NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF LIFE SCIENCES

INÊS MOUTINHO FERREIRA FIGUEIREDO CABRAL

BSc in Molecular and Cellular Biology

A COMPUTATIONAL APPROACH TO IDENTIFY TARGET RECEPTORS OF MARINE TOXINS IN THE HUMAN PROTEOME: POTENTIAL BIOTECHNOLOGICAL APPLICATIONS

MASTER IN MOLECULAR GENETICS AND BIOMEDICINE

NOVA University Lisbon

November. 2021

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF LIFE SCIENCES

INÊS MOUTINHO FERREIRA FIGUEIREDO CABRAL

BSc in Molecular and Cellular Biology

A COMPUTATIONAL APPROACH TO IDENTIFY TARGET RECEPTORS OF MARINE TOXINS IN THE HUMAN PROTEOME: POTENTIAL BIOTECHNOLOGICAL APPLICATIONS

MASTER IN MOLECULAR GENETICS AND BIOMEDICINE

NOVA University Lisbon

November, 2021

DEPARTAMENT OF
LIFE SCIENCES

A COMPUTATIONAL APPROACH TO IDENTIFY TARGET RECEPTORS OF MARINE TOXINS IN THE HUMAN PROTEOME: POTENTIAL BIOTECHNOLOGICAL APPLICATIONS

INÊS MOUTINHO FERREIRA FIGUEIREDO CABRAL

BSc in Molecular and Cellular Biology

Adviser: Pedro Manuel Brôa Costa
Assistant Professor, NOVA School of Science and Technology, NOVA University Lisbon

Co-adviser: Ana Rita Fialho Grosso
Researcher, NOVA School of Science and Technology, NOVA University Lisbon

Examination Committee:

Chair: Paula Maria Theriaga Mendes Bernardo Gonçalves
Associate Professor, NOVA School of Science and Technology, NOVA University Lisbon

Rapporteurs: Francisco Pina-Martins
Assistant Professor, Faculty of Science, University of Lisbon

Adviser: Pedro Manuel Brôa Costa
Assistant Professor, NOVA School of Science and Technology, NOVA University Lisbon

MASTER IN MOLECULAR GENETICS AND BIOMEDICINE

NOVA University Lisbon
November, 2021

A COMPUTATIONAL APPROACH TO IDENTIFY TARGET RECEPTORS OF MARINE TOXINS IN THE HUMAN PROTEOME: POTENTIAL BIOTECHNOLOGICAL APPLICATIONS

ACKNOWLEDGMENTS

First, I would like to thank Pedro Costa for all the support, motivation, advice and unconditional patience. I am very grateful to you for believing in me.

To Ana Rita, for always being available and for motivating me.

To Ana and Cátia, for all the patience and support as well as for stopping me from stressing out. To all the Seatox team, especially, Carla, Carolina and Mariaelena, for all the support and the warm welcome. To Marta, Sónia and Telma, for the incentive.

To all the Comics team, particularly, Daniel for helping me through the battle against the server and Mariana for all the support.

A special thank you to my family, especially, my parents and sister, for always being there for me and for believing in me.

Finally, I would also like to thank all my friends, mainly, Pernadas, Ana Sofia, Beatriz, Filipa, Phoebe, Júlia, Meggy, Mónica and Rita, for the motivation and great times. To all the friends that FCT gave to me, especially, Condez, Cat and Belinha.

ABSTRACT

The marine environment has a tremendous biodiversity, which implies an almost limitless source of bioreactives with potential biotechnological applications. Within these, proteinaceous toxins are highlighted as most evolved to interact with specific molecular targets. Toxins are part of complex cocktails, such as venoms, that are secreted for predation or defence. Recently, new drugs developed from marine natural products reached the market, like Prialt, a synthetic painkiller derived from a conotoxin. Altogether, marine invertebrates can be high-prized sources for novel bioreactives with biomedical applications, particularly the Polychaeta, due to their abundance and diversity albeit little explored. *Glycera alba* and *Hediste diversicolor* are two Polychaeta with distinct behaviours, but suspected to secrete toxins with different purposes. A comparative transcriptomic analysis between species and organs revealed distinct toxins and other bioreactives. Specifically, the venom apparatus of *G. alba* is localized in the proboscis and that neurotoxins and diffusing agents are secreted to overwhelm prey. On the other hand, *H. diversicolor*, an opportunistic forager, secretes fewer, less specific, toxins that are seemingly a defence measure against predators and pathogens. The analysis of protein-protein interactions between proteins predicted from the worms' transcriptome and the human proteome allowed unravelling novel toxins and bioreactives with potential biomedical applications, from which two full-coding sequences from both species were isolated. Among these, are proteins from *G. alba*'s venom that can interfere with regulatory pathways of apoptosis, for instance involving FADD, BAD and FAIM. In turn, *H. diversicolor* yielded proteins that can regulate the human innate immune response. These results show that omics and bioinformatics can be a powerful tool for bioprospecting and drug discovery, enabling mining through complex transcriptomes even of organisms with reduced genomic annotation. They also show that interactome-directed analysis against the druggable human proteome can be a highly efficient alternative to the design of synthetic drugs.

Keywords: *Glycera alba*, *Hediste diversicolor*, Polychaeta, RNA-Seq, Venom, Interactome-directed analysis

RESUMO

A enorme biodiversidade do ambiente marinho poderá implicar uma fonte quase ilimitada de bioreactivos com potencial para aplicações biotecnológicas. As toxinas peptídicas destacam-se, pois a maioria evoluiu para interagir com alvos molecular específicos. Estes constituintes de cocktails complexos, nomeadamente, de venenos, são secretados para predação ou defesa. Recentemente, novos fármacos baseados em produtos naturais marinhos chegaram ao mercado, como Prialt, um analgésico sintético desenvolvido a partir de uma conotoxina. Os invertebrados marinhos, nomeadamente poliquetas, podem ser fontes com elevado potencial para a descoberta de novos bioreactivos com interesse biomédico devido à sua abundância e diversidade, apesar de serem poucos estudados. *Glycera alba* e *Hediste diversicolor* são dois poliquetas com comportamentos distintos, mas poderão secretar toxinas com diferentes intuitos. A análise transcriptómica comparativa entre espécies e órgãos revelou diferentes toxinas e bioreactivos. Segundo os resultados, o sistema de secreção de veneno de *G. alba* está localizado no proboscis, onde ocorre a secreção de neurotoxinas e agentes permeabilizantes para paralisar as presas. No entanto, *H. diversicolor*, um oportunista em termos alimentares, secreta menos toxinas e com menor especificidade como uma medida de defesa contra predadores e patógenes. A análise de interações proteína-proteína entre proteínas destes anelídeos e o proteoma humano revelou novos bioreactivos com potenciais aplicações biomédicas, entre os quais duas regiões codificantes de ambas as espécies foram isoladas. Enquanto que proteínas do veneno de *G. alba*, como BAD, FAD e FAIM, podem interferir com a regulação da apoptose, proteínas de *H. diversicolor* poderão regular a resposta imunitária inata humana. Os resultados ilustram que as ómicas e a bioinformática são poderosas ferramentas para a bioprospecção marinha e descoberta de novos fármacos, mesmo através do estudo de transcriptomas complexos de organismos com reduzida anotação genómica. Uma análise dirigida contra o interactoma com proteoma humano poderá constituir uma alternativa ao desenho de fármacos sintéticos.

Palavras chave: *Glycera alba*, *Hediste diversicolor*, Poliquetas, RNA-Seq, Venenos, Análise dirigida contra o interactoma

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BAD	BCL2 associated agonist of cell death (gene).
BAD	Bcl2-associated agonist of cell death (protein).
BLAST	Basic Local Alignment Search Tool.
BLASTP	Basic Local Alignment Search Tool for Proteins.
CRISP	Cysteine-rich secretory protein.
DAVID	Database for Annotation, Visualization and Integrated Discovery.
DEG	Differentially-expressed genes.
DNA	Deoxyribonucleic acid.
***e*-value**	Expectation value or Expect value.
Expasy	SIB Expasy Bioinformatics Resources Portal.
FADD	Fas associated via death domain (gene).
FADD	FAS-associated death domain protein.
FAIM	Fas apoptotic inhibitory molecule (gene).
FAIM	Fas apoptotic inhibitory molecule 1 (protein).
FDR	False discovery rate.
GO	Gene ontology.
HuRI	The Human Reference Protein Interactome.
$\log_2$CPM	Average $\log_2$ counts per million.
$\log_2$FC	$\log_2$ fold change.
MEGA	Molecular Evolutionary Genetics Analysis.

NCBI	National Center for Biotechnology Information.
ORF	Open reading frame.
PCR	Polymerase chain reaction.
RNA	Ribonucleic acid.
RNA-Seq	Ribonucleic acid sequencing.
RT-qPCR	Quantitative reverse-transcription polymerase chain reaction.
Swiss-Prot	"UniProtKB/Swiss-Prot" (reviewed, manually annotated).
UniProt	Universal Protein Resource.
UniProtKB	UniProt Knowledgebase.

1

INTRODUCTION

1.1 Potential role of Marine-derived compounds in Human Health

Marine biotechnology has been attracting growing enthusiasm over the last few years due to the immense biodiversity of marine habitats that potentiates a unique chemical diversity (Montaser & Luesch, 2011). Additionally, the ancient radiation of marine life enabled marine compounds that are secreted for defence and predation, such as toxins and other venom components, to evolve towards higher specificity through more selective binding affinity for their molecular targets. Thus, these natural products are a more efficient and sustainable alternative to the design of new synthetic molecules, which is riskier and more time-consuming (see for instance Hong, 2011; Burgess, 2012). Consequently, bioprospecting for novel marine natural products has a key role in Blue Biotechnology.

The Blue Growth agenda was devised under the auspices of a sustainable trade-off between social-economic development and the marine environment and encompasses exploitation of renewable energy, aquaculture, tourism, seabed mining and marine biotechnology, which precludes bioprospecting for novel bioreactives (see, for instance, Lillebø et al., 2017; Burgess et al., 2018). Applications of marine bioproducts are expected to range between, for instance, nutritional supplements, eco-friendly pesticides and personal care products (PCPs), biopolymers to enzymes for research and diagnosis (for a review, see Imhoff et al., 2011; Freitas et al., 2012; Martins et al., 2014; Hamed et al., 2018). Most importantly, the pharmaceutical industry has also been investing in drug development from marine bioreactives. Even though marine products with potential biotechnological applications in biomedicine belong to various chemical classes (e.g., nucleosides, peptides, alkaloids, fatty acids and others), proteins and peptides are particularly interesting for drug development due to an easier synthesis process that is based on a heterologous expression strategy (Molinski et al., 2009; Martins et al., 2014; Hu et al., 2015). Among these, proteinaceous toxins engage special attention as there exists an almost limitless source of these molecules as well as they are a part of the chemical warfare developed by

marine animals and, consequently, they can interact with specific molecular targets and affect the target metabolism (Fry et al., 2009; Casewell et al., 2013).

As new efforts are being made to make marine bioprospecting more efficient, there are some drugs already on the market that are based on marine bioactive compounds, particularly those based on toxins (for a review, see Molinski et al., 2009; Montaser & Luesch, 2011; Martins et al., 2014). Briefly, Ziconotide, whose trade name is Prialt, is a synthetic form of the ω-conotoxin MVIIA present in the *Conus magus* venom, a marine cone snail, that was approved for management of severe chronic pain by the United States Food and Drug Administration (FDA) in 2004 (for a review, see Williams et al., 2008). Since this twenty-five amino acid peptide blocks N-type voltage-sensitive calcium-channels on the spinal cord dorsal horn, the triggering of neurotransmission by the nociceptive afferents nerves is inhibited (Olivera et al., 1987; Bowersox et al., 1996). Apart from proteinaceous toxins, there are drugs based on secondary metabolites that have already been commercialized, for instance, the Yondelis (trade name of Trabectedin), which was approved by the European Union in 2007 to treat soft tissue sarcoma (for a review, see Christinat & Leyvraz, 2009). This alkaloid is derived from a compound present in the marine tunicate *Ecteinascidia turbinata*, however the low yields from direct extraction rendered the necessity to develop a semi-synthetic process to obtain this molecule. It starts with the fermentation of *Pseudomonas fluorescens* for producing the antibiotic safracin B, followed by chemical reactions (Rinehart et al., 1990; Cuevas et al., 2000). This substance, which is used in conjunction with other drugs, such as pegylated liposomal doxorubicin, acts by bending the DNA towards the major groove through the alkylation of guanine at the N2 position in the minor groove and also by interacting with transition-coupled nucleotide excision repair system leading to cell death (Pommier et al., 1996; Zewail-Foote & Hurley, 1999; Takebayashi et al., 2001).

Despite the examples stated above and the promises of marine bioprospecting in the last two decades, the number of marine compounds effectively translated into commercial products is reduced (for a review, see Molinski et al., 2009; Burgess, 2012; Martins et al., 2014). Different explanations can be pointed out for this disappointing result. For instance, a supply problem could exist. This situation happens since the harvest yield of desirable marine product is very low or when the marine organism lives in nearly unreachable habitats, making their acquisition more challenging and expensive (Molinski et al., 2009; Martins et al., 2014). Moreover, the difficulty of characterising the different molecules present in poisons, venoms and other biological mixtures due to poor genomic annotation along with the challenges of figuring out their combined effect and their interaction with the targets are important bottlenecks (Martins et al., 2014; Rodrigo & Costa, 2019). Although biodiversity is a major booster of drug development from marine products, it can also become a challenge, since a substantial part of the oceans' immense biodiversity is still unexplored, as well as the physiology of many marine organisms, especially invertebrates (Martins et al., 2014; Rodrigo & Costa, 2019). Aquaculture, semi-synthesis or even complete chemical synthesis *in vitro* are being adopted as solutions to solve the supply problem, while omics approaches are facilitating new insights in the characterisation of the different molecules in

the biological samples as well as trying to predict their mode-of-action and effects (Molinski et al., 2009; Montaser & Luesch, 2011; Martins et al., 2019; Rodrigo & Costa, 2019).

For the purpose, omics can provide invaluable advantages and ultimately result in the isolation of the full coding sequences. The heterologous expression of these genes either in bacteria or in eukaryotic cells is an easier and more cost-effective process than the production of secondary metabolites since many enzymes and/or other chemical reactions are needed to obtain the non-proteinaceous compounds. The protein sequence isolated can then be used to predicted protein-protein interactions with the druggable human proteome (all the human proteins that could interact with drug-like molecules according to Hopkins & Groom (2002)). This analysis identifies the possible human targets and tries to predict the effects of the potential drugs on the pathways. Depending on the targets affected, a shortlist of proteins with more interest and potential to proceed with the drug discovery process is produced. The off-targets (non-therapeutic drug targets) can also be detected, which might facilitate the mitigation of the side-effects. Thus, the interactome analysis can be a more cost-efficient alternative for the drug discovery process as many studies and trials are cancelled due to misidentification of the targets and unexpected side-effects. Recent efforts are being made as new drug targets are continuously being added to the druggable human proteome and the tools for predicting protein-protein interactions used for the design of new synthetic drugs based on coevolutionary analysis and fragmental docking could potentially be employed in this novel approach for drug discovery (Bai et al., 2016; Wang et al., 2020). Therefore, the isolation of the full coding sequences and the following identification of the domains, which can influence the protein reactivity and the protein-protein interactions, are pivotal to this strategy.

1.2 Discovery of novel relevant compounds by transcriptome profiling

Next-generation RNA sequencing (RNA-Seq) is one of the most recent transcriptomic techniques that enables studying whole-transcriptomes and at same time detects isoforms and quantifies gene expression (for a review, see Wang et al., 2009; Martins et al., 2019). This method is based on total RNA extraction and fragmentation prior to reverse transcription double-stranded cDNA sequencing. The methodology relies heavily on bioinformatics tools due to the needs for massive computations, beginning with *de novo* transcriptome assembly whenever there is no reference transcriptome against which to map the sequenced reads, as common for non-conventional model and wild organisms (Wang et al., 2009; Martins et al., 2019). Although the large amount of data generated by this approach is a drawback, it can also pull out much more knowledge and information about the organism, pathways, mechanisms and even the identification of toxins (Wang et al., 2009; von Reumont et al. 2014a; Martins et al., 2019). Therefore, the RNA-seq technology allied with bioinformatics is being applied to study marine invertebrates, a very diverse group of unexplored and poorly studied animals that includes Phyla Cnidaria, Porifera, Annelida and Mollusca (for a review, see von Reumont et al. 2014a). Indeed, whole-

transcriptomes of the annelids *Eulalia* sp., *Glycera*, several amphinomids and the gastropod *Colubraria reticulata* were already sequenced to characterize the chemical warfare developed by marine invertebrates and to identify toxins (von Reumont et al. 2014b; Modica et al., 2015; Verdes et al., 2018; Rodrigo et al., 2021). As previously referred, the fact that toxins interact with specific molecular targets and interfere with biological pathways makes the venomous marine invertebrates an interesting source for the development of new drugs. Furthermore, as toxins and other compounds extracted from these organisms are showing remarkable antitumour, anti-inflammatory, antiviral, and antimicrobial properties, the interest on these molecules for potential biotechnological applications is increasing. Despite the current research on this group, many of its biodiversity is still unexplored. Lately, compounds resulting from the symbiosis between microorganism and marine invertebrates are also being captivating attention (Molinski et al., 2009; Martins et al., 2014; Rodrigo & Costa, 2019).

1.3 Polychaeta: the unexplored and rich resource of novel marine-derived compounds

Despite their abundance and ecological relevance, the Polychaeta are little studied for biotechnological purposes. The marine Polychaeta are diverse and widespread marine invertebrates, as they occupy very distinct habitats ranging from the deep-sea vents to the intertidal and their feeding habits vary from being predators, herbivores, deposited or even filter feeders (Hutchings, 1998). Moreover, a broad spectrum of substances of interest is produced by Polychaeta. While arenicin is a peptide from *Arenicola marina* that has antimicrobial proprieties against fungi and bacteria, nereistoxin is a non-proteinaceous toxin isolated from *Lumbriconereis heteropoda* that has neurotoxic effects as it is going to provoke a neuromuscular block and convulsions (Chiba et al., 1967; Ovchinnikova et al., 2004). *Eulalia* sp. secrets phyllotoxins, which are mainly cysteine-rich proteins that have a neuromuscular effect on the prey and are toxungenous components and not venomous, since they are not delivered via an inflicted wound but instead applied to the prey's body surface (Nelsen et al., 2014; Rodrigo et al., 2021). Altogether, the secretion of toxins and other bioreactives render these annelids as a potential source for biotechnological applications (Rodrigo & Costa, 2019). *Glycera* is a burrower marine Polychaeta with the ability of secreting venom to overwhelm its prey (mostly other Polychaeta and crustaceans), which are grabbed by *Glycera* due to a quick eversion of the proboscis after having detected small variations in the hydrostatic pressure (see for instance Ockelmann & Vahl, 1970; von Reumont et al. 2014a). The little knowledge about this genus was mainly obtained through the few studies of three species: *Glycera dibranchiata*, *Glycera tridactyla* and *Glycera fallax*. *Glycera* has four jaws containing a melanin-like network and the copper-based biomineral atacamite combined with proteins, which makes them highly resistant to abrasion (Gibbs & Bryan, 1980; Lichtenegger et al., 2002; Moses et al., 2006). This characteristic is pivotal since in order to inject the venom into the prey, the jaws have to penetrate the integument beforehand. In *Glycera*, each jaw is connected to a venom gland through a channel and has pores on the ventral side, allowing venom injection. The venom may then provoke progressive paralysis

of the prey, cardiac arrest and convulsions followed by death (Michel & Keil, 1975; Bon et al., 1985). When *Glycera* is mishandled as a fish bait, the jaws may puncture human skin and could cause localised itching and swelling (Smith, 2002). Previous studies have already indicated that *Glycera* venom is a complex and heterogeneous cocktail, which includes, for instance, neurotoxins and pore-forming toxins (Michel & Keil, 1975; Kagan et al., 1982; Bon et al., 1985; Meunier et al., 2002; von Reumont et al. 2014b). Glycerotoxin is a neurotoxin produced by *Glycera* that interacts with N-type calcium channels ($Ca_v2.2$) and reversibly up-regulated their activity (Meunier et al., 2002). By being an agonist of these pre-synaptic channels, the glycoprotein is going to increase the calcium influx and provoke a long-lasting spontaneous release of the neurotransmitters (Schenning et al., 2006; Meunier et al., 2010). The difference between this excitatory neurotoxin and the alpha-latrotoxin produced by the black widow spider is that glycerotoxin does not provoke a complete depletion of synaptic vesicles (Ceccarelli & Hurlbut, 1980; Meunier et al., 2010). A recent transcriptomic approach of venom composition from *G. dibranchiata*, *G. tridactyla* and *G. fallax* venom glands, complemented with phylogenetic analysis suggested that toxins could be divided into five groups: neurotoxins, pore-forming and membrane-disruptive toxins, proteases inhibitors, other enzymes and CAP domain proteins (von Reumont et al., 2014b). Whereas *Glycera* is an ambush predator, *Hediste* (*Nereis*) is an estuarine genus of omnivores that can build and inhabit in burrows on soft-bottom marine surfaces or become an opportunistic forager (for a review, see Scaps, 2002). This Polychaeta has an eversible proboscis with two robust jaws, which are used for grasping food (Bryan & Gibbs, 1979). In this family (Nereididae), namely, in the species *Nereis limbata*, the jaws are described to have zinc-chlorine compounds together with proteins rich in glycine and histidine (Lichtenegger et al., 2002; Lichtenegger et al., 2003). The feeding habits of *Hediste* range from being omnivorous, an active predator or a deposit feeder (Scaps, 2002). As phytoplankton concentration reaches a certain value, *Hediste diversicolor* (O.F. Müller, 1776) changes its feeding strategy and becomes a filter-feeder, as it ingests the food particles brought by the water column and caught by the mucous secreted from parapodial glands (Harley, 1950; Vedel et al., 1994). This annelid has a commercial and economical interest, since it can be sold as fish bait. Additionally, *H. diversicolor* is also used as a model in several ecotoxicological studies (Gomes et al., 2019). Recently, it was suggested that the Nereididae family, which *Hediste* belongs to, might secrete toxins by the skin to defend themselves (Gonçalves & Costa, 2020). Moreover, hedistin, which is an antimicrobial peptide with bromotryptophan residues secreted by *H. diversicolor*'s natural killer cells-like upon an immune challenge, appears to have activity against a large range of bacteria, which includes the methicillin resistant *Staphylococcus aureus* and *Vibrio alginolyticus* (Tasiemski et al., 2007). It must be noted, though, that the relationship between ecological traits and the properties of venoms is not well understood in Polychaeta, as in marine animals altogether. It is hypothesised that comparative transcriptomics can provide an overview of differences and similarities between these distinct but sympatric species as well as an opportunity to isolate novel marine bioreactives.

1.4 Objectives

Although there is already transcriptomic analysis of some species from the *Glycera* genus (*Glycera dibranchiata*, *Glycera tridactyla* and *Glycera fallax*), none is referent to *Glycera alba* (O.F. Müller, 1776). Therefore, this would be the first whole-transcriptome study of the species *G. alba* and also of *H. diversicolor*. These two distinct Polychaeta are suspected of secreting venom with different purposes: predation and defence, respectively. Thus, this work is focused on unravelling novel bioreactives from these two Polychaeta, preferably proteinaceous toxins that could interact with the druggable human proteome and, consequently, might have biomedical applications. The detailed objectives can be summarised as:

- Identify novel proteinaceous toxins secreted by *G. alba* and *H. diversicolor*.
- Perform a comparative analysis between *G. alba* and *H. diversicolor* and assess whether the different predatory behaviours (venom-injecting predator and opportunist forager, respectively) could be related with the type of biomolecules produced and secreted by these marine invertebrates.
- Discover and describe the organ or tissue used by these two marine invertebrates for the secretion of toxins and other bioreactives.
- Predict the main toxins' mechanisms of action.
- Isolate full coding sequences of interest.
- Identify potential targets for the toxins and other proteins of interest in the druggable human proteome.

2

MATERIAL AND METHODS

2.1 Animals and sampling

Glycera alba and *Hediste diversicolor* were hand-collected from Seixal (38°38'41.7" N, 9°06'08.2" W) and Alcochete (38°45'40.1" N, 8°56'08.1" W), Tagus Estuary, Western Portugal, from the muddy-sandy intertidal flats (Figure 2.1). The animals were acclimatised in the laboratory in a controlled mesocosm environment for approximately two weeks before processing. Worms were dissected, the proboscis and whole-body wall (mentioned onwards as skin) were excised from *G. alba*, as well as the glandular and muscular regions of *H. diversicolor*'s proboscis, here forth referred to as "glands" and "proboscis" (refer to Appendix Table A.1 for more specific details on sampling). To optimise RNA preservation, organ samples were immediately submerged in RNAlater (Sigma-Aldrich, St. Louis, MO, USA). The samples were stored at 4ºC considering RNA extraction was performed within a week. Otherwise, after an incubation overnight at 4°C, RNAlater was removed and samples were stored at -80ºC until RNA extraction.

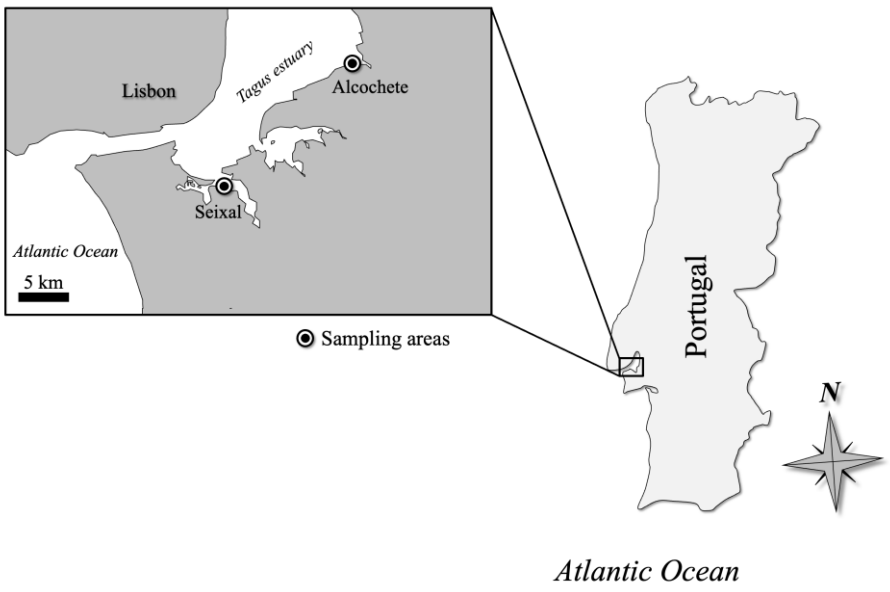

Figure 2.1. Map of the sampling areas. *Glycera alba* was collected from Seixal (38°38'41.7" N, 9°06'08.2" W), while *Hediste diversicolor* was collected from Alcochete (38°45'40.1" N, 8°56'08.1" W), two intertidal flats of the Tagus Estuary, Western Portugal. Map was adapted from Gonçalves & Costa (2020).

2.2 RNA extraction

Total RNA was isolated from 15-20 mg of organ per sample with the RNeasy Mini Kit (Qiagen, Hilden, Germany) following manufacturer's instructions. Residual DNA was eliminated on-column using the RNase-Free DNase set (Qiagen, Hilden, Germany). Preliminary quantification and quality assessment of RNAs were done using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Quantification of total RNA and estimation of the RNA Integrity Number (RIN) were obtained using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). All the samples had input of ≥ 1 µg total RNA and RIN ≈ 7 or higher, which are regarded as suitable for RNA-Seq (Schroeder et al., 2006). Both the collection of animals and RNA extraction were performed by C. Madeira (SeaTox Lab) in 2019-2020.

2.3 RNA sequencing and data pre-processing

Libraries were constructed by using the Stranded mRNA Library Preparation Kit, after which cDNA libraries were sequenced in the Illumina Novaseq platform (Illumina, San Diego, CA, USA) with 150bp paired-end reads with 100 or 40 million reads (Table A.1). The quality of raw sequence data obtained from the RNA-Seq was assessed with FastQC v0.11.9 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). TrimGalore v0.6.6 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was run with the default parameters in order to trim the Illumina adapter and the reads with length under 20 bp. A maximum of 0.75% of low-quality reads was removed from each sample.

2.4 Transcriptome assembly, quality and quantification

In absence of available transcriptomes for the species from public databases for read mapping, transcriptomes were *de novo* assembled using Trinity v2.6.6 (Grabherr et al., 2011; Haas et al., 2013). In the case of *G. alba*, all samples (40M reads) were used to construct the reference transcriptome, while in *H. diversicolor*, the reference transcriptome was assembled from two samples (one per organ) sequenced with 100M reads. Good quality of transcriptome assembly was verified through contigs N50, Ex90N50 and read content statistics by employing Trinity, Bowtie2 v2.4.1 and Samtools v1.11 (Langmead & Salzberg, 2012; Haas et al., 2013; Langmead et al., 2019; Danecek et al., 2021). To assess the individual expression levels, reads from all samples were mapped against the respective assembled transcriptome and transcript abundance quantified using Kallisto v0.43.0 (Bray et al., 2016).

2.5 Data analysis

All statistics were computed using R 4.0.3+ (refer to Appendix R Script for the code) (Ihaka & Gentleman, 1996). The identification of the transcripts that were differentially-expressed between two organs (proboscis and skin in *G. alba*; glands and proboscis in *H. diversicolor*) were achieved by employing the

edgeR v3.34.0 and limma v3.48.1 packages (Robinson et al., 2010; McCarthy et al., 2012; Ritchie et al., 2015). The R packages tximport v1.18.0 and seqnir v4.2-8 were used beforehand, so the data could be imported to R (Charif & Lobry, 2007; Soneson et al., 2015). The normalization factors were estimated by using the edgeR package. A linear model was fitted to the DGEList containing the RNA-Seq data normalized, allowing in that way to contrast the expression between tissues. Cut-offs to filter differentially-expressed genes between organs were set at $|\log_2 FC| > 1.5$ and FDR-adjusted $p < 0.05$. The packages gplots v3.1.1 (Warnes et al., 2020), RColorBrewer v1.1-2 (Neuwirth, 2014), ggplot2 v3.3.5 (Wickham, 2016), cowplot v1.1.1 (Wilke, 2020) and gridGraphics v0.5-1 (Murrell & Wen, 2020) were used for plotting.

2.6 Functional annotation

The predicted coding regions in the assembled transcriptomes were identified by using TransDecoder v5.5.0 (Haas et al., 2013). After, the predicted ORFs were functionally annotated by evaluating protein domains (*e*-value < 0.05) considering Pfam database (Pfam-A.hmm version 33.1, May 2020) through HMMER v3.3.1, and also by homology-matching against the Swiss-Prot database (version 18-11-2020 13:41, downloaded from https://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz) with BLASTP (*e*-value < 1E-5) from NCBI blast+ v2.10.0+ (Camacho et al., 2009; Eddy, 2009; Mistry et al., 2021). Since Swiss-Prot lacks poorly supported proteins, namely specific from these two Polychaeta species, a second BLASTP was performed to identify the transcripts coding for previously described glycerotoxin (Accession: A0A1U9VX91, A0A1U9VX95 and A0A1U9VX98) and hedistin (Accession: Q1PG44) from UniProt (UniProt release 2021_03) (Camacho et al., 2009; The UniProt Consortium, 2021). The sequences with homology-matching were annotated through the R package UniProtR v2.0.7 (UniProt release 2021_03), including the retrieval of the Gene Ontology (GO) Terms (Soudy et al., 2020; The UniProt Consortium, 2021). The GO and protein domain enrichment analysis was computed using R and the significance criterion was set at FDR-adjusted $p < 0.05$.

2.7 Validation of novel transcripts

Validation of transcriptome assembly and quantification of gene expression was done by isolation of transcripts using PCR methods, sequencing PCR products and quantitative reverse-transcription PCR (RT-qPCR). Transcripts were selected according to their possibility of encoding proteins that could have toxin activity or interfere with ion channels as well as according to the statistics associated with the differential expression and the annotation. Beta-actin and 18S were the housekeeping genes chosen as calibrators and the primers were designed (Table A.2). In brief, cDNA synthesis was done using the First-Strand cDNA Synthesis Kit (NZYTech, Lisbon, Portugal), following manufacturer instructions (initial incubation at room temperature for 10 min, followed by incubations at 50 ºC for 30 min, at 85 ºC for 5 min and a last one at 37 ºC for 20 min that was preceded by the addition of RNase H). Sequences were then isolated by PCR, which involved primer design to attempt amplification of the full coding sequence. The PCR was performed in Biometra Gradient Thermocycler96 (Analytik Jena, Jena, Germany) and the

PCR reagents were from Invitrogen (Thermo Fisher Scientific, Waltham, MA, USA). The annealing temperature ranged from 48 to 50ºC during 30s to 45s, while the extension lasted between 1min 20s to 1min 30s. The PCR products were resolved in an agarose gel and those whose amplification was a success were sequenced in an ABI 3730 xl sequencer (Thermo Fisher Scientific, Waltham, MA, USA). Following their alignment with the RNA-Seq sequences in MEGA X v10.2.6, the consensus sequences were translated and homology-matched against UniProt database by employing Expasy and their BLAST tool, respectively (Stecher et al., 2020; Duvaud et al., 2021). Specific primers were designed to amplify an expressed sequence tag (EST) in the coding region and verified with Primer Blast (Ye et al., 2012). Quantification was performed in Rotor-Gene Q (Qiagen, Hilden, Germany) and NZY qPCR Green Master Mix (NZYTech, Lisbon, Portugal) was used alongside. An initial denaturation at temperature 95ºC for 10 min preceded 55 cycles characterised by a denaturation at 94ºC during 45s, an annealing at 50ºC for 25s and an extension at 72ºC during 30s. The program finished with an incubation at 50ºC that lasted 1min, followed by a melting phase, where the temperature ranged from 50 to 99ºC (rising 1ºC every 5s). Expression was determined using the $2^{-\Delta\Delta Ct}$ method (Livak & Schmittgen, 2001). Normality of data and homogeneity of variances were assessed though Shapiro's and Levene's tests (R package car v3.0-11), respectively (Fox & Weisberg, 2019). Differences in gene expression between the two organs were assessed through the parametric Student's *t*-test. Significance was set at 5% for all analyses. Statistics were computed using R.

2.8 Human interactome matching analysis

A shortlist of proteins potentially acting as ligands of the druggable human proteome was produced based on the preceding results. The shortlist included proteins with potential toxin function, such as CRISPs (Cysteine-rich secretory proteins), neuropeptides, hormones and toxins retrieved from GO analysis. Secreted or cytosolic proteins were deemed priority, as opposed to membrane- or receptor-bound proteins. Broad-action enzymes such a zinc-dependent metalloproteinases (with the exception of peptidases M12A and M12B because of their relevance in venom signatures) and chitinases were removed from the list due to their low specificity. The shortlist was then hand-curated in order to include proteins that might be lost during filtering. Proteins able to interfere with the human proteome were then retrieved by contrasting against The Human Reference Protein Interactome (HuRi) platform (Luck et al., 2020), available at http://www.interactome-atlas.org/. Analysis was preceded by interconversion of worm genes into human homologs using the R package biomaRt v2.48.2 (Durinck et al., 2005; Durinck et al., 2009) and through BLAST searching against the Swiss-Prot database when necessary (Camacho et al., 2009). The list of potential human targets and the list of their potential interactors were then combined and analysed through the Database for Annotation, Visualization and Integrated Discovery (DAVID) v.6.8 for GO enrichment analysis (Huang et al., 2009a; Huang et al., 2009b). The significance threshold was set at FDR-adjusted $p < 0.05$ for the analyses.

3

RESULTS

3.1 Transcriptome assembly

Glycera alba and *Hediste diversicolor* assembled transcriptomes yielded 482 007 and 192 397 transcripts, respectively (Table 3.1 and refer to Appendix Table A.3 for the quality assessment of the transcriptome assembly). Of these, 100 340 and 88 796 showed potential open reading frames (ORFs), where 50 967 and 58 179 presented homology-matching against proteins from Swiss-Prot, respectively. Deep comparison of transcripts levels unveiled 3 075 transcripts differentially-expressed ($| \log_2 FC | > 1.5$, FDR-adjusted $p < 0.05$) between *G. alba*'s proboscis and skin and 3 122 between *H. diversicolor*'s glands and proboscis (Figures 3.1 and A.1, Table 3.1), which represented 0.64% and 1.62% of the transcripts, respectively (Figure 3.2). Altogether, assembled transcriptomes of *G. alba* and *H. diversicolor* yielded 976 and 1 893 ORFs annotated with differential expression, respectively (Table 3.1). Specifically, 566 ORFs were overexpressed and 410 underexpressed in *G. alba*'s proboscis relative to the skin, while, in *H. diversicolor*, 787 were overexpressed and 1 106 underexpressed in glands compared to the proboscis.

Table 3.1. Selection of the transcripts of interest after *Glycera alba* and *Hediste diversicolor* whole-transcriptome assembly. The table presents the number of sequentially-shortlisted transcripts after each stage of analysis.

Analytical stage	Stage objective	*Glycera alba*	*Hediste diversicolor*
1	Transcriptome Assembly	482 007	192 397
2	Transcripts with Coding Regions	100 340	88 976
3	Functional Annotation	50 967	58 179
4	Differentially-Expressed Transcripts	3 075	3 122
5	ORFs with Differential Expression	976	1 893

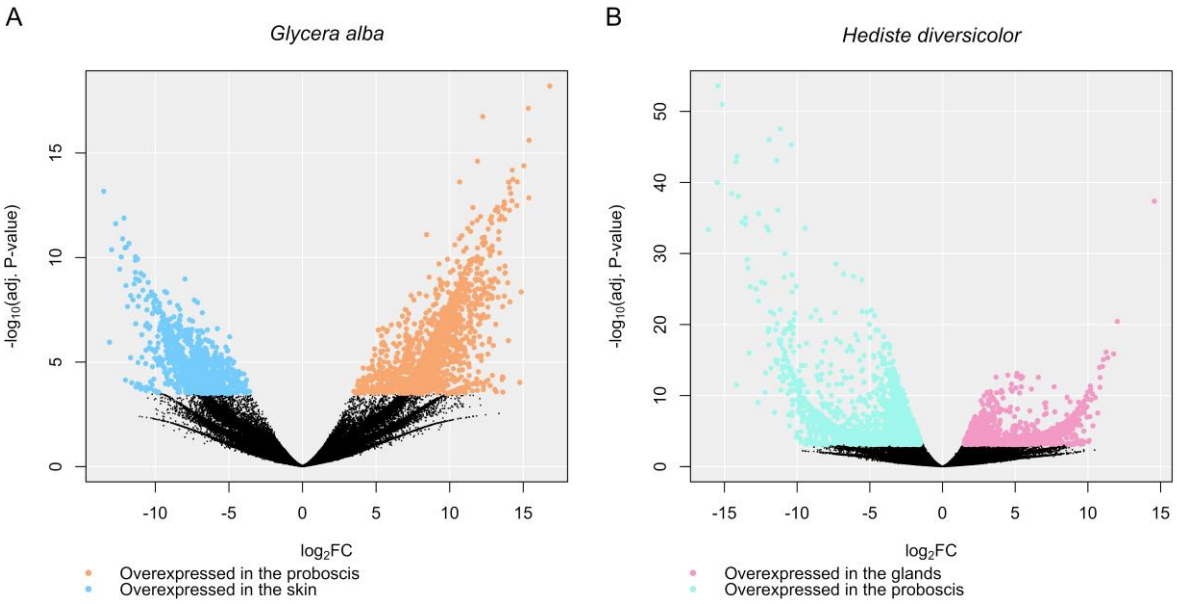

Figure 3.1. Volcano plot illustrating differentially-expressed transcripts. A) between *Glycera alba*'s proboscis and skin. B) between *Hediste diversicolor*'s glands and proboscis. The black dots represent transcripts that are not differentially-expressed between organs. The cut-off for differential expression was set at | log_2FC | > 1.5 and FDR-adjusted $p < 0.05$.

As expected, the differentially-expressed genes between the two *G. alba* organs were clearly separated according to their levels of expression, i.e., to the organ that they were overexpressed (Figure 3.3A and A.2A). This division could be related with different secretory functions. Moreover, the proboscis appeared to have more potential toxins and diffusing and permeabilising agents (flagged proteins) through the whole cluster (Figure 3.3A). On the other hand, the division according to the levels of expression was not so straightforward in *H. diversicolor* (Figure 3.3B and A.2B). Nevertheless, flagged proteins appeared to be more concentrated in three clusters: two containing proteins up-regulated in the glands and the other from the proboscis (Figure 3.3B).

Due to being absent from the Swiss-Prot database, the paradigmatic toxins glycerotoxin and hedistin (of genera *Glycera* and *Hediste*, respectively), were annotated by direct matching against records in UniProt (Accession: A0A1U9VX91, A0A1U9VX95, A0A1U9VX98 and Q1PG44). Nineteen transcripts of *G. alba* rendered significant homology against glycerotoxin (*e*-value <1E-131), all of which were overexpressed in the proboscis (log_2FC > 7). With the addition of these proteins, the total number of annotated proteins up-regulated in the proboscis rose to 576 (Figure 3.2A). Conversely, only twelve predicted coding regions could be matched against hedistin (*e*-value <1E-5), none of which were differentially-expressed between *H. diversicolor*'s organs under comparison.

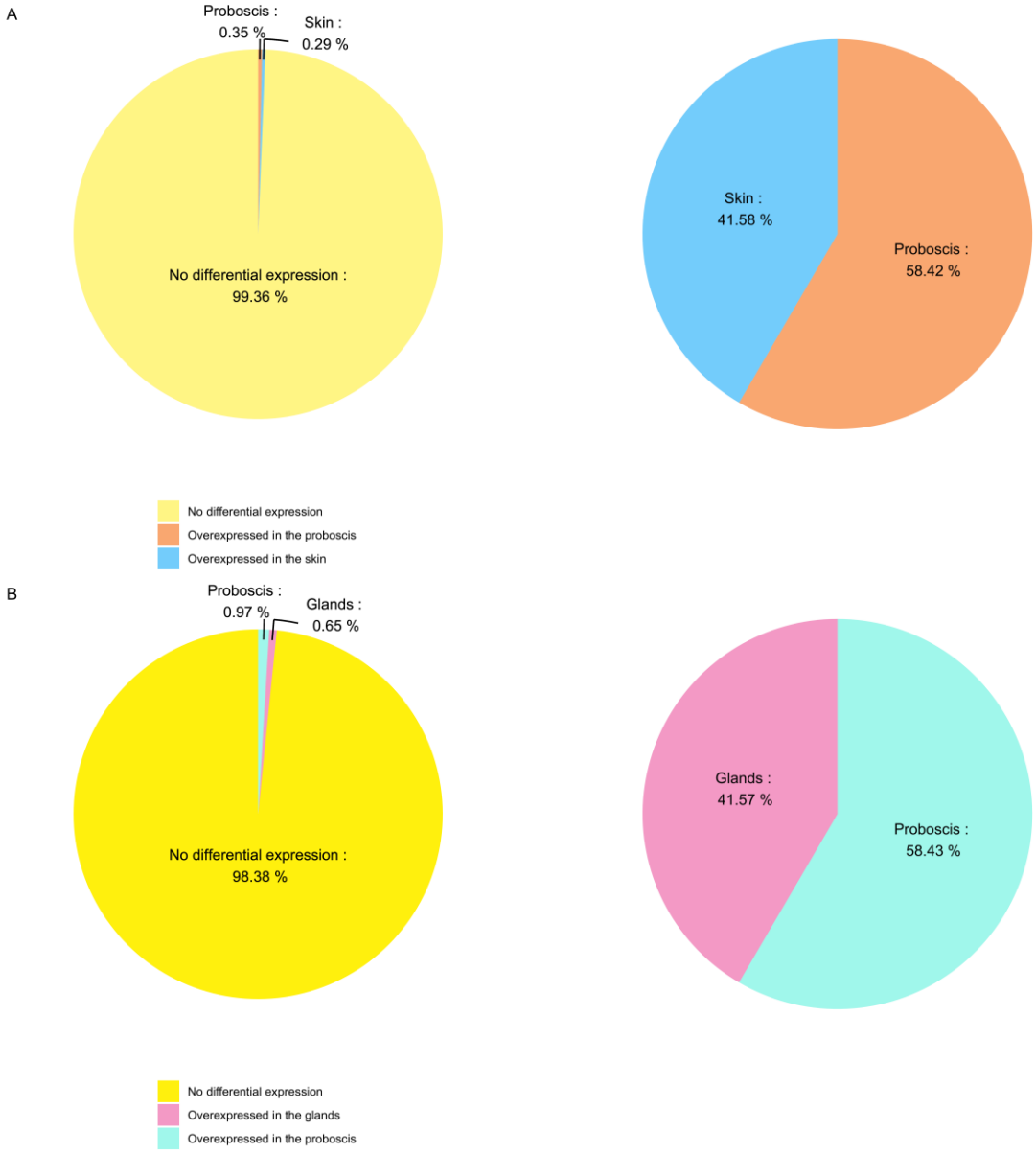

Figure 3.2. Relative proportion of differentially-expressed genes. A) between *Glycera alba*'s proboscis and skin. B) between *Hediste diversicolor*'s glands and proboscis. Significance thresholds were set at $|\log_2 FC| > 1.5$ and FDR-adjusted $p < 0.05$. The pie charts on the right illustrate the relative proportion of differentially-expressed genes that encode for protein with homology-matching against the Swiss-Prot and after the identification of the characteristic toxin of *Glycera* (glycerotoxin).

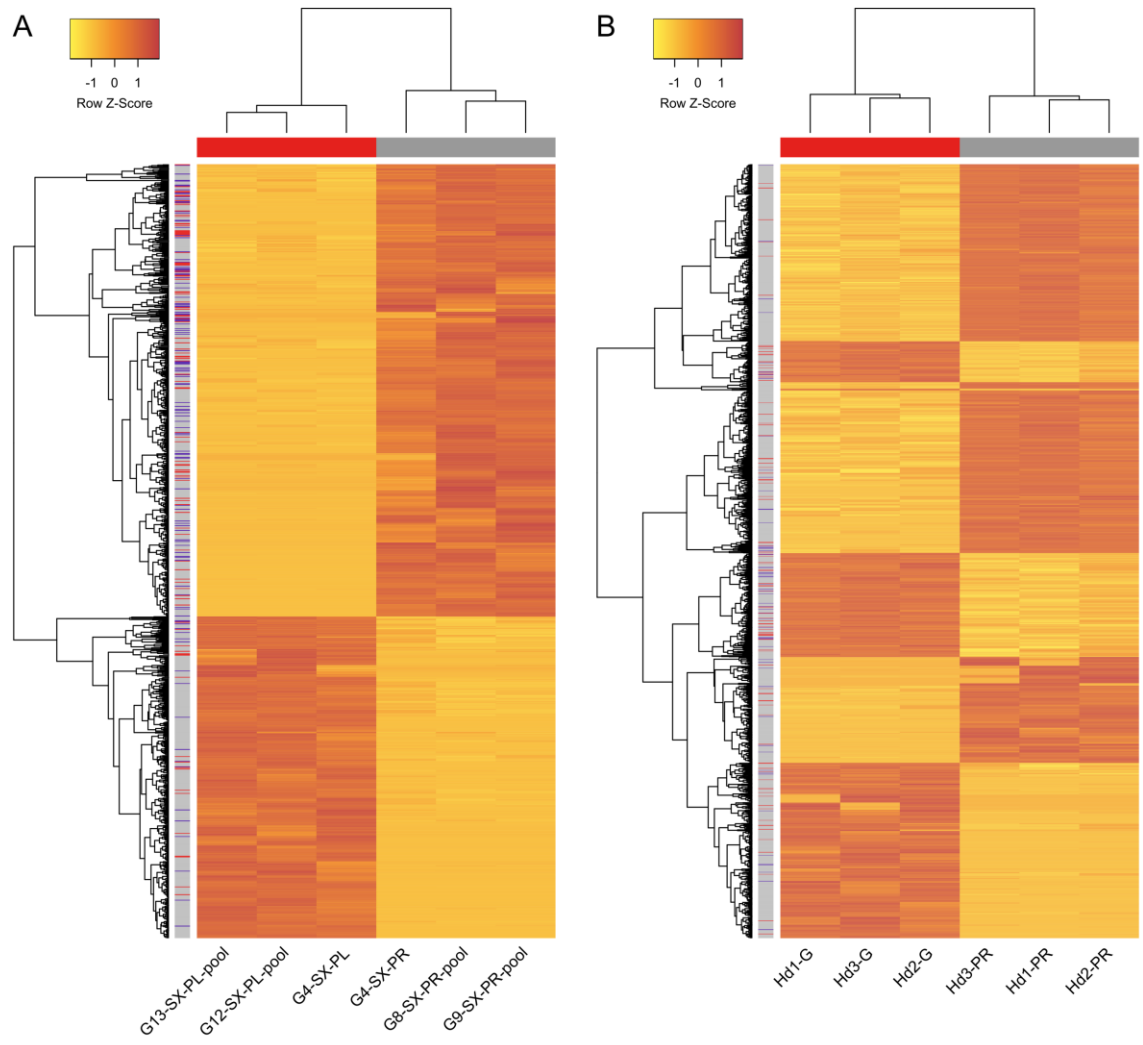

Figure 3.3. Heatmaps illustrating relative gene expression of the differentially-expressed genes. A) between *Glycera alba*'s proboscis (PR) and skin (PL). B) between *Hediste diversicolor*'s glands (G) and proboscis (PR). A | $\log_2$FC | > 1.5 and an FDR adjusted $p < 0.05$ were the cut-offs set for differential expression. The horizontal dendrogram illustrates the association between the three independent replicates for each organ, whereas the vertical dendrogram represents the association between the proteins. The metric and function of the cluster analysis are Euclidian distances and complete linkage, respectively. The side bar near the vertical dendrogram indicates the coding regions with homology-matching with potential toxins (blue) and permeabilizing and diffusing agents (red).

3.2 Identification of potential venom components in the transcriptomes of *Glycera alba* and *Hediste diversicolor*

The best annotated ORFs are provided in Appendix Tables A.4 and A.5. In *G. alba*, the ten most differentially-expressed genes in the proboscis included cysteine-rich venom proteins (CRVP) (Substrate-specific endoprotease Tex31) similar to those of *Conus textile*, glycerotoxin and zinc-dependent metalloproteinases, together with a ovoinhibitor-like protein, which inhibits the activity of potassium channels (Table 3.2). The majority of these top DEGs (differentially-expressed genes) were

secreted/extracellular. In turn, the ten most overexpressed genes in *G. alba*'s skin compared to the proboscis encoded proteins with homology-matching to proteins with various functions related with basal metabolism (Table A.6). While the ten higher differentially-expressed genes in *H. diversicolor*'s glands were mostly intracellular and linked with cell growth and metabolism, in the proboscis, the majority were secreted/extracellular and were directly or indirectly related to the extracellular matrix (Tables A.7 and A.8).

Gene ontology (GO) indicated that about 9% of genes overexpressed in *G. alba*'s proboscis compared to the skin were related to proteolysis (Figure 3.4A). Despite 1.04% of DEGs in this organ being associated with toxin activity, the GO Term based on the Molecular Function yielded a marginal corrected *p*-value of 0.085 (Figure 3.5). Nevertheless, other GO Terms connected with toxins and other venoms signatures components were enriched in the proboscis relative to the skin of this Polychaeta, namely, metallo(endo)peptidase activity, zinc ion binding, serine-type endopeptidase activity, extracellular region and positive regulation of apoptotic process. Although calcium binding was the GO Term based on Molecular Function most frequent in both *G. alba* organs, it was significantly enriched in the skin (Figure 3.4A, 3.4B and 3.5). Similar to the proboscis of *G. alba*, the proteolysis was also the GO term based on Biological Process most frequent in DEGs of *H. diversicolor*'s glands (Figure 3.4C). On the other hand, around 6% of DEGs in the proboscis were related with sarcomere organization (Figure 3.4D). In both organs of *H. diversicolor*, the percentage of DEGs annotated with toxin activity rose approximately 0.1%. In *H. diversicolor*'s glands, extracellular space, metallo(endo)peptidase activity, zinc ion binding and serine-type endopeptidase activity were significant enriched, while the proboscis had an enrichment of GO Terms connected with muscle (Figure 3.5).

Despite the reduced significance of the enrichment in both species, the GO Terms "neuropeptide signalling pathway", "hormone activity" and "neurohypophyseal hormone activity" also allowed the identification of neuropeptides and hormones (Figure 3.5). For instance, a protein similar to IDLSRF-like peptide was up-regulated in the proboscis of *G. alba*, while two neuropeptide variants and a hormone were found in the skin. In *H. diversicolor*'s glands, apart from the protein homologous to thyrostimulin beta-5 subunit, other four had homology-matching to neuropeptides or proteins with hormone activity, while in the proboscis, the number rose a total of four. Through manual curation, another neuropeptide similar to the prohormone 4 precursor was also found up-regulated in *G. alba*'s proboscis and in *H. diversicolor*'s glands.

Table 3.2. Top10 overexpressed genes in *Glycera alba*'s proboscis relative to the skin. The cut-offs established were $log_2FC > 1.5$ and an FDR-adjusted $p < 0.05$.

log_2FC	log_2CPM	FDR*p*	Protein	Accession	%ID	*e*-value	Organism
14.595	6.903	2.37E-14	Cysteine-rich venom protein (CRVP) (Substrate-specific endoprotease Tex31)	Q7YT83	35.567	1.59E-22	*Conus textile*
14.566	6.875	3.27E-13	Pancreatic triacylglycerol lipase (Fragment)	Q64425	34.921	1.75E-47	*Myocastor coypus*
14.093	6.402	1.31E-08	Plasma kallikrein (Plasma prekallikrein) (PKK) [Cleaved into: Plasma kallikrein heavy chain; Plasma kallikrein light chain]	P03952	39.382	3.43E-45	*Homo sapiens*
14.061	6.369	5.51E-13	Tenascin-R (Neural recognition molecule J1-160/180)	Q8BYI9	35.859	8.64E-26	*Mus musculus*
13.950	6.259	1.62E-09	Cysteine-rich venom protein (CRVP) (Substrate-specific endoprotease Tex31)	Q7YT83	37.008	6.51E-39	*Conus textile*
13.759	6.068	5.63E-10	Tenascin-R (Restrictin)	Q05546	36.869	2.78E-26	*Rattus norvegicus*
13.756	6.065	2.58E-10	Prolow-density lipoprotein receptor-related protein 1 [Cleaved into: Low-density lipoprotein receptor-related protein 1 85 kDa subunit (LRP-85); Low-density lipoprotein receptor-related protein 1 515 kDa subunit (LRP-515); Low-density lipoprotein receptor-related protein 1 intracellular domain (LRPICD)]	G3V928	34.848	2.02E-46	*Rattus norvegicus*
13.717	6.027	7.02E-09	Glycerotoxin (Fragment)	A0A1U9VX95	31.017	2.37E-172	*Glycera tridactyla*
13.613	5.923	2.73E-04	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	P10184	29.216	1.34E-30	*Gallus gallus*
13.595	5.905	2.24E-08	A disintegrin and metalloproteinase with thrombospondin motifs 7 (ADAM-TS 7)	Q1EHB3	30.000	1.50E-19	*Rattus norvegicus*

log_2FC – log_2 fold change; log_2CPM – Average log_2 counts per million; FDR*p* – False discovery rate adjusted *p*-value; %ID – Percentage of identity

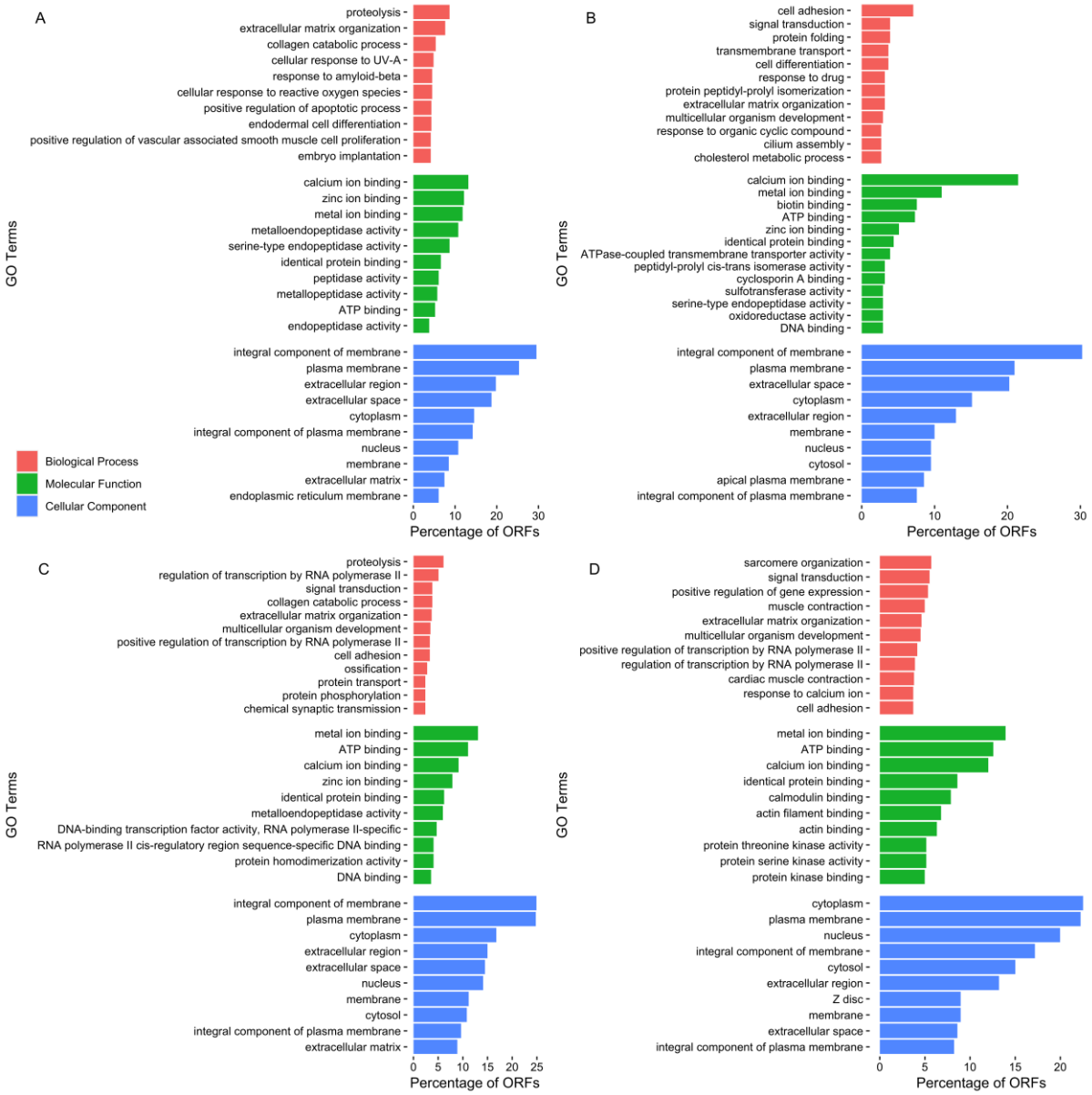

Figure 3.4. Top10 GO Terms of the differentially-expressed genes. The horizontal bar chart shows the percentage of the proteins (ORFs) up-regulated. A) in *Glycera alba*'s proboscis. B) in *Glycera alba*'s skin. C) in *Hediste diversicolor*'s glands. D) in *Hediste diversicolor*'s proboscis related to each GO Term. The same DEG can bear multiple GO Terms.

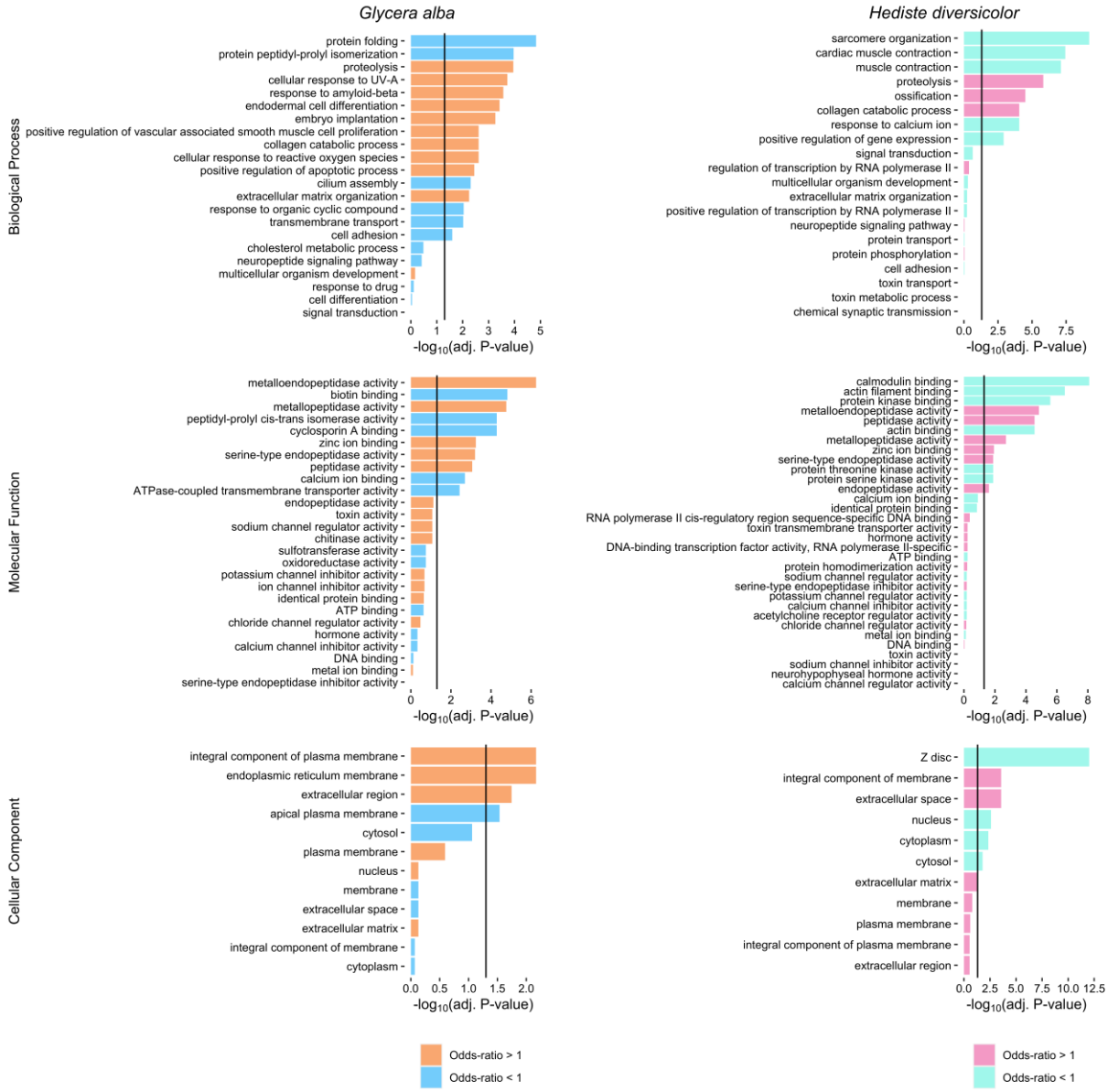

Figure 3.5. Fisher analysis of Top10 GO Terms and GO Terms of interest. in *Glycera alba* and *Hediste diversicolor*'s assembled transcriptomes. The odds-ratio higher than 1 correspond to a GO Terms enrichment in *Glycera alba*'s proboscis (orange) and in *Hediste diversicolor*'s glands (pink). On the other hand, the odds-ratio lower than 1 correspond to an enrichment in the skin of *G. alba* (dark blue) and in the proboscis of *H. diversicolor* (light blue). The horizontal bar chart shows FDR-adjusted *p*. Vertical lines represent the significance cut-off.

Altogether, toxins and other venom signature components such as zinc-dependent metalloproteinases, peptidase S1 (serine proteases), chitinase, C-type lectin, phospholipase A2 and B, serpin, sphingomyelinase and calcium-activated chloride channel regulator were flagged and shortlisted as proteins of interest. To these were added proteins with Kazal or Kunitz_BPTI domains according to the annotation in the UniProt or with the GO Term "serine-type endopeptidase inhibitors activity" or whose description included the terms "venom" or "toxin". Approximately 31% of predicted coding regions overexpressed in the proboscis of *G. alba* relative to the skin fulfilled these conditions (Figure 3.6A). Most of these proteins had homology-matching to peptidases (S1 and M10A), CRISPs, glycerotoxin and

proteins with astacin, reprolysin or Kazal domains. Twelve of the thirteen CRISPs up-regulated in this organ are cysteine-rich venom proteins. Both the astacin and reprolysin domains belong to zinc-dependent metalloproteinase that could have a toxin activity, more specifically, to the peptidase M12A and M12B, respectively. In *G. alba*'s skin, the proteins of interest accounted for less than 10% of the total number of proteins up-regulated (Figure 3.6B). Similar to the proboscis, peptidase S1 was the protein family with a higher number of sequences.

In both *H. diversicolor* organs, the percentage of proteins that might be linked to toxins and other venom components according to settings was lower than 15% of the total annotated DEGs (Figure 3.7). In the glands, the zinc-dependent metalloproteinases rose approximately 61% of the proteins of interest (Figure 3.7A). Peptidases S1, proteins with C-type lectins and protease inhibitors, such as, serpin and proteins with Kazal domains were also up-regulated in this organ. In turn, protein with the Kazal and Kunitz_BPTI domains as well as peptidase M10A and CRISPs were up-regulated in the proboscis of *H. diversicolor* (Figure 3.7B).

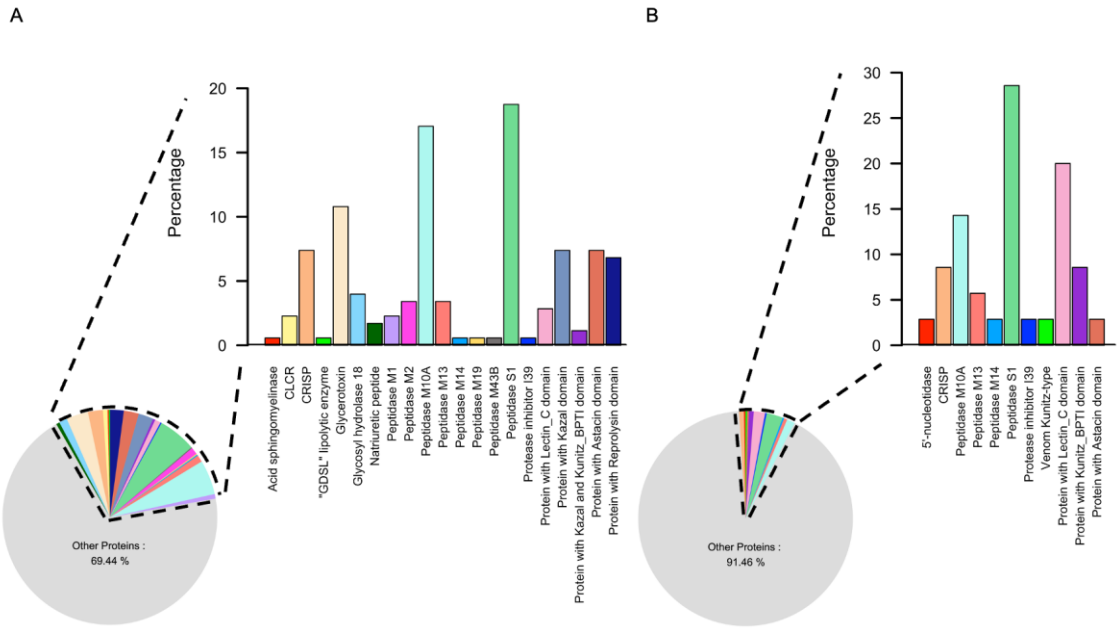

Figure 3.6. Proteins of interest encoded by *Glycera alba*'s assembled transcriptome. A) Proteins up-regulated in the proboscis. B) Proteins up-regulated in the skin. Pie and bar chart illustrate the frequency of the proteins of interest in the assembled transcriptome and inside the protein group, respectively. One protein up-regulated in the skin, that belongs to the protease inhibitor 39 (alpha-2-macroglobulin) family, has a Kazal domain. Proteins described by conserved domains (Lectin_C, Kazal, Kunitz_BPTI, Astacin, Reprolysin and Peptidase_M60 domains) were identified by matching against PROSITE.

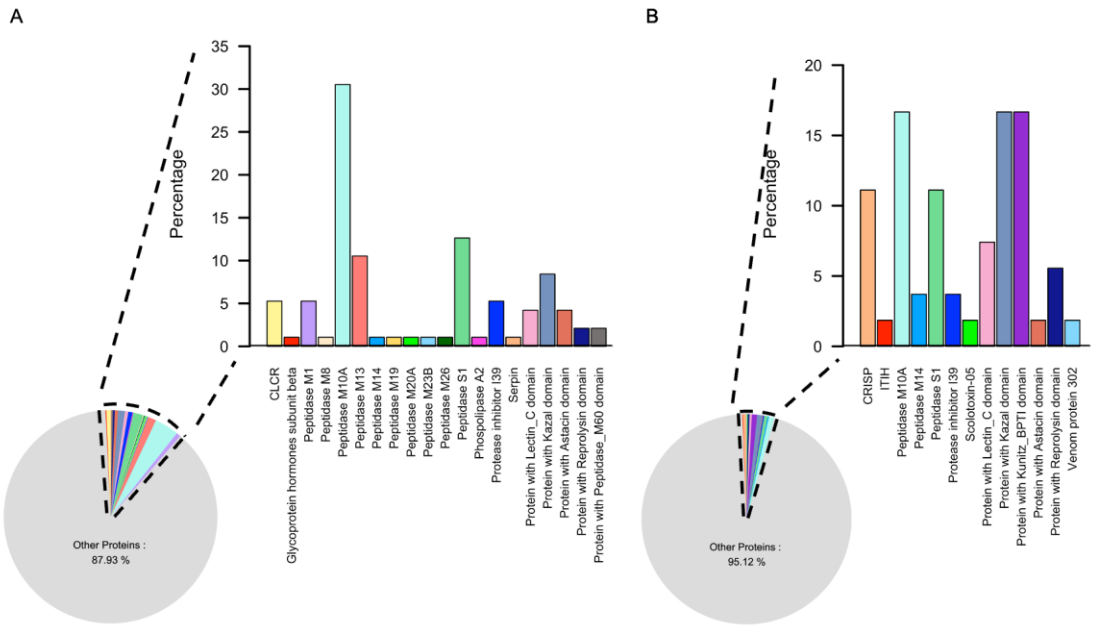

Figure 3.7. Proteins of interest encoded by *Hediste diversicolor*'s assembled transcriptome. A) Proteins up-regulated in the glands. B) Proteins up-regulated in the proboscis. Pie and bar chart illustrate the frequency of the proteins of interest in the assembled transcriptome and inside the protein group, respectively. One protein up-regulated in both organs, that belongs to the protease inhibitor 39 (alpha-2-macroglobulin) family, has a Kazal domain. Proteins described by conserved domains (Lectin_C, Kazal, Kunitz_BPTI, Astacin, Reprolysin and Peptidase_M60 domains) were identified by matching against PROSITE.

Glycera alba and *Hediste diversicolor* differentially-expressed genes encoded proteins that yielded 6 601 and 26 262 domains, respectively. More specifically, 2 992 and 3 609 conserved domains were found in predicted translated transcripts up-regulated in *G. alba*'s proboscis and skin, respectively. On the other hand, the predicted translated transcripts up-regulated in *H. diversicolor*'s glands and proboscis had 9 175 and 17 087 conserved domains, respectively. From all the conserved domains found, only 5 635 and 24 746 were in proteins encoded by DEGs in *G. alba* and *H. diversicolor*'s assembled transcriptomes that were annotated (Figure A.3). The Kazal domains (Kazal_1 and Kazal_2) were two of the ten domains with more hits against predicted translated transcripts up-regulated in *G. alba*'s proboscis, whilst the most frequent conserved domains in *H. diversicolor*'s proboscis were related with immunoglobulins (I-set, Ig_3, V-set, ig, Ig_2) (Figure 3.8). The proboscis of *G. alba* relative to the skin had a significant enrichment in conserved domains that might be present in toxins and other venoms components, such as CAP, ShK, Tox-GHH, astacin, reprolysin (reprolysin, reprolysin _2, reprolysin_3, reprolysin_4 and reprolysin_5), trypsin and CUB (Figure 3.9A and 3.9C). In turn, the enrichment in conserved domains that might be linked to toxins and other venom components was divided between both organs of *H. diversicolor*. While the glands had an enrichment in Kazal (Kazal_1 and Kazal_2), Lectin_C, and peptidase M13 (peptidase_M13 and peptidase_M13_N) domains, the proboscis had in Kunitz_BPTI conserved domain (Figure 3.9B and 3.9D).

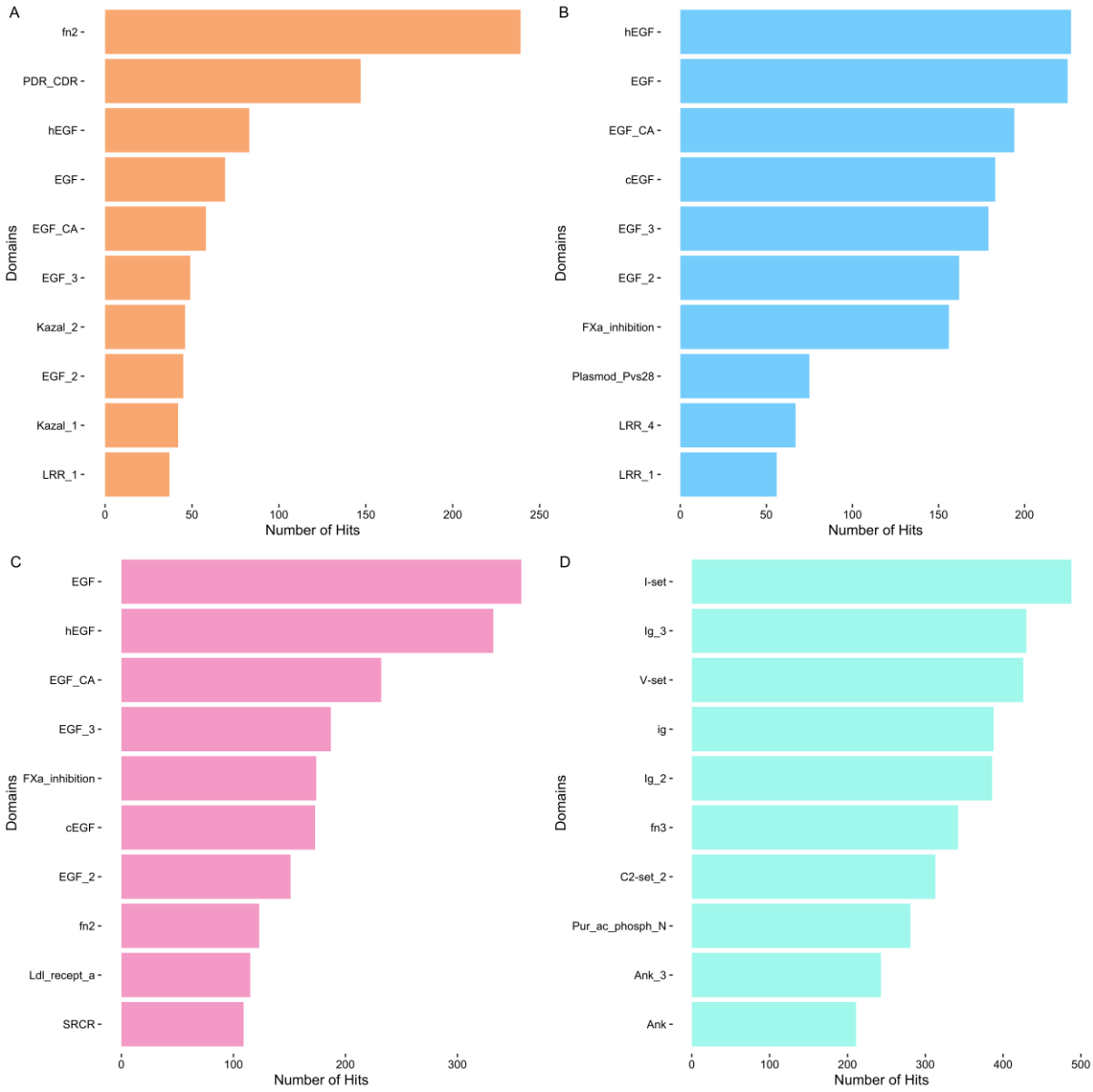

Figure 3.8. Top10 conserved domains in predicted translated transcripts encoded by differentially-expressed genes. The horizontal bar chart indicates the number of hits of domains against Pfam in the predicted translated transcripts up-regulated A) in *Glycera alba*'s proboscis. B) in *Glycera alba*'s skin. C) in *Hediste diversicolor*'s glands. D) in *Hediste diversicolor*'s proboscis. The same predicted translated transcript can bear multiple domains.

From all the 576 proteins up-regulated in *G. alba*'s proboscis relative to the skin, only ovoinhibitor homologous to four different variants was described to have an allergenic propriety (Table A.9). Moreover, these biomolecules also have biotechnological applications along with other four proteins (Table A.10). Only one protein up-regulated in *G. alba*'s skin and *H. diversicolor*'s glands had homology-matching to a protein with biotechnological applications (Table A.11 and A.12). Moreover, an additional protein up-regulated in the glands was similar to a protein which has already a pharmaceutical use, as it is sold under the name Infuse (Table A.13). Apart from the venom allergen 5, four other proteins up-

regulated in *H.* diversicolor's proboscis have allergenic proprieties (Table A.14). Two other proteins might have biotechnological applications (Table A.15).

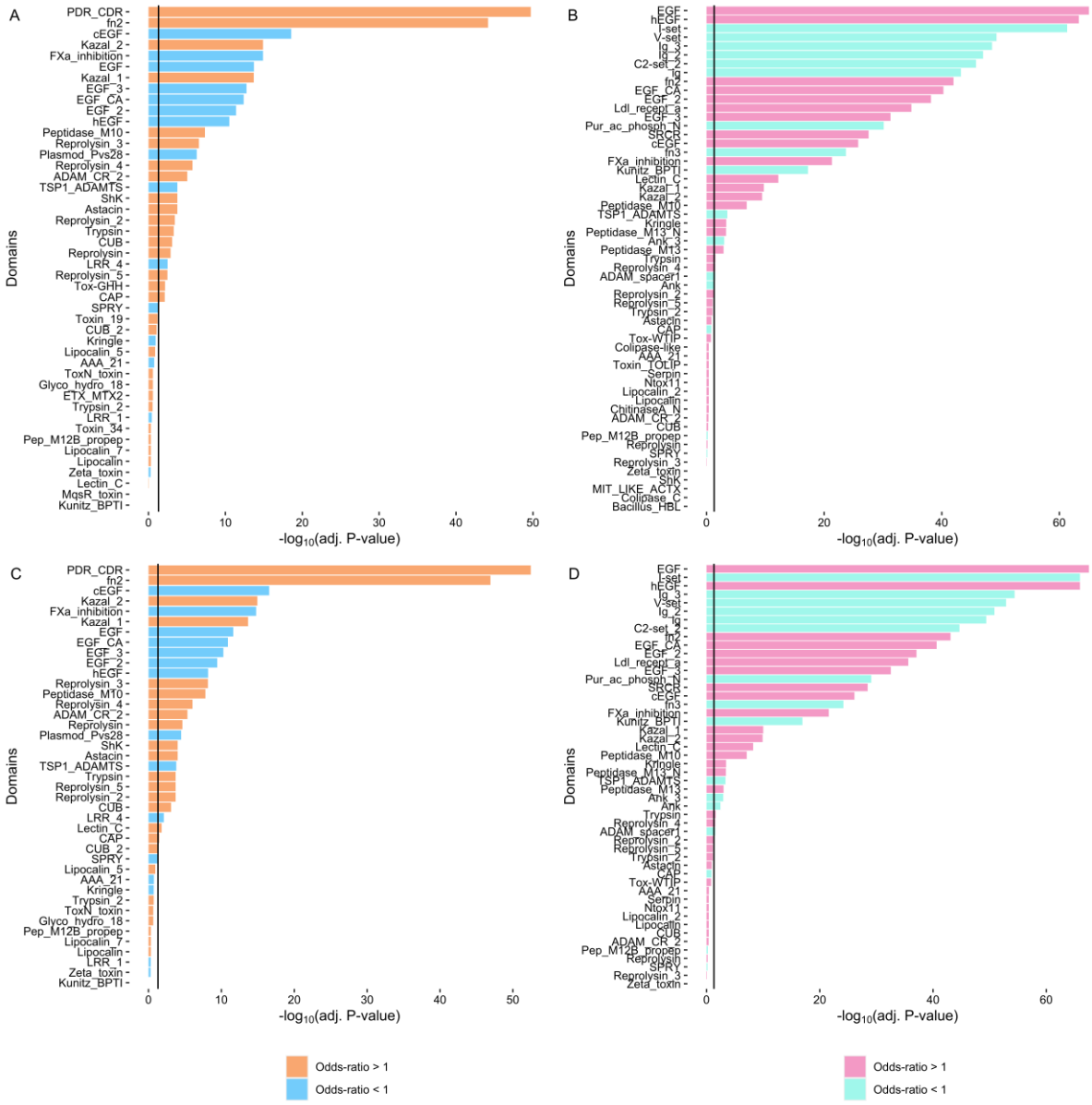

Figure 3.9. Fisher analysis of Top10 domains and domains of interest. A) and C) in *Glycera alba*'s assembled transcriptomes. B) and D) in *Hediste diversicolor*'s assembled transcriptomes. The horizontal bar chart shows the FDR-adjusted *p*. In A) and B), the analyses were performed with all the predicted translated transcripts encoded by differentially-expressed genes, whereas in C) and D), were only performed with the proteins that were annotated. Vertical lines represent the significance cut-off.

3.3 Validation of novel transcripts

For the validation of transcriptome assembly and quantification, the sequences of a few selected mRNAs of interest were isolated by PCR and their expression verified by RT-qPCR (Table A.2, Figure

A.4 and A.5). The main selection criterion was the potential to encode a protein with toxin function, followed by the statistics of the differential expression analysis and functional annotation. Therefore, the genes that encoded a cysteine-rich venom protein and ovoinhibitor-like protein that were overexpressed in *G. alba*'s proboscis compared to the skin were chosen. On the other hand, the genes selected from *H. diversicolor* encoded a thyrostimulin beta-5 subunit and a pathogenesis-related protein (CRISP). These genes were overexpressed and underexpressed in the glands compared to the proboscis, respectively. All four sequences were successful amplified and sequenced. Moreover, two polymorphisms were detected in the isolated transcript that encoded a cysteine-rich venom protein and one polymorphism was found in the ovoinhibitor-like sequence.

The relative expression of the gene that encoded a pathogenesis-related protein (CRISP) from *H. diversicolor* and a cysteine-rich venom protein from *G. alba* were significantly different between the organs, similarly to the expression pattern obtained from RNA-Seq (Figure 3.10). Although the relative expression of gene that encoded the ovoinhibitor-like protein was not found to be significantly different, likely due to relatively high error and low *n*, the trend is similar to RNA-Seq-based quantification, i.e., a higher expression in the proboscis relative to the skin. The relative expression of the gene that encodes thyrostimulin beta-5 subunit was not either found to be significantly different between the *H. diversicolor*'s organs. This might be due to the $\log_2$FC value, which was 2.33. Unless the housekeeping gene was beta-actin, the relative expression of the genes was not significantly different, except for the pathogenesis-related protein (CRISP). The isolated sequences were submitted to the GenBank under the accession number: OL606744-OL606747.

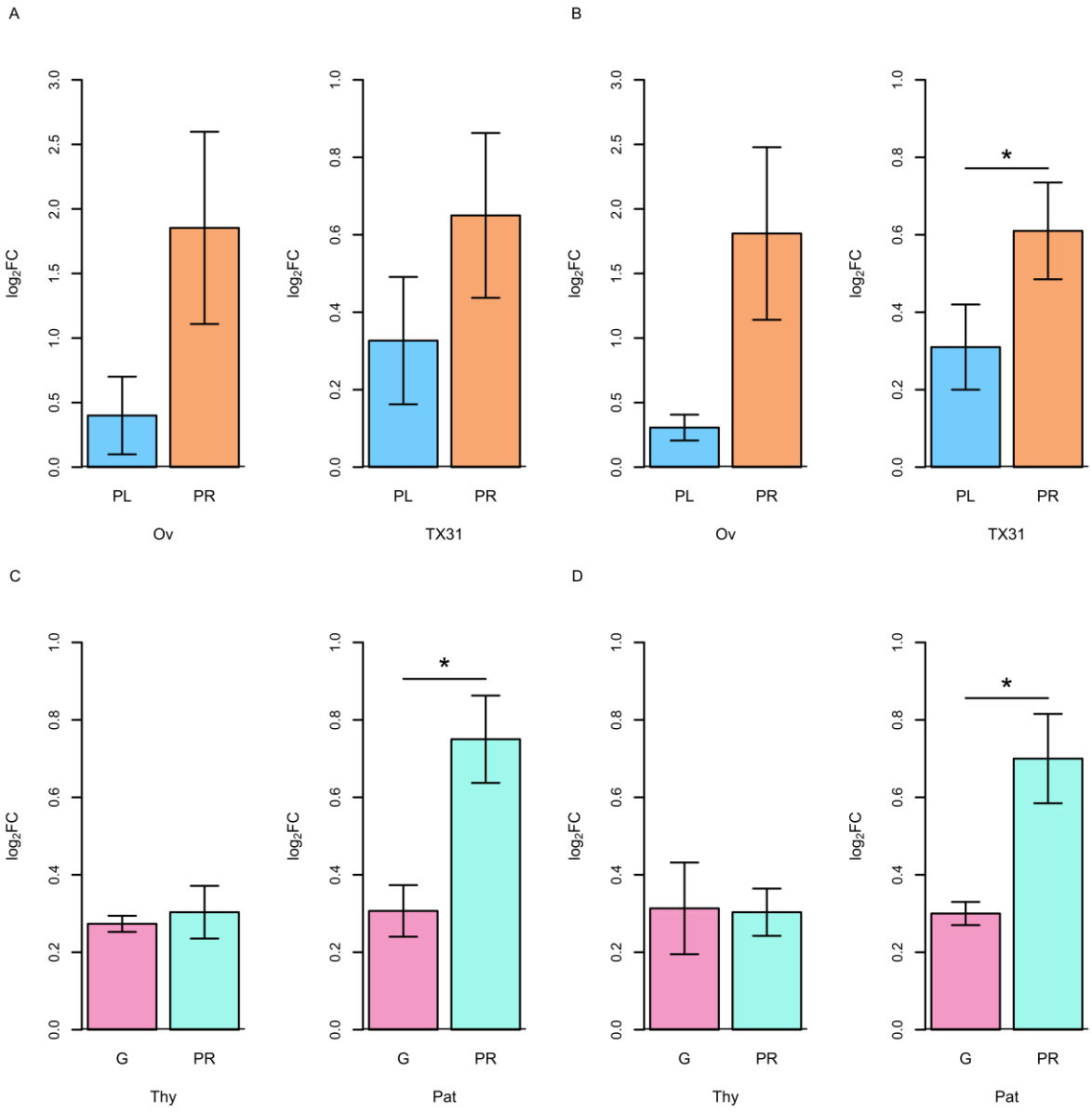

Figure 3.10. Relative expression of selected genes in *Glycera alba* and *Hediste diversicolor*. The expressed sequence tags (EST) detected by RT-qPCR were in mRNAs that encoded for an ovoinhibitor-like protein (Ov), a cysteine-rich venom protein (TX31), a pathogenesis-related protein (CRISP) (Pat) and thyrostimulin beta-5 subunit (Thy). Relative levels of expression of Ov and TX31 were compared between *Glycera alba*'s proboscis (PR) and skin (PL) (A and B), while, for Thy and Pat, the comparison was between *Hediste* diversicolor's glands (G) and proboscis (PR) (C and D). In A) and C), the levels of expression were normalised with the house-keeping gene 18S, while in B) and D) were with beta-actin. *indicates significant differences between organs (Student's t-test, $p < 0.05$)

3.4 Interactome analysis of potential venom compounds

The set of overexpressed genes in *G. alba*'s proboscis (compared to the skin) encoding proteins of interest comprised a total of 201 sequences. Of these, only twelve human proteins that were homologous to 24 Polychaeta proteins were present in the HuRI databases and, consequently, could be introduced into the platform (Table A.16). One variant of glycerotoxin had fibrillin-1 as its potential

human homolog, which likely results from the presence of EGF domains in either protein. The human homologs of the proteins of interest from Polychaeta will henceforth be termed as interactors. In turn, from an initial set of 107 proteins up-regulated in the skin, only nine interactors fulfilled the selection criteria and could be submitted (Table A.17). In *H. diversicolor*'s glands, the initial shortlist of 189 up-regulated proteins was reduced to 26 interactors after filtering through the restrictions of the HuRI datasets (Table A.18). On the other hand, 32 interactors from the proboscis of *H. diversicolor* could be analysed for protein-protein interactions, from an initial set of 197 sequences (Table A.19). The HuRI platform then allowed the identification of the protein-protein interactions between the human proteome (here forth referred to as targets) and the interactors (Figure 3.11 and 3.12). The potential interactors from *G. alba*'s proboscis and skin could bind with 112 and 45 proteins, respectively, whereas a total of 247 and 323 targets were discovered for the interactors from *H. diversicolor*'s glands and proboscis, respectively.

More than 50% of the 29 biological processes enriched within the group constituted by potential targets and interactors from *G. alba*'s proboscis were related with cell death, specially, apoptosis (Table A.20). Indeed, the majority of the ten most significantly enriched biological processes were linked with apoptosis and the apoptotic signalling pathway (Figure 3.13A). None of the biological process that were enriched in *G. alba*'s proboscis had the human homolog of glycerotoxin involved. Less than 10 biological processes were found enriched in the group joined by the potential targets and interactors from the skin of *G. alba* (Figure 3.13B and Table A.21). A total of 15 and 62 pathways were enriched in the shortlist that combined the targets and interactors from *H. diversicolor*'s glands and proboscis, respectively (Figure 3.13C and Table A.22 and A.23). Although the ten most significantly enriched biological process in *H. diversicolor*'s proboscis were mostly related with gene expression (Figure 3.13D), the regulation of tumour necrosis factor-mediated signalling pathway was also present in this top.

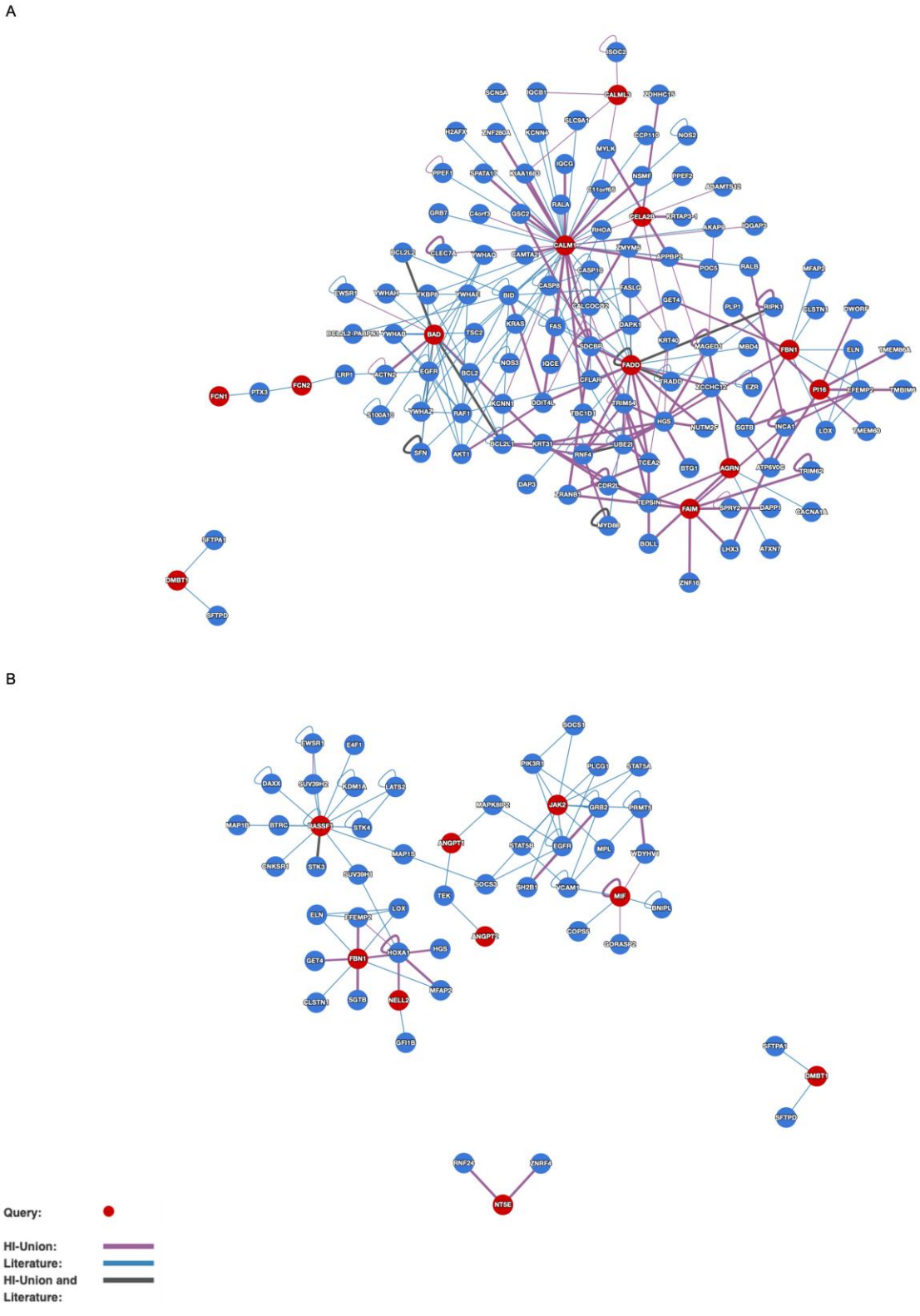

Figure 3.11. Protein-protein interactions between the potential interactors from the proboscis and skin of *Glycera alba* and the human proteome. The interactors were the human homologs of the proteins up-regulated A) in the proboscis. B) in the skin. The interactions were retrieved from the HuRI platform.

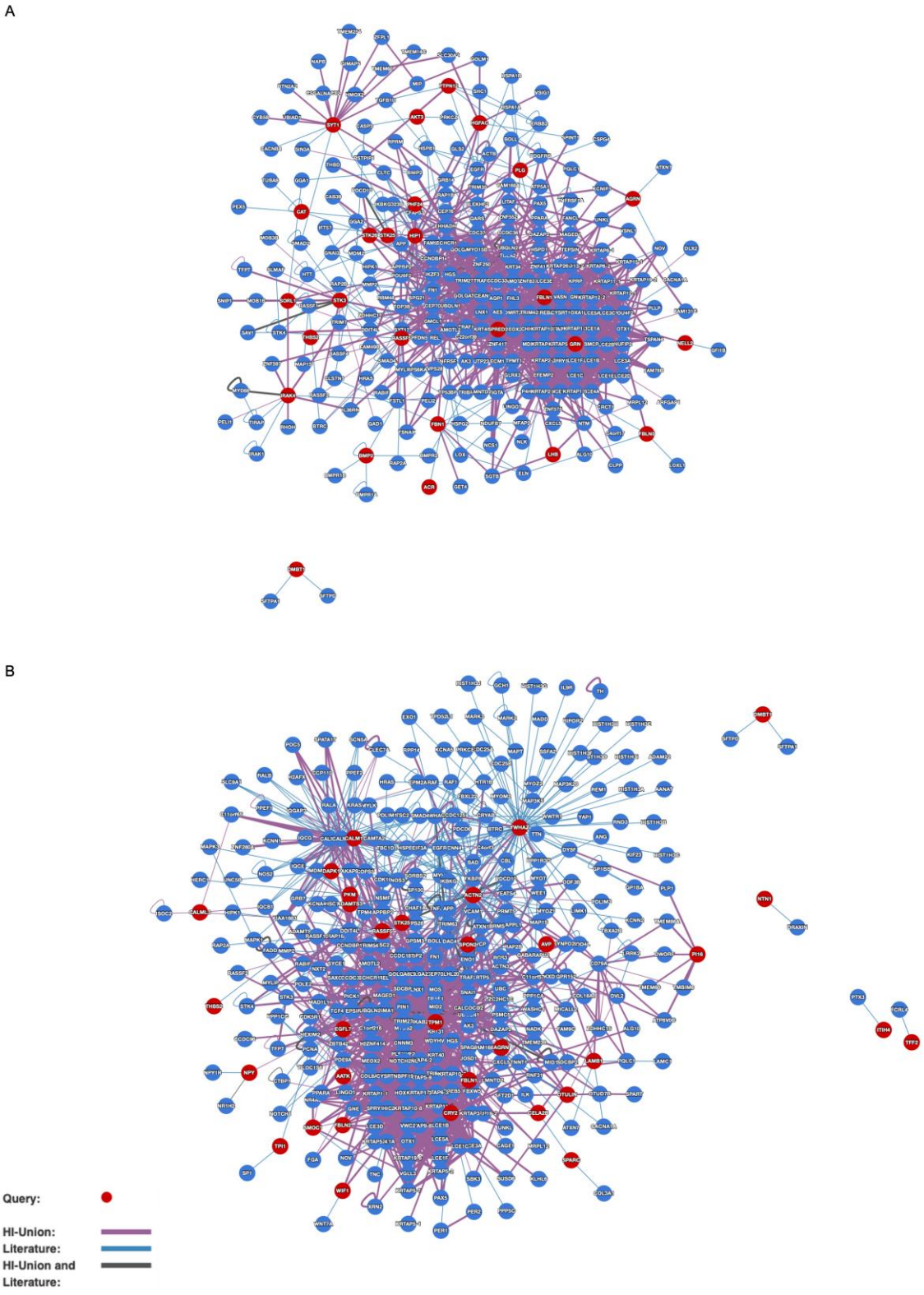

Figure 3.12. Protein-protein interactions between the potential interactors from the glands and proboscis of *Hediste diversicolor* and the human proteome. The interactors were human homologs of the proteins up-regulated A) in the glands. B) in the proboscis. The interactions were retrieved from the HuRI platform.

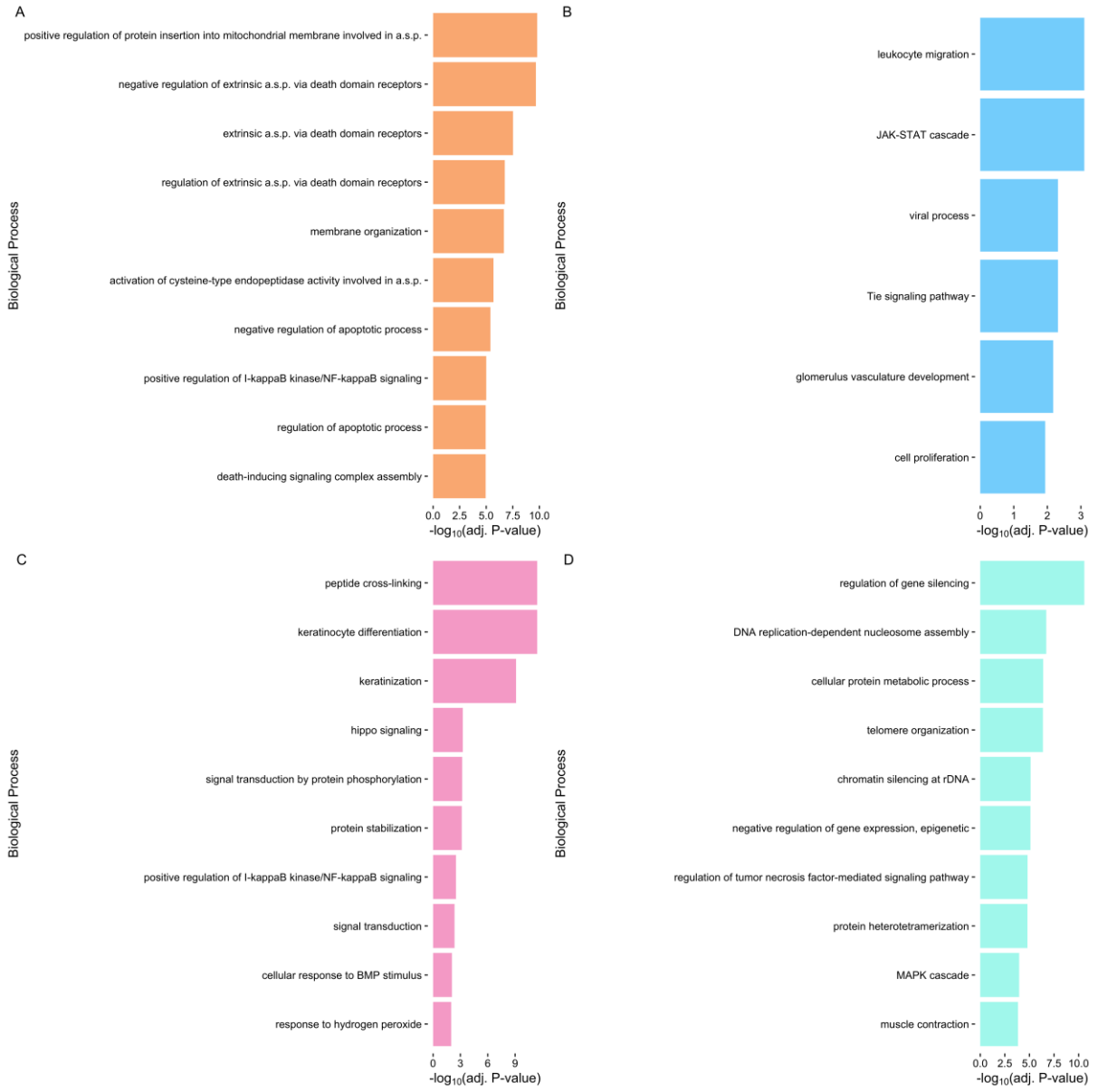

Figure 3.13. GO enrichment of the biological processes in the group combining the potential interactors from Polychaeta and their human targets. The interactors were the human homologs of the proteins up-regulated A) in *Glycera alba*'s proboscis. B) in *G. alba*'s skin. C) in *Hediste diversicolor*'s glands. D) in *H. diversicolor*'s proboscis. The human targets were the human proteins retrieved from the HuRI platform. The horizontal bar chart shows the FDR-adjusted *p*. The GO enrichment analysis was performed in the Database for Annotation, Visualization and Integrated Discovery (DAVID). a.s.p – apoptotic signalling pathway

4

DISCUSSION

Glycera alba and *Hediste diversicolor* are two Polychaeta with distinct behaviours, but suspected to secrete toxins as part of their predatorial or defensive behaviours, which provides them with high potential for biotechnological applications, especially drug discovery. Comparative transcriptomics indeed confirmed that both species secrete proteinaceous toxins and other bioreactives for feeding or defence, however, the results also indicate distinct biochemical nature and function of the substances.

The toxins and other bioreactives are secreted by distinct organs in either species, which most likely relates to the different ecological functions of the proteins. While *G. alba* injects proteinaceous toxins through the proboscis to overpower its prey, *H. diversicolor* might secrete noxious biomolecules through the skin for defensive purposes directed towards potential predators, parasites or even competitors. The expression of these substances could be a measure of protection as *H. diversicolor* is more exposed to pathogens and predators when foraging for food, whilst *G. alba* is a burrower that ambushes its prey. The current transcriptomic profiling of *G. alba*, together with a previous work (Gonçalves & Costa, 2020) in which the presence of glandular tissue capable of secreting toxins was identified due to the presence of thiol-rich compounds (a signature of animal venoms), supports the proboscis housing the venom production delivery apparatus in this species. These findings are in accordance with the expression of toxins in venom glands that has been described in three species of the same genus (von Reumont et al., 2014b). In turn, the transcriptomic analysis indicates that neither the pharyngeal glands nor the proboscis of *H. diversicolor* appeared to secrete a very significant number of toxins. The transcriptomic result together with the detection of thiol-rich compounds in epidermal cells (e.g. mucocytes) of a member from *H. diversicolor*'s family allow raising the hypothesis that this Polychaeta might secrete toxins through the skin, as proposed by Gonçalves & Costa (2020). The same authors also suggested that the noxious substances might not be proteinaceous toxins, but instead secondary metabolites with defence purposes. Therefore, the transcriptome of *H. diversicolor*'s skin should be analysed and could indicate enzymatic networks that might facilitate the choice of the dedicated untargeted metabolomics approaches to identify secondary metabolites of unknown biochemical nature.

The distinct venom apparatus and predatory behaviours can also be indicative of differential specificity of toxins towards molecular and cellular targets. Since *G. alba* injects toxins and other venom components to overwhelm its prey, this could suggest that these toxins might be more specific towards their targets, whereas *H. diversicolor*'s bioreactives may have a broader action as they might be release to the water through the skin, as a protective measure against potential predators or parasites. Similarly, *Bonellia viridis*, another Polychaeta, also secretes bioreactives, such as the chlorin pigment Bonellin, through the integument as a defence against pathogens and predators (de Nicola Giudici, 1984). In turn, the results suggest that *G. alba*'s neurotoxins are mainly secreted by the proboscis and their delivery is facilitated by a series of permeabilising agents, such as broad-subtract metalloproteinases, since these proteins digest the extracellular matrix thus facilitating the infiltration of toxins. There are, however, other venom components that could potentially facilitate the spreading of toxins through the prey, such as proteins that affect the blood either by altering the blood pressure or by having anti-coagulant activity. Indeed, while *G. alba* has a venom directed towards the predation and immobilization of the prey, *H. diversicolor* appears to secrete inhibitors, toxins and other proteins that can provoke inflammatory and allergenic responses. Unlike the diffusing and permeabilizing agents that are secreted by the proboscis of *G. alba* to facilitate the toxin infiltration and diffusion, the permeabilizing agents in *H. diversicolor* might have a role in assisting the digestion. In fact, amphinomids (also known as fireworms) were reported to also secrete toxins for defence (Verdes et al. 2018). The same authors suggested that while proteases, protease inhibitors and C-type lectins might interfere with blood coagulation and the inflammatory response, proteins with CAP domains could potentially provoke allergenic responses.

The interactome-directed analysis corroborates the distinct transcriptomic profiles of these two marine invertebrates, as the human homologs of the proteins up-regulated in *G. alba*'s proboscis are more linked with cell death, whereas for *H. diversicolor*, a portion of the proteins may potentially interfere with a wide range of processes, from cell death to the immune response. Nevertheless, both Polychaeta species appeared to be good candidates for biotechnological applications. The interference with apoptosis and its signalling pathway inferred from human homologs of the proteins up-regulated in *G. alba*'s proboscis indicates potential for biomedical applications, despite difficulties resulting from reduced genomic resources for Polychaeta and other marine invertebrates in general. In fact, several of the worms' proteins could not be analysed regarding protein-protein interaction with the druggable human proteome as the tools and databases are set for human proteins instead of the non-model organism. Particular attention is given to *G. alba* neurotoxins and their predicted interaction with neuronal ion channels. In their natural environment, the neurotoxins should result in the immobilization of the prey, but their interference with the nervous system can indicate novel bioproducts for drug discovery, as they may hold potential for managing and treating neurological diseases. See, for instance, Shen et al. (2000) that discussed the potential of various conopeptides with neurotoxin activity and secreted by *Conus* (cone snails) in the treatment of diseases from the nervous system. In contrast, the bioreactives from *H. diversicolor* could reach the markets as antibiotics or biocides, since they are

homologs to proteins with antimicrobial proprieties as they might inhibit proteases essential to invasion of pathogens.

The CRISPs secreted by *G. alba* might provoke prey paralysis, i.e., be neurotoxins (cysteine-rich venom proteins), whereas the CRISPs expressed by *H. diversicolor* could have a more defensive role. The neurotoxins similar to cysteine-rich venom proteins in *G. alba*'s proboscis could have a neuromuscular toxic effect identical to other venoms and toxungens of several species of marine invertebrates (from Gastropoda to Polychaeta), where these proteins with CAP domains block potassium and calcium channels and, consequently, muscle contraction is prevented (see for instance, von Reumont et al., 2014b; Modica et al., 2015; Verdes et al., 2018; Rodrigo et al., 2021). The proboscis of *G. alba* also secretes other neurotoxins, namely, glycerotoxin. While cysteine-rich venom protein might block ion channels, glycerotoxin interacts with N-type calcium channels and provokes long-lasting spontaneous release of neurotransmitters as well as the depletion of synaptic vesicles. Despite being characteristic of the *Glycera* genus, glycerotoxin was not found in the transcriptomic analysis of other three species (*G. dibranchiata, G. tridactyla and G. fallax*) performed by von Reumont et al. (2014b). Even though at the time of publication of the von Reumont et al. (2014b) work, the sequences had not been deposited in public databases, if the analysis were repeated today, the results would probably be identical (Richter et al., 2017). This would happen since von Reumont et al. (2014b) used a subset of the UniProt that was gathered by only considering the proteins that were secreted and the glycerotoxin does not have yet any annotation concerning the GO based on the Cellular Component or in Subcellular location section of database. The reduced annotation can bring challenges when characterizing the transcriptome as it makes difficult to know the function of proteins. It also highlights the need to continue to bioprospect the oceans, specially, marine invertebrates.

The transcriptomic profile of *G. alba* revealed a third potential group of neurotoxins in addition to the cysteine-rich venom protein and glycerotoxin, which were proteins with ShK domains. The conserved domain resembles the neurotoxic peptide from *Stichodactyla helianthus* that is characterized by being able to interfere with the voltage-gated potassium channels (Castañeda et al., 1995). The majority of the transcripts that encoded for proteins characterised by having ShK domain in *G. alba*'s proboscis also have other conserved domains characteristic of zinc-dependent metalloproteinase, for instance, astacin domains. This conjugation also happens in other venoms, such as, in *Colubraria reticulata*, *G. tridactyla* and *G. dibranchiata* and could originate enzymes capable of modulating the activity of ion channels (Rangaraju et al., 2010; von Reumont et al., 2014b; Modica et al., 2015). The proteins' potential for interacting with ion channels makes them strong candidates for further research as possible therapeutic drugs. However, the interactome-directed analysis of these molecules toward the druggable human proteome to identify potential therapeutical targets did not produce results as human homologs of these proteins were not present in the HuRI platform, which again highlights the pitfalls of reduced genomic annotation of non-conventional model species, especially for bioprospecting.

While the paradigmatic toxin of *Glycera* is a neurotoxin termed glycerotoxin, hedistin is a peptide exclusively secreted by *H. diversicolor* that has antimicrobial properties. This biomolecule was not found overexpressed in any of *H. diversicolor*'s organs, as expected, as it is secreted from the natural killer cells-like (the type 3 granulocytes), which are present in the coelomic fluid (Tasiemski et al., 2007). In turn, the hemerythrin (MPII) (non-metallothionein cadmium-binding protein) is another antimicrobial peptide paradigmatic of this Polychaeta. Despite being released from the type 1 granulocytes upon bacterial challenge (Deloffre et al., 2003), Salzet-Raveillon et al. (1993) found that some muscular tissue can also express the peptide, which might explain its up-regulation in the proboscis.

Glycera alba's proboscis and both surveyed organs of *H. diversicolor* have proteins with Kazal domains. Some serine protease inhibitors bearing Kazal domains have already been described in various animal poisons, venoms and toxungenous (e.g. *Glycera*, the phyllodocid Polychaeta *Eulalia*, fireworms and the honeybee *Apis cerana*), where they possess antimicrobial or anticoagulation activity (Kim et al., 2013; von Reumont et al., 2014b; Verdes et al., 2018; Rodrigo et al., 2021). A venom signature present in both Polychaeta transcriptomes are the transcripts that encode for peptidases M12A and M12B, which are enzymes with astacin and reprolysin domains, respectively. Apart from digestion of the extracellular matrix, some authors also suggested that these peptidases could have a pro-haemorrhagic activity and cause inflammation in numerous venoms (Trevisan-Silva et al., 2010; von Reumont et al., 2014b; Modica et al., 2015; Verdes et al., 2018; Rodrigo et al., 2021). The transcriptomic analysis of *H. diversicolor* also reveals proteins related with venoms that can also inhibit the blood coagulation and influence inflammation, see for instance, phospholipase A2 and serpin upregulated in the glands and proteins with the Kunitz_BPTI domain in the proboscis. Some authors raised the hypothesis that the phospholipase A2 can also induce cytotoxicity and neurotoxicity, while serpin could also influence the immune defence (von Reumont et al., 2014b; Verdes et al., 2018). Apart from the interference with the blood coagulation and inflammation due to a serine protease inhibitor activity, other proteins with Kunitz_BPTI domains have already been reported to inhibit ion channels, namely, potassium (Imredy & MacKinnon, 2000; Tsujimoto et al., 2012). The interactome-directed analysis revealed an enrichment in pathways involved with inflammatory responses, thus suggesting that the proteins of *H. diversicolor* may indeed interfere with specific targets of the human immune system. As an example, the IRAK4 (Interleukin-1 receptor-associated kinase 4) is recruited upon the activation of the toll-like receptors and leads to a cascade of reactions that culminates with the activation of a transcription factor involved in the regulation of the expression of genes related to the cell survival, immune and inflammatory response (Li et al., 2002).

Unlike those of *Glycera*, which are likely associated to neurotoxic action, the CRISPs up-regulated in the proboscis of *H. diversicolor* appear to be linked to a defensive role and with the immune system. Nevertheless, a protein similar to cysteine-rich venom protein Venom allergen 5 (Antigen 5) was up-regulated in *H. diversicolor*'s proboscis. This CRISP is present in *Polybia paulista*'s venom and despite the unknown function, it is involved in an allergenic response (dos Santos et al., 2010; dos Santos-Pinto

et al., 2014). The transcriptome profile of the proboscis also indicates other proteins belonging to the CRISP family, namely, pathogenesis-related and glioma pathogenesis-related-like proteins. The pathogenesis-related protein 1C is secreted by *Nicotiana tabacum* upon infection by pathogens, while the glioma pathogenesis-related proteins were first isolated in glioblastoma multiforme/astrocytoma and their high level of homology to the pathogenesis-related protein raised the hypothesis that they could have an equivalent role and, consequently, be involved in the immune response (Pfitzner & Goodman, 1987; Murphy et al., 1995; Szyperski et al., 1998).

Besides the aforementioned toxins, other potential proteins of interest were also found in the secretomes of the worms that may target the druggable human proteome. It is the case, for instance, of the ovoinhibitor-like protein and thyrostimulin beta-5 subunit that are up-regulated in *G. alba*'s proboscis and *H. diversicolor*'s glands, respectively. Despite its still unknown function, the ovoinhibitor protein has multiple Kazal domains and is characterized by the ability to inhibit the large conductance calcium-activated potassium channels from the bovine aortic smooth muscle cells (Moss et al., 1996). This protein was also described to have antimicrobial activity as it can inhibit serine proteases essential for microorganisms to invade the host (Bourin et al., 2011). Since the ovoinhibitor in *Gallus gallus* can interfere with ion channels and is secreted, the similar Polychaeta protein might be a novel toxin targeting these channels, which may indicate potential biotechnological applications, e.g., directed towards the neuromuscular response. In turn, thyrostimulin beta-5 subunit binds non-covalently and confers the hormone activity to the alpha-2 subunit of thyrostimulin, an ancient glycoprotein hormone that has also been reported in cone snails and in the gastropod *Aplysia* (Heyland et al., 2012; Robinson et al., 2017). In humans, due to the ability to interact with thyrotropin G protein-coupled receptor, the corresponding homolog hormone could influence the thyroid cell metabolism, whereas in *Aplysia* it could have neuronal modulatory activity (Nakabayashi et al., 2002; Heyland et al., 2012). Recent studies raise the hypothesis that the monomers of the thyrostimulin alone or as a homodimer could also have an activity since the subunits have different levels of expression (Dos Santos et al., 2009; Heyland et al., 2012), which was also verified in *H. diversicolor*.

Pathway enrichment in biological processes related with the regulation of apoptosis highlighted a few overexpressed genes in *G. alba*'s proboscis that might hold particular interest for biomedical applications. Among these, it must be emphasised the group of genes coding for proteins that are homologous to the human BAD (Bcl2-associated agonist of cell death), FADD (FAS-associated death domain protein) and FAIM (Fas apoptotic inhibitory molecule 1). In the first two cases, the respective annelid homologs were chiefly annotated through the identification of conserved domains. However, the genes homologous to *BAD* (BCL2 associated agonist of cell death) and *FADD* (Fas associated via death domain) presented relatively low levels of expression in either organ under analysis, i.e., the proboscis and the skin (log_2CPM = -0.595 and 0.180), which can mean further challenges to isolate mRNAs and resulting proteins from the worm. The FAIM protein is particularly noteworthy since it is an inhibitor of FAS-mediated apoptosis, whose underexpression has been associated with Alzheimer's disease

(Schneider et al., 1999; Carriba et al., 2015). In turn, *H.diversicolor* also yielded genes that encode for proteins able to interfere with programmed cell death, see for instance, the human proteins OTULIN (Ubiquitin thioesterase otulin) and DAPK1 (Death-associated protein kinase 1). The interactome-directed analysis also flagged the human protein STK3 (Serine/threonine-protein kinase 3) that is connected with the pathway involved in the regulation of organ size through the regulation of cellular proliferation and apoptosis in mammals and in *Drosophila* (Dong et al., 2007). Additionally, similar to Polychaeta genes homologous to *BAD* and *FADD*, the identification of domains unravelled several transcripts encoding proteins without homologs in the public hand-curated database and could be indicative of novel proteins, namely, toxins. Several of these predicted translated transcripts up-regulated in the proboscis of *G. alba* had conserved domains related with toxins, such as CAP, ETX_MTX2, Tox-GHH, Toxin_19 and Toxin_34 domains. This difference is simultaneously another indicator of the reduced annotation of the marine invertebrate and the potential for discovering novel bioreactives that could have potential biotechnological applications in drug discovery.

5

CONCLUSION

The transcriptomic analysis of *G. alba* and *H. diversicolor* not only allowed the identification of multiple proteins of interest for potential biotechnological applications in both Polychaeta, but it also highlighted the tremendous diversity of bioreactives to be bioprespected from marine invertebrates, with emphasis on proteins and peptides, which may offer advantageous synthesis. *Glycera alba* secretes proteinaceous neurotoxins and other venoms components through the proboscis, while for *H. diversicolor* has been hypothesized to secrete non-proteinaceous toxins through the skin. The toxins secreted by these two marine invertebrates might be linked to their feeding strategy. While *G. alba* is a burrower that ambushes its prey and secretes toxins to overpower them, *H. diversicolor* might secrete toxins for defence and repellent as it is more exposed to pathogens and predators when searching for food on the soft-bottom marine surface. Indeed, *G. alba*'s venom appears to have toxins that mostly targets the nervous system of the prey, whereas *H. diversicolor* has some inhibitors and peptides linked to immune defence (e.g. antimicrobial) as well as others that provoke inflammatory and allergenic reactions. Moreover, the transcriptomic analysis together with the interactome-directed analysis indicates that the proteins from *G. alba* might be used as drugs for managing or treating disease from the nervous system (e.g. neurodegenerative), whilst the bioreactives from *H. diversicolor* would be applied to modulate immune defence and inflammatory response and ultimately, be used as antibiotics and biocides. Therefore, the current study unveils that combining functional annotation through Gene Ontology, homology search and identification of domains allows to discover novel proteins with potential for biotechnological application. Furthermore, the interactome-directed analysis provides possible targets in the druggable human proteome of the biomolecules based on their human homologs. Most importantly, it must be highlighted that "omics", particularly transcriptomics, allowed to circumvent problems related to reduced or even absent genomic annotation for marine annelids, as for most invertebrates in general. Despite potential biases arising, for instance, from the interconversion of the worms' secretome into human homologs while searching for potential interactors with the druggable human proteome, the analyses generated a plausible map of potential molecules of interest. The next stage of research with any of such proteins of interest will involve obtaining laboratory evidence for protein-protein interactions, assessing toxicity effects and mechanism and optimising production via

DNA recombinant technology using adequate prokaryote or eukaryote models. Assessing the influence of polymorphisms of the sequences already isolated for cloning is also important. Due to the biotechnological potential, a cysteine-rich venom protein and an ovoinhibitor-like protein from the proboscis of *G. alba* as well as a pathogenesis-related protein (CRISP) and a thyrostimulin beta-5 subunit from the organs of *H. diversicolor* were isolated.

6

REFERENCES

Bai, F., Morcos, F., Cheng, R. R., Jiang, H., & Onuchic, J. N. (2016). Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis. *Proc Natl Acad Sci U S A*, *113*(50), E8051-E8058. https://doi.org/10.1073/pnas.1615932113

Bon, C., Saliou, B., Thieffry, M., & Manaranche, R. (1985). Partial purification of α-glycerotoxin, a presynaptic neurotoxin from the venom glands of the polychaete annelid *glycera convoluta*. *Neurochem Int*, *7*(1), 63-75. https://doi.org/10.1016/0197-0186(85)90009-9

Bourin, M., Gautron, J., Berges, M., Attucci, S., Le Blay, G., Labas, V., . . . Rehault-Godbert, S. (2011). Antimicrobial potential of egg yolk ovoinhibitor, a multidomain Kazal-like inhibitor of chicken egg. *J Agric Food Chem*, *59*(23), 12368-12374. https://doi.org/10.1021/jf203339t

Bowersox, S. S., Gadbois, T., Singh, T., Pettus, M., Wang, Y. X., & 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*(3), 1243-1249.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, *34*(5), 525-527. https://doi.org/10.1038/nbt.3519

Bryan, G. W., & Gibbs, P. E. (1979). Zinc - a major inorganic component of nereid polychaete jaws. *J Mar Biol Assoc U K*, *59*(4), 969-973. https://doi.org/10.1017/S0025315400036961

Burgess, J. G. (2012). New and emerging analytical techniques for marine biotechnology. *Curr Opin Biotechnol*, *23*(1), 29-33. https://doi.org/10.1016/j.copbio.2011.12.007

Burgess, M. G., Clemence, M., McDermott, G. R., Costello, C., & Gaines, S. D. (2018). Five rules for pragmatic blue growth. *Mar Policy, 87*, 331-339. http://doi.org/10.1016/j.marpol.2016.12.005

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Carriba, P., Jimenez, S., Navarro, V., Moreno-Gonzalez, I., Barneda-Zahonero, B., Moubarak, R. S., . . . Comella, J. X. (2015). Amyloid-β reduces the expression of neuronal FAIM-L, thereby shifting the inflammatory response mediated by TNFα from neuronal protection to death. *Cell Death Dis*, *6*, e1639. https://doi.org/10.1038/cddis.2015.6

Casewell, N. R., Wüster, W., Vonk, F. J., Harrison, R. A., & Fry, B. G. (2013). Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol Evol*, *28*(4), 219-229. https://doi.org/10.1016/j.tree.2012.10.020

Castañeda, O., Sotolongo, V., Amor, A. M., Stöcklin, R., Anderson, A. J., Harvey, A. L., . . . Karlsson, E. (1995). Characterization of a potassium channel toxin from the Caribbean Sea anemone *Stichodactyla helianthus*. *Toxicon*, *33*(5), 603-613. https://doi.org/10.1016/0041-0101(95)00013-c

Ceccarelli, B., & Hurlbut, W. P. (1980). Ca^{2+}-dependent recycling of synaptic vesicles at the frog neuromuscular junction. *J Cell Biol*, *87*(1), 297-303. https://doi.org/10.1083/jcb.87.1.297

Charif, D., & 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 Bastolla, U., Porto, M., Roman, H. E., & Vendruscolo, M. (Eds.), *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (pp. 207-232). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35306-5_10

Chiba, S., Saji, Y., Takeo, Y., Yui, T., & Aramaki, Y. (1967). Nereistoxin and its derivatives, their neuromuscular blocking and convulsive actions. *Jpn J Pharmacol*, *17*(3), 491-492. https://doi.org/10.1254/jjp.17.491

Christinat, A., & Leyvraz, S. (2009). Role of trabectedin in the treatment of soft tissue sarcoma. *Onco Targets Ther*, *2*, 105-113. https://doi.org/10.2147/ott.s4454

Cuevas, C., Pérez, M., Martín, M. J., Chicharro, J. L., Fernández-Rivas, C., Flores, M., . . . Manzanares, I. (2000). Synthesis of ecteinascidin ET-743 and phthalascidin Pt-650 from cyanosafracin B. *Org Lett*, *2*(16), 2545-2548. https://doi.org/10.1021/ol0062502

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

de Nicola Giudici, M. (1984). Defence mechanism of *Bonellia viridis*. *Mar Biol, 78*(3), 271-273. https://doi.org/10.1007/BF00393013

Deloffre, L., Salzet, B., Vieau, D., Andries, J. C., & Salzet, M. (2003). Antibacterial properties of hemerythrin of the sand worm *Nereis diversicolor*. *Neuro Endocrinol Lett*, *24*(1-2), 39-45.

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S. A., . . . Pan, D. (2007). Elucidation of a universal size-control mechanism in *Drosophila* and mammals. *Cell*, *130*(6), 1120-1133. https://doi.org/10.1016/j.cell.2007.07.019

dos Santos, L. D., Santos, K. S., Pinto, J. R. A., Dias, N. B., de Souza, B. M., dos Santos, M. F., . . . Palma, M. S. (2010). Profiling the proteome of the venom from the social wasp *Polybia paulista*: a clue to understand the envenoming mechanism. *J Proteome Res*, *9*(8), 3867-3877. https://doi.org/10.1021/pr1000829

Dos Santos, S., Bardet, C., Bertrand, S., Escriva, H., Habert, D., & Querat, B. (2009). Distinct expression patterns of glycoprotein hormone-α2 and -β5 in a basal chordate suggest independent developmental functions. *Endocrinology*, *150*(8), 3815-3822. https://doi.org/10.1210/en.2008-1743

dos Santos-Pinto, J. R. A., dos Santos, L. D., Andrade Arcuri, H., Castro, F. M., Kalil, J. E., & Palma, M. S. (2014). Using proteomic strategies for sequencing and post-translational modifications assignment of antigen-5, a major allergen from the venom of the social wasp *Polybia paulista*. *J Proteome Res*, *13*(2), 855-865. https://doi.org/10.1021/pr4008927

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*(16), 3439-3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, *4*(8), 1184-1191. https://doi.org/10.1038/nprot.2009.97

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., & Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res*, *49*(W1), W216-W227. https://doi.org/10.1093/nar/gkab225

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform*, *23*(1), 205-211.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third ed.). Sage, Thousand Oaks, CA, USA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Freitas, A. C., Rodrigues, D., Rocha-Santos, T. A. P., Gomes, A. M. P., & Duarte, A. C. (2012). Marine biotechnology advances towards applications in new functional foods. *Biotechnol Adv*, *30*(6), 1506-1515. https://doi.org/10.1016/j.biotechadv.2012.03.006

Fry, B. G., Roelants, K., Champagne, D. E., Scheib, H., Tyndall, J. D. A., King, G. F., . . . de la Vega, R. C. R. (2009). The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annu Rev Genomics Hum Genet*, *10*, 483-511. https://doi.org/10.1146/annurev.genom.9.081307.164356

Gibbs, P. E., & Bryan, G. W. (1980). Copper-the Major Metal Component of Glycerid Polychaete Jaws. *J Mar Biol Assoc U K*, *60*(1), 205-214. https://doi.org/10.1017/S0025315400024267

Gomes, A., Correia, A. T., & Nunes, B. (2019). Worms on drugs: ecotoxicological effects of acetylsalicylic acid on the Polychaeta species *Hediste diversicolor* in terms of biochemical and histological alterations. *Environ Sci Pollut Res Int*, *26*(13), 13619-13629. https://doi.org/10.1007/s11356-019-04880-1

Gonçalves, C., & Costa, P. M. (2020). Histochemical detection of free thiols in glandular cells and tissues of different marine Polychaeta. *Histochem Cell Biol*, *154*(3), 315-325. https://doi.org/10.1007/s00418-020-01889-3

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, *29*(7), 644-652. https://doi.org/10.1038/nbt.1883

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

Hamed, S. M., Abd El-Rhman, A. A., Abdel-Raouf, N., & Ibraheem, I. B. M. (2018). Role of marine macroalgae in plant protection & improvement for sustainable agriculture technology. *Beni Suef Univ J Basic Appl Sci, 7*(1), 104-110. https://doi.org/10.1016/j.bjbas.2017.08.002

Harley, M. B. (1950). Occurrence of a filter-feeding mechanism in the polychaete *Nereis diversicolor*. *Nature*, *165*(4201), 734-735. https://doi.org/10.1038/165734b0

Heyland, A., Plachetzki, D., Donelly, E., Gunaratne, D., Bobkova, Y., Jacobson, J., . . . Moroz, L. L. (2012). Distinct expression patterns of glycoprotein hormone subunits in the lophotrochozoan *Aplysia*: implications for the evolution of neuroendocrine systems in animals. *Endocrinology*, *153*(11), 5440-5451. https://doi.org/10.1210/en.2012-1677

Hong, J. (2011). Role of natural product diversity in chemical biology. *Curr Opin Chem Biol*, *15*(3), 350-354. https://doi.org/10.1016/j.cbpa.2011.03.004

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nat Rev Drug Discov*, *1*(9), 727-730. https://doi.org/10.1038/nrd892

Hu, Y., Chen, J., Hu, G., Yu, J., Zhu, X., Lin, Y., . . . Yuan, J. (2015). Statistical research on the bioactivity

of new marine natural products discovered during the 28 years from 1985 to 2012. *Mar Drugs*, *13*(1), 202-221. https://doi.org/10.3390/md13010202

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, *37*(1), 1-13. https://doi.org/10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, *4*(1), 44-57. https://doi.org/10.1038/nprot.2008.211

Hutchings, P. (1998). Biodiversity and functioning of polychaetes in benthic sediments. *Biodivers Conserv*, *7*, 1133-1145. https://doi.org/10.1023/A:1008871430178

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *J Comput Graph Stat*, *5*(3), 299-314. https://doi.org/10.1080/10618600.1996.10474713

Imhoff, J. F., Labes, A., & Wiese, J. (2011). Bio-mining the microbial treasures of the ocean: new natural products. *Biotechnol Adv*, *29*(5), 468-482. https://doi.org/10.1016/j.biotechadv.2011.03.001

Imredy, J. P., & MacKinnon, R. (2000). Energetic and structural interactions between δ-dendrotoxin and a voltage-gated potassium channel. *J Mol Biol,* *296*(5), 1283-1294. https://doi.org/10.1006/jmbi.2000.3522

Kagan, B. L., Pollard, H. B., & Hanna, R. B. (1982). Induction of ion-permeable channels by the venom of the fanged bloodworm *Glycera dibranchiata*. *Toxicon*, *20*(5), 887-893. https://doi.org/10.1016/0041-0101(82)90076-9

Kim, B. Y., Lee, K. S., Zou, F. M., Wan, H., Choi, Y. S., Yoon, H. J., . . . Jin, B. R. (2013). Antimicrobial activity of a honeybee (*Apis cerana*) venom Kazal-type serine protease inhibitor. *Toxicon*, *76*, 110-117. https://doi.org/10.1016/j.toxicon.2013.09.017

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, *35*(3), 421-432. https://doi.org/10.1093/bioinformatics/bty648

Li, S., Strelow, A., Fontana, E. J., & Wesche, H. (2002). IRAK-4: a novel member of the IRAK family with the properties of an IRAK-kinase. *Proc Natl Acad Sci U S A*, *99*(8), 5567-5572. https://doi.org/10.1073/pnas.082100399

Lichtenegger, H. C., Schöberl, T., Bartl, M. H., Waite, H., & Stucky, G. D. (2002). High abrasion resistance with sparse mineralization: copper biomineral in worm jaws. *Science*, *298*(5592), 389-392. https://doi.org/10.1126/science.1075433

Lichtenegger, H. C., Schöberl, T., Ruokolainen, J. T., Cross, J. O., Heald, S. M., Birkedal, H., . . . Stucky, G. D. (2003). Zinc and mechanical prowess in the jaws of *Nereis*, a marine worm. *Proc Natl Acad Sci U S A*, *100*(16), 9144-9149. https://doi.org/10.1073/pnas.1632658100

Lillebø, A. I., Pita, C., Garcia Rodrigues, J., Ramos, S., & Villasante, S. (2017). How can marine ecosystem services support the Blue Growth agenda? *Mar Policy*, *81*, 132-142. http://doi.org/10.1016/j.marpol.2017.03.008

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ Method. *Methods*, *25*(4), 402-408. https://doi.org/10.1006/meth.2001.1262

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., . . . Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature*, *580*(7803), 402-408. https://doi.org/10.1038/s41586-020-2188-x

Martins, A., Vieira, H., Gaspar, H., & Santos, S. (2014). Marketed marine natural products in the pharmaceutical and cosmeceutical industries: tips for success. *Mar Drugs*, *12*(2), 1066-1101. https://doi.org/10.3390/md12021066

Martins, C., Dreij, K., & 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*(23), 4718. https://doi.org/10.3390/ijerph16234718

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*, *40*(10), 4288-4297. https://doi.org/10.1093/nar/gks042

Meunier, F. A., Feng, Z. P., Molgó, J., Zamponi, G. W., & Schiavo, G. (2002). Glycerotoxin from *Glycera convoluta* stimulates neurosecretion by up-regulating N-type Ca^{2+} channel activity. *EMBO J*, *21*(24), 6733-6743. https://doi.org/10.1093/emboj/cdf677

Meunier, F. A., Nguyen, T. H., Colasante, C., Luo, F., Sullivan, R. K. P., Lavidis, N. A., . . . Schiavo, G. (2010). Sustained synaptic-vesicle recycling by bulk endocytosis contributes to the maintenance of high-rate neurotransmitter release stimulated by glycerotoxin. *J Cell Sci*, *123*(7), 1131-1140. https://doi.org/10.1242/jcs.049296

Michel, C., & Keil, B. (1975). Biologically active proteins in the venomous glands of the polychaetous annelid, *Glycera convoluta* Keferstein. *Comp Biochem Physiol B*, *50*(1), 29-33. https://doi.org/10.1016/0305-0491(75)90294-1

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., . . . Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res*, *49*(D1), D412-D419. https://doi.org/10.1093/nar/gkaa913

Modica, M. V., Lombardo, F., Franchini, P., & Oliverio, M. (2015). The venomous cocktail of the vampire snail *Colubraria reticulata* (Mollusca, Gastropoda). *BMC Genomics*, *16*, 441. https://doi.org/10.1186/s12864-015-1648-4

Molinski, T. F., Dalisay, D. S., Lievens, S. L., & Saludes, J. P. (2009). Drug development from marine natural products. *Nat Rev Drug Discov*, *8*(1), 69-85. https://doi.org/10.1038/nrd2487

Montaser, R., & Luesch, H. (2011). Marine natural products: a new wave of drugs? *Future Med Chem*, *3*(12), 1475-1489. https://doi.org/10.4155/fmc.11.118

Moses, D. N., Harreld, J. H., Stucky, G. D., & Waite, J. H. (2006). Melanin and *Glycera* jaws: emerging dark side of a robust biocomposite structure. *J Biol Chem*, *281*(46), 34826-34832. https://doi.org/10.1074/jbc.M603429200

Moss, G. W. J., Marshall, J., Morabito, M., Howe, J. R., & Moczydlowski, E. (1996). An evolutionarily conserved binding site for serine proteinase inhibitors in large conductance calcium-activated potassium channels. *Biochemistry*, *35*(50), 16024-16035. https://doi.org/10.1021/bi961452k

Murphy, E. V., Zhang, Y., Zhu, W., & Biggs, J. (1995). The human glioma pathogenesis-related protein is structurally related to plant pathogenesis-related proteins and its gene is expressed specifically in brain tumors. *Gene*, *159*(1), 131-135. https://doi.org/10.1016/0378-1119(95)00061-a

Murrell, P., & Wen, Z. (2020). *gridGraphics: Redraw Base Graphics Using 'grid' Graphics*. (R package version 0.5-1) https://CRAN.R-project.org/package=gridGraphics

Nakabayashi, K., Matsumi, H., Bhalla, A., Bae, J., Mosselman, S., Hsu, S. Y., & Hsueh, A. J. W. (2002). Thyrostimulin, a heterodimer of two new human glycoprotein hormone subunits, activates the thyroid-stimulating hormone receptor. *J Clin Invest*, *109*(11), 1445-1452. https://doi.org/10.1172/JCI14340

Nelsen, D. R., Nisani, Z., Cooper, A. M., Fox, G. A., Gren, E. C. K., Corbit, A. G., & Hayes, W. K. (2014). Poisons, toxungens, and venoms: redefining and classifying toxic biological secretions and the organisms that employ them. *Biol Rev Camb Philos Soc*, *89*(2), 450-465. https://doi.org/10.1111/brv.12062

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. (R package version 1.1-2) https://CRAN.R-project.org/package=RColorBrewer

Ockelmann, K. W., & Vahl, O. (1970). On the biology of the polychaete *Glycera alba*, especially its burrowing and feeding. *Ophelia*, *8*(1), 275-294. https://doi.org/10.1080/00785326.1970.10429564

Olivera, B. M., Cruz, L. J., de Santos, V., LeCheminant, G. W., Griffin, D., Zeikus, R., . . . Rivier, J. (1987). Neuronal calcium channel antagonists. Discrimination between calcium channel subtypes using ω-conotoxin from *Conus magus* venom. *Biochemistry*, *26*(8), 2086-2090. https://doi.org/10.1021/bi00382a004

Ovchinnikova, T. V., Aleshina, G. M., Balandin, S. V., Krasnosdembskaya, A. D., Markelov, M. L., Frolova, E. I., . . . Kokryakov, V. N. (2004). Purification and primary structure of two isoforms of arenicin, a novel antimicrobial peptide from marine polychaeta *Arenicola marina*. *FEBS Lett*, *577*(1-2), 209-214. https://doi.org/10.1016/j.febslet.2004.10.012

Pfitzner, U. M., & Goodman, H. M. (1987). Isolation and characterization of cDNA clones encoding pathogenesis-related proteins from tobacco mosaic virus infected tobacco plants. *Nucleic Acids Res*, *15*(11), 4449-4465. https://doi.org/10.1093/nar/15.11.4449

Pommier, Y., Kohlhagen, G., Bailly, C., Waring, M., Mazumder, A., & Kohn, K. W. (1996). DNA sequence- and structure-selective alkylation of guanine N2 in the DNA minor groove by ecteinascidin 743, a potent antitumor compound from the Caribbean tunicate *Ecteinascidia turbinata*. *Biochemistry*, *35*(41), 13303-13309. https://doi.org/10.1021/bi960306b

Rangaraju, S., Khoo, K. K., Feng, Z. P., Crossley, G., Nugent, D., Khaytin, I., . . . Chandy, K. G. (2010). Potassium channel modulation by a toxin domain in matrix metalloprotease 23. *J Biol Chem*, *285*(12), 9124-9136. https://doi.org/10.1074/jbc.M109.071266

Richter, S., Helm, C., Meunier, F. A., Hering, L., Campbell, L. I., Drukewitz, S. H., . . . Bleidorn, C. (2017). Comparative analyses of glycerotoxin expression unveil a novel structural organization of the bloodworm venom system. *BMC Evol Biol*, *17*(1), 64. https://doi.org/10.1186/s12862-017-0904-4

Rinehart, K. L., Holt, T. G., Fregeau, N. L., Stroh, J. G., Keifer, P. A., Sun, F., . . . Martin, D. G. (1990). Ecteinascidins 729, 743, 745, 759A, 759B, and 770: potent antitumor agents from the Caribbean tunicate *Ecteinascidia turbinata*. *J Org Chem*, *55*(15), 4512-4515. https://doi.org/10.1021/jo00302a007

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, S. D., Li, Q., Bandyopadhyay, P. K., Gajewiak, J., Yandell, M., Papenfuss, A. T., . . . Safavi-Hemami, H. (2017). Hormone-like peptides in the venoms of marine cone snails. *Gen Comp Endocrinol*, *244*, 11-18. https://doi.org/10.1016/j.ygcen.2015.07.012

Rodrigo, A. P., & Costa, P. M. (2019). The hidden biotechnological potential of marine invertebrates: The Polychaeta case study. *Environ Res*, *173*, 270-280. https://doi.org/10.1016/j.envres.2019.03.048

Rodrigo, A. P., Grosso, A. R., Baptista, P. V., Fernandes, A. R., & Costa, P. M. (2021). A Transcriptomic Approach to the Recruitment of Venom Proteins in a Marine Annelid. *Toxins (Basel)*, *13*(2), 97. https://doi.org/10.3390/toxins13020097

Salzet-Raveillon, B., Rentier-Delrue, F., & Dhainaut, A. (1993). Detection of mRNA encoding an antibacterial-metalloprotein (MPII) by *in situ* hybridization with a cDNA probe generated by polymerase chain reaction in the worm *Nereis diversicolor*. *Cell Mol Biol (Noisy-le-grand)*, *39*(1), 105-114.

Scaps, P. (2002). A review of the biology, ecology and potential use of the common ragworm *Hediste diversicolor* (O.F. Müller) (Annelida: Polychaeta). *Hydrobiologia, 470*(1), 203-218. https://doi.org/10.1023/A:1015681605656

Schenning, M., Proctor, D. T., Ragnarsson, L., Barbier, J., Lavidis, N. A., Molgó, J. J., . . . Meunier, F. A. (2006). Glycerotoxin stimulates neurotransmitter release from N-type Ca^{2+} channel expressing neurons. *J Neurochem*, *98*(3), 894-904. https://doi.org/10.1111/j.1471-4159.2006.03938.x

Schneider, T. J., Fischer, G. M., Donohoe, T. J., Colarusso, T. P., & Rothstein, T. L. (1999). A novel gene coding for a Fas apoptosis inhibitory molecule (FAIM) isolated from inducibly Fas-resistant B lymphocytes. *J Exp Med*, *189*(6), 949-956. https://doi.org/10.1084/jem.189.6.949

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., . . . Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol*, *7*, 3. https://doi.org/10.1186/1471-2199-7-3

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

Smith, M. L. (2002). Cutaneous problems related to coastal and marine worms. *Dermatol Ther*, *15*(1), 34-37. https://doi.org/10.1046/j.1529-8019.2002.01505.x

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

Soudy, M., Anwar, A. M., Ahmed, E. A., Osama, A., Ezzeldin, S., Mahgoub, S., & Magdeldin, S. (2020). UniprotR: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase). *J Proteomics*, *213*, 103613. https://doi.org/10.1016/j.jprot.2019.103613

Stecher, G., Tamura, K., & Kumar, S. (2020). Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol*, *37*(4), 1237-1239. https://doi.org/10.1093/molbev/msz312

Szyperski, T., Fernández, C., Mumenthaler, C., & Wüthrich, K. (1998). Structure comparison of human glioma pathogenesis-related protein GliPR and the plant pathogenesis-related protein P14a indicates a functional link between the human immune system and a plant defense system. *Proc Natl Acad Sci U S A*, *95*(5), 2262-2266. https://doi.org/10.1073/pnas.95.5.2262

Takebayashi, Y., Pourquier, P., Zimonjic, D. B., Nakayama, K., Emmert, S., Ueda, T., . . . Pommier, Y. (2001). Antiproliferative activity of ecteinascidin 743 is dependent upon transcription-coupled nucleotide-excision repair. *Nat Med*, *7*(8), 961-966. https://doi.org/10.1038/91008

Tasiemski, A., Schikorski, D., Le Marrec-Croq, F., Pontoire-Van Camp, C., Boidin-Wichlacz, C., & Sautière, P. E. (2007). Hedistin: A novel antimicrobial peptide containing bromotryptophan constitutively expressed in the NK cells-like of the marine annelid, *Nereis diversicolor*. *Dev Comp Immunol*, *31*(8), 749-762. https://doi.org/10.1016/j.dci.2006.11.003

The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, *49*(D1), D480-D489. https://doi.org/10.1093/nar/gkaa1100

Trevisan-Silva, D., Gremski, L. H., Chaim, O. M., da Silveira, R. B., Meissner, G. O., Mangili, O. C., . . . Senff-Ribeiro, A. (2010). Astacin-like metalloproteases are a gene family of toxins present in the venom of different species of the brown spider (genus *Loxosceles*). *Biochimie*, *92*(1), 21-32. https://doi.org/10.1016/j.biochi.2009.10.003

Tsujimoto, H., Kotsyfakis, M., Francischetti, I. M. B., Eum, J. H., Strand, M. R., & Champagne, D. E. (2012). Simukunin from the salivary glands of the black fly *Simulium vittatum* inhibits enzymes that regulate clotting and inflammatory responses. *PLoS One*, *7*(2), e29964. https://doi.org/10.1371/journal.pone.0029964

Vedel, A., Andersen, B. B., & Riisgård, H. U. (1994). Field investigations of pumping activity of the facultatively filter-feeding polychaete *Nereis diversicolor* using an improved infrared phototransducer system. Mar Ecol Prog Ser, *103*(1/2), 91-101. https://doi.org/10.3354/meps103091

Verdes, A., Simpson, D., & Holford, M. (2018). Are Fireworms Venomous? Evidence for the Convergent Evolution of Toxin Homologs in Three Species of Fireworms (Annelida, Amphinomidae). *Genome Biol Evol*, *10*(1), 249-268. https://doi.org/10.1093/gbe/evx279

von Reumont, B. M., Campbell, L. I., & Jenner, R. A. (2014a). *Quo vadis* venomics? A roadmap to neglected venomous invertebrates. *Toxins (Basel)*, *6*(12), 3488-3551. https://doi.org/10.3390/toxins6123488

von Reumont, B. M., Campbell, L. I., Richter, S., Hering, L., Sykes, D., Hetmank, J., . . . Bleidorn, C. (2014b). A Polychaete's powerful punch: venom gland transcriptomics of *Glycera* reveals a complex cocktail of toxin homologs. *Genome Biol Evol*, *6*(9), 2406-2423. https://doi.org/10.1093/gbe/evu190

Wang, J., Yazdani, S., Han, A., & Schapira, M. (2020). Structure-based view of the druggable genome. *Drug Discov Today*, *25*(3), 561-567. https://doi.org/10.1016/j.drudis.2020.02.006

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10*(1), 57-63. https://doi.org/10.1038/nrg2484

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., . . . Venables, B. (2020). *gplots: Various R Programming Tools for Plotting Data*. (R package version 3.1.1) https://CRAN.R-project.org/package=gplots

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-24277-4

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. (R package version 1.1.1) https://CRAN.R-project.org/package=cowplot

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

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, *13*, 134. https://doi.org/10.1186/1471-2105-13-134

Zewail-Foote, M., & Hurley, L. H. (1999). Ecteinascidin 743: a minor groove alkylator that bends DNA toward the major groove. *J Med Chem*, *42*(14), 2493-2497. https://doi.org/10.1021/jm990241l

A

APPENDIX

A.1 Figures and Tables

Table A1. Sampling.

	Sample Name	Species	Organ	Number of Individuals	Number of Raw Reads
1	G4-SX-PR	*Glycera alba*	Proboscis	1	39 749 500
2	G8-SX-PR-pool	*Glycera alba*	Proboscis	2 (pool)	43 177 154
3	G9-SX-PR-pool	*Glycera alba*	Proboscis	3 (pool)	45 489 590
4	G4-SX-PL	*Glycera alba*	Skin	1	43 073 170
5	G12-SX-PL-pool	*Glycera alba*	Skin	3 (pool)	46 741 642
6	G13-SX-PL-pool	*Glycera alba*	Skin	3 (pool)	38 915 178
7	Hd1-G	*Hediste diversicolor*	Glands	15 (pool)	31 099 868
8	Hd2-G	*Hediste diversicolor*	Glands	10 (pool)	123 202 926
9	Hd3-G	*Hediste diversicolor*	Glands	10 (pool)	49 858 858
10	Hd1-PR	*Hediste diversicolor*	Proboscis	15 (pool)	42 056 540
11	Hd2-PR	*Hediste diversicolor*	Proboscis	10 (pool)	94 650 066
12	Hd3-PR	*Hediste diversicolor*	Proboscis	10 (pool)	42 195 514

Table A.2. The primers sequences used for PCR and RT-qPCR. The genes of interest amplified encoded for an ovoinhibitor (Ov), a cysteine-rich venom protein (TX31), a pathogenesis-related protein (CRISP) (Pat) and a thyrostimulin beta-5 subunit (Thy), whereas the housekeeping genes were 18S and beta-actin (BAct). The primer forward of the thyrostimulin beta-5 subunit was used for PCR and RT-qPCR.

Target	Primers	PCR or RT-qPCR	Primer Sequences	Amplicon Size
18S	Foward	PCR/RT-qPCR	CGATGGTACGTGATATGCC	176
	Reverse	PCR/RT-qPCR	CGAATGAGTCCCGTATTGT	
BAct	Foward	PCR/RT-qPCR	CGGTATCGTGCTGGATTC	163
	Reverse	PCR/RT-qPCR	CGTGGTGGTGAAGCTGTA	
Ov	Foward_1	PCR	GCTTACATCTTATCATGCTC	639
	Reverse_1	PCR	CTGTATTGCACTCAGGTTC	
	Foward_2	PCR	CTACAAGTCGTGTCATGC	789
	Reverse_2	PCR	CTGCTTCTATTGGTTGGC	
TX31	Foward	PCR	GAATCCTGAACCTGTCTGTG	954
	Reverse	PCR	CAACCTGTTCTTAACATACTCC	
Pat	Foward	PCR	GTGGTCAAGATCAGAACTGC	667
	Reverse	PCR	GGAAACCATTTACAGGAGAGG	
Thy	Foward	PCR/RT-qPCR	GGACAGGCATGAAGGACA	465
	Reverse	PCR	CTCACAACGATTCGCAACT	
Ov	Foward	RT-qPCR	GTGTACCTACGAATACAACC	150
	Reverse	RT-qPCR	CCATCAGATCCGCAAACT	
TX31	Foward	RT-qPCR	AGGCTTCTTGAATGATGCT	172
	Reverse	RT-qPCR	TGTGTTGGACAGTGGTT	
Pat	Foward	RT-qPCR	ACCTCATCATCCTCTTTGC	140
	Reverse	RT-qPCR	TCTGCTGCTCAGTCTTGC	
Thy	Reverse	RT-qPCR	AGGATACAGACGGCAGTG	167

Table A.3. Transcriptome assembly. The Contig N50, Contig Ex90N50 and Read Content were the statistics used to evaluate the quality of the transcriptome assembly.

		Good Assembly Values	***Glycera alba* Transcriptome Assembly**	***Hediste diversicolor* Transcriptome Assembly**
Contig N50	Stats based on all transcript contigs:	>1000	1122	3176
	Stats based on only longest isoform per 'gene':		886	2033
Contig Ex90N50		>1000	1129	3327
Read Content		~70-80%	78.83 ± 2.36%	81.44 ± 0.96%

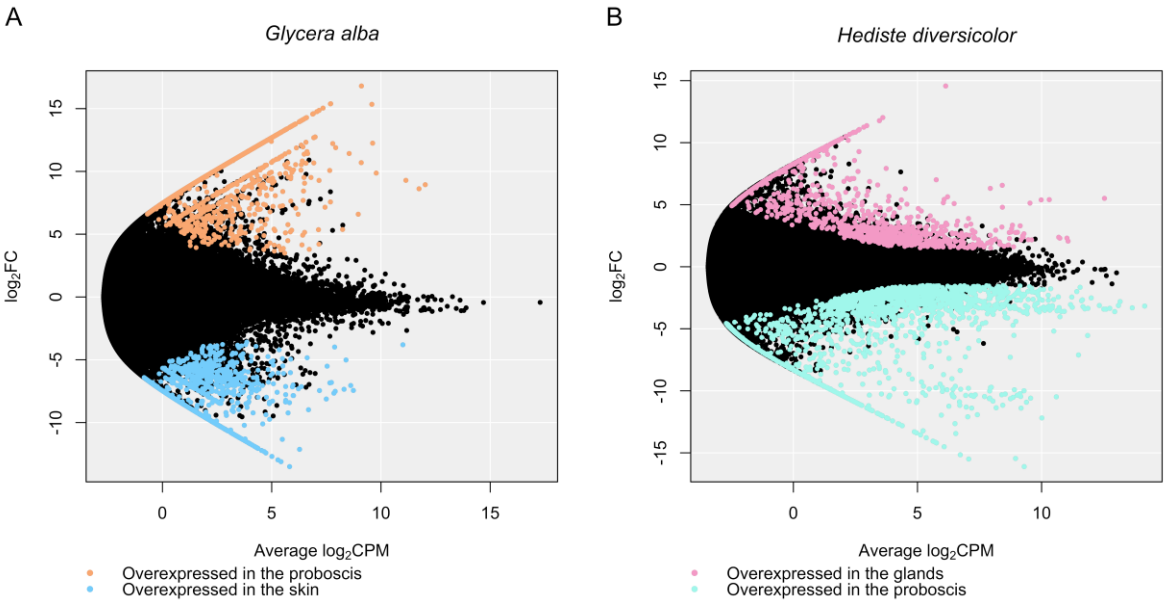

Figure A.1. Smear plot illustrating differentially-expressed transcripts. A) between *Glycera alba*'s proboscis and skin. B) between *Hediste diversicolor*'s glands and proboscis. The black dots represent transcripts that are not differentially-expressed between organs. The cut-off for differential expression was set at A | log_2FC | > 1.5 and FDR-adjusted $p < 0.05$.

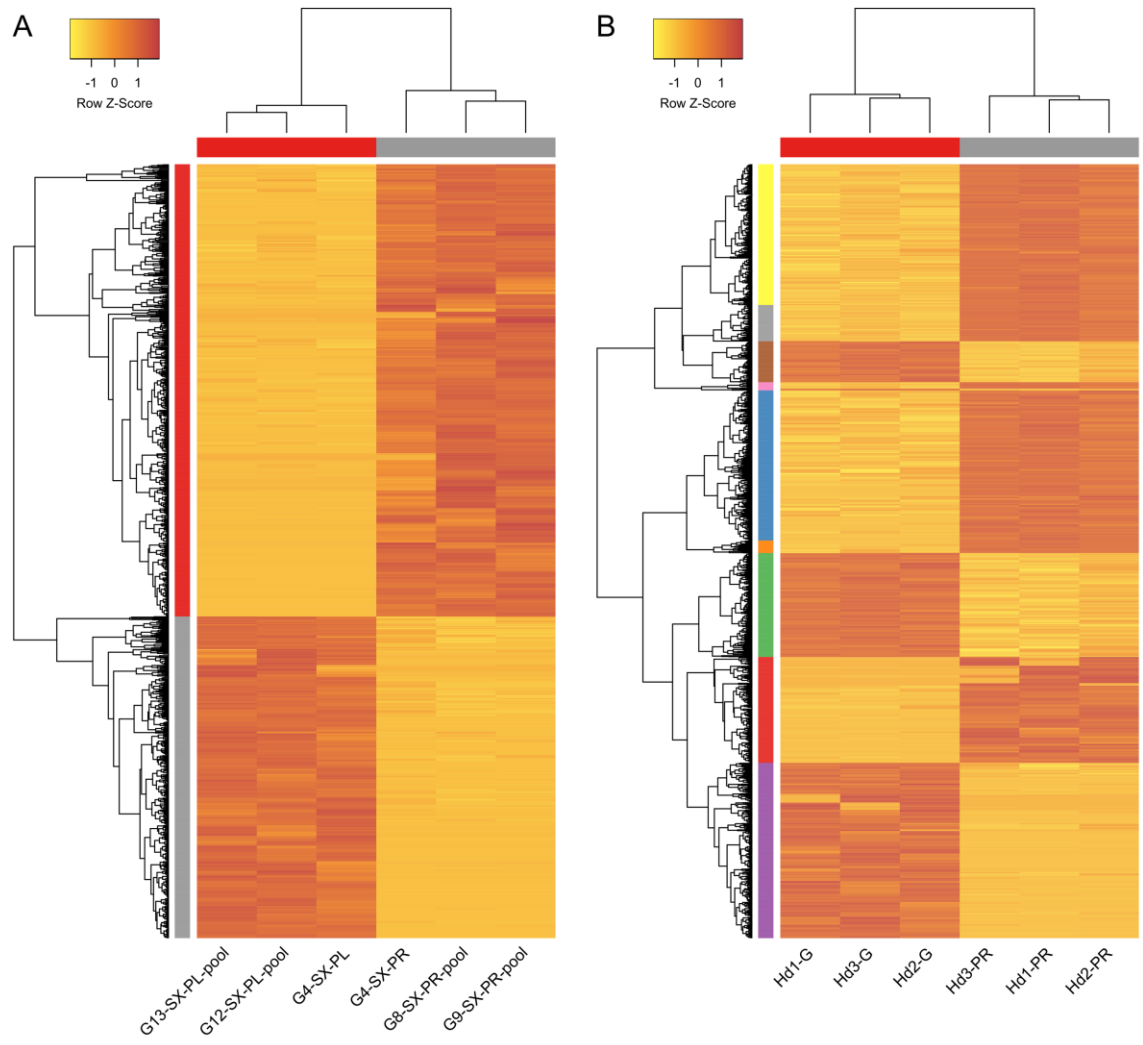

Figure A.2. Heatmaps illustrating relative gene expression of the differentially-expressed genes. A) between *Glycera alba*'s proboscis (PR) and skin (PL). B) between *Hediste diversicolor*'s glands (G) and proboscis (PR). A | log2FC | > 1.5 and an FDR adjusted $p < 0.05$ were the cut-offs set for differential expression. The horizontal dendrogram illustrates the association between the three independent replicates for each organ, whereas the vertical dendrogram represents the association between the proteins. The metric and function of the cluster analysis are Euclidian distances and complete linkage, respectively.

Table A.4. Top10 differentially-expressed genes in *Glycera alba*'s proboscis relative to the skin. The cut-offs established were | log_2FC | > 1.5 and an FDR-adjusted *p* < 0.05 as cut-off. The log_2FC > 1.5 is indicative of overexpressed genes in *G. alba*'s proboscis compared to the skin, while log_2FC < - 1.5 is for underexpressed.

log_2FC	log_2CPM	FDR*p*	Protein	Accession	%ID	*e*-value	Organism
7.973	0.450	1.58E-04	Glycerotoxin (Fragment)	A0A1U9VX95	98.693	0.00E+00	*Glycera tridactyla*
7.432	-0.020	1.10E-04	Glycerotoxin (Fragment)	A0A1U9VX95	97.416	0.00E+00	*Glycera tridactyla*
9.488	1.856	3.26E-06	Glycerotoxin (Fragment)	A0A1U9VX95	97.386	0.00E+00	*Glycera tridactyla*
7.137	-0.266	9.78E-05	Tubulin alpha-1 chain	P06603	97.143	0.00E+00	*Drosophila melanogaster*
-6.512	1.465	4.39E-05	Tubulin alpha-3 chain [Cleaved into: Detyrosinated tubulin alpha-3 chain]	P05214	96.919	0.00E+00	*Mus musculus*
10.435	2.774	8.60E-07	Glycerotoxin (Fragment)	A0A1U9VX95	96.894	0.00E+00	*Glycera tridactyla*
9.009	4.611	4.86E-08	Actin, cytoplasmic 1 (Beta-actin) [Cleaved into: Actin, cytoplasmic 1, N-terminally processed]	O93400	93.050	0.00E+00	*Xenopus laevis*
10.324	2.665	1.13E-06	Glycerotoxin (Fragment)	A0A1U9VX91	92.36	0.00E+00	*Glycera tridactyla*
8.977	3.252	1.46E-04	Glycerotoxin (Fragment)	A0A1U9VX95	85.657	0.00E+00	*Glycera tridactyla*
10.297	4.560	3.63E-08	Glycerotoxin (Fragment)	A0A1U9VX91	83.717	0.00E+00	*Glycera tridactyla*

log_2FC – log_2 fold change; log_2CPM – Average log_2 counts per million; FDR*p* – False discovery rate adjusted *p*-value; %ID – Percentage of identity

Table A.5. Top10 differentially-expressed genes in *Hediste diversicolor*'s glands relative to the proboscis. The | log_2FC | > 1.5 and an FDR-adjusted $p < 0.05$ were set as cut-off. The $log_2FC > 1.5$ is indicative of overexpressed genes in *H. diversicolor*'s glands compared to the proboscis, while $log_2FC < -1.5$ is for underexpressed.

log_2FC	log_2CPM	FDR*p*	Protein	Accession	%ID	*e*-value	Organism
-3.312	13.652	4.11E-14	Actin, cytoplasmic 1 (Beta-actin) [Cleaved into: Actin, cytoplasmic 1, N-terminally processed]	Q6NVA9	98.932	0.00E+00	*Xenopus tropicalis*
-2.054	8.644	2.40E-07	Glycocyamine kinase	P51546	97.861	0.00E+00	*Hediste diversicolor*
-3.351	12.747	2.03E-13	Glycocyamine kinase	P51546	97.861	0.00E+00	*Hediste diversicolor*
-3.271	11.237	6.98E-13	Glycocyamine kinase	P51546	94.906	0.00E+00	*Hediste diversicolor*
-1.577	7.468	1.14E-04	Tubulin alpha-1C chain (Alpha-tubulin 3) [Cleaved into: Detyrosinated tubulin alpha-1C chain]	P68365	94.286	0.00E+00	*Cricetulus griseus*
9.523	1.149	7.21E-04	cAMP-dependent protein kinase catalytic subunit 1 (PKA C)	P12370	88.988	0.00E+00	*Drosophila melanogaster*
5.388	4.350	3.48E-06	26S proteasome regulatory subunit 6A (Tat-binding protein 1)	P17980	88.124	0.00E+00	*Homo sapiens*
-1.587	6.880	1.93E-04	cAMP-dependent protein kinase regulatory subunit (N4 subunit of protein kinase A)	P31319	87.541	0.00E+00	*Aplysia californica*
-3.211	5.963	3.22E-05	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform (CAM-PRP catalytic subunit) (Calmodulin-dependent calcineurin A subunit alpha isoform) (CNA alpha)	P63328	86.579	0.00E+00	*Mus musculus*
-2.443	6.054	2.35E-05	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform (CAM-PRP catalytic subunit) (Calmodulin-dependent calcineurin A subunit alpha isoform) (CNA alpha)	P63328	86.579	0.00E+00	*Mus musculus*

log_2FC – log_2 fold change; log_2CPM – Average log_2 counts per million; FDR*p* – False discovery rate adjusted *p*-value; %ID – Percentage of identity

Table A.6. Top10 overexpressed genes in *Glycera alba*'s skin relative to the proboscis. The cut-offs established were $log_2FC > 1.5$ and an FDR-adjusted $p < 0.05$.

log_2FC	log_2CPM	FDR*p*	Protein	Accession	%ID	*e*-value	Organism
13.508	5.818	6.73E-14	Fibropellin-1 (Epidermal growth factor-related protein 1)	P10079	35.443	1.11E-68	*Strongylocentrotus purpuratus*
12.694	5.007	2.42E-12	Endothelin-converting enzyme homolog	Q8IS64	33.623	1.19E-122	*Locusta migratoria*
12.318	4.663	9.34E-11	Protein jagged-1b	Q90Y54	25.783	1.76E-26	*Danio rerio*
11.971	4.288	3.14E-11	Arrestin domain-containing protein 17 (Calcineurin-interacting protein 1)	O45782	24.390	1.73E-18	*Caenorhabditis elegans*
11.789	4.108	2.10E-11	Sodium-driven chloride bicarbonate exchanger (Solute carrier family 4 member 10)	Q32LP4	53.776	0.00E+00	*Bos taurus*
11.688	4.008	6.17E-06	Complement component C8 beta chain	Q9PVW7	40.385	4.54E-07	*Paralichthys olivaceus*
11.351	3.674	1.35E-10	Fibropellin-1 (Epidermal growth factor-related protein 1)	P10079	49.351	1.63E-102	*Strongylocentrotus purpuratus*
11.309	3.782	1.54E-04	Collagen alpha-1(IV) chain	P17139	41.613	1.59E-128	*Caenorhabditis elegans*
11.215	3.539	1.10E-04	Peptidyl-prolyl cis-trans isomerase B (Cyclophilin B) (Rotamase B)	P24367	46.774	3.19E-54	*Gallus gallus*
11.179	3.504	1.06E-05	Deoxyribonuclease-1 (DNase I)	P49183	38.725	7.75E-36	*Mus musculus*

log_2FC – log_2 fold change; log_2CPM – Average log_2 counts per million; FDR*p* – False discovery rate adjusted *p*-value; %ID – Percentage of identity

Table A.7. Top10 overexpressed genes in *Hediste diversicolor*'s glands relative to the proboscis. The cut-offs established were $log_2FC > 1.5$ and an FDR-adjusted $p < 0.05$.

log_2FC	log_2CPM	FDR*p*	Protein	Accession	%ID	*e*-value	Organism
14.568	6.136	4.36E-38	Serine palmitoyltransferase 2 (Long chain base biosynthesis protein 2)	P97363	77.043	9.76E-139	*Mus musculus*
12.024	3.602	3.71E-21	APOBEC1 complementation factor	Q923K9	55.660	0.00E+00	*Rattus norvegicus*
11.770	3.465	1.36E-16	NAD kinase (Poly(P)/ATP NAD kinase)	P58058	59.686	5.21E-156	*Mus musculus*
11.370	2.955	4.79E-16	Serine/threonine-protein kinase 26	Q9P289	80.000	3.06E-153	*Homo sapiens*
11.062	2.652	8.58E-16	Serine/threonine-protein kinase 26	Q9P289	80.000	5.61E-153	*Homo sapiens*
11.009	2.623	8.21E-15	APOBEC1 complementation factor	Q923K9	55.451	0.00E+00	*Rattus norvegicus*
10.845	2.500	2.85E-13	WD repeat-containing protein 90	A0JP70	45.756	0.00E+00	*Xenopus tropicalis*
10.845	2.500	2.85E-13	DNA-dependent protein kinase catalytic subunit	Q8QGX4	40.034	0.00E+00	*Gallus gallus*
10.800	2.395	1.10E-14	Matrix metalloproteinase-9	P41246	34.098	3.05E-33	*Oryctolagus cuniculus*
10.673	2.267	2.89E-08	Collagen alpha-4(VI) chain	A2AX52	24.388	1.03E-68	*Mus musculus*

log_2FC – log_2 fold change; log_2CPM – Average log_2 counts per million; FDR*p* – False discovery rate adjusted *p*-value; %ID – Percentage of identity

Table A.8. Top10 overexpressed genes in *Hediste diversicolor*'s proboscis relative to the glands. The cut-offs established were $log_2FC > 1.5$ and an FDR-adjusted $p < 0.05$.

log_2FC	log_2CPM	FDR*p*	Protein	Accession	%ID	*e*-value	Organism
15.497	7.065	1.09E-40	Matrix metalloproteinase-24 [Cleaved into: Processed matrix metalloproteinase-24]	Q9Y5R2	38.132	9.30E-44	*Homo sapiens*
14.506	6.074	3.66E-39	Sorbin and SH3 domain-containing protein 1	Q9BX66	31.017	6.91E-51	*Homo sapiens*
14.195	5.764	1.25E-43	Collagen alpha-6(VI) chain	A6NMZ7	26.471	7.18E-08	*Homo sapiens*
14.138	5.715	2.17E-44	Collagen alpha-1(XXVII) chain B	A0MSJ1	42.718	3.61E-105	*Danio rerio*
13.553	5.122	7.72E-35	Peroxidasin	Q9VZZ4	38.715	2.19E-103	*Drosophila melanogaster*
13.550	5.120	9.76E-36	Matrix metalloproteinase-9	P14780	43.429	1.11E-32	*Homo sapiens*
13.385	4.955	1.09E-28	Cartilage matrix protein (Matrilin-1)	P05099	28.191	1.11E-07	*Gallus gallus*
13.223	5.224	4.74E-26	Collagen alpha-1(XXVII) chain B	A0MSJ1	42.718	3.61E-105	*Danio rerio*
12.814	4.386	9.92E-26	Collagen alpha-6(VI) chain	Q8C6K9	24.229	2.41E-07	*Mus musculus*
12.651	4.224	4.93E-24	Putative tyrosinase-like protein tyr-1	P34269	39.607	1.14E-67	*Caenorhabditis elegans*

log_2FC – log_2 fold change; log_2CPM – Average log_2 counts per million; FDR*p* – False discovery rate adjusted *p*-value; %ID – Percentage of identity

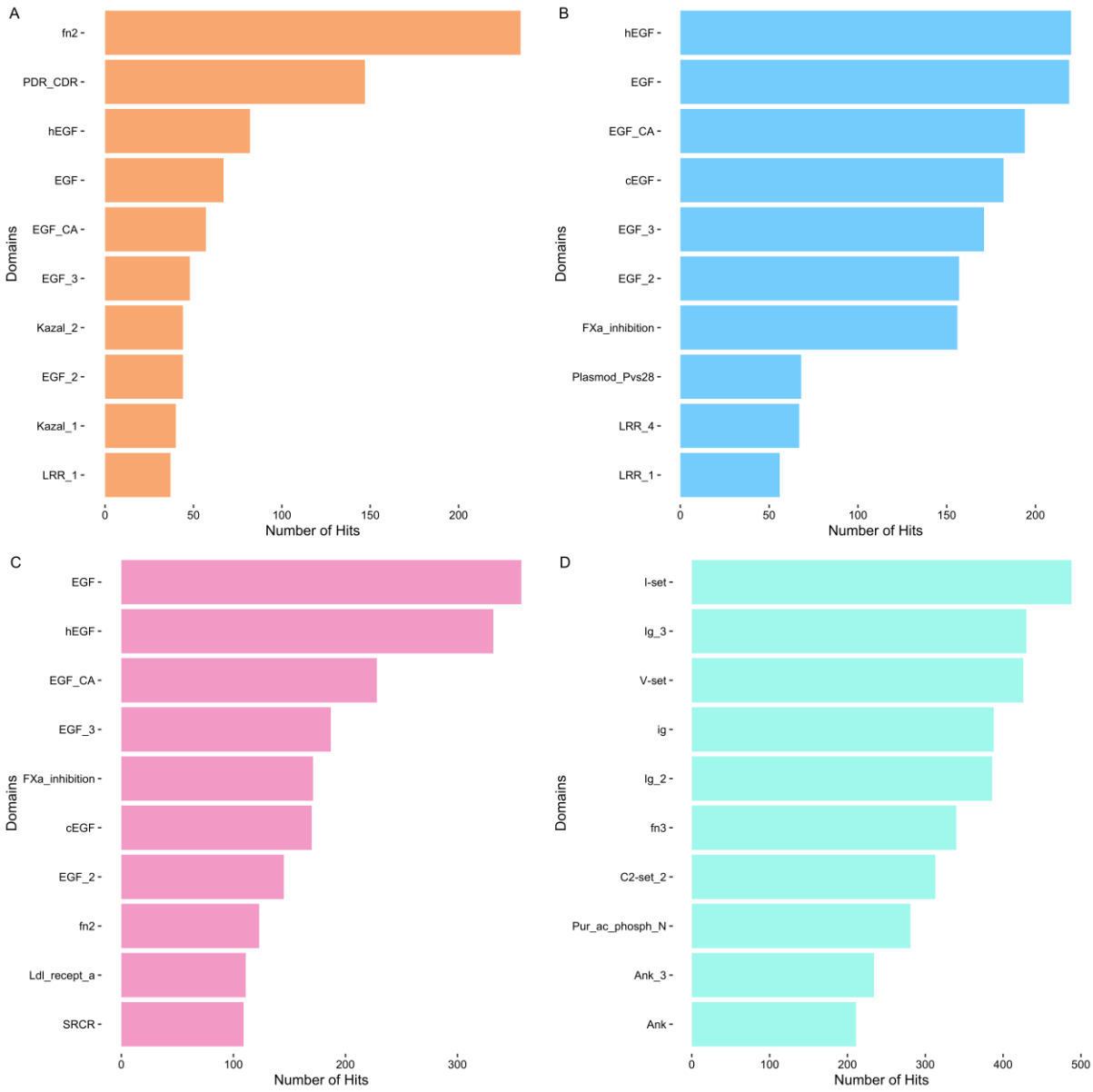

Figure A.3. Top10 domains in the differentially-expressed genes that encoded annotated proteins. The horizontal bar chart indicates number of hits of domains against Pfam in the overexpressed genes A) in *Glycera alba*'s proboscis. B) in *Glycera alba*'s skin. C) in *Hediste diversicolor*'s glands. D) in *Hediste diversicolor*'s proboscis. The same predicted coding region can have multiple domains.

Table A.9. Proteins up-regulated in *Glycera alba*'s proboscis with allergenic properties.

Accession	Protein	Allergenic Properties
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	Causes an allergic reaction in humans. Binds to IgE of egg-allergic patients. Immunoreactivity is lost by simulated gastric and gastroduodenal digestion (PubMed:23122126). Binds to rabbit anti-ovomucoid IgG antibody indicating the cross-reactivity between this protein and the ovomucoid protein from egg white (PubMed:6838526). {ECO:0000269|PubMed:23122126, ECO:0000269|PubMed:6838526}.
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	Causes an allergic reaction in humans. Binds to IgE of egg-allergic patients. Immunoreactivity is lost by simulated gastric and gastroduodenal digestion (PubMed:23122126). Binds to rabbit anti-ovomucoid IgG antibody indicating the cross-reactivity between this protein and the ovomucoid protein from egg white (PubMed:6838526). {ECO:0000269|PubMed:23122126, ECO:0000269|PubMed:6838526}.
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	Causes an allergic reaction in humans. Binds to IgE of egg-allergic patients. Immunoreactivity is lost by simulated gastric and gastroduodenal digestion (PubMed:23122126). Binds to rabbit anti-ovomucoid IgG antibody indicating the cross-reactivity between this protein and the ovomucoid protein from egg white (PubMed:6838526). {ECO:0000269|PubMed:23122126, ECO:0000269|PubMed:6838526}.
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	Causes an allergic reaction in humans. Binds to IgE of egg-allergic patients. Immunoreactivity is lost by simulated gastric and gastroduodenal digestion (PubMed:23122126). Binds to rabbit anti-ovomucoid IgG antibody indicating the cross-reactivity between this protein and the ovomucoid protein from egg white (PubMed:6838526). {ECO:0000269|PubMed:23122126, ECO:0000269|PubMed:6838526}.

Table A.10. Proteins up-regulated in *Glycera* alba's proboscis with biotechnological use.

Accession	Protein	Biotechnological Use
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	The galactomannan conjugate of this protein prepared through the Maillard reaction shows almost the same inhibitory activity toward trypsin, chymotrypsin and elastase, with stronger heat and emulsion stability, and better emulsifying properties than the untreated protein. The conjugate can therefore be useful for industrial application. {ECO:0000269|PubMed:14519973}.
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	The galactomannan conjugate of this protein prepared through the Maillard reaction shows almost the same inhibitory activity toward trypsin, chymotrypsin and elastase, with stronger heat and emulsion stability, and better emulsifying properties than the untreated protein. The conjugate can therefore be useful for industrial application. {ECO:0000269|PubMed:14519973}.
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	The galactomannan conjugate of this protein prepared through the Maillard reaction shows almost the same inhibitory activity toward trypsin, chymotrypsin and elastase, with stronger heat and emulsion stability, and better emulsifying properties than the untreated protein. The conjugate can therefore be useful for industrial application. {ECO:0000269|PubMed:14519973}.
P10184	Ovoinhibitor (Serine protease inhibitor Kazal-type 5) (allergen Gal d OIH)	The galactomannan conjugate of this protein prepared through the Maillard reaction shows almost the same inhibitory activity toward trypsin, chymotrypsin and elastase, with stronger heat and emulsion stability, and better emulsifying properties than the untreated protein. The conjugate can therefore be useful for industrial application. {ECO:0000269|PubMed:14519973}.
K0DZA0	2-oxoglutarate-dependent dioxygenase htyE (L-homotyrosine biosynthetic cluster protein E)	Due to their effectiveness as antifungal agents, echinocandin derivatives can be used for the treatment of human invasive candidiasis (PubMed:22998630). {ECO:0000269|PubMed:22998630}.
P37610	Alpha-ketoglutarate-dependent taurine dioxygenase (2-aminoethanesulfonate dioxygenase) (Sulfate starvation-induced protein 3)	Taurine dioxygenase can be used for enzymatic determination of taurine concentration in food quality control, biological research, and medical diagnosis. {ECO:0000269|PubMed:22595347}.
P86148	Reticulocyte-binding protein PFD0110w	Possible candidate for an effective malaria vaccine as determined by epitope response in sera. {ECO:0000269|PubMed:17653272}.
Q7SIG4	Diisopropyl-fluorophosphatase	Has potential application in bio-remediation by detoxification of organo-phosphates used in insecticides or nerve agents used in chemical warfare such as soman, tabun and sarin. {ECO:0000269|PubMed:11134940}.

Table A.11. Proteins up-regulated in *Glycera* alba's skin with biotechnological use.

Accession	Protein	Biotechnological Use
B3FWU0	Reducing polyketide synthase rdc5 (Hypothemycin biosynthesis cluster protein rdc5)	Radicicol is an important pharmacophore as an inhibitor of heat shock protein 90 (Hsp90), an ATP-dependent chaperone involved in the post-translational maturation and stabilization of over one hundred proteins, and which activity has been implicated in diverse pathologies ranging from oncology to neurodegenerative and infectious diseases (PubMed:19860733).

Table A.12. Proteins up-regulated in *Hediste diversicolor*'s glands with biotechnological use.

Accession	Protein	Biotechnological Use
S3DQP3	Nonribosomal peptide synthetase gloA (Pneumocandin biosynthesis cluster protein A)	Pneumocandin B0 is the starting molecule for the first semisynthetic echinocandin antifungal drug, caspofungin acetate (PubMed:25527531). Pneumocandin B0 is a minor fermentation product, and its industrial production was achieved by a combination of extensive mutation and medium optimization (PubMed:25527531). Inactivation of three of gloP/GLP450-1, gloO/GLP450-2, and gloM/GLOXY1 generates 13 different pneumocandin analogs that lack one, two, three, or four hydroxyl groups on 4R,5R-dihydroxy-ornithine and 3S,4S-dihydroxy-homotyrosine of the parent hexapeptide (PubMed:25879325). All of these cyclic lipopeptides show potent antifungal activities, and two new metabolites pneumocandins F and G are more potent in vitro against *Candida* species and *Aspergillus fumigatu*s than the principal fermentation products, pneumocandins A0 and B0 (PubMed:25879325). Moreover, feeding alternative side chain precursors yields acrophiarin and 4 additional pneumocandin congeners with straight C14, C15, and C16 side chains. One of those compounds, pneumocandin I, has elevated antifungal activity and similar hemolytic activity compared to pneumocandin B0, the starting molecule for caspofungin, demonstrating the potential for using gloD/GLligase for future engineering of new echinocandin analogs (PubMed:27494047). {ECO:0000269|PubMed:25527531, ECO:0000269|PubMed:25879325, ECO:0000269|PubMed:27494047}.

Table A.13. Proteins up-regulated in *Hediste diversicolor*'s glands with pharmaceutical use.

Accession	Protein	Pharmaceutical Use
P12643	Bone morphogenetic protein 2	Available under the name Infuse (Medtronic Sofamor Danek). Used for treating open tibial shaft fractures.

Table A.14. Proteins up-regulated in *Hediste diversicolor*'s proboscis with allergenic proprieties.

Accession	Protein	Allergenic properties
A0A1L5YRA2	Triosephosphate isomerase (Allergen Scy p 8) (Methylglyoxal synthase) (Triose-phosphate isomerase)	Causes an allergic reaction in human. Binds to IgE of patients allergic to crabs (PubMed:30187588, PubMed:31668066). Binds to IgE in 23% of the 30 patients tested (PubMed:30187588). Binds to IgE in all 12 patients tested. Induces activation of human basophils (PubMed:31668066). {ECO:0000269|PubMed:30187588, ECO:0000269|PubMed:31668066}.
A0T2M3	Polcalcin Cup a 4 (Calcium-binding pollen allergen Cup a 4)	Causes an allergic reaction in human. Binds to IgE in 9.6% of 177 patients allergic to Arizona cypress pollen. {ECO:0000269|PubMed:20869950}.
A1KYZ2	Tropomyosin (Tropomyosin, fast isoform) (Tm-Penm-fast)	Causes an allergic reaction in human. {ECO:0000269|PubMed:17263503, ECO:0000269|Ref.3}.
L7UZ85	Alpha-actinin (F-actin cross-linking protein) (allergen Der f 24)	Causes an allergic reaction in human. Binds to IgE. Activates basophils. {ECO:0000269|PubMed:24324688}.
P86686	Venom allergen 5 (Cysteine-rich venom protein)	Causes an allergic reaction in human (Ref.1, PubMed:24308509). Binds to IgE (PubMed:24308509). {ECO:0000269|PubMed:24308509, ECO:0000269|Ref.1}.

Table A.15. Proteins up-regulated in *Hediste diversicolor*'s proboscis with biotechnological use.

Accession	Protein	Biotechnological Use
C0H5F4	Reticulocyte binding protein 2 homolog b	Possible candidate for an effective malaria vaccine as determined by epitope response in sera. {ECO:0000269|PubMed:17653272}.
Q96524	Cryptochrome-2 (Blue light photoreceptor) (Protein PHR homolog 1) (Protein SUPPRESSOR OF elf3 20)	The rapid blue light-mediated reversible interaction between CRY2 and BHLH63/CIB1 is used to design an optogenetic control of target proteins or organelles. {ECO:0000269|PubMed:21037589, ECO:0000269|PubMed:22847441, ECO:0000269|PubMed:23833191, ECO:0000269|PubMed:24718798, ECO:0000269|PubMed:25963241}.

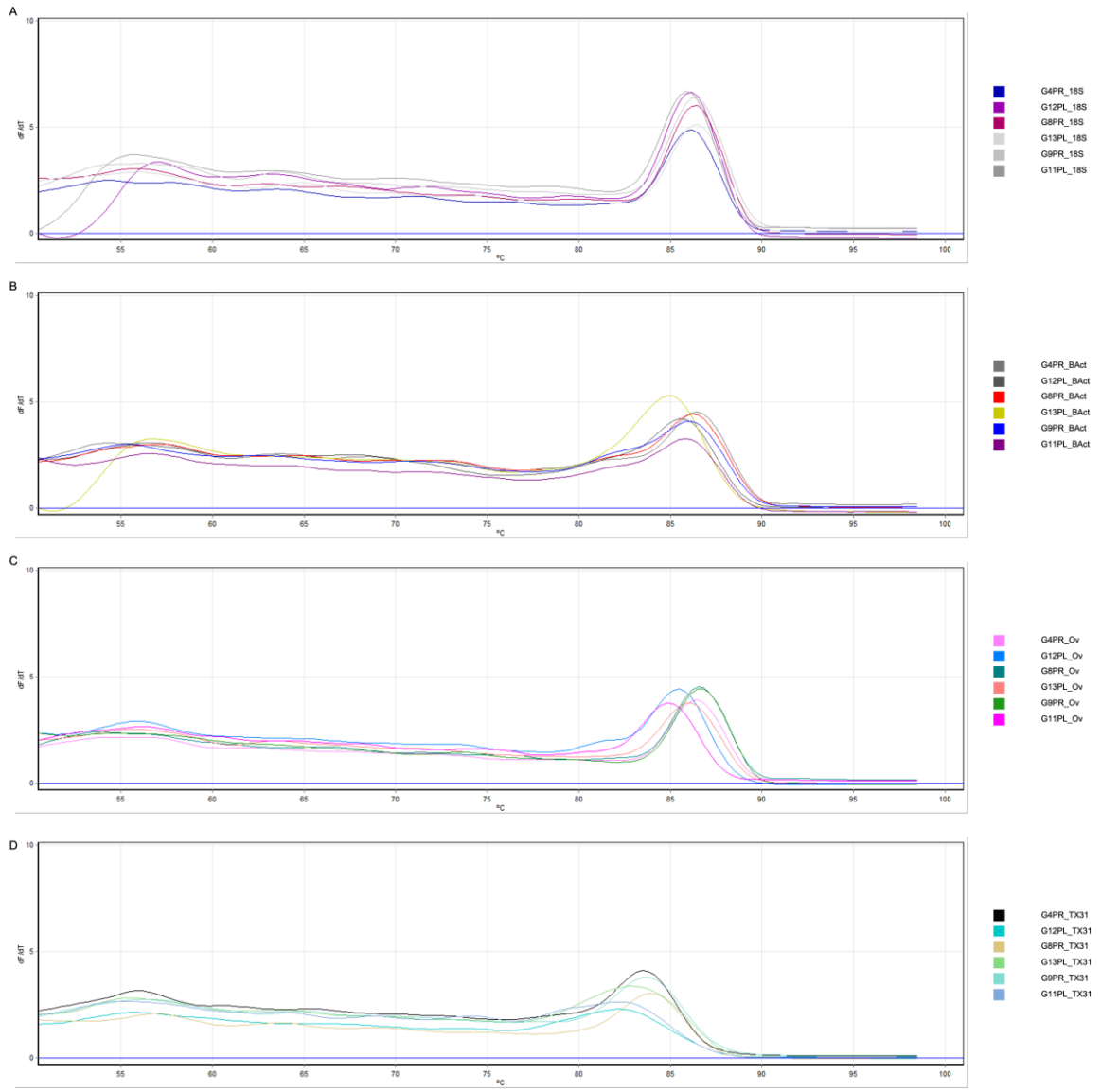

Figure A.4. Melting curves of *Glycera alba*'s expressed sequence tags amplified by RT-qPCR. The expressed sequence tags amplified were from A) the housekeeping gene 18S. B) the housekeeping gene beta-actin (BAct). C) the transcript encoding for an ovoinhibitor-like protein (Ov). D) the transcript encoding for a cysteine-rich venom protein (TX31). Three replicates were used for the proboscis (PR) and skin (PL) samples.

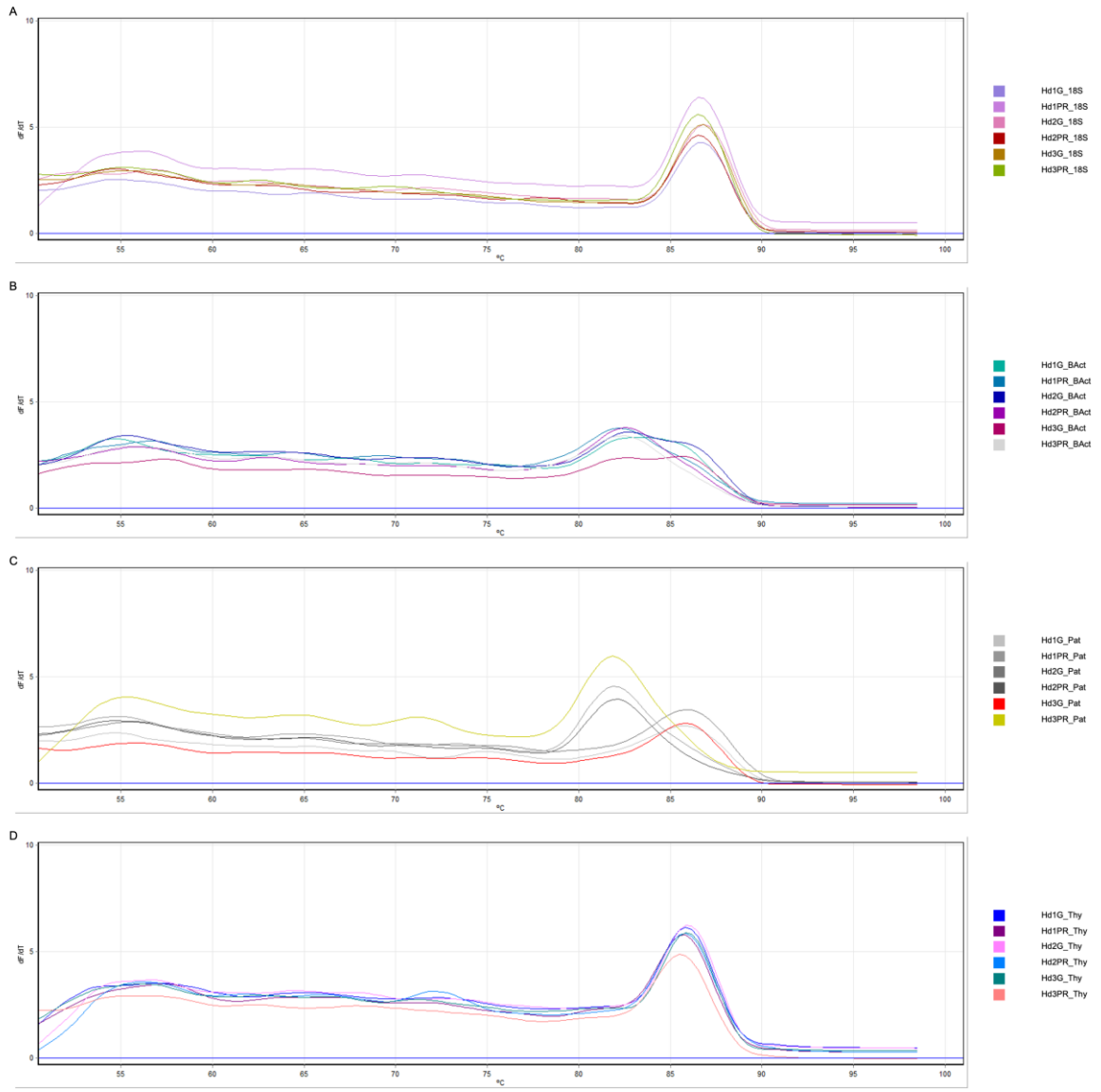

Figure A.5. Melting curves of *Hediste diversicolor*'s expressed sequence tags amplified by RT-qPCR. The expressed sequence tags amplified were from A) the housekeeping gene 18S. B) the housekeeping gene beta-actin (BAct). C) the transcript encoding a pathogenesis-related protein (CRISP) (Pat). D) the transcript encoding a thyrostimulin beta-5 subunit (Thy). Three replicates were used for the glands (G) and the proboscis (PR) samples.

Table A.16. Match between the overexpressed genes in the proboscis of *Glycera alba* relative to the skin and human interactor. The interactor was the human protein homologous to the proteins up-regulated in the proboscis of *G. alba*. When the Polychaeta gene was chiefly annotated through the identification of conserved domains in the encoded protein, it was described in the protein column which conserved domains were found present in the sequence.

Interactor	Gene	Protein
AGRN	NA	Four-domain proteases inhibitor
AGRN	NA	Four-domain proteases inhibitor
BAD	NA	Protein with Bcl-2_BAD, Bcl-2_BAD, Bclx_interact, BMF and BMF domains
CALM1	NA	Calmodulin
CALML3	cam1	Calmodulin
CELA2B	CELA2B	Chymotrypsin-like elastase family member 2B (Elastase-2B)
CELA2B	NA	Trypsin
CELA2B	NA	Trypsin
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
FADD	NA	Protein with Death, Death, Rax2, Rax2, NB, Shisa domains
FAIM	Faim	Fas apoptotic inhibitory molecule 1 (rFAIM)
FBN1	GLTx	Glycerotoxin (Fragment)
FCN1	NA	Fibrinogen-like protein A (FREP-A)
FCN2	FCN2	Ficolin-2 (Collagen/fibrinogen domain-containing protein 2)
PI16	Pi16	Peptidase inhibitor 16 (Cysteine-rich protease inhibitor) (CD antigen CD364)

Table A.17. Match between the overexpressed genes in the skin of *Glycera alba* relative to the proboscis and human interactor. The interactor was the human protein homologous to the proteins up-regulated in the skin of *G. alba*. When the Polychaeta gene was chiefly annotated through the identification of conserved domains in the encoded protein, it was described in the protein column which conserved domains were found in the sequence.

Interactor	Gene	Protein
ANGPT1	ANGPT1	Angiopoietin-1
ANGPT2	ANGPT2	Angiopoietin-2
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
FBN1	FBN1	Fibrillin-1 [Cleaved into: Asprosin]
JAK2	Jak2	Tyrosine-protein kinase JAK2
MIF	MIF	Macrophage migration inhibitory factor (Glycosylation-inhibiting factor) (L-dopachrome isomerase) (Phenylpyruvate tautomerase)
MIF	MIF	Macrophage migration inhibitory factor (Glycosylation-inhibiting factor) (L-dopachrome isomerase) (Phenylpyruvate tautomerase)
NELL2	NA	Protein with EGF_3, EGF_3, VWD, EGF_MSP1_1, EGF_MSP1_1 domains
NT5E	NA	Snake venom 5'-nucleotidase (Ecto-5'-nucleotidase) (Fragment)
RASSF1	RASSF1	Ras association domain-containing protein 1

Table A.18. Match between the overexpressed genes in the glands of *Hediste diversicolor* relative to the proboscis and human interactor. The interactor was the human protein homologous to the proteins up-regulated in the glands of *H. diversicolor*.

Interactor	Gene	Protein
ACR	Acr	Acrosin [Cleaved into: Acrosin light chain; Acrosin heavy chain]
AGRN	NA	Serine protease inhibitor dipetalogastin (Dipetalin) (Fragment)
AGRN	NA	Serine protease inhibitor dipetalogastin (Dipetalin) (Fragment)
AGRN	NA	Serine protease inhibitor dipetalogastin (Dipetalin) (Fragment)
AGRN	NA	Serine protease inhibitor dipetalogastin (Dipetalin) (Fragment)
AKT3	AKT3	RAC-gamma serine/threonine-protein kinase (Protein kinase Akt-3) (STK-2)
BMP2	BMP2	Bone morphogenetic protein 2
CAT	Cat	Catalase
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
FBLN1	fbl-1	Fibulin-1
FBLN5	Fbln5	Fibulin-5 (Developmental arteries and neural crest EGF-like protein)
FBN1	Fbn1	Fibrillin-1 [Cleaved into: Asprosin]
FBN1	FBN1	Fibrillin-1 [Cleaved into: Asprosin]
GRN	Grn	Progranulin (Acrogranin) [Cleaved into: Paragranulin; Granulin-1 (Granulin G); Granulin-2 (Granulin F); Granulin-3 (Epithelin-2) (Granulin B); Granulin-4 (Epithelin-1) (Granulin A); Granulin-5 (Granulin C); Granulin-6 (Granulin D); Granulin-7 (Granulin E)]
GRN	GRN	Progranulin (Acrogranin) (Glycoprotein of 88 Kda) (PC cell-derived growth factor) [Cleaved into: Paragranulin; Granulin-1 (Granulin G); Granulin-2 (Granulin F); Granulin-3 (Epithelin-2) (Granulin B); Granulin-4 (Epithelin-1) (Granulin A); Granulin-5 (Granulin C); Granulin-6 (Granulin D); Granulin-7 (Granulin E)]
HGFAC	HGFAC	Hepatocyte growth factor activator [Cleaved into: Hepatocyte growth factor activator short chain; Hepatocyte growth factor activator long chain]
HIP1	HIP1	Huntingtin-interacting protein 1
IRAK4	Irak4	Interleukin-1 receptor-associated kinase 4
LHB	NA	Thyrostimulin beta-5 subunit
NELL2	NELL2	Protein kinase C-binding protein NELL2
PHF24	Phf24	PHD finger protein 24
PLG	PLG	Plasminogen [Cleaved into: Plasmin heavy chain A; Activation peptide; Plasmin heavy chain A, short form; Plasmin light chain B]

Interactor	Gene	Protein
PLG	Plg	Plasminogen [Cleaved into: Plasmin heavy chain A; Activation peptide; Angiostatin; Plasmin heavy chain A, short form; Plasmin light chain B]
PTPN12	PTPN12	Tyrosine-protein phosphatase non-receptor type 12 (Protein-tyrosine phosphatase G1)
RASSF5	RASSF5	Ras association domain-containing protein 5 (Regulator for cell adhesion and polarization enriched in lymphoid tissues)
SORL1	SORL1	Sortilin-related receptor (Low-density lipoprotein receptor relative with 11 ligand-binding repeats) (Sorting protein-related receptor containing LDLR class A repeats)
SPRED2	Spred2	Sprouty-related, EVH1 domain-containing protein 2
STK25	STK25	Serine/threonine-protein kinase 25 (Sterile 20/oxidant stress-response kinase 1)
STK26	STK26	Serine/threonine-protein kinase 26 (MST3 and SOK1-related kinase) (Mammalian STE20-like protein kinase 4)
STK26	STK26	Serine/threonine-protein kinase 26 (MST3 and SOK1-related kinase) (Mammalian STE20-like protein kinase 4)
STK3	STK3	Serine/threonine-protein kinase 3 (Mammalian STE20-like protein kinase 2) [Cleaved into: Serine/threonine-protein kinase 3 36kDa subunit (MST2/N); Serine/threonine-protein kinase 3 20kDa subunit (MST2/C)]
STK3	STK3	Serine/threonine-protein kinase 3 (Mammalian STE20-like protein kinase 2) [Cleaved into: Serine/threonine-protein kinase 3 36kDa subunit (MST2/N); Serine/threonine-protein kinase 3 20kDa subunit (MST2/C)]
SYT1	SYT1	Synaptotagmin-1
THBS2	THBS2	Thrombospondin-2

Table A.19. Match between the overexpressed genes in the proboscis of *Hediste diversicolor* relative to the glands and human interactor. The interactor was the human protein homologous to the proteins up-regulated in the proboscis of *H. diversicolor*.

Interactor	Gene	Protein
AATK	Aatk	Serine/threonine-protein kinase LMTK1 (Apoptosis-associated tyrosine kinase) (Brain apoptosis-associated tyrosine kinase) (Lemur tyrosine kinase 1)
AATK	Aatk	Serine/threonine-protein kinase LMTK1 (Apoptosis-associated tyrosine kinase) (Brain apoptosis-associated tyrosine kinase) (Lemur tyrosine kinase 1)
ACTN2	NA	Alpha-actinin (F-actin cross-linking protein) (allergen Der f 24)
ADAMTS3	ADAMTS3	A disintegrin and metalloproteinase with thrombospondin motifs 3 (Procollagen II N-proteinase) (Procollagen II amino propeptide-processing enzyme)
AGRN	AGRN	Agrin [Cleaved into: Agrin N-terminal 110 kDa subunit; Agrin C-terminal 110 kDa subunit; Agrin C-terminal 90 kDa fragment (C90); Agrin C-terminal 22 kDa fragment (C22)]
AGRN	AGRN	Agrin [Cleaved into: Agrin N-terminal 110 kDa subunit; Agrin C-terminal 110 kDa subunit; Agrin C-terminal 90 kDa fragment (C90); Agrin C-terminal 22 kDa fragment (C22)]
AGRN	NA	Ovomucoid
AGRN	NA	Ovomucoid
AVP	NA	Conopressin/neurophysin [Cleaved into: Lys-conopressin G; Neurophysin]
CALM1	CALM1	Calmodulin
CALM1	NA	Calmodulin
CALM1	NA	Calmodulin
CALM1	NA	Calmodulin
CALM1	calm	Calmodulin
CALM1	CML12	Calmodulin-like protein 12 (Touch-induced calmodulin-related protein 3)
CALM1	CML12	Calmodulin-like protein 12 (Touch-induced calmodulin-related protein 3)
CALM1	NA	Calmodulin
CALM1	CML8	Calmodulin-like protein 8
CALML3	CML4	Calmodulin-like protein 4
CALML3	NA	Polcalcin Cup a 4 (Calcium-binding pollen allergen Cup a 4) (allergen Cup a 4)
CELA2B	Prss29	Serine protease 29 (Implantation serine proteinase 2) (Strypsin-2) (Tryptase-like proteinase)
CRY2	CRY2	Cryptochrome-2 (Blue light photoreceptor) (Protein SUPPRESSOR OF elf3 20)
DAPK1	DAPK1	Death-associated protein kinase 1
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
DMBT1	DMBT1	Deleted in malignant brain tumors 1 protein (Glycoprotein 340) (Hensin) (Salivary agglutinin) (Surfactant pulmonary-associated D-binding protein)
EGFL7	EGFL7	Epidermal growth factor-like protein 7 (NOTCH4-like protein) (Vascular endothelial statin)
FBLN1	FBLN1	Fibulin-1
FBLN2	FBLN2	Fibulin-2

Interactor	Gene	Protein
ITIH4	ITIH4	Inter-alpha-trypsin inhibitor heavy chain H4 (Plasma kallikrein sensitive glycoprotein 120) [Cleaved into: 70 kDa inter-alpha-trypsin inhibitor heavy chain H4; 35 kDa inter-alpha-trypsin inhibitor heavy chain H4]
LAMB1	LAMB1	Laminin subunit beta-1
LAMB1	LAMB1	Laminin subunit beta-1
LAMB1	LanB1	Laminin subunit beta-1
LAMB1	LanB1	Laminin subunit beta-1
NPY	NPY	Pro-neuropeptide Y [Cleaved into: Neuropeptide Y (Neuropeptide tyrosine) (NPY); C-flanking peptide of NPY (CPON)]
NTN1	NTN1	Netrin-1
NTN1	NTN1	Netrin-1
OTULIN	Otulin	Ubiquitin thioesterase otulin
OTULIN	Otulin	Ubiquitin thioesterase otulin
OTULIN	Otulin	Ubiquitin thioesterase otulin
OTULIN	Otulin	Ubiquitin thioesterase otulin
PI16	PI16	Peptidase inhibitor 16 (CD antigen CD364)
PI16	PI16	Peptidase inhibitor 16 (CD antigen CD364)
PKM	PKM	Pyruvate kinase PKM (Cytosolic thyroid hormone-binding protein) (Opa-interacting protein 3) (Pyruvate kinase muscle isozyme) (Tumor M2-PK)
PKM	PKM	Pyruvate kinase PKM (Cytosolic thyroid hormone-binding protein) (Opa-interacting protein 3) (Pyruvate kinase muscle isozyme) (Tumor M2-PK)
RASSF5	RASSF5	Ras association domain-containing protein 5 (Regulator for cell adhesion and polarization enriched in lymphoid tissues)
SMOC1	Smoc1	SPARC-related modular calcium-binding protein 1
SMOC1	Smoc1	SPARC-related modular calcium-binding protein 1
SMOC1	SMOC1	SPARC-related modular calcium-binding protein 1
SPARC	ost-1	SPARC (Basement-membrane protein 40) (Osteonectin) (Secreted protein acidic and rich in cysteine)
SPARC	ost-1	SPARC (Basement-membrane protein 40) (Osteonectin) (Secreted protein acidic and rich in cysteine)
SPON2	SPON2	Spondin-2 (Differentially expressed in cancerous and non-cancerous lung cells 1) (Mindin)
STK25	STK25	Serine/threonine-protein kinase 25 (Sterile 20/oxidant stress-response kinase 1)
TFF2	NA	Integumentary mucin C.1 (Fragment)
THBS2	THBS2	Thrombospondin-2
TPI1	NA	Triosephosphate isomerase (Allergen Scy p 8) (Methylglyoxal synthase)
TPM1	TM1	Tropomyosin (Tropomyosin, fast isoform) (Tm-Penm-fast)
WIF1	Wif1	Wnt inhibitory factor 1
YWHAZ	YWHAZ	14-3-3 protein zeta/delta (Factor activating exoenzyme S) (Protein kinase C inhibitor protein 1)

Table A.20. Enriched biological processes affected by the potential interactors from the proboscis of *Glycera alba* and their human targets. The interactors were the human homologs of the proteins up-regulated in the proboscis. The FDR-adjusted $p < 0.05$ was set as the cut-off. The GO enrichment analysis was performed in the Database for Annotation, Visualization and Integrated Discovery (DAVID).

	GO Terms	FE	FDR*p*	Human targets	Interactor
1	positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway [GO:1900740]	52.312	1.60E-10	BCL2, BID, CASP8, SFN, YWHAB, YWHAE, YWHAH, YWHAZ, YWHAQ	BAD
2	negative regulation of extrinsic apoptotic signaling pathway via death domain receptors [GO:1902042]	47.556	2.12E-10	FAS, FASLG, CASP8, DAPK1, NOS3, RAF1, TRADD, RIPK1, CFLAR	FADD
3	extrinsic apoptotic signaling pathway via death domain receptors [GO:0008625]	37.169	2.98E-08	FAS, FASLG, BCL2, BID, CASP10, DAPK1, TRADD	BAD, FADD
4	regulation of extrinsic apoptotic signaling pathway via death domain receptors [GO:1902041]	64.620	1.78E-07	FAS, FASLG, CASP8, TRADD, RIPK1, CFLAR	FADD
5	membrane organization [GO:0061024]	39.234	2.14E-07	SFN, RALA, YWHAB, YWHAE, YWHAH, YWHAZ, YWHAQ, TBC1D1	
6	activation of cysteine-type endopeptidase activity involved in apoptotic signaling pathway [GO:0097296]	72.431	2.08E-06	FAS, FASLG, CASP8, TRADD, RIPK1	FADD
7	negative regulation of apoptotic process [GO:0043066]	5.863	4.05E-06	AKT1, FAS, BCL2, BCL2L1, BCL2L2, EGFR, MYD88, RAF1, SLC9A1, TMBIM6, YWHAZ, ZNF16, LHX3, CFLAR, SPRY2, FKBP8	FAIM
8	positive regulation of I-kappaB kinase/NF-kappaB signaling [GO:0043123]	10.722	9.95E-06	FASLG, RHOA, CASP8, CASP10, MYD88, UBE2I, TRADD, RIPK1, CFLAR, TRIM62	FADD
9	regulation of apoptotic process [GO:0042981]	8.841	1.14E-05	ACTN2, FAS, BCL2L1, BCL2L2, BID, CASP8, CASP10, DAPK1, RAF1, CFLAR, MAGED1	FADD
10	death-inducing signaling complex assembly [GO:0071550]	98.084	1.14E-05	CASP8, RAF1, TRADD, RIPK1	FADD
11	extrinsic apoptotic signaling pathway [GO:0097191]	26.156	2.41E-05	FAS, FASLG, CASP8, TRADD, RIPK1	BAD, FADD
12	cellular response to mechanical stimulus [GO:0071260]	17.683	2.66E-05	AKT1, FAS, CASP8, EGFR, MYD88, SLC9A1	BAD, FADD
13	apoptotic signaling pathway [GO:0097190]	17.683	2.66E-05	FAS, FASLG, CASP8, CASP10, DAPK1, DAP3, RIPK1	FADD

	GO Terms	FE	FDR*p*	Human targets	Interactor
14	apoptotic process [GO:0006915]	4.705	4.06E-05	FAS, FASLG, BCL2, CASP8, CASP10, DAPK1, MYD88, RAF1, RALB, DAP3, TRADD, RIPK1, CFLAR, FKBP8	BAD, FADD, FAIM
15	activation of cysteine-type endopeptidase activity involved in apoptotic process [GO:0006919]	15.126	6.78E-05	FAS, FASLG, BID, CASP8, TRADD, RIPK1	BAD, FADD
16	extrinsic apoptotic signaling pathway in absence of ligand [GO:0097192]	27.694	1.52E-04	FAS, BCL2, BCL2L1, BCL2L2	BAD, FADD
17	necroptotic signaling pathway [GO:0097527]	104.623	3.15E-04	FAS, FASLG, RIPK1	FADD
18	release of cytochrome c from mitochondria [GO:0001836]	34.116	7.47E-04	BCL2, BCL2L1, BID, SFN	BAD
19	substantia nigra development [GO:0021762]	19.216	8.14E-04	RHOA, PLP1, YWHAE, YWHAH, YWHAQ	CALM1
20	viral process [GO:0016032]	5.774	1.08E-03	RHOA, ATP6V0C, H2AFX, RALA, TSC2, UBE2I, YWHAB, YWHAE, CFLAR, CALCOCO2, FKBP8	
21	signal transduction [GO:0007165]	2.703	5.74E-03	AKT1, FAS, FASLG, DAPK1, EGFR, SFN, GRB7, MYD88, RAF1, RALA, RALB, YWHAZ, TRADD, HGS, AKAP9, DAPP1, TRIM54, IQGAP3	FADD, AGRN, AGRN
22	positive regulation of apoptotic process [GO:0043065]	5.231	5.98E-03	AKT1, FAS, FASLG, BCL2L1, BID, SLC9A1, TRADD, RIPK1	BAD, FADD
23	positive regulation of cell growth [GO:0030307]	11.210	9.14E-03	AKT1, BCL2, EGFR, SFN, SDCBP, SLC9A1	
24	intrinsic apoptotic signaling pathway in response to DNA damage [GO:0008630]	16.695	1.00E-02	BCL2, BCL2L1, BCL2L2, SFN	BAD
25	regulation of nitric-oxide synthase activity [GO:0050999]	24.144	2.54E-02	AKT1, EGFR, NOS3	CALM1
26	protein heterooligomerization [GO:0051291]	11.712	3.62E-02	CASP8, YWHAB, TRADD, RIPK1	FADD
27	positive regulation of peptidyl-serine phosphorylation [GO:0033138]	11.210	3.96E-02	AKT1, BCL2, RAF1, AKAP9, SPRY2	
28	Ras protein signal transduction [GO:0007265]	11.210	3.96E-02	KRAS, RALA, RALB, SDCBP, IQGAP3	
29	positive regulation of intrinsic apoptotic signaling pathway [GO:2001244]	19.022	4.45E-02	BCL2, BCL2L1, BID	BAD

FDR*p* – False discovery rate adjusted *p*-value; FE – Fold enrichment

Table A.21. Enriched biological processes affected by the potential interactors from the skin of *Glycera alba* and their human targets. The interactors were the human homologs of the proteins up-regulated in the skin. The FDR-adjusted $p < 0.05$ was set as the cut-off. The GO enrichment analysis was performed in the Database for Annotation, Visualization and Integrated Discovery (DAVID).

	GO Terms	FE	FDR*p*	Human targets	Interactor
1	leukocyte migration [GO:0050900]	18.528	7.88E-04	GRB2, PIK3R1, PLCG1, TEK	ANGPT1, ANGPT2, MIF, MIF
2	JAK-STAT cascade [GO:0007259]	50.457	7.88E-04	STAT5A, STAT5B, SOCS1, SOCS3	JAK2
3	Tie signaling pathway [GO:0048014]	322.923	4.84E-03	TEK	ANGPT1, ANGPT2
4	viral process [GO:0016032]	8.640	4.84E-03	DAXX, E4F1, GRB2, PIK3R1, PLCG1, SUV39H1, VCAM1, BTRC	
5	glomerulus vasculature development [GO:0072012]	242.192	6.68E-03	TEK	ANGPT1, ANGPT2
6	cell proliferation [GO:0008283]	7.058	1.15E-02	E4F1, EGFR, ELN, MPL, GFI1B, PRMT5, KDM1A	MIF, MIF

FDR*p* – False discovery rate adjusted *p*-value; FE – Fold enrichment

Table A.22. Enriched biological processes affected by the potential interactors from the glands of *Hediste diversicolor* and their human targets.
The interactors were the human homologs of the proteins up-regulated in the glands. The FDR-adjusted $p < 0.05$ was set as the cut-off. The GO enrichment analysis was performed in the Database for Annotation, Visualization and Integrated Discovery (DAVID).

	GO Terms	FE	FDR*p*	Target	Interactor
1	peptide cross-linking [GO:0018149]	21.714	3.80E-12	FN1, LCE2B, CRCT1, LCE4A, LCE5A, LCE1A, LCE1B, LCE1C, LCE1D, LCE1E, LCE1F, LCE2D, LCE3A, LCE3C, LCE3E	
2	keratinocyte differentiation [GO:0030216]	16.190	3.80E-12	CASP3, STK4, LCE2B, CRCT1, SAV1, LCE4A, LCE5A, LCE1A, LCE1B, LCE1C, LCE1D, LCE1E, LCE1F, LCE2D, LCE3A, LCE3C, LCE3E	
3	keratinization [GO:0031424]	19.603	7.75E-10	LCE2B, LCE4A, LCE5A, LCE1A, LCE1B, LCE1C, LCE1D, LCE1E, LCE1F, LCE2D, LCE3A, LCE3C, LCE3E	
4	hippo signaling [GO:0035329]	18.765	5.52E-04	CASP3, STK4, AMOTL2, SAV1, MOB1B, AMOT	STK3, STK3
5	signal transduction by protein phosphorylation [GO:0023014]	13.160	6.33E-04	BMPR1A, BMPR1B, BMPR2, STK4, CAB39	STK3, STK3, STK25, STK26, STK26
6	protein stabilization [GO:0050821]	6.386	7.07E-04	HSPA1A, HSPA1B, HSPD1, STK4, RASSF2, CDC37, RASSF1, PDCD10, NLK, SAV1	HIP1, STK3, STK3
7	positive regulation of I-kappaB kinase/NF-kappaB signaling [GO:0043123]	5.395	3.07E-03	ECM1, IRAK1, MYD88, REL, TNFRSF1A, IKBKG, LITAF, ZDHHC17, PELI2, PELI1, TIRAP	IRAK4
8	signal transduction [GO:0007165]	2.182	4.36E-03	ECM1, EGFR, ERBB2, GRB14, HRAS, IRAK1, MYD88, PDGFRB, PRKCZ, RAP2B, CXCL5, SHC1, STK4, THBD, TNFRSF1A, TRAF1, TRAF2, BTRC, PSTPIP1, HGS, LITAF, ZDHHC17, RPS6KA6, SAV1, RASSF4, LINGO1, TIRAP	GRN, GRN, LHB, SORL1, STK3, STK3, AKT3, STK25, IRAK4, AGRN, AGRN, AGRN, AGRN
9	cellular response to BMP stimulus [GO:0071773]	14.476	8.45E-03	BMPR1A, BMPR1B, BMPR2, SMAD4, SPINT1	BMP2
10	response to hydrogen peroxide [GO:0042542]	9.934	1.02E-02	CASP3, HSPD1, PDGFRB, PDCD10	CAT, STK25, STK26, STK26

	GO Terms	FE	FDR*p*	Target	Interactor
11	extracellular matrix organization [GO:0030198]	4.431	1.17E-02	APP, ELN, FN1, HSPG2, LOX, LOXL1, MFAP2, SPINT1	FBLN1, FBN1, FBN1, FBLN5, AGRN, AGRN, AGRN, AGRN
12	positive regulation of NF-kappaB transcription factor activity [GO:0051092]	5.442	1.17E-02	HSPA1A, HSPA1B, IRAK1, MYD88, PRKCZ, TRAF1, TRAF2, IKBKG, TIRAP	CAT
13	positive regulation of MAP kinase activity [GO:0043406]	8.587	1.79E-02	EGFR, ERBB2, HRAS, IRAK1, PDGFRB, MAGED1, PDCD10	
14	protein phosphorylation [GO:0006468]	2.857	2.13E-02	APP, BMPR1A, BMPR1B, ERBB2, IRAK1, SMAD2, PRKCZ, STK4, RASSF2, RPS6KA6, NLK, TRIB3, HIPK1	BMP2, STK3, STK3, AKT3, STK25, STK26, STK26
15	cellular protein metabolic process [GO:0044267]	5.520	2.26E-02	APP, HSPG2, MMP2, SFTPD, GGA2, GGA1, SFTPA1	DMBT1, DMBT1, DMBT1, DMBT1, DMBT1, PLG, PLG

FDR*p* – False discovery rate adjusted *p*-value; FE – Fold enrichment

Table A.23. Enriched biological processes affected by the potential interactors from the proboscis of *Hediste diversicolor* and their human targets. The interactors were the human homologs of the proteins up-regulated in the proboscis. The FDR-adjusted $p < 0.05$ was set as the cut-off. The GO enrichment analysis was performed in the Database for Annotation, Visualization and Integrated Discovery (DAVID).

	GO Terms	FE	FDR*p*	Target	Interactor
1	regulation of gene silencing [GO:0060968]	48.771	2.58E-11	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
2	DNA replication-dependent nucleosome assembly [GO:0006335]	18.442	1.97E-07	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F, CHAF1A	
3	cellular protein metabolic process [GO:0044267]	7.729	4.00E-07	APP, FGA, MMP2, SFTPD, UBC, HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F, SFTPA1	DMBT1, DMBT1
4	telomere organization [GO:0032200]	19.870	4.25E-07	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
5	chromatin silencing at rDNA [GO:0000183]	14.500	7.66E-06	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
6	negative regulation of gene expression, epigenetic [GO:0045814]	11.803	7.79E-06	TRIM27, HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
7	regulation of tumor necrosis factor-mediated signaling pathway [GO:0010803]	16.095	1.57E-05	TNFAIP3, TRAF1, TRAF2, UBC, IKBKG, MADD, RNF31, HIPK1	OTULIN, OTULIN, OTULIN, OTULIN
8	protein heterotetramerization [GO:0051290]	12.773	1.58E-05	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	

	GO Terms	FE	FDR*p*	Target	Interactor
9	MAPK cascade [GO:0000165]	4.095	1.10E-04	ARAF, CALM2, CALM3, EGFR, HRAS, KRAS, MARK3, MAP3K5, MOS, PPP5C, MAPK1, MAPK3, PSMA1, PSMC5, RAF1, UBC, AKAP9, LRRK2	ACTN2, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
10	muscle contraction [GO:0006936]	6.518	1.45E-04	ACTN3, CALM2, CALM3, CRYAB, MYLK, TPM4, TTN, DYSF, MYOM2, MYOT, TRIM63	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, TPM1
11	response to calcium ion [GO:0051592]	9.250	2.12E-04	AANAT, BAD, CALM2, CALM3, EGFR, FGA, TTN, PDCD6	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, SPARC, SPARC
12	positive regulation of gene expression, epigenetic [GO:0045815]	8.653	3.46E-04	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
13	positive regulation of protein serine/threonine kinase activity [GO:0071902]	12.263	4.88E-04	CALM2, CALM3, RALB, STK3, STK4, CDK5R1, PDCD10	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1

	GO Terms	FE	FDR*p*	Target	Interactor
14	cell-cell adhesion [GO:0098609]	3.761	4.88E-04	CBL, MARK2, ENO1, GOLGA2, SDCBP, YWHAE, HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F, PDLIM1	PKM, PKM, YWHAZ
15	peptide cross-linking [GO:0018149]	9.657	4.88E-04	COL3A1, FN1, LCE3D, LCE5A, LCE1A, LCE1B, LCE1C, LCE1F, LCE3A	
16	protein phosphorylation [GO:0006468]	2.941	6.33E-04	APP, CDC25B, CTBP1, MARK2, ILK, LIMK1, MAP3K5, MYLK, CDK16, PRKAB2, PRKCE, MAPK1, MAPK3, RAF1, STK3, STK4, PICK1, RASSF2, MAP3K20, LRRK2, HIPK1, SBK3	DAPK1, AATK, AATK, STK25
17	hippo signaling [GO:0035329]	13.909	1.01E-03	DVL2, STK3, STK4, YWHAE, YAP1, WWTR1, AMOTL2	
18	blood coagulation [GO:0007596]	4.374	1.01E-03	FGA, GP1BA, GP1BB, HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F, BLOC1S6, AK3	
19	G2/M transition of mitotic cell cycle [GO:0000086]	5.091	1.01E-03	CALM2, CALM3, CDC25A, CDC25B, TPD52L1, UBC, WEE1, YWHAE, BTRC, CCP110, AKAP9, CEP70	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
20	circadian regulation of gene expression [GO:0032922]	8.471	1.01E-03	PER1, PPARA, PPP1CA, PPP1CC, PER2, USP2, MAGED1, PRMT5	CRY2
21	keratinocyte differentiation [GO:0030216]	7.059	1.11E-03	NOTCH1, STK4, YAP1, LCE3D, LCE5A, LCE1A, LCE1B, LCE1C, LCE1F, LCE3A	
22	wound healing [GO:0042060]	6.706	1.61E-03	COL3A1, EGFR, FN1, TNC, MAP3K5, PPARA, RAF1	SPARC, SPARC, TFF2, TPM1

	GO Terms	FE	FDR*p*	Target	Interactor
23	negative regulation of extrinsic apoptotic signaling pathway via death domain receptors [GO:1902042]	11.380	2.56E-03	FGA, NOS3, RAF1, TNFAIP3, TRAF2, FADD	DAPK1
24	regulation of circadian rhythm [GO:0042752]	8.759	2.71E-03	PER1, PPARA, PPP1CA, PPP1CC, PER2, BTRC, MAGED1	CRY2
25	positive regulation of protein autophosphorylation [GO:0031954]	15.328	2.83E-03	CALM2, CALM3, RAP2A, RAP2B, RASSF2	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
26	nucleosome assembly [GO:0006334]	4.959	6.01E-03	H2AFX, HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
27	response to hypoxia [GO:0001666]	4.055	7.02E-03	ANG, BAD, CRYAB, KCNA5, SMAD4, MMP2, NOS2, PPARA, RAF1, TH, VCAM1, PDLIM1	PKM, PKM
28	regulation of nitric-oxide synthase activity [GO:0050999]	12.380	7.56E-03	CALM2, CALM3, EGFR, GCH1, NOS3	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
29	positive regulation of MAP kinase activity [GO:0043406]	7.274	7.60E-03	EGFR, HRAS, ILK, KRAS, TPD52L1, MAGED1, PDCD10, LRRK2	
30	positive regulation of epithelial cell proliferation [GO:0050679]	7.153	8.18E-03	BAD, EGFR, HRAS, LAMC1, NOTCH1, SCN5A, NR4A3	LAMB1, LAMB1, LAMB1, LAMB1

	GO Terms	FE	FDR*p*	Target	Interactor
31	platelet degranulation [GO:0002576]	5.209	8.53E-03	APP, CALM2, CALM3, FGA, FN1, TTN	ACTN2, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, ITIH4, SPARC, SPARC
32	positive regulation of cell migration [GO:0030335]	3.790	1.12E-02	EGFR, GRB7, HRAS, ILK, MYLK, NOTCH1, MAPK1, SDCBP, SNAI1, PDCD10, COL18A1, WASHC1	LAMB1, LAMB1, LAMB1, LAMB1
33	positive regulation of axon extension [GO:0045773]	10.730	1.31E-02	FN1, ILK, LIMK1, MAPT, DISC1	NTN1, NTN1
34	gene silencing by RNA [GO:0031047]	4.833	1.38E-02	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	
35	keratinization [GO:0031424]	7.824	1.47E-02	LCE3D, LCE5A, LCE1A, LCE1B, LCE1C, LCE1F, LCE3A	
36	angiogenesis [GO:0001525]	3.368	1.61E-02	ANG, COL8A1, FN1, MEOX2, MMP2, MYH9, NOS3, NOV, WNT7A, PDCD6, PDCD10, COL18A1, UNC5B	EGFL7
37	response to corticosterone [GO:0051412]	14.902	1.61E-02	AANAT, CALM2, CALM3, TH	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
38	platelet activation [GO:0030168]	4.665	1.61E-02	COL3A1, FGA, GP1BA, GP1BB, PRKCE, MAPK1, MAPK3, RAF1, RAP2B	YWHAZ

	GO Terms	FE	FDR*p*	Target	Interactor
39	extracellular matrix organization [GO:0030198]	3.558	1.65E-02	APP, COL3A1, COL8A1, FGA, FN1, TNC, LAMC1, VCAM1, COL18A1	FBLN1, LAMB1, LAMB1, LAMB1, SPARC, SPARC, AGRN, AGRN, AGRN
40	microtubule cytoskeleton organization [GO:0000226]	6.045	1.78E-02	MARK2, MAPT, MARK3, MID1, ATXN7, WEE1, DISC1, MAP1S	
41	entrainment of circadian clock by photoperiod [GO:0043153]	13.412	2.26E-02	PER1, PPP1CA, PPP1CC, USP2	CRY2
42	viral process [GO:0016032]	2.871	2.42E-02	ATP6V0C, H2AFX, MDM2, MAP3K5, MAPK1, MAPK3, RALA, SP1, SP100, TSC2, VCAM1, YWHAE, BTRC, CALCOCO2, FKBP8	FBLN1
43	Fc-epsilon receptor signaling pathway [GO:0038095]	3.617	2.42E-02	CALM2, CALM3, HRAS, KRAS, MAPK1, MAPK3, PSMA1, PSMC5, UBC, IKBKG, BTRC	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
44	regulation of cell cycle [GO:0051726]	4.326	2.42E-02	CDC25A, CTBP1, CDK16, TSC2, WEE1, MADD, PER2, BTRC, COPS5, CCNDBP1	
45	positive regulation of protein dephosphorylation [GO:0035307]	12.773	2.50E-02	CALM2, CALM3, PIN1, NSMF	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
46	cerebral cortex development [GO:0021987]	6.706	2.62E-02	BAD, COL3A1, H2AFX, PAX5, TH, YWHAE	NPY
47	positive regulation of protein phosphorylation [GO:0001934]	4.224	2.66E-02	DVL2, EGFR, HRAS, KRAS, PIN1, MAPK3, RALB, RAP2A, STK4, LRRK2	

	GO Terms	FE	FDR*p*	Target	Interactor
48	negative regulation of protein binding [GO:0032091]	6.588	2.66E-02	GOLGA2, PIN1, PPARA, PPP1CA, RALB, TMBIM6, LRRK2	
49	muscle filament sliding [GO:0030049]	8.471	2.66E-02	ACTN3, TNNT1, TPM4, TTN	ACTN2, TPM1
50	positive regulation of nitric-oxide synthase activity [GO:0051000]	12.193	2.66E-02	CALM2, CALM3, GCH1, KRAS	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
51	regulation of cytokinesis [GO:0032465]	12.193	2.66E-02	CALM2, CALM3, PIN1, CCP110	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
52	stimulatory C-type lectin receptor signaling pathway [GO:0002223]	4.598	2.97E-02	HRAS, KRAS, PSMA1, PSMC5, RAF1, UBC, IKBKG, BTRC, CLEC7A	
53	response to ethanol [GO:0045471]	4.598	2.97E-02	BAD, CBL, HTR1B, TNC, TBXA2R, TH, VCAM1	AVP, SPARC, SPARC
54	axon guidance [GO:0007411]	3.712	3.03E-02	HRAS, KRAS, SMAD4, MAPK1, MAPK3, NR4A3, CDK5R1, ZNF280A, DRAXIN	NTN1, NTN1, SPON2
55	negative regulation of neuron death [GO:1901215]	8.047	3.11E-02	PPARA, PPP5C, REL, TRAF2, IKBKG, LRRK2	
56	platelet aggregation [GO:0070527]	7.851	3.35E-02	FGA, GP1BA, HSPB1, ILK, MYH9, RAP2B	
57	signal transduction [GO:0007165]	1.756	3.35E-02	EGFR, FGA, GRB7, HRAS, IL9R, LIMK1, PDE9A, PRKAB2, PRKCE, MAPK1, RAF1, RALA, RALB, RAP2B, CXCL5, STK3, STK4, TRAF1, TRAF2, FADD, BTRC, HGS, AKAP9, APPL1, TRIM54, TRIM63, LINGO1, IQGAP3, UNC5B, RASSF10	AVP, DAPK1, SPARC, SPARC, YWHAZ, STK25, WIF1, SMOC1, SMOC1, SMOC1, AGRN, AGRN, AGRN, AGRN
58	neuron death [GO:0070997]	19.509	3.35E-02	CBL, MEOX2, SLC9A1, LRRK2	

	GO Terms	FE	FDR*p*	Target	Interactor
59	endocardium development [GO:0003157]	53.649	3.51E-02	NOTCH1, STK3, STK4	
60	positive regulation of cyclic nucleotide metabolic process [GO:0030801]	53.649	3.51E-02	CALM2, CALM3	CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1, CALM1
61	positive regulation of cell proliferation [GO:0008284]	2.303	4.06E-02	CDC25B, EGFR, FN1, HRAS, TNC, ILK, KRAS, MDM2, NOTCH1, MAPK1, CXCL5, SDCBP, HDAC4, YAP1, PDCD10, WWTR1, COL18A1, HIPK1	AVP, NTN1, NTN1
62	positive regulation of JNK cascade [GO:0046330]	5.778	4.30E-02	HRAS, MAP3K5, SDCBP, STK3, TPD52L1, WNT7A, RASSF2	

FDR*p* – False discovery rate adjusted *p*-value; FE – Fold enrichment

A.2 R Script

Please refer to https://drive.google.com/file/d/14Wrh-_rvy8ilhGtMAOgSHpxVhdR87fs7/view?usp=sharing for the complete R script and the command lines used for the transcriptome assembly and annotation.

```
      ##Load a R package
      library(tximport)
      library(limma)
      library(edgeR)
      library(seqinr)
      library(car)

      ##Differentially-expressed analysis
      degGenes<-function(Kallistofolder, Trinityfile, species, pathDEG, pathDEGaux,
sampleName, Level, Organ, numberOfIndividuals){
        #Experimental design
        dataMatrix<-as.data.frame(
          cbind(
            sampleName,
            Level,
            Organ,
            numberOfIndiviuals
          )
        )
        write.table(dataMatrix, paste(pathDEG, species, "dataMatrix.csv", sep = ""), sep =
";", col.names = NA)
        #Data table
        ##Importing the abundance results and estimate the counts
        ##Merging the counts with the respective cDNA sequence
        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")
        print(head(Tsv$counts))
        print(dim(Tsv$counts))
        print(colnames(Tsv$counts))
        Fa<-read.fasta(Trinityfile, as.string = TRUE)
        print(head(Fa))
        TxID<-getName(Fa)
        TxSEQ<-unlist(getSequence(Fa, as.string=TRUE))
        TxSEQ<-as.data.frame(cbind(TxID, TxSEQ))
        print(head(TxSEQ))
        colnames(TxSEQ)[1]<-"ID"
        print(nrow(TxSEQ))
```

```
        Counts<-Tsv$counts
        Counts<-cbind(as.data.frame(row.names(Counts)), Counts)
        colnames(Counts)[1]<-"ID"
        Full<-merge(Counts, TxSEQ, by="ID")
        colnames(Full)<-c("ID", sampleName, "sequence")
        write.table(Full, paste(pathDEGaux, species, "Full.csv", sep = ""), sep = ";",
col.names = NA)
        #Statistics
        ##Create a DGEList object from a table of counts
        Data<-DGEList(counts = Full[,2:7], group = Level, genes = Full[,1])
        ##Calculating the normalization factors
        Data<-calcNormFactors(Data)
        ##Creating the generalized linear models for comparing the expression levels from two
organs
        Design<-model.matrix(~0+Level, data = Data$samples)
        colnames(Design)<-levels(Data$samples$group)
        Data<-estimateDisp(Data, Design)
        fit<-glmFit(Data, Design)
        ##Estimating the relative expression
        assign(paste("lrt", colnames(Design)[1], sep = ""),
relativeExpression(colnames(Design)[1], colnames(Design)[2], Design, fit))
        assign(paste("lrt", colnames(Design)[2], sep = ""),
relativeExpression(colnames(Design)[2], colnames(Design)[1], Design, fit))
        print(c("organs:", colnames(Design)))
        inputOrgan1<-readline("Please enter the name of organ of interest: ")
        inputOrgan2<-readline("Please enter the name of reference organ: ")
        Results<-cbind(Full$ID, get(paste("lrt", inputOrgan1, sep = ""))$table,
get(paste("lrt", inputOrgan2, sep = ""))$table)
        colnamesResults<-c("log2FC", "log2CPM","LR","FDRp")
        colnamesResults<-rep(colnamesResults, times = 2)
        organ<-rep(c(inputOrgan1, inputOrgan2), each = 4)
        colnamesResults<-paste(colnamesResults, organ, sep = "_")
        colnames(Results)<-c("ID", colnamesResults)
        save(Full, Results, file = paste(pathDEGaux, species, "FullResults.RData", sep = ""))
        save(Data, file = paste(pathDEG, species, "Data.RData", sep = ""))
        ##decideTest identifies the differentially-expressed genes according to the cut-offs
        ##0 means no differential expression
        ##1 means overexpressed in the organ of interest relative to the reference organ
        ##-1 means underexpressed in the organ of interest relative to the reference organ
        expressionTable<-decideTests(Results[,grepl("FDRp", colnames(Results))], coefficients
= Results[,grepl("log2FC", colnames(Results))], lfc = 1.5, adjust.method = "fdr")
        colnames(expressionTable)<-c(inputOrgan1, inputOrgan2)
        head(expressionTable)
        nrow(expressionTable)
        DEG<-cbind(Results, expressionTable)
        DEG0<-subset(DEG, DEG[,inputOrgan1] == 0)
        assign(paste("DEG", inputOrgan1, sep = ""), subset(DEG, DEG[,inputOrgan1] == 1))
        assign(paste("DEG", inputOrgan2, sep = ""), subset(DEG, DEG[,inputOrgan1] == -1))
        degTable<-DEG[which(abs(DEG[inputOrgan1]) == 1 | abs(DEG[inputOrgan2]) == 1),]
```

```
        print(head(degTable))
        nrow(degTable)
        write.table(degTable, paste(pathDEG, species, "degTable.csv", sep = ""), sep = ";",
col.names = NA)
        write.table(DEG, paste(pathDEGaux, species, "DEG.csv", sep = ""), sep=";",col.names =
NA)
        save(list = c(paste("lrt", colnames(Design)[1], sep = ""), paste("lrt",
colnames(Design)[2], sep = "")), file = paste(pathDEGaux, species, "lrt.RData", sep = ""))
        save(degTable, file = paste(pathDEG, species, "degTable.RData", sep = ""))
        save(list = c(paste("DEG", sep = ""), paste("DEG0", sep = ""), paste("DEG",
inputOrgan1, sep = ""), paste("DEG", inputOrgan2, sep = "")), file = paste(pathDEG, species,
"DEG.RData", sep = ""))
        #Annotation
        write.fasta(sequences = as.list(Full$sequence), names = Full$ID, file.out =
paste(pathDEG, species, "Sequences.fasta", sep = ""))
}

      relativeExpression<-function(organ1, organ2, Design, fit){
        contrast<-paste(organ1, "-", organ2, sep = "")
        organ1vorgan2<-makeContrasts(contrasts = contrast, levels = Design)
        print(organ1vorgan2)
        lrt<-glmLRT(fit, contrast = organ1vorgan2)
        print(topTags(lrt))
        return(lrt)
}

      ##Enrichment analysis
      ##Fisher analysis for the GO Terms and Domains
      ##The data in analysis is the TOP10 GO Terms/Domains together with GO Terms/Domains of
Interest, respectively
      ##The p-value was adjusted by FDR method
      Fisher<-function(path, pathaux, species, OrganOfInterest, ReferenceOrgan,
totalOrganOfInterest, totalReferenceOrgan, FisherOrganOfInterest, FisherReferenceOrgan, Terms,
Analysis, Category, colour){
        if(Analysis == "GOTerms"){
          FisherAux<-merge(FisherOrganOfInterest, FisherReferenceOrgan, by = "GOTerms", all =
TRUE)
          rownames(FisherAux)<-FisherAux$GOTerms
        }
        else if(Analysis == "Domains"){
          FisherAux<-merge(FisherOrganOfInterest, FisherReferenceOrgan, by = "Domains", all =
TRUE)
          rownames(FisherAux)<-FisherAux$Domains
        }
        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 = "", pathaux, "FisherAux", Category, ".RData"))
```

```
        save(FisherAux, file = paste(sep = "", pathaux, "FisherAux", Category, ".RData"))
        if(Analysis == "GOTerms"){
          rownames(FisherAux)<-sapply(rownames(FisherAux), function(x) unlist(strsplit(x," [",
fixed = TRUE))[1])
          Analysis<-Category
        }
        FisherAux<-FisherAux[rownames(FisherAux) %in% Terms,]
        print(Terms)
        print(rownames(FisherAux))
        Fisher<-data.frame(matrix(data = NA, nrow = nrow(FisherAux), ncol = 5, dimnames =
list(NULL, c(paste(Analysis, sep = ""), 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 = "", pathaux, species, Category, "Fisher.RData"))
        FisherFormatted(path, species, Category, Fisher)
        FisherGraph(Fisher, Category, Analysis, colour, species)
}
      ##Please refer to:
https://drive.google.com/file/d/14Wrh- rvy8ilhGtMAOgSHpxVhdR87fs7/view?usp=sharing for the
code of FisherFormatted and FisherGraph functions

      ##RT-qPCR: Statistics
      ##Normality test
      ##Test for homogeneity of variance
      NormalHomogeneity<-function(file){
        ##The first column of data has the name of the organ, while the remaining columns have
the expression values obtained from the -2ddt method
        data<-read.table(file, sep=";", header=TRUE, row.names = 1)
        Variables<-colnames(data[,2:ncol(data)])
        nFactors<-nlevels(as.factor(data$Organ))
        Factors<-levels(as.factor(data$Organ))
        for(i in 2:ncol(data)){
          for(j in 1:nFactors){
            print(paste(sep=" ","Variable:",colnames(data)[i],";","Factor:",Factors[j]))
            print(shapiro.test(data[data$Organ == Factors[j],i]))
          }
          print(paste(sep=" ","Variable:",colnames(data)[i]))
          print(leveneTest(data[,i]~as.factor(data$Organ),center=median))
        }
      }
```

```
      ##Student's t-test
      StudentTtest<-function(data, j, OrganOfInterest, ReferenceOrgan){
        print("Student's t-test (parametric)")
        print(paste(sep=" ","Variable: ",colnames(data)[j]))
        Studentaux<-t.test(data[data$Organ == ReferenceOrgan,j], data[data$Organ ==
OrganOfInterest,j])
        print(Studentaux)
        return(Studentaux$p.value)
}

      ##Interactome-directed analysis
      ##Retrieving the domains that have a specific term in their designation/definition
      ##The shortlist is to be hand-curated
      domainsInteractome<-c("binding", "inhibitor", "inhibitory", "antibacterial",
                            "antimicrobial", "virus", "hormone", "orexin", "secreted",
                            "regulatory", "ligand", "signal", "toxin", "lectin",
                            "antigen", "inflammation", "interleukin", "death",
                            "apoptosis", "necrosis", "neurotransmitter", "bacterial",
                            "bacteria", "peptide", "interferon", "invasin", "immune",
                            "regulator", "virulence", "rifin", "dickkopf", "antagonist",
                            "antibiotic", "mucin", "stress", "rap", "adam", "astacin",
                            "cap", "cub", "kazal", "kringle", "kunitz", "m12b",
                            "reprolysin", "shk", "plasmod", "egf", "trypsin", "sprouty",
                            "spry", "lipocalin")
      domainsInterest<-data.frame()
      for(i in 1:length(species)){
        load(paste(pathPfam, species[i], "degpfam.RData", sep = ""))
        for(l in 1:nrow(organ)){
          degpfam<-get(paste("degpfam", organ[l, species[i]], sep = ""))
          for(j in 1:length(domainsInteractome)){
            domainsInterestaux1<-subset(degpfam[,c("Domain", "DescriptionTarget")],
sapply(degpfam$Domain, function(x) grepl(domainsInteractome[j], tolower(x))))
            domainsInterestaux2<-subset(degpfam[,c("Domain", "DescriptionTarget")],
sapply(degpfam$DescriptionTarget, function(x) grepl(domainsInteractome[j], tolower(x))))
            domainsInterest<-rbind(domainsInterest, domainsInterestaux1, domainsInterestaux2)
          }
        }
      }
      domainsInterest<-unique(domainsInterest)
      write.table(domainsInterest, paste(pathInteractome, "domainsInterest.csv", sep = ""),
sep=";", col.names=NA)
```