



HAL
open science

DeTox: a pipeline for the detection of toxins in venomous organisms

Allan Ringeval, Sarah Farhat, Alexander Fedosov, Marco Gerdol, Samuele Greco, Lou Mary, Maria Vittoria Modica, Nicolas Puillandre

► To cite this version:

Allan Ringeval, Sarah Farhat, Alexander Fedosov, Marco Gerdol, Samuele Greco, et al.. DeTox: a pipeline for the detection of toxins in venomous organisms. *Briefings in Bioinformatics*, 2024, 25, 10.1093/bib/bbae094 . hal-04548190

HAL Id: hal-04548190


<https://hal.science/hal-04548190>

Submitted on 16 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DeTox: a pipeline for the detection of toxins in venomous organisms

Allan Ringeval , Sarah Farhat , Alexander Fedosov, Marco Gerdol, Samuele Greco, Lou Mary, Maria Vittoria Modica and Nicolas Puillandre

Corresponding author. Ringeval Allan, Institut Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, 75005 Paris, France. E-mail: allan.ringeval@mnhn.fr

Abstract

Venomous organisms have independently evolved the ability to produce toxins 101 times during their evolutionary history, resulting in over 200 000 venomous species. Collectively, these species produce millions of toxins, making them a valuable resource for bioprospecting and understanding the evolutionary mechanisms underlying genetic diversification. RNA-seq is the preferred method for characterizing toxin repertoires, but the analysis of the resulting data remains challenging. While early approaches relied on similarity-based mapping to known toxin databases, recent studies have highlighted the importance of structural features for toxin detection. The few existing pipelines lack an integration between these complementary approaches, and tend to be difficult to run for non-experienced users. To address these issues, we developed DeTox, a comprehensive and user-friendly tool for toxin research. It combines fast execution, parallelization and customization of parameters. DeTox was tested on published transcriptomes from gastropod mollusks, cnidarians and snakes, retrieving most putative toxins from the original articles and identifying additional peptides as potential toxins to be confirmed through manual annotation and eventually proteomic analysis. By integrating a structure-based search with similarity-based approaches, DeTox allows the comprehensive characterization of toxin repertoire in poorly-known taxa. The effect of the taxonomic bias in existing databases is minimized in DeTox, as mirrored in the detection of unique and divergent toxins that would have been overlooked by similarity-based methods. DeTox streamlines toxin annotation, providing a valuable tool for efficient identification of venom components that will enhance venom research in neglected taxa.

Keywords: bioinformatic pipeline; venom; toxin detection; transcriptomics

INTRODUCTION

The ability to produce a venom, a mix of molecules used to disrupt the normal physiological processes in another organism [1], is a trait that evolved independently over 101 times across the animal kingdom [2], leading to a total of more than 200 000 described venomous species on Earth [2]. Since each species is able to produce from a few tens to a few hundreds of unique toxins [3–5], venomous organisms represent a fantastic reservoir of millions of toxins. Some of these organisms also represent a serious threat for human health: snakes only are responsible for more than 100 000 casualties per year [6], and developing effective and specific antivenoms is thus a critical task in venom research. At the same time, some of the toxins isolated from venomous animals have proved efficient in human therapeutics, and 12 toxins from snakes, lizards, leeches, bees and cone snails are now approved for clinical use in humans [7]. In addition to their potential practical

value, toxin peptides constitute a relevant model to understand the molecular processes through which animals have evolved new adaptations to interact with their environments and successfully diversify.

The key step of venom research is the detection of toxins and the characterization of their peptide sequences. With the development of Next-Generation Sequencing techniques, RNA-seq of venom-producing tissues is now the favored and most efficient approach to quickly access the toxin repertoire of a venomous animal [8]. Compared with previously used methods (Sanger sequencing of Expressed Sequence Tags or proteomics), RNA-seq techniques, thanks to their higher throughput and dynamic range, cover more exhaustively the transcript diversity and allow an estimation of the relative quantity of each transcript. On the other hand, a single RNA-seq run can provide millions of reads, leading to tens of thousands of transcripts after assembly and navigating through the data to quickly and efficiently

Allan Ringeval is a PhD student at the Muséum National d'Histoire Naturelle, Paris (MNHN).

Sarah Farhat is a postdoc at the MNHN.

Alexander Fedosov is senior curator at the Swedish Museum of Natural History.

Marco Gerdol is a researcher at the University of Trieste.

Samuele Greco is a postdoc at the University of Trieste.

Lou Mary is a post-doc at the MNHN.

Maria Vittoria Modica is a researcher at the Stazione Zoologica Anton Dohrn.

Nicolas Puillandre is an assistant professor at the MNHN.

Received: November 14, 2023. Revised: January 29, 2024. Accepted: February 16, 2024

© The Author(s) 2024. Published by Oxford University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

target the contigs that potentially correspond to toxins remains challenging.

When the first transcriptomes from venom-producing tissues were obtained, researchers relied on available toxin databases to detect significant sequence similarity between the assembled transcripts and a reference database, based on arbitrarily defined similarity threshold [9–12]. Later on, methods based on probabilistic approaches and machine learning were introduced to enhance the detection process in transcriptomic data: unlike similarity methods, these approaches can even detect toxins that significantly diverge from previously described ones [13–16]. Nevertheless, all these reference-based approaches are limited to the detection of toxins that are similar to the ones already characterized in the toxin database or included in the training data set [17].

More recently, several studies demonstrated the utility of using structural features in addition to primary sequence similarity to detect potential toxins in transcriptomes. For example, Fassio et al. [18] and Fedosov et al. [17] relied on the detection of characteristic structural features of the secreted toxins to guide *de novo* annotation of the potential toxin transcripts. These features are (i) the presence of an N-terminal signal peptide sequence, which directs processing of the secretory peptide precursor, (ii) the absence of transmembrane domain(s) and (iii) the presence of a cysteine framework—a conservative arrangement of cysteine residues in the mature peptide that define its spatial conformation through disulfide bridges [19, 20]. This strategy has proven efficient, allowing identification of those toxins that remained undetected with similarity approaches [17]. However, if similarity-based approaches are formalized in several pipelines, such as ToxCodAn or Venomix [21–24], they rarely implement a structural approach. Furthermore, the available tools are not always maintained, often lacking updates, or, in case of web-hosted tools, losing accessibility. Another common issue is the lack of a user-friendly interface that makes the use of these tools not always straightforward for non-experienced users. In addition, none of these pipelines integrate the whole process, from the transcriptome assembly to the annotation, nor a gene expression quantification step, implying that these bioinformatic analyses should be carried out separately using different tools.

Here, we present DeTox, a new tool for the Detection of peptide Toxins in venomous organisms. DeTox integrates both the similarity- and structure-based approaches, relying on a toxin database for the former and on the presence of structural features commonly reported in toxins for the latter. DeTox was primarily developed for the detection of toxins in neogastropod mollusks, but it can be used for any group of venomous organisms. The main goal of developing DeTox is to assist researchers in a rational selection of toxin candidates based on a set of customizable criteria. It provides a user the opportunity to emphasize some criteria over others, depending on the species of interest and quality of input data, to efficiently identify toxin candidates. DeTox is built as a user-friendly tool, overcoming the limitation of the previously available pipelines. To illustrate its use, we applied DeTox to a range of published transcriptomic data sets from neogastropods, cnidarians and snakes.

MATERIAL AND METHODS

Overview of the pipeline

DeTox is conceived as a pipeline for *de novo* toxin detection in transcriptomes (Figure 1). The approach is based on a three-step methodology that allows the identification of similar toxins

and potential assignation of functional annotations. First, DeTox identifies all transcripts that encode peptides showing typical structural characteristics of toxins. Second, a similarity search is performed between the predicted proteins and a database of known toxins specific to the taxonomic group of interest, provided by the user in fasta format (thereafter referred to as the 'toxin database'). Third, it optionally provides gene expression levels as a critical information for toxin/non-toxin discrimination (when toxin-producing tissues are specifically targeted). Importantly, only predicted proteins that lack the expected structural features or without any hit on the toxin database are excluded: all the others are retained for further investigation. Although the pipeline is designed to work on preassembled transcriptomes, it also provides users the option to run the *de novo* assembly step, by accepting raw RNA-seq data generated with Illumina platforms as an input. The pipeline is developed using snakemake [25]. If the assembled transcriptome is not provided, the pipeline will automatically run the *de novo* assembly step using the paired-end ('r1' and 'r2' options) or single-end ('r1' only) file(s) provided by the user. The toxin database is provided by the user in the 'toxin_db' option. The pipeline generates a table summarizing all the information for each putative toxin detected, including its functional annotation (Table 1).

Reads quality cleaning and assembly

This part of the pipeline provides a simple and straightforward approach for Illumina reads cleaning and assembly. The raw reads are filtered based on quality and the adapters are removed using Trimmomatic [26] v0.39 (default parameter: 'ILLUMINACLIP:Adapters.fa:2:40:15 LEADING:15 TRAILING:15 MINLEN:36 SLIDINGWINDOW:4:15'). The reads are then assembled using Trinity [27] v2.15.0 (with default parameters). Since the implemented transcriptome assembly option is not tailored for RNA-seq data generated with other platforms (e.g. ONT, PacBio, etc.), we advise the users to adopt gold-standard methods for reads trimming and assembly, specific to their technology of choice. Regardless of the origin of the transcriptome assembly, we strongly advise the users to filter the contigs from possible contamination by providing a contaminant database. This functionality can be accessed through the 'contaminants' parameter in the DeTox configuration file. We recommend including sequences of bacteria, protists and fungi from GenBank [28], as well as ribosomal RNAs from, e.g. Silva database [29] (Supplementary Material 1 available online at <http://bib.oxfordjournals.org/>). Putative contaminant sequences are removed based on the detection of significant sequence homology with BLASTn v2.14.1 [30] (with the option 'contamination_evalue') against the reference database specified by the user. Then, DeTox identifies the open reading frames (ORF) in all contigs using orfipy v0.0.4 [31]. The user can set the 'minlen' option defining the minimum length of the reading frame (33 amino acids by default), 'maxlen' option defining the maximum length of the reading frame (not defined by default) with the addition of 'partial-3' and 'partial-5' to retrieve ORFs without the N-terminal Methionine and/or stop codon. Because of possible redundancy in the resulting translated transcripts, sequences are clustered using CD-HIT v4.8.1 [32] with an identity threshold ('clustering_threshold' option, 0.99 by default), and retaining the longest sequence used as a basis for alignment in each cluster. Please note that this threshold may need to be adjusted depending on the quality of raw RNA-seq input data, assembly strategy and genetic background of the target species.

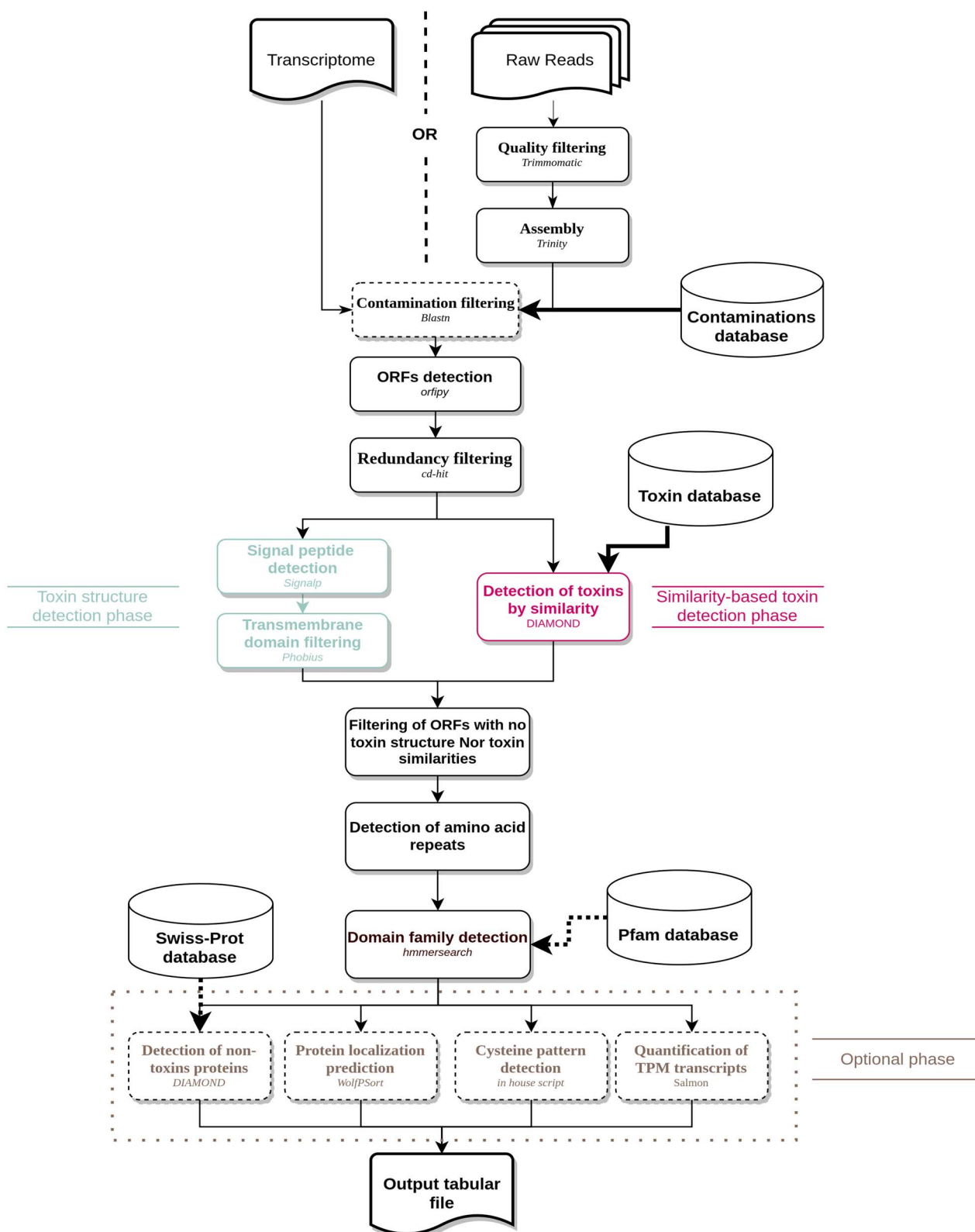


Figure 1. Detox pipeline outline. Diagram of all the steps of the pipeline with the corresponding tools used, including the necessary input files and databases. Dotted rectangles represent optional features of the pipeline.

Secretome and toxin peptides detection

The identified peptide sequences are analyzed based on their structural features. The presence of a signal peptide in the sequence indicates that the newly synthesized protein is targeted

to the classical secretory pathway, thereby being either secreted outside the cell, resident in the endoplasmic reticulum, Golgi or endosomes or (if transmembrane regions are also present) inserted in cell membranes. To detect signal peptides, Signalp

Table 1: General summary of DeTox outputs for the 11 analyzed transcriptomes, with comparison with the results from the original articles

Species	<i>P. vaubani</i>	<i>P. neocaledonicus</i>	<i>C. ebraeus</i>	<i>C. reticulata</i>	<i>A. elegantissima</i>	<i>C. andreae</i>	<i>C. cantherigerus</i>	<i>T. variabilis</i>
Toxins identified in the original publication	159	86	298	136	128	35	4	11
Among them, toxins identified by DeTox	141 (89%)	81 (94%)	286 (96%)	124 (91%)	112 (88%)	34 (97%)	4 (100%)	10 (91%)
Total number of peptides detected by DeTox	22 101	12 795	8040	11 654	35 069	6261	4185	7663
Number of DeTox peptides with a TPM >1000	38 (11)	30 (4)	36 (23)	53 (3)	104 (8)	15 (0)	7 (0)	8 (0)
Number of DeTox peptides with a hit against the Toxin database (PID > 95)	3 (2)	3 (1)	112 (76)	78 (48)	5 (2)	4 (0)	1 (0)	3 (0)
Number of transcriptomes	2	2	1	1	2	1	1	1
Read count	129 629 574	78 096 694	58 508 016	329 576 088	272 940 126	54 593 038	50 762 546	56 115 370
Transcripts count	274 068	134 537	68 809	144 380	380 109	59 031	26 751	57 467
DeTox runtime	2 h 24 min	2 h 5 min	45 min	2 h 4 min	2 h 38 min	49 min	46 min	42 min

Note: Percentages into parenthesis correspond to the proportion of toxins recovered by DeTox among the toxins identified in the original publication. Numbers into parenthesis correspond to the number of secreted peptides detected by DeTox but not found in the original articles. PID: percentage of identity. The last line corresponds to the execution time of DeTox on 16 cores/32Gb RAM, without performing the assembly step.

v5.0b [33] is used with the arguments '-org euk -format short -verbose -batch 5000'. Since this program is increasingly time-consuming with large input files, only the first 50aa of each translated transcript are distributed into multiple files, each containing 5000 peptide sequences to serve as an input for SignalP in individual runs. Then, the output files are merged into one file and only sequences with a signal peptide having a D-value higher than a user-specific threshold ('signalp_dvalue' option, 0.8 by default) are kept. Please note that a small number of eukaryote peptides, subject to leaderless secretion and lacking canonical signal peptides, might be discarded by DeTox; if the users suspect that a few toxins in the target organism may fall within this category, they are advised to recover additional matches using, e.g. SecretomeP [34]. The ORFs encoding peptides having one or more transmembrane domains and therefore associated with cell or organelle membranes are detected using Phobius v1.01 [35] and discarded.

All the peptides (whether they display toxin structural features or not) are aligned against the toxin database to ensure the detection of already known or low divergence toxins, even if they lack a signal peptide or if they possess a transmembrane domain. This ensures the identification of toxins recovered in partially assembled contigs lacking their N-terminal region, deriving from the translation of incomplete or misassembled transcripts. The diamond program v2.1.8 [36] aligns the peptides against the toxin database provided by the user ('toxin_db' option), which is required. Only hits, with an E-value below the threshold provided by the 'toxin_evalue' option (1E-10 by default), are kept for each predicted protein. Then, the best hit for each sequence is reported in the final output table. At this step, the peptides are retained only if they have a signal peptide and lack a transmembrane domain, or have a hit against the provided toxin database.

Some predicted proteins may contain repeats of one or several amino acids which may indicate a sequencing or assembly artifact. We included an in-house code to automatically detect repeats of one, two or three amino-acids, reporting them in the final output file. The minimum number of repeated units can be changed with the 'repeat_length' option (default is 5).

The remaining peptides are then analyzed against the Hidden Markov Models database to detect functional domains. DeTox employs HMMER v3.3.2 [37] with the '-cut_ga' argument (set by

default) against the Pfam database [38]. If the user did not provide the database path using the options 'pfam_database_path' and/or 'swissprot_database_path', DeTox automatically downloads the latest version. For each predicted protein, all domain predictions are reported according to the E-value (the HMMER output file can also be found in the working directory).

The analysis of the putative toxin repertoire is completed by four additional steps that are integrated as optional in the pipeline. (1) With the 'swissprot' option set to 'True', an alignment is carried out using diamond program v2.1.8 [36] on SwissProt database [39] to provide information on the inferred function. The E-value threshold can be changed ('swissprot_evalue' option, 1E-10 by default). The SwissProt database is automatically downloaded if its path is missing in the options. This step can prove valuable when a putative secreted peptide lacks a match in the toxin database. Indeed, a match in SwissProt might reveal a toxin-related function, even in the absence of a match in the toxin database, or, on the contrary, suggest that the peptide is not related to a toxin function if it matches another function in SwissProt. (2) A prediction of the cell location of the peptide using the WolfPSort program [40] ('wolfpsort' option to 'True') is used to estimate the probability for a sequence to be secreted. (3) The detection of cysteine patterns by counting the number of cysteine residues and their relative position, determining whether they are vicinal or not in the protein sequence, is carried out when the 'cys_pattern' parameter is set to 'True'. Cysteine patterns are indicators of the spatial conformation of the protein as defined by disulfide bridges, a common feature in toxins from gastropods and other venomous organisms. (4) The computation of gene expression levels, reported as Transcript Per Million (TPM) metric for each transcript, is performed using Salmon v1.10.2 [41] with the 'quant' parameter set to 'True' and required the provision of RNA-seq illumina reads ('r1' and 'r2'). The TPM metric efficiently compares transcript expression levels within and between samples, correcting for transcript length and sequencing depth [42, 43]. Although TPM can be estimated using different approaches, DeTox uses the k-mer-based counting strategy implemented by Salmon to minimize computation time, using raw (or trimmed) RNA-seq data as an input. Salmon can be used with all type of input data, regardless of the sequencing platform used to generate the reads.

Output file

At the end of a successful run, all the results are summarized in a tsv formatted table provided in the current directory unless the user sets the absolute path of the directory by setting the option 'output_dir'. In this table, each row corresponds to a peptide identified by DeTox, with the following information reported: an identifier marked as 'ID' (a concatenation of the transcript identifier and the ORF number), the complete amino acid sequence marked as 'Sequence', the peptide signal sequence marked as 'signal_prediction' with the information about the probability of the signal ('prob_signal'), the position of the signal ('cutsite'), the predicted localization of the peptide in the cell ('wolfsort_prediction'), all Pfam domains separated by ';' marked as 'pfam_domains', three columns including the best match against the toxin database with the identifier, name and function ('toxinDB_sseqid'), the identity percent ('toxinDB_pident') and the evalue ('toxin_DB_evalue'), three columns including the best match against the SwissProt database with the identifier ('uniprot_sseqid'), the identity percent ('uniprot_pident') and the evalue ('uniprot_evalue'), two columns for the repeats of one, two or three amino acids with the type of repeat ('RepeatsTypes') and its sequence length ('RepeatsLengths'), the mature peptide if present ('mature_peptide'), the cysteine pattern if present ('Cys_pattern'), the identifier of the transcript from which the protein originates, its length and effective length ('Length'), the expression level of the associated gene ('TPM') and the Rating system ('Rating'). Only the best match, according to the score, is reported in the final table for the alignments of the peptides against the different databases, but the original output files of the intermediate steps can be found in the working directories if further information is needed.

It should be noted that not all the peptides retrieved by DeTox correspond to toxins, as our focus is on minimizing filtering steps to reduce false negatives. Results are likely to include a certain number of false positives, requiring users to further explore them at their discretion. To help determining the peptides' putative function, DeTox provides a flag system. Flags such as 'S' denote the presence of a signal peptide without transmembrane domains, '*' indicates the absence of a peptide signal, 'B' signifies a positive hit in the toxin database, 'C' denotes a cysteine pattern, 'T' indicates a TPM metric above 1000, 'D' represents a predicted domain and '!' signifies a hit in the UniProt database but not in the toxin database. Consequently, a peptide with the 'SBCT' flag can be considered as a strong toxin candidate, whereas 'SD!' suggests a secreted protein with no toxic function. Peptides flagged 'S' are likely false positive lacking additional evidence besides secretion. Despite this, they are retained in results to provide users with valuable information for informed decision-making and application of personalized filtering criteria.

Test on empirical data sets

A set of 11 transcriptomes was selected to test the effectiveness of the DeTox pipeline. These transcriptomes are derived from different major animal lineages, comprising species known to produce toxins, including *Profundiconus* spp. (Neogastropoda, Conidae; [18]), *Conus ebraeus* (Neogastropoda, Conidae; [44]), *Cumia reticulata* (Neogastropoda, Colubrariidae; [45]), *Anthopleura elegantissima* (Anthozoa, Actiniidae; [46]) and *Caraiba andreae*, *Cubophys cantherigerus* and *Tretanorhinus variabilis* (Serpentes, Dipsadidae; [47]).

De novo transcriptome assembly was omitted as assemblies were obtained from original articles or provided by authors.

Likewise, contamination removal was not performed because of potential bias in benchmarking, as toxins were inadvertently removed in preliminary DeTox tests. *A. elegantissima*'s transcriptome was sourced from NCBI. For the two *Profundiconus*, *C. ebraeus* and *A. elegantissima*, transcriptomes, DeTox was run with the databases and parameters used in the original studies to enable comparison. The *C. reticulata* transcriptome was analyzed with an in-house database, created using Conoserver [48], supplemented by all the SwissProt entries containing the words 'conopeptide' and 'conotoxin', the entries of SwissProt for Neogastropoda (taxonomy_id:6479) with the keyword 'toxin', the Tox-prot database entries as well as published sequences of neogastropod toxins not available in Conoserver nor Uniprot [39]. For *C. andreae*, *C. cantherigerus* and *T. variabilis*, Venomzone database (<https://venomzone.expasy.org/>) was used, focusing on toxins related to Serpentes.

To compare the list of peptides detected by DeTox and the putative toxins published in the original articles, an alignment of the two lists was performed using BLASTp, with an E-value threshold of $1E-5$ and max_target_seq=1. For *C. reticulata*, the toxin sequences were not provided in the original study, and we used the contig identifiers to compare the list of toxins candidates provided by DeTox and the list of toxins from the original article. When a contig from the original study was clustered with a peptide detected by DeTox, the toxin from the original study was considered as detected by DeTox.

To follow the methodology used in the original article, DeTox was run separately on the two transcriptomes of *A. elegantissima* which represent two distinct tissues in the same specimen (aggressive and nonaggressive polyps) and the results were then merged into a single final output file.

RESULTS

DeTox benchmark

DeTox's performance was assessed by comparing its detected peptide lists with putative toxins in five original articles [18, 44–47] across different taxonomic classes (Gastropoda, Reptilia and Anthozoa). To ensure consistency, toxin databases were sourced directly from publications or reconstructed following their methods. DeTox, with all optional steps, recovered 88–100% of the toxins found in each original article. Additionally, it identified extra putative toxins, including highly expressed secreted peptides in toxin-producing tissues or those with homology to toxin database entries (Table 1). Detailed results for each data set are provided below.

Profundiconus spp.

Fassio et al. [18] identified 245 putative toxins in both transcriptomes of *Profundiconus vaubani* and *Profundiconus neocaledonicus*. DeTox identified 34 896 putative secreted peptides, including 53 peptides encoded by highly expressed transcripts (TPM > 1000) in the venom gland not detected in the original study. Additionally, three newly identified peptides matched the toxin database. DeTox missed 23 toxins; seven had a transmembrane domain, seven others had a different ORF probably longer than the one selected in the original paper and nine lacked peptide signals, even though the right ORF was identified (Table 1 and Figure 2).

Most of the sequences detected by DeTox had only a signal peptide detected (29 033). A set based on structural features retained 4396 sequences (signal peptide and a cysteine pattern). Then, additional sequences were detected either with the similarity-based criterion only (26) or with a combination of structural and

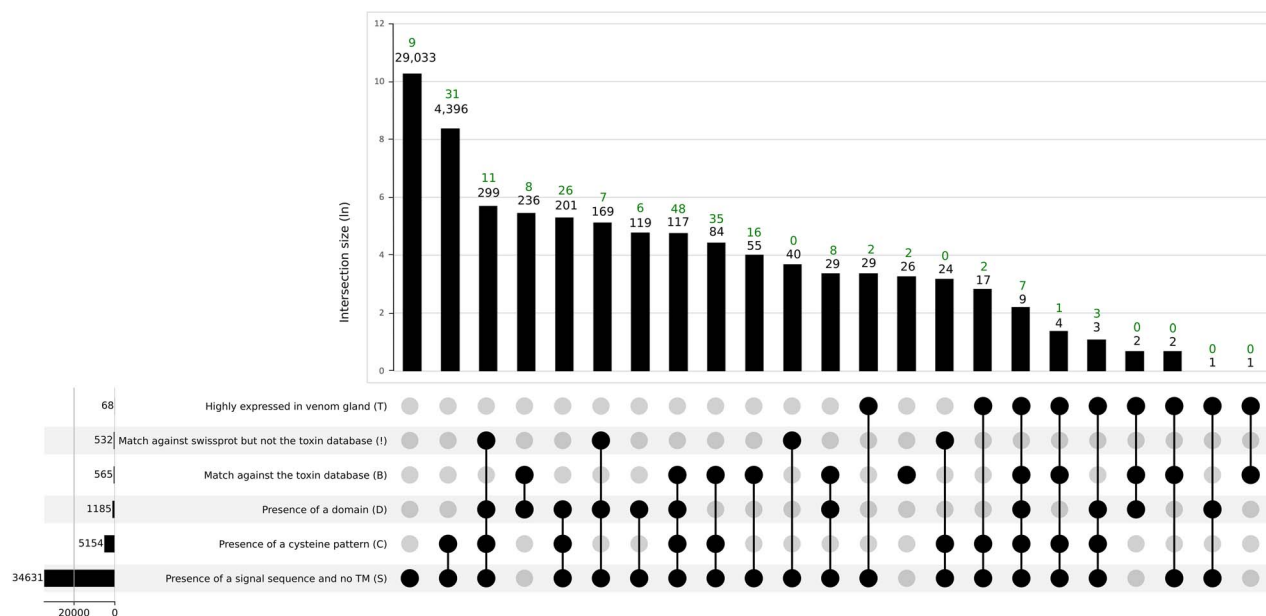


Figure 2. UpsetR diagram for *Profundiconus*. Distribution (barplot on the upper part) of all the sequences detected by DeTox with the various (combinations of) criteria (dots). Highly expressed putative toxins have a TPM > 1000. Top numbers above each bar represent the sequences also found in the original article.

similarity-based criteria. In this context, the potential toxins of particular interest are those that demonstrate a combination of high expression levels in the venom gland, structural indications and a match in the toxin database, with no corresponding entry in the SwissProt database and absence of functional domains except those related to toxins. If we consider the flag SCBT, for example, four sequences fulfill all four criteria, whereas only one was annotated as a toxin in the original study.

Conus ebraeus

DeTox identified 96% (286) of the 298 toxins in the original study [44]. Among the 12 undetected toxins, three had a transmembrane domain, four lacked signal peptide (probably because of a difference in the use of SignalP) and five had different ORFs. Moreover, 23 venom gland-expressed peptides and 76 matching toxin database entries were newly discovered. Similar to *Profundiconus*, most DeTox-detected sequences had only a signal peptide (6344), whereas 946 were based on structural features only. Peptides with only a signal peptide or a combination of a signal peptide and a cysteine pattern are more likely to be false positives, as well as peptides with the '!' flag. Notably, nine SCBT-flagged, including one new discovery and six TB-flagged peptides not previously detected emerged as potential toxin candidates. Further investigation, including toxin database curation and expression pattern analysis, is warranted to validate these promising findings (Table 1 and Figure 3).

Cumia reticulata

The transcriptome of *C. reticulata* is the only one analyzed using DeTox with an in-house toxin database optimized for neogastropods. The pipeline was able to find 124 of the 136 originally reported toxins. Among the 12 undetected toxins, seven did not return the same ORF when analyzed with orfipy, including one removed because of the sequence length falling below the minimum size threshold. Also, one sequence contains a transmembrane domain. For the four remaining ones, neither signal peptide nor match in the toxin database were detected by DeTox maybe because of differences in the toxin database and parameters of

signal that are different from the original study. But DeTox was able to identify additional 47 highly expressed transcripts, potentially related to toxins functions (Flag T). Further 566 transcripts, not detected in the original study, were annotated with the toxin database (Flag B), and may correspond to toxins as well (Table 1 and Figure 4).

Anthopleura elegantissima

DeTox, applied to the two original study's transcriptomes [46], successfully identified 88% of initially detected toxins (112 of 128). Among the 16 undetected, three had different ORF and 13 lacked both a peptide signal and a toxin database match. Challenges in reconstructing the reference database contributed to retrieval issues, as some GenBank identifiers used in toxin database construction were untraceable. Discrepancies in SignalP versions (4.0 in the original study, 5.0 in DeTox) impacted signal region probabilities estimates. Additionally, 104 secreted peptides with high expression levels, and five with toxin database annotations, were revealed. Most sequences (25 208) were detected only with a signal sequence, and thus are likely to be false positives. However, peptides flagged SCBT (two new sequences), SB (five new sequences) and B (16 new sequences), present promising toxin candidates, assuming that the toxin database is reliable (Table 1 and Figure 5).

Caraiba andreae, Cubophis cantherigerus and Tretanorhinus variabilis

The transcriptomes of *C. andreae*, *C. cantherigerus* and *T. variabilis* were analyzed using DeTox separately recovering 18 109 peptides, including 48 of the 50 original toxins (96%) [47]. DeTox excluded a sequence of *T. variabilis* with a transmembrane domain. The exclusion of another sequence in *C. andreae* is likely attributed to the utilization of the Venomzone database instead of a blast approach on the non-redundant NCBI and Uniref100 nucleotide database in the original study. Some flagged peptides, like SD!T, might not be toxins, possibly representing highly expressed, non-toxic secreted peptides. However, three newly detected peptides, absent in the original study, exhibit high expression and a signal

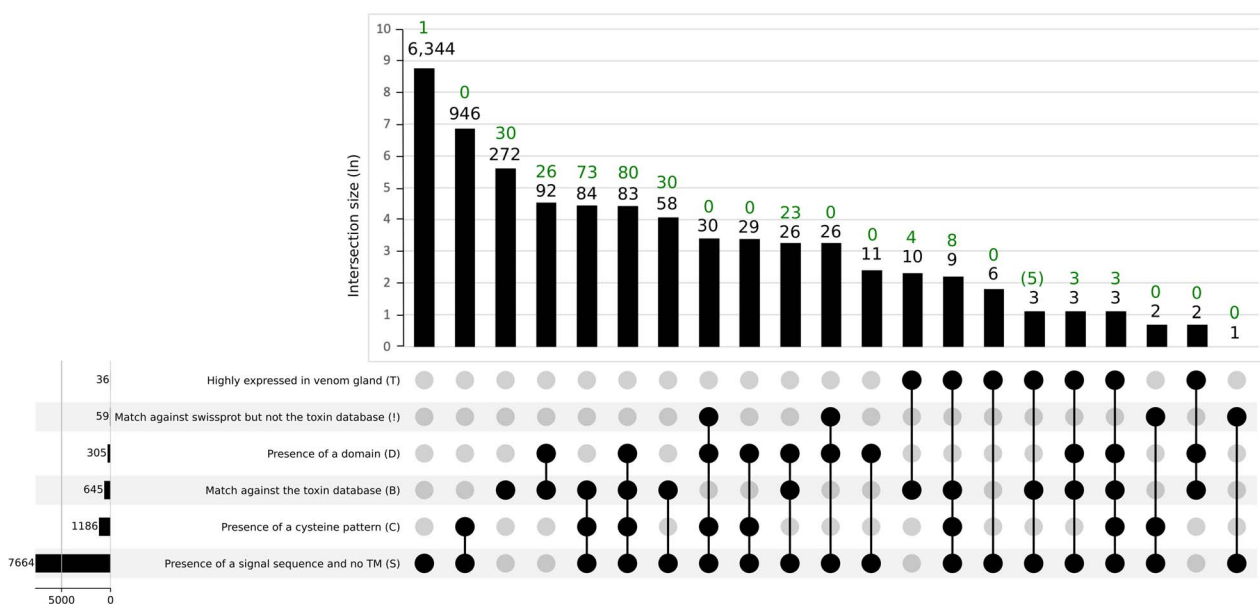


Figure 3. UpsetR diagram for *C. ebraeus*. Distribution (barplot on the upper part) of all the sequences detected by DeTox with the various (combinations of) criteria (dots). Highly expressed putative toxins have a TPM > 1000. Top numbers above each bar represent the number of sequences also found in the original article. The number in parenthesis indicates that multiple original sequences are included in one of the putative toxins identified by DeTox.

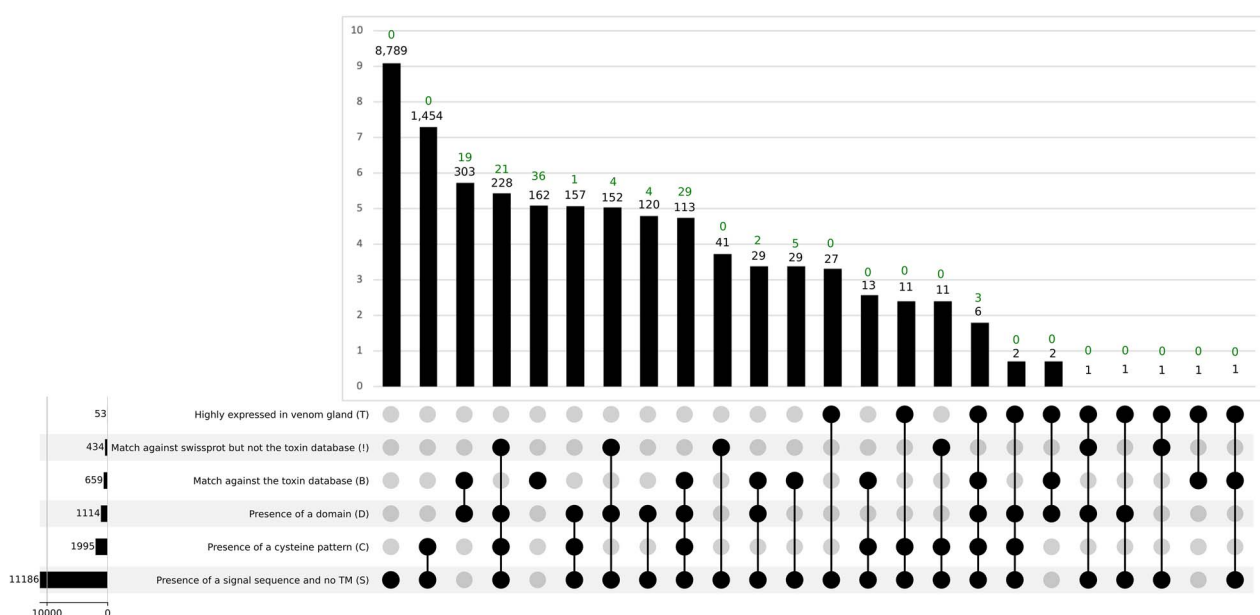


Figure 4. UpsetR diagram for *C. reticulata*. Distribution (barplot on the upper part) of all the sequences detected by DeTox with the various (combinations of) criteria (dots). Highly expressed putative toxins have a TPM > 1000. Top numbers above each bar represent the sequences also found in the original article.

peptide, suggesting they may be toxins. Additionally, 10 peptides with both a signal peptide and cysteine pattern, coupled with either a toxin database match (five peptides) or high expression in the venom gland (five peptides), represent further potential toxin candidates (Table 1 and Figure 6).

DISCUSSION

Over the past 15 years, RNA-seq has replaced Sanger sequencing of EST libraries, enabling comprehensive characterization of toxin repertoires in several venomous organisms [4, 10, 49–52]. However, with increasing accessibility to transcriptomic data, the bottleneck has shifted from sequencing to data analysis. Toxins are encoded by relatively short coding sequences (CDSs),

and while animal genomes typically contain no more than 40–50 000 ORFs, transcriptomes predict far more ORFs, making it challenging to distinguish real ORFs from bioinformatics artifacts. Existing tools like Conodictor 2 [14], Venomix [23] or ToxCodAn [22] are unsuitable for handling large data sets, face issues with maintenance and software compatibility, and lack integration across all steps, from assembly to toxin identification, requiring manual execution of multiple programs. For instance, Venomix necessitates three input files, including alignment against a toxin database, an assembled transcriptome and a gene expression quantification file, complicating the process. The time required for cleaning the data and integrating the assembly runs into an analysis pipeline is significant when searching for toxins in multiple transcriptomes, ranging from tens to hundreds. To address

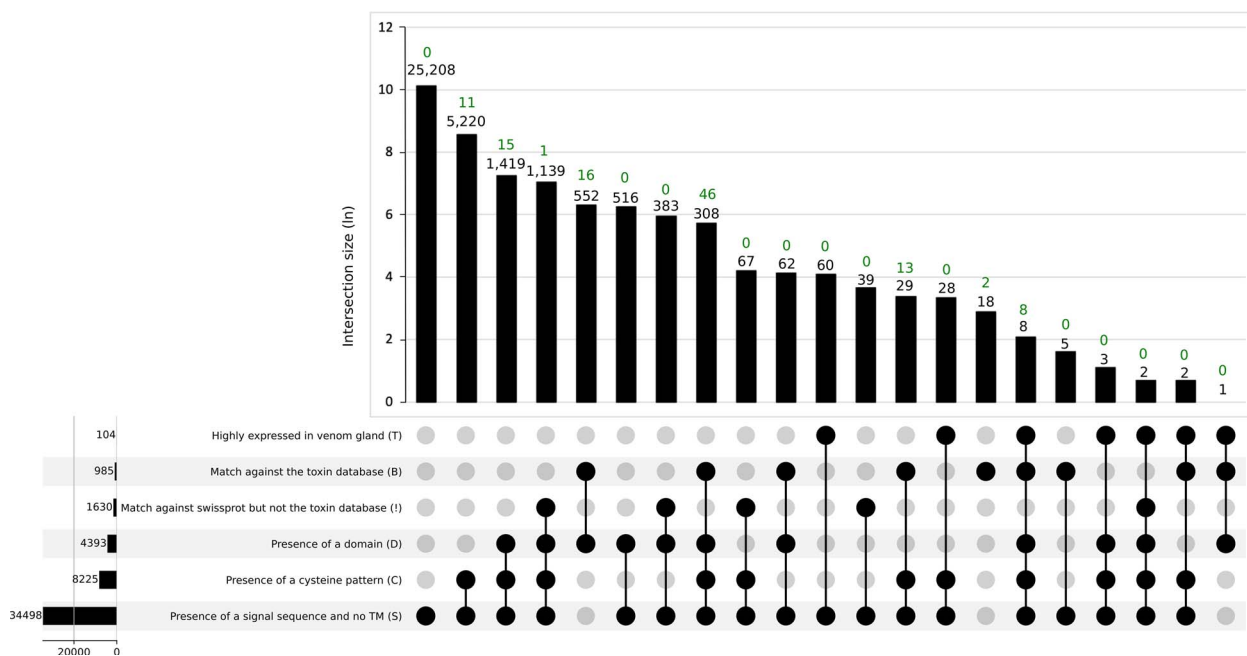


Figure 5. UpsetR diagram for *A. elegantissima*. Distribution (barplot on the upper part) of all the sequences detected by DeTox with the various (combinations of) criteria (dots). Highly expressed putative toxins have a TPM > 1000. Top numbers above each bar represent the sequences also found in the original article.

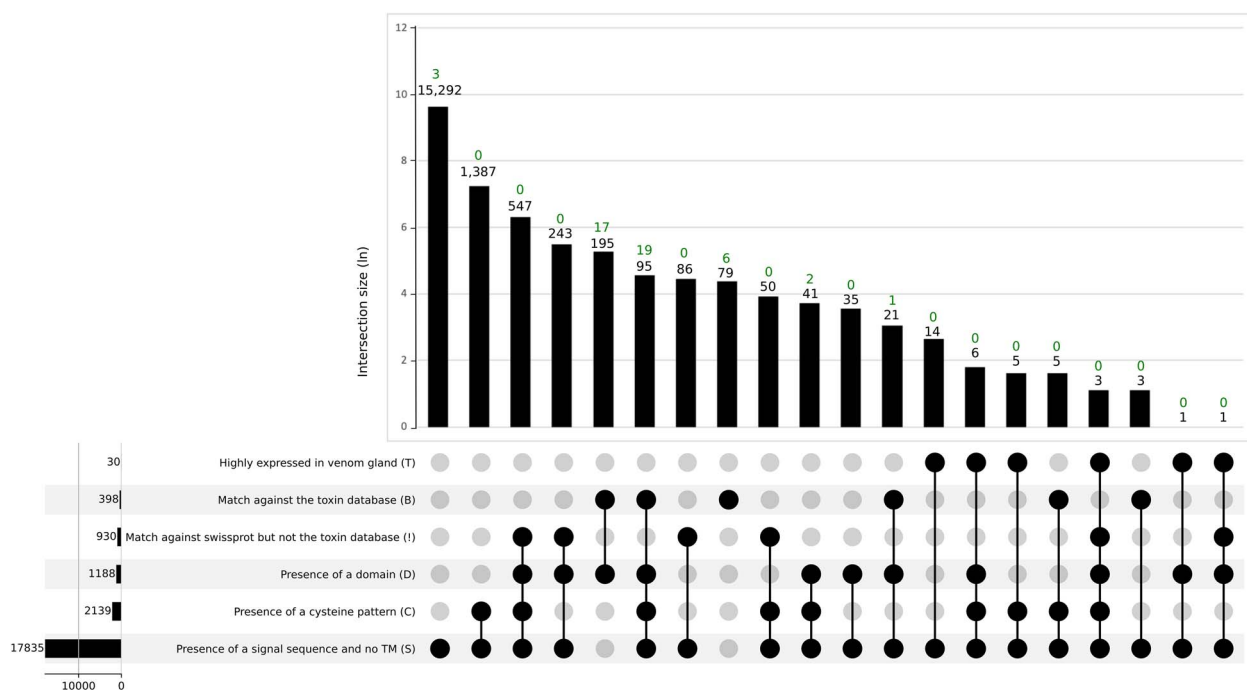


Figure 6. UpsetR diagram for *C. andreae*, *C. cantherigerus* and *T. variabilis*. Distribution (barplot on the upper part) of all the sequences detected by DeTox with the various (combinations of) criteria (dots). Highly expressed putative toxins have a TPM > 1000. Top numbers above each bar represent the sequences also found in the original article.

these limitations, we present DeTox as an integrated tool designed to streamline toxin research. DeTox is user-friendly, requiring minimal bioinformatics skills, and integrates fast methods for *de novo* toxin research, employing the latest versions of each tool. It addresses the time-consuming aspects by parallelizing numerous steps through the use of snakemake [25]. DeTox is modular, allowing customization of parameters to meet specific toxin search requirements. It is accessible (freely distributed),

portable (snakemake), easy to install (Anaconda) and executes quickly. While the pipeline runs without user intervention, the output furnishes a list of toxin candidates based on customizable parameters and detailed information for further exploration. A flag system facilitates result interpretation, helping users gauge the strength of evidence supporting a detected putative toxin.

DeTox stands out from its predecessors also by introducing a structure-based search alongside the conventional similarity

search. This approach incorporates filtering steps to concentrate on potential toxins without known counterparts in existing databases. In its current implementation, DeTox filters translated CDSs based on the presence of a signal sequence and the absence of transmembrane domains—the features prevalent in most known toxins. Additional information, such as the cysteine pattern often found in gastropod toxins, is provided. By combining structure-based and similarity-based approaches, along with optional consideration of gene expression levels, DeTox enhances the likelihood of detecting both toxins with sequences similar to known ones and putative toxins structurally congruent with known toxins but overlooked by similarity-based methods. Moreover, by relying on structural features rather than solely on database similarities, DeTox mitigates biases associated with the over-representation of certain taxa in toxin databases. For example, in the ATDB toxin database, neogastropod toxins are predominantly those from the genus *Conus* [53]. Similarly, the VenomZone database mainly provides toxins from cone snail, represented by 137 species out of the 156 included neogastropod species (<http://venomzone.expasy.org/>). While similarity-based approaches easily detect toxins similar to known cone snail toxins, divergent toxins may go unnoticed. For the same reason, we also refrained from providing automatically created databases. We thus leave it to the user to create their own database, even if it can be quite challenging, ensuring that it contains all the available toxins (from public databases and the literature) for the targeted taxa, and the entries are properly curated—the two steps that are difficult to automatize.

DeTox was tested on previously published transcriptomes from four species of neogastropods, one cnidarian and three snakes. The primary objectives included assessing DeTox's capability to detect toxins reported in the original studies (e.g. a quality control) and identifying new putative toxins not initially found in the original research. In all cases, DeTox successfully detected most of the originally reported toxins. However, some challenges arose because of methodological differences, such as the inability to use the same toxin database or variations in software versions, particularly SignalP. Indeed, discordant results between SignalP versions 4 and 5 have been documented [54]. Notably, the original pipeline for *C. ebraeus* [44] aligned transcripts against the toxin database before identifying ORFs, potentially missing toxins distant from the species under consideration. In contrast, DeTox starts by detecting the longest ORF and then aligns the predicted peptide to the toxin database. Moreover, orfipy is designed to identify the largest ORF within a nucleotide sequence [31], occasionally leading to deduced amino acid sequences with N-terminal regions longer than the encoded protein sequences. This may significantly impact the detection of signal regions, as SignalP is designed to start searching for signal regions at the beginning of a provided sequence. Similar to other tools in previous pipelines, DeTox employs thresholds to determine the presence of structures or infer similarity to toxins in the database. Consequently, the marking of a peptide as a toxin depends on such threshold values, making them a key factor in the entire analysis. DeTox allows users to customize many of these settings, offering flexibility to detect toxins potentially overlooked in the original publications.

Remarkably, DeTox was able to detect a higher number of putative toxins than originally described in all the cases here tested. This result can be explained by the design of DeTox, which minimizes false negatives reducing the probability of missing putative highly divergent toxins. The other face of the coin is that DeTox is prone to report false positives, i.e. peptides that passed all the filters but are not toxins. For *Conus*, *Profundiconus*,

Anthopleura and the snakes (Figures 2, 3 and 5), most of the detected peptides possess a signal sequence, a feature of secreted proteins. However, many of these are likely not toxins. Nevertheless, among them, some show high TPM values in venom-producing tissues, suggesting (together with other lines of evidence) that they might indeed be relevant venom components.

The number of new putative toxins detected thanks to the presence of a signal sequence is also high for *Cumia*, but in this case most of the previously unidentified putative toxins were detected by their similarity with the toxin database (Figures 4 and 6). The use of a more complete toxin database than the original article may also explain the higher number of detected putative toxins. At this step, users should consider the various criteria provided by DeTox's final output. These include putative annotations, gene expression levels and the presence of a cysteine pattern. Additionally, specific information available to the users, such as the tissue type from which the transcriptome was derived, can aid in the identification of true toxins. Indeed, in all cases, DeTox was able to detect putative toxins, not reported in the original articles, matching most—if not all—the available criteria to consider them as toxins (Supplementary Material 2 available online at <http://bib.oxfordjournals.org/>). For example, in *C. ebraeus*, DeTox found a putative toxin with a signal sequence, a VI/VII cys-pattern, a hit (E -value = $3E-10$) with a I1-superfamily conotoxin (Vc11) from *Conus victoriae* and a TPM value of 2155.2. Similarly, in *A. elegantissima*, DeTox detected a new putative toxin with a signal sequence, a match (E -value = $1.09E-44$) with a U-actitoxin (SwissProt ID: P0DMZ3) from *Anemonia viridis* and a TPM of 3289.57. Our results highlight DeTox as a valuable tool to detect novel putative toxins. However, it is always necessary to verify their actual function by confirming the presence of matching proteins in the proteomes of the venom gland or the injected venom [55], and by performing *in vitro/in vivo* experiments to identify the molecular target and the exact function of the toxin [8, 56, 57].

Key Points

- DeTox is a fully integrated pipeline to identify putative toxins in transcriptomic data.
- DeTox is proved efficient in detecting toxins and in identifying secreted peptides displaying the characteristics expected for a toxin candidate for any venomous organism by providing the suitable toxin database.
- DeTox summarizes into a single output file all the information available for each detected putative peptide.
- Its modular nature allows integration of additional criteria of toxin detection in future versions, as well as tools that would facilitate transcriptome comparisons, such as the possibility to cluster similar toxins found in different transcriptomes.

SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

ACKNOWLEDGEMENTS

We would like to thank Rafael Zardoya, José Ramon Pardos-Blas, Samuel Abalde, Giulia Zancolli and Nicholas Casewell for their

help with their original data set we used to test DeTox. Biocomputing were done on the MNHN cluster: Plateforme de Calcul Intensif et Algorithmique PCIA, Muséum national d'histoire naturelle, Centre national de la recherche scientifique, UAR 2700 2AD, CP 26, 57 rue Cuvier, F-75231 Paris Cedex 05, France. We also thank the associate editor and the reviewers for their careful reading and thoughtful comments.

FUNDING

European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 865101) to N.P.

DATA AVAILABILITY

The source codes of this study can be found on GitHub at <https://github.com/Hyperdiverseproject/DeTox>

REFERENCES

- Fry BG, Wroe S, Teeuwisse W, et al. A central role for venom in predation by *Varanus komodoensis* (Komodo Dragon) and the extinct giant *Varanus (Megalania) priscus*. *Proc Natl Acad Sci* 2009;**106**:8969–74.
- Schendel V, Rash LD, Jenner RA, Undheim EAB. The diversity of venom: the importance of behavior and venom system morphology in understanding its ecology and evolution. *Toxins* 2019;**11**:666.
- Ferraz CR, Arrahman A, Xie C, et al. Multifunctional toxins in snake venoms and therapeutic implications: from pain to hemorrhage and necrosis. *Front Ecol Evol* 2019;**7**:218.
- Li R, Bekaert M, Wu L, et al. Transcriptomic analysis of marine gastropod *Hemifusus tuba* provides novel insights into conotoxin genes. *Mar Drugs* 2019;**17**:466.
- Lüddecke T, Herzig V, Von Reumont BM, et al. The biology and evolution of spider venoms. *Biol Rev* 2022;**97**:163–78.
- Casewell NR, Jackson TNW, Laustsen AH, Sunagar K. Causes and consequences of snake venom variation. *Trends Pharmacol Sci* 2020;**41**:570–81.
- de Castro Figueiredo Bordon K, Cologna CT, Fornari-Baldo EC, et al. From animal poisons and venoms to medicines: achievements, challenges and perspectives in drug discovery. *Front Pharmacol* 2020;**11**:1132.
- Verdes A, Anand P, Gorson J, et al. From mollusks to medicine: a Venomics approach for the discovery and characterization of therapeutics from Terebridae peptide toxins. *Toxins* 2016;**8**:117.
- Torres AFC, Huang C, Chong C-M, et al. Transcriptome analysis in venom gland of the predatory Giant ant *Dinoponera quadriceps*: insights into the polypeptide toxin arsenal of hymenopterans. *PLoS One* 2014;**9**:e87556.
- Hwang HJ, Patnaik BB, Chung JM, et al. De novo transcriptome sequencing of triton shell *Charonia lampas sauliae*: identification of genes related to neurotoxins and discovery of genetic markers. *Marine Genomics* 2021;**59**:100862.
- Valente RH, Luna MS, De Oliveira UC, et al. *Bothrops jararaca* accessory venom gland is an ancillary source of toxins to the snake. *J Proteomics* 2018;**177**:137–47.
- Zhao H-Y, Wen L, Miao Y-F, et al. Venom-gland transcriptomic, venom, and antivenomic profiles of the spine-bellied sea snake (*Hydrophis curtus*) from the South China Sea. *BMC Genomics* 2021;**22**:520.
- Li Q, Watkins M, Robinson S, et al. Discovery of novel conotoxin candidates using machine learning. *Toxins* 2018;**10**:503.
- Koua D, Ebou A, Dutertre S. Improved prediction of conopeptide superfamilies with ConoDicator 2.0. *Bioinform Adv* 2021;**1**:vbab011.
- Dao F-Y, Yang H, Su Z-D, et al. Recent advances in conotoxin classification by using machine learning methods. *Molecules* 2017;**22**:1057.
- Cole TJ, Brewer MS. TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ* 2019;**7**:e7200.
- Fedosov A, Zaharias P, Puillandre N. A phylogeny-aware approach reveals unexpected venom components in divergent lineages of cone snails. *Proc R Soc B* 2021;**288**:20211017.
- Fassio G, Modica MV, Mary L, et al. Venom diversity and evolution in the most divergent cone snail genus *Profundiconus*. *Toxins* 2019;**11**:623.
- Kaas Q, Westermann J-C, Craik DJ. Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicol* 2010;**55**:1491–509.
- Watkins M, Hillyard DR, Olivera BM. Genes expressed in a Turrid venom duct: divergence and similarity to conotoxins. *J Mol Evol* 2006;**62**:247–56.
- Yao G, Peng C, Zhu Y, et al. High-throughput identification and analysis of novel conotoxins from three vermivorous cone snails by transcriptome sequencing. *Mar Drugs* 2019;**17**:193.
- Nachtigall PG, Rautsaw RM, Ellsworth SA, et al. ToxCodAn: a new toxin annotator and guide to venom gland transcriptomics. *Brief Bioinform* 2021;**22**:bbab095.
- Macrander J, Panda J, Janies D, et al. Venomix: a simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data. *PeerJ* 2018;**6**:e5361.
- Aili SR, Touchard A, Hayward R, et al. An integrated proteomic and transcriptomic analysis reveals the venom complexity of the bullet ant *Paraponera clavata*. *Toxins* 2020;**12**:324.
- Mölder F, Jablonski KP, Letcher B, et al. Sustainable data analysis with Snakemake. *F1000Res* 2021;**10**:33.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
- Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52.
- Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res* 2012;**41**:D36–42.
- Quast C, Priesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2012;**41**:D590–6.
- Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
- Singh U, Wurtele ES. Orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* 2021;**37**:3019–20.
- Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
- Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019;**37**:420–3.
- Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. *BMC Microbiol* 2005;**5**:58.
- Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004;**338**:1027–36.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.

37. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**:e1002195.
38. Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;**44**:D279–85.
39. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
40. Horton P, Park K-J, Obayashi T, et al. WoLF PSORT: protein localization predictor. *Nucleic Acids Res* 2007;**35**:W585–7.
41. Patro R, Duggal G, Love MI, et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**:417–9.
42. Phuong MA, Mahardika GN, Alfaro ME. Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics* 2016;**17**:401.
43. Abalde S, Tenorio MJ, Afonso CML, Zardoya R. Conotoxin diversity in *Chelyconus ermineus* (born, 1778) and the convergent origin of piscivory in the Atlantic and indo-Pacific cones. *Genome Biol Evol* 2018;**10**:2643–62.
44. Pardos-Blas JR, Tenorio MJ, Galindo JCG, Zardoya R. Comparative venomomics of the cryptic cone snail species *Virroconus ebraeus* and *Virroconus judaeus*. *Mar Drugs* 2022;**20**:149.
45. Modica MV, Lombardo F, Franchini P, Oliverio M. The venomous cocktail of the vampire snail *Colubraria reticulata* (Mollusca, Gastropoda). *BMC Genomics* 2015;**16**:1–21.
46. Macrander J, Brugler MR, Daly M. A RNA-seq approach to identify putative toxins from acrorhagi in aggressive and non-aggressive *Anthopleura elegantissima* polyps. *BMC Genomics* 2015;**16**:221.
47. Domínguez-Pérez D, Durban J, Agüero-Chapin G, et al. The Harderian gland transcriptomes of *Caraiba andreae*, *Cubophis cantherigerus* and *Tretanorhinus variabilis*, three colubroid snakes from Cuba. *Genomics* 2019;**111**:1720–7.
48. Kaas Q, Westermann J-C, Halai R, et al. ConoServer, a database for conopeptide sequences and structures. *Bioinformatics* 2008;**24**:445–6.
49. Cheng T-C, Long R-W, Wu Y-Q, et al. Identification and characterization of toxins in the venom gland of the Chinese bird spider, *Haplopelma hainanum*, by transcriptomic analysis: toxins of Chinese bird spider. *Insect Science* 2016;**23**:487–99.
50. Barassé V, Touchard A, Téné N, et al. The peptide venom composition of the fierce stinging ant *Tetraponera aethiops* (Formicidae: Pseudomyrmecinae). *Toxins* 2019;**11**:732.
51. Bose U, Wang T, Zhao M, et al. Multiomics analysis of the giant triton snail salivary gland, a crown-of-thorns starfish predator. *Sci Rep* 2017;**7**:6000.
52. Pardos-Blas JR, Irisarri I, Abalde S, et al. Conotoxin diversity in the venom gland transcriptome of the Magician's cone, *Pionoconus magus*. *Mar Drugs* 2019;**17**:553.
53. He Q-Y, He Q-Z, Deng X-C, et al. ATDB: a uni-database platform for animal toxins. *Nucleic Acids Res* 2007;**36**:D293–7.
54. Garcion C, Béven L, Foissac X. Comparison of current methods for signal peptide prediction in Phytoplasmas. *Front Microbiol* 2021;**12**:661524.
55. Abalde S, Dutertre S, Zardoya R. A combined transcriptomics and proteomics approach reveals the differences in the predatory and defensive venoms of the molluscivorous cone snail *cylinder ammiralis* (*Caenogastropoda*: Conidae). *Toxins* 2021;**13**:642.
56. Eriksson A, Anand P, Gorson J, et al. Using *Drosophila* behavioral assays to characterize terebrid venom-peptide bioactivity. *Sci Rep* 2018;**8**:15276.
57. Moon J, Gorson J, Wright M, et al. Characterization and recombinant expression of terebrid venom peptide from *Terebra guttata*. *Toxins* 2016;**8**:63.