; Bordon et al., 2020). Transcriptomics results also highlighted that the ASG of *O. vulgaris* presented higher abundance of homologs that encode for proteins containing chitin-binding domains, phospholipases, peptidases and hyaluronidases, with less abundance of toxins in comparison with the PSG. These classes of enzymes such as chitinases and phospholipases, act in the extracellular region promoting, for instance, chitin and lipid degradation and therefore aid in the digestive processes. The anterior salivary gland also displayed a significant enrichment of pathways associated with protein glycosylation. This *post*-translational modification plays an important role in protein folding, stability and solubility, which can promote biological effectiveness in venom proteins (He et al., 2024).

Our recent omics-based research pipelines have already disclosed that the venoms and poisons of marine invertebrates, from cnidarians to annelids and gastropods, contain a broad span of putative bioactives including cysteine-rich neurotoxins characterised by CAP domains and by being potential blockers or inhibitors of potassium-, sodium- and calcium-channels, therefore affecting muscle function of prey (Rodrigo et al., 2021; D'Ambrosio et al., 2021; Moutinho Cabral et al., 2022; Campos et al., 2023; Alcaide et al., 2024). These findings are accordant with those on cephalopods, which shows that these proteinaceous toxins are ubiquitous among Eumetazoa and that marine taxa are a wide source of these valuable biotechnological assets. Whitelaw et al. (2016) identified nine protein families behind the 23 putative proteins profiled by a combination of transcriptomics and proteomics of the PSG of the octopuses *Hapalochlaena maculosa* and *Octopus kaurna*. Among others, these included CAP proteins plus extracellular matrix-directed enzymes such as chitinases, hyaluronidases, metalloproteases, phospholipases and serine proteases. However, Hartman et al. (1960), through chromatography and radioactive methods, had already alerted for significant

between-taxa variability of endogenous amines and aminoacids decarboxylases content in the PSG of *O. apollyon* (now *Enteroctopus dofleini*) and *O. bimaculatus*. In fact, we verified that proteins being secreted in the octopus and cuttlefish PSG were mainly linked to neurotoxin homologs or proteins that mediate the entry and diffusion of toxins into prey. However, we detected a higher number of cephalotoxin homologs in PSG of the cuttlefish *Sepia officinalis* (Gonçalves et al., 2024). Interestingly, a cephalotoxin was the first proteinaceous toxin to be successful isolated from the posterior salivary gland, not only in this species but in coleoids altogether (Ghiretti, 1959). It must be noted that phylogenetic analyses suggest that full coding sequences of ubiquitous proteins of interest (recall Figure 14D), present small variations within venomous taxa, however without prejudice of a certain degree of similarity between orthologs from other Cephalopoda. Both chitinase and hyaluronidase are closely related to their homologs from octopodids, still *O. vulgaris* cysteine-rich venom proteins (CRVP and TX31) present higher homology with *Sepia*, forming a distinct clade. In the latter case, the cysteine-rich venom proteins may be distinct from better characterised forms from *Conus* and *Aplysia*, which reinforces the need for further investigation and characterisation of venom proteins in coleoids. Altogether, whereas the phylogeny of permeabilising proteins on the venom of the common octopus may appear relatively straightforward, the phylogeny of toxins may be more intricate and reflect multiple moments of convergent adaptation within the Cephalopoda.

Despite the known richness of compounds in salivary glands, with special emphasis on proteinaceous toxins, studies on the toxicity of these compounds to molluscs are still lacking. Our preliminary assays with crude posterior salivary gland extracts of *O. vulgaris* and *S. officinalis* onto primary mussel gill cells resulted in no significant impairment in cellular viability, albeit significant variation. It should be noticed that mussel gill cells are attractive models in ecotoxicological studies because gills are the first uptake site for many toxicants in the aquatic environment, thus being expectedly affected by exposure to waterborne chemicals (Gómez-Mendikute et al., 2005). However, the activity of the octopus' venom may depend on entry, as injection via buccal apparatus is quite different from surface contact. On the other hand, we may suggest that molluscan cells may not be the most adequate target for the toxins and enzymes therein as the common octopus is known to prefer crustaceans and even fish (Villanueva et al., 2017). In addition, gills cells are not primarily neuromuscular, which could explain the lack of effects of the extracts neurotoxic fraction. Altogether, testing the bioactivity of crude protein extracts and isolated proteins need to be undertaken before we may reason the full biotechnological potential of these proteins.

4.6 Conclusions

As in other coleoid cephalopods, both salivary glands in *O. vulgaris* are responsible for secreting a mucin-based saliva that acts as vehicle for substances (especially proteins) involved in the capture and digestion of prey. The anterior salivary gland secretes mainly mucus and other proteinaceous substances with neurotoxic (e.g., ShKT domain-containing proteins), immunosuppressive and inhibitors of hydrolytic enzymes that act in blood clotting and platelet function processes. This gland seemingly holds an important role in protein functionalisation, as it bears enzymes involved in *post*-translational modifications such as glycosylation. The more intricate microanatomy of the posterior salivary gland, with respect to diversity of cell types and finer neuromotor control, suggests higher specificity in the production and release of venom proteins (including cysteine-rich venom proteins, toxin-like domain-containing proteins and neurotoxins) and accompanying permeabilising agents.

Altogether, the anterior salivary gland is likely the activating vehicle of the posterior gland's toxins and enhancers when the animal is feeding. Both glands secrete potentially valuable bioactive compounds. We isolated and validated the full coding sequences of proteins of the octopus's salivary glands, including two cysteine-rich venom proteins, a chitinase, and a hyaluronidase. The comparison between the posterior (venom) glands of the octopus and the cuttlefish revealed that the cuttlefish may be more specialised in producing more neurotoxic substances than the octopus. Nonetheless, the extracts from the salivary glands of both species did not reveal cytotoxic effects on mollusc primary gill cells, indicating that this biological model may resist to this cocktail of bioactive compounds. Still, further tests are needed to investigate the effects of crude and isolated toxin extracts.

5

HUMAN G PROTEIN-COUPLED RECEPTORS AS TARGETS OF BIOACTIVES FROM THE CEPHALOPOD VENOM GLAND

5.1 Abstract

Cephalopod venoms are complex mixtures of protein-based bioactive substances optimised through evolution to hold significant levels of potency and specificity toward their molecular targets. This characteristic is an important asset for drug discovery against the human druggable proteome, which now expands to G protein-coupled receptors (GPCRs) due to their involvement in almost every biological function, thus pushing their momentum as therapeutical targets. Even though early research for venom peptides as potential ligands for GPCRs already yielded some peptides that interact with these receptors; the potential of marine toxins remains unknown, albeit being increasingly acknowledged as prized resources for drug discovery. In the present work eight human GPCRs were screened for ligands within crude protein extracts of the salivary glands of the common octopus (*Octopus vulgaris*) and the common cuttlefish (*Sepia officinalis*). The extracts of the posterior salivary glands (PSG) or "venom gland" of both species were tested, plus the extract from the anterior salivary gland (ASG) or "mucus gland" of *O. vulgaris*. The activity of selected GPCRs was measured through the G protein–independent β-arrestin–recruitment assay in human HTLA cells. The results revealed that PSG extracts from both *O. vulgaris* and *S. officinalis* activated significantly all tested GPCRs, however with different efficacies. Overall, the PSG extract from *Octopus* appeared to be more specific for receptors that regulate both cardiovascular function and the blood glucose levels, whereas the PSG extract from *Sepia* presented higher affinity for five out of eight GPCRs. However, the ASG extract from *Octopus* yielded either absence of interaction or the inhibition of receptor activity. Although additional research is required, this study highlighted the potential of cephalopod bioactives in bioprospecting.

5.2 Introduction

Venoms are composed of a complex cocktail of proteinaceous compounds that have been sequentially optimised through the evolutionary process. The genetic diversity of venoms has mostly resulted from the recruitment of bioactives and its evolution led to an ultimate efficacy and specificity of these highly bioactive substances towards their targets. These substances, which can be delivered in a wide range of different structures (e.g. jaws, teeth or fangs, spurs, proboscises, nematocysts and stings), are secreted, stored and released as mechanisms for the capture of prey or when self-defence is triggered. These substances, which form an intricate

cocktail commonly known as venom or poison, can directly affect the neuromuscular and the cardiovascular systems of target organisms, generating uncontrolled proinflammatory responses, while often possessing antimicrobial and antiviral properties as well (Calvete et al., 2009; Zhang, 2015; Karthik et al., 2019).

Despite the lack of studies on the composition of venoms, it is known that they can interfere with several physiological and biochemical processes due to their affinity to more or less specific molecular targets, with emphasis on cell surface receptors such as ion channels and transporters (Lewis and Garcia 2003; Stonik et al., 2016; Bordon et al., 2020). As such, toxins and other bioactives in venoms are prime targets for drug discovery and development. In addition, their composition is comprised mostly of small peptides and proteins bearing *post*-translational modifications (e.g., glycosylation), which provides them with biochemically stable and resistant structures, to which is added relatively ease of synthesis by using DNA recombinant technology through convenient models such as *E. coli* and yeast (King, 2011; Van Baelen et al., 2022).

Even though the number of marketed products might still be disappointing, there are important examples of approved pharmaceuticals derived from animal venoms. These were pioneered by the painkiller ziconotide (commercialised as Prialt®), a recombinant protein designed from a neurotoxin (ω-conotoxin) isolated from the cone snail *Conus magus* that effectively blocks calcium channels (Williams et al., 2008). Other approved drugs with different modes of action have also been marketed, such as captopril and bivalirudin. Captopril was developed from a bradykinin-potentiating peptide isolated from the venom of the Brazilian pit viper *Bothrops jararaca* and became the first oral angiotensin-converting enzyme (ACE) inhibitor used for the treatment of hypertension and some types of congestive heart failure (Cushman and Ondetti, 1991). On the other hand, bivalirudin, a synthetic analog of hirudin, which is an anticoagulant found in the saliva of medicinal leeches (*Hirudo medicinalis*), directly inhibits thrombin, an enzyme involved in blood clotting and is used as an anticoagulant that is mostly applied during coronary interventions (Bates and Weitz, 1998). Overall, it is of general consensus that the pharmaceutical potential of venoms is just beginning to be understood. Within this context, new approaches for drug development and the discovery of new therapies have been gaining attention, particularly those focused on G protein-coupled receptors (GPCRs). GPCRs represent the largest family of membrane proteins that function as cell surface receptors and modulate important physiological processes such as hormonal signalling, neuronal transmission, sensory transduction, and cell-cell

communication (Prashanth et al., 2017). These signalling cascades of events are triggered by chemically diverse ligands, including ions, photons, odorant molecules, hormones, amino acids, lipids, nucleotides, peptides and proteins (Lagerström and Schiöth, 2008). Due to these characteristics, GPCRs have been highlighted as important therapeutic targets, with already marketed drugs, including exenatide, the synthetic version of exendin-4, a glucagon-like peptide isolated from the saliva of the Gila monster (*Heloderma suspectum*), which is used for the treatment of type 2 diabetes mellitus (Eng et al., 1992; Furman, 2012). With the purpose of drug development, efforts have been made to discover new GPCR allo- and orthosteric ligands, both agonists and antagonists, derived from animal venoms and toxins, yet much progress is still needed. (Van Baelen et al., 2022).

Marine bioprospecting has been paying attention to the high diversity of venomous marine animals, especially invertebrates, due to their unsurmountable biodiversity. Coleoids, in particular, a subclass of the Cephalopoda represented by octopuses, squids and cuttlefishes, are prized targets for bioprospecting, especially cuttlefishes and octopuses, because they are long known to secrete a venomous saliva. Current research has been revealing the complexity of cephalopod venoms, i.e., salivary secretions, which are mainly composed of proteinaceous bioactives, including compounds with potential neurotoxic and endocrine-disrupting effects, permeabilising enzymes, such as metalloproteases and serine proteases, antimicrobial peptides and even hormones (Caruana et al., 2016; Cooke et al., 2017; Almeida et al., 2020; Gonçalves et al., 2024).The objective of the present work is to screen for human GPCRs ligands in the crude protein extracts of the salivary glands of the common octopus (*Octopus vulgaris*) and the common cuttlefish (*Sepia officinalis*). For the case, the extracts of the posterior salivary glands or "venom gland" of both species were tested, plus the extract from the anterior salivary gland or "mucus gland" of *O. vulgaris*. The tested GPCRs were chosen from PRESTO-Tango GPCR kit and the modulation of their activity by salivary extracts was measured through the G protein–independent β-arrestin–recruitment assay in human HTLA cells.

5.3 Materials and Methods

5.3.1 Animal collection and tissue homogenisation

Seven *Octopus vulgaris* (mantle length 10.78 ± 2.32 cm; weight 487.43 ± 272.69 g) and one *Sepia officinalis* (mantle length 17.26 cm; weight 450.00 g) were collected from Praia do Norte

(38°39′0.153″ N, 9°14′43.876″W) beach on the Portuguese West coast. After euthanasia by sectioning the central brain, the organs of interest, i.e., anterior (ASG) and posterior salivary glands (PSG) of *O. vulgaris* and the PSG of *S. officinalis*, were harvested and immediately stored at –80°C until further processing. Crude protein extracts from the three different tissues were obtained from approximately 300 mg of pooled samples per target organ. Pooled samples were homogenised in 1 mL of cold and sterile Dulbecco's phosphate-buffered saline (PBS) pH 7.4, followed by centrifuging at >10,000 × g for 20 min at 4°C, to remove cellular debris. The crude extracts' total protein content was determined using the Pierce 660 reagent (Thermo Fisher Scientific, Waltham, MA, USA), with bovine serum albumin as a standard.

5.3.2 GPCRs of interest

The panel of G protein–coupled receptors (GPCRs) of interest was customised from the list of available receptors in the PRESTO-Tango plasmid kit (kit #1000000068 from Addgene, Watertown, MA, USA) and receptor activation was assessed via the modified Tango β-arrestin recruitment assay (Zeghal et al., 2020). The GPCRs were selected based on the match of their reported functions from available literature and the functions of the bioactive compounds that compose the cocktail of salivary glands' secretions of both *O. vulgaris* and *S. officinalis*. A total of eight GPCRs of interest were selected and are listed in Table 4. The PRESTO-Tango plasmid kit was developed by Bryan Roth (Kroeze et al., 2015).

Table 4 – Families of G protein–coupled receptors (GPCRs) of interest with their reported functions and related toxin studies. The GPCRs were selected from the PRESTO-Tango plasmid kit based on the match of their functions and the functions of the bioactive compounds that compose the cocktail of salivary glands' secretions of both *O. vulgaris* and *S. officinalis*.

Functions	Toxin studies	Reference
Vasoactive intestinal peptide receptor (VIPR1)		
Modulates smooth muscles relaxation, exocrine and endocrine secretion, immune response, regulation of water/ion flux in lungs and intestinal epithelia.	Helodermin, venom peptide from the Gila monster (*Heloderma suspectum* and *Heloderma horridum*) interacts with VIP receptors stimulating cyclic AMP formation and amylase secretion.	Amiranoff et al., 1983; Robberecht et al., 1984; Latek et al., 2019; Lu et al., 2022

Functions	Toxin studies	Reference
Glucagon-like peptide receptor (GLP1R)		
Regulation of blood glucose levels by triggering insulin release and inhibiting glucagon secretion. It also displays neuroprotective effects in animal models.	Exendin-4, venom peptide produced by the salivary glands of the Gila monster (*Heloderma suspectum*). Act as agonist of GLP1R, stimulating insulin synthesis. Synthetic version Exenatide ER approved for the treatment of type 2 diabetes mellitus.	Van Baelen et al., 2022; Syed et al., 2015.
Neuromedin U receptor (NMUR2)		
Regulation of smooth-muscle contraction, blood pressure and local blood flow, ion transport, stress responses, cancer, gastric acid secretion, pronociception and feeding behaviour.	-	Brighton et al., 2004; Malendowicz and Rucinski, 2021
Melanocortin receptor (MC1R)		
Interfere in different physiological functions: obesity, inflammation, sexual function, cardiovascular tone and steroidogenesis. MC1R, in particular, regulates pigmentation and protection against UV damage.	Spider (*Poecilotheria regalis*) and scorpion (*Parabuthus transvaalicus*) toxins exhibit agonist activity on MC1R.	Reynaud et al., 2020; Van Baelen et al., 2022
Angiotensin II receptor (AGTR2)		
Control blood pressure and regulate cardiovascular function. Potential therapeutic target for glioblastoma through the inhibition of AGTR2.	A-CTX-cMila, a toxin from the cone snail *Conus miliaris*, exhibits antagonist activity on AGTR1.	Singh and Karnik, 2016; Perryman et al., 2022; Van Baelen et al., 2023
Lysophosphatidic acid receptor (LPAR1)		
Mediation of several pathophysiological effects, such platelet aggregation; endothelial hyperpermeability, pro-inflammatory responses and tumour progression.	-	Rivera et al., 2015
Neurotensin receptor (NTSR2)		
Regulates neurodevelopment and neurodegeneration, body temperature, gastrointestinal motility. Modulates pain signal transmission and perception, pituitary hormone secretion and inflammatory responses.	Contulakin-G, a venom peptide from *Conus geographus*, acts as agonist on human and non-human neurotensin receptors (human NTSR1, rat NTSR2, rat NTSR3).	Craig et al., 1999; Kleczkowska et al., 2013; Van Baelen et al., 2022
Arginine vasopressin receptor (AVPR2)		
Regulates fluid homeostasis in renal structures, promoting water retention and facilitating the urine concentration mechanism	Mambaquaretin-1, venom peptide from the snake *Dendroaspis angusticeps* shows high selective antagonist activity on AVPR2.	Juul et al., 2014; Ciolek et al., 2017; Van Baelen et al., 2022

[-] Data not available/unknown.

5.3.3 Cell culturing

Human embryonic kidney cells (HEK293T), provided, as courtesy, by Dr. Bryan Roth, stably expressing β-arrestin2-TEV and tTA-driven luciferase (HTLA cells) were grown in Dulbecco's modified Eagle's medium (DMEM) purchased from Sigma-Aldrich, (St. Louis, MO, USA) supplemented with 10% (v/v) fetal bovine serum (Sigma-Aldrich), 4 mM L-glutamine (Sigma-Aldrich), 1 mM sodium pyruvate (Sigma-Aldrich), 100U/100μg mL^{-1} penicillin/streptomycin (Sigma-Aldrich), 2.5 μg mL^{-1} puromycin (Invitrogen) and 50 μg mL^{-1} hygromycin (Sigma-Aldrich). After a confluency of 80−90% cells were dissociated with trypsin-EDTA solution (Sigma-Aldrich). Cell cultures were incubated and maintained at 37°C in a humidified atmosphere containing 5% (v/v) CO_2.

5.3.4 β-arrestin (Tango) recruitment assay

The activation of GPCRs was measured through the β-arrestin (PRESTO-Tango) recruitment assay as described previously (Zeghal et al., 2020). In brief, cells were seeded into poly-L-lysine hydrobromide (Sigma-Aldrich) treated 96-well plates (35 000 cells/well) and incubated for 24 h before transfection with the plasmid mixture (175 ng/well). The plasmid pcDNA™3.1/Zeo$^{(+)}$ and melatonin were used as negative and positive controls, respectively. The plasmid mixtures were transfected using the TransIT®-LT1 transfection reagent (Mirus Bio, Madison, WI, USA) and incubated for 24 h. Cells were then exposed to different concentrations (from 0 to 1.3 mg mL^{-1}) of crude protein extracts of the ASG and PSG of *O. vulgaris*, and the PSG of *S. officinalis* for a 24-h period. During this period, cells were maintained in starving medium, specifically, DMEM supplemented with 1% (v/v) dialysed fetal bovine serum (Thermo Fisher Scientific). A control composed of starving media and PBS was employed for each GPCR. Afterwards, the cells were lysed and luciferase activity was measured for determination of receptor activation.

5.3.5 Cell viability

Cytotoxic effects produced by the crude protein extracts of the cephalopods salivary glands were assessed by monitoring the HTLA cell metabolism and cell membrane integrity with the resazurin reduction and the 5-carboxyfluorescein diacetate acetoxymethyl ester (CFDA-AM) assays, respectively, as described previously by Pérez-Albaladejo et al. (2016). HTLA cells were incubated for 24 h with successive dilutions of each protein extract (0 to 1.3 mg mL^{-1}).

Fluorescence was measured using a EnSpire 2300 Multilabel microplate reader (PerkinElmer, Waltham, Massachusetts, USA), at 530/590 nm and at 485/530 nm for the cell metabolism and the cell membrane integrity assays, respectively. The assays were performed in triplicate. Each experimental replicate (n = 3) was performed with three technical replicates. 1% Triton X-100 was used in all assays as positive control for reduced cell viability. Results are expressed as percentage change in cell viability relative to control.

5.3.6 Statistics

Homoscedasticity and normality of data were evaluated using Levene's and Shapiro-Wilk's tests, respectively. After the invalidation of at least one of these assumptions for parametric tests, the non-parametric Kruskal-Wallis H test was employed, followed by Dunn's test for multiple comparisons. A significance level $\alpha = 0.05$ was set for all analyses. Statistics were computed using R (Ihaka and Gentleman, 1996), version 4.*x*.

5.4 Results

Cytotoxicity and GPCRs activation were determined in HTLA cells after a 24-h period of exposure to crude protein extracts of the ASG and PSG of *O. vulgaris* and the PSG of *S. officinalis*. The viability of HTLA cells, as assessed by the cell metabolism and the cell membrane integrity assays, was consistently above 80% (in comparison to control) after exposure to increasing concentrations of the cephalopod salivary gland extracts (Figure 16). However, a non-significant decrease of the cell viability percentage to 56%, as measured by the cell membrane integrity assay, was recorded in cells after exposure to the highest concentration (1.3 mg mL^{-1}) of the *O. vulgaris* ASG crude protein extract.

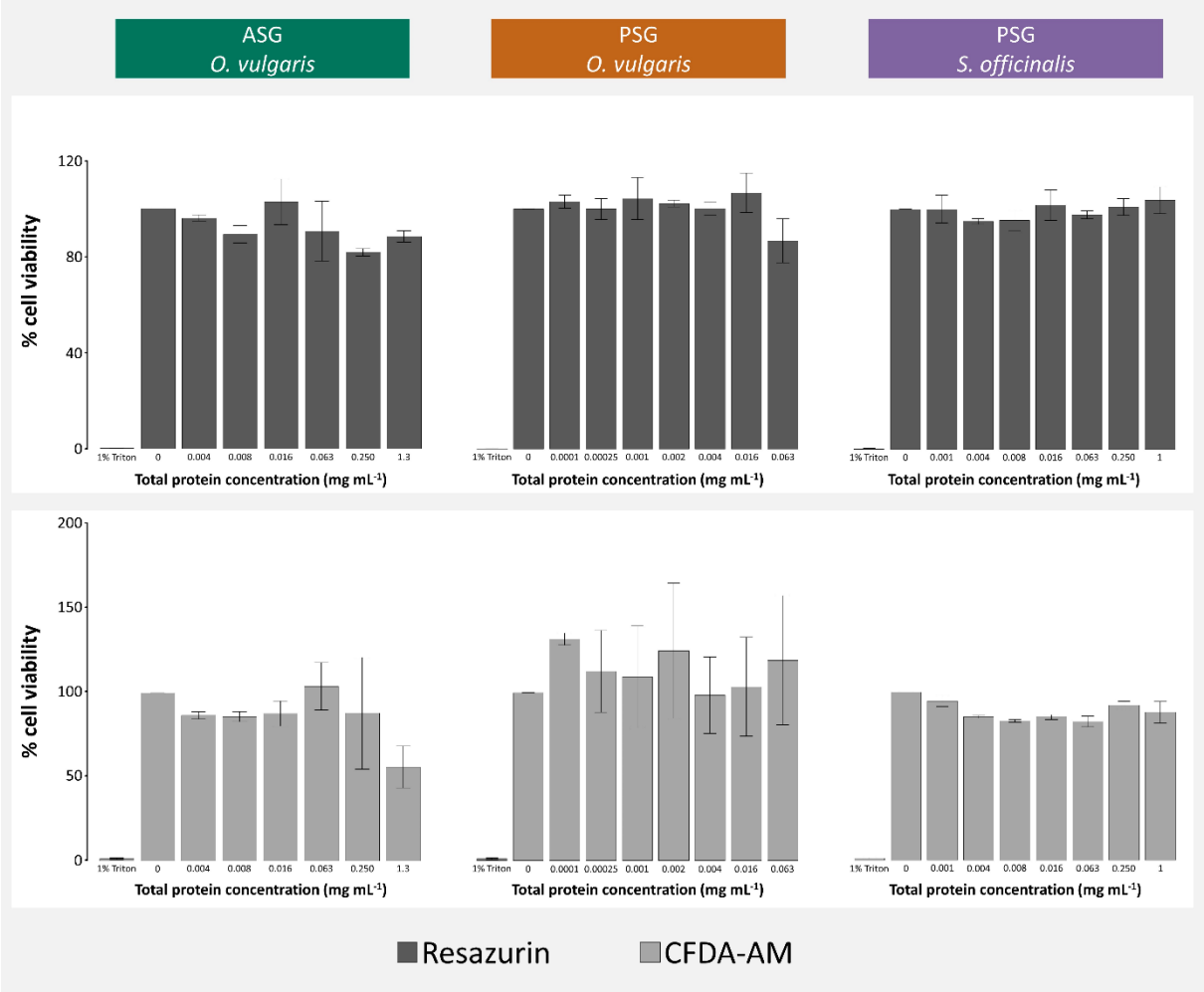


Figure 16 – Viability of HTLA cells after exposure to crude protein extracts from the salivary glands of *Octopus vulgaris* and *Sepia officinalis*. Cells were exposed to different concentrations of each extract for 24 hours as indicated. The metabolic activity and the cell membrane integrity were assessed through the resazurin and the CFDA-AM assays, respectively. Triton X-100 (1%) was used as positive control of reduced cell viability. Data is presented as mean values ± SEM.

A significant dose-dependent activation of several GPCRs (Dunn's test $p < 0.05$) was observed following exposure of HTLA cells to the three toxin-bearing extracts of cephalopods (Figure 17). Cells exposed to the *Octopus vulgaris* PSG extracts exhibited different responses from those exposed to the *Sepia* PSG extract. Specifically, the HTLA cells exposed to the highest concentration of *Octopus*' PSG extract (0.063 mg ml^{-1}), for instance, detained a fold increase of 2.5, 1.9 and 1.5 for the angiotensin II receptor (AGTR2), the glucagon-like peptide receptor (GLP1R) and the melanocortin receptor (MC1R), respectively (Figure 17A). The neuromedin U receptor (NMUR2) and the arginine vasopressin receptor (AVPR2) obtained lower fold

activation responses of 1.1 and 1.3, respectively. Similar responses were registered for the vasoactive intestinal peptide (VIPR1) and the lysophosphatidic acid (LPAR1) receptors, where a maximal fold activation of 1.2 were attained for the second highest tested concentration (0.016 mg mL^{-1}), however these receptors showed a decreasing activation trend, along with the neurotensin receptor (NTSR2). Conversely, the PSG extract from *Sepia officinalis* (Figure 17B) activated all the GPCRs, with the maximal and significant fold activation being registered for the highest concentrations (0.250 and 1 mg mL^{-1}). Despite being significantly activated, the AVPR2, LPAR1 and NTSR2 receptors registered lower activation responses in comparison with the remain activated receptors. The latter, namely, the MC1R and the AGTR2 receptors had a maximal fold activation of 3.2 and 3.4, respectively for the highest concentration (1 mg mL^{-1}), whereas the VIPR1 did not exceed 2.9-fold activation at the highest concentration (1 mg mL^{-1}). The GLP1R and the NMUR2 receptors demonstrated also a significant activation of up to 2.4-fold and 2.2-fold, respectively. In contrast, the HTLA cells exposed to the ASG extract, had a significant (Dunn's test $p < 0.05$) decrease in luciferase activity for all GPCRs in all tested concentrations (Figure 17C).

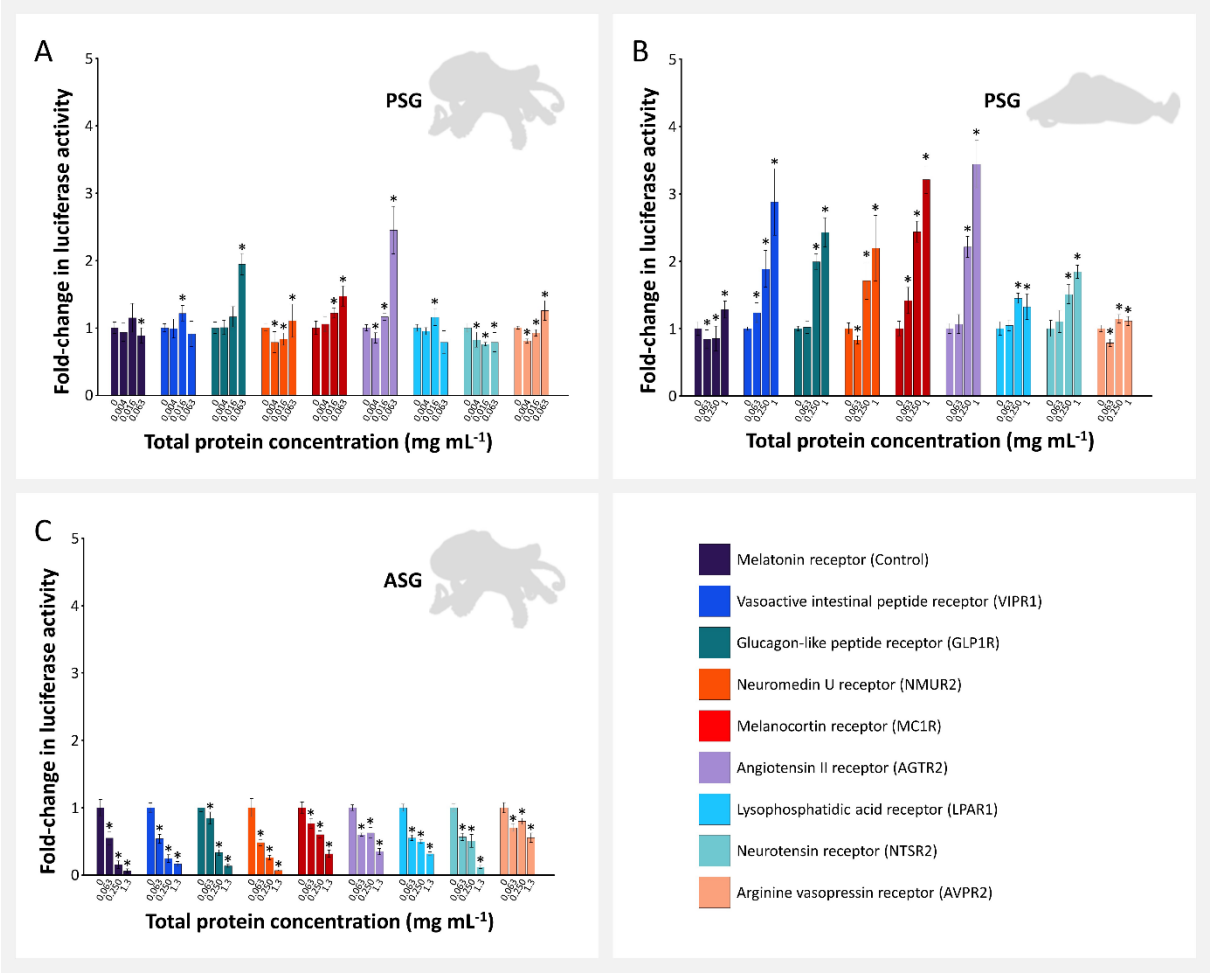


Figure 17 – Responses of G protein–coupled receptors (GPCRs) activity in HTLA cells exposed to extracts from anterior (ASG) and posterior (PSG) salivary glands of *Octopus vulgaris* (**A** and **C**) and the PSG extract of *Sepia officinalis* **(B)**. Cells were exposed to different concentrations of each extract for 24 hours as indicated. Responses in exposed cells are presented as mean fold change in luciferase activity ± SEM. Asterisks indicate significant differences between the control (0 mg ml^{-1}) of each GPCRs (Dunn's test $p < 0.05$).

5.5 Discussion

The salivary gland extracts of *Octopus vulgaris* and *Sepia officinalis* were investigated for bearing potential GPCR ligands for bioprospection. For this purpose, the cytotoxicity and the modulatory activity of these extracts were assessed with eight GPCRs of interest, including receptors that regulate blood pressure, cardiovascular function, vasoconstriction, muscle function, blood glucose levels, fluid homeostasis and inflammatory responses, via a modified Tango β-arrestin recruitment assay in a human cell line. The results showed significant

activation responses for GPCRs in cells exposed to the crude protein extracts from the posterior salivary glands (PSG) of *Octopus* and *Sepia*, in comparison with the *Octopus* anterior salivary gland (ASG). Most importantly, such was achieved without significant cytotoxicity (i.e., unhindered cell viability), which is an important outcome to safeguard efficacy and safety as prospect therapeutics. Nevertheless, direct comparison with other studies is challenging because research on the discovery of new GPCR targets in crude animal venoms is lacking, more even considering venoms from marine invertebrates. It must be noted that the *modus operandi* of studies to discover new ligands for GPCRs involves purification of single venom peptides, which of course gives detailed information regarding the biochemical properties and the structure of the testing peptide, but it makes the analyses more demanding in terms of time and effort (Amiranoff et al., 1983; Robberecht et al., 1984; Craig et al., 1999; Ciolek et al., 2017; Reynaud et al., 2020).

Testing crude protein extracts from cephalopods allowed an initial screening for responses rather than depending on the stochasticity associated to a single molecule when looking for agonists or antagonists of GPCRs. Indeed, the highest agonistic activity was reported for the angiotensin II (AGTR2) and the glucagon-like peptide (GLP1R) receptors following exposure to PSG extracts from both *Octopus* and *Sepia*. The activation of these receptors, whose functions are linked to the regulation of blood pressure, cardiovascular function and blood glucose levels, is in accordance with our previous studies on the transcriptomes of the PSG of both *Octopus vulgaris* and *Sepia officinalis* (Gonçalves et al., 2024; Chapter 4). Specifically, we revealed that this gland, also known as the 'cephalopod venom gland', secretes high diversity of bioactives ranging from neurotoxins to, in the case of *Sepia*, venom components with endocrine disrupting effects, namely insulin-like peptides. These latter have been found in the venoms of other molluscs, presumably to stun prey by inducing hypoglycaemic shock (Safavi-Hemami et al., 2015). The secretion of hormone-like toxins with insulin domains could be a possible explanation to the activation of the GLP1R, since this receptor is involved in glucose homeostasis, by enhancing insulin secretion upon activation. Even though these hormone-like toxins could be cephalopod-specific, previous studies on the insulin-like peptides from cone snails (such as *C. geographus* insulins 1 and 2) injected in zebrafish elicited physiological responses similar to the GLP1R activation, as a decrease in the blood glucose levels was registered in treated fishes (Safavi-Hemami et al., 2015). Although our homology-matching analyses based on transcriptomics of the PSG of *Octopus* did not reveal any insulin-like hormone, the results in the present study showed higher receptor affinity for this particular

species in comparison to the cuttlefish, which could be verified by higher fold-activation of the highest concentration of PSG extracts of *Octopus* (0.063 mg mL^{-1}), comparatively to the same concentration in *Sepia*. This highlights the need to keep conducting further research in this area. However, homology-matching is cumbersome in species with reduced genomic annotation, potentially leading to gaps in the identification of toxins (see Moutinho Cabral et al. (2024), for a review).

Besides the secretion of insulin, GLP1R also modulates cardiovascular function, similarly to AGTR2, which exhibited here the highest fold-activation in PSG extracts of both species, comparatively to the other tested GPCRs. Receptor AGTR2 was indeed activated by both PSG extracts, however, that from *Sepia* required higher concentrations to activate it, indicating that the secretions from the venom gland of *Octopus* may contain more specific ligands for this receptor. Besides the role on cardiovascular function, further research on AGTR2 functions have also identified that it controls various biological processes involved in development, cell differentiation, tissue repair and even apoptosis (for a review, see Kaschina and Unger, 2003). Accordingly, we found transcripts encoding for proteins linked to cell growth stimulation and differentiation in the *Octopus* PSG, such as the epidermal growth factor-related protein 1. Even though the identification of ligands for this specific receptor is still in its infancy, the recent study of Van Baelen et al. (2023) identified the first animal toxin that is agonist of angiotensin II receptors. After purifying the peptide A-CTX-cMila, a toxin from the cone snail *Conus miliaris*, and measuring the G-protein activation and β-arrestin 2 recruitment using bioluminescence resonance energy transfer-based biosensors, the same authors unravelled that this toxin has antagonist activity on AGTR1. However, AGTR2 continues to lack ligands, even though it has already been recognised as a potential therapeutic target for glioblastoma.

Comparatively, cells exposed to *Sepia* PSG crude protein extract presented higher fold-activation responses in GPCRs associated with the modulation of both smooth-muscle contraction and relaxation, namely the neuromedin U receptor 2 (NMUR2) and the vasoactive intestinal peptide receptor (VIPR1), respectively. Similarly to *Octopus*, *S. officinalis* is known to secrete effective neurotoxins, including cephalotoxins, that act as immobilisers on crustaceans and therefore, affect the prey neuromuscular system (Ghiretti, 1959). Although both PSG extracts hold toxin transcripts, our transcriptomic analyses reported higher number of cephalotoxin homologs in *Sepia*, comparatively to *Octopus*, which may have increased agonist responses in cells exposed to the PSG extract of the cuttlefish.

The melanocortin receptor (MC1R) is also of great importance as therapeutic target, in this case against melanoma, since this particular receptor regulates the production of eumelanin, a UV radiation-absorbing pigment (Kobayashi et al., 1998). In this work, the exposure of cells to PSG extracts of *O. vulgaris* and *S. officinalis* resulted in the activation of the MC1R, however with a higher response for the PSG extract from *Sepia*. Recently, Reynaud et al. (2020) discovered two toxins, one form from the spider *Poecilotheria regalis* (N-TRTX-Preg1a) and another from the scorpion *Parabuthus transvaalicus* (N-BUTX-Ptr1a), that exhibited agonist activity on MC1R. The experiments involved recombinant expression of both toxins using *E. coli* and subsequent receptor activation, which was assessed through a luciferase assay after 3h of cell stimulation. In addition, previous authors identified an inhibitory cystine knot motif in the spider toxin and the CS$\alpha\beta$ structure in a toxin from the same scorpion, which can be interpreted as these toxins are both cysteine-rich peptides (Craik et al., 2001). These latter findings are in accordance with our preceding transcriptomics approach, which revealed the presence of several cysteine-rich venom proteins in PSG of both *O. vulgaris* and *S. officinalis*, three of which were isolated and validated. Although the higher agonist response was registered for the PSG extract from *Sepia*, it is important to note that similar activation responses were attained for these two extracts at the concentration of 0.063 mg mL^{-1}.

Albeit with lower fold-activation response, the PSG extract of *Sepia* exhibited significant agonistic properties in receptors which functions are, among others, related to pro-inflammatory responses, such as the lysophosphatidic acid receptor (LPAR1) and neurotensin receptor (NTSR2), as well as regulation of fluid homeostasis in renal structures, such as the arginine vasopressin receptor (AVPR2). These lower responses should indicate reduced affinity of constituents of the PSG venomous cocktail of both species for these receptors. However, the PSG extract from *Sepia* revealed a slight increase of the fold-activation response in both LPAR1 and NTSR2 at the highest concentration (1 mg mL^{-1}), comparatively to the highest concentration of the PSG extract from *Octopus* (0.063 mg mL^{-1}), which can indicate that the agonist activity could be dose-dependent. Previous studies on human and non-human neurotensin receptors already revealed that the venom peptide contulakin-G extracted and purified from the venom gland duct of the *Conus geographus*, act as agonist on the human NTSR1, however, for the NTSR2, the agonist activity could only be recorded for the rat neurotensin receptor (Craig et al., 1999). In the case of the AVPR2, antagonist activity was verified with the mambaquaretin-1, a venom peptide from the snake *Dendroaspis angusticeps*, through the study of several pathways, such as cAMP production, β-arrestin interaction, and

MAP kinase activity (Ciolek et al., 2017). Further investigations of these authors revealed that, after injected in mice, mambaquaretin-1 inhibits renal cysts progression without significant toxicity, being a promising therapeutic target against polycystic kidney diseases.

Despite our previous transcriptomic data on the ASG of *O. vulgaris* had revealed the presence of proteins with neurotoxic effects (i.e., ShkT domain-containing proteins), the results of the present study may indicate either no interaction with the receptors or inhibition of activity. In addition to the large number of matrix-permeabilising enzymes, capable of degrading chitin and lipids in prey tissues, this gland also holds protein functionalisation roles, which could explain the lack of specific ligands for these receptors. This gland also contains proteins that inhibit the activation of essential enzymes for blood clotting and platelet functions, which may explain the lack of activation of, for example, LPAR1, which is linked to the mediation of platelet aggregation. Altogether, even though both salivary glands hold significant variety of bioactive compounds it is the PSG that secretes and conveys the majority of compounds capable of interfering with GPCRs. However, even though the *Sepia* ASG could not be retrieved because it is deeply embedded in the buccal mass, this same gland in other cephalopods may be an interesting target organ for the bioprospecting for mucins and molecule stabilisers, protectors and artefacts activity.

5.6 Conclusions

Despite the immense potential of therapeutics targeting GPCRs, the identification of potential ligands within the vast biochemical diversity of marine animal toxins is still lagging. We demonstrated that cephalopods may become a major contribution to this field since their saliva has been found to be composed of bioactives with agonistic properties for specific human GPCRs, without significant cytotoxic effects on human HTLA cells (determined as cell viability), at least under the tested dilutions of crude protein extracts. The results from the present study were compared with our previous results from transcriptomics on the salivary glands of both *O. vulgaris* and *S. officinalis*. Overall, the PSG extracts from either species revealed higher affinity for the shortlisted GPCRs (selected for their relevance as targets for animal toxins), compared to the ASG extract of *Octopus*. The clear distinction of the interaction between GPCRs and the extracts from the 'venom glands' *versus* the 'mucous gland' validates our previous results on the gland's transcriptomes. Whereas the PSG (the "venom gland of cephalopods") holds a wider range of bioactives, the ASG contains inhibitory and

permeabilising enzymes and is more dedicated to the functionalisation of proteins, which resulted in absent interaction with the tested GPCRs, including potential inhibition of their activity. Comparing the two species, we found that the octopus possess more venom peptides with higher affinity with the GPCRs, since even a higher dilution of the PSG extract was able to activate with greater expression two out of the eight receptors studied, whereas are included receptors that regulate blood pressure, cardiovascular function and blood glucose levels. The cuttlefish, on the other hand, showed greater diversity of ligands, with the activation of most of the tested GPCRs, with particular emphasis on receptors that modulate smooth muscles relaxation and contraction, regulate immune response and ion transport. Despite the need for further analyses, this was one of the first studies with crude protein extracts from cephalopods as an attempt to search for ligands for GPCRs and aid in bioprospecting.

6

FINAL CONSIDERATIONS

6.1 Concluding remarks

The immense biodiversity of the oceans is reflected in its vast biomolecular diversity of marine bioactives, which have long been highlighted for their biotechnological potential on various applications ranging from novel therapeutical alternatives and even ecologically friendly anti-foulants and pesticides. Toxins, in particular, are found in the crosshairs of biotechnologists as targets for bioprospecting due to their bioreactivity, potential specificity potency and, especially if peptidic, relative simplicity to modify and produce *in vitro*. These are effectively novel resources for biotechnologists that have been optimised by evolutionary processes. They may offer greener, safer and more cost-effective alternatives to the risky and laborious process of designing of synthetic compounds for the development of new drugs, for instance. However, the process spanning from marine bioprospecting (in which most of this thesis falls within) to pre-clinical trials entails several challenges. To begin with, researchers must shortlist the immeasurable diversity of the oceans according to the needs of industry and society. Then there is the problem of reduced genomic resources for the vast majority of marine species, besides chronic knowledge gaps on their anatomy, physiology and often even ecology. Tacking all these challenges and attempting to construct roadmaps first to species of interest, then to target molecules is essential for sustainable and responsible marine bioprospecting.

The present work aimed to understand the ecophysiology, evolution and toxicology of cephalotoxins and other coleoid toxins to ascertain the potential of the common octopus and the common cuttlefish as providers of high-value bioactives and to predict their molecular targets as a means to indicate potential biotechnological applications. It must be emphasised that the common octopus and the common cuttlefish, both common and high-value species in Portuguese waters, were chosen as representatives of coleoids secreting putative bioactives of interest in posterior salivary glands, which are commonly referred to as 'the cephalopod venom gland'. With this particular respect, the current thesis is also a contribution to shift the exploitation of marine resources in Portugal from fishery, which still stands as a traditional economic enterprise, to sustainable and more innovative bioprospecting.

The present work demonstrated that ecology and ecophysiology are essential landmarks in the roadmap to find bioactives of interest. Targeting toxin-secreting organisms is an example strategy to shortlist biodiversity. Then, understanding the adaptive value of crude secretions and individual compounds (here inferred with the scaffold of receptor binding assays and

transcriptomics), combined with knowledge on the animals' ecology and even the secreting organ's microanatomy indicated specialisation and therefore specificity and function of the venom. However, the detailed information on the morphoanatomy of the toxin-delivery apparatus is often absent among marine animals, especially invertebrates. Here, it became a crucial step for understanding how, where and why these bioactives are produced. The comparative description of *Octopus* and *Sepia* salivary glands enabled us to infer about the differences of the proteinaceous content between gland and species. Thus, the *Sepia* PSG and the *Octopus* ASG are anatomically simpler, than the *Octopus* PSG, which hold structural complexity by bearing two different epithelia. Altogether, microanatomy alone showed that the two species secrete different venoms, which makes them both interesting targets for bioprospecting.

The transcriptomic pipeline confirmed the high diversity of bioactives secreted by the salivary glands of both species, even though the *Sepia* ASG could not be analysed as it is deeply embedded into the buccal mass and could not be effectively isolated. Despite the absence of transcriptomic data on ASG of the cuttlefish, which again highlights the difficulties of working with marine invertebrates even if relatively well-known, the results showed in both species the presence of venom components with potential neurotoxic and endocrine-disrupting effects and proteins with inhibitory functions. In addition, peptides and proteins with proteolytic properties (that in a venom seemingly act as permeabilisers) were also found. The transcriptomic results also shed some light about potential challenges of these compounds in future steps along the biotechnology pipeline, since several putative proteins were identified as potentially responsible for *post*-translational modifications. This is an important finding, as they permit better *a priori* choice of model, since these modifications may hinder heterologous expression in prokaryote and perhaps even yeast models. The set of validated full coding sequences of several bioactive proteins that are venom components, namely three cysteine-rich venom proteins, a venom insulin, a cephalotoxin, a chitinase and a hyaluronidase, is also an important biotechnological asset.

Assisted by phylogenetic models, we produced an image of the homology between the full coding sequences of the seven abovementioned venom bioactives secreted by the target species. These exhibit different patterns of sequence variation within venomous taxa, depending on the type of bioactive. Permeabilising enzymes and venom insulins, for instance, have a certain level of conservation between Cephalopoda, whereas cephalotoxins and other venom proteins rich in cysteines show some variability between different cephalopods and

also differ considerably from their homologs in cnidarians and even other molluscs. Venom insulin, for instance, is a particularly interesting target following the recent discoveries in Cone snail venom insulin, which is more potent than human insulin and evolved to stun prey by glycaemic shock. This represents a novel path for coleoid venom components and is a potentially promising biotechnology area of research. Combined with the findings from the identification of new specific ligands of G protein–coupled receptors (GPCRs), these results show that cuttlefish venom can be able to interact with a wider range of receptors, however the venom of the octopus can hold substances that are more potent interacting agents with some of the receptors (such as AGTR2 and GLP1R) without significant cytotoxic effects on human cells. Although additional research is required, these secretions proved to interact with several GPCRs of interest, of which we highlight receptors that control blood pressure and regulate both cardiovascular function and blood glucose levels. This GPCR interaction thus emphasises the advantage to use venom-derived peptides to expedite the drug-discovery process by increasing the efficacy of new pharmacological drugs.

Altogether, this work revealed the potential of exploring common toxin-bearing animals from marine sources to uncover novel biotechnological applications. Through the integration of ecophysiology, molecular biology, bioinformatics and toxicology this work contributed to unravel the diversity of bioactives from these species, of which we emphasise cephalotoxins, cysteine-rich venom proteins, permeabilising enzymes and pro-insulins, whose study still needs to be deepened. This work also contributed to increasing awareness to preserve marine resources as they comprise the widest biodiversity in promising substances with interest for novel applications, thus in accordance with the European Union's ambitious prospects for leading the 'Blue Growth Strategy' in the near future.

6.2 Future perspectives

The richness of venom proteins and peptides that were identified in the salivary glands of *Octopus* and *Sepia* in this work, represents the high potential of toxin-secreting marine invertebrates for blue biotechnology. Most importantly, the outcomes show that even cosmopolitan organisms from temperate waters like *Sepia* and *Octopus* can be a source of bioactives that can interfere with specific molecular pathways, from neuronal activity to glycaemia and a range of nuclear receptor-mediated processes. The next steps along the biotechnological pipeline towards effective drug development from these bioactives must

involve the complete characterisation of the venoms, as venom components produce a combined outcome, which may involve integration of multiple omics, especially proteomics and un-targeted metabolomics (as not all venom components are peptidic). Most importantly, the present work produced the full coding sequences of relevant bioactives whose production can be conducted using DNA recombinant methods, for which the most cost-effective models must be selected, from *E. coli* to animal cells. These methods must be attempted and adapted for optimal activity, safety and eventual the scaling up of bioactive production *in vitro* to become compliant with bioindustry. Evidently such critical steps must be followed by purification and testing for toxicity and bioactivity of each protein, ultimately seeking dose- and time-responsiveness. The activity and functionality of recombinant proteins of interest should then be ascertained, to gain a comprehensive understanding of how these proteins affect cellular processes. This includes refining research of target receptors as privileged drug targets, using, e.g., GPCR methods or immune co-precipitation of ligand-receptor heterodimers for a more thorough generalisation of the approach in human druggable proteome. Evidently non-human targets can also be of importance to find environmental applications of these products, for instance, as natural pesticides or anti-foulants in maritime paints.

7

REFERENCES

Albertin, C. B. and Simakov, O. (2020). Cephalopod biology: At the intersection between genomic and organismal novelties. Annu. Rev. Anim. Biosci. 8, 71–90. doi: 10.1146/annurev-animal-021419-083609

Alcaide, M., Moutinho Cabral, I., Carvalho, L., Mendes, V.M., Alves de Matos, A.P., Manadas, B., Saúde, L., D'Ambrosio, M. and Costa, P.M. (2024). A comparative analysis of the venom system between two morphotypes of the sea anemone *Actinia equina*. Animals (Basel). 14, 981. doi: 10.3390/ani14060981.

Almeida, D., Domínguez-Pérez, D., Matos, A., Agüero-Chapin, G., Osório, H., Vasconcelos, V., Campos, A. and Antunes, A. (2020). Putative antimicrobial peptides of the posterior salivary glands from the cephalopod *Octopus vulgaris* revealed by exploring a composite protein database. Antibiotics. 9(11), 757. doi: 10.3390/antibiotics9110757

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Ambrosino, L., Tangherlini, M., Colantuono, C., Esposito, A., Sangiovanni, M., Miralto, M. et al. (2019). Bioinformatics for marine products: an overview of resources, bottlenecks, and perspectives. Mar. Drugs 17, 576. doi: 10.3390/md17100576

Amiranoff, B., Vauclin-Jacques, N., Boige, N., Rouyer-Fessard, C. and Laburthe, M. (1983). Interaction of gila monster venom with VIP receptors in intestinal epithelium of human. A comparison with rat. FEBS Lett. 164(2), 299–302. doi: 10.1016/0014-5793(83)80305-6

Anadón, R. (2019). Functional histology: The tissues of common coleoid cephalopods. In: Gestal, C., Pascual, S., Guerra, Á., Fiorito, G., Vieites, J. (eds) Handbook of Pathogens and Diseases in Cephalopods. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-11330-8_4

Avella, I., Calvete, J.J., Sanz, L., Wüster, W., Licata, F., Quesada-Bernat, S., et al. (2022). Interpopulational variation and ontogenetic shift in the venom composition of Lataste's viper (*Vipera latastei*, Boscá 1878) from northern Portugal. J. Proteomics 263, 104613. doi: 10.1016/j.jprot.2022.104613

Baden, T., Briseño, J., Coffing, G., Cohen-Bodénès, S., Courtney, A., Dickerson, D., et al. (2023). Cephalopod-omics: Emerging fields and technologies in cephalopod biology. Integr. Comp. Biol. 63, 1226–1239 doi: 10.1093/icb/icad087

Bates, S.M. and Weitz, J.I. (1998). Direct thrombin inhibitors for treatment of arterial thrombosis: Potential differences between bivalirudin and hirudin. Am. J. Cardiol. 82, 12–18. doi: https://doi.org/10.1016/S0002-9149(98)00660-2

Bordon, K.C.F., Cologna, C.T., Fornari-Baldo, E.C., Pinheiro-Júnior, E.L., Cerni, F.A., Amorim, F.G., et al. (2020). From animal poisons and venoms to medicines: achievements, challenges and perspectives in drug discovery. Front. Pharmacol. 11:1132. doi: 10.3389/fphar.2020.01132

Bottazzi, F. and Valentini, V. (1924). Nuove ricerche sul veleno della "saliva" di *Octopus macropus*. Arch. Sci. Biol. 6, 153–168.

Boucaud-Camou, E. (1968). Etude histologique et histochimique de l'appareil digestif de Sepiola atlantica d'Orbigny et *Sepia officinalis* L. Bull. Soc. Linn. Normandie 9, 220–243.

Boucaud-Camou, E. and Boucher-Rodoni, R. (1983). "Feeding and digestion in cephalopods," in The Mollusca Physiology Part 2, Vol. 5, eds A.S.M. Saleuddin and K.M. Wilburk (New York, NY: Academic Press), 149–187. doi: 10.1016/b978-0-12-751405-5.50011-7

Bowersox, S.S., Gadbois, T., Singh, T., Pettus, M., Wang, Y.X. and Luther, R.R. (1996). Selective N-type neuronal voltage-sensitive calcium channel blocker, SNX-111, produces spinal antinociception in rat models of acute, persistent and neuropathic pain. J. Pharmacol. Exp. Ther. 279, 1243–1249.

Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527. doi: 10.1038/nbt.3519

Brighton, P.J., Szekeres, P.G. and Willars, G.B. (2004). Neuromedin U and its receptors: Structure, function, and physiological roles. Pharmacol. Rev. 56(2), 231–48. doi: 10.1124/pr.56.2.3

Budelmann, B.U., Schipp, R. and von Boletzky, S. (1997). "Cephalopoda," in Microscopic Anatomy of Invertebrates Mollusca II, Vol. 6A, eds F.W. Harrison and A.J. Kohn (New York, NY: Wiley-Liss), 119–414.

Butler, M.S. (2005). Natural products to drugs: natural product derived compounds in clinical trials. Nat. Prod. Rep. 22, 162–195. doi: 10.1039/b402985m

Calvete, J.J., Sanz, L., Angulo, Y., Lomonte, B. and Gutiérrez, J.M. (2009). Venoms, venomics, antivenomics. FEBS Lett. 583, 1736–1743. doi: 10.1016/j.febslet.2009.03.029

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2008). BLAST+: architecture and applications. BMC Bioinform. 10, 421. doi: 10.1186/1471–2105-10–421

Campos, S., Rodrigo, A.P., Moutinho Cabral, I., Mendes, V.M., Manadas, B., D'Ambrosio, M. and Costa, P.M. (2023). An exploration of novel bioactives from the venomous marine annelid *Glycera alba*. Toxins. *15*, 655. doi: https://doi.org/10.3390/toxins15110655

Cappello, E. and Nieri, P. (2021). From life in the sea to the clinic: The marine drugs approved and under clinical trial. Life 11, 1390. doi: 10.3390/life11121390

Cariello, L. and Zanetti, L. (1977). α- and β-cephalotoxin: two paralysing proteins from posterior salivary glands of *Octopus vulgaris*. Comp. Biochem. Physiol. 57, 169–173. doi: 10.1016/0306-4492(77)90066-1

Caruana, N.J., Cooke, I.R., Faou, P., Finn, J., Hall, N.E., Norman, M., et al. (2016). A combined proteomic and transcriptomic analysis of slime secreted by the southern bottletail squid, *Sepiadarium austrinum* (Cephalopoda). J. Proteomics 148, 170–182. doi: 10.1016/j.jprot.2016.07.026

Casewell, N.R., Jackson, T.N.W., Laustsen, A.H. and Sunagar, K. (2020). Causes and consequences of snake venom variation. Trends Pharmacol. Sci. 41, 570–581. doi: 10.1016/j.tips.2020.05.006

Castañeda, O., Sotolongo, V., Amor, A.M., Stöcklin, R., Anderson, A.J., Harvey, A.L., Engström, A., Wernstedt, C. and Karlsson, E. (1995). Characterization of a potassium channel toxin from the Caribbean Sea anemone *Stichodactyla helianthus*. Toxicon. 33, 603–13. doi: 10.1016/0041-0101(95)00013-c

Castro e Silva, J., Lopes do Couto, L., de Oliveira Amaral, H., Gomes, F.M.M., Campos, G.A.A., Silva, L.P., et al. (2020). Neuropolybin: a new antiseizure peptide obtained from wasp venom. Biochem. Pharmacol. 181:114119. doi: 10.1016/j.bcp.2020.114119

Charif, D. and Lobry, J. R. (2007). "SeqinR 1.0–2: A contributed package to the R Project for statistical computing devoted to biological sequences retrieval and analysis," in Structural approaches to sequence evolution: Molecules, networks, populations. Eds. U. Bastolla, M. Porto, H.E. Roman and M. Vendruscolo (Springer-Verlag, Berlin Heidelberg), 207–232.

Chichery, M.P. and Chichery, R. (1988). Manipulative motor activity of the cuttlefish *Sepia officinalis* during prey-capture. Behav. Process 17, 45–56. doi: 10.1016/0376-6357(88)90049-6

Ciolek, J., Reinfrank, H., Quinton, L., Viengchareun, S., Stura, E.A., Vera, L. et al. (2017). Green mamba peptide targets type-2 vasopressin receptor against polycystic kidney disease. Proc. Natl. Acad. Sci. U.S.A. 114(27), 7154–7159. doi: 10.1073/pnas.1620454114

Colwell, R.R. (1983). Biotechnology in the marine sciences. Science. 222, 19–24. doi: 10.1126/science.222.4619.19

Cooke, I.R., Whitelaw, B., Norman, M., Caruana, N. and Strugnell, J.M. (2017). Toxicity in Cephalopods. In: Malhotra, A. (eds) Evolution of Venomous Animals and Their Toxins. Toxinology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6458-3_7

Cornet, V., Henry, J., Corre, E., Le Corguille, G., Zanuttini, B. and Zatylny-Gaudin, C. (2014). Dual role of the cuttlefish salivary proteome in defense and predation. J. Proteomics. 108, 209–222. doi: 10.1016/j.jprot.2014.05.019

Costa, P.M. and Costa, M.H. (2012). Development and application of a novel histological multichrome technique for clam histopathology. J. Invert. Pathol. 110, 411–414. doi: 10.1016/j.jip.2012.04.013

Costa, P.M., Rodrigo, A.P. and Costa, M.H. (2014). Microstructural and histochemical advances on the digestive gland of the common cuttlefish, *Sepia officinalis* L. Zoomorphology. 133, 59–69. doi: 10.1007/s00435-013-0201-8

Craig, A.G., Norberg, T., Griffin, D., Hoeger, C., Akhtar, M., Schmidt, K. et al. (1999). Contulakin-G, an *O*-glycosylated invertebrate neurotensin. J. Biol. Chem. 274(20), 13752–9. doi: 10.1074/jbc.274.20.13752

Craik, D.J., Daly, N.L. and Waine, C. (2001). The cystine knot motif in toxins and implications for drug design. Toxicon. 39(1), 43–60. doi: 10.1016/s0041-0101(00)00160-4

Cuevas, C. and Francesch, A. (2009). Development of Yondelis® (trabectedin, ET-743). a semisynthetic process solves the supply problem. Nat. Prod. Rep. 26, 322–337. doi: 10.1039/b808331m

Cushman, D.W. and Ondetti, M.A. (1991). History of the design of captopril and related inhibitors of angiotensin converting enzyme. Hypertension. 17(4), 589–92. doi: 10.1161/01.hyp.17.4.589

D'Ambrosio, M., Gonçalves, C., Calmão, M., Rodrigues, M. and Costa, P.M. (2021). Localization and bioreactivity of cysteine-rich secretions in the marine gastropod *Nucella lapillus*. Mar. Drugs 19, 276. doi: 10.3390/md19050276

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10, 1–4. doi: 10.1093/gigascience/giab008

De Meyts, P. (2016). Structural basis for the poisonous activity of a predator's venom insulin. Nat. Struc. Mol. Biol. 23, 872–874. doi: 10.1038/nsmb.3304

Domínguez-Pérez, D., Campos, A., Alexei Rodríguez, A., Turkina, M.V., Ribeiro, T., Osorio, H., et al. (2018). Proteomic analyses of the unexplored sea anemone *Bunodactis verrucosa*. Mar. Drugs. 16, 42. doi: 10.3390/md16020042

Ducros, C. (1972). Etude ultrastructurale de l'innervation des glandes salivaires postérieures chez *Octopus vulgaris*. Z. Zellforsch. Mik. Ana. 132, 35–49. doi: https://doi.org/10.1007/bf00310296

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform. 23, 205–211. doi: 10.1142/p715

Eng, J., Kleinman, W.A., Singh, L., Singh, G. and Raufman, J.P. (1992). Isolation and characterization of exendin-4, an exendin-3 analogue, from *Heloderma suspectum* venom. Further evidence for an exendin receptor on dispersed acini from guinea pig pancreas. J. Biol. Chem. 267(11), 7402–5.

Erspamer, V. (1948). Active substances in the posterior salivary glands of octopoda. II. tyramine and octopamine (oxyoctopamine). Acta Pharmacol. Toxicol. 4:224. doi: 10.1111/j.1600-0773.1948.tb03345.x

Erspamer, V. and Anastasi, A. (1962). Structure and pharmacological actions of Eledoisin, the active endecapeptide of the posterior salivary glands of Eledone. Experientia 15, 58–59. doi: 10.1007/bf02138250

Erspamer, V. and Asero, B. (1953). Isolation of enteramine from extracts of posterior salivary glands of *Octopus vulgaris* and of *Discoglossus pictus* skin. J. Biol. Chem. 200, 311–318. doi: 10.1016/s0021-9258(18)38466-7

Fahmy, S.R. and Soliman, A.M. (2013). *In vitro* antioxidant, analgesic and cytotoxic activities of *Sepia officinalis* ink and *Coelatura aegyptiaca* extracts. Afr. J. Pharm. Pharmacol. 7(22), 1512 – 1522. doi: 10.5897/AJPP2013.3564

Fassio, G., Modica, M.V., Mary, L., Zaharias, P., Fedosov, A.E., Gorson, J., et al. (2019). Venom diversity and evolution in the most divergent cone snail genus *Profundiconus*. Toxins 11, 623. doi: 10.3390/toxins11110623

Fedosov, A., Zaharias, P. and Puillandre, N. (2021). A phylogeny-aware approach reveals unexpected venom components in divergent lineages of cone snails. Proc. Biol. Sci. 288, 20211017. doi: 10.1098/rspb.2021.1017

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 39, 783–791. doi: 10.2307/2408678

Fernández-Gago, R., Molist, P. and Rocha, F. (2018). Anatomical and histochemical features of the digestive system of *Octopus vulgaris* Cuvier, 1797 with a special focus on secretory cells. Acta Zool. 100, 1–16. Doi: https://doi.org/10.1111/azo.12257

Fingerhut, L.C.H.W., Strugnell, J.M., Faou, P., Labiaga, Á.R., Zhang, J. and Cooke, I.R. (2018). Shotgun proteomics analysis of saliva and salivary gland tissue from the common octopus *Octopus vulgaris*. J. Proteome Res. 17, 3866–3876. doi: 10.1021/acs.jproteome.8b00525

Fry, B.G., Roelants, K. and Norman, J.A. (2009). Tentacles of venom: toxic protein convergence in the kingdom animalia. J. Mol. Evol. 68, 311–321. doi: 10.1007/s00239-009-9223-8

Fry, B.G., Roelants, K., Champagne, D.E., Scheib, H., Tyndall, J.D., King, G.F. et al. (2009). The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. Annu. Rev. Genomics Hum. Genet. 10, 483–511. doi: 10.1146/annurev.genom.9.081307.164356.

Furman, B.L. (2012). The development of Byetta (exenatide) from the venom of the Gila monster as an anti-diabetic agent. Toxicon. 59(4), 464–71. doi: 10.1016/j.toxicon.2010.12.016

Fusetani, N. (2009). "Marine toxins: An overview," in Marine Toxins as Research Tools. Progress in Molecular and Subcellular Biology, vol. 46. Eds. N. Fusetani and W. Kem (Springer, Berlin, Heidelberg), 1–44. doi: 10.1007/978–3-540–87895-7_1

Gallego Romero, I., Pai, A.A., Tung, J. and Gilad, Y. (2014). RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 12, 42. doi: 10.1186/1741–7007-12–42

Gayral, P., Weinert, L., Chiari, Y., Tsagkogeorga, G., Ballenghien, M. and Galtier, N. (2011). Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Mol. Ecol. Resour. 11, 650–661. doi: 10.1111/j.1755-0998.2011.03010.x

Ghiretti, F. (1959). Cephalotoxin: the crab-paralysing agent of the posterior salivary glands of cephalopods. Nature 183, 1192–1193. doi: 10.1038/1831192b0

Ghiretti, F. (1960). Toxicity of octopus saliva against crustacea. Ann. N. Y. Acad. Sci. 90, 726–741. doi: 10.1111/j.1749-6632.1960.tb26417.x

Gibbs, G.M., Roelants, K. and O'Bryan, M.K. (2008). The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins – roles in reproduction, cancer, and immune defense. Endocr. Ver. 29, 865–897. doi: 10.1210/er.2008-0032

Gómez-Mendikute, A., Elizondo, M., Venier, P. and Cajaraville, M.P. (2005). Characterization of mussel gill cells *in vivo* and *in vitro*. Cell Tissue Res. 321, 131–140. https://doi.org/10.1007/s00441-005-1093-9

Gonçalves, C. and Costa, P.M. (2021). Cephalotoxins: A hotspot for marine bioprospecting? Front. Mar. Sci. 8. doi: 10.3389/fmars.2021.647344

Gonçalves, C., Moutinho Cabral, I., Alves de Matos, A.P., Grosso, A.R. and Costa, P.M. (2024) Transcriptome profiling of the posterior salivary glands of the cuttlefish *Sepia officinalis* from the Portuguese West coast. Front. Mar. Sci. 11:1362824. doi: 10.3389/fmars.2024.1362824

Gonzalez, S.N., Sulzyk, V., Weigel Muñoz, M. and Cuasnicu, P.S. (2021). Cysteine-rich secretory proteins (CRISP) are key players in mammalian fertilization and fertility. Front. Cell. Dev. Biol. 9. doi: 10.3389/fcell.2021.800351

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. doi: 10.1038/nbt.1883

Greener, M. (2020). The next generation of venom-based drugs. Prescriber 31, 28–32. doi: 10.1002/psb.1837

Guerra, Á. (2019). Functional anatomy: Macroscopic anatomy and post-mortem examination. In: Gestal, C., Pascual, S., Guerra, Á., Fiorito, G., Vieites, J. (eds) Handbook of Pathogens and Diseases in Cephalopods. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-11330-8_3

Guryanova, S.V., Balandin, S.V., Belogurova-Ovchinnikova, O.Y. and Ovchinnikova, T.V. (2023). Marine invertebrate antimicrobial peptides and their potential as novel peptide antibiotics. Mar. Drugs 21, 503. doi: 10.3390/md21100503

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hagen, N.A., Fisher, K.M., Lapointe, B., du Souich, P., Chary, S., Moulin, D., et al. (2007). An open-label, multi-dose efficacy and safety study of intramuscular tetrodotoxin in patients with severe cancer-related pain. J. Pain Symptom Manage. 34, 171–182. doi: 10.1016/j.jpainsymman.2006.11.008

Hartman, W.J., Clark, W.G., Cyr, S.D., Jordon, A.L. and Leibhold, R.A. (1960). Pharmacologically active amines and their biogenesis in the octopus. Ann. N. Y. Acad. Sci. 90, 637–66. doi: 10.1111/j.1749-6632.1960.tb26411.x

He, M., Zhou, X. and Wang, X. (2024). Glycosylation: mechanisms, biological functions and clinical implications. Sig. Transduct. Target Ther. 9, 194. doi: https://doi.org/10.1038/s41392-024-01886-1

Henze, M. (1913). p-Oxyphenyläthylamin, das speicheldrüsengift der cephalopoden. Biol. Chem. 87, 51–58. doi: https://doi.org/10.1515/bchm2.1913.87.1.51

Hosseini, H., Al-Jabri, H.M., Moheimani, N.R., Siddiqui, S.A. and Saadaoui, I. (2022). Marine microbial bioprospecting: Exploitation of marine biodiversity towards biotechnological applications-a review. J. Basic Microbiol. 62, 1030–1043. doi: 10.1002/jobm.202100504

House, C.R. (1980). Physiology of invertebrate salivary glands. Biol. Rev. 55, 417–473. doi: https://doi.org/10.1111/j.1469-185X.1980.tb00700.x

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5, 299–314. doi: 10.1080/10618600.1996.10474713

Imran, M.A.S., Carrera, M., Pérez-Polo, S., Pérez, J., Barros, L., Dios, S., et al. (2023). Insights into common octopus (*Octopus vulgaris*) ink proteome and bioactive peptides using proteomic approaches. Mar. Drugs 21, 206. doi: 10.3390/md21040206

Jiráček, J. and Žáková, L. (2019). From venom peptides to a potential diabetes treatment. eLife 8, e44829. doi: 10.7554/eLife.44829

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Joseph, S.M., George, M.C., Nair, J.R., Senan, V.P., Pillai, D., and Sherief, P.M. (2005). Effect of feeding cuttlefish liver oil on immune function, inflammatory response and platelet aggregation in rats. Curr. Sci. 88, 507 - 510.

Juul, K.V., Bichet, D.G., Nielsen, S. and Nørgaard, J.P. (2014). The physiological and pathophysiological functions of renal and extrarenal vasopressin V2 receptors. Am. J. Physiol. Renal Physiol. 306(9), F931–40. doi: 10.1152/ajprenal.00604.2013

Kanda, A., Iwakoshi-Ukena, E., Takuwa-Kuroda, K. and Minakata, H. (2003). Isolation and characterization of novel tachykinins from the posterior salivary gland of the common octopus *Octopus vulgaris*. Peptides 24, 35–43. doi: 10.1016/s0196-9781(02)00274-7

Karthik, R., Manigandan, V., Ebenezar, K.K., Kavitha, M. and Saravanan, R. (2019). Structural characterization, teratogenicity and in vitro avian antimicrobial activity of posterior salivary gland (PSG) toxin from cuttlefish, *Sepia prashadi*. Int. J. Biol. Macromol. 124, 1145–1155. doi: 10.1016/j.ijbiomac.2018.12.040

Karthik, R., Manigandan, V., Ebenezar, K.K., Vijayashree, R. and Saravanan, R. (2017). *In vitro* and *in vivo* anticancer activity of posterior salivary gland toxin from the cuttlefish *Sepia pharaonis*, Ehrenberg (1831). Chem. Biol. Interact. 272, 10–20. doi: 10.1016/j.cbi.2017.04.002

Karthik, R., Saravanan, R., Ebenezar, K.K. and Sivamalai, T. (2015). Isolation, purification, and characterization of avian antimicrobial glycopeptide from the posterior salivary gland of *Sepia pharaonis*. Appl. Biochem. Biotechnol. 175, 1507–1518. doi: 10.1007/s12010-014-1370-8

Kaschina, E. and Unger, T. (2003). Angiotensin AT1/AT2 receptors: Regulation, signalling and function. Blood Press. 12(2), 70–88. doi: 10.1080/08037050310001057

Kem, W.R. (2005). "Marine organisms," in Encyclopedia of Toxicology, 2nd Edn, ed. P. Wexler (Cambridge, MA: Academic Press), 16–21. doi: 10.1017/cbo9780511803567.003

Kem, W.R. and Scott, J.D. (1980). Partial purification and characterization of a cytotoxic protein from squid (*Loligo pealei*) posterior salivary glands. Biol. Bull. 158:475.

King, G.F. (2011). Venoms as a platform for human drugs: translating toxins into therapeutics. Expert Opin. Biol. Ther. 11(11), 1469–84. doi: 10.1517/14712598.2011.621940

Kleczkowska, P. and Lipkowski, A.W. (2013). Neurotensin and neurotensin receptors: Characteristic, structure–activity relationship and pain modulation—A review. Eur. J. Pharmacol. 716, 54–60. doi: https://doi.org/10.1016/j.ejphar.2013.03.004

Kobayashi, N., Nakagawa, A., Muramatsu, T., Yamashina, Y., Shirai, T., Hashimoto, M.W., Ishigaki, Y., Ohnishi, T. and Mori, T. (1998). Supranuclear melanin caps reduce ultraviolet induced DNA photoproducts in human epidermis. J. Invest. Dermatol. 110(5), 806–10. doi: 10.1046/j.1523-1747.1998.00178.x

Kroeze, W.K., Sassano, M.F., Huang, X.P., Lansu, K., McCorvy, J.D., Giguère, P.M., Sciaky, N. and Roth, B.L. (2015). PRESTO-Tango as an open-source resource for interrogation of the druggable human GPCRome. Nat. Struct. Mol. Biol. 22, 362–9. doi: 10.1038/nsmb.3014.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549. doi: 10.1093/molbev/msy096

Kussmann, M., Abe Cunha, D.H. and Berciano, S. (2023). Bioactive compounds for human and planetary health. Front. Nutr. 10. doi: 10.3389/fnut.2023.1193848

Lagerström, M.C. and Schiöth, H.B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. Nat. Rev. Drug Discov. 7(4), 339–57. doi: 10.1038/nrd2518. Erratum in: Nat. Rev. Drug Discov. (2008). 7(6), 542.

Lago, J., Rodríguez, L.P., Blanco, L., Vieites, J.M. and Cabado, A.G. (2015). Tetrodotoxin, an extremely potent marine neurotoxin: distribution, toxicity, origin and therapeutical uses. Mar. Drugs 13, 6384–6406. doi: 10.3390/md13106384

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2, Nat. Methods. 9, 357–359. doi: 10.1038/nmeth.1923

Langmead, B., Wilks, C., Antonescu, V. and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 35, 421–432. doi: 10.1093/bioinformatics/bty648

Latek, D., Langer, I., Krzysko, K.A. and Charzewski, L. (2019). A molecular dynamics study of vasoactive intestinal peptide receptor 1 and the basis of its therapeutic antagonism. Int. J. Mol. Sci. 20(18), 4348. doi: 10.3390/ijms20184348.

Leal, M.C., Puga, J., Serôdio, J., Gomes, N.C. and Calado, R. (2012). Trends in the discovery of new marine natural products from invertebrates over the last two decades-where and what are we bioprospecting? PloS One. 7, e30580. doi: 10.1371/journal.pone.0030580

Leal, M.N., Anaya-Rojas, J.M., Munro, M.H. G., Blunt, J.W., Melian, C.J., Calado, R., et al. (2020). Fifty years of capacity building in the search for new marine natural products. Proc. Natl. Acad. Sci. U.S.A. 117, 24165–24172. doi: 10.1073/pnas.2007610117

Lewis, R.J. and Garcia, M.L. (2003). Therapeutic potential of venom peptides. Nat. Rev. Drug Discov. 2(10), 790–802. doi: 10.1038/nrd1197.

Lindequist, U. (2016). Marine-derived pharmaceuticals - challenges and opportunities. Biomol. Ther. (Seoul) 24, 561–571. doi: 10.4062/biomolther.2016.181

Livak, K.J. and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. Methods, 25, 402–408. doi: 10.1006/meth.2001.1262

Lo Bianco, S. (1888). Notizie biologiche riguardanti specialmente il periodo di maturità sessuale degli animali del Golfo di Napoli. Mitth. Z. Stat. Neapel. 8:385.

Lobo-da-Cunha, A. (2019). Structure and function of the digestive system in molluscs. Cell Tissue Res. 377, 475–503. doi: 10.1007/s00441-019-03085-9

Lu, J., Piper, S.J., Zhao, P., Miller, L.J., Wootten, D. and Sexton, P.M. (2022). Targeting VIP and PACAP receptor signaling: New insights into designing drugs for the PACAP subfamily of receptors. Int. J. Mol. Sci. 23(15), 8069. doi: 10.3390/ijms23158069

Luft, J. H. (1961). Improvements in epoxy resin embedding methods. J. Biophys. Biochem. Cytol. 9, 409–414. doi: 10.1083/jcb.9.2.409

Maghembe, R., Damian, D., Makaranga, A., Nyandoro, S.S., Lyantagaye, S.L., Kusari, S., et al. (2020). Omics for bioprospecting and drug discovery from bacteria and microalgae. Antibiotics (Basel) 9, 229. doi: 10.3390/antibiotics9050229

Malendowicz, L.K., Rucinski, M. (2021). Neuromedins NMU and NMS: An updated overview of their functions. Front. Endocrinol. 12, 713961. doi: 10.3389/fendo.2021.713961

Martins, A., Vieira, H., Gaspar, H. and Santos, S. (2014). Marketed marine natural products in the pharmaceutical and cosmeceutical industries: tips for success. Mar. Drugs 12, 1066–1101. doi: 10.3390/md12021066

Martins, C., Dreij, K. and Costa, P.M. (2019). The state-of-the art of environmental toxicogenomics: Challenges and perspectives of "omics" approaches directed to toxicant mixtures. Int. J. Environ. Res. Public Health. 16, 4718. doi: 10.3390/ijerph16234718

Matus, A.I. (1971). Fine structure of the posterior salivary gland of *Eledone cirrosa* and *Octopus vulgaris*. Z. Zellforsch. Mik. Ana. 122, 111–21. doi: 10.1007/BF00936120.

McDonald, N.M. and Cottrell, G.A. (1972). Purification and mode of action of toxin from *Eledone cirrosa*. Comp. Gen. Pharmacol. 3, 243–248. doi: 10.1016/0010-4035(72)90033-x

Ménez, A., Stöcklin, R., Mebs, D. (2006). 'Venomics' or : The venomous systems genome project. Toxicon. 47(3), 255 – 9. doi: 10.1016/j.toxicon.2005.12.010.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419. doi: 10.1093/nar/gkaa913

Molinski, T.F., Dalisay, D.S., Lievens, S.L. and Saludes, J.P. (2009). Drug development from marine natural products. Nat. Rev. Drug Discov. 8, 69–85. doi: 10.1038/nrd2487

Montaser, R. and Luesch, H. (2011). Marine natural products: a new wave of drugs? Future Med. Chem. 3, 1475–1489. doi: 10.4155/fmc.11.118

Montesanto, G. (2015). A fast GNU method to draw accurate scientific illustrations for taxonomy. Zookeys 515, 191–206. doi: 10.3897/zookeys.515.9459

Moutinho Cabral, I., Madeira, C., Grosso, A.R., and Costa, P.M. (2022). A drug discovery approach based on comparative transcriptomics between two toxin-secreting marine annelids: *Glycera alba* and *Hediste diversicolor*. Mol. Omics. 18, 731–744. doi: 10.1039/D2MO00138A

Moutinho Cabral, I., Gonçalves, C., Grosso, A.R. and Costa, P.M. (2024). Bioprospecting and marine 'omics': surfing the deep blue sea for novel bioactive proteins and peptides. Front. Mar. Sci. 11:1362697. doi: 10.3389/fmars.2024.1362697

Moya, A., Huisman, L., Ball, E.E., Hayward, D.C., Grasso, L.C., Chua, C.M., et al. (2012). Whole transcriptome analysis of the coral *Acropora millepora* reveals complex responses to CO2 - driven acidification during the initiation of calcification. Mol. Ecol. 21, 2440–2454. doi: 10.1111/j.1365-294X.2012.05554.x

Newman, D.J. and Cragg, G.M. (2014). Marine-sourced anti-cancer and cancer pain control agents in clinical and late preclinical development. Mar. Drugs. 12, 255–278. doi: 10.3390/md12010255

Norman, N. and Reid, A. (2000). A guide to squid, cuttlefish and octopuses of Australasia. Clayton, Vic: CSIRO Publishing. doi: 10.1071/9780643101098

Olivera, B.M., Cruz, L.J., de Santos, V., LeCheminant, G.W., Griffin, D., Zeikus, R., et al. (1987). Neuronal calcium channel antagonists. Discrimination between calcium channel subtypes using ω-conotoxin from *Conus magus* venom. Biochemistry 26, 2086–2090. doi: 10.1021/bi00382a004

Pardos-Blas, J.R., Irisarri, I., Abalde, S., Tenorio, M.J. and Zardoya, R. (2019). Conotoxin diversity in the venom gland transcriptome of the magician's cone, *Pionoconus magus*. Mar. Drugs. 17, 553. doi: 10.3390/md17100553

Pennington, M.W., Czerwinski, A. and Norton, R.S. (2018). Peptide therapeutics from venom: current status and potential. Bioorg. Med. Chem. 26, 2738–2758. doi: 10.1016/j.bmc.2017.09.029

Pérez-Albaladejo, E., Rizzi, J., Fernandes, D., Lille-Langøy, R., Karlsen, O. A., Goksøyr, A., Oros, A., Spagnoli, F. and Porte, C. (2016). Assessment of the environmental quality of coastal sediments by using a combination of in vitro bioassays. Mar. Pollut. Bull. 108 (1-2), 53– 61. doi: 10.1016/j.marpolbul.2016.04.063

Pérez-Polo, S., Imran, M.A.S., Dios, S., Pérez, J., Barros, L., Carrera, M., et al. (2023). Identifying natural bioactive peptides from the common octopus (*Octopus vulgaris* Cuvier 1797) skin mucus by-products using proteogenomic analysis. Int. J. Mol. Sci. 24, 7145. doi: 10.3390/ijms24087145

Perryman, R., Renziehausen, A., Shaye, H., Kostagianni, A.D., Tsiailanis, A.D., Thorne, T. et al. (2022). Inhibition of the angiotensin II type 2 receptor AT_2R is a novel therapeutic strategy for glioblastoma. Proc. Natl. Acad. Sci. U.S.A. 119(32), e2116289119. doi: 10.1073/pnas.2116289119

Ponte, G. and Modica, M.V. (2017). Salivary glands in predatory mollusks: evolutionary considerations. Front. Physiol. 8:580. doi: 10.3389/fphys.2017.00580

Prashanth, J.R., Hasaballah, N. and Vetter, I. (2017). Pharmacological screening technologies for venom peptide discovery. Neuropharmacology. 127, 4–19. doi: https://doi.org/10.1016/j.neuropharm.2017.03.038.

Prentis, P.J., Pavasovic, A. and Norton, R.S. (2018). Sea anemones: quiet achievers in the field of peptide toxins. Toxins 10:36. doi: 10.3390/toxins10010036

Puchta, M., Boczkowska, M. and Groszyk, J. (2020). Low RIN value for RNA-Seq library construction from long-term stored seeds: A case study of barley seeds. Genes. 11, 1190. doi: 10.3390/genes11101190

Ramos-Silva, P., Kaandorp, J., Huisman, L., Marie, B., Zanella-Cléon, I., Guichard, N., et al. (2013). The skeletal proteome of the coral *Acropora millepora*: the evolution of calcification by co-option and domain shuffling. Mol. Biol. Evol. 30, 2099–2112. doi: 10.1093/molbev/mst109

Rashmi, U., Khochare, S., Attarde, S., Laxme, R.R.S., Suranse, V., Martin, G., et al. (2021). Remarkable intrapopulation venom variability in the monocellate cobra (*Naja kaouthia*) unveils neglected aspects of India's snakebite problem. J. Proteomics. 242, 104256. doi: 10.1016/j.jprot.2021.104256

Reynaud, S., Ciolek, J., Degueldre, M., Saez, N.J., Sequeira, A.F., Duhoo, Y. et al. (2020). A venomics approach coupled to high-throughput toxin production strategies identifies the first venom-derived melanocortin receptor agonists. J. Med. Chem. 63(15), 8250–8264. doi: 10.1021/acs.jmedchem.0c00485

Rivera, I.-G., Ordoñez, M., Presa, N., Gomez-Larrauri, A., Simón, J., Trueba, M. and Gomez-Muñoz, A. (2015). Sphingomyelinase D/ceramide 1-phosphate in cell survival and inflammation. Toxins. 7(5), 1457–66. doi: 10.3390/toxins7051457

Robberecht, P., Waelbroeck, M., Dehaye, J.-P., Winand, J., Vandermeers, A., Vandermeers-Piret, M.-C. and Christophe, J. (1984). Evidence that helodermin, a newly extracted peptide from Gila monster venom, is a member of the secretin/VIP/PHI family of peptides with an original pattern of biological properties. FEBS Lett. 166, 277–282. doi: https://doi.org/10.1016/0014-5793(84)80095-2

Rodrigo, A.P. and Costa, P.M. (2019). The hidden biotechnological potential of marine invertebrates: the polychaeta case study. Environ. Res. 173, 270–280. doi: 10.1016/j.envres.2019.03.048

Rodrigo, A.P., Grosso, A.R., Baptista, P.V., Fernandes, A.R. and Costa, P.M. (2021). A transcriptomic approach to the recruitment of venom proteins in a marine annelid. Toxins. 13, 97. doi: 10.3390/toxins13020097

Rodrigo, A.P., Martins, C., Costa, M.H., Alves de Matos, A.P. and Costa, P. M. (2018). A morphoanatomical approach to the adaptive features of the epidermis and proboscis of a marine Polychaeta: *Eulalia viridis* (Phyllodocida: Phyllodocidae). J. Anat. 233, 567–579. doi: 10.1111/joa.12870

Romanini, M.G. (1952). Osservazioni sulla ialuronidasi delle ghiandole salivari enteriorie posteriori degli octopodi. Pubbl. Staz. Zool. Napo. 23, 251–270.

Ruder, T., Sunagar, K., Undheim, E.A.B., Ali, S.A., Wai, T., Low, D.H.W., et al. (2013). Molecular phylogeny and evolution of the proteins encoded by coleoid (cuttlefish, octopus, and squid) posterior venom glands. J. Mol. Evol. 76, 192–204. doi: 10.1007/s00239-013-9552-5

Safavi-Hemami, H., Gajewiak, J., Karanth, S., Robinson, S.D., Ueberheide, B., Douglass, A.D., et al. (2015). Specialized insulin is used for chemical warfare by fish-hunting cone snails. Proc. Natl. Acad. Sci. U.S.A. 112, 1743–1748. doi: 10.1073/pnas.1423857112

Sample, C.J., Hudak, K.E., Barefoot, B.E., Koci, M.D., Wanyonyi, M.S., Abraham, S., et al. (2013). A mastoparan-derived peptide has broad-spectrum antiviral activity against enveloped viruses. Peptides. 48, 96–105. doi: 10.1016/j.peptides.2013.07.014

Savage, I.V.E. and Howden, M.E.H. (1977). Hapalotoxin, a second lethal toxin from the octopus *Hapalochlaena maculosa*. Toxicon. 15, 463–466. doi: 10.1016/0041-0101(77)90127-1

Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods. 9, 671–675. doi: 10.1038/nmeth.2089

Shen, G.S., Layer, R.T. and McCabe, R.T. (2000). Conopeptides: From deadly venoms to novel therapeutics. Drug Discov. Today. 5, 98–106. doi: 10.1016/s1359-6446(99)01454-3

Sheng, Y.H. and Hasnain, S.Z. (2022). Mucus and mucins: The underappreciated host defense system. Front. Cell. Infect. Microbiol. 12, 856962. doi: 10.3389/fcimb.2022.856962

Sheumack, D.D., Howden, M.E.H., Spence, I. and Quinn, R.J. (1978). Maculotoxin: a neurotoxin from the venom glands of the octopus *Hapalochlaena maculosa* identified as tetrodotoxin. Science. 199, 188–189.

Shi, J., Liu, T.T., Wang, X., Epstein, D.H., Zhao, L.Y., Zhang, X.L., et al. (2009). Tetrodotoxin reduces cue-induced drug craving and anxiety in abstinent heroin addicts. Pharmacol. Biochem. Behav. 92, 603–607.

Shiomi, K., Honma, T., Ide, M., Nagashima, Y., Ishida, M., and Chino, M. (2003). An epidermal growth factor-like toxin and two sodium channel toxins from the sea anemone *Stichodactyla gigantea*. Toxicon. 41, 229–236.

Silva, F., Huang, Y., Yang, V., Mu, X., Shi, Q. and Antunes, A. (2018). Transcriptomic characterization of the South American freshwater stingray *Potamotrygon motoro* venom apparatus. Toxins. 10, 544. doi: 10.3390/toxins10120544

Silver, N., Best, S., Jiang, J. and Thein, S.L. (2006). Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. BMC Mol. Biol. 7, 33. doi: 10.1186/1471-2199-7-33

Singh, K.D., and Karnik, S.S. (2016). Angiotensin receptors: Structure, function, signaling and clinical applications. J. Cell. Signal. 1(2), 111. doi: 10.4172/jcs.1000111.

Sirakov, M., Zarrella, I., Borra, M., Rizzo, F., Biffali, E., Arnone, M.I. and Fiorito, G. (2009). Selection and validation of a set of reliable reference genes for quantitative RT-PCR studies in the brain of the Cephalopod Mollusc *Octopus vulgaris*. BMC Mol. Biol. 10, 70. doi: 10.1186/1471-2199-10-70.

Smith, E.G., Surm, J.M., Macrander, J., Simhi, A., Amir, G., Sachkova, M.Y., et al. (2023). Micro and macroevolution of sea anemone venom phenotype. Nat. Commun. 14, 249. doi: 10.1038/s41467-023-35794-9

Soneson, C., Love, M.I. and Robinson, M.D. (2015). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Res 4, 1521. doi: 10.12688/f1000research.7563.2

Song, H., Li, J., Lu, C.L., Kang, L., Xie, L., Zhang, Y.Y., et al. (2011). Tetrodotoxin alleviates acute heroin withdrawal syndrome: a multicentre, randomized, double-blind, placebo-controlled study. Clin. Exp. Pharmacol. Physiol. 38, 510–514.

Songdahl, J.H. and Shapiro, B.I. (1974). Purification and composition of a toxin from the posterior salivary gland of *Octopus dofleini*. Toxicon. 12, 109–115.

Stonik, V.A. and Stonik, I.V. (2016). Toxins produced by marine invertebrate and vertebrate animals: A short review. In: Gopalakrishnakone, P., Haddad Jr., V., Tubaro, A., Kim, E., Kem, W. (eds) Marine and Freshwater Toxins. Toxinology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6419-4_5

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 28, 1102–1104. doi: 10.2144/00286ir01

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 28, 1102–1104. doi: https://doi.org/10.2144/00286ir01

Syed, Y.Y. and McCormack, P.L. (2015). Exenatide extended-release: An updated review of its use in type 2 diabetes mellitus. Drugs. 75, 1141–1152. doi: https://doi.org/10.1007/s40265-015-0420-z

Tadokoro, T., Modahl, C.M., Maenaka, K. and Aoki-Shioi, N. (2020). Cysteine-rich secretory proteins (CRISPs) from venomous snakes: An overview of the functional diversity in a large and underappreciated superfamily. Toxins. 12, 175. doi: 10.3390/toxins12030175

Turner, A., Kaas, Q. and Craik, D.J. (2020). Hormone-like conopeptides – new tools for pharmaceutical design. RSC Med. Chem. 11, 1235–1251. doi: 10.1039/D0MD00173B

Ueda, A., Nagai, H., Ishida, M., Nagashima, Y. and Shiomi, K. (2008). Purification and molecular cloning of SE-cephalotoxin, a novel proteinaceous toxin from the posterior salivary gland of cuttlefish *Sepia esculenta*. Toxicon. 52, 574–581. doi: 10.1016/j.toxicon.2008.07.007

UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489. doi: 10.1093/nar/gkaa1100

Van Baelen, A.-C., Iturrioz, X., Chaigneau, M., Kessler, P., Llorens-Cortes, C., Servant, D., Gilles, N. and Robin, P. (2023). Characterization of the first animal toxin acting as an antagonist on AT1 receptor. Int. J. Mol. Sci. 24(3):2330. doi: 10.3390/ijms24032330

Van Baelen, A.-C., Robin, P., Kessler, P., Maïga, A., Gilles, N. and Servant, D. (2022). Structural and functional diversity of animal toxins interacting with GPCRs. Front. Mol. Biosci. 9:811365. doi: 10.3389/fmolb.2022.811365.

Venable, J.H. and Coggeshall, R. (1965). A simplified lead citrate stain for use in electron microscopy. J. Cell Biol. 25, 407–408. doi: 10.1083/jcb.25.2.407

Vicens, A. and Treviño, C.L. (2018). Positive selection in the evolution of mammalian CRISPs. J. Mol. Evol. 86, 635–645. doi: 10.1007/s00239-018-9872-6

Vieira, H., Leal, M.C. and Calado, R. (2020). Fifty shades of blue: How Blue Biotechnology is shaping the bioeconomy. Trends Biotechnol. 38, 940–943. doi: 10.1016/j.tibtech.2020.03.011

Vila-Farrés, X., Giralt, E. and Vila, J. (2012). Update of peptides with antibacterial activity. Curr. Med. Chem. 19, 6188–6198.

Villanueva, R., Perricone, V. and Fiorito, G. (2017). Cephalopods as predators: A short journey among behavioral flexibilities, adaptions, and feeding habits. Front. Physiol. 8, 598. doi: 10.3389/fphys.2017.00598.

von Euler, U.S. (1952). Presence of catechol amines in visceral organs of fish and invertebrates. Acta Physiol. Scand. 28, 297–305.

von Reumont, B.M., Campbell, L.I. and Jenner, R.A. (2014a). *Quo Vadis* venomics? A roadmap to neglected venomous invertebrates. Toxins 6, 3488–3551. doi: https://doi.org/10.3390/toxins6123488

von Reumont, B.M., Campbell, L.I., Richter, S., Hering, L., Sykes, D., Hetmank, J., et al. (2014b). A Polychaete's powerful punch: venom gland transcriptomics of *Glycera* reveals a complex cocktail of toxin homologs. Genome Biol. Evol. 6, 2406–2423. doi: 10.1093/gbe/evu190

von Reumont, B.M., Undheim, E.A.B., Jauss, R.T. and Jenner, R.A. (2017). Venomics of remipede crustaceans reveals novel peptide diversity and illuminates the venom's biological role. Toxins. 9, 234. doi: 10.3390/toxins9080234

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303. doi: 10.1093/nar/gky427

Watters, D.J. (2018). Ascidian toxins with potential for drug development. Mar. Drugs 16:162.

Whitelaw, B.L., Strugnell, J.M., Faou, P., da Fonseca, R.R., Hall, N.E., Norman, M., et al. (2016). Combined transcriptomic and proteomic analysis of the posterior salivary gland from the southern blue-ringed octopus and the southern sand octopus. J. Proteome Res. 15, 3284–3297. doi: 10.1021/acs.jproteome.6b00452

Williams, J.A., Day, M. and Heavner, J.E. (2008). Ziconotide: an update and review. Expert Opin. Pharmacother. 9, 1575–1583. doi: 10.1517/14656566.9.9.1575

Wozniak, K.M., Rojas, C., Wu, Y. and Slusher, B.S. (2012). The role of glutamate signaling in pain processes and its regulation by GCP II inhibition. Curr. Med. Chem. 19, 1323–1334.

Yotsu-Yamashita, M., Mebs, D. and Flachsenberger, W. (2007). Distribution of tetrodotoxin in the body of the blue-ringed octopus (*Hapalochlaena maculosa*). Toxicon. 49, 410–412.

Young, J.Z. (1965). The buccal nervous system of octopus. Phil. Trans. R. Soc. Lond. B. 24927–44. doi: http://doi.org/10.1098/rstb.1965.0007

Zeghal, M., Laroche, G. and Giguère, P. M. (2020). Parallel interrogation of β-Arrestin2 recruitment for ligand screening on a GPCR-wide scale using PRESTO-Tango assay. J. Vis. Exp. 157, e60823. doi:10.3791/60823

Zhang, Q., Xu, J., Zhou, X. and Liu, Z. (2022). CAP superfamily proteins from venomous animals: Who we are and what to do? Int. J. Biol. Macromolecules. 221, 691–702. doi: 10.1016/j.ijbiomac.2022.09.079

Zhang, Y. (2015). Why do we study animal toxins? Zool Res. 36(4), 183–222. doi: 10.13918/j.issn.2095-8137.2015.4.183.

A

Appendix to Chapter 3

A.1 *R* Script

```
setwd("/[PATH]/R")

##Loading a R package
library(cowplot)
library(edgeR)
library(ggplot2)
library(limma)
library(packcircles)
library(psych)
library(seqinr)
library(tximport)
library(UniprotR)

##Diretories
dir.create("Annot")
dir.create("TopHit")
dir.create("Circles")
dir.create("TPM+1")
dir.create("Pfam")
dir.create("Pfam/SeaTox")

pathExpression<-"Expression/"
pathAnnot<-"Annot/"
pathTopHit<-"TopHit/"
pathCircles<-"Circles/"
pathTPM<-"TPM+1/"
pathPfam<-"Pfam/"
pathPfamSeaTox<-"Pfam/SeaTox/"


##Files
filefull<-"/[PATH]/sepiaFull.csv"
fileBlastp<-
data.frame(SeaTox="/[PATH]/SepiaTrim_Blastp_uniprot_database_seatox.csv",Toxin="/[P
ATH]/SepiaTrim_Blastp_uniprot_toxin.csv",Enzyme="/[PATH]/SepiaTrim_Blastp_enzyme.cs
v",Secreted       =       "/[PATH]/SepiaTrim_Blastp_uniprot_secreted.csv",PTM       =
"/[PATH]/SepiaTrim_Blastp_PTM.csv")
database<-c("SeaTox", "Toxin", "Enzyme", "Secreted", "PTM")
Kallistofolder<-"/[PATH]/Kallisto/"
filepfam<-data.frame(pfam = "/[PATH]/Pfam/SepiaTrim_pfam_cor.domtblout")


##Expression####
##Function
##Calculating the average counts across all samples
Expression<-function(pathExpression, filefull, AnnotTable){
  dir.create(pathExpression)
  full<-read.csv(filefull, sep=";", header = TRUE, row.names = 1)
  for(i in 1:nrow(full)){
    full$Mean[i]<-mean(as.numeric(full[i,2:4]))
  }
  save(full, file = paste(pathExpression, "full.RData", sep = ""))}

##R Console
Expression(pathExpression, filefull, AnnotTable)

##Annotation: Analysing the results from the blast against different databases####
##Function
Annotate<-function(fileBlastp, pathAnnot){
  Annot<-read.csv(fileBlastp, sep=";", header = FALSE)
  print(fileBlastp)
```

```
  print(nrow(Annot))
  Annot[,(ncol(Annot)+1)]<-unlist(sapply(Annot[,2],function(x)unlist(strsplit(x,
"|", fixed = TRUE))[2]))
  Annot<-Annot[,c(1,16,3,11)]
  colnames(Annot)<-c("TrinityID", paste(sep = ": ", names(fileBlastp), c("Accession",
"Blastp_%ID", "Blastp_Evalue")))
  Annot<-Annot[which(grepl("TRINITY", Annot$TrinityID)),]
  print(nrow(Annot))
  ##Sorting by e-value and %ID
  Annot<-Annot[order(Annot[,paste(sep = ": ", names(fileBlastp), "Blastp_Evalue")],
-Annot[,paste(sep = ": ", names(fileBlastp), "Blastp_%ID")]),]
  rownames(Annot)<-NULL
  write.table(Annot,  paste(pathAnnot, "/", names(fileBlastp), ".csv", sep = ""),
sep=";", col.names=NA)
  return(Annot)}

##RConsole
for(i in 1:ncol(fileBlastp)){
  assign(database[i],Annotate(setNames(fileBlastp[1,i],    colnames(fileBlastp)[i]),
pathAnnot))}
save(SeaTox, Toxin, Enzyme, Secreted, PTM, file = paste(pathAnnot, "5DB.RData", sep
= ""))

##Creating an extra column in each dataframe with 1
SeaTox$SeaTox<-1
Toxin$Toxin<-1
Enzyme$Enzyme<-1
Secreted$Secreted<-1
PTM$PTM<-1

##Merge the results from the blasts against different dabatases: SeaTox, Toxin,
Enzyme, Secreted and PTM
blastpAll<-merge(SeaTox, Toxin, by = "TrinityID", all = TRUE)
blastpAll<-merge(blastpAll, Enzyme, by = "TrinityID", all = TRUE)
blastpAll<-merge(blastpAll, Secreted, by = "TrinityID", all = TRUE)
blastpAll<-merge(blastpAll, PTM, by = "TrinityID", all = TRUE)

save(blastpAll, file = paste(pathAnnot, "blastpAll.RData", sep = ""))

##It is going to analyse each trinity and check if it has homology-matching against
a specific database
##If it does not have, it is going to add a 0 in the column of the respective database
for(i in 1:nrow(blastpAll)){
  if(is.na(blastpAll$SeaTox[i]))
    blastpAll$SeaTox[i]<-0
  if(is.na(blastpAll$Toxin[i]))
    blastpAll$Toxin[i]<-0
  if(is.na(blastpAll$Enzyme[i]))
    blastpAll$Enzyme[i]<-0
  if(is.na(blastpAll$Secreted[i]))
    blastpAll$Secreted[i]<-0
  if(is.na(blastpAll$PTM[i]))
    blastpAll$PTM[i]<-0}
save(blastpAll, file = paste(pathAnnot, "blastpAll2.RData", sep = ""))
write.table(blastpAll,    paste(pathAnnot,  "blastpAll2.csv",  sep  =  ""),  sep=";",
col.names=NA)


##Top hits####
##Analysing the results from the blast against SeaTox database
##Do a top 10 of the best hits
##Function
TopHit<-function(fileBlastp, pathTopHit){
  Annot<-read.csv(fileBlastp, sep=";", header = FALSE)
  print(fileBlastp)
```

```
  print(nrow(Annot))
  Annot[,(ncol(Annot)+1)]<-unlist(sapply(Annot[,2], function(x) unlist(strsplit(x,
"|", fixed = TRUE))[2]))
  ##Retrieving the different information present in blastp_Title, specifically,
Protein_name and organism
  for(i in 1:nrow(Annot)){
    Annot[i,17]<-unlist(strsplit(Annot[i,15], paste(Annot[i,13], " ", sep = ""),
fixed = TRUE))[2]
    Annot[i,18]<-unlist(strsplit(Annot[i,17], "OS=", fixed = TRUE))[2]
    Annot[i,17]<-unlist(strsplit(Annot[i,17], "OS=", fixed = TRUE))[1]
    Annot[i,18]<-unlist(strsplit(Annot[i,18], "OX=", fixed = TRUE))[1] }
  Annot<-Annot[,c(1,16,3,11,17,18)]
  colnames(Annot)<-c("TrinityID", paste(sep = ": ", names(fileBlastp), c("Accession",
"Blastp_%ID", "Blastp_Evalue", "Protein", "Organism")))
  ##Sometimes the blast_title cuts short and continues in a different line
  ##So, we are going to erase those lines
  Annot<-Annot[which(grepl("TRINITY", Annot$TrinityID)),]
  print(nrow(Annot))
  ##Sorting by e-value and %ID
  Annot<-Annot[order(Annot[,paste(sep = ": ", names(fileBlastp), "Blastp_Evalue")],
-Annot[,paste(sep = ": ", names(fileBlastp), "Blastp_%ID")]),]
  rownames(Annot)<-NULL
  write.table(Annot, paste(pathTopHit, names(fileBlastp), ".csv", sep = ""),
sep=";", col.names=NA)
  Same<-which((Annot[,4] == Annot[10,4]) & (Annot[,3] == Annot[10,3]))
  write.table(Annot[1:Same[length(Same)],], paste(pathTopHit, "Top10",
names(fileBlastp), ".csv", sep = ""), sep=";", col.names=NA)
  save(Annot, file = paste(pathTopHit, names(fileBlastp), ".RData", sep = ""))
  return(Annot)}

##R Console
TopHitSeatox<-TopHit(setNames(fileBlastp[1,1], colnames(fileBlastp)[1]), pathTopHit)

##Retrieving information from uniprot####
##RConsole
##Retrieve the keywords and the protein families of the specific accessions
load(paste(pathTopHit, "Seatox.RData", sep = ""))
Accession<-unique(Annot$`SeaTox: Accession`)
AccessionInfo<-GetProteinAnnontate(Accession, c("keyword", "protein_families"))
save(AccessionInfo, file = paste(pathCircles, "AccessionInfo.RData", sep = ""))

##Calculate the cpms####
##RConsole
load(paste(pathAnnot, "blastpAll2.RData", sep = ""))
load(paste(pathExpression, "full.RData", sep = ""))

Level<-c("PSG", "PSG", "PSG")
##Create a DGEList object from a table of counts
Data<-DGEList(counts = full[, 2:4], group = Level, genes = full[, 1])
##Calculate the normalization factors
Data<-calcNormFactors(Data)
save(Data, file = paste(pathExpression, "Data.RData", sep = ""))
##Calculate the cpms
cpms<-cpm(Data, normalized.lib.sizes = TRUE, log = T)
rownames(cpms)<-as.matrix(Data$genes)
save(cpms, file = paste(pathExpression,"cpms.RData", sep = ""))

##Merging the dataframe with the cpms with the dataframe with results from all blasts
blastpAll$blastpID<-unlist(sapply(blastpAll[,1],function(x)unlist(strsplit(x,".",
fixed=TRUE))[1]))

blastcpm<-merge(cpms, blastpAll, by.x = "row.names", by.y = "blastpID")
for(i in 1:nrow(blastcpm)){
  blastcpm$Mean[i]<-mean(as.numeric(blastcpm[i,2:4]))}
save(blastcpm, file = paste(pathExpression, "blastcpm.RData", sep = ""))
```

```
##Building a dataframe with only the log2CPM and evalue
blastcpmGraph<-as.data.frame(matrix(data = NA, nrow = nrow(blastcpm), ncol = 2,
dimnames = list(blastcpm$TrinityID, c("log2CPM", "Evalue"))))
blastcpmGraph$log2CPM<-blastcpm$Mean

for(i in 1:nrow(blastcpm)){
  if(is.na(blastcpm$`SeaTox: Blastp_Evalue`[i]))
    blastcpm$`SeaTox: Blastp_Evalue`[i]<-1
  if(is.na(blastcpm$`Toxin: Blastp_Evalue`[i]))
    blastcpm$`Toxin: Blastp_Evalue`[i]<-1
  if(is.na(blastcpm$`Enzyme: Blastp_Evalue`[i]))
    blastcpm$`Enzyme: Blastp_Evalue`[i]<-1
  if(is.na(blastcpm$`Secreted: Blastp_Evalue`[i]))
    blastcpm$`Secreted: Blastp_Evalue`[i]<-1
  if(is.na(blastcpm$`PTM: Blastp_Evalue`[i]))
    blastcpm$`PTM: Blastp_Evalue`[i]<-1}
blastcpmGraph$Evalue<-apply(blastcpm[,c("SeaTox: Blastp_Evalue", "Toxin:
Blastp_Evalue", "Enzyme: Blastp_Evalue", "Secreted: Blastp_Evalue", "PTM:
Blastp_Evalue")], 1, min)
save(blastcpmGraph, file = "Expression/blastcpmGraph.RData")

##Check if the transcripts of interest belong to other DB apart from SeaTox####
IDs<-c("TRINITY_DN23540_c1_g6_i2.p1",
       "TRINITY_DN24442_c10_g1_i1.p1",
       "TRINITY_DN24442_c10_g1_i4.p1",
       "TRINITY_DN23515_c5_g1_i13.p1")
TranscriptsOfInterest<-blastpAll[which(blastpAll$TrinityID %in% IDs),]
write.table(TranscriptsOfInterest, paste(pathTopHit, "TranscriptsOfInterest.csv",
sep = ""), sep=";", col.names=NA)

##Venn Diagrams####
##Functions
##Venn Diagram with only the four databases: Toxin, Enzyme, Secreted and PTM
VennDiagram4<-function(blastpAll){
  quartz()
  vennDiagram(blastpAll[,c("Toxin", "Enzyme", "Secreted", "PTM")],
circle.col=c("tomato", "orange", "pink", "gold2"), show.include = FALSE,
cex=c(0.8,0.7,0.7), names = c("Toxin", "Enzyme", "Secreted", "PTM"))
  recordPlot()}

##Venn Diagram comparing the hits obtain in the blasts against SeaTox database and
the merged result of the other four databases
VennDiagram<-function(blastpAll){
  quartz()
  vennDiagram(blastpAll[,c("SeaTox", "Databases")],
circle.col=c("peachpuff","tomato"), show.include = FALSE, cex=c(0.8,0.7,0.7), names
= c("SeaTox", "Previous Databases"))
  recordPlot()}

##RConsole
##Venn Diagram comparing the hits obtain in the different blast against four
databases: Toxin, Enzyme, Secreted and PTM
quartz()
venn<-VennDiagram4(blastpAll)
venngraph<-plot_grid(venn)
ggsave(paste(pathAnnot, "VennDiagram4Databases.svg", sep = ""), height = 15, width =
15,  units = 'cm', dpi=600)

##Creating a column that has the value 1 if that ORF has homology-matching agaisnt
one of the following databases: Toxin, Enzyme, Secreted and PTM
for(i in 1:nrow(blastpAll)){
  if(blastpAll$Toxin[i] == 1 | blastpAll$Enzyme[i] == 1 | blastpAll$Secreted[i] == 1
| blastpAll$PTM[i] == 1)
    blastpAll$Databases[i]<-1
```

```
  else
    blastpAll$Databases[i]<-0}

##Venn Diagram comparing the hits obtain in the blasts against SeaTox database and
the merged result of the other four databases
quartz()
venn<-VennDiagram(blastpAll)
venngraph<-plot_grid(venn)
ggsave(paste(pathAnnot, "VennDiagramSeatoxOtherDatabases.svg", sep = ""), height =
15, width = 15,  units = 'cm', dpi=600)
write.table(blastpAll,  paste(pathAnnot,  "blastpAll.csv",  sep  =  ""),  sep=";",
col.names=NA)

##SeaTox: Discovering the ORFs####
##Retrieving that are only annotated agaisnt SeaTox database
##as well as the ORFs that annotated against both SeaTox database and one of the
following databases: Toxin, Enzyme, Secreted and PTM
##RConsole
uniqueSeatox<-blastpAll[which(blastpAll$SeaTox == 1 & blastpAll$Databases == 0),]
Interset<-blastpAll[which(blastpAll$SeaTox == 1 & blastpAll$Databases == 1),]

load(paste(pathTopHit, "SeaTox.RData", sep = ""))
uniqueSeatox<-merge(uniqueSeatox, Annot[,c(1,5:6)], by = "TrinityID")
uniqueSeatox<-uniqueSeatox[,c(1:4,23,24,5:22)]

Interset<-merge(Interset, Annot[,c(1,5:6)], by = "TrinityID")
Interset<-Interset[,c(1:4,23,24,5:22)]

save(uniqueSeatox, Interset, file = paste(pathTopHit, "SeatoxVenn.RData", sep = ""))
write.table(uniqueSeatox,  paste(pathTopHit, "uniqueSeatox.csv", sep = ""), sep=";",
col.names=NA)
write.table(Interset,    paste(pathTopHit,  "Interset.csv",  sep  =  ""),  sep=";",
col.names=NA)

##Top10 of the dataframe unique
uniqueSeatox<-uniqueSeatox[order(uniqueSeatox[,"SeaTox:      Blastp_Evalue"],      -
uniqueSeatox[,"SeaTox: Blastp_%ID"]),]
rownames(uniqueSeatox)<-NULL
Same<-which((uniqueSeatox[,4]   ==   uniqueSeatox[10,4])   &   (uniqueSeatox[,3]   ==
uniqueSeatox[10,3]))
write.table(uniqueSeatox[1:Same[length(Same)],],                    paste(pathTopHit,
"Top10_uniqueSeatox.csv", sep = ""), sep=";", col.names=NA)

##Top10 of the dataframe Interset
Interset<-Interset[order(Interset[,"SeaTox:   Blastp_Evalue"],   -Interset[,"SeaTox:
Blastp_%ID"]),]
rownames(Interset)<-NULL
Same<-which((Interset[,4] == Interset[10,4]) & (Interset[,3] == Interset[10,3]))
write.table(Interset[1:Same[length(Same)],],                        paste(pathTopHit,
"Top10_Interset.csv", sep = ""), sep=";", col.names=NA)

save(uniqueSeatox, Interset, file = paste(pathTopHit, "Top10uniqueSeatox.RData", sep
= ""))

##Calculates TPMs####
##RConsole
load(paste(pathTopHit, "Seatox.RData", sep = ""))
load(paste(pathCircles, "AccessionInfo.RData", sep = ""))

##Reading the output of kallisto by importing the counts and abundance
samples<-dir(Kallistofolder)
file<-c(paste(sep = "", Kallistofolder, samples, "/", "abundance.h5"))
names(file)=samples
print(names(file))
```

```
Tsv<-tximport(file, type = "kallisto", txOut = TRUE, countsFromAbundance =
"lengthScaledTPM")

##TPMs
##We added 0.1 to the TPMs so we do not have negative values afterwards
TPM<-as.data.frame(Tsv$abundance)
TPM<-TPM+0.1

##Calculate the mean of TPMs
for(i in 1:nrow(TPM)){
  TPM$Mean[i]<-mean(as.numeric(TPM[i,1:3]))}
save(Tsv, TPM, file = paste(pathTPM, "CountsAbundance.RData", sep = ""))

##Circles with expression levels (TPMs)####
##Function
NumberOfCircles<-function(Term, Column, Annot, full){
  aux<-grep(Term, Annot[,Column])
  ##To erase serine protease inhibitors from the group of serine proteases
  if(Term == "Serine protease"){
    aux1<-grep("Serine protease inhibitor", Annot[,Column])
    aux<-setdiff(aux, aux1)  }
  AnnotTerm<-Annot[aux,]
  ID<-unlist(sapply(AnnotTerm[,1], function(x) unlist(strsplit(x,".",
fixed=TRUE))[1]))
  AnnotTerm<-cbind(AnnotTerm, ID)
  AnnotTerm<-merge(AnnotTerm, full, by.x = "ID", by.y = "row.names")
  ##Calculates the mean of the expression value, e.g., TPM, from all ORFs belong to
a protein family
  Mean<-mean(AnnotTerm$Mean)
  ##Calculates the number of ORFs that belong to a protein family
  Circle<-length(aux)
  ##It joins the number of ORFs with mean of
  Circle<-c(Circle, Mean)}

##RConsole
##Protein families of interest
##It also says in where a specific protein family is classified, whether in keywords
or in protein families
Circles<-data.frame(Term = c("Toxin", "Hormone", "Serine protease",
"Metalloprotease", "Glycosyl hydrolase 18 family", "Glycosyl hydrolase 56 family"),
Column = c("SeaTox: Keywords", "SeaTox: Keywords", "SeaTox: Keywords", "SeaTox:
Keywords", "SeaTox: Protein Families", "SeaTox: Protein Families"))

##It merges the Annot (a dataframe in this case with the results of the blast against
SeaTox database) with information retrived from Uniprot
AnnotSeaTox<-merge(Annot, AccessionInfo, by.x = "SeaTox: Accession", by.y =
"row.names", all = TRUE)
AnnotSeaTox<-AnnotSeaTox[,c(2,1,3:8)]
colnames(AnnotSeaTox)[7:8]<-c("SeaTox: Keywords", "SeaTox: Protein Families")
AnnotSeaTox<-AnnotSeaTox[order(AnnotSeaTox[,"SeaTox: Blastp_Evalue"], -
AnnotSeaTox[,"SeaTox: Blastp_%ID"]),]

##Circles with expression levels (TPMs)
for(i in 1:nrow(Circles)){
  assign(Circles[i,1], NumberOfCircles(Circles[i,1], Circles[i,2], AnnotSeaTox,
TPM))}

data<-data.frame(rbind(Toxin, Hormone, `Serine protease`, Metalloprotease, `Glycosyl
hydrolase 18 family`, `Glycosyl hydrolase 56 family`))
rownames(data)[5:6]<-c("Chitinase", "Hyaluronidase")
colnames(data)<-c("Circles", "Mean")

# Function that return a dataframe. 1 line per bubble with the center, radius is
proportional of to the expression value
packing <- circleProgressiveLayout(data$Mean, sizetype= 'area')
```

```
# Add packing data to the object "data"
data <- cbind(data, packing)
#Coordinates of a circle that is drawn by a "npoints" of straight lines (10000)
dat.gg <- circleLayoutVertices(packing, npoints=10000)
# Create plot
quartz()
ggplot() +
  # Create the bubbles
  geom_polygon(data = dat.gg, aes(x, y, group = id, fill=as.factor(id)), alpha = 0.7)
+
  # Add text in each bubble and control the size of the text
  #fontface=2 put the letters in bold type
  geom_text(data = data, aes(x, y, size=Mean, label = rownames(data), fontface=2)) +
  geom_text(data = data, aes(x, y-1.1, size=80, label = Circles, fontface=2)) +
  scale_size_continuous(range = c(1,4)) +
  theme_void() +
  scale_fill_manual(values = c("#F8766D", "gold1", "#00BA38", "#00BFC4", "orange",
"purple3"))+
  theme(legend.position= "none") +
  coord_equal()
ggsave(paste(pathCircles, "CirclesExpression.svg", sep = ""), height = 15, width =
15,  units = 'cm', dpi=600)

##Plot with expression (with TPMs)####
##Function
ExpressionPlot<-function(blastcpmGraph){
  log2TPM<-expression(paste("Average log"[2], "(TPM + 0.1)"))
  log10evalue<-expression(paste("-log"[10], "(", italic("e"), "-value)"))
  colour<-c("#83d6fc", "#A855B0", "#fcb683")
  legend<-c("CRVP", "CTX", "VIns")
  blastcpmGraph$Evalue<- -log10(blastcpmGraph$Evalue)
  ##Inf does not appear in the graph, so we changed it to 180
  for(i in 1:nrow(blastcpmGraph)){
    if(blastcpmGraph$Evalue[i] == "Inf")
      blastcpmGraph$Evalue[i]<-180  }
  quartz()
  plot(
    blastcpmGraph$Evalue,
    blastcpmGraph$log2TPM,
    pch = ".",
    xlab = log10evalue,
    ylab = log2TPM)
  ##Highlighting the points that correspond to the transcripts of interest

points(blastcpmGraph["TRINITY_DN23540_c1_g6_i2.p1",2],blastcpmGraph["TRINITY_DN2354
0_c1_g6_i2.p1",1],pch = 20,cex = 2,col = c("#9BBB59"))
points(blastcpmGraph["TRINITY_DN24442_c10_g1_i1.p1",2],blastcpmGraph["TRINITY_DN244
42_c10_g1_i1.p1",1],pch = 20,cex = 2,col = c("#CCC1DA"))
points(blastcpmGraph["TRINITY_DN24442_c10_g1_i4.p1",2],blastcpmGraph["TRINITY_DN244
42_c10_g1_i4.p1",1],pch = 20,cex = 2,col = c("#CCC1DA"))
points(blastcpmGraph["TRINITY_DN23515_c5_g1_i13.p1",2],blastcpmGraph["TRINITY_DN235
15_c5_g1_i13.p1",1],pch = 20,cex = 2,col = c("#F79646"))
recordPlot()}

##RConsole
load(paste(pathExpression, "blastcpmGraph.RData", sep = ""))
load(paste(pathTPM, "CountsAbundance.RData", sep = ""))
blastID<-unlist(sapply(rownames(blastcpmGraph), function(x) unlist(strsplit(x, ".",
fixed = TRUE))[1]))
blastcpmGraph<-cbind(blastcpmGraph, blastID, rownames(blastcpmGraph))

BlastTPMGraph<-merge(TPM, blastcpmGraph, by.x = "row.names", by.y = "blastID")
rownames(BlastTPMGraph)<-BlastTPMGraph[,8]

##Calculate log2 of TPMs
```

```
BlastTPMGraph$logSP6<-log2(BlastTPMGraph$KallistoALL_SP6)
BlastTPMGraph$logSP7<-log2(BlastTPMGraph$KallistoALL_SP7)
BlastTPMGraph$logSP8<-log2(BlastTPMGraph$KallistoALL_SP8)
##Calculate the mean of log2TPM for each ORFs
for(i in 1:nrow(BlastTPMGraph)){
  BlastTPMGraph$log2TPM[i]<-mean(as.numeric(BlastTPMGraph[i,9:11]))}
BlastTPMGraphaux<-BlastTPMGraph
rownames(BlastTPMGraph)<-BlastTPMGraph[,8]
##BlastTPMGraph becames a dataframe only with two collums: log2TPM e evalue (from
blast)
BlastTPMGraph<-BlastTPMGraph[,c(12,7)]

##Expression Plot with TPMs
ExpressionGraph<-ExpressionPlot(BlastTPMGraph)
plot_grid(ExpressionGraph)
dev.print(tiff, paste(pathExpression, "ExpressionPlotAve.tif", sep = ""), height =
15, width = 15,  units = 'cm', res=600)
ggsave(paste(pathExpression, "ExpressionPlotAve.svg", sep = ""), height = 15, width
= 15,  units = 'cm', dpi=600)
save(BlastTPMGraphaux,     BlastTPMGraph,     file     =     paste(pathExpression,
"BlastTPMGrapAve.RData", sep = ""))

##Pairs Plot####
##Plot several genes at once (pair plot)
##RConsole
load(paste(pathExpression, "BlastTPMGrapAve.RData", sep = ""))
quartz()
##BlastTPMPairs is a dataframe with logTPMs for every ORFs in each sample
BlastTPMPairs<-BlastTPMGraphaux[,c(9:11)]
BlastTPMPairs$TrinityID<-BlastTPMGraphaux[,8]
colnames(BlastTPMPairs)[1:3]<-c("SP6", "SP7", "SP8")

##Creating a new collumn where the transcripts of interest are highlighted
for(i in 1:nrow(BlastTPMPairs)){
  if(BlastTPMPairs$TrinityID[i] == "TRINITY_DN23540_c1_g6_i2.p1"){
    BlastTPMPairs$Transcripts[i]<-"CRISP"}
  else if(BlastTPMPairs$TrinityID[i] == "TRINITY_DN24442_c10_g1_i1.p1"){
    BlastTPMPairs$Transcripts[i]<-"CTX"}
  else if(BlastTPMPairs$TrinityID[i] == "TRINITY_DN24442_c10_g1_i4.p1"){
    BlastTPMPairs$Transcripts[i]<-"CTX"}
  else if(BlastTPMPairs$TrinityID[i] == "TRINITY_DN23515_c5_g1_i13.p1"){
    BlastTPMPairs$Transcripts[i]<-"VIns"}
  else
    BlastTPMPairs$Transcripts[i]<-"Other"}
##Change the order of ORFs, so we could decide the colours better
BlastTPMPairs<-BlastTPMPairs[c(1:8701,
8703:8769,8771:11354,11356,11358:16158,8770,11355,11357,8702),]
BlastTPMPairs<-subset(BlastTPMPairs, !is.na(BlastTPMPairs$Transcripts))
colors<-c("#9BBB59",                    "#CCC1DA",                    "black",
"#F79646")[unclass(as.factor(BlastTPMPairs$Transcripts))]

##Pairs Plot
quartz()
pairs.panels(BlastTPMPairs[,c(1,2,3)], scale = TRUE, pch=21, bg = colors, ellipses =
FALSE, smooth = FALSE)
dev.print(tiff, paste(pathTPM, "pairsTPM.tiff", sep = ""), height = 15, width = 15,
units = 'cm', res=600)
save(BlastTPMPairs, file = paste(pathTPM, "BlastTPMPairsTPM.RData", sep = ""))

##Retrieving TPM of transcripts of interest####
##RConsole
TxInterest<-BlastTPMGraph[c("TRINITY_DN23540_c1_g6_i2.p1",
"TRINITY_DN24442_c10_g1_i1.p1",                    "TRINITY_DN24442_c10_g1_i4.p1",
"TRINITY_DN23515_c5_g1_i13.p1"),1]
TxInterest<-as.data.frame(TxInterest)
```

```
rownames(TxInterest)<-c("TRINITY_DN23540_c1_g6_i2.p1",
"TRINITY_DN24442_c10_g1_i1.p1",                    "TRINITY_DN24442_c10_g1_i4.p1",
"TRINITY_DN23515_c5_g1_i13.p1")
colnames(TxInterest)<-"log2(TPM+0.1)"
write.table(TxInterest,    paste(pathTPM,  "TxInterest.csv",  sep  =  ""),  sep=";",
col.names=NA)

##Percentage of complete ORFs####
##RConsole
##Save the sequences of ORFs in a RData object
TransDecoderfile<-c("/[PATH]/TransDecoder/Trinity.fasta.transdecoder.pep")
Fa<-read.fasta(TransDecoderfile, seqtype = "AA", as.string = TRUE)
ID<-getName(Fa)
pepSEQ<-unlist(getSequence(Fa, as.string=TRUE))
pepSEQ<-as.data.frame(cbind(ID, pepSEQ))
save(pepSEQ, file = paste(pathAnnot, "pepSEQ.RData", sep = ""))

##Analysing the size of the ORFs, if they have start codon and a stop codon
pep<-pepSEQ
pep$length<-nchar(pep$pepSEQ)
pep$M<-ifelse(substring(pep$pepSEQ, 1,1) == "M", "M", "")
pep$Stop<-ifelse(grepl("*", pep$pepSEQ, fixed = TRUE), "STOP", "")
pep$length<-ifelse(pep$Stop == "STOP", pep$length - 1, pep$length)
CompletePep<-pep[which(pep$M == "M" & pep$Stop == "STOP"),]
PercentageCompletePep<-nrow(CompletePep)/nrow(pep)*100
save(pep,    CompletePep,    PercentageCompletePep,    file    =    paste(pathAnnot,
"Completepep.RData", sep = ""))

##Domain analysis:Pfam####
##Function
##Reading the Pfam results
readpfam<-function(pathPfam, filepfam){
  pfam<-read.table(filepfam, sep="\t", header=F, fill = T, quote="")
  pfam<-DescriptionTarget(pfam)
  save(pfam, file = paste(pathPfam, "pfamAll.RData", sep = ""))
  ##Applying a cut-off: e-value < 0.05
  pfam<-pfam[which(pfam$Evalue < 0.05),]
  pfam<-pfam[order(pfam$Evalue),]
  rownames(pfam)<-NULL
  print(nrow(pfam))
  save(pfam, file = paste(pathPfam, "pfamEvalue.RData", sep = ""))
  write.table(pfam,     paste(pathPfam,   "pfamEvalue.csv",   sep   =   ""),   sep=";",
col.names=NA)}

##The last column (Description factors) from the output of the pfam analysis is not
formatted. Each word is in a different column, so this function is going to search
for all the words from a domain and put them all together in a single column
DescriptionTarget<-function(pfam){
  TrinityID<-grepl("TRINITY", pfam[,4])
  descriptionTarget<-matrix(data = NA, nrow = nrow(pfam), ncol = 1)
  for (i in 1:nrow(pfam)){
    if(TrinityID[i]){
      lastCol = 23
      for(j in 24:ncol(pfam)){
        if(pfam[i,j] == "")
          break
        else
          lastCol = lastCol + 1}
      descriptionTarget[i,1]<-paste(pfam[i,23:lastCol], collapse = " ")
      if(i != nrow(pfam) && TrinityID[i+1] == FALSE){
        lastCol1 = 1
        for(j in 2:ncol(pfam)){
          if(is.na(pfam[i+1,j]) || pfam[i+1,j] == "")
            break
          else
```

```
            lastCol1 = lastCol1 + 1}
        descriptionAux<-paste(pfam[i+1,1:lastCol1], collapse = " ")
        descriptionTarget[i,1]<-paste(descriptionTarget[i,1], descriptionAux)}}}
  pfam<-cbind(pfam, descriptionTarget)
  TrinityIDaux<-which(TrinityID == FALSE)
  pfam<-pfam[-TrinityIDaux,-c(23:(ncol(pfam)-1)),drop=FALSE]
  colnames(pfam)<-c("Domain", "Accession", "tlen", "TrinityID", "accession", "qlen",
"Evalue", "score", "bias", "#", "of", "cevalue", "ievalue", "score", "bias", "from(hmm
coord)", "to(hmm coord)", "from(ali coord)", "to(ali coord)", "from(env coord)",
"to(env coord)", "acc", "DescriptionTarget")
  pfam$Evalue<-as.numeric(pfam$Evalue)
  return(pfam)}

domainRelative<-function(pathPfamSeaTox, pfam, pepSEQ, domainsList){
  ##Selects the domains of interest from all the results
  pfam<-pfam[pfam$Domain %in% domainsList,]
  ##Calculate the percentage of ORFs that a specific domain
  domainsUnique<-unique(pfam[,c("Domain", "Accession", "DescriptionTarget")])
  domains<-sapply(domainsUnique$Domain,                                    function(x)
length(unique(pfam[which(pfam$Domain %in% x), "TrinityID"])))
  domains<-domains[order(-domains)]
  domainsPercentage<-as.data.frame(domains/nrow(pepSEQ)*100)
  colnames(domainsPercentage)<-c("Percentage")
  ##Calculates the mean of the evalues of each domain
  domainsEvalue<-aggregate(pfam$Evalue, list(pfam$Domain), mean, na.rm=TRUE)
  colnames(domainsEvalue)<-c("domains", "E-value")
  domainsAll<-merge(domainsPercentage, domainsEvalue, by.x = "row.names", by.y = 1)
  write.table(domainsAll, paste(pathPfamSeaTox, "PfamCountProteinEvalue.csv", sep =
""), sep=";", col.names=NA)
  save(domains, domainsPercentage, domainsEvalue, domainsAll, file =
paste(pathPfamSeaTox, "PfamCountProteinEvalue.RData", sep = ""))
  return(domainsAll)}

##RConsole
##Converting the output of hmmer in a file that can be read by the function read.table
system("tr -s ' ' '\t' < /[PATH]/SepiaTrim_pfam.domtblout >
/[PATH]/Pfam/SepiaTrim_pfam_cor.domtblout")
##Reading the Pfam results
readpfam(pathPfam, filepfam[1,1])

load(paste(pathAnnot, "pepSEQ.RData", sep = ""))
load(paste(pathPfam,"pfamEvalue.RData", sep = ""))
load(paste(pathAnnot, "5DB.RData", sep = ""))

##Merge the results pfam with SeaTox dataframe
pfamSeaTox<-merge(SeaTox, pfam, by = "TrinityID")
save(pfamSeaTox, file = paste(pathPfamSeaTox, "pfamSeaTox.RData", sep = ""))

##Domains of interest
domainsList<-c("Astacin", "CAP", "CUB", "CUB_2", "Glyco_hydro_18", "Kazal_1",
"Kazal_2","Kunitz_BPTI", "Lectin_C","Pep_M12B_propep","Peptidase_M13",
"Peptidase_M13_N", "Reprolysin", "Reprolysin_2", "Reprolysin_3", "Reprolysin_4",
"Reprolysin_5", "Serpin", "ShK", "Trypsin", "Trypsin_2",
"TSP1_ADAMTS","Bee_toxin","Chi-conotoxin","Toxin_13", "Toxin_16",
"Toxin_18","Toxin_23", "Toxin_29", "Toxin_32", "Toxin_37", "Toxin_7", "Toxin_9",
"Glyco_hydro_56", "DUF316","ADAMTS_spacer1", "ADAMTS_CR_2", "ADAM17_MPD",
"ADAMTS_CR_3", "Insulin", "TSP_1", "Ldl_recept_a")

##Calculates the Percentage of ORFs that have a specific domain and the mean of
evalues
domains<-domainRelative(pathPfamSeaTox, pfamSeaTox, pepSEQ, domainsList)

##Circles plot with the domains of interest, where the size is proportional to the
percentage of ORFs that have that domain and are annotated against SeaTox
load(paste(pathPfamSeaTox, "PfamCountProteinEvalue.RData", sep = ""))
```

```
domains<-cbind(rownames(domainsPercentage), domainsPercentage)
domains2<-domains[sample(nrow(domains)),]
domains<-domains2

# Function that return a dataframe. 1 line per bubble with the center, radius
(proportional of the number of occurrences. Area that should be proportionnal to the
occurrences number
packing <- circleProgressiveLayout(domains$Percentage, sizetype= 'area')
# Add packing data to the object “data”
domains <- cbind(domains, packing)
#Coordinates of a circle that is drawn by a “npoints” of straight lines (10000)
dat.gg <- circleLayoutVertices(packing, npoints=10000)
# Create plot
quartz()
ggplot() +
  # Create the bubbles
  geom_polygon(data = dat.gg, aes(x, y, group = id, fill=as.factor(id)), alpha = 0.7)
+
  # Add text in each bubble and control the size of the text
  #fontface=2 put the letters in bold type
  geom_text(data = domains, aes(x, y, size = Percentage, label = rownames(domains),
fontface=2)) +
  scale_size_continuous(range = c(1,4)) +
  theme_void() +
  theme(legend.position= "none") +
  coord_equal()
dev.print(tiff, paste(pathPfamSeaTox, "CirclesPfamRandom.tif", sep = ""), height =
15, width = 15,  units = 'cm', res=600)
ggsave(paste(pathPfamSeaTox, "CirclesPfamRandom.svg", sep =""), height = 15, width =
15,  units = 'cm', dpi=600)
save(domains, file = paste(pathPfamSeaTox, "domainsRandom.RData", sep = ""))

##Server####
##Running Hmmer
##Building databases
/[PATH]/hmmer/bin/hmmpress /[PATH]/databases/pfam/Pfam-A.hmm
##Searching against Pfam
/[PATH]/hmmer/bin/hmmscan --cpu 20 --seed 1 --domtblout SepiaTrim_pfam.domtblout --
tblout            SepiaTrim_pfam.tblout            /[PATH]/databases/pfam/Pfam-A.hmm
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep

##Running Blast
##Building databases
/[PATH]/ncbi-blast-2.13.0+/bin/makeblastdb                                    -in
/[PATH]/Databases/uniprot_SwissProt.fasta -out /[PATH]/Databases/uniprot_SwissProt -
dbtype prot
/[PATH]/ncbi-blast-2.13.0+/bin/makeblastdb                                    -in
/[PATH]/Databases/uniprot_database_seatox_2023_03.fasta                      -out
/[PATH]/Databases/uniprot_database_seatox_2023_03 -dbtype prot
/[PATH]/ncbi-blast-2.13.0+/bin/makeblastdb                                    -in
/[PATH]/Databases/uniprotkb_secreted.fasta -out /[PATH]/Databases/uniprotkb_secreted
-dbtype prot
/[PATH]/ncbi-blast-2.13.0+/bin/makeblastdb                                    -in
/[PATH]/Databases/uniprotkb_Toxin.fasta  -out  /[PATH]/Databases/uniprotkb_Toxin  -
dbtype prot
/[PATH]/ncbi-blast-2.13.0+/bin/makeblastdb                                    -in
/[PATH]/Databases/uniprotkb_enzyme.fasta  -out  /[PATH]/Databases/uniprotkb_enzyme  -
dbtype prot
/[PATH]/ncbi-blast-2.13.0+/bin/makeblastdb -in /[PATH]/Databases/uniprotkb_PTM.fasta
-out /[PATH]/Databases/uniprotkb_PTM -dbtype prot

##Running BlastP against different databases
/[PATH]/ncbi-blast-2.13.0+/bin/blastp                                      -query
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep         -db
/[PATH]/Databasesuniprot_SwissProt -max_target_seqs 1 -max_hsps 1 -outfmt '10 delim=;
```

```
qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore
sacc      qcovs      stitle'      -evalue      1e-5      -num_threads      20      >
/[PATH]/SepiaTrim_Blastp_uniprot_SwissProt.csv
/[PATH]/ncbi-blast-2.13.0+/bin/blastp                                          -query
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep            -db
/[PATH]/Databases/uniprot_database_seatox_2023_03 -max_target_seqs 1 -max_hsps 1 -
outfmt '10 delim=; qseqid sseqid pident length mismatch gapopen qstart qend sstart
send evalue bitscore sacc qcovs stitle' -evalue 1e-5 -num_threads 20 >
/[PATH]/SepiaTrim_Blastp_uniprot_database_seatox.csv
/[PATH]/ncbi-blast-2.13.0+/bin/blastp                                          -query
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep            -db
/[PATH]/Databases/uniprotkb_secreted -max_target_seqs 1 -max_hsps 1 -outfmt '10
delim=; qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue
bitscore    sacc    qcovs    stitle'    -evalue    1e-5    -num_threads    20    >
/[PATH]/SepiaTrim_Blastp_uniprot_secreted.csv
/[PATH]/ncbi-blast-2.13.0+/bin/blastp                                          -query
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep            -db
/[PATH]/Databases/uniprotkb_Toxin -max_target_seqs 1 -max_hsps 1 -outfmt '10 delim=;
qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore
sacc      qcovs      stitle'      -evalue      1e-5      -num_threads      20      >
/[PATH]/SepiaTrim_Blastp_uniprot_toxin.csv
/[PATH]/ncbi-blast-2.13.0+/bin/blastp                                          -query
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep            -db
/[PATH]/Databases/uniprotkb_enzyme -max_target_seqs 1 -max_hsps 1 -outfmt '10 delim=;
qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore
sacc qcovs stitle' -evalue 1e-5 -num_threads 20 > /[PATH]/SepiaTrim_Blastp_enzyme.csv
/[PATH]/ncbi-blast-2.13.0+/bin/blastp                                          -query
/[PATH]/TransDecoder_out/ALL_SepiaORFs/Trinity.fasta.transdecoder.pep            -db
/[PATH]/Databases/uniprotkb_PTM -max_target_seqs 1 -max_hsps 1 -outfmt '10 delim=;
qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore
sacc qcovs stitle' -evalue 1e-5 -num_threads 20 > /[PATH]/SepiaTrim_Blastp_PTM.csv
```

B

Appendix to Chapter 4

B.1 Materials and Methods: Total RNA extraction and RNA-Seq

Table B.1 - Number of reads per sample.

Salivary gland	Sample	Number of reads	Mean read length (bp)
Anterior	1	43,819,778	150
	2	40,654,198	
	3	42,552,744	
Posterior	1	43,468,766	
	2	44,934,238	
	3	44,457,874	

B.2 *R* Script

```
##Load necessary R packages
library(tximport)
library(limma)
library(edgeR)
library(seqinr)
library(UniprotR)
library(ggplot2)
library(gplots)
library(biomaRt)
library(cowplot)
library(dplyr)
library(viridis)
library(hrbrthemes)
library(scales)
library(RColorBrewer)
display.brewer.all()


folder<-"C:\\Users\\Cátia\\Documents\\PhD_Bioinformatics\\Kallisto\\"
setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary")

#Samples
sampleName<-c(
  "OA3",
  "OA5",
  "OA8",
  "OP3",
  "OP6",
  "OP9")

organType<-c(   "Anterior Salivary Gland",   "Anterior Salivary Gland",   "Anterior
Salivary Gland",  "Posterior Salivary Gland",  "Posterior Salivary Gland",  "Posterior
Salivary Gland")

octopLevel<-c("ASG","ASG","ASG","PSG","PSG","PSG")


dataMatrix<-as.data.frame(
  cbind(
    sampleName,
    organType  ))

write.table(dataMatrix, "dataMatrix.csv", sep=";", col.names=NA)

#Data table
samples<-dir(folder)
file<-c(paste(sep="",folder,samples,"\\","abundance.h5"))
names(file)=samples
names(file)

#tximport:
octopTsv<-tximport(file,  type  =  "kallisto",  txOut  =  TRUE,  countsFromAbundance  =
"lengthScaledTPM")
head(octopTsv$counts)

dim(octopTsv$counts) ##302475
colnames(octopTsv$counts) ##6

#Transcriptome
file<-"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Trinity.fasta\\"
```

```
octopFa<-read.fasta(file, as.string = TRUE)
head(octopFa)
octopTxID<-getName(octopFa)
octopTxSEQ<-unlist(getSequence(octopFa,as.string=TRUE))
octopTxSEQ<-as.data.frame(cbind(octopTxID,octopTxSEQ))
head(octopTxSEQ)
colnames(octopTxSEQ)[1]<-"ID"
nrow(octopTxSEQ)
octopCounts<-octopTsv$counts
octopCounts<-cbind(as.data.frame(row.names(octopCounts)),octopCounts)
colnames(octopCounts)[1]<-"ID"
octopFull <- merge(octopCounts,octopTxSEQ,by="ID")
colnames(octopFull)<-c("ID", sampleName, "sequence")
write.table(octopFull, "octopFull.csv", sep=";", col.names=NA)

#Statistics
octopData<-DGEList(counts=octopFull[,2:7], group=octopLevel,genes=octopFull[,1])
octopData<-calcNormFactors(octopData)
octopDesign<-model.matrix(~0+octopLevel,data=octopData$samples)
colnames(octopDesign) <- levels(octopData$samples$group)
octopData<-estimateDisp(octopData,octopDesign)
octopfit <- glmFit(octopData,octopDesign)

ASGvPSG <- makeContrasts(ASG-PSG, levels=octopDesign)
PSGvASG <- makeContrasts(PSG-ASG, levels=octopDesign)
octoplrtASG <- glmLRT(octopfit,contrast=ASGvPSG)
topTags(octoplrtASG)
octoplrtPSG <- glmLRT(octopfit,contrast=PSGvASG)
topTags(octoplrtPSG)
octopResults<-cbind(octopFull[,1], octoplrtASG$table, octoplrtPSG$table)
colnames(octopResults)<-c("ID",       "logFC-ASG",       "logCPM-ASG","LR-ASG","FDRp-
ASG","logFC-PSG","logCPM-PSG","LR-PSG","FDRp-PSG")
write.table(octopResults,"octopResults.csv",sep=";",col.names=NA)

expressionTable<-decideTests(octopResults[,grepl("FDRp",
colnames(octopResults))],coefficients=octopResults[,grepl("logFC",
colnames(octopResults))],lfc=1.5,adjust.method="fdr")
colnames(expressionTable)<-c("ASG","PSG")
head(expressionTable)
nrow(expressionTable)
degTable<-octopDEG[which(abs(octopDEG$ASG) == 1 | abs(octopDEG$PSG) == 1),]
head(degTable)
nrow(degTable)
write.table(degTable,"degTable.csv",sep=";",col.names=NA)
write.table(octopDEG,"octopDEG.csv",sep=";",col.names=NA)


################################################################################
#####################       Annotation SwissProt     ###########################
################################################################################


#Read Blastp for swissprot results
blastp_swissP                                                                  <-
as.matrix(read.table("C:\\Users\\Cátia\\Documents\\PhD_Bioinformatics\\Octopus_ORFs
_blastp_uniprot_sprot.outfmt6", sep="\t", header=F))
blastpID  <-  unlist(sapply(blastp_swissP[,1],  function(x)  unlist(strsplit(x,".",
fixed=TRUE))[1]))
blastp_swissP <- cbind(blastpID, blastp_swissP)
rownames(blastp_swissP) <- NULL
blastp_swissP <- blastp_swissP[order(as.numeric(blastp_swissP[,12])),]
```

```
##Read degTable
file1<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\degTable.csv
"
degTable<-as.matrix(read.table(file1,sep=";",header=T))
colnames(degTable)<-c("X",     "ID",     "logFC-ASG",     "logCPM-ASG","LR-ASG","FDRp-
ASG","logFC-PSG","logCPM-PSG","LR-PSG","FDRp-PSG", "ASG", "PSG")

##Merge degTable & Blastp_swissP tables
degAnnotAll<-merge(degTable,blastp_swissP,by.x="ID",by.y="blastpID")
write.table(degAnnotAll, "OctopdegAnnotAll.csv", sep=";", col.names=NA)

####For PSG####
##PSG Annot Table
degAnnotPSG<-matrix(data=NA,             nrow=nrow(degAnnotAll),             ncol=5,
dimnames=list(degAnnotAll[,13],  c("logFC-PSG","FDRp-PSG",  "Blastp_AccessionNumber",
"Blastp_%ID", "Blastp_Evalue")))
for(i in 1:nrow(degAnnotAll)){
  degAnnotPSG[i,1:2]<-c(degAnnotAll[i,7],degAnnotAll[i,10])
  degAnnotPSG[i,3:5]<-c(degAnnotAll[i,14],degAnnotAll[i,15],degAnnotAll[i,23])}
write.table(degAnnotPSG, "OctopdegAnnotPSG.csv", sep=";", col.names=NA)

####For PSG####
##Sorted by e-value and %ID: PSG
degAnnotOrderedPSG<-degAnnotPSG[order(as.numeric(degAnnotPSG[,5]),-
as.numeric(degAnnotPSG[,4])),]
write.table(degAnnotOrderedPSG,       "OctopusPSGdegAnnotOrdered.csv",       sep=";",
col.names=NA)

####For PSG####
##Unique Trinity-ID and Top10Match: PSG
TrinityIDsPSG<-unique(rownames(degAnnotOrderedPSG))
degAnnotTop10MatchPSG<-matrix(data=NA,      nrow=length(TrinityIDsPSG),      ncol=5,
dimnames=list(TrinityIDsPSG,  c("logFC-PSG",  "FDRp-PSG",  "Blastp_AccessionNumber",
"Blastp_%ID", "Blastp_Evalue")))
for(j in 1:length(TrinityIDsPSG)){
  allHits<-degAnnotOrderedPSG[rownames(degAnnotOrderedPSG) == TrinityIDsPSG[j],]
  if(is.vector(allHits) == F) degAnnotTop10MatchPSG[j,]<-allHits[1,]
  else degAnnotTop10MatchPSG[j,]<-allHits}
nrow(degAnnotTop10MatchPSG)
write.table(degAnnotTop10MatchPSG,        "OctoPSGdegAnnotMatch.csv",        sep=";",
col.names=NA)
write.table(degAnnotTop10MatchPSG[1:10,], "OctoPSGdegAnnotTop10Match.csv", sep=";",
col.names=NA)

####For ASG####
##ASG Annot Table
degAnnotASG<-matrix(data=NA,             nrow=nrow(degAnnotAll),             ncol=5,
dimnames=list(degAnnotAll[,13],  c("logFC-ASG","FDRp-ASG",  "Blastp_AccessionNumber",
"Blastp_%ID", "Blastp_Evalue")))
for(i in 1:nrow(degAnnotAll)){
  degAnnotASG[i,1:2]<-c(degAnnotAll[i,3],degAnnotAll[i,6])
  degAnnotASG[i,3:5]<-c(degAnnotAll[i,14],degAnnotAll[i,15],degAnnotAll[i,23])}
write.table(degAnnotASG, "OctopdegAnnotASG.csv", sep=";", col.names=NA)

####For ASG####
##Sorted by e-value and %ID: ASG
degAnnotOrderedASG<-degAnnotASG[order(as.numeric(degAnnotASG[,5]),-
as.numeric(degAnnotASG[,4])),]
write.table(degAnnotOrderedASG,       "OctopusASGdegAnnotOrdered.csv",       sep=";",
col.names=NA)

####For ASG####
##Unique Trinity-ID and Top10Match: ASG
TrinityIDsASG<-unique(rownames(degAnnotOrderedASG))
```

```
degAnnotTop10MatchASG<-matrix(data=NA,      nrow=length(TrinityIDsASG),      ncol=5,
dimnames=list(TrinityIDsASG,  c("logFC-ASG",  "FDRp-ASG",  "Blastp_AccessionNumber",
"Blastp_%ID", "Blastp_Evalue")))
for(j in 1:length(TrinityIDsASG)){
  allHits<-degAnnotOrderedASG[rownames(degAnnotOrderedASG) == TrinityIDsASG[j],]
  if(is.vector(allHits) == F) degAnnotTop10MatchASG[j,]<-allHits[1,]
  else degAnnotTop10MatchASG[j,]<-allHits}
nrow(degAnnotTop10MatchASG)
write.table(degAnnotTop10MatchASG,        "OctoASGdegAnnotMatch.csv",        sep=";",
col.names=NA)
write.table(degAnnotTop10MatchASG[1:10,],  "OctoASGdegAnnotTop10Match.csv",  sep=";",
col.names=NA)


################################################################################
##############################  UniprotR  ######################################
################################################################################


##Get from UniprotR: Entry_Name, Gene_Name, Protein_Name, Organism, Subcell_Location,
GOTerms
##Unique Accession
##Uniprot: Accession without version

####For PSG####
file2<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\OctoPSGdegAn
notMatch.csv"
degAnnotPSG<-as.matrix(read.table(file2, sep=";", header=T))
colnames(degAnnotPSG)<-c("TrinityID","logFC_PSG","FDRp_PSG",     "Blastp_Accession",
"Blastp_%ID", "Blastp_Evalue")
AccessionPSG<-sapply(degAnnotPSG[,4],       function(x)       unlist(strsplit(x,".",
fixed=TRUE))[1])
degAnnotPSG<-cbind(degAnnotPSG,AccessionPSG)
AccessionUniquePSG<-unique(degAnnotPSG[,7])
length(AccessionUniquePSG)
NameTaxa<-GetNamesTaxa(AccessionUniquePSG)
SubcellLocation<-GetSubcellular_location(AccessionUniquePSG)
GOTerms<-GetProteinGOInfo(AccessionUniquePSG)
AccessionInfoAllPSG<-cbind(NameTaxa, SubcellLocation, GOTerms)
nrow(AccessionInfoAllPSG)
write.table(AccessionInfoAllPSG,       "OctoPSGAccessionInfoAll.csv",       sep=";",
col.names=NA)
save(NameTaxa,    SubcellLocation,    GOTerms,    AccessionInfoAllPSG,    file    =
"OctoPSGAnnot.RData")

####For ASG####
file3<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\OctoASGdegAn
notMatch.csv"
degAnnotASG<-as.matrix(read.table(file3, sep=";", header=T))
colnames(degAnnotASG)<-c("TrinityID","logFC_ASG","FDRp_ASG",     "Blastp_Accession",
"Blastp_%ID", "Blastp_Evalue")
AccessionASG<-sapply(degAnnotASG[,4],       function(x)       unlist(strsplit(x,".",
fixed=TRUE))[1])
degAnnotASG<-cbind(degAnnotASG,AccessionASG)
AccessionUniqueASG<-unique(degAnnotASG[,7])
length(AccessionUniqueASG)
NameTaxa<-GetNamesTaxa(AccessionUniqueASG)
SubcellLocation<-GetSubcellular_location(AccessionUniqueASG)
GOTerms<-GetProteinGOInfo(AccessionUniqueASG)
AccessionInfoAllASG<-cbind(NameTaxa, SubcellLocation, GOTerms)
nrow(AccessionInfoAllASG)
write.table(AccessionInfoAllASG,       "OctoASGAccessionInfoAll.csv",       sep=";",
col.names=NA)
```

```
save(NameTaxa, SubcellLocation, GOTerms, AccessionInfoAllASG, file =
"OctoASGAnnot.RData")

####For PSG####
##Merge
degAnnotTotalPSG<-merge(degAnnotPSG, AccessionInfoAllPSG, by.x="AccessionPSG",
by.y="row.names", sort=FALSE)
degAnnotTotalOrderedPSG<-degAnnotTotalPSG[order(as.numeric(degAnnotTotalPSG[,7]),-
as.numeric(degAnnotTotalPSG[,6])),]
nrow(degAnnotTotalPSG)
nrow(degAnnotTotalOrderedPSG)
write.table(degAnnotTotalPSG, "OctopusPSGdegAnnotTotal.csv", sep=";", col.names=NA)
write.table(degAnnotTotalOrderedPSG, "OctopusPSGdegAnnotTotalSorted.csv", sep=";",
col.names=NA)

degAnnotInfoPSG<-matrix(data=NA, nrow=nrow(degAnnotTotalOrderedPSG), ncol=13,
dimnames=list(degAnnotTotalOrderedPSG[,2], c("logFC_PSG","FDRp_PSG",
"Blastp_Accession", "Blastp_%ID", "Blastp_Evalue", "Entry_Name", "Gene_Name",
"Protein_Name", "Organism", "Subcell_Location", "GO_BiologicalProcess",
"GO_MolecularFunction", "GO_CelularComponent")))
for(i in 1:nrow(degAnnotTotalOrderedPSG)){
  degAnnotInfoPSG[i,1:5]<-c(degAnnotTotalOrderedPSG[i,3],
degAnnotTotalOrderedPSG[i,4], degAnnotTotalOrderedPSG[i,5],
degAnnotTotalOrderedPSG[i,6], degAnnotTotalOrderedPSG[i,7])
  degAnnotInfoPSG[i,6:9]<-c(degAnnotTotalOrderedPSG[i,8],
degAnnotTotalOrderedPSG[i,10], degAnnotTotalOrderedPSG[i,16],
degAnnotTotalOrderedPSG[i,14])
  degAnnotInfoPSG[i,10]<-sapply(degAnnotTotalOrderedPSG[i,20], function(x)
unlist(strsplit(x,"SUBCELLULAR LOCATION: "))[2])
  degAnnotInfoPSG[i,11:13]<-c(degAnnotTotalOrderedPSG[i,26],
degAnnotTotalOrderedPSG[i,27], degAnnotTotalOrderedPSG[i,28])}
nrow(degAnnotInfoPSG)
write.table(degAnnotInfoPSG, "OctopusPSGdegAnnotInfo.csv", sep=";", col.names=NA)


####For ASG####
##Merge
degAnnotTotalASG<-merge(degAnnotASG, AccessionInfoAllASG, by.x="AccessionASG",
by.y="row.names", sort=FALSE)
degAnnotTotalOrderedASG<-degAnnotTotalASG[order(as.numeric(degAnnotTotalASG[,7]),-
as.numeric(degAnnotTotalASG[,6])),]
nrow(degAnnotTotalASG)
nrow(degAnnotTotalOrderedASG)
write.table(degAnnotTotalASG, "OctopusASGdegAnnotTotal.csv", sep=";", col.names=NA)
write.table(degAnnotTotalOrderedASG, "OctopusASGdegAnnotTotalSorted.csv", sep=";",
col.names=NA)

degAnnotInfoASG<-matrix(data=NA, nrow=nrow(degAnnotTotalOrderedASG), ncol=13,
dimnames=list(degAnnotTotalOrderedASG[,2], c("logFC_ASG","FDRp_ASG",
"Blastp_Accession", "Blastp_%ID", "Blastp_Evalue", "Entry_Name", "Gene_Name",
"Protein_Name", "Organism", "Subcell_Location", "GO_BiologicalProcess",
"GO_MolecularFunction", "GO_CelularComponent")))
for(i in 1:nrow(degAnnotTotalOrderedASG)){
  degAnnotInfoASG[i,1:5]<-c(degAnnotTotalOrderedASG[i,3],
degAnnotTotalOrderedASG[i,4], degAnnotTotalOrderedASG[i,5],
degAnnotTotalOrderedASG[i,6], degAnnotTotalOrderedASG[i,7])
  degAnnotInfoASG[i,6:9]<-c(degAnnotTotalOrderedASG[i,8],
degAnnotTotalOrderedASG[i,10], degAnnotTotalOrderedASG[i,16],
degAnnotTotalOrderedASG[i,14])
  degAnnotInfoASG[i,10]<-sapply(degAnnotTotalOrderedASG[i,20], function(x)
unlist(strsplit(x,"SUBCELLULAR LOCATION: "))[2])
  degAnnotInfoASG[i,11:13]<-c(degAnnotTotalOrderedASG[i,26],
degAnnotTotalOrderedASG[i,27], degAnnotTotalOrderedASG[i,28])}
nrow(degAnnotInfoASG)
write.table(degAnnotInfoASG, "OctopusASGdegAnnotInfo.csv", sep=";", col.names=NA)
```

```
####For PSG####
#Top11 PSG Annoted UniProtR
write.table(degAnnotInfoPSG[1:11,],    "OctopusPSGdegAnnotInfoTop11.csv",    sep=";",
col.names=NA)


####For ASG####
#Top11 ASG Annoted UniProtR
write.table(degAnnotInfoASG[1:11,],    "OctopusASGdegAnnotInfoTop11.csv",    sep=";",
col.names=NA)


#ORFs characterisation::::  Read peptide sequence
library(seqinr)
TransDecoderfile<-
c("C:\\Users\\Cátia\\Documents\\PhD_Bioinformatics\\Trinity.fasta.transdecoder.pep"
)
Fa<-read.fasta(TransDecoderfile, seqtype = "AA", as.string = TRUE)
ID<-getName(Fa)
pepSEQ<-unlist(getSequence(Fa, as.string=TRUE))
pep<-as.data.frame(cbind(ID, pepSEQ))
save(pep,                                    file                                    =
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\ORFsCharacterisationFiles\\pep.
RData")  ### Este passo serve para quando abrir o R outra vez, correr apenas o obj
ficheiro pep.RData e todos os obj associados são abertos (neste caso o pep)


##Check for full coding sequences
pep$length<-nchar(pep$pepSEQ)
pep$M<-ifelse(substring(pep$pepSEQ, 1,1) == "M", "M", "")
pep$Stop<-ifelse(grepl("*", pep$pepSEQ, fixed = TRUE), "STOP", "")
pep$length<-ifelse(pep$Stop == "STOP", pep$length - 1, pep$length)
CompletePep<-pep[which(pep$M == "M" & pep$Stop == "STOP"),]
PercentageCompletePep<-nrow(CompletePep)/nrow(pep)*100
save(pep, CompletePep, PercentageCompletePep, file = "Completepep.RData")

#############################################################################
############################  VolcanoPlot  ##################################
#############################################################################

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary")

###Read file octopDEG:
file<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\octopDEG.csv
"
octopDEG<-as.data.frame(read.table(file, sep=";",header=T))
log2FC <- expression(paste("log"[2], "FC"))
log10p <- expression(paste("-log"[10], "(", italic("p"), "-value)"))
octopDEG$signif<- -log10(octopDEG$FDRp.PSG)

#Order "signif". Subsets PSG=1 e ASG=1
OrderSignif<-octopDEG[order(-octopDEG$signif, octopDEG[,12] == 1, octopDEG[,11] ==
1),]
subset1_PSG<-subset(OrderSignif, OrderSignif$PSG ==1)
subset1_ASG<-subset(OrderSignif, OrderSignif$PSG ==-1)

###Load "ORFunico"
blastp_swissP                                                                    <-
read.table("C:\\Users\\Cátia\\Documents\\PhD_Bioinformatics\\Octopus_ORFs_blastp_un
iprot_sprot.outfmt6", sep="\t", header=F)
blastpID  <-  unlist(sapply(blastp_swissP[,1],  function(x)  unlist(strsplit(x,".",
fixed=TRUE))[1]))
blastp_swissP <- cbind(blastpID, blastp_swissP)
rownames(blastp_swissP) <- NULL
```

```
#Clean entries from Swissprot blast (30818)
A<-blastp_swissP[order(blastp_swissP$V11,-blastp_swissP$V3),]
ORFunico<-A[match(unique(A$V1), A$V1),]
ORFunico<-ORFunico[,c(1:4,12)]
colnames(ORFunico)[2:5]<-c("TrinityID", "Accession", "Blastp_%ID", "Blastp_Evalue")

###Retrive gene
#subset1_PSG e subset1_ASG
GenePSG<-merge(subset1_PSG, ORFunico, by.x="ID", by.y="blastpID", all=TRUE)
GeneASG<-merge(subset1_ASG, ORFunico, by.x="ID", by.y="blastpID", all=TRUE)

#Top10 genes
GenePSGTop10<-GenePSG[order(-GenePSG$signif),][1:10,]
GeneASGTop10<-GeneASG[order(-GeneASG$signif),][1:10,]

##############################
      #  VolcanoPlot  #
##############################

windows()
signif<- -log10(octopDEG$FDRp.PSG)
plot(octopDEG$logFC.PSG,signif,pch=".", xlab= log2FC, ylab= log10p)
points(octopDEG[which(octopDEG[,12]                ==                1),"logFC.PSG"],-
log10(octopDEG[which(octopDEG[,12] == 1),"FDRp.PSG"]),pch=20,cex=0.8, col="#fc8b38")
points(octopDEG[which(octopDEG[,12]                ==                -1),"logFC.PSG"],-
log10(octopDEG[which(octopDEG[,12]        ==        -1),"FDRp.PSG"]),pch=20,cex=0.8,
col="#579e68")

dev.print(tiff,"VolcanoPlot.tiff", height = 15, width = 15, unit="cm", res=600)


#############################################################################
##############################  Pie Chart  ##################################
#############################################################################

#Swissprot annotation: %DEGs PSG annotated, %DEGs ASG annotated, %DEGs PSG not
annotated, %DEGs ASG not annotated

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses")
library(ggplot2)

###Read file
blastp_swissP                                                                  <-
read.table("C:\\Users\\Cátia\\Documents\\PhD_Bioinformatics\\Octopus_ORFs_blastp_un
iprot_sprot.outfmt6", sep="\t", header=F)
blastpID  <-  unlist(sapply(blastp_swissP[,1],  function(x)  unlist(strsplit(x,".",
fixed=TRUE))[1]))
blastp_swissP <- cbind(blastpID, blastp_swissP)
rownames(blastp_swissP) <- NULL

#Clean entries of swissprot blast(30818)
A<-blastp_swissP[order(blastp_swissP$V11,-blastp_swissP$V3),]
ORFunico<-A[match(unique(A$V1), A$V1),]
ORFunico<-ORFunico[,c(1:4,12)] #30818
colnames(ORFunico)[2:5]<-c("TrinityID", "Accession", "Blastp_%ID", "Blastp_Evalue")

###Read transdecoder (ORFs predicted, 51090)
library(seqinr)
TransDecoderfile<-
c("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Trinity.fasta.transdecoder.pe
p")
Fa<-read.fasta(TransDecoderfile, seqtype = "AA", as.string = TRUE)
ID<-getName(Fa)
pepSEQ<-unlist(getSequence(Fa, as.string=TRUE))
pep<-as.data.frame(cbind(ID, pepSEQ))
```

```
###Merge transdecoder (51090) with ORFunico (30818 ORFs)
ORFAnnot<-merge(pep, ORFunico, by.x="ID", by.y="TrinityID", all = TRUE)

ORFAnnot$blastpID <- unlist(sapply(ORFAnnot[,1], function(x) unlist(strsplit(x,".",
fixed=TRUE))[1]))


###Read data file (DEGs):
degTable<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\deg
Table.csv", sep=";", header=TRUE, row.names=1)
degTable<-degTable[,c("ID", "ASG", "PSG")]

###Merge degTable with ORFAnnot
degTableAnnotated<-merge(degTable, ORFAnnot, by.x="ID", by.y="blastpID")
#write.table(degTableAnnotated, "degTableAnnotated.csv", sep=";", col.names=NA)

#Conditions (subsets) and calculate %:
##  annotated ORFs PSG (PSG=1 & Accession different de "NA")
##  annotated ORFs ASG (ASG=1 & Accession different de "NA")
##  non annotated ORFs PSG (PSG=1 & Accession = "NA")
##  non annotated ORFs ASG (ASG=1 & Accession = "NA")


#Subset de degTableAccession: ASG=1 e PSG=1
PSG_ORF_Annot<-subset(degTableAnnotated, as.numeric(degTableAnnotated[,3]) == 1 &
is.na(degTableAnnotated[,6]) == FALSE) #1351
ASG_ORF_Annot<-subset(degTableAnnotated, as.numeric(degTableAnnotated[,2]) == 1 &
is.na(degTableAnnotated[,6]) == FALSE) #803
PSG_ORF_NA<-subset(degTableAnnotated, as.numeric(degTableAnnotated[,3]) == 1 &
is.na(degTableAnnotated[,6])) #917
ASG_ORF_NA<-subset(degTableAnnotated, as.numeric(degTableAnnotated[,2]) == 1 &
is.na(degTableAnnotated[,6])) #510
nrow(PSG_ORF_Annot) #1351
nrow(ASG_ORF_Annot) #803
nrow(PSG_ORF_NA) #917
nrow(ASG_ORF_NA) #510

#Calculate %:
condition<-c("PSG_ORF_Annot","ASG_ORF_Annot","PSG_ORF_NA","ASG_ORF_NA")
number<-c(nrow(PSG_ORF_Annot), nrow(ASG_ORF_Annot),
nrow(PSG_ORF_NA),nrow(ASG_ORF_NA))
ORFTotal<-data.frame(cbind(condition, number))
ORFTotal$Percentage<-
as.numeric(ORFTotal$number)*100/sum(as.numeric(ORFTotal$number))
ORFTotal$condition<-as.factor(ORFTotal$condition)
ORFTotal<-ORFTotal[c(1,3,2,4),]


windows()
ggplot(ORFTotal, aes(x = "", y = Percentage, fill = condition))+
  geom_col(show.legend = TRUE, size = 8) +
  geom_text(aes(label = paste("PSG Annotated:\n", round(Percentage[1], digits=2),
"%")), x = -0.45, y = 0, size = 4) +
  geom_text(aes(label = paste("PSG Not Annotated:\n", round(Percentage[2], digits=2),
"%")), x = 1.7, y = 12, size = 4) +
  geom_text(aes(label = paste("ASG Annotated:\n", round(Percentage[3], digits=2),
"%")), x = 1.7, y = -12, size = 4) +
  geom_text(aes(label = paste("ASG Not Annotated:\n", round(Percentage[4], digits=2),
"%")), x = 0.7, y = -10, size = 4) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("#00765c", "#7dbe66", "#c63a69", "#de7388"),
       breaks = c("PSG_ORF_Annot","PSG_ORF_NA", "ASG_ORF_Annot", "ASG_ORF_NA"),
       labels = c("PSG Annotated", "PSG Not Annotated", "ASG Annotated", "ASG Not
Annotated"),
```

```
       name = "")+
  theme_void()+
  theme(legend.text = element_text(size = 11))
dev.print(tiff, "OctopusPieChartDEGsAnnot&notAnnot.tif", height = 20, width = 20,
units = 'cm', res=600)
ggsave("OctopusPieChartDEGsAnnot&notAnnot.svg", height = 15, width = 15,  units =
'cm', dpi=600)


################################################################################
################################  GOTerms  #####################################
################################################################################


#Load octopDEG (302 475 transcritos)
octopDEG<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\oct
opDEG.csv", sep=";", header=TRUE, row.names=1)

#Remove isoforms
TrinityGenes<-unlist(sapply(octopDEG$ID, function(x) unlist(strsplit(x, "i", fixed =
TRUE))[1]))
TrinityGenes<-data.frame(octopDEG, TrinityGenes)
TrinityGenes<-TrinityGenes[,c(1,12,2:11)]  #302475
length(unique(TrinityGenes$TrinityGenes)) #207109 genes

#Order TrinityGenes
TrinityGenesOrd<-TrinityGenes[order(-abs(TrinityGenes$logFC.PSG)),] #302475
length(unique(TrinityGenesOrd$TrinityGenes)) #207109 genes

#Match best gene
GeneUnico<-TrinityGenesOrd[match(unique(TrinityGenesOrd$TrinityGenes),
TrinityGenesOrd$TrinityGenes),] #207109
length(unique(GeneUnico$TrinityGenes)) #207109 genes
write.table(GeneUnico, "GeneUnico.csv", sep=";", col.names=NA)


########## SWISSPROT ANALYSES ##########

##Merge data with Swissprot info
#Load swissprot file and remove .p's
#Read Blastp for swissprot results (207 mil)
blastp_swissP                                                                  <-
read.table("C:\\Users\\Cátia\\Documents\\PhD_Bioinformatics\\Octopus_ORFs_blastp_un
iprot_sprot.outfmt6", sep="\t", header=F)
blastpID  <-  unlist(sapply(blastp_swissP[,1],  function(x)  unlist(strsplit(x,".",
fixed=TRUE))[1]))
blastp_swissP <- cbind(blastpID, blastp_swissP)
rownames(blastp_swissP) <- NULL

#Clean entries of swissprot blast(30818)
A<-blastp_swissP[order(blastp_swissP$V11,-blastp_swissP$V3),]
ORFunico<-A[match(unique(A$V1), A$V1),]
ORFunico<-ORFunico[,c(1:4,12)]
colnames(ORFunico)[2:5]<-c("TrinityID", "Accession", "Blastp_%ID", "Blastp_Evalue")

#Merge 207 mil genes with 30818 ORFs from blast
GeneAnnot<-merge(GeneUnico, ORFunico, by.x="ID", by.y="blastpID")
write.table(GeneAnnot, "GeneAnnot.csv", sep=";", col.names=NA)

#More than 1 ORFs per gene! Remove by e-value and %ID #at the end: 10441 annotated
genes
B<-GeneAnnot[order(GeneAnnot$Blastp_Evalue,-GeneAnnot$`Blastp_%ID`),]
GeneAnnotUnico<-B[match(unique(B$TrinityGenes), B$TrinityGenes),]
write.table(GeneAnnotUnico, "GeneAnnotUnico.csv", sep=";", col.names=NA)
```

```
#Merge GOTerms with GeneAnnotUnico
AccessionSimples<-sapply(GeneAnnotUnico[,14],   function(x)   unlist(strsplit(x,".",
fixed=TRUE))[1])
GeneAnnotUnico<-cbind(GeneAnnotUnico, AccessionSimples)
AccessionUnico<-unique(GeneAnnotUnico[,17])
length(AccessionUnico) #8177
NameTaxa<-GetNamesTaxa(AccessionUnico)
SubcellLocation<-GetSubcellular_location(AccessionUnico)
GOTerms<-GetProteinGOInfo(AccessionUnico)
GeneAnnotUnicoInfoTerms<-cbind(NameTaxa, SubcellLocation, GOTerms)
nrow(GeneAnnotUnicoInfoTerms)
rownames(GeneAnnotUnicoInfoTerms)<- NULL
colnames(GeneAnnotUnicoInfoTerms)[1]<-c("AccessionSimples")
write.table(GeneAnnotUnicoInfoTerms,     "GeneAnnotUnicoInfoTerms.csv",     sep=";",
col.names=NA)

DEGInfoTerms<-merge(GeneAnnotUnico, GeneAnnotUnicoInfoTerms, by="AccessionSimples")


#Get DEGs from 'GeneAnnotUnicoInfoTerms' (PSG e ASG)
PSG_DEGInfoTerms<-subset(DEGInfoTerms, as.numeric(DEGInfoTerms[,13]) > 0)
ASG_DEGInfoTerms<-subset(DEGInfoTerms, as.numeric(DEGInfoTerms[,13]) < 0)
nrow(PSG_DEGInfoTerms)
nrow(ASG_DEGInfoTerms)
write.table(PSG_DEGInfoTerms,  "PSG_DEGInfoTerms.csv",  sep=";",  col.names=NA)  #773
entries
write.table(ASG_DEGInfoTerms,  "ASG_DEGInfoTerms.csv",  sep=";",  col.names=NA)  #462
entries


setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses\\GOTermsO
utubro2024")

for(l in 1:2){
  if(l == 1)
    degAnnotTissue<-PSG_DEGInfoTerms
  else
    degAnnotTissue<-ASG_DEGInfoTerms
  for(j in 36:38){
    GOTerms<-matrix()
    NACount<-0
    for(i in 1:nrow(degAnnotTissue)){
      GOAux<-sapply(degAnnotTissue[i,j],    function(x)    unlist(strsplit(x,";    ",
fixed=TRUE)))
      if(!is.na(GOAux[1]))
        GOTerms<-rbind(GOTerms, GOAux)
      else
        NACount<-NACount+1    }
    colnames(GOTerms)<-NULL
    GOTerms<-as.data.frame(GOTerms[-1,,drop=FALSE])
    GOUnique<-unique(GOTerms)
    GO<-sapply(GOUnique[,1],   function(x)   length(GOTerms[which(GOTerms[,1]   %in%
x),]))
    if(j == 36 && l == 1){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOBiologicalProcessPSG.RData")
      write.table(GO, "OctopusGOBiologicalProcessPSG.csv", sep=";", col.names=NA)}
    else if(j == 36 && l == 2){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOBiologicalProcessASG.RData")
      write.table(GO, "OctopusGOBiologicalProcessASG.csv", sep=";", col.names=NA)}
    else if(j == 37 && l == 1){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOMolecularFunctionPSG.RData")
      write.table(GO, "OctopusGOMolecularFunctionPSG.csv", sep=";", col.names=NA)}
```

```
    else if(j == 37 && l == 2){
      save(GOTerms,                GOUnique,                NACount,                GO,
file="OctopusGOMolecularFunctionASG.RData")
      write.table(GO, "OctopusGOMolecularFunctionASG.csv", sep=";", col.names=NA)}
    else if(j ==38 && l == 1){
      save(GOTerms,                GOUnique,                NACount,                GO,
file="OctopusGOCellularComponentPSG.RData")
      write.table(GO, "OctopusGOCellularComponentPSG.csv", sep=";", col.names=NA)}
    else{
      save(GOTerms,                GOUnique,                NACount,                GO,
file="OctopusGOCellularComponentASG.RData")
      write.table(GO, "OctopusGOCellularComponentASG.csv", sep=";", col.names=NA)}}}


###TFisher Contingency table:
##Enrichment analysis
##Fisher analysis for the GO Terms
##The data in analysis is the "FisherOrganOfInterest" and "FisherReferenceOrgan"
##The p-value was adjusted by FDR method

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses\\GOTermsO
utubro2024\\TesteFisher")

Fisher<-function(OrganOfInterest,        ReferenceOrgan,        totalOrganOfInterest,
totalReferenceOrgan, FisherOrganOfInterest, FisherReferenceOrgan, Category){
  FisherAux<-merge(FisherOrganOfInterest, FisherReferenceOrgan, by = "GOTerms", all
= TRUE)
  rownames(FisherAux)<-FisherAux$GOTerms
  FisherAux<-FisherAux[,-1,drop=FALSE]
  colnames(FisherAux)<-c(OrganOfInterest, ReferenceOrgan)
  FisherAux[which(is.na(FisherAux[,OrganOfInterest])),OrganOfInterest]<-0
  FisherAux[which(is.na(FisherAux[,ReferenceOrgan])),ReferenceOrgan]<-0
  print(paste(sep = "", "FisherAux", Category, ".RData"))
  save(FisherAux, file = paste(sep = "", "FisherAux", Category, ".RData"))

  rownames(FisherAux)<-sapply(rownames(FisherAux),  function(x)  unlist(strsplit(x,"
[", fixed = TRUE))[1])

  print(nrow(FisherAux))
  FisherAux<-subset(FisherAux,           FisherAux[,OrganOfInterest]>=10           |
FisherAux[,ReferenceOrgan]>=10)
  print(nrow(FisherAux))

  Fisher<-data.frame(matrix(data = NA, nrow = nrow(FisherAux), ncol = 5, dimnames =
list(NULL,   c("GOTerms",   paste("Upregulated",   OrganOfInterest,   sep   =   ""),
paste("Upregulated", ReferenceOrgan, sep = ""), "odds_ratio", "pvalue"))))
  for(i in 1:nrow(Fisher)){
    contTable<-rbind(FisherAux[i,],             c(totalOrganOfInterest             -
FisherAux[i,OrganOfInterest], totalReferenceOrgan - FisherAux[i,ReferenceOrgan]))
    fisherTest<-fisher.test(contTable)
    Fisher[i,]<-data.frame(rownames(FisherAux)[i],     FisherAux[i,OrganOfInterest],
FisherAux[i,ReferenceOrgan], fisherTest$estimate, fisherTest$p.value)}
  Fisher$FDRp<-p.adjust(Fisher$pvalue, method = "fdr")
  Fisher<-Fisher[order(Fisher$FDRp, -Fisher$odds_ratio),]
  save(Fisher, file=paste(sep = "", "Octopus", Category, "Fisher.RData"))
  FisherFormatted(Category, Fisher)}

##Formatting the output table from the fisher analysis
FisherFormatted<-function(Category, FisherFormatted){
  FisherFormatted$odds_ratio<-sprintf("%.3f",
as.numeric(format(FisherFormatted$odds_ratio, digits = 3)))
  FisherFormatted$pvalue<-format(FisherFormatted$pvalue, digits = 3)
  FisherFormatted$FDRp<-format(FisherFormatted$FDRp, digits = 3, scientific = TRUE)
  save(FisherFormatted,    file    =    paste(sep    =    "",    "Octopus",    Category,
"FisherFormatted.RData"))
```

```
  write.table(FisherFormatted,    paste(sep    =        "",    "Octopus",    Category,
"FisherFormatted.csv"), sep=";", col.names=NA)}


################################
## Fisher Analysis: GO Terms ##
#           Objects            #
################################

pathGO<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses\\GOTermsOutubro
2024\\"
GOCategory<-c("GOBiologicalProcess","GOMolecularFunction", "GOCellularComponent")
organ<-c("PSG", "ASG")
PSG_DEGInfoTerms<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses\\PSG_D
EGInfoTerms.csv", sep=";", header=TRUE, row.names=1)
ASG_DEGInfoTerms<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses\\ASG_D
EGInfoTerms.csv", sep=";", header=TRUE, row.names=1)

for(k in 1:length(GOCategory)){
  for(j in 1:length(organ)){
    load(paste(pathGO, "Octopus", GOCategory[k], organ[j], ".RData", sep = ""))
    assign(paste("Fisher",  organ[j],  sep  =""),  data.frame(GOTerms  =  names(GO),
NumberORFs = GO))}
  Fisher(organ[1],    organ[2],    nrow(PSG_DEGInfoTerms),    nrow(ASG_DEGInfoTerms),
FisherPSG, FisherASG, GOCategory[k])}

##############################################################################
######################        Annotation SeaTox DB        ####################
##############################################################################

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast")

Annotate<-function(fileBlastp, fileAnnot){
  Annot<-read.csv(fileBlastp, sep=";", header = FALSE)
  print(fileBlastp)
  print(nrow(Annot))
  Annot[,(ncol(Annot)+1)]<-unlist(sapply(Annot[,2],  function(x)  unlist(strsplit(x,
"|", fixed = TRUE))[2]))
  Annot<-Annot[,c(1,16,3,11,15)]
  colnames(Annot)<-c("TrinityID", "Accession", "Blastp_%ID", "Blastp_Evalue")
  Annot<-Annot[which(grepl("TRINITY", Annot$TrinityID)),]
  print(nrow(Annot))

  ##Sorting by e-value and %ID
  Annot<-Annot[order(Annot[, "Blastp_Evalue"], -Annot[, "Blastp_%ID"]),]
  rownames(Annot)<-NULL
  write.table(Annot, fileAnnot, sep=";", col.names=NA)
  return(Annot)}

BlastOctopSeaToxDB<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\BlastOctopSeaTo
xDB.csv"
Output_Allgene<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\BlastOctopSeaTo
xDB_Annot.csv"

BlastOctopDEGsSeaToxDB<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\BlastOctopDEGsS
eaToxDB.csv"
Output_DEGs<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\BlastOctopDEGsS
eaToxDB_Annot.csv"
```

```
Allgene_Annot<-Annotate(BlastOctopSeaToxDB, Output_Allgene)
DEGs_Annot<-Annotate(BlastOctopDEGsSeaToxDB, Output_DEGs)

#########################################################
###                      UniProt                      ###
###        Uniprot version: Release 2024_05           ###
#########################################################

Accessions<-unique(c(Allgene_Annot[,2], DEGs_Annot[,2]))
AccessionInfo<-GetProteinAnnontate(Accessions,    c("keyword",    "protein_families",
"cc_biotechnology"))
AccessionInfo$Accession <- rownames(AccessionInfo)

##Merge AccessionInfo with Allgene_Annot
Allgene_AnnotMergeInfo<-merge(Allgene_Annot, AccessionInfo, by="Accession")
write.table(Allgene_AnnotMergeInfo, "AllgeneInfo.csv", sep=";", col.names=NA)

##Remove p's from'trinity's ID'
Allgene_AnnotMergeInfo<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\Allgen
eInfo.csv", sep=";", header=TRUE, row.names=1)
ID<-unlist(sapply(Allgene_AnnotMergeInfo[,2],   function(x)   unlist(strsplit(x,".",
fixed=TRUE))[1]))
Allgene_AnnotMergeInfo<-cbind(ID, Allgene_AnnotMergeInfo)
rownames(Allgene_AnnotMergeInfo) <- NULL #3004, all ID's

#Order ID's by por p-value and %BlastID
P<-Allgene_AnnotMergeInfo[order(Allgene_AnnotMergeInfo$Blastp_Evalue,-
Allgene_AnnotMergeInfo$Blastp_.ID),]
ORFunico<-P[match(unique(P$ID), P$ID),] #2940
length(unique(ORFunico$ID)) #2940

#Remove i's from 'ID'
GeneID<-unlist(sapply(ORFunico$ID,  function(x)  unlist(strsplit(x,  "i",  fixed  =
TRUE))[1]))
GeneID<-data.frame(ORFunico, GeneID) #2940
GeneID<-GeneID[,c(10,1:9)]
length(unique(GeneID$GeneID)) #1115

#Load octopDEG #302475
octopDEG<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\octopFilesJanuary\\oct
opDEG.csv", sep=";", header=TRUE, row.names=1)

#Merge 'GeneUnico' with 'GeneID'
Allgene_AnnotMergeInfo_GeneUnico<-merge(GeneID,  octopDEG,  by="ID")  #2940,  continuo
com genes repetidos

#Order 'GeneID' by log FC
Ord_logFC<-Allgene_AnnotMergeInfo_GeneUnico[order(-
abs(Allgene_AnnotMergeInfo_GeneUnico$logFC.PSG)),] #2940

#final table
FINAL_Allgene_AnnotMergeInfo_GeneUnico<-Ord_logFC[match(unique(Ord_logFC$GeneID),
Ord_logFC$GeneID),]
length(unique(FINAL_Allgene_AnnotMergeInfo_GeneUnico$GeneID)) #1115 genes
write.table(FINAL_Allgene_AnnotMergeInfo_GeneUnico,
"FINAL_Allgene_AnnotMergeInfo_GeneUnico.csv", sep=";", col.names=NA)

#########################################
#######         Circles Plot         #######
#######  (All transcripts--> 1 ORF)   #######
#######      (selected categories)    #######
#######    Two organs: ASG and PSG    #######
#########################################
```

```
FINAL_Allgene_AnnotMergeInfo_GeneUnico<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\FINAL_
Allgene_AnnotMergeInfo_GeneUnico.csv", sep=";", header=TRUE, row.names=1)
Data_subsetPSG<-
subset(FINAL_Allgene_AnnotMergeInfo_GeneUnico,FINAL_Allgene_AnnotMergeInfo_GeneUnic
o$PSG ==1)
Data_subsetASG<-
subset(FINAL_Allgene_AnnotMergeInfo_GeneUnico,FINAL_Allgene_AnnotMergeInfo_GeneUnic
o$ASG ==1)


##############################
####   Circles Plot PSG   ####
##############################

NumberOfCircles<-function(Term, Column, Data_subsetPSG){
  aux<-grep(Term, Data_subsetPSG[,Column])
  if(Term == "Serine protease"){
    aux1<-grep("Serine protease inhibitor", Data_subsetPSG[,Column])
    aux<-setdiff(aux, aux1)  }
  AnnotTerm<-Data_subsetPSG[aux,]
  Circle<-length(aux)}

Circles<-data.frame(Term    =    c("Toxin",    "Hormone",    "Serine    protease",
"Metalloprotease", "Glycosyl hydrolase 18 family", "Glycosyl hydrolase 56 family",
"Peptidase", "Phospholipase"),
                    Column  =  c("Keywords",  "Keywords",  "Keywords",  "Keywords",
"Protein.families", "Protein.families", "Protein.families", "Protein.families"))


for(i in 1:nrow(Circles)){
  assign(Circles[i,1], NumberOfCircles(Circles[i,1], Circles[i,2], Data_subsetPSG))}

data<-data.frame(rbind(Toxin, Hormone, `Serine protease`, Metalloprotease, `Glycosyl
hydrolase 18 family`, `Glycosyl hydrolase 56 family`, Peptidase, Phospholipase))
rownames(data)[5:6]<-c("Chitinase", "Hyaluronidase")
colnames(data)<-c("Circles")

##Circles Plot
#Function that return a dataframe. 1 line per bubble with the center, radius
(proportional of the number of occurrences. Area that should be proportional to the
occurrences number
packing <- circleProgressiveLayout(data$Circles, sizetype= 'area')
# Add packing data to the object "data"
data <- cbind(data, packing)
#Coordinates of a circle that is drawn by a "npoints" of straight lines (10000)
dat.gg <- circleLayoutVertices(packing, npoints=10000)
# Create plot
windows()
ggplot() +
  # Create the bubbles
  geom_polygon(data = dat.gg, aes(x, y, group = id, fill=as.factor(id)), alpha = 0.7)
+
  # Add text in each bubble and control the size of the text
  #fontface=2 put the letters in bold type
  geom_text(data = data, aes(x, y, size=Circles, label = rownames(data), fontface=2))
+
  geom_text(data = data, aes(x, y-0.95, size=80, label = Circles, fontface=2)) +
  scale_size_continuous(range = c(1,4)) +
  # General theme:
  theme_void() +
  scale_fill_manual(values = c("#70AD47", "#FFD700", "#CC6677", "#648FFF", "#FFA500",
"#7D26CD", "#E87F36", "#BEBEBE"))+
  theme(legend.position= "none") +
  coord_equal()
```

```
dev.print(tiff, "Circles/CirclesALL_PSG.tif", height = 15, width = 15,  units = 'cm',
res=600)
ggsave("Circles/CirclesALL_PSG.svg", height = 15, width = 15,  units = 'cm', dpi=600)


##############################
####   Circles Plot ASG   ####
##############################

NumberOfCircles<-function(Term, Column, Data_subsetASG){
  aux<-grep(Term, Data_subsetASG[,Column])
  if(Term == "Serine protease"){
    aux1<-grep("Serine protease inhibitor", Data_subsetASG[,Column])
    aux<-setdiff(aux, aux1)  }
  AnnotTerm<-Data_subsetASG[aux,]  Circle<-length(aux)}

Circles<-data.frame(Term    =    c("Toxin",    "Hormone",    "Serine    protease",
"Metalloprotease", "Glycosyl hydrolase 18 family", "Glycosyl hydrolase 56 family",
"Peptidase", "Phospholipase"),
                    Column  =  c("Keywords",  "Keywords",  "Keywords",  "Keywords",
"Protein.families", "Protein.families", "Protein.families", "Protein.families"))


for(i in 1:nrow(Circles)){
  assign(Circles[i,1], NumberOfCircles(Circles[i,1], Circles[i,2], Data_subsetASG))}

data<-data.frame(rbind(Toxin, Hormone, `Serine protease`, Metalloprotease, `Glycosyl
hydrolase 18 family`, `Glycosyl hydrolase 56 family`, Peptidase, Phospholipase))
rownames(data)[5:6]<-c("Chitinase", "Hyaluronidase")
colnames(data)<-c("Circles")


##Circles Plot
#Function that return a dataframe. 1 line per bubble with the center, radius
(proportional of the number of occurrences. Area that should be proportional to the
occurrences number
packing <- circleProgressiveLayout(data$Circles, sizetype= 'area')
# Add packing data to the object "data"
data <- cbind(data, packing)
#Coordinates of a circle that is drawn by a "npoints" of straight lines (10000)
dat.gg <- circleLayoutVertices(packing, npoints=10000)
# Create plot
windows()
ggplot() +
  # Create the bubbles
  geom_polygon(data = dat.gg, aes(x, y, group = id, fill=as.factor(id)), alpha = 0.7)
+
  # Add text in each bubble and control the size of the text
  #fontface=2 put the letters in bold type
  geom_text(data = data, aes(x, y, size=Circles, label = rownames(data), fontface=2))
+
  geom_text(data = data, aes(x, y-0.95, size=80, label = Circles, fontface=2)) +
  scale_size_continuous(range = c(1,4)) +
  # General theme:
  theme_void() +
  scale_fill_manual(values = c("#70AD47", "#FFD700", "#CC6677", "#648FFF", "#FFA500",
"#7D26CD", "#E87F36", "#BEBEBE"))+
  theme(legend.position= "none") +
  coord_equal()
dev.print(tiff, "Circles/CirclesALL_ASG.tif", height = 15, width = 15,  units = 'cm',
res=600)
ggsave("Circles/CirclesALL_ASG.svg", height = 15, width = 15,  units = 'cm', dpi=600)
```

```
##############################################################################
############################  Top 10 PSG  ####################################
##############################################################################
## Order: e-value then %ID

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast")
FINAL_Allgene_AnnotMergeInfo_GeneUnico<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\FINAL_
Allgene_AnnotMergeInfo_GeneUnico.csv", sep=";", header=TRUE, row.names=1)
Subset_PSG<-
subset(FINAL_Allgene_AnnotMergeInfo_GeneUnico,FINAL_Allgene_AnnotMergeInfo_GeneUnic
o$PSG ==1)
Top10_OV<-Subset_PSG[order(Subset_PSG$Blastp_Evalue,-Subset_PSG$Blastp_.ID),]
FINAL_Top10_OV<-Top10_OV[1:10,]
write.table(FINAL_Top10_OV, "FINAL_Top10_OV.csv", sep=";", col.names=NA)


##############################################################################
########################  GOTerms SeaTox DB  #################################
##############################################################################

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTermsOu
t2024_SeatoxDB")

# 'Allgene_AnnotMergeInfo'
#Merge 207 mil genes with 3004 ORFs from blast SeaTox
BlastOctopSeaToxDB_Annot<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\BlastO
ctopSeaToxDB_Annot.csv", sep=";", header=TRUE, row.names=1)
blastpID       <-       unlist(sapply(BlastOctopSeaToxDB_Annot[,1],       function(x)
unlist(strsplit(x,".", fixed=TRUE))[1]))
BlastOctopSeaToxDB_Annot <- cbind(blastpID, BlastOctopSeaToxDB_Annot)
rownames(BlastOctopSeaToxDB_Annot) <- NULL

GeneUnico<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\Out2024Analyses\\GeneU
nico.csv", sep=";", header=TRUE, row.names=1)
GeneAnnot<-merge(GeneUnico, BlastOctopSeaToxDB_Annot, by.x="ID", by.y="blastpID")
write.table(GeneAnnot, "GeneAnnot.csv", sep=";", col.names=NA)

#Remove ORFs by e-value and %ID #860 annotated genes
B<-GeneAnnot[order(GeneAnnot$Blastp_Evalue,-GeneAnnot$`Blastp_.ID`),]
GeneAnnotUnico<-B[match(unique(B$TrinityGenes), B$TrinityGenes),]
write.table(GeneAnnotUnico, "GeneAnnotUnico.csv", sep=";", col.names=NA)

AccessionUnico<-unique(GeneAnnotUnico[,14])
length(AccessionUnico) #302
NameTaxa<-GetNamesTaxa(AccessionUnico)
SubcellLocation<-GetSubcellular_location(AccessionUnico)
GOTerms<-GetProteinGOInfo(AccessionUnico)
GeneAnnotUnicoInfoTerms<-cbind(NameTaxa, SubcellLocation, GOTerms)
nrow(GeneAnnotUnicoInfoTerms)
rownames(GeneAnnotUnicoInfoTerms)<- NULL
colnames(GeneAnnotUnicoInfoTerms)[1]<-c("AccessionSimples")
write.table(GeneAnnotUnicoInfoTerms,     "GeneAnnotUnicoInfoTerms.csv",     sep=";",
col.names=NA)

DEGInfoTerms<-merge(GeneAnnotUnico,    GeneAnnotUnicoInfoTerms,    by.x="Accession",
by.y="AccessionSimples")


#Subset 'GeneAnnotUnicoInfoTerms'
PSG_DEGInfoTerms<-subset(DEGInfoTerms, as.numeric(DEGInfoTerms[,13]) > 0)
ASG_DEGInfoTerms<-subset(DEGInfoTerms, as.numeric(DEGInfoTerms[,13]) < 0)
nrow(PSG_DEGInfoTerms)
```

```
nrow(ASG_DEGInfoTerms)
write.table(PSG_DEGInfoTerms, "PSG_DEGInfoTerms.csv", sep=";", col.names=NA)
write.table(ASG_DEGInfoTerms, "ASG_DEGInfoTerms.csv", sep=";", col.names=NA)

##GOTerms - Calculate the percentage of genes related with a specific GO Term

for(l in 1:2){
  if(l == 1)
    degAnnotTissue<-PSG_DEGInfoTerms
  else
    degAnnotTissue<-ASG_DEGInfoTerms
  for(j in 36:38){
    GOTerms<-matrix()
    NACount<-0
    for(i in 1:nrow(degAnnotTissue)){
      GOAux<-sapply(degAnnotTissue[i,j],    function(x)    unlist(strsplit(x,";    ",
fixed=TRUE)))
      if(!is.na(GOAux[1]))
        GOTerms<-rbind(GOTerms, GOAux)
      else
        NACount<-NACount+1    }
    colnames(GOTerms)<-NULL
    GOTerms<-as.data.frame(GOTerms[-1,,drop=FALSE])
    GOUnique<-unique(GOTerms)
    GO<-sapply(GOUnique[,1],   function(x)   length(GOTerms[which(GOTerms[,1]   %in%
x),]))
    if(j == 36 && l == 1){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOBiologicalProcessPSG.RData")
      write.table(GO, "OctopusGOBiologicalProcessPSG.csv", sep=";", col.names=NA)}
    else if(j == 36 && l == 2){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOBiologicalProcessASG.RData")
      write.table(GO, "OctopusGOBiologicalProcessASG.csv", sep=";", col.names=NA)}
    else if(j == 37 && l == 1){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOMolecularFunctionPSG.RData")
      write.table(GO, "OctopusGOMolecularFunctionPSG.csv", sep=";", col.names=NA)}
    else if(j == 37 && l == 2){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOMolecularFunctionASG.RData")
      write.table(GO, "OctopusGOMolecularFunctionASG.csv", sep=";", col.names=NA)}
    else if(j ==38 && l == 1){
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOCellularComponentPSG.RData")
      write.table(GO, "OctopusGOCellularComponentPSG.csv", sep=";", col.names=NA)}
    else{
      save(GOTerms,               GOUnique,               NACount,               GO,
file="OctopusGOCellularComponentASG.RData")
      write.table(GO, "OctopusGOCellularComponentASG.csv", sep=";", col.names=NA)}}}

###TFisher Contingency table:
##Enrichment analysis
##Fisher analysis for the GO Terms
##The data in analysis is the "FisherOrganOfInterest" and "FisherReferenceOrgan"
##The p-value was adjusted by FDR method

setwd("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTermsOu
t2024_SeatoxDB\\TesteFisher")

Fisher<-function(OrganOfInterest,       ReferenceOrgan,       totalOrganOfInterest,
totalReferenceOrgan, FisherOrganOfInterest, FisherReferenceOrgan, Category){
  FisherAux<-merge(FisherOrganOfInterest, FisherReferenceOrgan, by = "GOTerms", all
= TRUE)
  rownames(FisherAux)<-FisherAux$GOTerms
```

```
  FisherAux<-FisherAux[,-1,drop=FALSE]
  colnames(FisherAux)<-c(OrganOfInterest, ReferenceOrgan)
  FisherAux[which(is.na(FisherAux[,OrganOfInterest])),OrganOfInterest]<-0
  FisherAux[which(is.na(FisherAux[,ReferenceOrgan])),ReferenceOrgan]<-0
  print(paste(sep = "", "FisherAux", Category, ".RData"))
  save(FisherAux, file = paste(sep = "", "FisherAux", Category, ".RData"))

  rownames(FisherAux)<-sapply(rownames(FisherAux),  function(x)  unlist(strsplit(x,"
[", fixed = TRUE))[1])


  FisherAux<-subset(FisherAux,           FisherAux[,OrganOfInterest]>=2          |
FisherAux[,ReferenceOrgan]>=2) #Condição definida para esta análise. Cada categoria
tem de ter pelo menos 3 genes, ou as duas tem que ter 3.
  # Contingency table (contTable)
  Fisher<-data.frame(matrix(data = NA, nrow = nrow(FisherAux), ncol = 5, dimnames =
list(NULL,   c("GOTerms",   paste("Upregulated",   OrganOfInterest,   sep   =   ""),
paste("Upregulated", ReferenceOrgan, sep = ""), "odds_ratio", "pvalue"))))
  for(i in 1:nrow(Fisher)){
    contTable<-rbind(FisherAux[i,],             c(totalOrganOfInterest            -
FisherAux[i,OrganOfInterest], totalReferenceOrgan - FisherAux[i,ReferenceOrgan]))
    fisherTest<-fisher.test(contTable)
    Fisher[i,]<-data.frame(rownames(FisherAux)[i],     FisherAux[i,OrganOfInterest],
FisherAux[i,ReferenceOrgan], fisherTest$estimate, fisherTest$p.value)  }
  Fisher$FDRp<-p.adjust(Fisher$pvalue, method = "fdr")
  Fisher<-Fisher[order(Fisher$FDRp, -Fisher$odds_ratio),]
  save(Fisher, file=paste(sep = "", "Octopus", Category, "Fisher.RData"))
  FisherFormatted(Category, Fisher)}

##Formatting the output table from the fisher analysis
FisherFormatted<-function(Category, FisherFormatted){
  FisherFormatted$odds_ratio<-sprintf("%.3f",
as.numeric(format(FisherFormatted$odds_ratio, digits = 3)))
  FisherFormatted$pvalue<-format(FisherFormatted$pvalue, digits = 3)
  FisherFormatted$FDRp<-format(FisherFormatted$FDRp, digits = 3, scientific = TRUE)
  save(FisherFormatted,    file    =    paste(sep    =    "",    "Octopus",    Category,
"FisherFormatted.RData"))
  write.table(FisherFormatted,    paste(sep    =       "",    "Octopus",    Category,
"FisherFormatted.csv"), sep=";", col.names=NA)}


################################
## Fisher Analysis: GO Terms ##
#          Objects             #
################################

pathGO<-
"P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTermsOut2024_
SeatoxDB\\"
GOCategory<-c("GOBiologicalProcess","GOMolecularFunction", "GOCellularComponent")
organ<-c("PSG", "ASG")
PSG_DEGInfoTerms<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTerm
sOut2024_SeatoxDB\\PSG_DEGInfoTerms.csv", sep=";", header=TRUE, row.names=1)
ASG_DEGInfoTerms<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTerm
sOut2024_SeatoxDB\\ASG_DEGInfoTerms.csv", sep=";", header=TRUE, row.names=1)

for(k in 1:length(GOCategory)){
  for(j in 1:length(organ)){
    load(paste(pathGO, "Octopus", GOCategory[k], organ[j], ".RData", sep = ""))
    assign(paste("Fisher",  organ[j],  sep  =""),  data.frame(GOTerms  =  names(GO),
NumberORFs = GO))  }
  Fisher(organ[1],    organ[2],    nrow(PSG_DEGInfoTerms),    nrow(ASG_DEGInfoTerms),
FisherPSG, FisherASG, GOCategory[k])}
```

```
##############################
######   Bubble Plot    ######
##############################

#Data
OctopusGOBiologicalProcessFisherFormatted<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTerm
sOut2024_SeatoxDB\\TesteFisher\\OctopusGOBiologicalProcessFisherFormatted.csv",
sep=";", header=TRUE, row.names=1)
OctopusGOCellularComponentFisherFormatted<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTerm
sOut2024_SeatoxDB\\TesteFisher\\OctopusGOCellularComponentFisherFormatted.csv",
sep=";", header=TRUE, row.names=1)
OctopusGOMolecularFunctionFisherFormatted<-
read.csv("P:\\Cátia\\Documents\\.PhD\\Bioinformática\\octop\\SeaToxDB_Blast\\GOTerm
sOut2024_SeatoxDB\\TesteFisher\\OctopusGOMolecularFunctionFisherFormatted.csv",
sep=";", header=TRUE, row.names=1)

#OddsRatio= Inf.
OctopusGOBiologicalProcessFisherFormatted[4:9,4]<-6.99
OctopusGOCellularComponentFisherFormatted[c(2,6,7,10),4]<-6.99

DataBio<-OctopusGOBiologicalProcessFisherFormatted
DataBio$log<--log10(DataBio$FDRp)
DataBio$NumberGenes<-DataBio$UpregulatedPSG + DataBio$UpregulatedASG

DataCel<-OctopusGOCellularComponentFisherFormatted
DataCel$log<--log10(DataCel$FDRp)
DataCel$NumberGenes<-DataCel$UpregulatedPSG + DataCel$UpregulatedASG

DataMol<-OctopusGOMolecularFunctionFisherFormatted
DataMol$log<--log10(DataMol$FDRp)
DataMol$NumberGenes<-DataMol$UpregulatedPSG + DataMol$UpregulatedASG

#Create columns BP, CC, MF
DataBio$Type<-c("BP")
DataCel$Type<-c("CC")
DataMol$Type<-c("MF")

#rbind
Data<-rbind(DataBio,DataCel,DataMol)

Data$TypeNumeric<-ifelse(Data$Type=="BP", 1,ifelse(Data$Type=="CC",2,3))

DataTop10<-
rbind(Data[which(Data$Type=="BP")[1:10],],Data[which(Data$Type=="CC")[1:10],],Data[
which(Data$Type=="MF")[1:10],])
log10FDR <- expression(paste("-log"[10], "(adj. ", italic("p"), "-value)"))


windows()
ggplot(DataTop10, aes(y = reorder(GOTerms, TypeNumeric), x = log, size = NumberGenes))
+
  geom_point(aes(color = odds_ratio), alpha = 1.0) +
    xlim(0,3.9) +
  theme_minimal() +
  scale_color_gradientn(colours=c("black",                          "white","blue"),
values=rescale(c(0,1,7)), limits=c(0,7))+
  labs(title= "Enriched pathways", y= "GO terms", x = log10FDR)
dev.print(tiff, "DataTop10SeaTox_Bio_Cel_Mol.tif", height = 20, width = 20,  units =
'cm', res=600)
ggsave("DataTop10SeaTox_Bio_Cel_Mol.svg", height = 20, width = 20,  units = 'cm',
dpi=600)
```

B.3 Melting curves

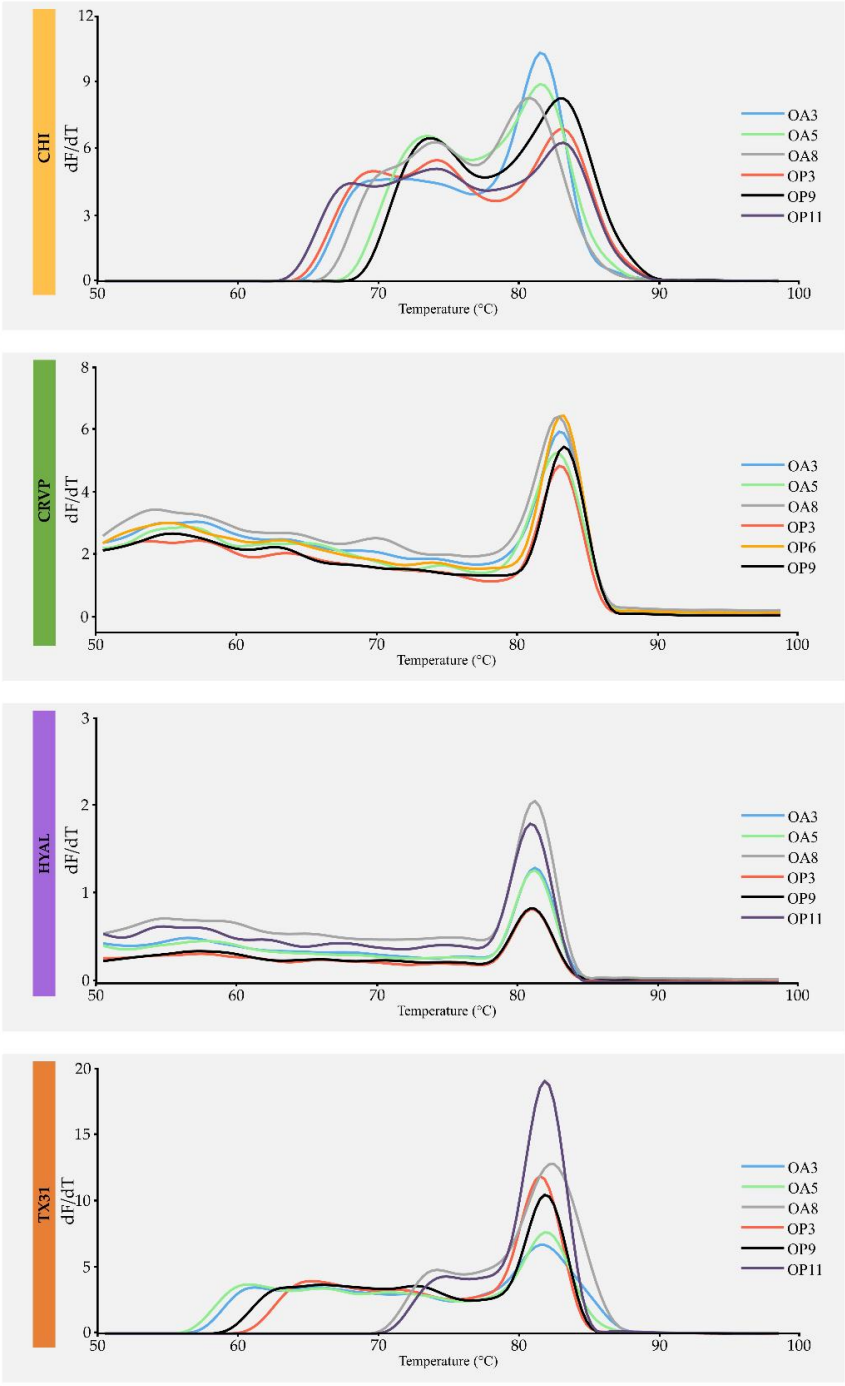


Figure B.3 - Melting curve analyses of the four primer pairs used in RT-qPCR for four shortlisted toxin full coding mRNAs isolated from the anterior and posterior salivary glands of *Octopus vulgaris*. CHI, chitinase; CRVP, cysteine-rich venom protein; HYAL, hyaluronidase ; TX31, cysteine-rich venom protein. Individual replicates of the anterior salivary gland are identified as OA3, OA5, OA8. Individual replicates of the posterior salivary gland are identified as OP3, OP6, OP9, OP11.

B.4 Results: Transcriptomic profiles unveiled different functions between salivary glands

Table B.4 - Gene Ontology (GO) analysis of the differentially expressed genes (DEGs) with hit in Swiss-Prot database, of anterior (ASG) and posterior (PSG) salivary glands. Statistics were based on Fisher analysis of Top10 GO terms and GO terms of interest (FDR-adjusted $p < 0.05$). Odds-ratios higher than 1 correspond to a GO term enrichment in the posterior salivary gland and lower than 1 correspond to a GO term enrichment in the anterior salivary gland. Note that the same DEG can bear multiple GO terms. Asterisk corresponds to significantly enriched pathways.

GO Terms		Upregulated PSG	Upregulated ASG	Odds ratio	FDRp	
Molecular Function						
1	serine-type endopeptidase activity	33	5	4.072	3.92E-02	*
2	protein kinase activity	16	1	9.734	8.70E-02	
3	signaling receptor binding	14	1	8.495	1.62E-01	
4	protein tyrosine kinase activity	11	1	6.649	2.07E-01	
5	actin binding	26	6	2.643	2.07E-01	
6	protein serine kinase activity	30	8	2.29	2.07E-01	
7	zinc ion binding	39	12	1.991	2.07E-01	
8	calcium ion binding	54	20	1.659	2.52E-01	
9	ATP hydrolysis activity	11	14	0.462	2.52E-01	
10	protein serine/threonine kinase activity	28	8	2.132	2.58E-01	
11	RNA binding	25	25	0.585	2.58E-01	
12	microtubule binding	14	3	2.82	3.56E-01	
13	protein kinase binding	14	3	2.82	3.56E-01	
14	metalloendopeptidase activity	11	2	3.318	3.56E-01	
15	mRNA binding	11	2	3.318	3.56E-01	
16	guanyl-nucleotide exchange factor activity	22	7	1.903	3.91E-01	
17	protein-macromolecule adaptor activity	12	3	2.411	4.01E-01	
18	ubiquitin-protein transferase activity	10	2	3.012	4.47E-01	
19	metal ion binding	101	72	0.814	4.47E-01	
20	actin filament binding	15	5	1.808	5.81E-01	
21	calmodulin binding	23	9	1.543	5.81E-01	
22	DNA binding	25	20	0.739	5.81E-01	

	GO Terms	Upregulated PSG	Upregulated ASG	Odds ratio	FDRp	
Molecular Function						
23	enzyme binding	11	4	1.652	6.55E-01	
24	protein homodimerization activity	29	13	1.346	6.55E-01	
25	DNA-binding transcription factor activity, RNA polymerase II-specific	25	18	0.825	7.29E-01	
26	GTPase activity	13	10	0.773	7.29E-01	
27	transcription coactivator activity	10	4	1.5	7.84E-01	
28	magnesium ion binding	17	8	1.276	8.65E-01	
29	identical protein binding	43	23	1.124	8.65E-01	
30	ubiquitin protein ligase activity	12	6	1.198	9.72E-01	
31	RNA polymerase II cis-regulatory region sequence-specific DNA binding	27	15	1.078	9.81E-01	
32	ATP binding	82	51	0.956	9.81E-01	
33	protein-containing complex binding	14	8	1.047	1.00E+00	
34	GTP binding	19	11	1.033	1.00E+00	
35	chromatin binding	13	8	0.971	1.00E+00	
36	small GTPase binding	13	8	0.971	1.00E+00	
	Cellular Component					
1	Golgi membrane	31	48	0.361	1.23E-03	*
2	Golgi apparatus	26	37	0.4	1.51E-02	*
3	membrane raft	11	1	6.649	2.73E-01	
4	mitochondrial outer membrane	14	2	4.239	2.73E-01	
5	collagen-containing extracellular matrix	20	4	3.039	2.73E-01	
6	glutamatergic synapse	31	7	2.714	2.73E-01	
7	neuron projection	22	5	2.676	2.73E-01	
8	cytoskeleton	25	6	2.538	2.73E-01	
9	axon	34	10	2.078	2.73E-01	
10	mitochondrion	54	19	1.75	2.73E-01	
11	endoplasmic reticulum membrane	32	32	0.581	2.73E-01	
12	focal adhesion	17	4	2.573	4.54E-01	
13	neuronal cell body	24	7	2.082	4.54E-01	
14	extracellular region	82	36	1.404	4.54E-01	
15	plasma membrane	207	104	1.259	4.54E-01	
16	endoplasmic reticulum	34	31	0.64	4.54E-01	
17	sarcolemma	11	2	3.318	5.04E-01	

	GO Terms	Upregulated PSG	Upregulated ASG	Odds ratio	FDRp
	Cellular Component				
18	Z disc	16	4	2.419	5.04E-01
19	intracellular membrane-bounded organelle	18	5	2.178	5.04E-01
20	centrosome	16	16	0.589	5.04E-01
21	endosome	13	14	0.548	5.04E-01
22	extracellular matrix	17	5	2.054	5.09E-01
23	postsynaptic density	17	5	2.054	5.09E-01
24	extracellular space	65	29	1.37	5.09E-01
25	basement membrane	10	2	3.012	5.60E-01
26	cytoplasm	197	103	1.192	5.60E-01
27	nucleoplasm	44	35	0.737	5.60E-01
28	apical plasma membrane	22	19	0.683	5.73E-01
29	lysosomal membrane	15	14	0.634	5.73E-01
30	nuclear body	11	3	2.207	6.02E-01
31	lamellipodium	18	6	1.811	6.10E-01
32	adherens junction	10	3	2.004	7.45E-01
33	GABA-ergic synapse	10	3	2.004	7.45E-01
34	presynaptic membrane	16	6	1.606	7.45E-01
35	cell cortex	13	5	1.563	7.45E-01
36	postsynapse	13	5	1.563	7.45E-01
37	mitochondrial inner membrane	21	9	1.405	7.45E-01
38	postsynaptic membrane	21	9	1.405	7.45E-01
39	protein-containing complex	25	11	1.37	7.45E-01
40	dendrite	34	16	1.282	7.45E-01
41	cytosol	152	82	1.134	7.45E-01
42	lysosome	21	17	0.731	7.45E-01
43	trans-Golgi network	10	9	0.66	7.45E-01
44	chromatin	15	6	1.503	7.48E-01
45	extracellular exosome	30	14	1.292	7.72E-01
46	nucleolus	20	9	1.337	7.97E-01
47	basolateral plasma membrane	17	13	0.777	7.97E-01
48	actin cytoskeleton	12	5	1.441	8.16E-01
49	synapse	29	14	1.247	8.16E-01
50	cilium	10	8	0.744	8.16E-01

	GO Terms	Upregulated PSG	Upregulated ASG	Odds ratio	FDRp	
	Cellular Component					
51	synaptic vesicle membrane	10	8	0.744	8.16E-01	
52	cell surface	20	14	0.85	9.14E-01	
53	synaptic vesicle	11	5	1.319	9.57E-01	
54	mitochondrial matrix	14	7	1.199	9.57E-01	
55	nucleus	155	95	0.969	9.57E-01	
56	endosome membrane	12	8	0.895	9.57E-01	
57	microtubule	10	7	0.852	9.57E-01	
58	nuclear envelope	10	5	1.198	1.00E+00	
59	perikaryon	10	5	1.198	1.00E+00	
60	receptor complex	10	5	1.198	1.00E+00	
61	early endosome	15	8	1.123	1.00E+00	
62	presynapse	15	8	1.123	1.00E+00	
63	nuclear speck	11	6	1.097	1.00E+00	
64	membrane	120	73	0.979	1.00E+00	
65	perinuclear region of cytoplasm	29	18	0.962	1.00E+00	
66	cytoplasmic vesicle	16	10	0.955	1.00E+00	
	Biological Process					
1	cell adhesion	31	3	6.384	8.76E-03	*
2	proteolysis	51	11	2.894	8.80E-03	*
3	endoplasmic reticulum to Golgi vesicle-mediated transport	3	12	0.146	8.80E-03	*
4	protein glycosylation	2	11	0.107	8.80E-03	*
5	intracellular signal transduction	21	3	4.269	7.73E-02	
6	intracellular protein transport	7	13	0.316	1.13E-01	
7	potassium ion transmembrane transport	11	1	6.649	1.85E-01	
8	protein phosphorylation	17	3	3.438	1.85E-01	
9	chemical synaptic transmission	10	1	6.037	2.67E-01	
10	extracellular matrix organization	12	2	3.624	3.10E-01	
11	microtubule cytoskeleton organization	12	2	3.624	3.10E-01	
12	positive regulation of gene expression	20	5	2.426	3.10E-01	
13	inflammatory response	11	2	3.318	4.41E-01	
14	homophilic cell adhesion via plasma membrane adhesion molecules	15	4	2.265	4.41E-01	

	GO Terms	Upregulated PSG	Upregulated ASG	Odds ratio	FDRp
	Biological Process				
15	negative regulation of transcription by RNA polymerase II	25	9	1.682	5.48E-01
16	sarcomere organization	10	2	3.012	5.58E-01
17	signal transduction	25	10	1.51	6.75E-01
18	heart development	10	3	2.004	7.28E-01
19	positive regulation of apoptotic process	10	3	2.004	7.28E-01
20	G protein-coupled receptor signaling pathway	16	6	1.606	7.28E-01
21	nervous system development	16	6	1.606	7.28E-01
22	actin cytoskeleton organization	15	6	1.503	8.83E-01
23	positive regulation of transcription by RNA polymerase II	25	18	0.825	8.93E-01
24	actin filament organization	10	4	1.5	9.55E-01
25	apoptotic process	19	9	1.268	9.66E-01
26	regulation of transcription by RNA polymerase II	33	22	0.892	9.66E-01
27	positive regulation of DNA-templated transcription	15	11	0.811	9.66E-01
28	spermatogenesis	10	8	0.744	9.66E-01
29	protein transport	21	14	0.894	9.78E-01
30	negative regulation of apoptotic process	13	6	1.3	1.00E+00
31	cell-cell adhesion	10	5	1.198	1.00E+00
32	positive regulation of cell population proliferation	15	8	1.123	1.00E+00
33	axon guidance	13	7	1.112	1.00E+00
34	cell differentiation	24	13	1.107	1.00E+00
35	endocytosis	11	6	1.097	1.00E+00
36	regulation of DNA-templated transcription	11	6	1.097	1.00E+00
37	cell division	14	8	1.047	1.00E+00
38	protein localization to plasma membrane	10	6	0.996	1.00E+00
39	innate immune response	11	7	0.938	1.00E+00

C

LIST OF PUBLICATIONS AND COMMUNICATIONS

Publications

Gonçalves, C. and Costa, P.M. (2021) Cephalotoxins: A hotspot for marine bioprospecting? Frontiers in Marine Science 8, 1 – 7. (DOI: 10.3389/fmars.2021.647344)

Gonçalves, C., Moutinho Cabral, I., Alves de Matos, A.P., Grosso, A.R. and Costa, P.M. (2024) Transcriptome profiling of the posterior salivary glands of the cuttlefish *Sepia officinalis* from the Portuguese West coast. Frontiers in Marine Science 11:1362824. (DOI: 10.3389/fmars.2024.1362824)

Moutinho Cabral, I., **Gonçalves, C.**, Grosso, A.R. and Costa, P.M. (2024) Bioprospecting and marine 'omics': surfing the deep blue sea for novel bioactive proteins and peptides. Frontiers in Marine Science 11:1362697. (DOI: 10.3389/fmars.2024.1362697)

Gonçalves, C., Alves de Matos, A.P., Grosso, A.R. and Costa, P.M. (*in prep.*). Salivary gland function, complexity and venom secretion in cephalopods: the *Octopus vulgaris* case study and a comparison with *Sepia officinalis*

Gonçalves, C., Jacobsen, R.G., Goksøyr, A., Costa, P.M. and Karlsen, O.A. (*in prep*.) Human G protein–coupled receptors as targets of bioactives from the cephalopod venom gland.

Poster communications

Gonçalves, C., Moutinho Cabral, I., Rodrigo, A.P., Madeira, C., Alves de Matos, A.P., Grosso, A.R. and Costa, P.M. (2023) Uncovering the potential of Cephalopoda and Polychaeta toxins for drug discovery. Encontro Ciência 2023, Aveiro (Portugal), July 2023.

Gonçalves, C., Grosso, A.R. and Costa, P.M. (2022) Advances on the venom glands of Cephalopoda: Targets for the prospecting of novel bioactives. UCIBIO Scientific Advisory Board Meeting, Caparica (Portugal), July 2022.

Gonçalves, C. and Costa, P.M. (2022) Unravelling cephalopod venom glands through histological and transcriptomic approaches. Encontro Ciência'22, Lisboa (Portugal), May 2022

Gonçalves, C., Alves de Matos, A.P. and Costa, P.M. (2022) On the quest for venoms: Clues from the microanatomy of cephalopod salivary glands. Cephalopod International Advisory Council (CIAC 2022), Sesimbra (Portugal), April 2022.

Gonçalves, C., Fonseca, D., Silva, F., Grosso, A.R., Costa, P.M. (2021) New insights into cephalopods venoms – A promising target for bioprospection. European Conference on Marine Natural Products (ECMNP2021), Galway (Ireland), August 2021.

Molecular resources

The datasets obtained within my PhD can be found in online repositories. As such, the full coding sequences of shortlisted genes of interest were publicly available from GenBank (https://www.ncbi.nlm.nih.gov/genbank/), under accession numbers OP198203 to OP198205 for *Sepia officinalis* and OP209720 to OP209723 for *Octopus vulgaris*:

Gonçalves, C., Moutinho Cabral, I., Grosso, A.R. and Costa, P.M. Cysteine-rich venom protein from *Sepia officinalis*. Direct submission to GenBank – Accession number: OP198203

Gonçalves, C., Moutinho Cabral, I., Grosso, A.R. and Costa, P.M. SE-cephalotoxin from *Sepia officinalis*. Direct submission to GenBank – Accession number: OP198204

Gonçalves, C., Moutinho Cabral, I., Grosso, A.R. and Costa, P.M. Venom Insulin from *Sepia officinalis*. Direct submission to GenBank – Accession number: OP198205

Gonçalves, C., Grosso, A.R. and Costa, P.M. Cysteine-rich venom protein from *Octopus vulgaris*. Direct submission to GenBank – Accession number: OP209720 (under publication embargo)

Gonçalves, C., Grosso, A.R. and Costa, P.M. Cysteine-rich venom protein (latisemin-like) from *Octopus vulgaris*. Direct submission to GenBank – Accession number: OP209721 (under publication embargo)

Gonçalves, C., Grosso, A.R. and Costa, P.M. Chitinase from *Octopus vulgaris*. Direct submission to GenBank – Accession number: OP209722 (under publication embargo)

Gonçalves, C., Grosso, A.R. and Costa, P.M. Hyaluronidase from *Octopus vulgaris*. Direct submission to GenBank – Accession number: OP209723 (under publication embargo)

The transcriptome data of *Sepia officinalis* and *Octopus vulgaris* can be freely accessible at Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), under the following accession numbers:

Gonçalves, C., Moutinho Cabral, I., Grosso, A.R. and Costa, P.M. Transcriptome profiling of the posterior salivary gland of the cuttlefish *Sepia officinalis*. Direct submission to GEO – Accession GSE251667

Gonçalves, C., Moutinho Cabral, I., Grosso, A.R. and Costa, P.M. Transcriptome profiling of the posterior and anterior salivary glands of the common octopus, *Octopus vulgaris*. Direct submission to GEO – Accession GSE287337 (under publication embargo)

2025

CÁTIA VANESSA CAETANO GONÇALVES

Coleoid venoms: predicting cephalotoxin function and biotechnological applications from ecological and evolutionary traits