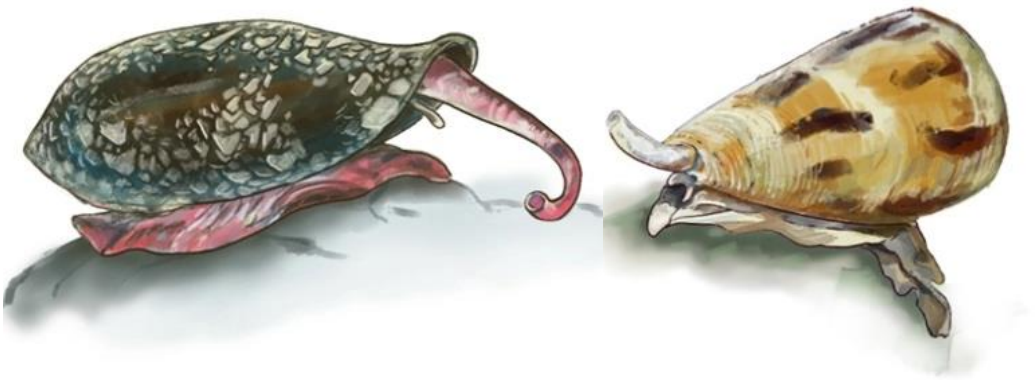


Deciphering the transcriptomics of the *Conus* species' natural venoms

José Manuel Ramos Morim

Mestrado em Aplicações em Biotecnologia e Biologia Sintética

Departamento de Química e Bioquímica e Departamento de Biologia

2021-2022

Orientador

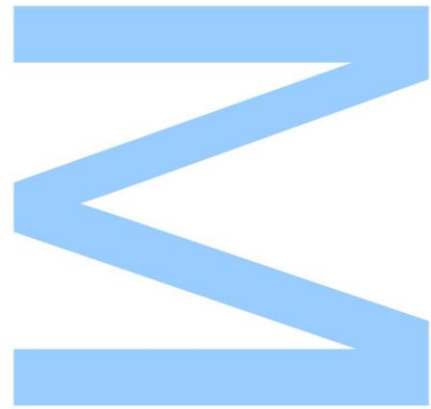
Prof. Agostinho Antunes, FCUP

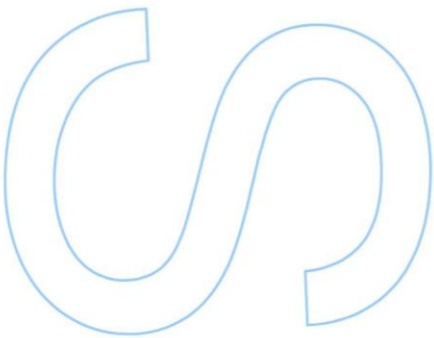
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

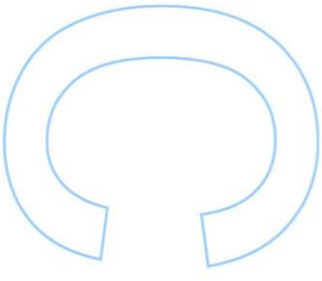
Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____







"Farà male il dubio di non essere nessuno, sarai qualcuno se resterai diverso dagli altri."

– Måneskin

Sworn Statement

I, José Manuel Ramos Morim, enrolled in the Master Degree Applications in Biotechnology and Synthetic Biology at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

José Manuel Ramos Morim

29/09/2022

Acknowledgments

My effort for this Master's degree began well before even entering it. Every day I remember the strife to get the grades back in the bachelors, the endeavour of being accepted for matriculation and the amazing feeling of accomplishment after I succeeded. I recall all the labour for the completion of every task, of every exam, of every objective in it. From the very first day I felt challenged. Often, I was pushed to my limit. After all my work, I finally pose to end this chapter. Completing this Master's program will be the greatest academic achievement of my life up until this point.

And I did not reach this point alone. Along my path, there were those who stood by my side when it most mattered, did not leave then and are here now. They supported me, guided me, helped me throughout the way. I do not forget it, will not forget. Anything recorded by writing is permanent, immutable, and therefore immortal. In this way I want to dedicate my achievements to the people I most profoundly trust and admire.

First, I would like to thank Professor Agostinho Antunes for accepting me in this project, giving me the opportunity to embrace this dissertation in a theme I was fascinated about. In a more human aspect, I want to thank for the sympathy and concern the Professor always demonstrated towards me and for the focus in directing my work.

Next, I want to thank my laboratorial colleague and friend, Yihe Zhao. Nearly everything I learned or did for this dissertation was as a member of a team I formed with him leading. Thank you for always helping, tutoring, and directing me to where it was best.

To Professor Paula Gomes, I want to thank for the opportunity of entering the Masters, as without her advice I would not have made it.

To my father, my mother, my sister and my brother I want to thank for always being there as a true family, supporting and backing my efforts, educating me as a person with the values I will forever hold and be true to.

To my three truest friends, I want to dedicate a very special thank you. In everything we have been and will continue to go through together, I will always have you as true brothers.

To my dearest, truest love, Sofia, in my life you represent love, passion, virtue, justice and balance. You have rendered my life more beautiful than I ever imagined it could be. Thank you for everything.

At last, I would also like to express an important acknowledgement to everyone from FCUP for helping me whenever I needed and supporting the project. This work was supported in part by the Portuguese Foundation for Science and Technology (FCT) project PTDC/CTA-AMB/31774/2017 (POCI-01-0145-FEDER/031774/2017).

Resumo

Os gastrópodes *Conus* são espécies icónicas e conhecidas há séculos principalmente pelas suas bonitas conchas cheias de padrões coloridos, as quais são muito procuradas para troféus e efeitos de decoração. Desde tempos ancestrais, as conchas, que são prezados objetos de joalharia, são também um autêntico tesouro para muitos colecionadores, para os quais ainda existe um grande mercado. Apenas recentemente especial atenção foi dedicada às capacidades do veneno destas espécies. Além de terem conchas muito belas estes gastrópodes marinhos produzem um veneno predatório com imensos compostos complexos, cada qual com extrema afinidade para um grupo específico de recetores. Estes compostos atuam de formas muito diversas, desde alívio de dor a paralisia, sendo, portanto, muito relevantes para várias áreas de aplicação biomédica se mais estudos lhes fossem dedicados. Nesse sentido, o objetivo desta dissertação é contribuir para decifrar a diversificação molecular do veneno predatório natural destes gastrópodes marinhos ao estudar todos os transcriptomas disponíveis das espécies do género *Conus* utilizando uma abordagem de genómica comparativa e bioinformática.

Nesta dissertação, foi desenvolvida uma metodologia para descodificar o contexto genómico e as relações entre as proteínas do veneno destes predadores marinhos. Após a recolha e estudo de toda a informação transcriptómica fidedigna disponível, a montagem dos transcriptomas e a anotação funcional das proteínas revelaram a função dos genes expressos. Análises e comparações baseadas nos resultados obtidos permitiram retirar várias conclusões. Resumidamente, neste trabalho é reportado (1) o número e função dos genes partilhados apenas expressos na glândula do veneno das 20 espécies de *Conus* analisadas, (2) uma relação entre o tamanho da montagem do transcriptoma e o número de genes únicos encontrados, (3) evidência de relações simbióticas de microrganismos na glândula venenosa e (4) uma congruência com estudos prévios relativamente aos níveis detetados de duplicação no transcriptoma.

Adicionalmente, a crescente evidência científica a apontar para um papel central por parte de nAChRs na infeção do vírus SARS-CoV-2 deu azo à ideia de que conotoxinas podem ter um papel no tratamento da doença Covid-19. Por conseguinte, neste trabalho também se fez uma tentativa preliminar de anotar todas as sequências provenientes dos transcriptomas de espécies *Conus* e do SARS-Cov-2 que revelassem algum grau de correspondência entre si, mas a evidência de relação genómica está ainda por ser encontrada.

Palavras-chave: *Conus*, transcriptómica, glândula venenosa, GO IDs partilhados, simbiose, Covid-19.

Abstract

Cone snails are iconic species known for centuries for their beautifully coloured pattern conical seashells, which were and still are very sought after for trophies and decorative assets. Since ancient times, the shells which are prized objects used for jewellery, are also a treasure for many collectors, for who there is still a thriving exchanging market. It was only recently that special attention was diverted to the capacities of these species' venom. As it turns out, other than having good-looking seashells, these snails produce a predatory venom with many complex compounds, each with incredible affinity for a certain class of receptors. These compounds act in a variety of ways from pain relieving to paralysation, thus being highly relevant for a broad field of biomedical applications if more effort is directed to their studies. With such goal in mind, this dissertation aims to contribute .to deciphering the molecular diversification of natural predatory venoms from these marine gastropods by studying available transcriptomes of Conus genus' species using comparative genomics and bioinformatics assessments.

For this dissertation a methodology was developed to decode the genomic background as well as the relationships among the proteins of these marine predators' venoms. After gathering and polishing all trustworthy transcriptomic data at disposal, transcriptome assembly and functional annotation of proteins revealed the functions of the expressed genes. Multiple-step analysis and comparisons based on the results obtained enabled the weaving of several conclusions. In short, in this work it is reported (1) the number and functions of the shared genes found to be uniquely expressed in the venom glands of all 20 *Conus* species analysed, (2) a correlation of assembly size with the number of unique genes found, (3) evidence for symbiotic microorganism relationships within the venom gland, and (4) an agreement with previous works regarding transcriptomic duplication levels.

Furthermore, increasing scientific evidence pointing for a central role of nAChRs in the SARS-CoV-2 infection gave rise to the idea of possible applications for conotoxins in the Covid-19 disease treatment. Therefore, in this work a preliminary attempt was made at reporting any matching sequences between the transcriptomes of *Conus* species and SARS-Cov-2, but evidence of a genomic relation is yet to be found.

Key words: *Conus*, transcriptomics, venom gland, shared GO IDs, symbiosis, Covid-19.

Table of contents

List of tables

List of figures

List of abbreviations

ACE2 – Angiotensin-Converting Enzyme 2

BAM – Binary Alignment Map

BLAST – Basic Local Alignment Search Tool

BUSCO – Benchmarking Universal Single-Copy Orthologs

COI – Cytochrome c oxidase subunit I

CPU – Central Processing Unit

DE – Differential expression

DNA – Deoxyribonucleic Acid

EggNOG – Evolutionary genealogy of genes: Non-supervised Orthologous Groups

FCUP – Faculdade de Ciências da Universidade do Porto/ Faculty of Sciences of the University of Porto

GO – Gene Ontology knowledgebase

HMMs – Hidden Markov Models

HTML – HyperText Markup Language

IO – Input/Output

M – Megabytes

MEGA – Molecular Evolutionary Genetics Analysis

mRNA – Messenger Ribonucleic Acid

nAChRs – nicotinic acetylcholine receptors

NCBI – National Center for Biotechnology Information

PD – Peptidase Domain

PTM – Posttranscriptional Modification

RBD – Receptor Binding Domain

RNA – Ribonucleic Acid

RNA-seq – High-throughput Sequencing of Transcriptomes

rRNA – Ribossomal Ribonucleic Acid

S protein – Spike glycoprotein

SAM – Sequence Alignment Map

SARS-CoV-2 – Severe Acute Respiratory Syndrome Coronavirus-2

SRA – Sequence Read Archive

UP – Universidade do Porto/ University of Porto

1. Introduction

1.1. Natural venoms: overview of the research field

Scientific interest in natural venoms started many decades ago mainly for health reasons in a search for medicines (1). The logic behind this being that the understanding of one venom's mechanism of action provides a perfect insight on how to effectively counter its poisonous nature and ultimately render it harmless. Beyond this simplistic counter-poison reasoning however, the study of a natural venom serves purposes other than helping with the finding of a medicine treatment for that specific venom.

Integrated in a defensive or offensive strategy, each natural venom is a very sophisticated biological weapon. Possessing a singular mixture of often unique components, mainly proteins and peptides, a venom interferes in the functioning of specific or diverse biological targets. Consequently, the accurate cataloguing and description of a venom component's nature and function potentially unveils previously unknown molecules and mechanisms. In turn, this discovery of novel molecules and its functions provides knowledge of incalculable value for a variety of disciplines and scientific fields such as health, genomics, proteomics and cellular, molecular and neurobiology (2) (3) (4).

Over the years, venom originating from all biological kingdoms was isolated and characterized. Fruitfully, it already prompted the creation and development of biochemical tools for health treatments, illustrating well how venom's studies are far reaching, pushing for the development of treatments for diseases other than the poisoning by the original venom from which the components originated from (2) (3) (4) (5) (6) (7) (8) (9) (10).

1.2. *Conus* species ecology

Particularly fascinating for this field of research is the venom of *Conus* species. From an ecological point of view, the *Conus* genus is a highly diverse natural group with more than 700 species of sea snails living preferentially in the intertidal zone of tropical and subtropical regions worldwide [Fig. 1] (11) (12) (13) (14).

Being natural predators, they can be found in rocky shores and coral reefs hunting with a specialized harpoon-like radular tooth to inject their shuddering, paralyzing venom onto fishes (piscivorous), molluscs (molluscivorous) or worms (vermivorous) (15) (16). Although not as widely studied, the *Conus* genus also produces a defensive venom for

situations where the animal perceives to be in imminent danger or under attack (17). The defensive venom is also produced in the same tissues as the offensive one, but as it is harder to obtain, it is not as well studied and is not subject of this dissertation.

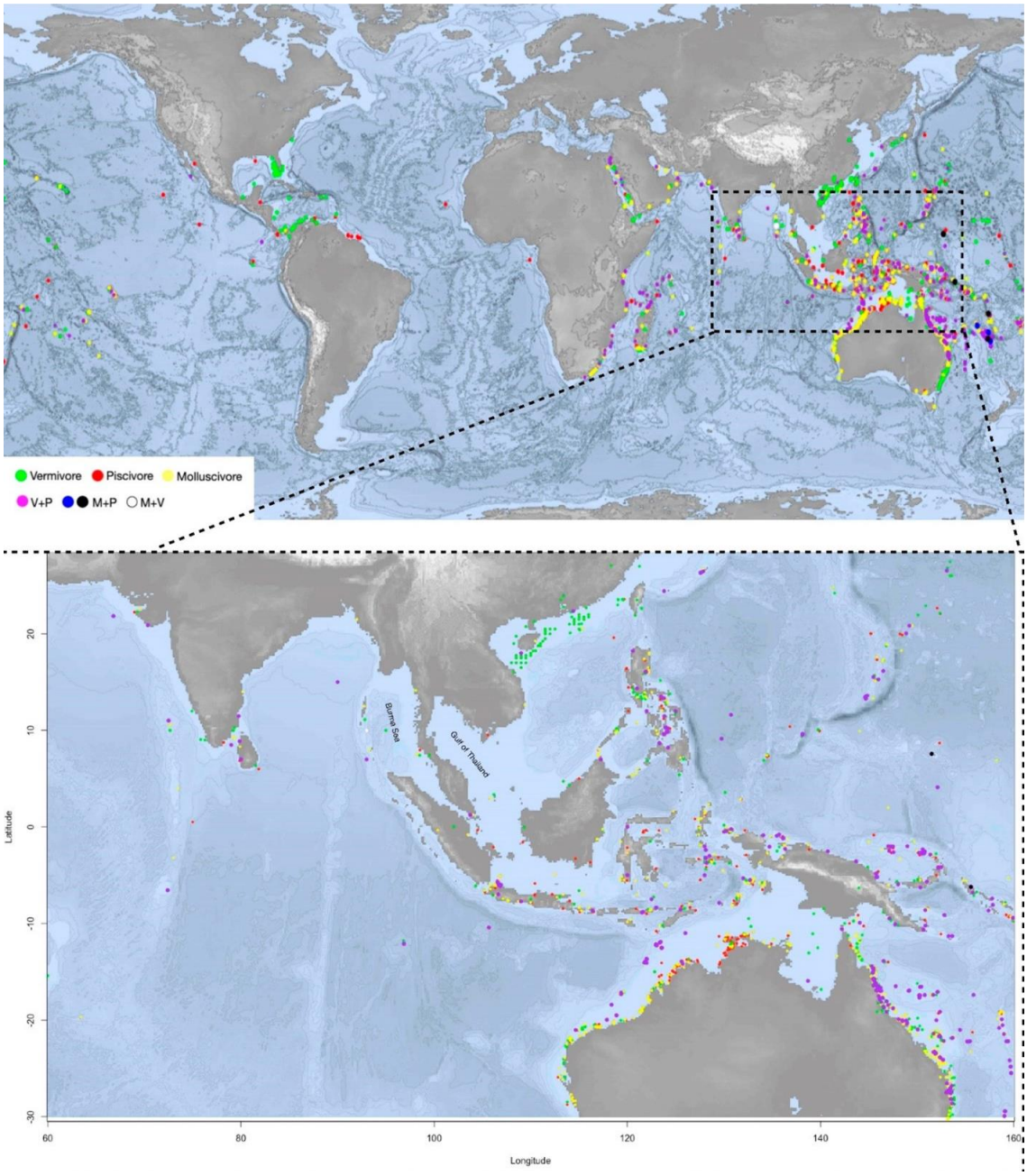


Fig. 1 – Distribution of *Conus* species throughout the world. The top map pictures the worldwide distribution of cone snails: molluscivous (M) in orange dots, piscivorous (P) in red, and vermivorous (V) in green. Dots in purple indicate the feeding habit of V+P; blue and black dots do it for the M+P; and the solely white dot represents the V+M. The lower map zooms in on all Southeast Asia and the northern coast of Australia, spotlighting the collection of habitats with the greatest diversification of *Conus* species. Image adapted from *(56)*.

1.3. Characteristics of the venom and the genome

The predatory venom is a cocktail composed of roughly a couple hundred bioactive well-structured polypeptides (~10-40 amino acid residues) known as conotoxins (18). They are secreted in the venom duct – a long and convoluted tubular structure [Fig. 2] – by the epithelial secretory cells before being pushed by muscle peristalsis of the venom bulb to be loaded into the harpoon-like tooth (19) (20). Conotoxins are synthesized as precursors with a three-domain structure: one is a conserved signal region; another is a pro-peptide region involved in the processing of the precursor; and lastly there is a highly variable, cysteine-rich mature region, which is the functional toxin (18) (21) (22). After their synthesis, these mostly disulphide-rich peptides suffer a series of PTMs enabling each of them to interact specifically with their intended target – an important aspect since it means minor side effects in disease treatment (18) (23) (24) (25) (26) (27) (28).

Targets of conopeptides include the presynaptic membrane calcium channel or G protein receptor, voltage-gated potassium and calcium channels, the receptors of serotonin, somatostatin, norepinephrine and adrenal hormone, and others (18) (29). The conserved signal region is currently used to classify precursors into "superfamilies" – a superfamily may consist of several families, each family targeting a specific ion channel and/or receptor (23) – although there is some degree of debate as to whether or not the current classification is well suited and pertinent (30).

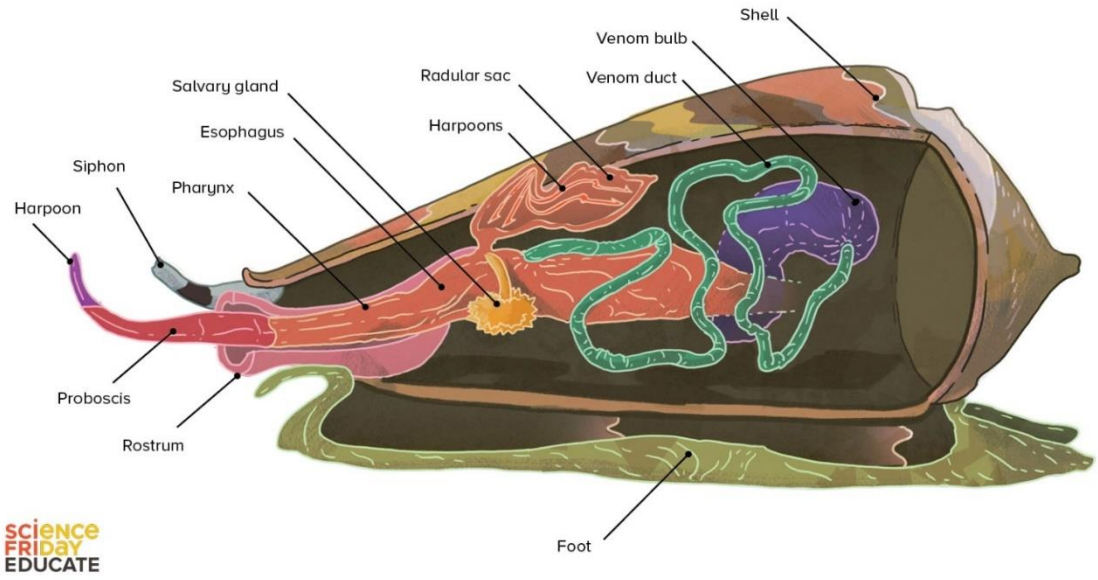


Fig. 2 – Macroscopic anatomy of a cone snail. Image taken from *(139)*.

Nowadays, there are 30 gene superfamilies and more than 7,000 conopeptides discovered out of an estimated 50,000 to 1 million different bioactive conotoxins across all these beautifully patterned shell species [Fig. 3] (31) (32) (33) (34) (35). Although only a few percent of this polypeptides have been sequenced and studied, the genus *Conus*' venoms are already well established in the grand venom's research field. Thoroughly reflecting the great potential as drug precursors are the proven results and direct practical applications in various scientific contexts such as neuroscience and pharmacology (36) (37) (38). For instance, in neuroscience, there are some venom's molecules in use as molecular tools, several are in clinical trials, and one is even already approved as a drug against chronic pain (13) (39) (40) (41).

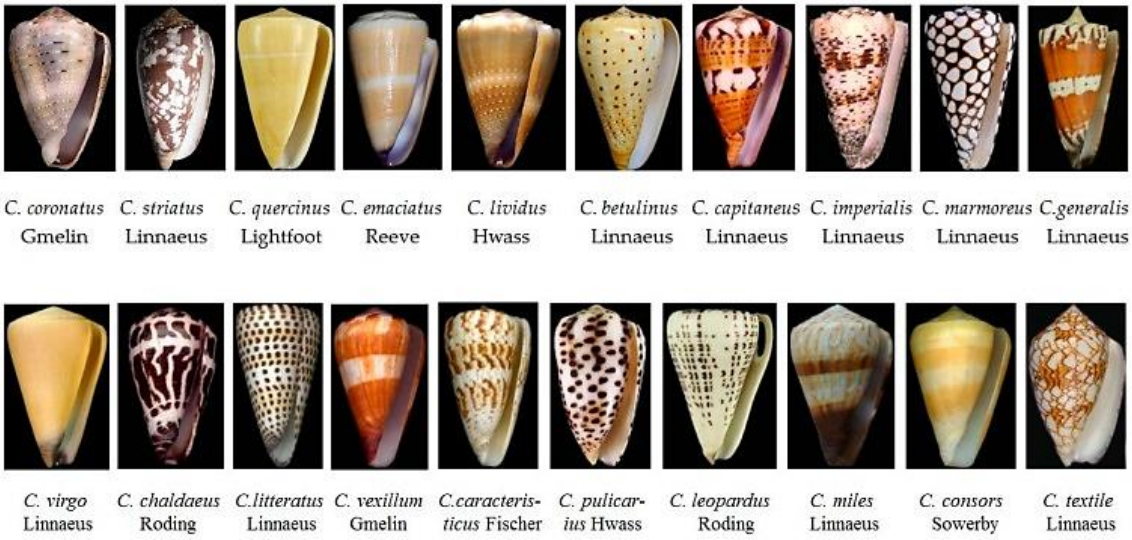


Fig. 3 – Beautiful cone snail shell patterns belonging to the 20 most abundant species in the South China Sea *(13)*.

Concerning the genome, *Conus* specimen's total genome size tends to have around 3.5Gb, being smaller than expected from previous fluorometric assays and flow cytometry (42) (43). Assemblies conducted for *C. betulinus* (the first intact genome assembled) and *C. ventricosus* showcase this as they have 86% and 87.6% of their corresponding size estimation, respectively (42) (43). As for genome structure, gastropods in general have an ample range of chromosome numbers and Conidae genomes are no different, varying from 16 pairs in *C. magus* to 35 pairs in *C. coronatus* (44) (45) (46).

Regarding the conotoxin genes it was revealed by the venom gland transcriptome that genes scattered in different chromosomes and located within repetitive regions encode conotoxin precursors, hormones, and other venom-related proteins (42) (43). The genes encoding conotoxin precursors are normally structured into 3 exons (mean~85

base pairs) which do not necessarily coincide with the 3 structural domains of the corresponding conotoxins – generally the boundaries of the first and second exons do not always coincide with the boundaries of signal and pro-peptide domains, but the third exon does exclusively encode the mature domain (43). Curiously, there is a big discrepancy between number of conotoxin genes versus transcripts and this may be due to natural variation among individuals coupled with the PTMs as while highly expressed transcripts are common to both specimens, variation of moderately and weakly expressed precursor sequences is surprisingly big (26) (42) (43) (47) (48) (49). Further illustrating this is the fact that, when conotoxin variability is compared quantitatively, highly expressed peptides show a strong correlation between transcription and translation, whereas peptides expressed at lower levels show a poor correlation (42) (43) (47). This in turn suggests that major transcripts are subject to stabilizing selection, while minor transcripts are subject to diversifying selection (47).

Additionally, on account of gene duplication, accelerated substitution rates, recombination, alternative splicing and differential expression, these neuroactive toxins are strikingly structurally diverse from species to species even before undergoing extremely complex PTMs (42) (43). Furthermore, some venom-related proteins have expression levels one order of magnitude higher in the foot than in the venom gland indicating that these hormones and proteins may be endogenous, having a physiological function common to different tissues and not restricted to the venom gland (43). The detection of low expression levels of toxin genes in different tissues outside the venom gland has also been demonstrated in snakes and in platypus (50) (51). To explain the evolutionary origin of this, it is suggested that toxin genes emerge through one out of two main paths: either by gene duplication and adaptive neofunctionalization of physiological genes in the venom gland coupled with reduction of expression levels in other tissues; or instead by sub functionalization through neutral evolution and restriction to the venom gland (47) (50) (52) (53).

At organ level, previous studies on dissected venom ducts also revealed that modern *Conus* species can produce venom peptides with different functions at different parts of the venom gland and even for different parts of the venom duct (17) (30). All along the length of it, qualitative and quantitative differences in conotoxin components were found, suggesting specialization of duct sections for biosynthesis of conotoxins (30) (54) (55). Further evidence for this specialization lies in the variation in epithelial composition in the proximal, central and distal portions of the duct (30) (54). These anatomic variations and differential expressions explain the compositional differences noted in multiple injections during single feeding events (30) (54) (55). Moreover, the

defence conotoxin group has been identified and distinguished from the predation-evoked conotoxin group, the former being present in the bulb-proximal and the latter in the bulb-distal end of the specialized venom ducts (17) (30). It is this astonishing capacity to diversify the arsenal of conotoxins that allowed the *Conus* snails to become specialized in attacking, defending, hunting and intimidating throughout all geographic locations of their habitats. Because of this remarkable flexibility, conotoxin compositions vary and are expected to do so from natural conditions to laboratorial cultures, causing difficulty in research assessments (56).

During predation some peptides assume greater importance than others, materializing the diverse conotoxin compositions throughout successive venom injections. Studies on the function of conotoxins and biomechanics of prey capture attribute a higher potency to the venom from the posterior end of the duct (the end attached to the venom bulb) comparatively to the venom extracted from the anterior portion near the proboscis (30). As the first injections tend to have peptides predominantly found in regions of the duct proximal to its insertion, whereas later injections are composed of peptides found in more distal regions near a large muscular venom bulb connected to the end of the duct, it can be concluded that the last injections are the most dangerous (16) (19) (42) (43). In general, compounds from A, M, O1 and T gene superfamilies account for the bulk of venom cocktails (26) (42) (43) (48). Peptides from the O-superfamily are prominent in all venom profiles and are present in multiple shots from single feeding events, indicating the importance of O-superfamily members throughout prey capture (13) (26) (42) (43) (54) (57) (58). These peptides block calcium and potassium channels and slowly inactivate sodium channels (13) (57) (58). In subsequent injections, peptides of the M- and T-superfamilies are more prevalent (54). Members of the T-superfamily target the presynaptic membrane calcium channel or G protein receptor, the somatostatin receptor and block noradrenaline transporters or voltage-gated sodium channels while venom peptides from the M-superfamily are known blockers of sodium channels, potassium channels and acetylcholine receptors (13) (57) (58). In this context, it is also important to note that the expression of some conopeptides is so low that their presence cannot be detected by traditional proteomic experiments (56).

1.4. The transcriptomics approach

Analysis of transcriptomes in a transcriptomics (transcriptome + bioinformatics) approach provided all this knowledge as in order to understand the mechanism of

action of a venom, first there is the need to discover what kind of molecules are composing it.

In this regard, comprehension of the transcriptome – being the set of all RNA transcripts including mRNA – is fundamental as the RNA sequences in the sample provide knowledge of the genes being actively expressed while also indicating the composition and structure of the proteins composing the venom. Thus, data processing and analysis is done utilising the raw RNA sequences' samples through software programs based on RNA-seq (59) (60).

In essence, RNA-Seq enables the study of genetic and functional information regarding any organism at an unprecedented speed and scale. Together, these two features immensely facilitate functional genomics research in species, especially in those for which genetic or financial resources are limited. This less spotlighted group includes many non-model organisms — organisms that, although they have not been extensively studied in a research setting, are nevertheless of substantial ecological, evolutionary or therapeutic importance — such as *Conus* and other venomous species (13) (59) (61).

In light of this dissertation, the most relevant aspect of RNA-Seq is that it makes it possible to simultaneously study the transcript structure and expression with a high resolution and a broad dynamic range. In this wise, investigation is conducted following predetermined strategies made in accordance with the species of interest. Generally, the primary workflow can be resumed to three steps: sample collection and polishing, transcriptome assembly and functional annotation. Upon annotation completion, the dataset composed of putative toxins, novel genes, and already known venom proteins and genes is then open to a variety of further analysis (59) (62) (63).

1.5. Possible relevance in the Covid-19 disease treatment

Since late 2019, a novel strain of coronavirus has enveloped our world in a pandemic. As of September of 2022, the severe acute respiratory syndrome coronavirus-2 that causes the Covid-19 disease has infected a confirmed 518 million people worldwide, claiming the lives of over 6 million of them (64).

This coronavirus is a positive-strand RNA virus, and its infection starts in the epithelial cells of the respiratory system, mainly in the lungs and trachea, via its S protein. Firstly, the trimeric viral S protein has its 3 heterodimers cleaved into its subunits: the S1s and S2s. Afterwards, while the S2s play a membrane fusion role, the RBD of the S1 subunits binds with the PD of the host ACE2. The establishment of this early

connection allows the virus to enter the host cells and is therefore crucial for the viral infection (65) (66).

Frequently, the infection reaches organs beyond the respiratory system ones as, among other aspects, there are ACE2 receptors in the heart, kidneys and intestine (67) (68) (69). Thus, the infection might and often does interfere in the functioning of the circulatory, renal, urogenital, digestive and even central nervous system as through the vascular system's blood vessels the virus is able to attack the peripheral nerves and disrupt the blood-brain barrier (70) (71) (72) (73).

In the present, the universally adopted counter measure against this contagious virus is vaccines (74). Through various doses of vaccination, the obvious aim is to render the populace ultimately immune to the Covid-19 disease. However, there are multiple possible side effects related to the treatment with all vaccines currently being administrated. For example, the Pfizer vaccine, being one of the most given vaccines, has an enormous list of possible side effects (75) (76). For these reasons, it is important to conduct further research in order to raise the standard of safety, comfort and trustworthiness of the methods used to combat the pandemic.

Interestingly, as Covid-19 is primarily classified as a respiratory disease, the theoretically expected main risk group would logically include anyone with a less healthy respiratory system, such as the elderly, the asthmatics or the smokers (77) (78). However, while this is true for the first two aforementioned groups of people, the reported low prevalence of smoking patients hospitalized due to Covid-19 really stood out and caught the attention of the professionals (79) (80) (81). On account of this early observation, it was suggested that nicotine might mitigate or even prevent the virus' infection (79). This hypothesis is being currently studied under different perspectives, with one of them being the direct administration of medicinal nicotine – already undergoing clinical trials (with no results posted until now) (82) (83).

Following this important suggestion, another one was made that stated a central role for nAChRs in the SARS-CoV-2 infection (70). This link was first based upon the discovery of a sequence homology between the furin cleavage site of the S protein and a neurotoxin motif that targets nAChRs (84). Reasoning that if Covid-19 can be, in some degree, controlled using nicotine to compete with the virus for binding to nAChRs, these receptors could be central in the process of infection (72) (84) (85). Indeed, evidence shows that the viral glycoprotein possesses a favourable affinity towards nAChRs, thus supporting this view and validating the importance of the initial link (70) (86) (87). Therefore, the neurotoxin in question assumes great relevance.

Curiously, it belongs to a family of powerful neurotoxins only present in the natural venom of a fascinating yet neglected group of species: the *Conus* species.

1.6. Objectives and strategy

For this Masters' thesis, a multitude of databases and bioinformatic procedures were respectively accessed and applied with the ultimate objective of contributing to the deciphering of the evolutionary genomics and transcriptomics of the *Conus* genus' natural venom.

In addition, as the possible application in the treatment of COVID-19 disease gradually took shape along the research work, the objective of helping in the endeavour of the fight against the pandemic also materialized. As stated previously, the sequence similarity between some conotoxins and a key virus protein (the S protein) gave the impression of an interaction of the latter with nAChRs – now confirmed. Aforestated studies unequivocally demonstrate that the S protein promptly binds with the α9-nAChR subunit, just like α-conotoxin. In virtue of these facts, further investigation connecting the *Conus* genus' toxins and the SARS-CoV-2 virus is thoroughly necessary, not to say potentially essential.

In order to reach the objectives, a four-step pipeline was designed and is briefly illustrated bellow [Fig. 4]. As before mentioned, the first step is the collection of all possible transcriptomic data regarding cone snails as well as the genome sequences of both the SARS-Cov-2 virus and the viral Spike protein. Within this step, it is required to polish the *Conus*' dataset properly and carefully before advancing further. The assembly of the acquired transcriptome follows, being conducted exclusively in the *de novo* method on account of limitations of time and computer storage space. Next, the functional annotation of the assembled transcriptome is performed, starting by sorting the data into two cured datasets: one exclusively with samples recovered from the venom apparatus and another with samples retrieved from other tissues and organs. Afterwards, the finding of the biological functions of the genes encountered empowers the jump to the final gene and protein comparison and matching analysis. Thus, in the fourth and last step aims to find relationships amidst the *Conus* genome datasets, and with the coronavirus' genome. The results are presented in the form of narrow category heatmaps, intersection plots, expression charts and a phylogeny tree.

Hence, starting with all the *Conus*' transcriptomic data publicly available and utilizing the latest bioinformatic tools in a combined transcriptomics approach to process it, the main objective of this dissertation is to produce a cohesive and coherent functional

annotation of the *Conus* species natural predatory venom, as well as reporting any relationships found within the wider genome of *Conus* comparatively to the narrower venom gland transcriptome. Extending from the original purpose, research on matching sequences between the cone snails' venoms and the coronavirus that caused the terrible pandemic was conducted.

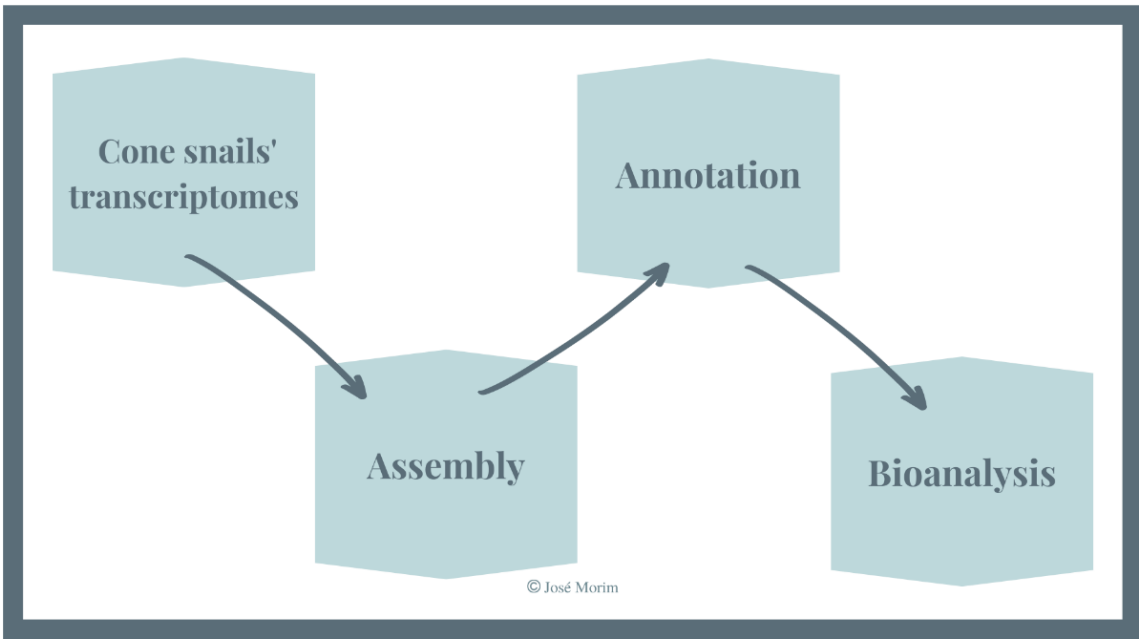


Fig. 4 – Diagram outlining the research' workflow of this dissertation.

2. Materials and Methods

2.1. Materials

All materials utilized in the performing of this research work were property of the UP. All procedures were conducted from one laboratorial computer located in the FCUP's laboratory 2.49. Conveniently, in times of mobility restriction or for progress security check-up, this computer was remotely accessed from my own personal computer using Anydesk – a closed-source remote desktop application (88). Being platform-independent, this software program provides remote access to other computers and devices running the host application.

The laboratorial computer operated on Linux system (for system specifications, see the Annexes section 7.1.), with every procedure made and all bioinformatic tools called for using a Konsole terminal – an open-source terminal emulator (89). All command functions employed are listed in the Annexes section (see 7.2.) exactly as they were written manually in the terminal.

2.2. Methods

The methodology of this dissertation is firstly illustrated bellow [Fig. 5] and then thoroughly explained in the following subchapters. In Fig. 5, all actions performed as well as all software utilized are listed in linear and chronological order.

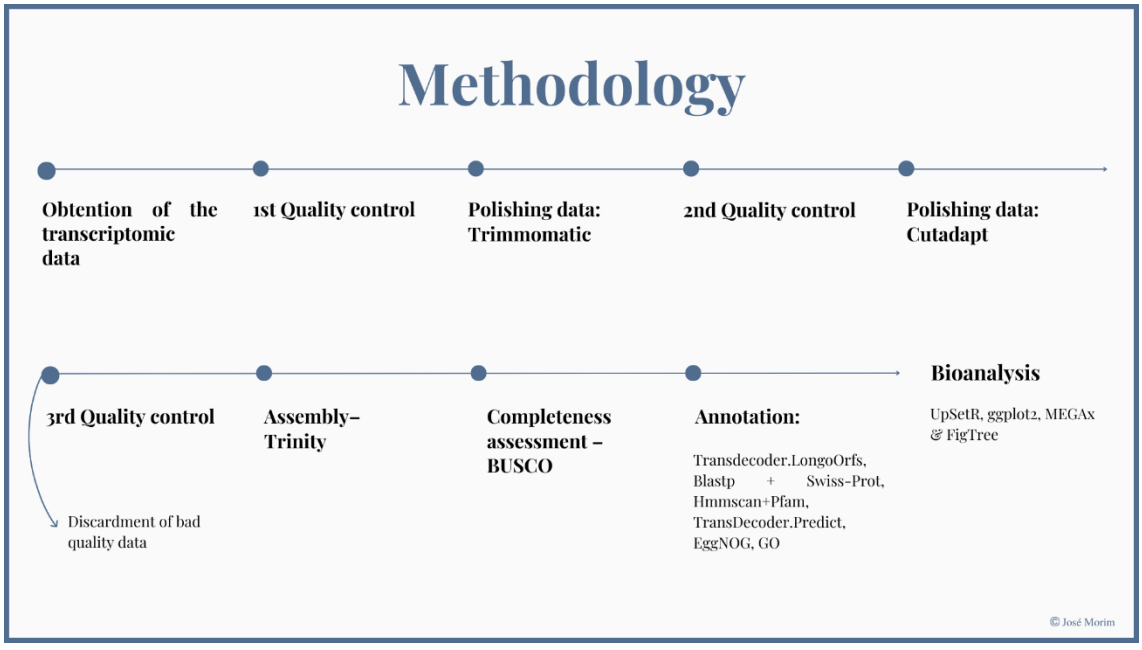


Fig. 5 – Comprehensive diagram illustrating the methodology workflow.

2.2.1. Obtention of the transcriptomic data

The transcriptomic data was acquired from the NCBI platform in the "SRA" format (see 7.2.1.). The SRA is a publicly available archive hosted by NCBI, providing a public repository for mostly raw DNA sequencing data (90) (91) (92). Previously known as Short Read Archive, this repository was initially for the "short reads" generated by high-throughput sequencing, which are typically less than 1,000 base pairs in length. It includes data submitted to NCBI, the European Bioinformatics Institute, and the DNA Data Bank of Japan. All data in the SRA format has full, per-base quality scores (93).

Originally, only 101 out of 105 total files were effectively downloaded as 4 of them had download access denied. Fortunately, after investigating the content of the files, it was verified that all the 4 missing ones were in fact equal to another downloaded file. Therefore, every single available data was obtained, considered, and processed.

2.2.2. Conversion of the dataset to the standard FASTQ format

Next there was the need to convert the downloaded data to a standard format, so, the data set was converted from SRA to FASTQ format – the *de facto* standard text-based format for storing the output of high-throughput sequencing instruments (see 7.2.2.) (94). The regular computational function can sometimes be – and indeed in this case was being – very slow, regardless of the technological resources (e.g., network, IO, CPU) available. To face this problem and increase the speed of the process, the tool was parallelized. Basically, this way the procedure is accelerated by first dividing the work into a requested, personalized number of threads, then running multiple functions in parallel and in the end concatenating the results back together (95). In practical terms, it makes the process much faster without affecting the result. The chosen thread number was 40 (out of 56 possible for the system used) and, just as intended, the parallelization delivered the end results much faster than normal functions previously attempted.

2.2.3. Quality control analysis

The tool used for quality control analysis was FastQC. This is a very popular bioinformatic instrument used to provide an overview of basic quality control metrics for raw next generation sequencing data. It imports data from a variety of file formats (any variant of BAM, SAM or, most importantly, FASTQ files) and exports the results to an HTML based permanent report. This report contains a set of summary analyses in a group of coloured bar charts and histograms to provide a clear impression on the state of the research data while also precisely indicating the areas where the quality falters (96). General statistics of the data such as an estimative percentage of duplicates, GC

content, read lengths and total sequences (in the millions for the samples used for this work) can be found right in the beginning of the report. Then, many other parameters are summarily presented but the most important are: the mean quality values across each base position in the read, the number of reads with average quality scores (per sequence quality scores), the sequence length distribution (which is expected to be around 100 base pairs) and the adapter content and per base N content (the percentage of each base position for which an N was called), both of which should be minimum to none. Other parameters such as duplication values and GC content are not very important since the former does not necessarily translate neither good or bad quality, and the latter should mostly present a normal distribution – a "bell shaped" graph – but deviations in the case of this research are to be expected since the data is of transcriptomic nature. Lastly, in the end of each report there is a heatmap compiling all the categories.

In this work, FastQC analysis was undertaken a total of three times: the first (see 7.2.3.) was done on the converted data as soon as the conversion from SRA to FASTQ format finished, the second was executed after the data was processed by the software Trimmomatic (see 7.2.4.), and in a similar way the third analysis was made after the dataset being further processed by the software Cutadapt (see 7.2.5.). The first analysis was a routine one made to generally access the state of the dataset without polishing. In contrast, the second and third quality controls were executed after examining the previous reports and submitting the dataset to trimming software a first and second time, respectively.

Following the third quality control, all data still not suitable for further analysis was discarded, but only after being properly investigated (see Results section 3.1.).

2.2.4. Polishing data with trimming software

The need for raw read filtering and trimming can be explain by demonstrations of previous studies that indicate these processes improve the quality of the future assembly (59) (97) (98) (99). Usually, this need is made clear by a quality control report on the newly acquired data. In this sense, after being acquired from NCBI and the routine quality control was made, the files were then processed by the software Trimmomatic. This software is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters (100). According to the problems found in the first quality report, Trimmomatic was run with adjusted parameters (see 7.2.6.) in the manner of the following reference:

❖ ILLUMINACLIP: <fastq_data> : <seed_mismatches> : <palindrome_clip_threshold> : <simple_clip_threshold> : <min_Adapter_Length> : <keep_both_reads> LEADING: <quality> TRAILING: <quality> SLIDINGWINDOW: <window_size> : <required_quality> MINLEN: <length> THREADS: <number>

The function of the options and the specified number attributed are thoroughly explained bellow (101):

- ILLUMINACLIP: <fastqData>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>
  - fastqData: specifies the path to the FASTQ files. In this work, the suggested adapter sequences used were the ones provided for TruSeq3 (as used by HiSeq and MiSeq machines); since the data is paired ended, this was also specified by writing "PE";
  - seed_mismatches: set as 2, it specifies the maximum mismatch count which will still allow a full match to be performed;
  - palindrome_clip_threshold: set as 30, specifies how accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment;
  - simple_clip_threshold: set as 10, specifies how accurate the match between any adapter etc. sequence must be against a read;
  - min_adapter_length: in addition to the alignment score, palindrome mode can confirm that a minimum length of adapter has been detected. Because the palindrome mode has a very low false positive rate, this number can be safely set even to 1 (reducing drastically from a default of 8, for historical reasons), as in this case, to allow shorter adapter fragments to be removed;
  - keep_both_reads: after read-though has been detected by palindrome mode, and the adapter sequence removed, the reverse read still contains the same sequence information as the forward read. For this reason, the default behaviour is to entirely drop the reverse read. This parameter avoids that, retaining the reverse read – which may be useful e.g., if the downstream tools cannot handle a combination of paired and unpaired reads;
- LEADING: <quality>
  - quality: set as 20, it specifies the minimum quality required to keep a base in the beginning of the sequence;

- TRAILING: <quality>
  - quality: set as 20, it specifies the minimum quality required to keep a base in the end of the sequence;
- SLIDINGWINDOW: <window_size>:<required_quality>
  - window_size: set as 4, it specifies the number of the group of bases permanently on the scanning window, across which quality is averaged;
  - required_quality: set as 15, it specifies the average quality required to keep the whole group of bases in the scanning window previously defined;
- MINLEN: <length>
  - length: set as 51, specifies the minimum length of reads to be kept;
- THREADS: <number>
  - set as 48 (from a maximum possible of 56).

In short, the previous parameters act in the removal of leading, trailing low quality (below or above quality 20, respectively) and N bases. They also dictate a scan of the read with a 4-base wide sliding window, cutting it when the average quality per base drops below 15 and dropping reads below the 51 bases long.

Following the Trimmomatic's editing, the second analysis report revealed the need for additional correction and so the software Cutadapt was used. In essence, Cutadapt is another trimming software that also finds and removes adapter sequences, primers, poly-A tails, and other types of unwanted sequences (102). In this case, it was used to just chop out the first and last 15 bases from the sequences (see 7.2.7.). This software also supports parallel processing – to enable it, the option "-j N" was used, where "N" is the number of cores to use and was set as 30.

As previously stated, after finishing the second trimming process, a third quality control analysis was called and subsequently all data at a less acceptable state was discarded.

2.2.5. Assembly of the transcriptome

The assembly of the transcriptome was done recurring to the software Trinity. This software processes large volumes of RNA-Seq reads through a combination of three independent modules which are applied consecutively to extract full-length splicing isoforms (103). The first module, Inchworm, generates transcript contigs for the second module, Chrysalis, to cluster together, which it does by constructing complete de Bruijn graphs for each cluster. Every single cluster represents the full transcriptional complexity for a given gene or groups of genes if they have the sequences in common. Then, Chrysalis partitions the full read set among these clusters that are in disjoint

Bruijn graphs. The last module, Butterfly, processes the individual graphs in parallel, eventually reporting full-length transcripts for alternatively spliced isoforms and setting aside transcripts that correspond to paralogous genes (103) (104).

There are two primary methods for assembly: through the guidance of previously assembled genomic sequences or via *de novo* assembly (104). For model organisms, the standard approach to transcriptome studies is the genome-guided one. However, as this approach cannot be applied to organisms for which a well-assembled genome does not exist – like most venomous species – a *de novo* transcriptome assembly is required (104) (105). For the genus *Conus*, the genome guided approach became possible only recently, with the intact assemblies conducted for *C. betulinus* and *C. ventricosus* made publicly available for reference in 2021 (42) (43).

In this work, although both methods of assembly are possible, only the *de novo* was carried out, partially due to time constrains but mainly as a consequence of limitations in computer storage space. The *de novo* assembly performed out of the whole transcriptomic data and without genome reference was therefore attempted (see 7.2.8.). The function ran on 25 CPUs and with an attached maximum memory of 25 Gigabytes. Unfortunately, the task could not be completed at first and stopped mid-way because the storage capacity of the laboratorial computer was all used up. Consequently, the dataset was compressed to highly compressed archive files. Fortunately, it succeeded in arranging enough storage space for the remainder of the operations to occur without any further delay of this sort.

2.2.6. Assembly completeness assessment

Evaluating the quality of the assembly and check the execution of the process before moving on is of major importance since every subsequent step utilizes the assembled dataset as a basis. In this regard, the BUSCO proved to be the elite choice to both quickly and reliably produce a quantitative and qualitative assessment of the transcriptome assembly completeness. Based on evolutionarily informed expectations of highly conserved gene content from near omnipresent single copy orthologs, this software performs an evaluation on genome, proteins, or, in this case, transcriptome even without a reference genome (106). Moreover, as its most time-consuming steps are parallelized, the software is very speedy. In the end, the assessment is presented in the form of a bar chart with the following categories:

- "Complete and single copy": for the orthologs whose aligned sequence length is within 2 standard deviations of the BUSCO group's mean (i.e., 95%).

- "Complete and duplicated": when multiple copies of the same orthologs are found in the gene set being assessed.
- "Fragmented": for orthologs whose length of their aligned sequence is beyond 2 standard deviations of the BUSCO group's mean length (i.e., <95%).
- "Missing": for any BUSCO without a BUSCO-matching gene meeting the 'expected score' cut-off.

For this research, two assessments were made in succession on the whole assembled transcriptome against two of the BUSCO's datasets: the "Mollusca" and the "Metazoa" (see 7.2.9.a and 7.2.9.b, respectively). The intention was to choose datasets of animal groups that included the *Conus* species but on a different magnitude. Therefore, the datasets chosen correspond to the phylum and kingdom where cone snails are inserted.

When analysing completeness assessments, the results should be interpreted by comparing both assessments given with the two BUSCO datasets. The reason for this lies with the fact of the completeness assessment being itself a comparison, searching the assemblies for highly conserved genes present in certain animal groups. Hence, as a relative value, it should always be analysed in perspective, meaning particular attention should not be placed on the percentages of completeness alone, but rather on which comparative dataset the tendency for a greater level of completeness is. Moreover, by focusing on specific tissues as in this case, a transcriptomic experiment is unlikely to produce a BUSCO-complete transcriptome since the level of differentiation is very high. For these reasons, the desired outcome is consistency across the assembled dataset.

After analysing the first two assessments, the dataset was split in two groups based on one category: the origin of the transcriptomic samples. Naturally, the transcriptome previously acquired was recovered from many organs and tissues of *Conus* species. As the interest of this work lies with the venom-related data, the dataset was divided in a group of venom-related transcriptomes and another for samples from other tissues. In this way, the assessment results are easier to interpret. Thus, after the origin of the raw data was inspected on NCBI, both groups had their completeness evaluated against each of the two BUSCO datasets used before (see 7.2.9.c, 7.2.9.d, 7.2.9.e, and 7.2.9.f). On a final note, as the graphs made by the software were very disfigured, the values were used to make better graphs with excel.

2.2.7. Annotation of the transcriptome

Annotation is a complex multi-step procedure that utilizes several different software and protein databases in sequential order to attain the functions of the proteins. The first stage is to identify and predict coding regions from the assembled transcriptome. Afterwards, the functionalities of the coding sequences are obtained through crosschecking and comparison across various reviewed databases. The perfected method chosen for annotation is illustrated bellow [Fig. 6] and is detailly explained in the following subchapters.

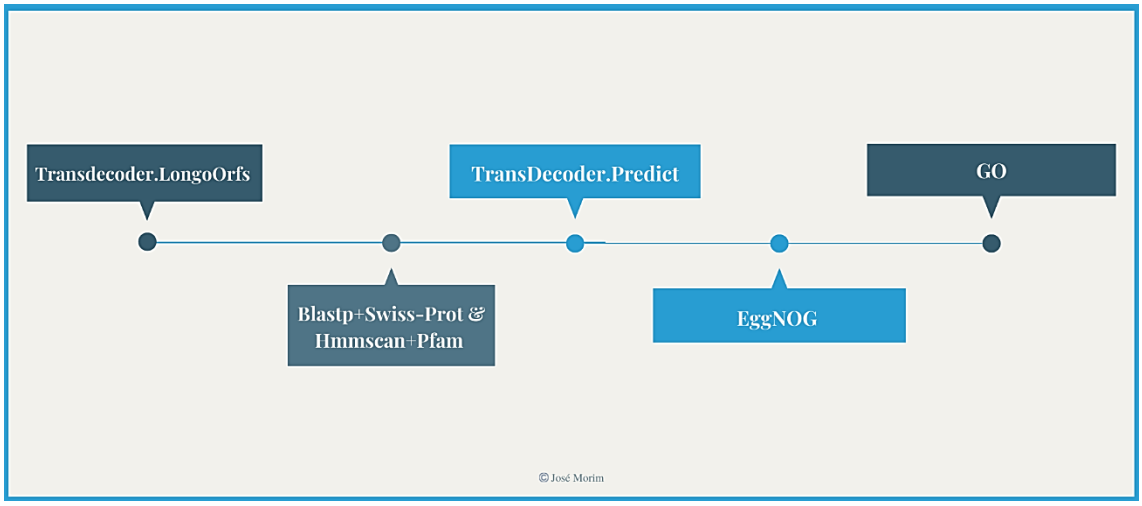


Fig. 6 – Annotation pipeline

2.2.7.1. Identifying likely coding sequences

The first of said software is TransDecoder, more specifically the TransDecoder.LongOrfs component. It serves to identify candidate coding regions within transcript sequences such as those generated by transcript assembly using Trinity (107). These likely coding regions are then confronted at various steps with one or more databases to provide an ever-more complete and accurate classification of protein families and domains. In this strategy, the most important output of TransDecoder.LongOrfs is the 'Trinity.fasta.transdecoder.pep' file, which contains the protein sequences corresponding to the predicted coding regions within the transcripts. This ".pep" file obtained for each of the assembled transcriptomes with TransDecoder.LongOrfs (see 7.2.10.) was used to elaborate the subsequent sequence homology and other bioinformatic analyses.

2.2.7.2. Support annotation with UniProt and Pfam

Having the intention of utilizing the greatest tools in conjunction with the best databases available, the strategy for annotation splits up at this point and follows two methods.

The first method combines the standard protein-protein BLAST (BlastP) with the UniProt database, or more specifically, its Swiss-Prot branch (108) (109). Used for both identifying a query amino acid sequence and for finding similar sequences in protein databases, BlastP is designed to find local regions of similarity like other BLAST programs (108) (110) (111). Regarding Swiss-Prot, it is part of the bigger UniProt database, but this section is reviewed, meaning that is a curated protein sequence database which strives to provide a high level of annotations, a minimal level of redundancy, and high level of integration with other databases (109) (112). Thus, the first method was used to report similar sequences existing in the previously obtained ".pep" files and Swiss-Prot, keeping only 1 aligned sequence and ignoring any hit with more than $1e^{-5}$ (e-value) chance of being random (see 7.2.11.a).

As for the second method, it features the combination of the "hmmscan" tool with the Pfam database. The tool offers a command line application to chart nucleic acid sequences, typically transcripts as in this case, in gene ontology. Based on the similarity of the query sequences, the aim is to implement methods using probabilistic models – called profile HMMs – to purify the sequence alignments by targeting protein families represented in Pfam (113) (114). As to Pfam, it is a protein database that consists of an immense collection of protein families, each represented by multiple sequence alignments and HMMs – which is why this method groups together this database with the referred tool (115) (116).

In similar fashion to the first method, with an e-value of $1e^{-5}$ any similar sequences between the Pfam database and the ".pep" files were reported (see 7.2.11.b). To improve the speed of the process, the dataset was temporarily sectioned in 8 lists (numbered from 1 to 8). In this way, 8 command functions were called to individually process each list (see 7.2.11.c – in it there is only the function called to process the list 1, the 7 other lists had the exact same script).

2.2.7.3. Prediction of coding sequences

The "Predict" function is the complementary part of the TransDecoder software and its function is one of prediction of likely coding regions. It is possible to use this software to ensure that the peptides with blast hits or domain hits (from Pfam) are retained in the set of previously reported likely coding regions. In this way, the results generated before with the two "supporting" methods can be leveraged by TransDecoder and this is exactly what was attempted (see 7.2.12.).

Once done, the procedure to obtain the coding information contained in the assembled transcriptome had also finished. Intending to check if it generated a viable output, the

results were compared to the Pfam database (see 7.2.13.). If the produced dataset had few to no hits in the Pfam database, it would mean that at least something in the methodology was incorrectly done or that the parameters were imprecise, and that it would be necessary to review the whole process. Alternatively, the existence of abundant recognitions from the reviewed database directly means the success of the methodology.

2.2.7.4. Annotation mapping

EggNOG is an extensive public database in which a dataset can be analysed with thousands of genomes to establish ortholog relationship (117). The multi-threaded EggNOG-mapper is ideal for the large-scale functional annotation needed for this dissertation as it uses the computed orthologous groups and phylogenies from the vast software database to provide the accession numbers of various annotation sources (117) (118). Essentially, across multiple platforms this software charts the obtained coding regions into accession numbers for various databases, notably for the Gene Ontology knowledgebase, which once provided enable the functional annotation of the transcriptome.

From the website of the EggNOG, the dataset of "Eukariota" was aquired and the subsequent run of the software was parallelly executed against it to obtain the beforementioned accession numbers for multiple platforms (see 7.2.14.) (119).

2.2.7.5. Functional annotation

The Gene Ontology (GO) is an excellent resource for developing a comprehensive model of biological systems across a multiplicity of species. As a matter of fact, the GO knowledgebase is the largest source of information on the functions of genes in the world and thus it is ideal for functional annotation (120) (121).

The previous mapper software provided information on the existence of GO terms in the dataset. Consequently, the next step in the research was to process the dataset through the GO knowledgebase itself (see 7.2.15.). Utilizing the latest database of GO (obtained in (122) and (123)), extremely reliable information on the functions of the genes encountered – whether they relate to cellular components, molecular functions or biological processes – is given in the form of a "GO ID". A GO ID is unique seven-digit identifier prefixed with the text "GO:" e.g., GO:0000001, GO:0010101, or GO: 1500489 (124).

2.2.7.6. Secondary annotation using locally built databases

Going even further in order to obtain a narrower picture of the functions the sequences within the ".pep" files from the transcriptomic dataset, these files were processed with BlastP against two locally-built databases: Tox-Prot and Conoserver. The Tox-Prot is a specific toxin protein database that is part of the Swiss-Prot, which is in turn part of the larger UniProt as previously explained (125). The Conoserver is a database specialized in the sequence and structures of conopeptides (126). Both databases were made locally with a BLAST function called "makeblastdb", that serves to make a local specific database of protein (or DNA) sequences. Its purpose is to speed up the search, being used automatically by an appropriate BLAST program. Hence, these "makeblastdb" commands were written to build a protein database keeping the original sequence identifiers – otherwise the command will generate its own identifiers (see 7.2.16.)

After creating the local databases, these BlastP functions were called following the example of earlier ones (see 7.2.17.).

2.2.8. Annotation of the viral Spike protein and full SARS-Cov-2 genome

After obtaining the genome of the Spike protein from NCBI, it was processed by TransDecoder in the same fashion as the before given example to obtain a ".pep" file necessary for protein-protein searches (see 7.2.18.a). Afterwards, the ".pep" file was used versus the Tox-Prot and Pfam databases (see 7.2.19.a and 7.2.19.b, respectively). Additionally, in almost identical manner as the Spike protein genome, the whole SARS-Cov-2 genome was downloaded from NCBI, submitted to TransDecoder (see 7.2.18.b), and compared against the Pfam database (see 7.2.19.c). The results from both annotations were then compared in search for any matching sequences between the venom-related dataset (with known toxins or conotoxins) and the coronavirus.

2.2.9. Statistical analysis, graphic visualization, and phylogeny study

Due to the massive size of the transcriptomic dataset, visualization of the relationships among the huge collection of data was only possible through statistical processing. Utilising a package from the R software denominated UpSetR, the previously acquired GO IDs from the Gene Ontology knowledgebase are used to generate static UpSet plots (127) (128). These plots allow the visualization of all the intersections of a given dataset in a clean matrix layout. Hence, UpSet plots were made to visualize the intersections of GO IDs among the samples of venom related data first, and data coming from other tissues and organs second (see 7.2.20.). Intersections among all samples from the dataset indicate the genes present in all samples, while intersection

values of zero represent genes only present in one sample. In addition to UpSet plots, another package of the R software was used to visualize the genomic relationships: ggplot2 (see 7.2.21.) (129). This system is specially designed to create graphs and charts for comprehensive visualization of large and complex datasets.

In the end, to contextualize the findings a phylogeny study for the 20 species present in the dataset was made using two 16S rRNA genes for each species collected from NCBI. The reason for using 16S rRNA genes and not 12S rRNA or COI genes is linked with the more accurate practical results yielded from using the 16S genes. Also, only two were used for each species because of the limiting maximum number of existing 16S rRNA genes for each of the present species. After a maximum likelihood test (using Tajima 2-parameter model) was conducted with MEGA software, a phylogeny tree was created with FigTree software (130) (131).

3. Results

3.1. Pre-assembly quality controls

The quality control performed on the acquired data from NCBI revealed a clear need for polishing and trimming as can be seen in the first quality control report (see 7.3.1.). From the report, however promising the basic statistics parameter was (Fig. 18, extreme left column), it was evident that several sequence' stretches along the whole dataset were polluting the true *Conus* transcriptomic data. This contamination was perceived to be mainly due to residues of adapters, which coupled with numerable mismatches and unspecified base pairs made the quality of the dataset drop significantly. At this point, the under-quality data was not yet attributed to inadequate or substandard sample sequencing.

Having under advisement the unsettling report, the Trimmomatic executed an orderly scan of the dataset, eliminating the first, last and any bases along the sequences with less than the defined threshold quality and discarding any read bellow 51 base pair length, just as intended. This succeeded in bringing the whole dataset to a good per base sequence quality level as evidenced in second quality control report (see 7.3.2.). Unfortunately, the result dataset still lacked sufficient quality to proceed for assembly. This time, the greatest problem was due singularly to the persistent low per base sequence content, as well low-quality base pairs in the start and in the end of the sequences, which was figured to be residues of adapter content.

In the face of this result, a second attempt at polishing the dataset was carried out with Cutadapt. This time, all first and last 15 base pairs of the reads were removed, ensuing at last that most of the dataset presented a good quality for assembly as seen in the third quality control report (see 7.3.3.). Beyond the basic statistics being solidly good, most of the data had excellent scores in the parameters that mattered the most, such as sequence quality, per base sequence quality and content, and minimum adapter content and "N" bases. However, there were some remaining transcriptomic files which were seemingly irremediably inadequate for the assembly process. From those files, 7 were discarded for bad per sequence content and 18 were removed on account of adapter content still polluting the samples, for a total of 25 files removed and discarded [Fig. 7].

Ultimately, the assembly process was executed on 76 approved transcriptomic samples [Fig. 8]. Altogether there were 20 different species of cone snails, the most represented being *C. miliaris* with 22 specimens, followed by *C. ermineus* with 9, *C.*

consors with 7, *C. betulinus* and *C. imperialis* both with 5, *C. coronatus* and *C. tribblei* both with 4, *C. lenavati* and *C. ventricosus* both with 3, *C. litteratus*, *C. magus* and *C. ebraeus* with 2 each, and *C. varius*, *C. virgo*, *C. marmoreus*, *C. lividus*, *C. judaeus*, *C. quercinus*, *C. rattus* and *C. sponsalis* with 1 specimen each. The 76 transcriptomes were collected from a total of 7 different tissues including not only the venom gland's tissues (venom duct and bulb), but also foot, osphradium, salivary gland, nervous ganglions, and proboscis.

Article	Data file	Body part	Species	Date of collection	Reason for removal
Transcriptomic-Proteomic Correlation in the Predation-Evoked Venom of the Cone Snail, *Conus imperialis*	SRR12186674	Venom gland	*C. imperialis*	2019	**Adapter content**
	SRR12186675				
Diversity of Conopeptides and Conoenzymes from the Venom Duct of the Marine Cone Snail *Conus bayani* as Determined from Transcriptomic and Proteomic Analyses	SRR13781584	Venom duct	*C. bayani*	2016	
Reticulate evolution in Conidae: Evidence of nuclear and mitochondrial introgression	SRR14407584	Venom duct	*C. abbreviatus*	2009	
	SRR14407585	Venom duct			
	SRR14407586	Venom duct	*C. aristophanes*	2010	
	SRR14407587	Venom duct			
	SRR14407577	Osphradium	*C. chaldaeus*	2015	
	SRR14407576	Venom duct	*C. ebraeus*	2015	
	SRR14407590	Venom duct			
	SRR14407582	Venom duct	*C. fulgetrum*	2015	
	SRR14407583	Venom duct			
	SRR14407588	Venom duct	*C. judaeus*	2015	
	SRR14407589	Venom duct			
	SRR14407578	Venom duct	*C. mordeirae*	2002	
	SRR14407579	Venom duct			
	SRR14407580	Venom duct	*C. regonae*	2002	
	SRR14407581	Venom duct			
Transcriptomic resources for three populations of *Conus miliaris* (Mollusca: Conidae) from Easter Island, American Samoa and Guam	SRR1544118	Venom duct	*C. miliaris*	2007	**Bad per sequence content**
Conus consors transcriptome sequencing	SRR1958824	Mantle	*C. consors*	2007	
The Venom Repertoire of *Conus gloriamaris* (Chemnitz, 1777), the Glory of the Sea	SRR827576	Muscle venomas	*C. gloriamaris*	2016	
	SRR827577	Central nervous system			
Optimized deep-targeted proteotranscriptomic profiling reveals unexplored *Conus* toxin diversity and novel cysteine frameworks	DRR034331	Venom apparatus	*C. episcopatus*	2015	
	DRR034332				
	DRR034333				

Fig. 7 – Description of the 25 transcriptomic samples removed from the dataset, including (from left to right) the samples' origin study, names, tissues, species, date of collection and reason for removal.

	SRR file	Tissue	Species	Assembly size (M)
Venom-related samples	SRR2124878	Venom duct	*C. betulinus*	63
	SRR2124879	Venom duct	*C. betulinus*	76
	SRR2124880	Venom duct	*C. betulinus*	87
	SRR2124881	Venom duct	*C. betulinus*	105
	SRR2124882	Venom duct	*C. betulinus*	106
	SRR1964035	Venom bulb	*C. consors*	139
	SRR1954994	Venom duct	*C. consors*	167
	SRR14407591	Venom duct	*C. coronatus*	12
	SRR2609545	Venom duct	*C. coronatus*	16
	SRR14407592	Venom duct	*C. coronatus*	18
	SRR2609544	Venom duct	*C. coronatus*	20
	SRR2609538	Venom duct	*C. ebraeus*	25
	SRR17653518	Venom duct	*C. ebraeus*	39
	SRR6983162	Venom duct	*C. ermineus*	29
	SRR6983167	Venom duct	*C. ermineus*	30
	SRR6983169	Venom duct	*C. ermineus*	32
	SRR6983168	Venom duct	*C. ermineus*	33
	SRR6983164	Venom duct	*C. ermineus*	36
	SRR6983163	Venom duct	*C. ermineus*	48
	SRR6983161	Venom duct	*C. ermineus*	51
	SRR6983165	Venom duct	*C. ermineus*	52
	SRR6983166	Venom duct	*C. ermineus*	91
	SRR2609542	Venom duct	*C. imperialis*	14
	SRR12186678	Venom duct	*C. imperialis*	22
	SRR12186679	Venom duct	*C. imperialis*	22
	SRR12186677	Venom duct	*C. imperialis*	24
	SRR12186676	Venom duct	*C. imperialis*	53
	SRR17653514	Venom duct	*C. judaeus*	68
	SRR1803942	Venom duct	*C. lenavati*	62
	SRR1803941	Venom duct	*C. lenavati*	67
	SRR1803940	Venom duct	*C. lenavati*	101
	SRR6381569	Venom duct	*C. litteratus*	66
	SRR6381570	Venom duct	*C. litteratus*	76
	SRR2609539	Venom duct	*C. lividus*	26
	SRR9831243	Venom duct	*C. magus*	38
	SRR9831255	Venom duct	*C. magus*	46
	SRR2609532	Venom duct	*C. marmoreus*	24
	SRR1548190	Venom duct	*C. miliaris*	10
	SRR1544120	Venom duct	*C. miliaris*	14
	SRR1544597	Venom duct	*C. miliaris*	15
	SRR1548188	Venom duct	*C. miliaris*	15
	SRR1544142	Venom duct	*C. miliaris*	16
	SRR1548185	Venom duct	*C. miliaris*	16
	SRR1548186	Venom duct	*C. miliaris*	16
	SRR1544600	Venom duct	*C. miliaris*	17
	SRR1544622	Venom duct	*C. miliaris*	17
	SRR1544692	Venom duct	*C. miliaris*	17
	SRR1548189	Venom duct	*C. miliaris*	19
	SRR1544690	Venom duct	*C. miliaris*	20
	SRR1544140	Venom duct	*C. miliaris*	21
	SRR1544595	Venom duct	*C. miliaris*	22
	SRR1548192	Venom duct	*C. miliaris*	22
	SRR1544119	Venom duct	*C. miliaris*	23
	SRR1544137	Venom duct	*C. miliaris*	23
	SRR1548187	Venom duct	*C. miliaris*	24
	SRR1542681	Venom duct	*C. miliaris*	25
	SRR1542424	Venom duct	*C. miliaris*	26
	SRR1544117	Venom duct	*C. miliaris*	28
	SRR1544627	Venom duct	*C. miliaris*	34
	SRR2609537	Venom duct	*C. quercinus*	25
	SRR2609540	Venom duct	*C. rattus*	23
	SRR2609541	Venom duct	*C. sponsalis*	21
	SRR1803939	Venom duct	*C. tribblei*	78
	SRR1803938	Venom duct	*C. tribblei*	83
	SRR1803937	Venom duct	*C. tribblei*	86
	SRR1799982	Venom duct	*C. tribblei*	150
	SRR2609543	Venom duct	*C. varius*	30
	SRR13740844	Venom gland	*C. ventricosus*	89
	SRR2608262	Venom duct	*C. virgo*	17
Samples from other tissues	SRR1958882	Foot	*C. consors*	48
	SRR13770976	Foot muscle	*C. ventricosus*	77
	SRR13757741	Foot muscle	*C. ventricosus*	98
	SRR1955039	Salivary glands	*C. consors*	115
	SRR1954996	Nervous ganglions	*C. consors*	178
	SRR1958823	Osphradium	*C. consors*	194
	SRR1964034	Proboscis	*C. consors*	208
Total:	**76**	**7**	**20**	

Fig. 8 – List of the assembled data, featuring (from left to right) the samples' names, tissues, species, and assembly size (in megabytes), as well as total numbers of samples (76), tissues (7), and species present (20).

3.2. Assembly results and completeness assessments

3.2.1. General numbers

In order to better interpret the results, the transcriptomic dataset with 76 files in total was split in two groups, as explained before. The group of venom-related data had 69 files from the venom ducts, bulb and gland of 20 species, and the group with data coming from other tissues and organs had 7 samples, which originated from 5 different organs and tissues of 2 different species: *C. consors* and *C. ventricosus*. By organizing the dataset in this fashion, it becomes easier to visualize the assessment results.

The smallest assembly size had 10M and belonged to a *C. miliaris* venom duct, while the largest had 208M and was from a proboscis of a *C. consors*. The mean assembly size for the whole dataset was 52M. However, the mean size for the 69 transcriptome assemblies from the venom apparatus complex was 44M, which means that all 7 samples from other parts of cone snails' body had an assembly size superior to the mean. In fact, the mean value for the assembly size among these samples was 131M, illustrating just how large these samples were comparatively to the venom-related ones.

From the 69 venom related assemblies, 24 (a third of the venom related data) had larger assemblies than the local mean of 44M, with 6 even having more than 100M, belonging to venom ducts of *C. lenavati* (101M), *C. betulinus* (105 and 106M), *C. tribblei* (150M) and *C. consors* (139 and 167M). Looking at the 7 samples from other body parts, the 3 foot transcriptome assemblies were the smallest having less than 100M, with the one from *C. consors* having 48M and the two from *C. ventricosus* having 77M and 98M. The remaining 4 samples were all from different body tissues of *C. consors* and all had more than 100M size – 115M for the salivary glands' sample, 178M for the nervous ganglions, 194M for the osphradium and 208M for the proboscis, the largest of them all.

3.2.2. Completeness Assessments

Aiming to find consistency across the whole dataset of transcriptomes assembled the results point to higher completeness levels with the "Metazoa" BUSCO dataset and lower with the "Mollusca" one (see 7.4.1.). The values for complete BUSCO-matching orthologs with both datasets are not by any means high, and most of the dataset has huge quantities of genes encountered without any BUSCO matching ortholog. Again, this is expected since most of the data is from highly differentiated tissues and organs. However, the assessments revealed a clear tendency for greater genomic familiarity outside of the Mollusca phylum. These results seem to suggest that, despite being

molluscs, the cone snails' transcriptome (not only from the venom apparatus) is significantly different from others encountered in the rest of the Mollusca phylum.

Analysing the reports more closely, in the assessment made with BUSCO's Mollusca dataset only 4 samples had a completeness near 50%: 1 from the venom duct of a *C. triblei*, and 3 from a *C. consors*, from the proboscis, osphradium and nervous ganglions (see 7.4.1., Fig. 21). In turn, by looking over to the assessment made with BUSCO's Metazoa dataset, there are 15 samples with completeness fully over 50%. Enumerating, 7 are from the venom ducts of *C. triblei* (1), *C. betulinus* (3), *C. litteratus* (1), *C. consors* (1) and *C. ermineus* (1), 5 are from the venom bulb, proboscis, salivary glands, osphradium and nervous ganglions of a *C. consors*, and the last three come from the venom gland (1) and from the foot (2) of a *C. ventricosus* (see 7.4.1., Fig. 22). From this second assessment, the possible deficit of quality and depth of the sequencing data can be steadily diminished to but a few samples, for indeed there is a considerably greater quantity of orthologs. In reality, the missing results must arise from the aforementioned reasons, plus the lack of genomes of *Conus* species available and studied.

Regarding the venom-related group (see 7.4.2., Fig. 23 and 24), the tendency of higher BUSCO-matching in the Metazoa set can easily be distinguished for there are several samples with near or higher than 50% completeness – mainly from the venom ducts of *C. ermineus*, *C. consors*, *C. betulinus*, *C. ventricosus*, and especially *C. triblei*. The suspiciously high duplication values observed in 3 samples from the venom ducts of *C. triblei* and *C. consors*, as well as in the venom bulb of *C. consors* are strange at first since BUSCO works with single copy orthologs. Nevertheless, they can be explained by the fact of both the venom duct and bulb being highly differentiated organs, where the genes produced are very closely related and expressed with great intensity.

As for the samples from other tissues and organs (see 7.4.3., Fig. 25 and 26), again the tendency for higher BUSCO-matching in the Metazoa dataset is again verified, with all but one sample having lower than 50% completeness: a lonely *C. consors* foot sample. Another 2 samples from a *C. ventricosus* foot had only a slightly higher completeness value, with the remaining 4 samples all being from *C. consors*. These revealed a both high single ortholog content and a high duplication value which in this case can be attributed to a potentially, naturally occurring duplication event in the species.

In all, the examining of the assembly completeness proved the expected consistence and coherent tendencies of the dataset, which in turn verify the success of the trimming and assembly of the transcriptome processes.

3.3. Annotation of the transcriptome

3.3.1. Obtention of coding sequences

Success in predicting coding regions would deeply validate the whole methodology utilized until this point, for only in the case of correct pre-assembly and assembly procedures there would be an abundance of coding sequences. In this wise, comparing the predicted coding regions with the Pfam database produced very interesting results (see 7.5., Fig. 27). Already at first glance it is manifestly clear that the intended method had succeeded, for there were domain hits in the order of thousands for every piece of transcriptomic data, regardless of how many coding sequences had been predicted for each one.

A more scrutinous examination reveals a minimum ratio of 24 domain hits for every 50 coding sequences predicted, meaning all data had at least 48% of its predicted coding sequences with some connection with a known domain on the public database. In reality, 68 samples (out of the 76 samples – 89% of the dataset) had at least 50% correlation with Pfam. For the whole dataset, the mean value of connection with the protein database sits at a very comfortable 56% with 45 samples having this or higher ratio. In fact, there were even 19 samples (25% of the dataset) having a 60% or higher ratio, proving that the method had in fact accomplished the desired objective of correctly identifying and predicting coding sequences.

As to the proper number of coding sequences, the range of them predicted by TransDecoder goes from a low of 2,922 (in the smaller size samples) to a staggering high of 45,578 (in the samples of larger size), with the hits in Pfam running from 1,705 to 25,184. In total, TransDecoder predicted 951,425 sequences, with Pfam recognizing 510,303 of them, meaning there were 441,122 predicted but not recognized sequences. Noticeably, as the size of the data becomes larger, more numerable become the sequences predicted and greater are the amount of Pfam hits attributed.

Concerning the venom-related data, 45 out of the 69 samples (65% of the group dataset) had a ratio of Pfam-hits/predicted-coding-sequences superior to the mean value (see 7.5., Fig. 28). In total, 749,112 sequences were predicted from these samples, with domain hits for 405,090, meaning 344,022 sequences were not recognized from this group. In all, the great number of coding sequences predicted with

TransDecoder and valued with Pfam provided a very strong basis for annotation, with many possible conotoxins and venom-related proteins being present. Markedly, the samples with the highest prediction and domain hit values belonged to *C. lenavati*, *C. betulinus*, *C. consors*, *C. triblei* and *C. litteratus*.

Interestingly, all 7 samples from tissues non-related to the venom gland had a ratio inferior to the mean value, however close they were (see 7.5., Fig. 29). With TransDecoder 202,313 sequences were predicted and Pfam recognized 105,213 of them. The mean value of the ratio of predicted sequences to Pfam hits for this group of samples is 51%. This number while still being positive lags behind the previous ones. This may be due to the smaller differentiation of these organs and tissues, coupled with a lack of *Conus* species genome studies, spiralling to higher diversity of possible coding sequences detected, especially when compared with the highly specialized venom apparatus. Alternatively, it may also indicate some over prediction by the software.

3.3.2. Gene Ontology results

The transcriptomic experiment concludes with the obtention of the precious GO IDs, the terms where precise information related to the cellular components, molecular functions, and biological processes of the genes encountered is stored.

3.3.2.1. Shared genes number and assembly size-unique genes correlation

Analysing the general numbers for the samples coming from tissues and organs not related to the venom apparatus – which is a smaller dataset – it is found that the 7 samples share 5,913 GO IDs in common [Fig. 9]. This group of shared genomic content is very important as it is perceived to represent housekeeping genes. Provided with this knowledge, the comparison with the venom-related data group can be done.

Nevertheless, analysing other particularities of this dataset can be of relevance in the future. Hence, per assembly, the number of GO IDs ranges from just over 7,500 in a sample of a *C. consors* foot to a maximum of little more than 15,000 in the samples of the proboscis and osphradium of the same specimen, with the other 4 samples having more than 10,000 GO IDs attributed. At an individual level [Fig. 10], the samples of the osphradium and proboscis of a *C. consors* have 536 and 434 unique GO IDs, respectively, being the ones with the greatest amount of unique GO IDs attributed. In the other side of the spectrum, there is the sample of the foot of *C. consors*, with just 36 unique. However, the samples from *C. venstricosus* have 308 and 162 unique, which is curious since they also come from the foot. The samples from the nervous ganglions and salivary glands show 312 and 170 unique GO IDs, respectively.

By examining these numbers, the suspicion of a correlation emerges: it seems the greater the assembly size along the 7 samples and the more GO IDs present, the more unique genes are detected. This is very pertinent as there was a correlation noticed previously in which the prediction of coding sequences grew with the increasing of assembly size.

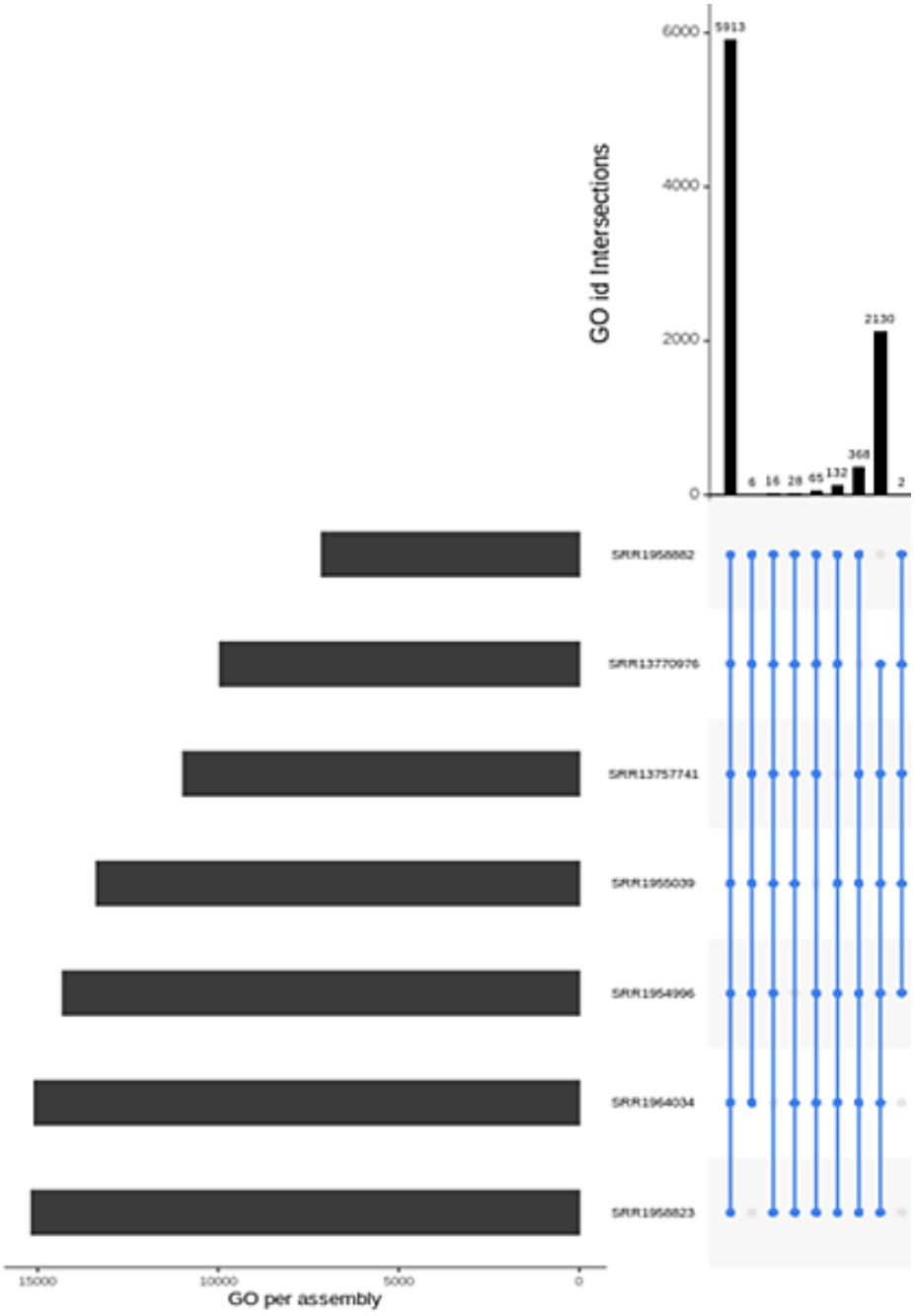


Fig. 9 – UpSet plot illustrating the common GO IDs for the 7 transcriptomes of various tissues in the top bar chart. The first vertical blue line intersecting all 7 samples shows the 5,913 IDs that are shared by these tissues from all over the body and are thus perceived as housekeeping genes. Additionally, in the side black-bar chart are the total GO IDs attributed to each sample individually.

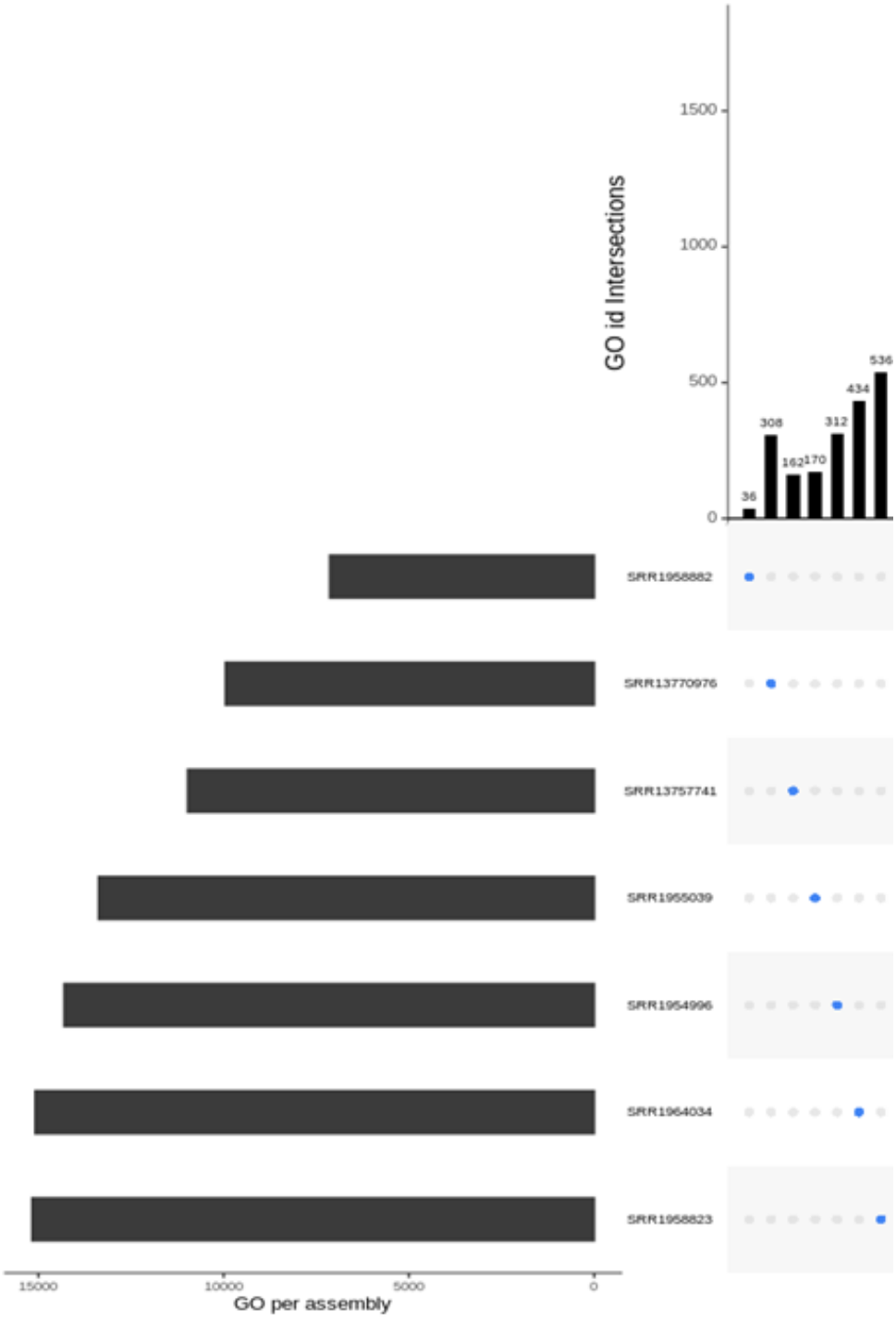


Fig. 10 – UpSet plot picturing the unique GO IDs for each of the 7 transcriptomes collected from various body parts of cone snails in the top bar chart, while also showing the total GO IDs attributed to each sample in the side bar chart. Organized in a crescent order of degree, this image illustrates the opposite edge of the sequence started in Fig. 8, where the IDs were organized in a decrescent order – the order meaning the number of intersections. Thus, in this plot it is possible to observe the genes with zero intersections (only dots without lines connecting), meaning the unique genes present in each of the transcriptomic samples.

Regarding the venom-related data, the 69 files possessed 2,104 IDs in common [Fig. 11]. These genes are not unique to the venom group, as they most of them are expected to be also expressed and found in the group of the GO IDs from the other tissues. The genes uniquely shared among venom-related transcriptomes are presented next, in 3.3.2.2.

At an individual level [Fig. 12], the striking realization is that not all samples have unique genes – only 57 of the 69 have. This set of 12 come from *C. miliaris* (7), *C. coronatus* (2), *C. ermineus* (1), *C. rattus* (1) and *C. ebraeus* (1). The sample with the greatest unique genes comes from the venom duct of a *C. lenavati* with 262. This is a very detached value, as the rest of the samples with the most numerous unique genes have around 50 to 90 and are from the venom ducts of various species including *C. betulinus*, *C. consors*, *C. triblei*, *C. litteratus*, as well as other *C. lenavati* venom duct samples. These were also the species with the higher number of predicted coding sequences and domain hits. So, by examining the results (see 7.6.), the suspected correlation becomes increasingly apparent: the increasing assembly size, which is accompanied by increasing GO IDs, is also accompanied by an increased number of unique GO IDs attributed. Indeed, by dividing the venom data in categories of size it is noticed:

a) all 12 assemblies without unique GO IDs have less than 50M;
b) only 9 out of 47 assemblies (meaning less than 20%) with less than 50M have unique GO ID content superior to the mean value;
c) 4 out of 8 assemblies (50%) with size between 50 and 75M have unique GO ID content superior to the mean value;
d) finally, all 14 assemblies (100%) with size superior to 75M have unique GO ID content superior to the mean value.

Thus, the correlation is confirmed. The greater the assembly size, the higher the number of predicted coding sequences and domain recognitions. Logically, with higher number of coding sequences detected, more GO IDs are attributed, and more unique genes are encountered. In turn, the smaller the assembly, the less unique genes are retrieved. Additionally, no foundation was encountered for a correlation between the number of unique genes and the sequencing instrument and technique utilized, as the sequencing techniques of the lesser size assemblies with less unique genes are mostly the same utilized in the data with the larger size and greatest number of unique genes found.

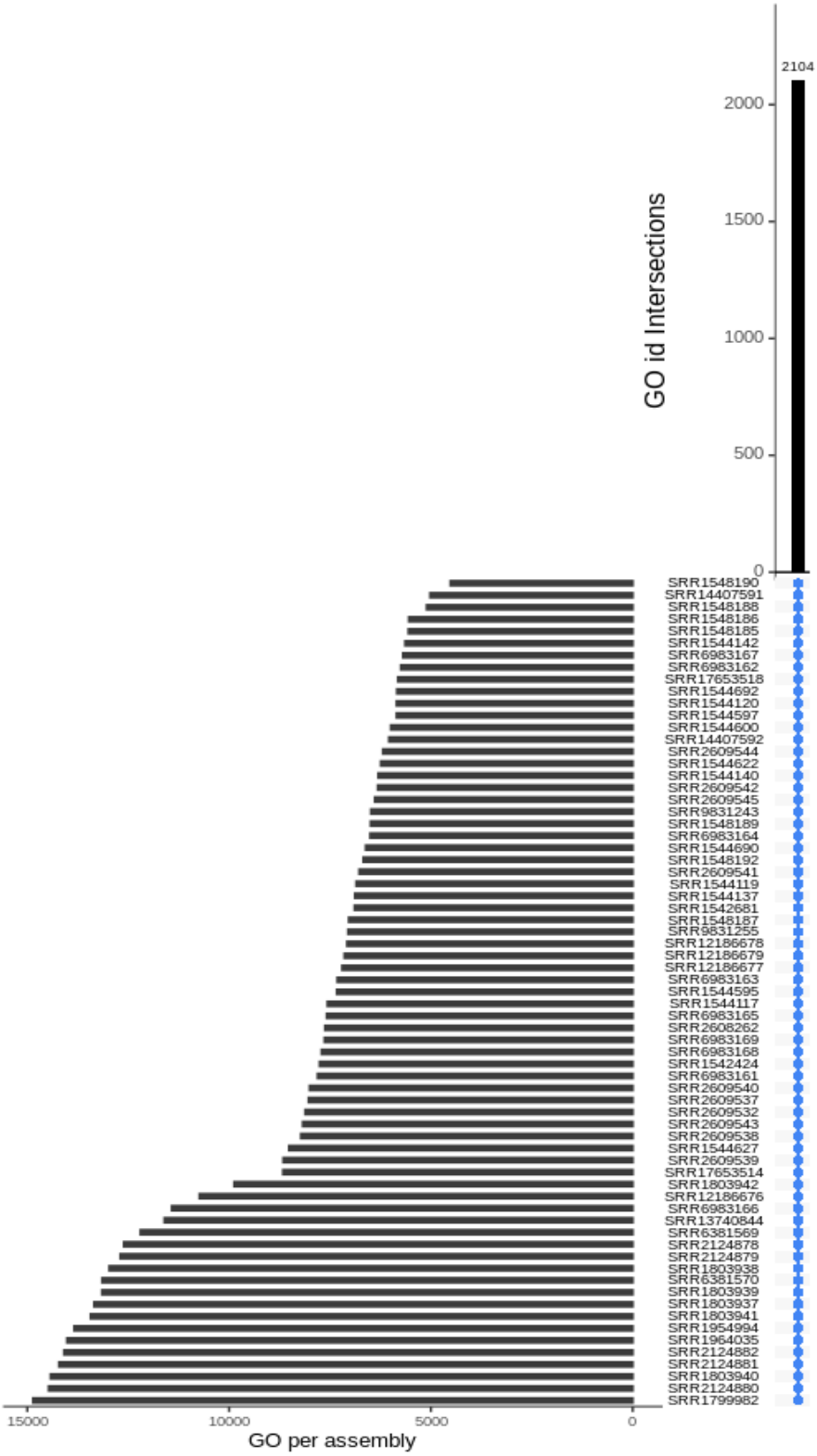


Fig. 11 – UpSet plot illustrating the common GO IDs for the 69 transcriptomes of the venom-related group in the top black vertical bar. The vertical blue line intersecting all 69 samples shows the 2,104 IDs shared by these transcriptomes. Additionally, in the side black bar chart are the total GO IDs attributed to each sample individually.

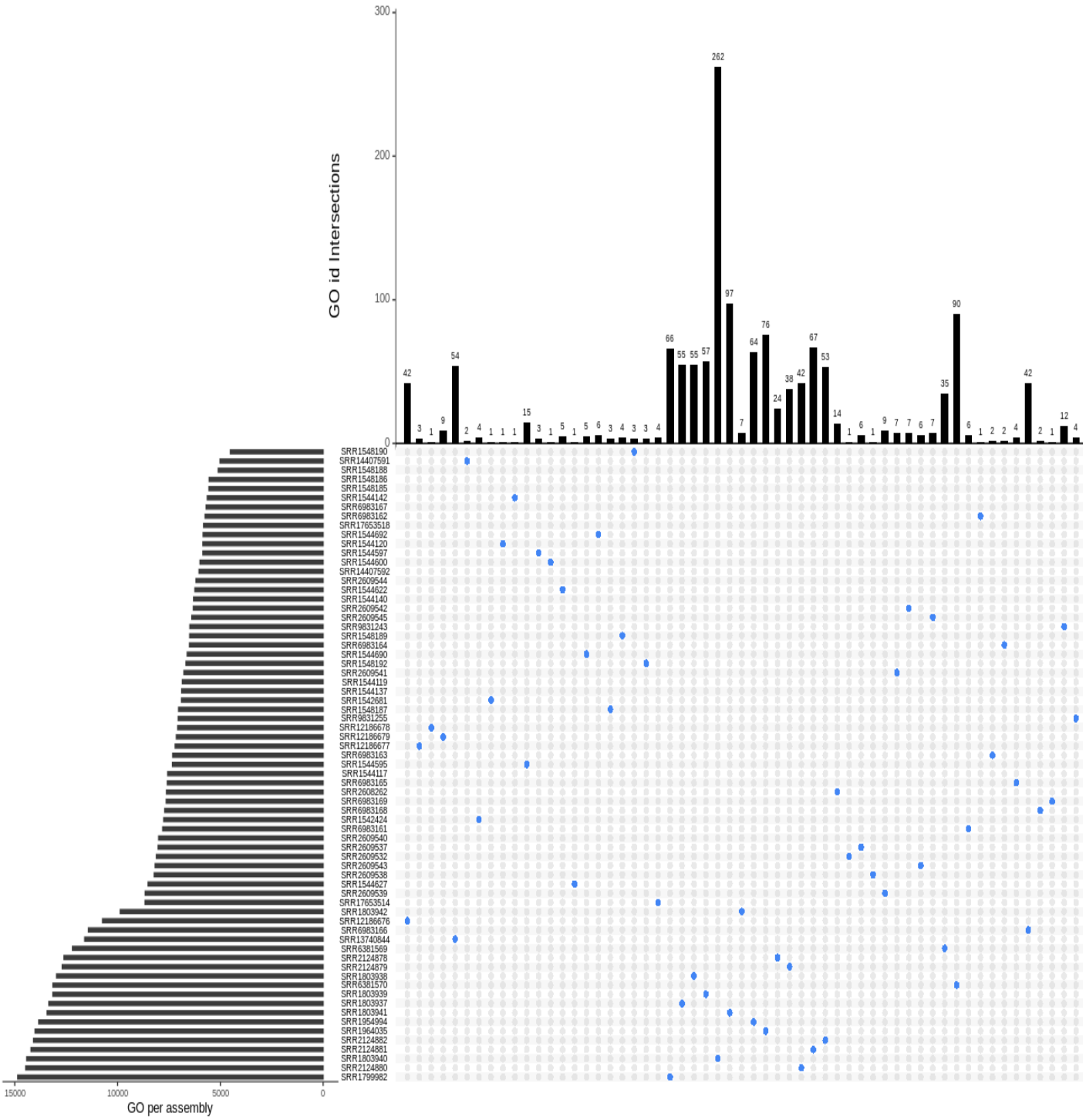


Fig. 12 – UpSet plot picturing the amount of unique GO IDs for each of the 69 transcriptomes collected from venom ducts, venom bulb and a venom gland of various cone snails in the top bar chart, while also showing the total GO IDs attributed to each sample in the side bar chart. Organized in a crescent order of degree, this image illustrates the opposite edge of the sequence started in Fig. 10, where the IDs were organized in a decrescent order – order meaning the number of intersections. Thus, in this plot it is possible to observe the genes with zero intersections (no lines connecting the blue dots), meaning the unique genes present in each of the samples.

3.3.2.2. Shared genes only among venom-related tissues

After excluding from the list of genes shared among venom related data (2,104) the genes shared among other tissues (5,913), the resulting list is composed of genes uniquely shared among venom related samples [Fig. 13]. Remarkably, this group numbers 29 genes. As listed below in the Table 1, of those 29, 19 are attributed to a biological process, 8 are associated with a molecular function, and 2 codify a cellular component.

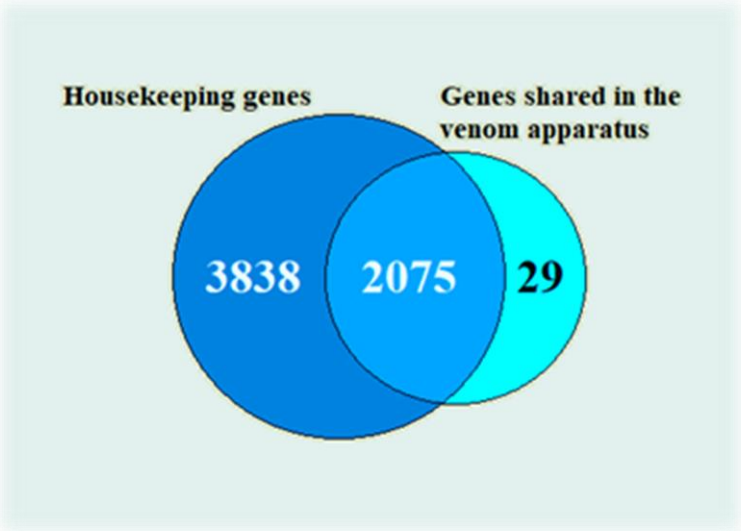


Fig. 13 – Venn diagram displaying the different groups of shared genes. In the left circle are the housekeeping genes (5,913) and in the right circle are the genes shared by the venom-related transcriptomes (2,104). The cross area indicates the number of genes shared in the venom related transcriptomes which are also expressed in other tissues (2,075). In this way, it is possible to visualise the group of shared genes expressed in various body parts but not the venom apparatus (3838) and most importantly the shared genes only expressed in the venom apparatus (29).

Table 1 – The 29 genes shared only by venom related tissues with their respective functions.

GO ID	GO Category	GO Term
GO:0000050	Biological process	Urea cycle
GO:0002003	Biological process	Angiotensin maturation
GO:0003081	Biological process	Regulation of systemic arterial blood pressure by renin-angiotensin
GO:0003084	Biological process	Positive regulation of systemic arterial blood pressure
GO:0006265	Biological process	DNA topological change
GO:0006807	Biological process	Nitrogen compound metabolic process
GO:0010157	Biological process	Response to chlorate
GO:0019882	Biological process	Antigen processing and presentation
GO:0030212	Biological process	Hyaluronan metabolic process
GO:0031288	Biological process	Sorocarp morphogenesis

GO:0042445	Biological process	Hormone metabolic process
GO:0051353	Biological process	Positive regulation of oxidoreductase activity
GO:0061515	Biological process	Myeloid cell development
GO:0071577	Biological process	Zinc ion transmembrane transport
GO:0097067	Biological process	Cellular response to thyroid hormone stimulus
GO:0140206	Biological process	Dipeptide import across plasma membrane
GO:1903052	Biological process	Positive regulation of proteolysis involved in protein catabolic process
GO:1903665	Biological process	Negative regulation of asexual reproduction
GO:1903669	Biological process	Positive regulation of chemorepellent activity
GO:0008239	Molecular function	Dipeptidyl-peptidase activity
GO:0008241	Molecular function	Peptidyl-dipeptidase activity
GO:0015174	Molecular function	Basic amino acid transmembrane transporter activity
GO:0016532	Molecular function	Superoxide dismutase copper chaperone activity
GO:0016671	Molecular function	Oxidoreductase activity; acting on a sulphur group of donors; disulphide as acceptor
GO:0031545	Molecular function	Peptidyl-proline 4-dioxygenase activity
GO:0043138	Molecular function	3'-5' DNA helicase activity
GO:0099106	Molecular function	Ion channel regulator activity
GO:0031597	Cellular component	Cytosolic proteasome complex
GO:1905103	Cellular component	Integral component of lysosomal membrane

The functions of these 29 genes are very much related to four main aspects:

a) DNA replication – e.g., GO:0006265 and GO:0043138 mark the presence of ubiquitous genes for replication of DNA in the venom tissues;
b) metabolism and blood pressure – e.g., GO:0003084, GO:0006807 and GO:1903052 show the tendency of shared genes to be related to energy management, mostly catalytic metabolic processes in addition to blood circulation;
c) ion channel and transportation regulation – e.g., GO:0071577 and GO:0099106 illustrate that ion channel and cell transportation is central in the shared genomics of venom tissues;
d) PTMs – e.g., GO:0019882, GO:0008239 and GO:0008241 indicate shared genes for posttranscriptional modifications.

In all, the most particular findings in this group of expressed genes shared across so diverse cone snails are the presence of numerous metabolic and transport regulators, as well as genes for the PTMs which are so characteristic of the venoms of *Conus* species. The only peculiar function is reported with GO:0031288 which is a gene for

sorocarp morphogenesis. The importance of all these findings, especially the bizarre GO:0031288 is debated in the Discussion section (see 4.3.).

3.3.2.3. DE of shared GO IDs in venom-related tissues

Although commonly expressed by all the venom tissues present in the dataset, the 29 GO IDs are not expressed to an equal extent in all samples. The Fig. bellow [Fig. 14] illustrates the quantitative expression for each of the 29 GO IDs normalized according to the 20 species present in the dataset (instead of quantified for each of the 69 venom-related transcriptomes). In this way, it is possible to accurately assess from a functional point of view the DE of the shared GO IDs across many *Conus* species.

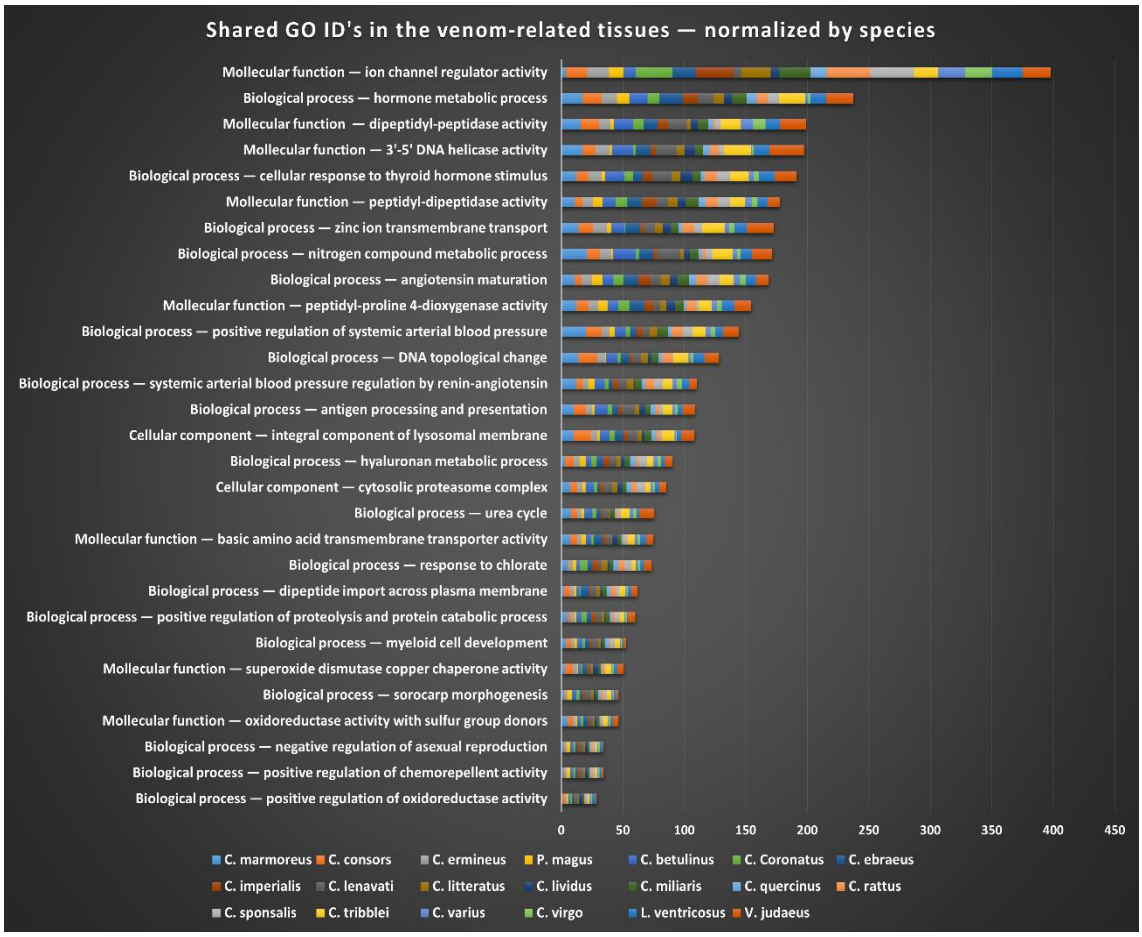


Fig. 14 – DE of the 29 genes commonly expressed in all venom-related transcriptomes normalized for the 20 species of cone snails present in the dataset.

The most strongly expressed gene in all *Conus* species studied is related to the regulation of ion channel activity, which is a molecular function consisting of modulating the activity of an ion channel via direct interaction with it. Next in line is a gene coding for hormone metabolism. This is a biological process related to chemical reactions involving any kind of hormone or naturally occurring substance secreted by specialized

cells, influencing the metabolism of other cells possessing functional receptors for the hormone. The third and fourth most expressed genes are again active on a molecular level: while the former is related to the catalytic process of N-terminal dipeptides hydrolysis from a polypeptide chain, the latter is involved in the anabolic process of unwinding the DNA helix in the direction 5' to 3' (which is driven by ATP hydrolysis). The fifth most expressed gene is involved in the complex biological process of cellular response to thyroid hormone stimulus. This is a process involving any change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) because of a stimulus from thyroid hormones. After these strongly expressed genes, there are moderately expressed genes related to: zinc ion transmembrane transport, chemical reactions involving organic or inorganic nitrogen containing compounds, angiotensin maturation, proline hydroxylase activity, positive regulation of systemic arterial blood pressure and DNA topological change. Lastly, the less expressed shared genes are involved in positive regulation of oxidoreductase and chemorepellent activity and negative regulation of asexual reproduction.

Furthermore, since each *Conus* species is mostly specialized in the predation of either fish, worms, or molluscs, it is in the greatest interest of this dissertation to discern the DE of the shared genes according to the three main feeding habits of cone snails. Luckily, in the 20 species of the dataset there are representatives of the three feeding habits as 16 are vermivorous (*C. betulinus*, *C. coronatus*, *C. ebraeus*, *C. imperialis*, *C. lenavati*, *C. litteratus*, *C. lividus*, *C. miliaris*, *C. quercinus*, *C. rattus*, *C. sponsalis*, *C. tribblei*, *C. varius*, *C. virgo*, *C. ventricosus*, and *C. judaeus*), 3 are piscivorous (*C. consors*, *C. ermineus* and *C. magus*) and 1 is molluscivorous (*C. marmoreus*). In this wise, the DE of the shared genes was normalized according to the feeding habits [Fig. 15] (also see 7.7.). From this second expression analysis, the general DE tendency for all genes observed in the previous species analysis is mostly maintained. In fact, most genes keep the same position in this second quantitative expression assessment. However, while the gene for ion channel regulation maintains the top expression position, now it is possible to discern the level of expression for this gene that each diet demands. This is important since now the evidence puts the top expressed gene in general for all species in the opposite end of the expression level for molluscivorous species (only about 10% of the top horizontal bar is from the molluscivorous species). Moreover, this gene's expression level is now closely followed by other expression levels of genes like the ones involved in hormone and nitrogen compound metabolic processes. The reason for this deviation might be due to most species (80%) of the dataset being vermivorous, with only three piscivorous and one molluscivorous. In this

way, the previous results were misleading in terms of biological importance in the broader spectrum of *Conus* species. Thus, these results indicate that a feeding habit analysis is important in the context of the of the *Conus* venom' general characteristics.

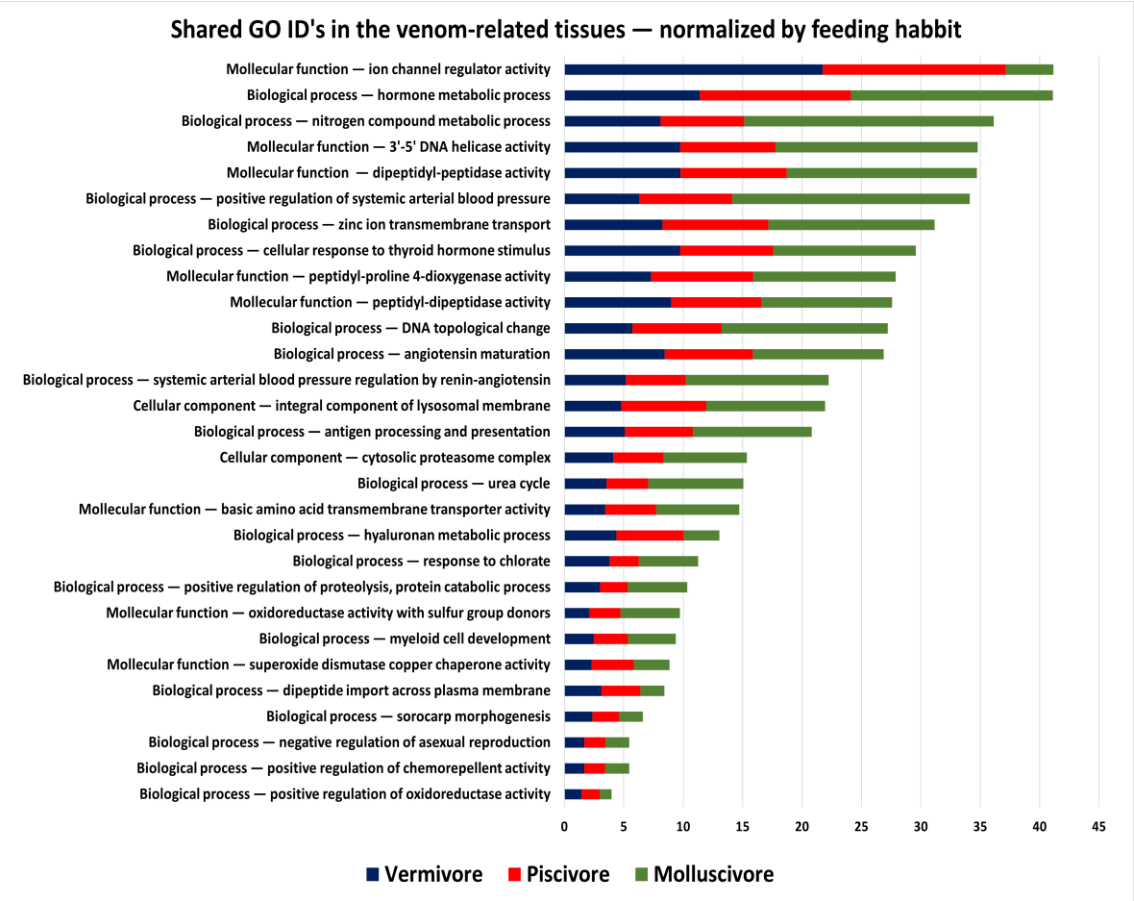


Fig. 15 – DE of the 29 GO IDs commonly expressed in all venom-related transcriptomes normalized for the 3 different feeding habits of cone snails.

3.4. Relationship between the *Conus'* venom and SARS-Cov-2

This research found no direct relationship between the venom of the 20 *Conus* species present in this study and the SARS-Cov-2 virus. Indeed, the annotation of the *Conus* transcriptome against Tox-Prot and the Conopeptides databases produced the desired output to compare with the annotated spike protein and full genome of the virus, but the comparison yielded no results at all.

Nevertheless, results for the annotation of the full genome of the coronavirus with the Pfam database (see 7.8.) show the success of the methodology in correctly finding and reporting proteins present in the genome of the virus. In this wise, if any major or direct relationship between these two genomes existed, it would have likely been found with the designed approach.

4. Discussion

4.1. Duplication levels in the assembled transcriptomes

Unusually high level of duplicated matching genes was detected with the completeness assessments made for the assembly of the transcriptome (see 7.4.). Since BUSCO's basis lies in research with single copy orthologs, the rampant presence of greater-than-expected numbers for duplication levels throughout the whole dataset may be an important finding. According to these results, the presence of many similar orthologs may be rooted in a series of naturally occurring duplication events in many *Conus* species.

A previous study made in 1999 by Thomas F. Duda, Jr. and Stephen R. Palumbi in "Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*" already indicated the presence of perpetually high duplication levels for certain venom-related genes in *Conus* species (132). This study was conducted on *C. abreviattus* and *C. lividus*. two (deemed at the time being) distant related species. An exemplar of *C. lividus* is present on this dissertation's dataset, however the highest level of duplicated genes is associated to samples of *C triblei*, *C. betulinus* and especially *C. consors*. More recently, two research articles titled "Evolution of *Conus* Peptide Genes: Duplication and Positive Selection in the A-Superfamily" (133), and "Extensive and Continuous Duplication Facilitates Rapid Evolution and Diversification of Gene Families" (134) further cemented the suspected multiple gene duplication events occurring in cone snails. Most importantly, these two studies were conducted with various species also present in the dataset of this dissertation including *C. magus*, *C. lividus,* and especially *C. consors*. The presence of *C. consors* in those studies explains the origin and supports the existence of the highest duplication levels. As an answer to these duplication levels, the authors of the latter articles suggested that these processes facilitated the rapid evolution and prompted the drastic difference found in the venom compositions of these snails, linking these duplication events to evolutionary responses of predator-prey interactions.

The findings in the completeness assessments for the transcriptome assembly made in the research work of this dissertation are aligned with previous results and conclusions of diverse studies regarding *Conus* species. This work may inclusively further corroborate the possibility of duplication events in these marine gastropods in general, and particularly in *C. consors*.

4.2. Correlation of assembly size and unique genes found

The results show that unique genes found per transcriptome assembly are far more numerable when the assembly size is greater [Fig. 16] (see 7.6.). Basically, as the assembly size grows, the number of predicted coding sequences and domain recognitions also increases sharply. Logically, with higher number of coding sequences detected, more complete genes are encountered. Naturally, the probability of finding a unique gene increases when the number of genes is greater.

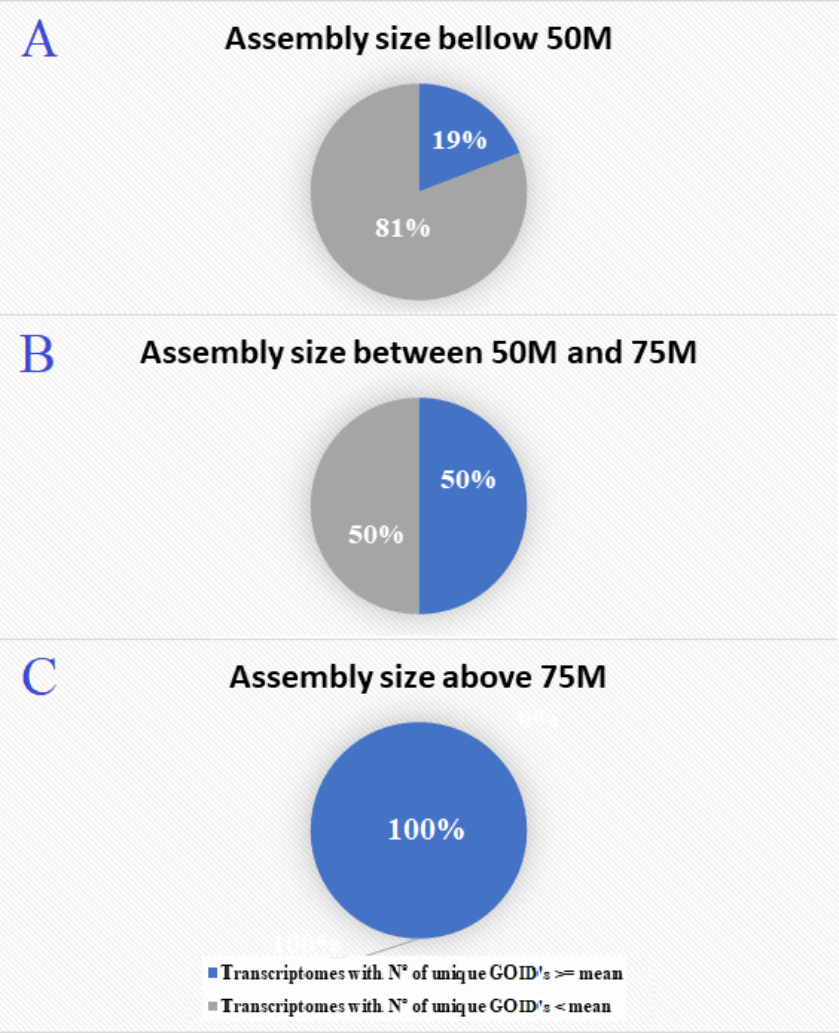


Fig. 16 – Three circular charts illustrating the percentage of transcriptomes, divided by assembly size, with their respective number of unique genes (GO IDs) compared to the mean value. A – chart representing the transcriptomes with assembly size below 50M and their respective variable number of unique GO IDs relatively to the mean; B – chart representing the transcriptomes with assembly size from 50 up to 75M and their respective variable number of unique GO IDs relatively to the mean; C – chart representing the transcriptomes with assembly size greater than 75M and their respective variable number of unique GO IDs relatively to the mean.

Thus, a larger assembly directly influences the chance of retrieving a unique gene from each sample. On the contrary, with smaller assemblies, less or possibly even no unique genes are retrieved. Unfortunately, this realization has a vastly negative effect

on the species with just one transcriptome assembly, as the risk of no unique genes retrieved from that species is greater. In fact, it happens on this dataset as there was only one transcriptome from *C. rattus* and it had no unique genes. There was a total of 12 samples from which no unique genes were found, but they belonged to species from which there were more samples available, significantly reducing the beforementioned danger. On a final note, this correlation is seemingly independent of sequencing instrument and technique, as the sequencing techniques of the assemblies a smaller size having less unique genes are mostly the same utilized in the sequencing of the data with the larger size having the greatest number of unique genes found. Indeed, the greater number of valuable transcriptomic content seems to depend solely on the greater size of the assembly.

Ultimately, the logical recommendation for future studies is clear: studies regarding the research for novel genes and proteins should focus primarily on larger sequencing and greater genomic content per sample. The recommended assembly size which should be aimed for according to the empirical results obtained in this dissertation is 75M, with the minimum recommended size being above 50M. Beyond the scope of the research efforts for *Conus* species, the principle underlined with this correlation assumes massive importance for the broader venom's research field. One of the main objectives – often the single objective (135) (136) (137) – in many projects and studies dealing with venom and venom-related subjects is the finding and report of novel genes, proteins, and toxins. It is in the interest of all scientific projects having this objective to strengthen all variables at play to ensure the discovery of novel genomic and proteomic content.

4.3. Venom genes shared by *Conus* species

The group of 29 GO IDs shared by the venom-related tissues of 20 *Conus* species possibly represents a conserved genomic repertory of venom synthesis for all cone snails. From the DE analysis normalized by the feeding habit [Fig. 15] it is now possible to understand which commonly expressed genes are more important in each predatory diet. Moreover, encountering quantitative differences in the expression levels of the shared genomic content by all *Conus* species provides another insight to the evolution of these marine snails. These results seem to point to a divergent evolution in which each predatory diet corresponds to a clear clade only including species with a specific feeding habit. A phylogeny study made for the 20 species present on the dataset [Fig. 17] further solidifies this conclusion.

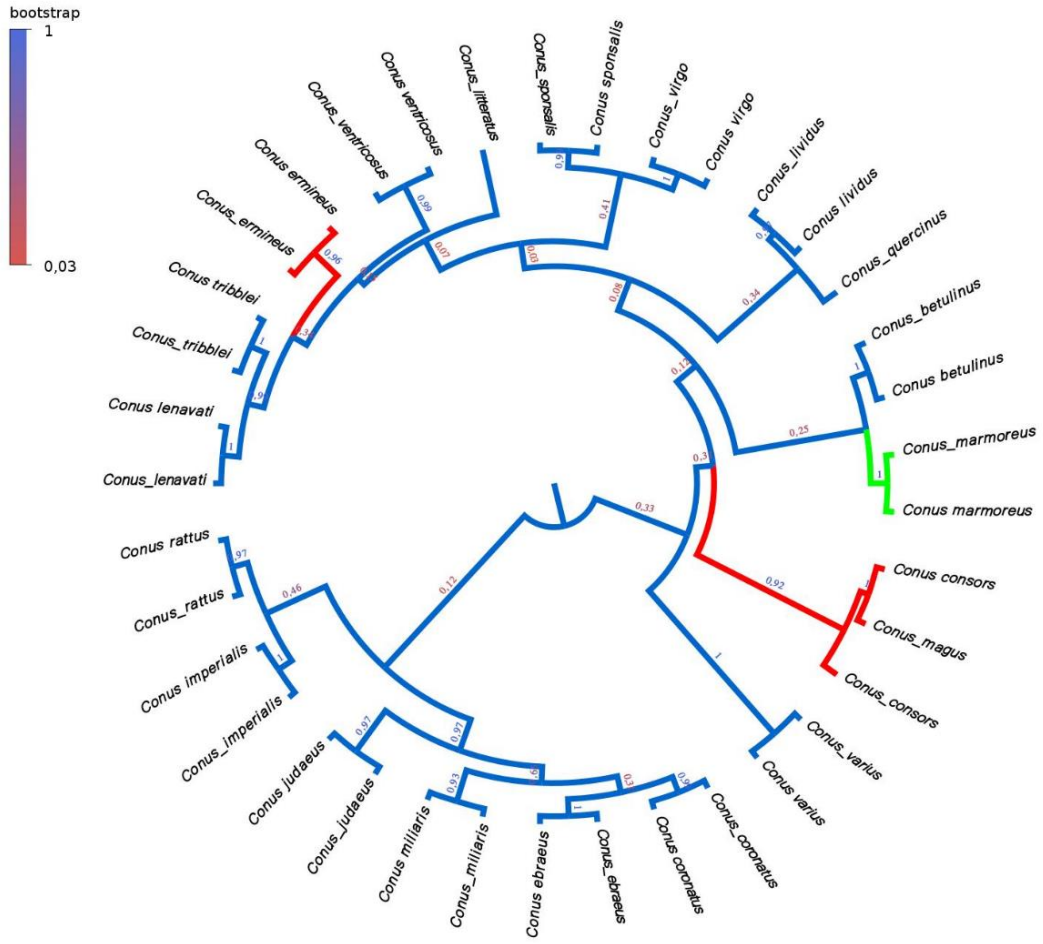


Fig. 17 – Maximum likelihood phylogenetic tree of 20 Conus species constructed using two rRNA 16S genes for each species. The branch colours represent feeding habits: blue for vermivorous, red for piscivorous, and green for molluscivorous. Bootstrap values are written next to the branches in a colour reflecting their number according to the legend in the top left corner.

One possible weak point for this study could be the number of species present in this dissertation's dataset, which corresponds roughly to 2% of the total existing species of cone snails (the genus *Conus* has more than 700 species). However, the phylogeny results are very much aligned with other phylogeny studies such as the one elaborated by Yihe Zhao and Agostinho Antunes in "Biomedical Potential of the Neglected Molluscivorous and Vermivorous *Conus* Species" which connected 350 *Conus* (56). A comparison of that phylogenetic tree with the one made in this work shows that in fact the 20 species present in this research provide a good overall coverage of the diversity of *Conus* species, having representatives for all different major clades. In addition, the lower bootstrap values on the phylogeny study made in this dissertation can be explained by the relatively small number of species present. As there is a huge number of species missing to link the more distant branches, the bootstrap value is compromised, but apparently the obtained results are not.

Concerning the proper gene functions, nearly all findings are fitted in the 4 categories previously stated: ion channel and transportation regulation, metabolism and blood pressure, DNA replication, and PTMs. Regulation of ion channel, molecular transportation, metabolism, and blood pressure all logically align in view of the biological logistics for venom synthesis. The gene for downregulating another metabolic-heavy process like reproduction also makes sense in this perspective. On the other hand, the genes related to posttranscriptional modifications are an integral part of the genomic background for the venom toxins refinery process. In all, most of the genes are related to metabolism – coding catabolic reactions to create energy for the anabolic reactions needed for venom synthesis (DNA transcription and translation, as well as PTMs) – and maturation of the venom's proteins.

However, the finding of a gene for sorocarp morphogenesis was completely unexpected. The finding of a gene related to sporulation in fungi would be intriguing enough even if it appeared only on one transcriptome, but it was a common gene expressed in all 69 venom samples. Surprisingly, a study made in 2009 revealed that "microhabitats within venomous cone snails contain diverse actinobacteria" (138). That research exposed the existence of a thriving actinomycete community living in a seemingly symbiotic relationship with cone snails. The results of that study suggest that certain species of symbionts may be commonly found in all snails, where they are needed to perform identical tasks required by all host snails. According to the research, the microhabitats within the venom gland where these symbioses happen are even suitable sources for studies aiming at drug discovery. Additionally, the study hypothesizes that while some are shared, other groups of symbionts may be specific to their host snail, having been selected by their host to fit its biology. Conclusively, studying more cone snails is essential to test these hypotheses. However, under this perspective, the discovery of the GO:0031288 in all samples assumes a vital importance as it supports the findings of symbiotic microorganisms existing in the venom apparatus of *Conus* species.

4.4. Relationship between *Conus'* venoms and SARS-Cov-2

Despite similarities between the viral the furin cleavage site of the Spike protein and certain conotoxins reported mainly in "A nicotinic hypothesis for Covid-19 with preventive and therapeutic implications" (84), "A potential interaction between the SARS-CoV-2 spike protein and nicotinic acetylcholine receptors" (70), and "Omicron and Alpha P680H block SARS-CoV2 spike protein from accessing cholinergic inflammatory pathway via α9-nAChR mitigating the risk of MIS-C" (86), no genomic link

connecting the two parts was found in this work. This result thoroughly indicates the absence of direct matching genes or proteins. However, since the described similarities in the aforementioned studies were on the protein domain level, further similarities between the venom of cone snails and the coronavirus may exist, but only on that level. In this work, the adopted methodology was not designed for protein domain level comparisons, but in contrast, the comparisons between the genomes of the two entities revealed no direct relationship.

In this sense, further studies should be conducted on this topic but using a proteomic approach, searching for matching conotoxin domains instead of a methodology for matching transcriptomics. Such approach should also focus on the role of the proteomic mechanisms behind the transformation of conotoxins, as well as the PTMs of the conotoxins with domains already found to match those of the Spike protein of SARS-Cov-2.

5. Conclusion

Cone snails are diverse predatory creatures with a sophisticated, powerful venom developed to hunt various prey in accordance with their habitat. Extremely complex and distinct, the treasured neuroactive toxins present in the venom cocktail are sought after for a variety of biomedical applications. Despite the high relevance for science and health however, the *Conus* species ecology and proper feeding habits are still insufficiently studied. The severe lack of studies in these themes coupled with deficient genomic resources undermines efforts made to better understand variability patterns in the venom's composition. Moreover, insufficient transcriptomic material may limit biological interpretations. Acknowledging these realities, the methodology was designed to highlight the relationships present in the venoms of all *Conus* species. This was accomplished by focusing on the transcriptomic sequences available in the entire genus, rather than only on those of just one species or one individual.

First and foremost, this research found a group of 29 GO IDs that 20 cone snail species express exclusively in the venom gland. This unprecedented discovery suggests a potentially preserved genomic repertory for venom synthesis present in all *Conus* species. The phylogenetic relationship among those 20 species supports this suggestion. Even though the number of species analysed amounts to just 2% of the total number of existing *Conus* species, there is a good overall coverage of all feeding habits and major phylogenetic clades. Furthermore, transcriptomic evidence of symbiotic relationships within the venom gland was detected with the surprising finding of the shared GO:0031288. This exiting discovery seems to point to the existence of symbiotic microorganisms in all species, as the GO ID is present in the venom glands of all cone snails. Additionally, by unveiling a correlation of assembly size with unique genes found, the research work on this dissertation provides empirical values for assisting future sequencing improvements. Finally, it is also reported that no direct genomic link was found between conotoxins and the SARS-Cov-2 virus, despite increasing studies connecting both parts at the protein domain level.

Future studies along this axis should also consider the expression levels of genes for all transcriptomes, which were not included in this research due to technical limitations (computer storage space). Following the assembly process, a DE analysis should be conducted on all transcriptomic samples to provide a strong basis to evaluate and validate further the findings of any unique and/or shared genes encountered. In addition to this, a genome-backed assembly should also be performed, ideally using the two *Conus* species – *C. betulinus* and *C. ventricosus* – genomes sequenced and

now available (42) (43). One or two genome-backed assemblies followed by a DE analysis conducted on all transcriptomes assembled coupled with the annotation and bioanalysis processes performed with the adopted methodology in this dissertation would be more accurate and balanced, ultimately strengthening the results.

In conclusion, through a transcriptomics approach and a strategy never attempted before in the study of *Conus* species, this work succeeded in contributing to further decipher the genomics behind the complex predatory venom of these fascinating marine gastropods.

6. Bibliography

1. *Cone Shell Stings. Recent Cases of Human Injury due to Venomous Marine Snails of the Genus Conus.* AJ, Kohn. s.l. : Hawaii Med. J., 1958, Vol. 17, p. 528.

2. *A Venomics Approach to the Identification and Characterization of Bioactive Peptides from Animal Venoms for Colorectal Cancer Therapy: Protocol for a Proof-of-Concept Study.* Shahzadi SK, Karuvantevida N, Banerjee Y. 12, s.l. : JMIR Res Protoc., 2021, Vol. 10.

3. *Nature and applications of scorpion venom: an overview.* Saadia Tobassum, Hafiz Muhammad Tahir, Muhammad Arshad, Muhammad Tariq Zahid, Shaukat Ali & Muhammad Mohsin Ahsan. 3, s.l. : Toxin Reviews, 2020, Vol. 39, pp. 214-225.

4. *Bee Venom: An Updating Review of Its Bioactive Molecules and Its Health Applications.* Carpena M, Nuñez-Estevez B, Soria-Lopez A, Simal-Gandara J. 11, s.l. : Nutrients, 2020, Vol. 12.

5. *Cellular targets and molecular activity mechanisms of bee venom in cancer: recent trends and developments.* Ayşegül Varol, Serap Sezen, Dilhan Evcimen, Atefeh Zarepour, Gönül Ulus, Ali Zarrabi, Gamal Badr, Sevgi Durna Daştan, Asya Gülistan Orbayoğlu, Zeliha Selamoğlu & Mehmet Varol. s.l. : Toxin Reviews, 2022.

6. *Overview of Apitherapy Products: Anti-Cancer Effects of Bee Venom Used in Apitherapy.* Şengül, F. & Vatansev, H. 1, s.l. : International Journal of Traditional and Complementary Medicine Research, 2021, Vol. 2, pp. 36-48.

7. *Snake venom components and their applications in biomedicine.* Koh, D.C.I., Armugam, A. & Jeyaseelan. s.l. : Cell. Mol. Life Sci., 2006, Vol. 63, pp. 3030–3041.

8. *Snake venom toxins: toxicity and medicinal applications.* Chan YS, Cheung RCF, Xia L, Wong JH, Ng TB, Chan WY. 14, s.l. : Appl Microbiol Biotechnol., 2016, Vol. 100, pp. 6165-6181.

9. *Treatment of venous ulcers with fibrin sealant derived from snake venom.* Gatti, M. A. N., et al. s.l. : Journal of Venomous Animals and Toxins including Tropical Diseases, 2011, Vol. 17, pp. 226-229.

10. *A New Fibrin Sealant from Crotalus durissus terrificus Venom: Applications in Medicine.* Barros LC, Ferreira RS Jr, Barraviera SR, et al. 8, s.l. : J Toxicol Environ Health B Crit Rev., 2009, Vol. 12, pp. 553-571.

11. *Origins of diverse feeding ecologies within Conus, a genus of venomous marine gastropods.* Duda Jr, Thomas F., Alan J. Kohn, and Stephen R. Palumbi. s.l. : Biol. J. Linn. Soc., 2001, Vol. 73, pp. 391–409.

12. *One, four or 100 genera? A new classification of the cone snails.* Puillandre, N., Duda, T. F., Meyer, C., Olivera, B. M., & Bouchet, P. s.l. : J. Molluscan Stud., 2015, Vol. 81, pp. 1–23.

13. *Cone snails: A big store of conotoxins for novel drug discovery.* Gao, B., Peng, C., Yang, J., Yi, Y., Zhang, J., & Shi, Q. 397, s.l. : Toxins, 2017, Vol. 9.

14. *The role of defensive ecological interactions in the evolution of conotoxins.* Prashanth, J. R., S. Dutertre, A. H. Jin, V. Lavergne, B. Hamilton, F. C. Cardoso, J. Griffin, D. J. Venter, P. F. Alewood, and R. J. Lewis. s.l. : Mol. Ecol., 2016, Vol. 25, pp. 598–615 .

15. *Panorama sur La Diversite des Conidae 110 Espèces Prédatrices des Plus Efficaces.* Richard Georges, Michael Rabiller. 2021.

16. *The venom apparatus of Conus magus.* Endean, R., and Claudine Duchemin. s.l. : Toxicon, 1967, Vol. 4, pp. 275–284.

17. *Evolution of separate predation- and defence-evoked venoms in carnivorous cone snails.* Dutertre, S., Jin, A.H., Vetter, I., Hamilton, B., Sunagar, K., Lavergne, V., Dutertre, V., Fry, B.G., Antunes, A., Venter, D.J. and Alewood, P.F. 3521, s.l. : Nat. Commun., 2014, Vol. 5.

18. *Conotoxins: Chemistry and Biology.* Jin, Ai-Hua, Markus Muttenthaler, Sebastien Dutertre, S. W. A. Himaya, Quentin Kaas, David J. Craik, Richard J. Lewis, and Paul F. Alewood. 119, s.l. : Chem. Rev., 2019, pp. 11510–11549.

19. *Anatomical correlates of venom production in Conus californicus.* Marshall, Jennifer, Wayne P. Kelley, Stanislav S. Rubakhin, Jon-Paul Bingham, Jonathan V. Sweedler, and William F. Gilly. 417, s.l. : Biol. Bull.; Mar. Drugs, 2002; 2018, Vol. 203; 16, pp. 27–41; 14 of 20.

20. *Venom kinematics during prey capture in Conus: The biomechanics of a rapid injection system.* Salisbury, S. Michael, Gary G. Martin, William M. Kier, and Joseph R. Schulz. s.l. : J. Exp. Biol., 2010, Vol. 213, pp. 673–682. .

21. *Constant and hypervariable regions in conotoxin propeptides.* Woodward, S. R., L. J. Cruz, B. M. Olivera, and D. R. Hillyard. s.l. : EMBO J., 1990, Vol. 9, pp. 1015–1020. .

22. *Optimized deep-targeted proteotranscriptomic profiling reveals unexplored Conus toxin diversity and novel cysteine frameworks.* Lavergne, Vincent, Ivon Harliwong, Alun Jones, David Miller, Ryan J. Taft, and Paul F. Alewood. 29 , s.l. : Proceedings of the National Academy of Sciences, 2015, Vol. 112.

23. *Discovery, synthesis, and structure: Activity relationships of conotoxins.* Akondi, Kalyana B., Markus Muttenthaler, Sébastien Dutertre, Quentin Kaas, David J. Craik, Richard J. Lewis, and Paul F. Alewood. s.l. : Chem. Rev., 2014, Vol. 114, pp. 5815–5847.

24. *Conotoxins and the posttranslational modification of secreted gene products.* Buczek, O., G. Bulaj, and B. M. Olivera. s.l. : Cell. Mol. Life Sci., 2005, Vol. 62, pp. 3067–3079.

25. *Small Molecules in the Cone Snail Arsenal.* Neves, Jorge LB, Zhenjian Lin, Julita S. Imperial, Agostinho Antunes, Vitor Vasconcelos, Baldomero M. Olivera, and Eric W. Schmidt. s.l. : Org. Lett., 2015, Vol. 17, pp. 4933–4935.

26. *Various Conotoxin Diversifications Revealed by a Venomic Study of Conus flavidus.* Lu, Aiping, Longjin Yang, Shaoqiong Xu, and Chunguang Wang. s.l. : Mol. Cell. Proteom., 2014, Vol. 13, pp. 105–118.

27. *Transcriptomic messiness in the venom duct of Conus miles contributes to conotoxin diversity.* Jin, Ai-hua, Sebastien Dutertre, Quentin Kaas, Vincent Lavergne, Petra Kubala, Richard J. Lewis, and Paul F. Alewood. s.l. : Mol. Cell. Proteom., 2013, Vol. 12, pp. 3824–3833.

28. *Screening for post-translational modifications in conotoxins using liquid chromatography/mass spectrometry: An important component of conotoxin discovery.* Jakubowski, Jennifer A., Wayne P. Kelley, and Jonathan V. Sweedler. s.l. : Toxicon, 2006, Vol. 47, pp. 688–699.

29. *Conopeptide characterization and classifications: An analysis using ConoServer.* Kaas, Quentin, Jan-Christoph Westermann, and David J. Craik. s.l. : Toxicon, 2010, Vol. 55, pp. 1491–1509.

30. *Conotoxin Diversity in Chelyconus ermineus (Born, 1778) and the Convergent Origin of Piscivory in the Atlantic and Indo-Pacific Cones.* Abalde, Samuel, Manuel J. Tenorio, Carlos ML Afonso, and Rafael Zardoya. 10, s.l. : Genome biology and evolution, 2018, Vol. 10, pp. 2643-2662.

31. *In Snails: Biology, Ecology and Conservation; Gotsiridze-Columbus, N., Ed.* Emil M Hämäläinen, Sofia Järvinen. New York : Nova Science Publishers Inc, 2012.

32. *Novel venom peptides from the cone snail Conus pulicarius discovered through next-generation sequencing of its venom duct transcriptome.* Lluisma, Arturo O., Brett A. Milash, Barry Moore, Baldomero M. Olivera, and Pradip K. Bandyopadhyay. s.l. : Mar Genomics. Mar. Genom., 2012, Vol. 5, pp. 43–51.

33. *Deep venomics reveals the mechanism for expanded peptide diversity in cone snail venom.* Dutertre, Sebastien, Ai-hua Jin, Quentin Kaas, Alun Jones, Paul F. Alewood, and Richard J. Lewis. s.l. : Mol. Cell. Proteom., 2013, Vol. 12, pp. 312–329.

34. *Evolutionary diversification of multigene families: Allelic selection of toxins in predatory cone snails.* Duda Jr, Thomas F., and Stephen R. Palumbi. s.l. : Mol. Biol. Evol., 2000, Vol. 17, pp. 1286–1293.

35. *Identifying novel conopepetides from the venom ducts of Conus literatus through integrating transcriptomics and proteomics.* Zhang, Han, Yonggui Fu, Lei Wang, Anwen Liang, Shangwu Chen, and Anlong Xu. s.l. : J. Proteom., 2019, Vol. 192, pp. 346–357.

36. *From deadly venoms to novel therapeutics.* Shen, Gregory S., Richard T. Layer, and R. Tyler McCabe. s.l. : Drug Discov. Today, 2000, Vol. 5, pp. 98–106.

37. *Cone venom—From accidental stings to deliberate injection.* McIntosh, J. Michael, and Robert M. Jones. s.l. : Toxicon, 2001, Vol. 39, pp. 1447–1451.

38. *Drugs from the sea: Conopeptides as potential therapeutics.* Livett, Bruce G., Ken R. Gayler, and Zeinab Khalil. s.l. : Curr. Med. Chem., 2004, Vol. 11, pp. 1715–1723.

39. *Cone snail venomics: From novel biology to novel therapeutics.* ashanth, Jutty Rajan, Andreas Brust, Ai-Hua Jin, Paul F. Alewood, Sebastien Dutertre, and Richard J. Lewis. s.l. : Future Med. Chem., 2014, Vol. 6, pp. 1659–1675.

40. *Conotoxins: Natural product drug leads.* Halai, Reena, and David J. Craik. s.l. : Nat. Prod. Rep., 2009, Vol. 26, pp. 526–536.

41. *Ziconotide: Neuronal calcium channel blocker for treating severe chronic pain.* Miljanich, G. P. s.l. : Curr. Med. Chem., 2004, Vol. 11, pp. 3029–3040.

42. *The first Conus genome assembly reveals a primary genetic central dogma of conopeptides in C. betulinus.* Peng, C., Huang, Y., Bian, C., Li, J., Liu, J., Zhang, K., You, X., Lin, Z., He, Y., Chen, J. and Lv, Y. 11, s.l. : Cell Discov., 2021, Vol. 7.

43. *The genome of the venomous snail Lautoconus ventricosus sheds light on the origin of conotoxin diversity.* Pardos-Blas, J.R., Irisarri, I., Abalde, S., Afonso, C.M., Tenorio, M.J. and Zardoya, R. s.l. : Gigascience, 2021, Vol. 10.

44. *Advances in chromosomal studies of gastropod molluscs.* THIRIOT-QUIÉVREUX, C.A.T.H.E.R.I.N.E. s.l. : J. Molluscan Stud., 2003, Vol. 69, pp. 187–202.

45. *Karyological analysis and FISH physical mapping of 18S rDNA genes, (GATA)n centromeric and (TTAGGG)n telomeric sequences in Conus magus Linnaeus, 1758.* Dalet, J.T., Saloma, C.P., Olivera, B.M. and Heralde, F.M. s.l. : J. Molluscan Stud., 2014, Vol. 81, pp. 274–289.

46. *Cytogenetic studies on metaphase chromosomes of eight gastropod species of orders Mesogastropoda and Neogastropoda from the Red Sea (Prosobranchia-Mollusca).* Ebied, A.M., Hassan, H.A., Abu-Almaaty, A.H. and Yasen, A.E. s.l. : J. Egypt. Ger. Soc. Zool., 2000, Vol. 33, pp. 317–336.

47. *Transcriptomic-Proteomic Correlation in the Predation-Evoked Venom of the Cone Snail, Conus imperialis.* Jin, A.H., Dutertre, S., Dutt, M., Lavergne, V., Jones, A., Lewis, R.J. and Alewood. 3, s.l. : Marine drugs, 2019, Vol. 17, p. p.177.

48. *Intraspecies variability and conopeptide profiling of the injected venom of Conus ermineus.* Rivera-Ortiz, J.A., Cano, H. and Marí, F. s.l. : Peptides, 2011, Vol. 32, pp. 306–316.

49. Dutertre, S., Biass, D., Stöcklin, R. and Favreau, P. *Dramatic intraspecimen variations within the injected venom of Conus consors: An unsuspected contribution to venom diversity.* s.l. : Toxicon, 2010. pp. 1453-1462. Vol. 55.

50. *Venom-related transcripts from Bothrops jararaca tissues provide novel molecular insights into the production and evolution of snake venom.* Junqueira-de-Azevedo, I.L., Bastos, C.M.V., Ho, P.L., Luna, M.S., Yamanouye, N. and Casewell, N.R. s.l. : Mol Biol Evol, 2015, Vol. 32, pp. 754–66.

51. *Platypus venom genes expressed in non-venom tissues.* Whittington, C.M. and Belov, K. s.l. : Aust J Zool, 2009, Vol. 57, pp. 199–202.

52. *Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom.* Reyes-Velasco, J., Card, D.C., Andrew, A.L., Shaney, K.J., Adams, R.H., Schield, D.R., Casewell, N.R., Mackessy, S.P. and Castoe, T.A. s.l. : Mol Biol Evol, 2015, Vol. 32, pp. 173–83.

53. *Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins.* Hargreaves, A.D., Swain, M.T., Hegarty, M.J., Logan, D.W. and Mulley, J.F. s.l. : Genome Biol Evol, 2014, Vol. 6, pp. 2088– 95.

54. *Venom variation during prey capture by the cone snail, Conus textile.* Prator, C.A., Murayama, K.M. and Schulz, J.R. s.l. : PLoS ONE, 2014, Vol. 9.

55. *Geographic variation in venom allelic composition and diets of the widespread predatory marine gastropod Conus ebraeus.* Duda Jr, T.F., Chang, D., Lewis, B.D. and Lee, T. s.l. : PLoS ONE, 2009, Vol. 4.

56. *Biomedical Potential of the Neglected Molluscivorous and Vermivorous Conus Species.* Zhao, Y. and Antunes, A. 2, s.l. : Marine drugs, 2022, Vol. 20, p. p.105.

57. *ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database.* Kaas, Q., Yu, R., Jin, A.H., Dutertre, S. and Craik, D.J. s.l. : Nucleic Acids Res., 2011, Vol. 40, pp. D325–D330.

58. *Therapeutic potential of conopeptides.* Schroeder, C.I. and Craik, D.J. s.l. : Future Med. Chem., 2012, Vol. 4, pp. 1243–1255.

59. *Studying Smaller and Neglected Organisms in Modern Evolutionary Venomics Implementing RNASeq (Transcriptomics)—A Critical Guide.* Von Reumont, B.M. 7, s.l. : Toxins, 2018, Vol. 10, p. p.292.

60. *Characterization of the Conus bullatus genome and its venom-duct transcriptome.* Hu, H., Bandyopadhyay, P.K., Olivera, B.M. and Yandell, M. 1, s.l. : BMC genomics, Vol. 12, pp. 1-15.

61. *Guidelines for RNA-Seq data analysis.* Delhomme, N., Mähler, N., Schiffthaler, B., Sundell, D., Mannapperuma, C., Hvidsten, T.R. and Street, N.R. s.l. : Epigenesys Protoc, 2014, Vol. 67, pp. 1-24.

62. *Discovery Methodology of Novel Conotoxins from Conus Species.* Fu, Y., Li, C., Dong, S., Wu, Y., Zhangsun, D. and Luo, S. 11, s.l. : Marine drugs, 2018, Vol. 16, p. p.417.

63. *Discovery of Novel Conotoxin Candidates Using Machine Learning.* Li, Q., Watkins, M., Robinson, S.D., Safavi-Hemami, H. and Yandell, M. 12, s.l. : Toxins, 2018, Vol. 10, p. p.503.

64. Coronavirus disease (COVID-19) Situation reports. [Online] World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

65. *Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2.* Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y. and Zhou, Q. 6485, s.l. : Science, 2020, Vol. 367, pp. 1444-1448.

66. *Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2.* Song, W., Gui, M., Wang, X. and Xiang, Y. 8, s.l. : PLoS pathogens, 2018, Vol. 14.

67. *New understanding of the damage of SARS-CoV-2 infection outside the respiratory system.* Zhang, Y., Geng, X., Tan, Y., Li, Q., Xu, C., Xu, J., Hao, L., Zeng, Z., Luo, X., Liu, F. and Wang, H. s.l. : Biomedicine & pharmacotherapy, 2020, Vol. 127.

68. *Outcomes of cardiovascular magnetic resonance imaging in patients recently recovered from coronavirus disease 2019 (COVID-19).* Puntmann, V.O., Carerj, M.L., Wieters, I., Fahim, M., Arendt, C., Hoffmann, J., Shchendrygina, A., Escher, F., Vasa-Nicotera, M., Zeiher, A.M. and Vehreschild, M. 11, s.l. : JAMA cardiology, 2020, Vol. 5, pp. 1265-1273.

69. *Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis.* Hamming, I., Timens, W., Bulthuis, M.L.C., Lely, A.T., Navis, G.V. and van Goor, H. s.l. : Journal of Pathology, 2004, Vol. 203, pp. 631–637.

70. *A potential interaction between the SARS-CoV-2 spike protein and nicotinic acetylcholine receptors.* Oliveira, A.S.F., Ibarra, A.A., Bermudez, I., Casalino, L., Gaieb, Z., Shoemark, D.K., Gallagher, T., Sessions, R.B., Amaro, R.E. and Mulholland, A.J. 6, s.l. : Biophysical journal, 2021, Vol. 120, pp. 983-993.

71. *Severe acute respiratory syndrome coronavirus 2 may be an underappreciated pathogen of the central nervous system.* Alam, S.B., Willows, S., Kulka, M. and Sandhu, J.K. 11, s.l. : European journal of neurology, 2020, Vol. 27, pp. 2348-2360.

72. *Nicotine and the nicotinic cholinergic system in COVID-19.* Tizabi, Y., Getachew, B., Copeland, R.L. and Aschner, M. 17, s.l. : The FEBS journal, 2020, Vol. 287, pp. 3656-3663.

73. *Evidence of the COVID-19 virus targeting the CNS: tissue distribution, host-virus interaction, and proposed neurotropic mechanisms.* Baig, A.M., Khaleeq, A., Ali, U. and Syeda, H. 8, s.l. : ACS chemical neuroscience, 2020, Vol. 11, pp. 995-998.

74. *https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-(covid-19)-vaccines?adgroupsurvey={adgroupsurvey}&gclid=Cj0KCQjwhLKUBhDiARIsAMaTLnH3cpI2JZ-r6BL17LmVLAXjEXhQSLtuCkIa6QskgH8vbOdQfkn4mzU.*

75. *STN-125742_0_0-Section-2.5-Clinical-Overview.* s.l. : Public Health and Medical Professionals for Transparency, 2021.

76. *Pfizer-BioNTech COVID-19 Vaccine (BNT162b2) Side Effects: A Systematic Review.* Dighriri, I.M., Alhusayni, K.M., Mobarki, A.Y., Aljerary, I.S., Alqurashi, K.A., Aljuaid, F.A., Alamri, K.A., Mutwalli, A.A., Maashi, N.A., Aljohani, A.M. and Alqarni, A.M. 3, s.l. : Cureus, 2022, Vol. 14.

77. *The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak.* Rothan, H.A. and Byrareddy, S.N. s.l. : Journal of autoimmunity, 2020, Vol. 109, p. p.102433.

78. *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.* Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. and Cheng, Z. 10223, s.l. : The lancet, 2020, Vol. 395, pp. 497-506.

79. *Systematic reviewof the prevalence of current smoking among hospitalized COVID-19patients in China: could nicotine be a therapeutic option?* Farsalinos, K., Barbouni, A. and Niaura, R. 5, s.l. : Internal and emergency medicine, 2020, Vol. 15, pp. 845-852.

80. *China Medical Treatment, Clinical characteristics of coronavirus Disease 2019 in China.* Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S. and Du, B. 18, s.l. : New England journal of medicine, 2020, Vol. 382, pp. 1708-1720.

81. *Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease.* Liu, W., Tao, Z.W., Wang, L., Yuan, M.L., Liu, K., Zhou, L., Wei, S., Deng, Y., Liu, J., Liu, H.G. and Yang, M. 09, s.l. : Chinese medical journal, 2020, Vol. 133, pp. 1032-1038.

82. *Cytokine release syndrome (CRS) and nicotine in COVID-19 patients: trying to calm the strom.* Gonzalez-Rubio, J., Navarro-Lopez, C., Lopez-Najera, E., Lopez-

Najera, A., Jimenez-Diaz, L., Navarro-Lopez, J.D. and Najera, A. s.l. : Frontiers in Immunology, 2020, p. p.1359.

83. *clinicaltrials.gov.* [Online] https://clinicaltrials.gov/ct2/show/study/NCT04429815.

84. *A nicotinic hypothesis for Covid-19 with preventive and therapeutic implications.* Changeux, J.P., Amoura, Z., Rey, F.A. and Miyara, M. 1, s.l. : Comptes Rendus. Biologies, 2020, Vol. 343, pp. 33-39.

85. *Cytokine release syndrome (CRS) and nicotine in COVID-19 patients: trying to calm the storm.* Gonzalez-Rubio, J., Navarro-Lopez, C., Lopez-Najera, E., Lopez-Najera, A., Jimenez-Diaz, L., Navarro-Lopez, J.D. and Najera, A. s.l. : Frontiers in Immunology, 2020, p. p.1359.

86. *Omicron and Alpha P680H block SARS-CoV2 spike protein from accessing cholinergic inflammatory pathway via α9-nAChR mitigating the risk of MIS-C.* Camacho, Ulises Santiago, Carlos J. s.l. : bioRxiv, 2022.

87. *Does nicotine prevent cytokine storms in COVID19?* Dratcu, L. and Boland, X. 10, s.l. : Cureus, 2020, Vol. 12.

88. [Online] AnyDesk. https://anydesk.com/en.

89. [Online] KDE Applications. https://konsole.kde.org/.

90. *The sequence read archive.* Rasko Leinonen, Hideaki Sugawara, Martin Shumway, on behalf of the International Nucleotide Sequence Database Collaboration. s.l. : Nucleic Acids Res., 2011, Vol. 39, pp. D19-D21.

91. *Database resources of the national center for biotechnology information.* Sayers EW, Bolton EE, Brister JR, et al. s.l. : Nucleic Acids Res., 2022, Vol. 50, pp. D20-D26.

92. The Sequence Read Archive (SRA). *NCBI.* [Online] SRA Toolkit Development Team. https://www.ncbi.nlm.nih.gov/sra/docs/.

93. SRA Data Formats. *NCBI.* [Online] https://www.ncbi.nlm.nih.gov/sra/docs/sra-data-formats/.

94. *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.* Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 6, s.l. : Nucleic Acids Res, 2010, Vol. 38, pp. 1767-1771.

95. parallel-fastq-dump. [Online] https://github.com/rvalieris/parallel-fastq-dump.

96. FastQC. *Babraham Bioinformatics.* [Online] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

97. *An extensive evaluation of read trimming effects on Illumina NGS data analysis.* s.l. : PLoS ONE, 2013, Vol. 8, p. e85024.

98. *Software for pre-processing Illumina next-generation sequencing short read sequences.* 8, s.l. : Source Code Biol. Med., 2014, Vol. 9.

99. *AdapterRemoval: Easy cleaning of next-generation sequencing reads.* s.l. : BMC Res. Notes, 2012, Vol. 5, p. 337.

100. *Trimmomatic: A flexible trimmer for Illumina sequence data.* s.l. : Bioinformatics, 2014, Vol. 30, pp. 2114–2120.

101. Trimmomatic: A flexible read trimming tool for Illumina NGS data. *USADELLAB.org.* [Online] http://www.usadellab.org/cms/?page=trimmomatic.

102. *Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads.* Martin, Marcel. 1, s.l. : EMBnet j., Vol. 17, pp. 10-12. 10.

103. *Full-length transcriptome assembly from RNA-seq data without a reference genome.* Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. and Chen, Z. 7, s.l. : Nature biotechnology, 2011, Vol. 29, pp. 644-652.

104. *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Haas, B., Papanicolaou, A., Yassour, M. s.l. : Nat Protoc, 2013, Vol. 8, pp. 1494–1512.

105. *Evaluating the Performance of De Novo Assembly Methods for Venom-Gland Transcriptomics.* Holding, M.L., Margres, M.J., Mason, A.J., Parkinson, C.L. and Rokyta, D.R. 6, s.l. : Toxins, 2018, Vol. 10, p. 249.

106. *BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.* 10, s.l. : Molecular Biology and Evolution, 2021, Vol. 38, pp. 4647–4654.

107. TransDecoder. *github.* [Online] https://github.com/TransDecoder/TransDecoder/releases/tag/TransDecoder-v5.5.0.

108. BLAST. *Blastp.* [Online] https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins.

109. *UniProt: the universal protein knowledgebase in 2021.* The UniProt Consortium. D1, s.l. : Nucleic Acids Res., 2021, Vol. 49, pp. D480–D489.

110. *Basic local alignment search tool.* Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. s.l. : J. Mol. Biol., 1990, Vol. 215, pp. 403-410.

111. *Magic-BLAST, an accurate RNA-seq aligner for long and short reads.* Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. 1, s.l. : BMC Bioinformatics, 2019, Vol. 20, p. 405.

112. UniProtKB/Swiss-Prot. *Expasy.* [Online] https://www.expasy.org/resources/uniprotkb-swiss-prot.

113. HMMER. [Online] http://hmmer.org/.

114. *HMMER.* [Online] http://hmmer.org/download.html.

115. *Pfam: The protein families database in 2021.* Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ, Finn RD. D1, s.l. : Nucleic Acids Res., 2021, Vol. 49, pp. D412–D419.

116. Xfam Blog. [Online] https://xfam.wordpress.com/2021/11/19/pfam-35-0-is-released/.

117. *eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.* Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C. 47, s.l. : Nucleic Acids Res., 2019, pp. D309-D314.

118. *eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.* Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 12, s.l. : Molecular Biology and Evolution, 2021, Vol. 38, pp. 5825–5829.

119. EggNOG 5.0.0. *Downloading EggNOG raw data.* [Online] http://eggnog5.embl.de/#/app/downloads.

120. *Gene ontology: tool for the unification of biology.* Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. 1, s.l. : Nat Genet, 2000, Vol. 25, pp. 25-29.

121. *The Gene Ontology resource: enriching a GOld mine.* The Gene Ontology Consortium . D1, s.l. : Nucleic Acids Res, 2021, Vol. 49, pp. D325-D334.

122. *geneontology.org.* [Online] http://purl.obolibrary.org/obo/go/go-basic.obo.

123. *UniProt databases.* [Online] https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz.

124. *UniProt databases.* [Online] https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz.

125. *Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase.* Jungo F, Bairoch A. 3, s.l. : Toxicon, 2005, Vol. 45, pp. 293-301.

126. *ConoServer: updated content, knowledge, and discovery tools in the conopeptide database.* Kaas Q, Yu R, Jin AH, Dutertre S, Craik DJ. D1, s.l. : Nucleic Acids Res., 2012, Vol. 40, pp. D325-30.

127. *UpSetR: an R package for the visualization of intersecting sets and their properties.* Jake R Conway, Alexander Lex, Nils Gehlenborg. 18, s.l. : Bioinformatics, 2017, Vol. 33, pp. 2938–2940.

128. *UpSet: Visualization of Intersecting Sets.* A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot and H. Pfister. 12, s.l. : IEEE Transactions on Visualization and Computer Graphics, 2014, Vol. 20, pp. 1983-1992.

129. *ggplot2: Elegant Graphics for Data Analysis.* Villanueva RA, Randle Aaron M., Chen ZJ. s.l. : Journal of Statistical Software, 2019, pp. 160-167.

130. *MEGA11: Molecular Evolutionary Genetics Analysis version 11.* Koichiro Tamura, Glen Stecher, and Sudhir Kumar. s.l. : Molecular Biology and Evolution, 2021, Vol. 38, pp. 3022-3027.

131. Molecular evolution, phylogenetics and epidemiology. *FigTree.* [Online] http://tree.bio.ed.ac.uk/software/figtree/.

132. *Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod Conus.* Thomas F. Duda, Jr., Stephen R. Palumbi. 12, s.l. : Biological Sciences, 1999, Vol. 96, pp. 6820-6823.

133. *Evolution of Conus Peptide Genes: Duplication and Positive Selection in the A-Superfamily.* Puillandre, N., Watkins, M., Olivera, B.M. s.l. : J Mol Evol, 2010, Vol. 70, pp. 190–202.

134. *Extensive and Continuous Duplication Facilitates Rapid Evolution and Diversification of Gene Families.* Dan Chang, Thomas F. Duda, Jr. 8, s.l. : Molecular Biology and Evolution, 2012, Vol. 29, pp. 2019–2029.

135. *Systematic interrogation of the Conus marmoreus venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene superfamilies.* Vincent Lavergne, Sébastien Dutertre, Ai-hua Jin, Richard J Lewis, Ryan J Taft, Paul F Alewood. 708, s.l. : BMC Genomics, 2013, Vol. 14.

136. *Novel ω-Conotoxins from Conus catus Discriminate among Neuronal Calcium Channel Subtypes.* Lewis RJ, Nielsen KJ, Craik DJ, Loughnan ML, Adams DA, Sharpe IA, Luchian T, Adams DJ, Bond T, Thomas L, Jones A, Matheson JL, Drinkwater R, Andrews PR, Alewood PF. 275, s.l. : J Biol Chem, 2000, Vol. 10, pp. 35335-44.

137. *Diversity and evolution of conotoxins based on gene expression profiling of Conus litteratus.* Pi C, Liu J, Peng C, Liu Y, Jiang X, Zhao Y, Tang S, Wang L, Dong M, Chen S, Xu A. 6, s.l. : Genomics, 2006, Vol. 88, pp. 809-819.

138. *Microhabitats within venomous cone snails contain diverse actinobacteria.* Peraud O, Biggs JS, Hughen RW, Light AR, Concepcion GP, Olivera BM, Schmidt EW. 21, s.l. : Appl Environ Microbiol, 2009, Vol. 75, pp. 6820-6826.

139. *Science Friday.* [Online] SciFri. https://www.sciencefriday.com/educational-resources/how-do-killer-snails-kill-their-victims/.

7. Annexes

7.1. System specifications

- Linux-4.15.0-189-generic-x86_64-with-debian-buster-sid (linux_64);
- Python version: 3.7.3;
- CPUs number: 56.

7.2. Command functions' scripts and software calls

7.2.1. Prefetch

```
for i in $(cat transcriptome_SRR_list.txt)

do

        prefetch "$i"

done
```

7.2.2. Parallel fastq-dump

```
for i in $(cat ../transcriptome_SRR_list.txt)

do

        parallel-fastq-dump --defline-seq '@$sn[_$rn]/$ri' --split-files --
        gzip -O fastq_from_sra --sra-id "$i" --threads 40

done
```

7.2.3. 1st FastQC

```
for i in $(cat ../../transcriptome_SRR_list.txt)

do

         ../../try/FastQC/fastqc --nogroup "$i".sra_1.fastq.gz -o 1st_qc

        ../../try/FastQC/fastqc --nogroup "$i".sra_2.fastq.gz -o 1st_qc

done
```

7.2.4. 2nd FastQC

```
for i in $(cat ../../../transcriptome_SRR_list.txt)

do

        ../../../try/FastQC/fastqc --nogroup "$i"_1_paired..fastq.gz -o
2nd_qc
```

../../../try/FastQC/fastqc --nogroup "$i"_2_paired..fastq.gz -o 2nd_qc

done

7.2.5. 3rd FastQC

for i in $(cat ../../../../transcriptome_SRR_list.txt)

do

../../../../try/FastQC/fastqc --nogroup "$i""_1_paired..fastq.gz -o 3rd_qc

../../../../try/FastQC/fastqc --nogroup "$i""_2_paired..fastq.gz -o 3rd_qc

done

7.2.6. Trimmomatic

for i in $(cat ../../../transcriptome_SRR_list.txt)

do

trimmomatic PE ../"$i"_1.fastq.gz ../"$i"_2.fastq.gz "$i"_1_paired.fastq.gz "$i"_1_unpaired.fq.gz "$i"_2_paired.fastq.gz "$i"_2_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:1:keepBothReads LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 -threads 48 MINLEN:51

done

7.2.7. Cutadapt

for i in $(cat ../list.txt)

do

cutadapt -j 30 -u 15 -U 15 -o "$i"_15bpcut.1.fastq.gz -p "$i"_15bpcut.2.fastq.gz ../"$i"_1_paired.fastq.gz ../"$i"_2_paired.fastq.gz

done

7.2.8. Trinity

a. cat trinity_script_1.sh

for i in $(cat first50list.txt)

do

../../../../../try/trinityrnaseq-v2.14.0/Trinity --seqType fq --left ../"$i"_15bpcut.1.fastq.gz --right "$i"_15bpcut.2.fastq.gz --output "$i"_trinity --25 --max_memory 25G

done

b. cat trinity_script_1.sh

for i in $(cat 49list.txt)

do

../../../../../try/trinityrnaseq-v2.14.0/Trinity --seqType fq --left ../"$i"_15bpcut.1.fastq.gz --right "$i"_15bpcut.2.fastq.gz --output "$i"_trinity --25 --max_memory 25G

done

7.2.9. BUSCO

a. for i in $(cat assembly_fasta_list.txt)
do
/home/labpc10c/Documents/Toby/tools/busco/bin/busco -c 40 -o BUSCO_result/"$i"_busco.out -i "$i"_trinity.Trinity.fasta -m transcriptome -l ~/Documents/BUSCO/busco_downloads/mollusca_odb10
done

b. for i in $(cat assembly_fasta_list.txt)
do
/home/labpc10c/Documents/Toby/tools/busco/bin/busco -c 40 -o BUSCO_result/"$i"_busco.out -i "$i"_trinity.Trinity.fasta -m transcriptome -l ~/Documents/BUSCO/busco_downloads/metazoa_odb9
done

c. for i in $(cat /home/labpc10c/Documents/Toby/Toby_2022/ncbi_transcriptome_conus/venom_list.txt)
do

```
    cp
    "$i"_busco.out/short_summary.specific.mollusca_odb10."$i"_bus
    co.out.txt venom_summary/
done
```

d.
```
for i in $(cat
/home/labpc10c/Documents/Toby/Toby_2022/ncbi_transcriptome_co
nus/other_tissue_list.txt)
do
    cp
    "$i"_busco.out/short_summary.specific.mollusca_odb10."$i"_bus
    co.out.txt other_tissue_summary/
done
```

e.
```
for i in $(cat
/home/labpc10c/Documents/Toby/Toby_2022/ncbi_transcriptome_co
nus/venom_list.txt)
do
    cp
    "$i"_busco.out/short_summary.specific.metazoa_odb10."$i"_bus
    co.out.txt venom_summary/
done
```

f.
```
for i in $(cat
/home/labpc10c/Documents/Toby/Toby_2022/ncbi_transcriptome_co
nus/other_tissue_list.txt)
do
    cp
    "$i"_busco.out/short_summary.specific.metazoa_odb10."$i"_bus
    co.out.txt other_tissue_summary/
done
```

7.2.10. TransDecoder.LongOrfs

```
for i in $(cat ../assembly_fasta_list.txt)

do

    TransDecoder.LongOrfs -t ../"$i"_trinity.Trinity.fasta

done
```

7.2.11. BlastP+Uniprot and Hmmscan+Pfam

a. for i in $(cat ../assembly_fasta_list.txt)
   do
   blastp -query "$i"_trinity.Trinity.fasta.transdecoder_dir/longest_orfs.pep -db ~/Documents/Annotation_db/UniProt/UniProt -num_threads 48 -max_target_seqs 1 -outfmt 6 -evalue 1e-5 > "$i"_UniProt_blastp.outfmt6
   done

b. for i in $(cat ../assembly_fasta_list.txt)
   do
   hmmscan --cpu 4 --domtblout "$i"_pfam.dom --tblout "$i"_pfam.tbl -o "$i"_pfam_out.txt -E 1e-5 ~/Documents/Annotation_db/Pfam/Pfam-A.hmm "$i"_trinity.Trinity.fasta.transdecoder_dir/longest_orfs.pep
   done

c. for i in $(cat pfam_list_1.txt)

   do

   hmmscan --cpu 2 --domtblout "$i"_pfam.dom --tblout "$i"_pfam.tbl -o "$i"_pfam_out.txt -E 1e-5 ~/Documents/Annotation_db/Pfam/Pfam-A.hmm "$i"_trinity.Trinity.fasta.transdecoder_dir/longest_orfs.pep

   done

7.2.12. TransDecoder.Predict

for i in $(cat ../assembly_fasta_list.txt)

do

TransDecoder.Predict -t ../"$i"_trinity.Trinity.fasta --retain_blastp_hits "$i"_UniProt_blastp.outfmt6 --retain_pfam_hits "$i"_pfam_out.txt

done

7.2.13. TransDecoder.Predict vs Pfam

a. for i in $(cat ~/Documents/Toby/Toby_2022/ncbi_transcriptome_conus/list_82.txt)

```
do
    echo"$i"
    grep > "$i"_trinity.Trinity.fasta.transdecoder.pep |awk 'BEGIN
    {FS="."}; {print $1}' | sort | uniq | wc -l
done
```

b.
```
for i in $(cat ~/Documents/Toby/Toby_2022/ncbi_transcriptome_conus/list_82.txt)
do
    echo"$i"
    awk '{print $4}' "$i"_pfam.dom | grep 'TRINITY_DN' |awk 'BEGIN
    {FS="."}; {print $1}' | sort | uniq | wc -l
done
```

7.2.14. EggNOG

a. download_eggnog_data.py -H -d 2759
b. download_eggnog_data.py -H -d 33208
c.
```
for i in $(cat ~/Documents/Toby/Toby_2022/ncbi_transcriptome_conus/list_82.txt)
do
    emapper.py -d euk -i TransDecoder_Predict/"$i"_trinity.Trinity.fasta.transdecoder.cds --itype CDS -o eggNOG_results/"$i"_eggnog_cds --cpu 48 --usemem --no_file_comments --override
done
```

7.2.15. GO

```
for i in $(cat ../../assembly_fasta_list.txt)
do
    python2 /home/labpc10c/Documents/Annotation_db/GO/Uniprot2GO_annotated.py /home/labpc10c/Documents/Annotation_db/GO/idmapping_selected.tab.gz ../Annotation/"$i"_UniProt_blastp.outfmt6 "$i"_UniProt2Go_for_stats.out
done
```

7.2.16. Makeblastdb

a. makeblastdb -in ToxProt.fasta -dbtype prot -out ToxProt -parse_seqids
b. makeblastdb -in conoserver_protein.fa -dbtype prot -out Cono_pep

7.2.17. BlastP + Locally built databases

a. for i in $(cat ../assembly_fasta_list.txt)

   do

   blastp -query "$i"_trinity.Trinity.fasta.transdecoder_dir/longest_orfs.pep -db ~/Documents/Annotation_db/Conoserver_pep/conoserver_protein – num_threads 1 -max_target_seqs 1 -outfat 6 -evalue 1e-5 > "$i"_ConoPep_blastp.outfat6

   done

b. for i in $(cat ../assembly_fasta_list.txt)

   do

   blastp -query "$i"_trinity.Trinity.fasta.transdecoder_dir/longest_orfs.pep -db ~/Documents/Annotation_db/ToxProt/Toxprot -num_threads 2 -max_target_seqs 1 -outfat 6 -evalue 1e-5 > "$i"_Toxprot_blastp.outfat6

   done

7.2.18. TransDecoder runs on Spike protein and SARS-Cov-2 genome

a. TransDecoder.LongOrfs -t <Covid_S.fasta>
b. TransDecoder.LongOrfs -t <Covid_full.fasta>

7.2.19. BlastP and Hmmscan of the Spike Protein vs ToxProt and Pfam, respectively + Full covid genome vs Pfam

a. blastp -query COVID_S.fasta.transdecoder_dir/longest_orfs.pep -db ~/Documents/Annotation_db/ToxProt/Toxprot -num_threads 20 -max_target_seqs 1 -outfat 6 -evalue 1e-5 > Spike_Toxprot_blastp.outfat6

b. Hmmscan --domtblout spike_fam.dom --tblout spike_fam.tbl -o spike_fam_out.txt -E 1e-5 ~/Documents/Annotation_db/Pfam/Pfam-A.hmm COVID_S.fasta.transdecoder_dir/longest_orfs.pep
c. Hmmscan --domtblout full_fam.dom --tblout full_fam.tbl -o full_fam_out.txt ~/Documents/Annotation_db/Pfam/Pfam-A.hmm COVID_full.fasta.transdecoder_dir/longest_orfs.pep

7.2.20. UpSetR

```
library(UpSetR)

A <- read.csv("69_GO_list_uniq.txt",sep = "\t", header = T)

x <- fromList(A)

upset(x, order.by="degree", decreasing = FALSE, nsets=69, nintersects = 100, keep.order = TRUE, main.bar.color = 'black', matrix.color="#4285F4", mainbar.y.label = "GO id Intersections", sets.x.label = "GO per assembly", text.scale = c(1.5, 1.2, 1.2, 1, 1, 1), point.size = 2,line.size = 0.75, number.angles = 0, mb.ratio = c(0.4,0.6))
```

7.2.21. ggplot2

```
library(ggplot2)

library(reshape2)

data <-read.csv('Table3.csv', stringsAsFactors = TRUE,header=TRUE,sep='\t')

df<-
melt(data,id.vars="GO",variable.name="Feeding_habit",value.name="count")

ggplot(df, aes(x = factor(GO,levels = unique(GO)),y = count,fill = Feeding_habit)) +

        geom_bar(stat = "identity") +

        coord_flip() +

        labs(x = "GO names", y = "GO numbers", fill="Feeding habits")
```

7.3. MultiQC reports' heatmaps

7.3.1. First multiQC report – First FastQC status check heatmap

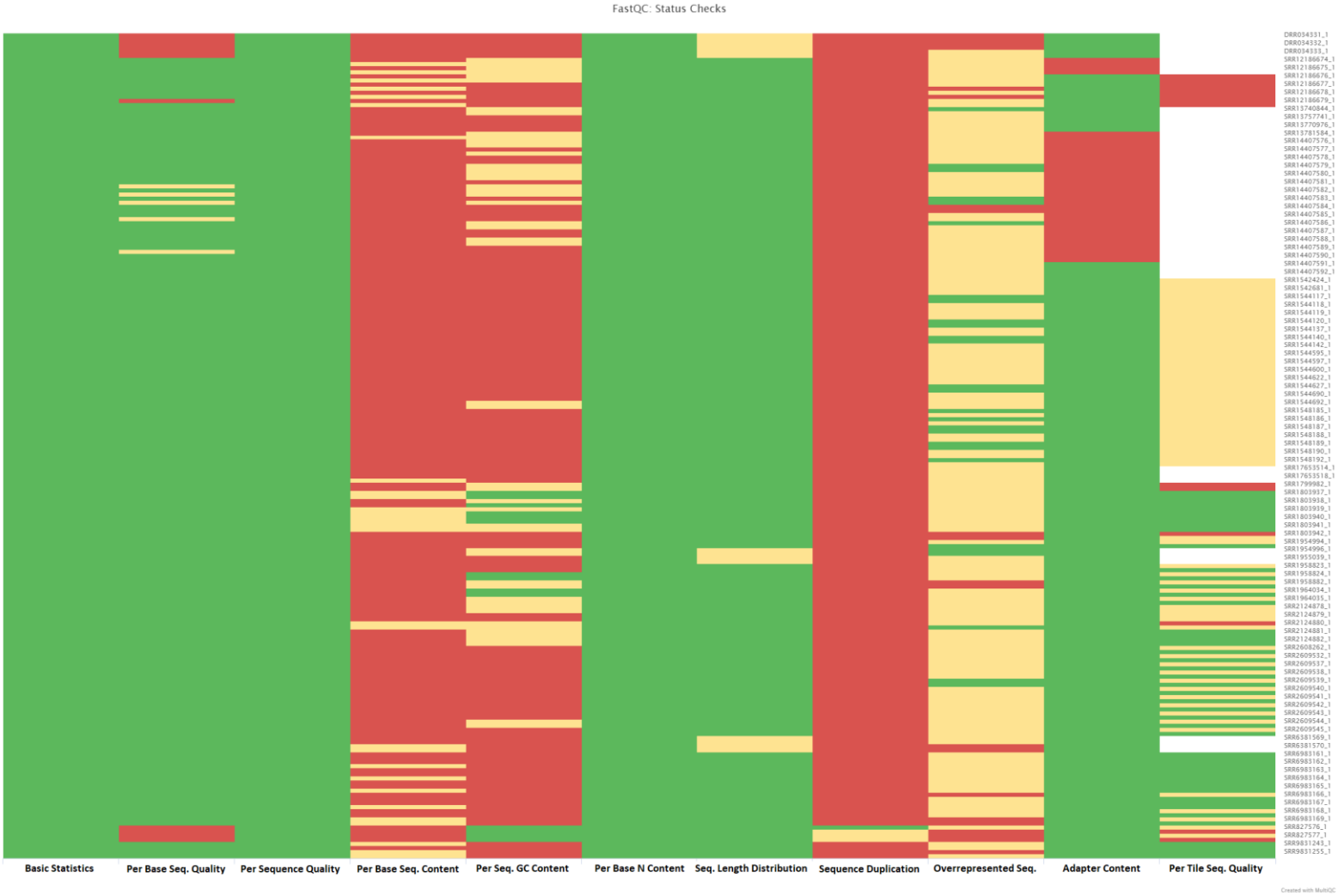


Fig. 18 – FastQC status check heatmap of the first MultiQC report; this heatmap illustrates the state of the transcriptomic data right after being acquired from the NCBI platform. Along the vertical axis in the right are the files with the transcriptomic data. Along the horizontal axis in the bottom are the categories evaluated, from left to right: Basic Statistics, Per Base Sequence Quality, Per Sequence Quality, Per Base Sequence content, Per Sequence GC Content, Per Base N Content, Sequence Length Distribution, Sequence Duplication, Overrepresented Sequences, Adapter Content and Per Tile Sequence Quality. The red colour represents bad quality, yellow is medium quality and the green represents good quality.

7.3.2. Second multiQC Report – Second FastQC status check heatmap

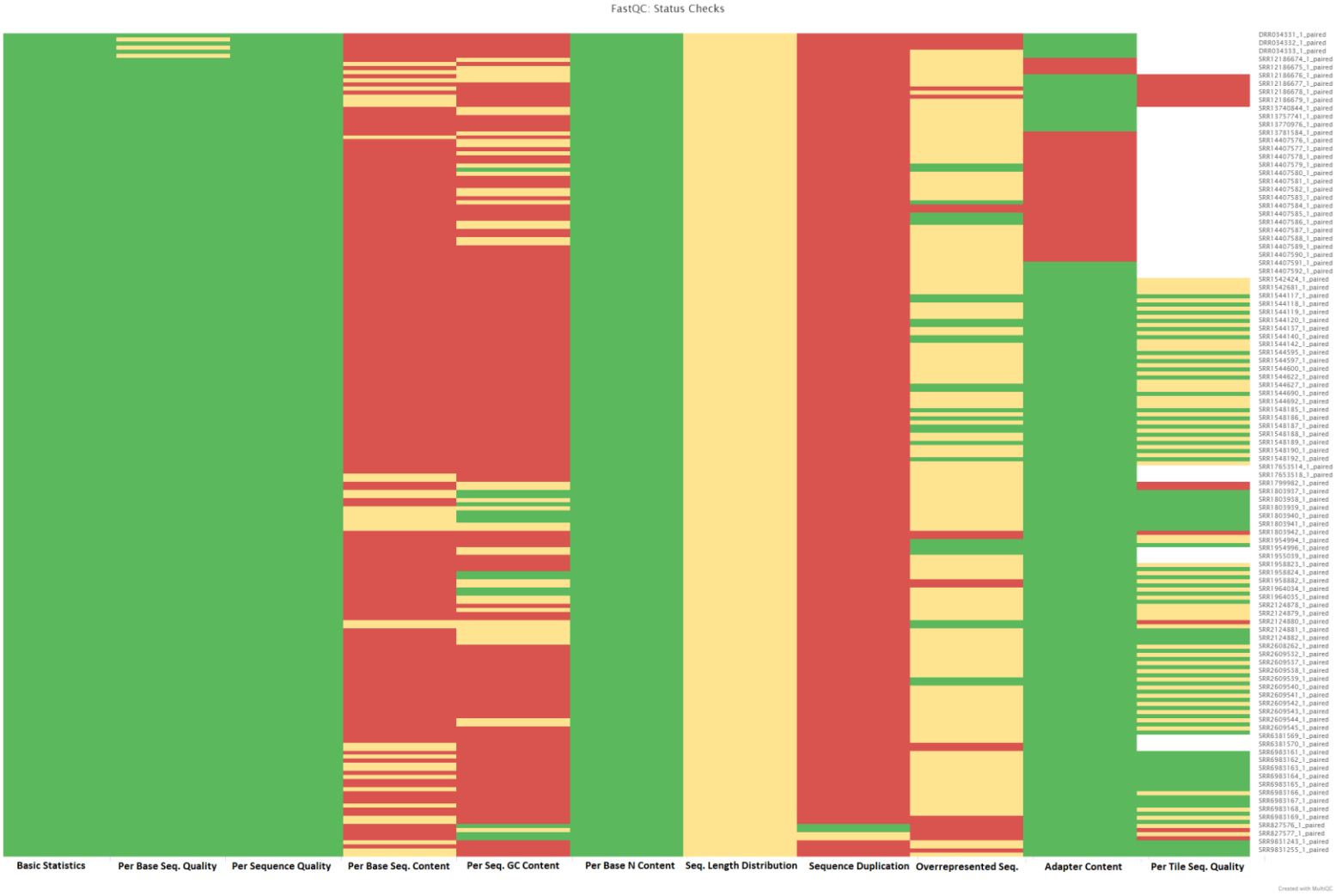


Fig. 19 – FastQC status check heatmap of the second MultiQC report; this heatmap illustrates the state of the transcriptomic data after being processed by the software Trimmomatic. Along the vertical axis in the right are the files with the transcriptomic data. Along the horizontal axis in the bottom are the categories evaluated, from left to right: Basic Statistics, Per Base Sequence Quality, Per Sequence Quality, Per Base Sequence content, Per Sequence GC Content, Per Base N Content, Sequence Length Distribution, Sequence Duplication, Overrepresented Sequences, Adapter Content and Per Tile Sequence Quality. The red colour represents bad quality, yellow is medium quality and the green represents good quality.

7.3.3. Third multiQC report – Third FastQC status check heatmap

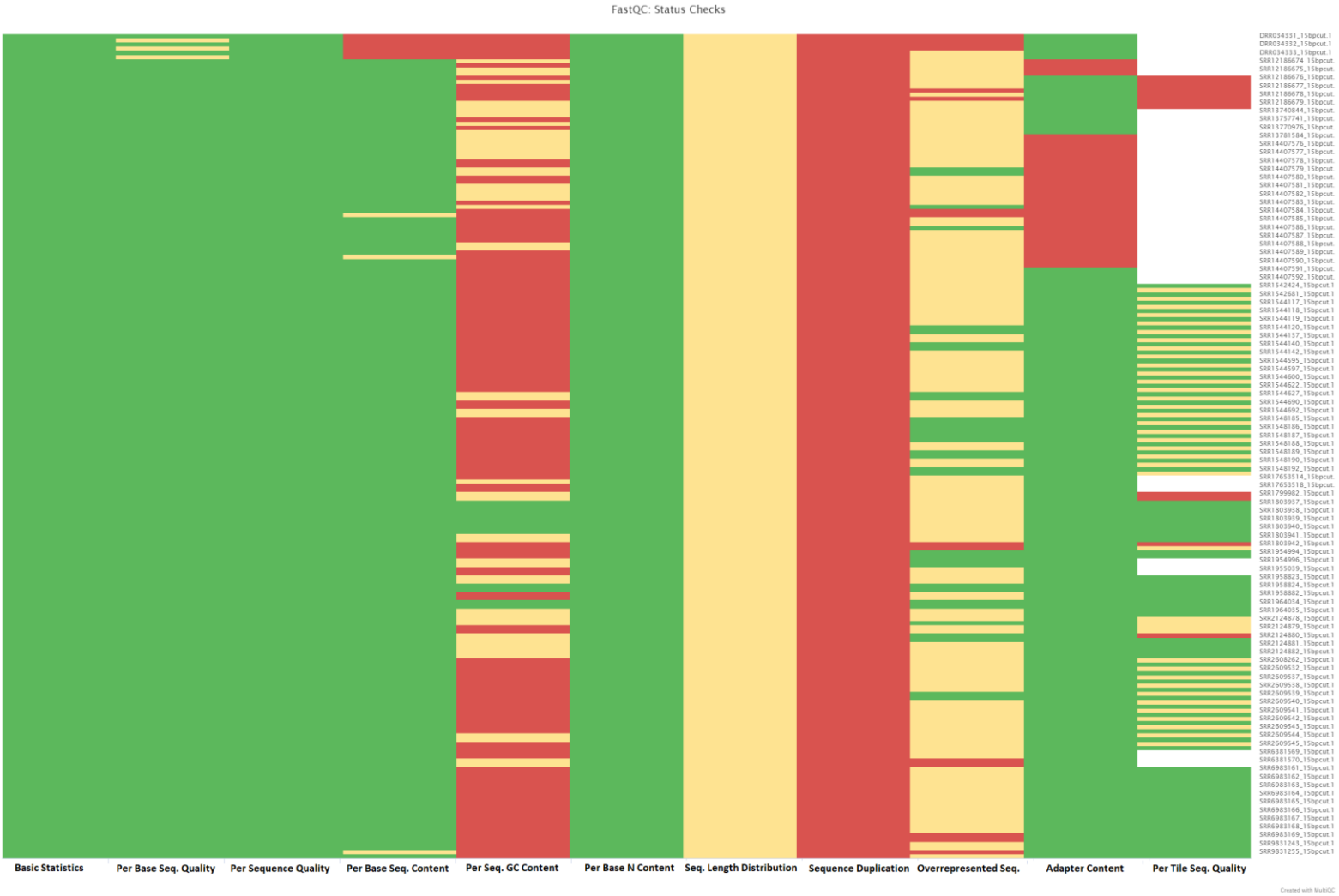


Fig. 20 – FastQC status check heatmap of the third MultiQC report; this heatmap illustrates the state of the transcriptomic data after being processed by the software Cutadapt. Along the vertical axis in the right are the files with the transcriptomic data. Along the horizontal axis in the bottom are the categories evaluated, from left to right: Basic Statistics, Per Base Sequence Quality, Per Sequence Quality, Per Base Sequence content, Per Sequence GC Content, Per Base N Content, Sequence Length Distribution, Sequence Duplication, Overrepresented Sequences, Adapter Content and Per Tile Sequence Quality. The red colour represents bad quality, yellow is medium quality and the green represents good quality.

7.4. BUSCO assessment charts

7.4.1. BUSCO assessments for the whole dataset

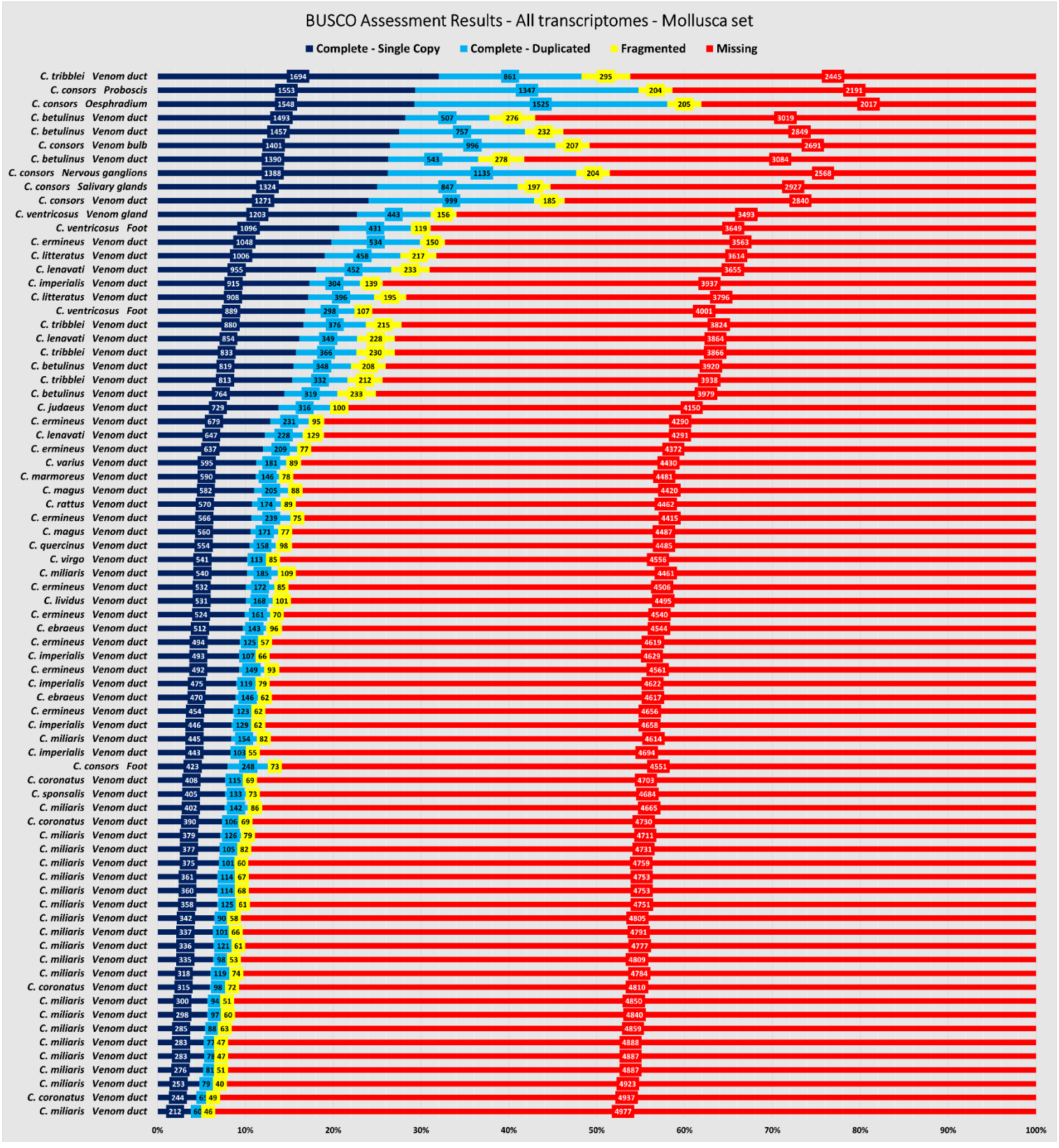


Fig. 21 – BUSCO assessment performed against BUSCO's Mollusca set for all 76 transcriptome assemblies.

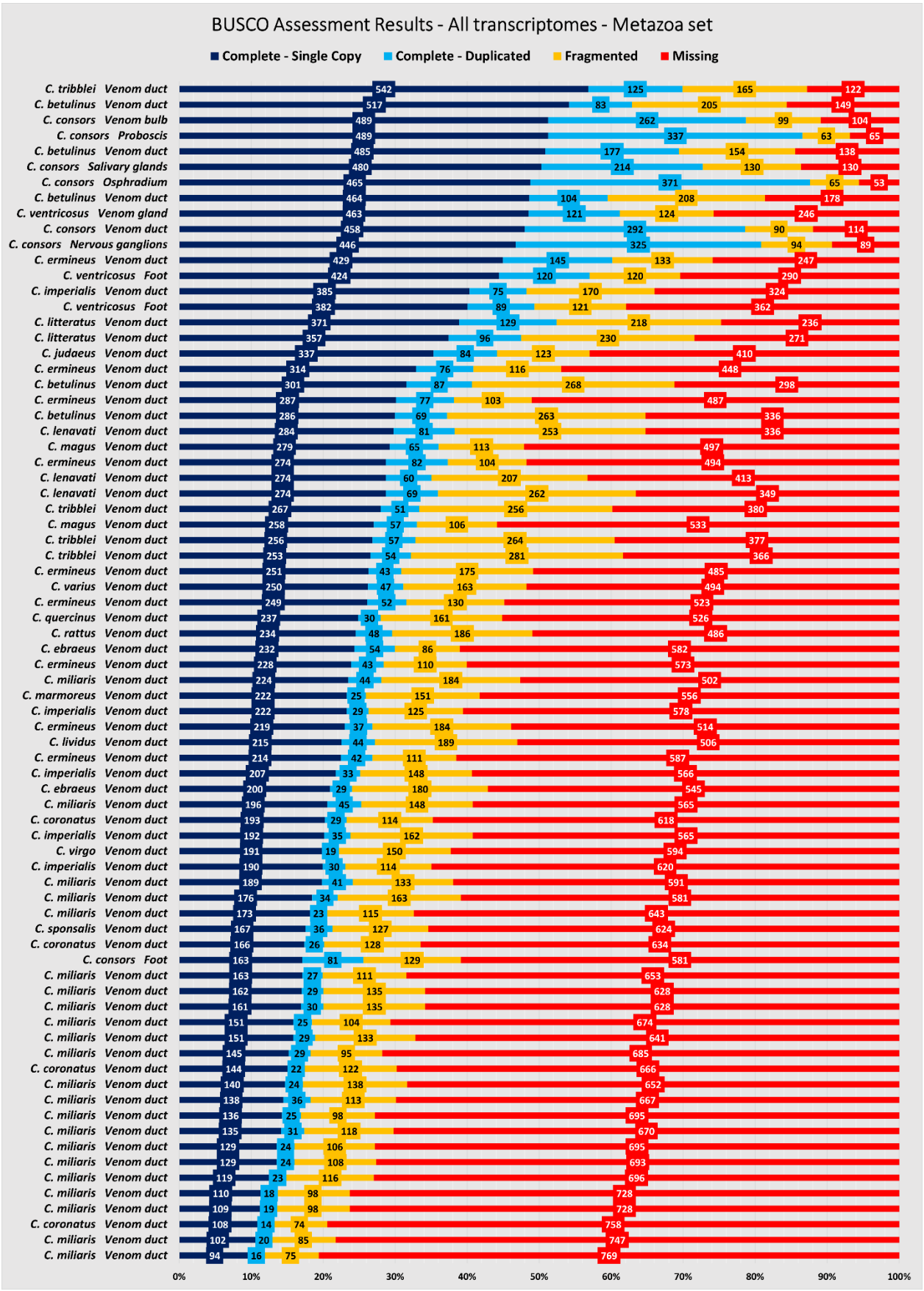


Fig. 22 – BUSCO assessment performed against BUSCO's Metazoa set for all 76 transcriptome assemblies.

7.4.2. BUSCO assessments on the venom-related transcriptomes

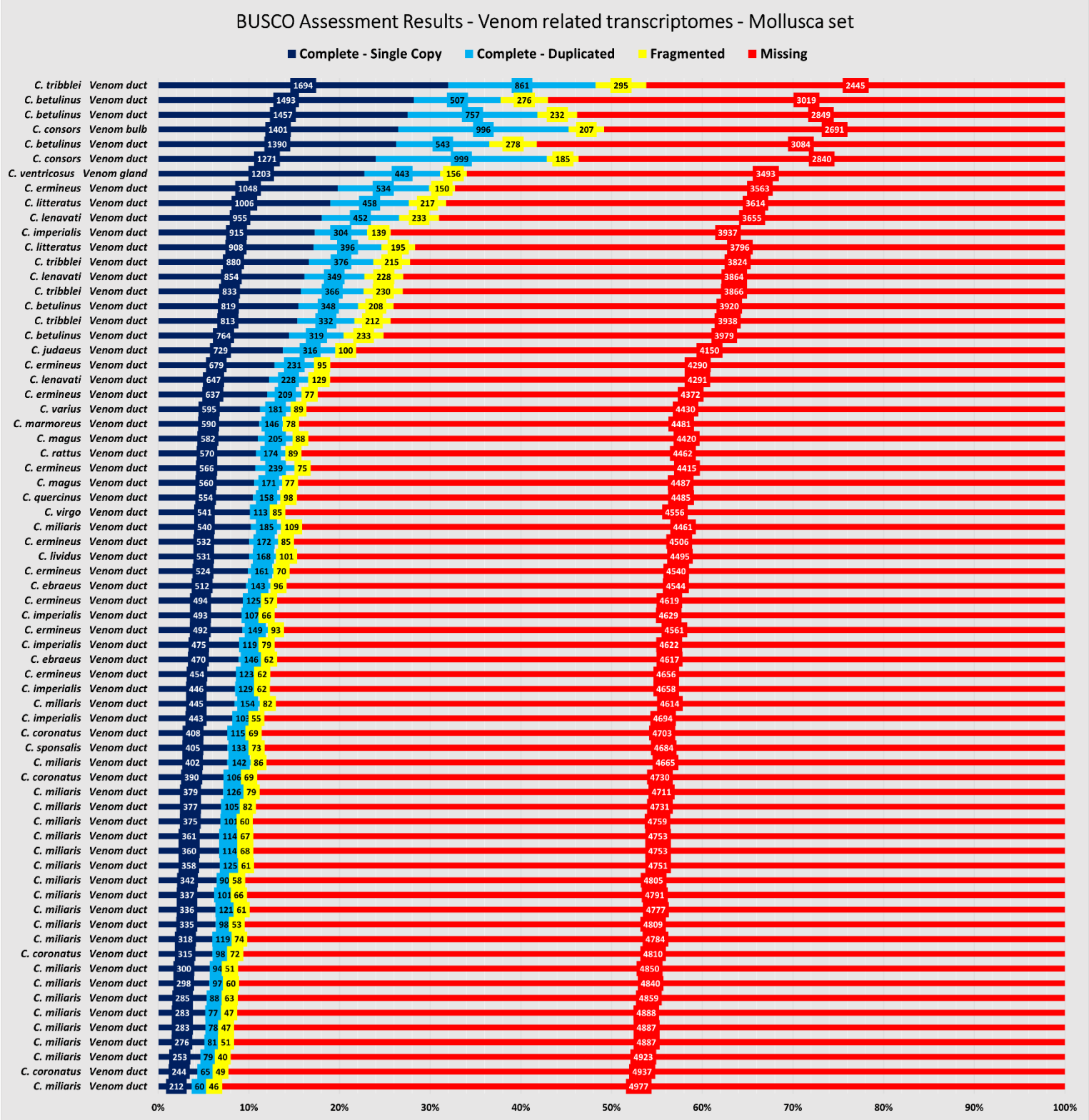


Fig. 23 – BUSCO assessment performed against BUSCO's Mollusca set for the 69 transcriptome assemblies from venom related tissues.

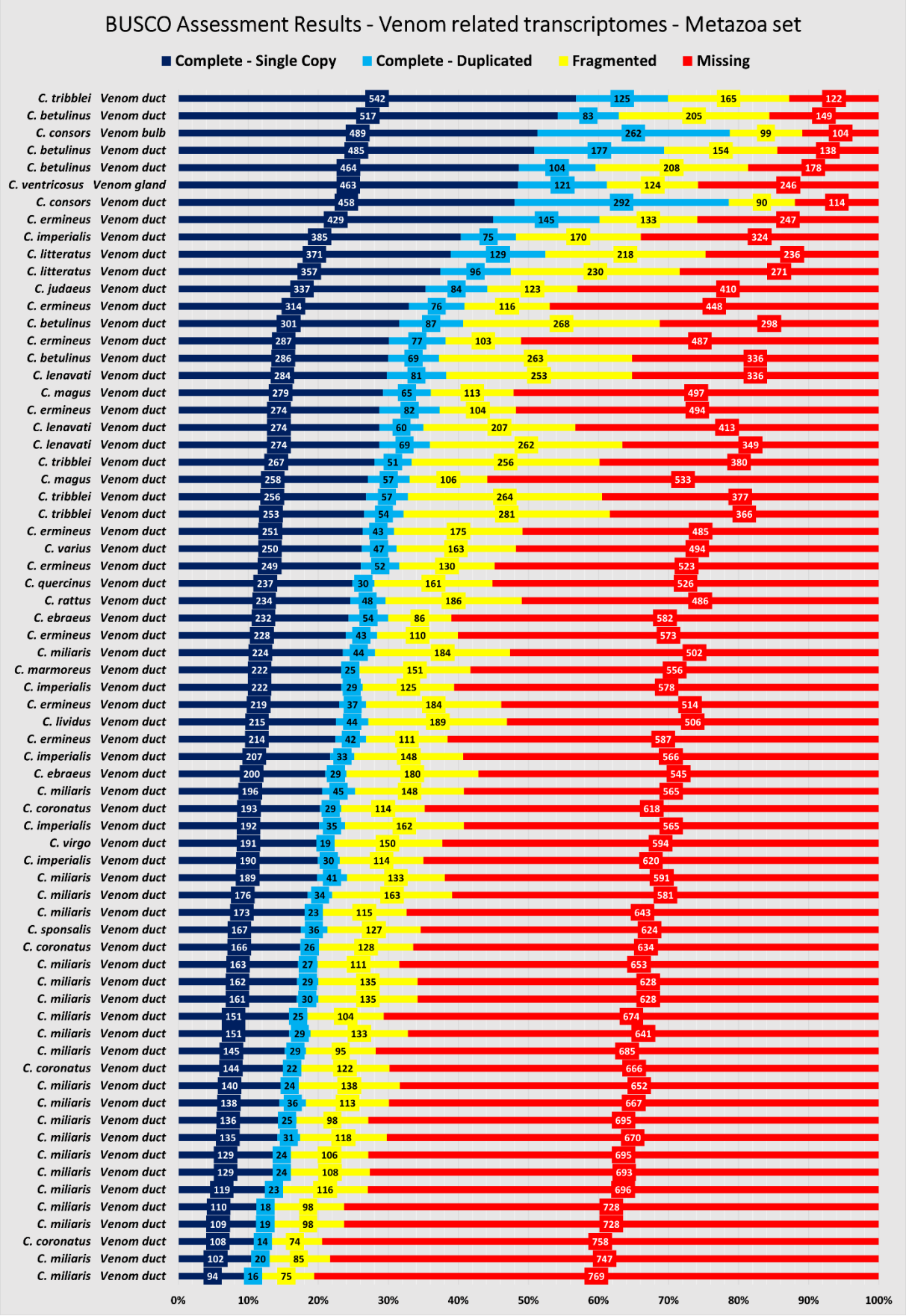


Fig. 24 – BUSCO assessment performed against BUSCO's Metazoa set for the 69 transcriptome assemblies from venom related tissues.

7.4.3. BUSCO assessments on transcriptomes from various tissues

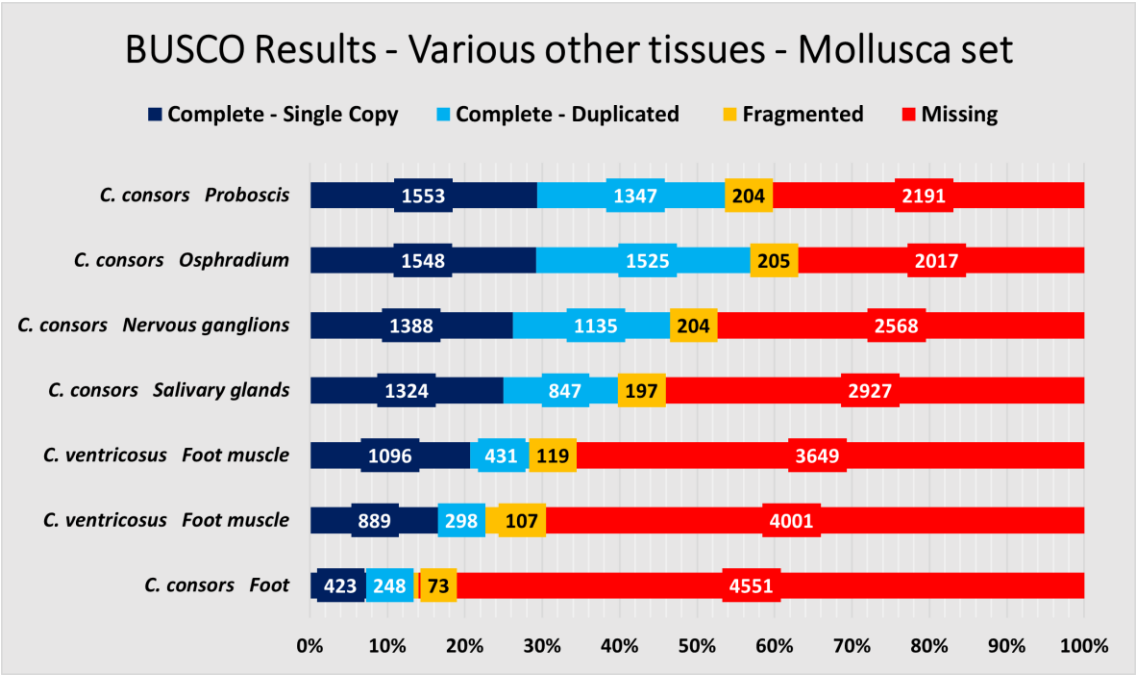


Fig. 25 – BUSCO assessment performed against BUSCO's Mollusca set for the 7 transcriptome assemblies from various tissues.

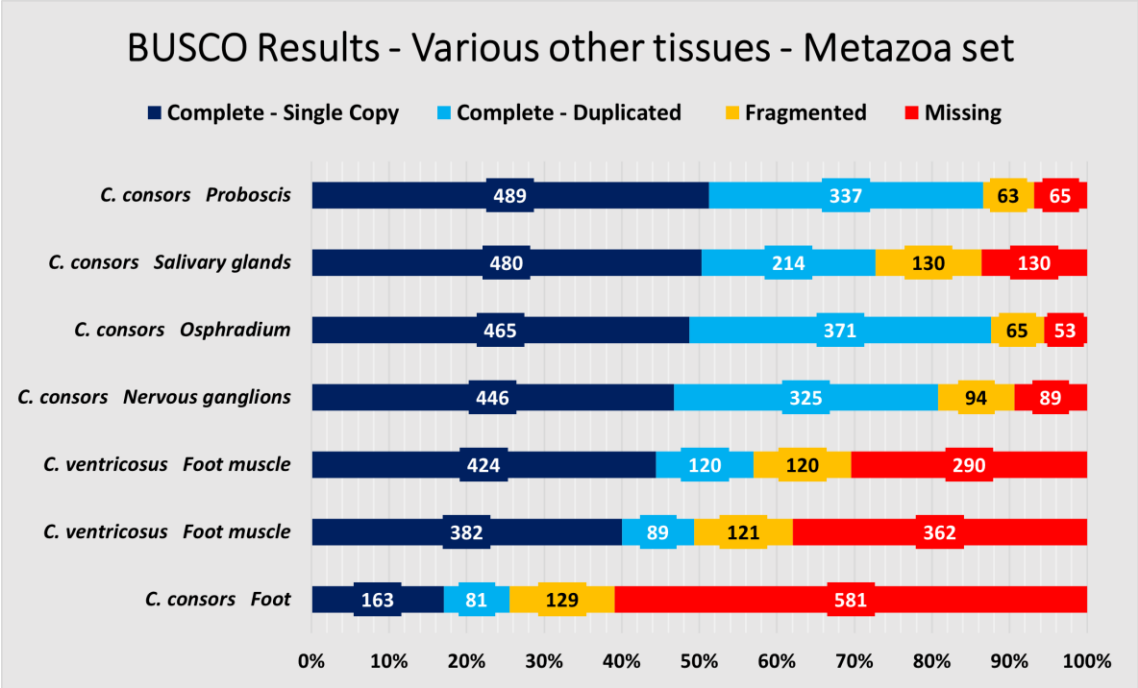


Fig. 26 – BUSCO assessment performed against BUSCO's Metazoa set for the 7 transcriptome assemblies from various tissues.

7.5. Annotation comparison charts

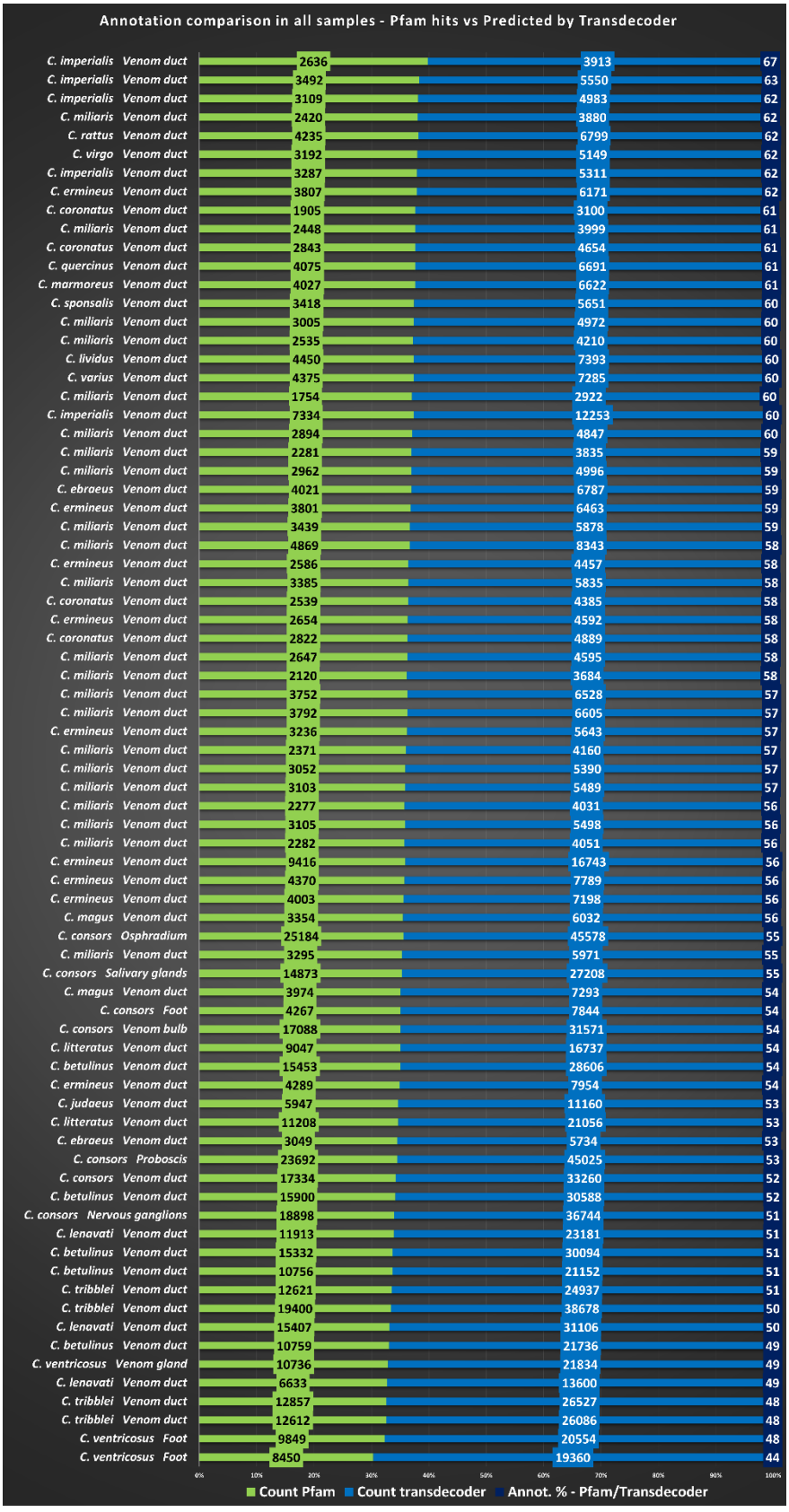


Fig. 27 – Chart illustrating the predicted coding sequences in blue, and the sequences recognized by the database in green for all the transcriptomes in decrescent order of ratio.

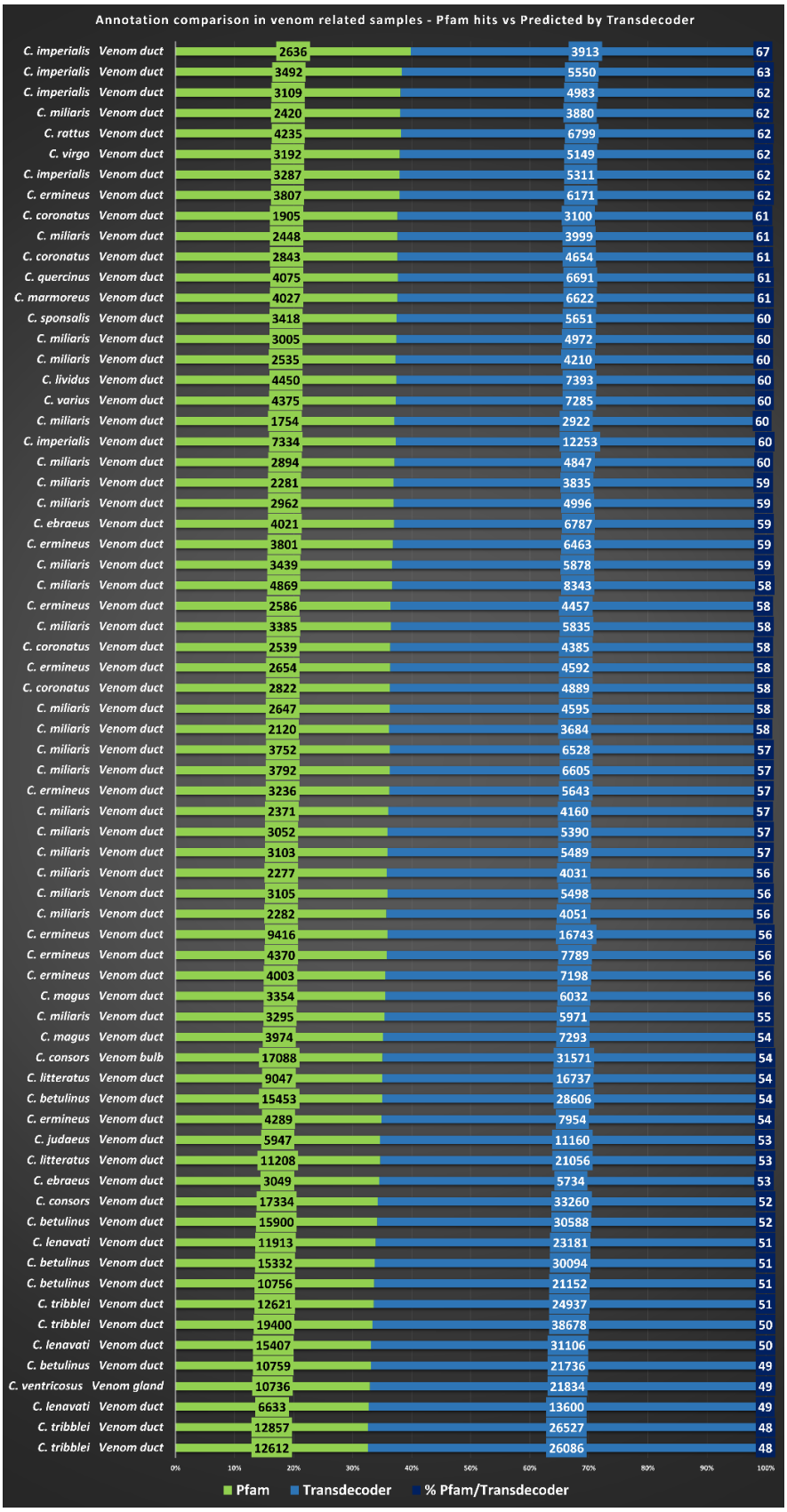


Fig. 28 – Chart showing the predicted and recognized coding sequences in blue and green respectively, for the venom-related transcriptomes in decrescent order of ratio.

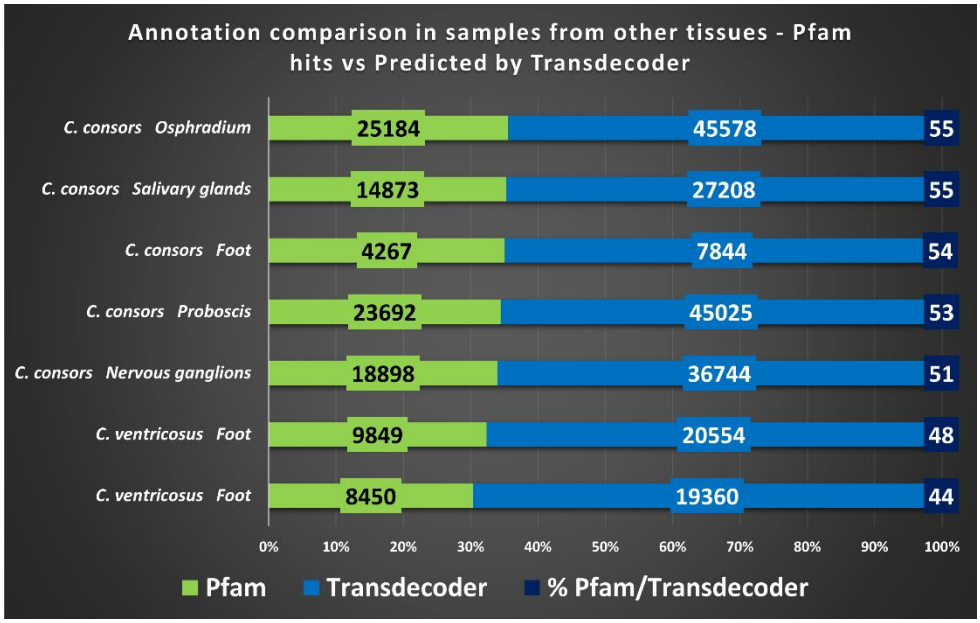


Fig. 29 – Chart showing the predicted and recognized coding sequences in blue and green respectively, for the samples of other body parts in decrescent order of ratio.

7.6. Table with assembly size correlation with gene number

	Venom data (sorted in crecent order of assembly size)									
	Species	Tissue	SRR file	Sequencing instrument	Assembly size (M)	Nº of unique GO ID's	Unique GO ID's per M	General mean for unique GO ID's per Megabite	Individual mean of unique GO ID's per M > General Mean	Individual mean > 1
Data with assembly size from 10M to 50M	*C. miliaris*	Venom duct	SRR1548190	Illumina HiSeq 2000	10	3	0,30		-----	-----
	C. coronatus	Venom duct	SRR14407591	Illumina HiSeq 4000	12	2	0,17		-----	-----
	C. miliaris	Venom duct	SRR1544120	Illumina HiSeq 2000	14	1	0,07		-----	-----
	C. imperialis	Venom duct	SRR2609542	Illumina HiSeq 2000	14	7	0,50		YES - Illumina HiSeq 2000	-----
	C. miliaris	Venom duct	SRR1544597	Illumina HiSeq 2000	15	3	0,20		-----	-----
	C. miliaris	Venom duct	SRR1548188	Illumina HiSeq 2000	15	0	0,00		-----	-----
	C. miliaris	Venom duct	SRR1544142	Illumina HiSeq 2000	16	1	0,06		-----	-----
	C. miliaris	Venom duct	SRR1548185	Illumina HiSeq 2000	16	0	0,00		-----	-----
	C. miliaris	Venom duct	SRR1548186	Illumina HiSeq 2000	16	0	0,00		-----	-----
	C. coronatus	Venom duct	SRR2609545	Illumina HiSeq 2000	16	7	0,44		YES - Illumina HiSeq 2000	-----
	C. miliaris	Venom duct	SRR1544600	Illumina HiSeq 2000	17	1	0,06		-----	-----
	C. miliaris	Venom duct	SRR1544622	Illumina HiSeq 2000	17	5	0,29		-----	-----
	C. miliaris	Venom duct	SRR1544692	Illumina HiSeq 2000	17	6	0,35		YES - Illumina HiSeq 2000	-----
	C. virgo	Venom duct	SRR2608262	Illumina HiSeq 2000	17	14	0,82		YES - Illumina HiSeq 2000	-----
	C. coronatus	Venom duct	SRR14407592	Illumina HiSeq 4000	18	0	0,00		-----	-----
	C. miliaris	Venom duct	SRR1548189	Illumina HiSeq 2000	19	4	0,21		-----	-----
	C. miliaris	Venom duct	SRR1544690	Illumina HiSeq 2000	20	5	0,25		-----	-----
	C. coronatus	Venom duct	SRR2609544	Illumina HiSeq 2000	20	0	0,00		-----	-----
	C. miliaris	Venom duct	SRR1544140	Illumina HiSeq 2000	21	0	0,00		-----	-----
	C. sponsalis	Venom duct	SRR2609541	Illumina HiSeq 2000	21	7	0,33		YES - Illumina HiSeq 2000	-----
	C. imperialis	Venom duct	SRR12186678	Illumina HiSeq 2000	22	1	0,05		-----	-----
	C. imperialis	Venom duct	SRR12186679	Illumina HiSeq 2000	22	9	0,41		YES - Illumina HiSeq 2000	-----
	C. miliaris	Venom duct	SRR1544595	Illumina HiSeq 2000	22	15	0,68		YES - Illumina HiSeq 2000	-----
	C. miliaris	Venom duct	SRR1548192	Illumina HiSeq 2000	22	3	0,14		-----	-----
	C. miliaris	Venom duct	SRR1544119	Illumina HiSeq 2000	23	0	0,00		-----	-----
	C. miliaris	Venom duct	SRR1544137	Illumina HiSeq 2000	23	0	0,00		-----	-----
	C. rattus	Venom duct	SRR2609540	Illumina HiSeq 2000	23	0	0,00		-----	-----
	C. imperialis	Venom duct	SRR12186677	Illumina HiSeq 2000	24	3	0,13		-----	-----
	C. miliaris	Venom duct	SRR1548187	Illumina HiSeq 2000	24	3	0,13		-----	-----
	C. marmoreus	Venom duct	SRR2609532	Illumina HiSeq 2000	24	1	0,04		-----	-----
	C. miliaris	Venom duct	SRR1542681	Illumina HiSeq 2000	25	1	0,04		-----	-----
	C. quercinus	Venom duct	SRR2609537	Illumina HiSeq 2000	25	6	0,24		-----	-----
	C. ebraeus	Venom duct	SRR2609538	Illumina HiSeq 2000	25	1	0,04		-----	-----
	C. miliaris	Venom duct	SRR1542424	Illumina HiSeq 2000	26	4	0,15		-----	-----
	C. lividus	Venom duct	SRR2609539	Illumina HiSeq 2000	26	9	0,35	0,31	YES - Illumina HiSeq 2000	-----
	C. miliaris	Venom duct	SRR1544117	Illumina HiSeq 2000	28	0	0,00		-----	-----
	C. ermineus	Venom duct	SRR6983162	Illumina HiSeq 2500	29	1	0,03		-----	-----
	C. varius	Venom duct	SRR2609543	Illumina HiSeq 2000	30	6	0,20		-----	-----
	C. ermineus	Venom duct	SRR6983167	Illumina HiSeq 2500	30	0	0,00		-----	-----
	C. ermineus	Venom duct	SRR6983169	Illumina HiSeq 1500	32	1	0,03		-----	-----
	C. ermineus	Venom duct	SRR6983168	Illumina HiSeq 1500	33	2	0,06		-----	-----
	C. miliaris	Venom duct	SRR1544627	Illumina HiSeq 2000	34	1	0,03		-----	-----
	C. ermineus	Venom duct	SRR6983164	Illumina HiSeq 2500	36	2	0,06		-----	-----
	P. magus	Venom duct	SRR9831243	Illumina HiSeq 2500	38	12	0,32		YES - Illumina HiSeq 2500	-----
	V. ebraeus	Venom duct	SRR17653518	Illumina HiSeq 2500	39	0	0,00		-----	-----
	P. magus	Venom duct	SRR9831255	Illumina HiSeq 2500	46	4	0,09		-----	-----
	C. ermineus	Venom duct	SRR6983163	Illumina HiSeq 2500	48	2	0,04		-----	-----
Data with assembly size from 50M to 75M	*C. ermineus*	Venom duct	SRR6983161	Illumina HiSeq 2500	51	6	0,12		-----	-----
	C. ermineus	Venom duct	SRR6983165	Illumina HiSeq 2500	52	4	0,08		-----	-----
	C. imperialis	Venom duct	SRR12186676	Illumina HiSeq 2000	53	42	0,79		YES - Illumina HiSeq 2000	-----
	C. lenavati	Venom duct	SRR1803942	Illumina HiSeq 2000	62	7	0,11		-----	-----
	C. betulinus	Venom duct	SRR2124878	Illumina HiSeq 2000	63	24	0,38		YES - Illumina HiSeq 2000	-----
	C. litteratus	Venom duct	SRR6381569	Illumina HiSeq 2500	66	35	0,53		YES - Illumina HiSeq 2500	-----
	C. lenavati	Venom duct	SRR1803941	Illumina HiSeq 2000	67	97	1,45		YES - Illumina HiSeq 2000	YES
	V. judaeus	Venom duct	SRR17653514	Illumina HiSeq 2500	68	4	0,06		-----	-----
Data with assembly size >75M	*C. betulinus*	Venom duct	SRR2124879	Illumina HiSeq 2000	76	38	0,50		YES - Illumina HiSeq 2000	-----
	C. litteratus	Venom duct	SRR6381570	Illumina HiSeq 2500	76	90	1,18		YES - Illumina HiSeq 2500	YES
	C. tribblei	Venom duct	SRR1803939	Illumina HiSeq 2000	78	57	0,73		YES - Illumina HiSeq 2000	-----
	C. tribblei	Venom duct	SRR1803938	Illumina HiSeq 2000	83	55	0,66		YES - Illumina HiSeq 2000	-----
	C. tribblei	Venom duct	SRR1803937	Illumina HiSeq 2000	86	55	0,64		YES - Illumina HiSeq 2000	-----
	C. betulinus	Venom duct	SRR2124880	Illumina HiSeq 2000	87	42	0,48		YES - Illumina HiSeq 2000	-----
	L. ventricosus	Venom gland	SRR13740844	Illumina HiSeq 2000	89	54	0,61		YES - Illumina HiSeq 2000	-----
	C. ermineus	Venom duct	SRR6983166	Illumina HiSeq 1500	91	42	0,46		YES - Illumina HiSeq 1500	-----
	C. lenavati	Venom duct	SRR1803940	Illumina HiSeq 2000	101	262	2,59		YES - Illumina HiSeq 2000	YES
	C. betulinus	Venom duct	SRR2124881	Illumina HiSeq 2000	105	67	0,64		YES - Illumina HiSeq 2000	-----
	C. betulinus	Venom duct	SRR2124882	Illumina HiSeq 2000	106	53	0,50		YES - Illumina HiSeq 2000	-----
	C. consors	Venom bulb	SRR1964035	Illumina Genome Analyzer II	139	76	0,55		YES - Illumina Genome Analyzer II	-----
	C. tribblei	Venom duct	SRR1799982	Illumina HiSeq 2000	150	66	0,44		YES - Illumina HiSeq 2000	-----
	C. consors	Venom duct	SRR1954994	Illumina Genome Analyzer II	167	64	0,38		YES - Illumina Genome Analyzer II	-----

Fig. 30 – Table with the venom-related transcriptomes divided in categories of assembly size: bellow 50M in light green, between 50 and 75M in light yellow, and above 75M in blue. Each type of sequencing instrument has his own colour: blue for Illumina HiSeq 1500, green for Illumina HiSeq 2000, orange for Illumina HiSeq 2500, orange for Illumina HiSeq 4000 and yellow for Illumina Genome Analizer II. Information based on the variables of assembly size and number of unique GO IDs is presented from column 7 up to 10.

7.7. ggplot2 – normalized GO category by feeding habit (for shared genes of venom-related tissues)

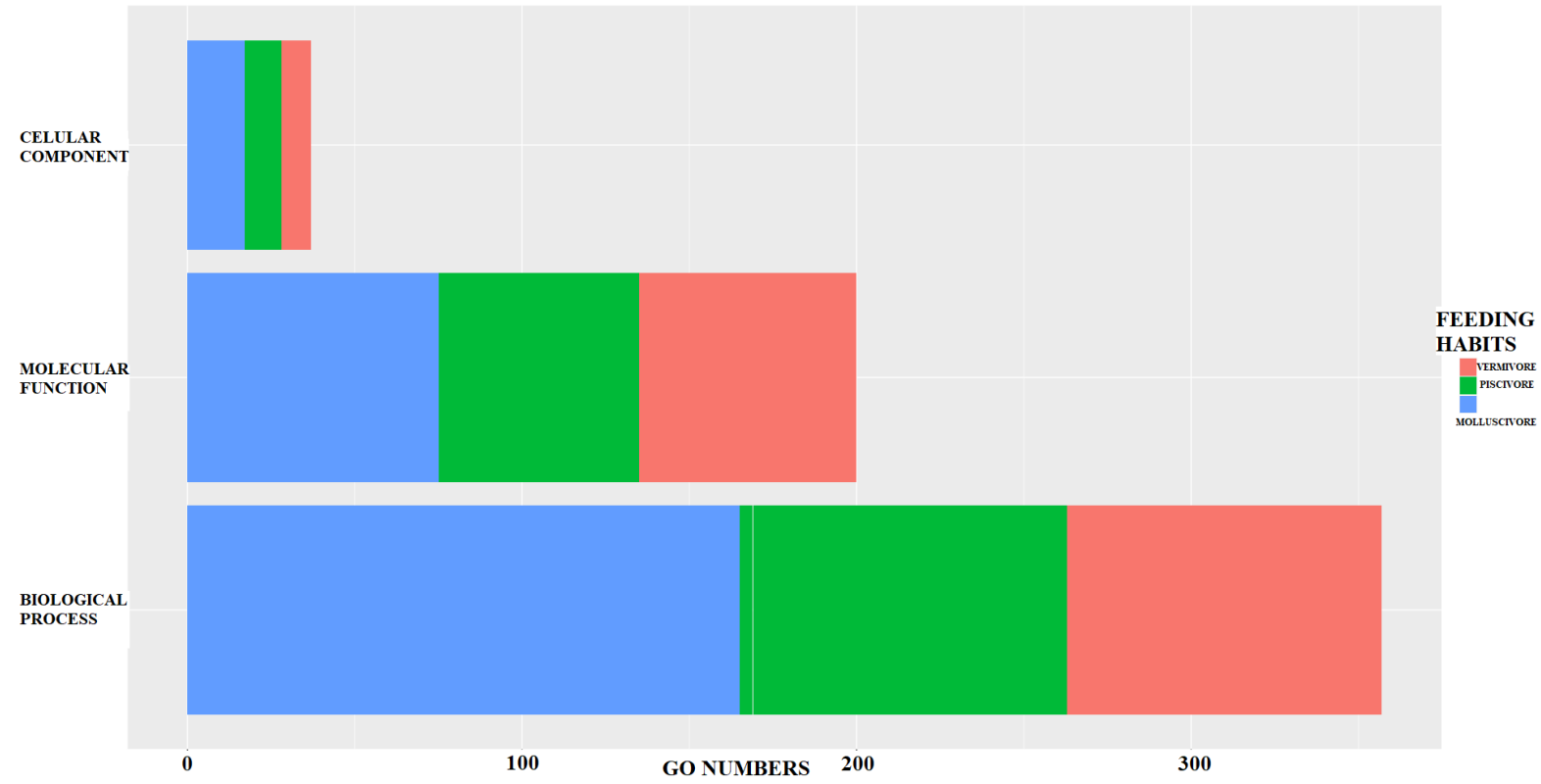


Fig. 31 – ggplot2 depicting the DE of the 29 shared GO IDs in the venom-related transcriptomes normalized by the GO category and feeding habit.

7.8. Table with full SARS-Cov-2 genome annotation

	Protein	Function
Cell' life cycle	Coronavirus replicase NSP3	cell cycle - an essential component of the replication/transcription complex
	Coronavirus papain-like peptidase	peptidase
	Coronavirus replicase NSP4	cell cycle - critical role in the replication of SARS-CoV through the rearrangements of host-derived membranes
	Coronavirus replicase NSP8	cell cycle
	Coronavirus RNA synthesis protein NSP10	RNA synthesis
	Coronavirus replicase NSP6	cell cycle
	Betacoronavirus replicase NSP3	cell cycle
	Coronavirus replicase NSP9	cell cycle
	Coronavirus replicase NSP7	cell cycle
	Betacoronavirus replicase NSP1	cell cycle
	Coronavirus replicase NSP2	cell cycle
	Coronavirus proofreading exoribonuclease	cell cycle
	Coronavirus RNA-dependent RNA polymerase	cell cycle
	Coronavirus replicase NSP15, uridylate-specific endoribonuclease	cell cycle
	Coronavirus replicase NSP15, N-terminal oligomerisation	cell cycle
	Coronavirus replicase NSP15, middle domain	cell cycle
	Viral RNA-dependent RNA polymerase	cell cycle
	AAA domain	cell-cycle regulation, protein degradation, organelle biogenesis and vesicle-mediated protein transport
	Viral (Superfamily 1) RNA helicase	cell cycle - bind and may even remodel nucleic acid or nucleic acid protein complexes
	UvrD-like helicase C-terminal domain	cell cycle - DNA repair, replication, and recombination
	MukF middle domain	Involved in chromosome condensation, segregation and cell cycle progression.
	Coronavirus nucleocapsid	interacts with the viral membrane protein during virion assembly; plays a critical role in virus transcription and assembly
	Ellis van Creveld protein 2 like protein	help regulate the signaling pathway Sonic Hedgehog (plays roles in cell growth and specialization, and the normal patterning of the body
	Adipogenin	stimulating adipocyte differentiation and development
	Betacoronavirus nucleic acid-binding	nucleic acid-binding
Mechanistic and metabolic functions	coronavirus endopeptidase	cysteine protease
	Betacoronavirus SUD-C domain	could be related to metal, adenylate and nucleic acid binding
	Coronavirus 2'-O-methyltransferase	O-methyltransferase
	Calcium-dependent calmodulin binding	Calcium-dependent calmodulin binding
	Protein of unknown function (DUF1664)	improved bacterial resistance to drought
	Tetramerisation domain of TRPM	temperature sensing, inflammation, insulin secretion, and redox sensing
	Syntaxin-like protein	bind synaptotagmin on SVs in response to calcium entry
	Biogenesis of lysosome-related organelles complex-1 subunit 2	genesis of organs to break down excess or worn-out cell parts even virus or bacteria
	EF-hand domain	binds calcium ions
	Betacoronavirus NS7A protein	transmembrane protein
	Intu longin-like domain 2	mechanistic functions and transport pathways
	Betacoronavirus lipid binding protein	lipid binding
	PHB de-polymerase C-terminus	degradation processes of a natural polyester Poly(3-hydroxyburate).
	GtrA-like protein	integral membrane proteins with three or four transmembrane spans
	FAM163 family	Predicted to be integral component of membrane
	NADH dehydrogenase subunit 2 N-terminal	converts NADH, the reduced form of nicotinamide adenine dinucleotide (NAD) to its oxidized form NAD+
	Transient receptor potential (TRP) ion channel	Ion channel
Pathogenical function	Coronavirus spike glycoprotein S2	viral infection
	Betacoronavirus-like spike glycoprotein S1, N-terminal	viral infection
	Betacoronavirus spike glycoprotein S1, receptor binding	viral infection
	Coronavirus spike glycoprotein S1, C-terminal	viral infection
	Coronavirus spike glycoprotein S2, intravirion	viral infection
	Baculovirus polyhedron envelope protein, PEP, C terminus	receptor binding and fusion
	Retroviral envelope protein	involved in several aspects of the virus' life cycle, such as assembly, budding, envelope formation, and pathogenesis
	Laminin Domain II	interact with receptors anchored in the plasma membrane of cells adjacent to basement membranes
	Betacoronavirus viroporin	to participate in virion morphogenesis and release from host cells
	Coronavirus M matrix/glycoprotein	transmembrane glycoproteins, defines the shape of the viral envelope, central organiser of coronavirus assembly
	M penetrans paralogue family 26	aids in cytadherence, the adherence to respiratory epithelium
	Betacoronavirus NS8 protein	might be involved in endoplasmic function
	Coronavirus small envelope protein E	involved in several aspects of the virus' life cycle, such as assembly, budding, envelope formation, and pathogenesis
	Betacoronavirus NS6 protein	to prevent both nuclear import and export, which renders host cells incapable of responding to SARS-CoV-2 infection
	Betacoronavirus NS7B protein	structural component of SARS-CoV virions and an integral membrane protein, Its transmembrane domain is essential for Golgi compartment localization
Unknown function	DUF2959	Proteins of unknown function
	SlyX	
	UPF0184	
	DUF1461	
	Betacoronavirus uncharacterised protein 14 (SARS-CoV-2 like)	
	Baculovirus 11 kDa family	
	DUF3587	

Fig. 32 – Full annotation result of the entire SARS-Cov-2 genome processed against the Pfam database. In yellow are the proteins related to cell's life cycle; in green are proteins with mechanistic and metabolic functions; in blue are proteins involved in the pathogenic pathways of the virus; finally, in orange are the proteins which are recognized but whose function is unknown.