Venomix: A simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data

Jason Macrander Corresp., [1, 2] , **Jyothirmayi Panda** [3] , **Daniel Janies** [3, 4] , **Marymegan Daly** [2] , **Adam M Reitzel** [1]

[1] Department of Biological Sciences, University of North Carolina at Charlotte, Charlotte, North Carolina, United States
[2] Department of Evolution, Ecology, and Organismal Biology, Ohio State University, Columbus, Ohio, United States
[3] College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, North Carolina, United States
[4] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina, United States

Corresponding Author: Jason Macrander
Email address: jmacrand@uncc.edu

The advent of next-generation sequencing has resulted in transcriptome-based approaches to investigate functionally significant biological components in a variety of non-model organism. This has resulted in the area of “venomics”: a rapidly growing field using combined transcriptomic and proteomic datasets to characterize toxin diversity in a variety of venomous taxa. Ultimately, the transcriptomic portion of these analyses follows very similar pathways after transcriptome assembly: candidate toxin identification using BLAST, expression level screening, protein sequence alignment, gene tree reconstruction, and characterization of potential toxin function. Here we describe the python package Venomix, which streamlines these processes using commonly used bioinformatic tools along with a public, annotated database comprised of characterized venom proteins. In this study, we use the Venomix pipeline to characterize candidate venom diversity in four phylogenetically distinct organisms, a cone snail (Conidae; *Conus sponsalis*), a snake (Viperidae; *Echis coloratus*), an ant (Formicidae; *Tetramorium bicarinatum*), and a scorpion (Scorpionidae; *Urodacus yaschenkoi*). Data on these organisms was sampled from public databases and thus different approaches to either transcriptome assembly, toxin identification, or gene expression quantification was used for each. Of the organisms used in our analysis, Venomix recovered numerically more candidate toxin transcripts for three of the four transcriptomes than the original analyses. In four of four organisms we identified new toxin candidates that were not reported in the original analysis. In summary, we show that the Venomix package is a useful tool to identify and characterize the diversity of toxin-like transcripts. Venomix is available at: https://bitbucket.org/JasonMacrander/Venomix/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26733v1 | CC BY 4.0 Open Access | rec: 19 Mar 2018, publ: 19 Mar 2018

Venomix: A Simple Bioinformatic Pipeline for Identifying and Characterizing Toxin Gene Candidates from Transcriptomic Data.

Jason Macrander[1,2], Jyothirmayi Panda[3], Daniel Janies[3,4], Marymegan Daly[2], Adam M. Reitzel[1] 1. University of North Carolina at Charlotte, Department of Biological Sciences, 9201 University City Blvd., 373 Woodward Hall, Charlotte, NC-28223, USA.

2. The Ohio State University, Department of Evolution, Ecology, and Organismal Biology, 318 W. 12th Ave. 300 Aronoff Laboratory, Columbus, OH-43215, USA.

3. University of North Carolina at Charlotte, College of Computing and Informatics, 9201 University City Blvd., Charlotte, NC-28223, USA.

4. University of North Carolina at Charlotte, Department of Bioinformatics and Genomics, 9201 University City Blvd., Charlotte, NC-28223, USA.

Corresponding Author

Jason Macrander[1,2]

E-mail address: jmacrand@uncc.edu

Abstract

The advent of next-generation sequencing has resulted in transcriptome-based approaches to investigate functionally significant biological components in a variety of non-model organism. This has resulted in the area of “venomics”: a rapidly growing field using combined transcriptomic and proteomic datasets to characterize toxin diversity in a variety of venomous taxa. Ultimately, the transcriptomic portion of these analyses follows very similar pathways after transcriptome assembly: candidate toxin identification using BLAST, expression level screening, protein sequence alignment, gene tree reconstruction, and characterization of potential toxin function. Here we describe the python package Venomix, which streamlines these processes using commonly used bioinformatic tools along with a public, annotated database comprised of characterized venom proteins. In this study, we use the Venomix pipeline to characterize candidate venom diversity in four phylogenetically distinct organisms, a cone snail (Conidae; *Conus sponsalis*), a snake (Viperidae; *Echis coloratus*), an ant (Formicidae; *Tetramorium bicarinatum*), and a scorpion (Scorpionidae; *Urodacus yaschenkoi*). Data on these organisms was sampled from public databases and thus different approaches to either transcriptome assembly, toxin identification, or gene expression quantification was used for each. Of the organisms used in our analysis, Venomix recovered numerically more candidate toxin transcripts for three of the four transcriptomes than the original analyses. In four of four organisms we identified new toxin candidates that were not reported in the original analysis. In summary, we show that the Venomix package is a useful tool to identify and characterize the diversity of toxin-like transcripts. Venomix is available at: https://bitbucket.org/JasonMacrander/Venomix/

Keywords: Venom, Transcriptome, Python, Transdecoder, SignalP, Protein

1. Introduction

Throughout the animal kingdom, venom has evolved independently multiple times to be used in prey capture, predatory defense, and intraspecific competition (Casewell et al., 2013). Venoms are toxic cocktails with remarkable diversity in protein action and specificity across animals. The evolutionary and ecological processes shaping this diversity are of major interest (Fry et al., 2009; Wong & Belov, 2012; Casewell, Huttley & Wüster, 2012; Sunagar et al., 2016; Rodríguez de la Vega & Giraud, 2016), with much of this focusing on characterizing protein and RNA composition expressed in the venom gland (Ménez, Stöcklin & Mebs, 2006). As sequencing costs decrease and assembly programs are becoming more efficient, the number of venom-focused studies is increasing at a dramatic rate. For some of the better studied venomous lineages (*e.g.* Colubroidea), comparative transcriptome and genome sequencing are being used to investigate processes involved with toxin gene recruitment and tissue specific gene expression (Vonk et al., 2013; Hargreaves et al., 2014a; Reyes-Velasco et al., 2015; Junqueira-deAzevedo et al., 2015). For many poorly studied taxonomic lineages (*e.g.* Cnidaria), similar techniques are being used to evaluate venom diversity using bioinformatic pipelines for a particular species or taxonomic group (Tan, Khan & Brusic, 2003; Reumont et al., 2014; Macrander, Brugler & Daly, 2015; Kaas & Craik, 2015; Prashanth & Lewis, 2015). Although these studies all take similar approaches to study diverse venoms across animal lineages, a streamlined systematic pipeline does not exist for rapid identification of candidate toxin genes from these datasets.

Bioinformatic tools that use transcriptomic, proteomic, and genomic data sets have emerged for a variety venomous taxa. Among these, programs founded in machine learning appear to be the most abundant tools currently available; these use a combination of lineage specific annotation datasets (Kaplan, Morpurgo & Linial, 2007; Fan et al., 2011; Wong et al., 2013) and identifiers based on residue frequency and protein domains of interest (Gupta et al., 2013). Although taxonomic and tissue specific application of these programs vary, the pipelines follow a similar mechanical path. Typically starting with millions of raw reads assembled *de novo* using Trinity (Grabherr et al., 2011) or similar. Next, expression values for each transcript calculated using RSEM (Li & Dewey, 2011) or similar. Ultimately with the resulting transcriptome assembly searched by some component of BLAST (Camacho et al., 2009; Neumann, Kumar & Shalchian-Tabrizi, 2014) or other motif-searching algorithms (Kozlov & Grishin, 2011). For many of the pipelines, these outputs are then screened using custom query datasets comprised of lineage specific toxin genes or the manually curated ToxProt dataset (Jungo et al., 2012), which includes all characterized/annotated animal venom proteins. Following candidate toxin gene identification, downstream analyses often involve predicting open-reading frames using Transdecoder (Haas et al., 2013), in combination with signal region prediction using SignalP (Petersen et al., 2011). These types of data are sometimes complemented with genome and proteome datasets (see Sunagar et al., 2016). However, the majority of studies that are exploratory in nature use transcriptomic approaches to describe overall toxin diversity for a variety of poorly studied taxa (Reumont et al., 2014; Macrander, Brugler & Daly, 2015; Barghi et al., 2015; Luna-Ramírez et al., 2015; Macrander, Broe & Daly, 2016; Lewis Ames et al., 2016). One major drawback to this approach, and using these self-constructed pipelines, is that downstream analyses begin to become quite cumbersome when trying to identify and characterize multiple toxin gene families for diverse toxin genes.

Here were present Venomix, a bioinformatic pipeline written in the programming languages Python and R that follows generally accepted methods for identifying and characterizing toxin-like genes from transcriptomic datasets. This is a single, easy-to-use bioinformatic pipeline that will screen transcriptomic datasets of diverse taxa for toxin-like genes. Venomix incorporates widely used programs into its pipeline, like BLAST (Camacho et al., 2009) for initial toxin-like transcript identification, Transdecoder (Haas et al., 2013) to translate transcripts into their proper reading frame, SignalP (Petersen et al., 2011) to predict toxin gene signaling regions, MAFFT (Katoh & Standley, 2013) for protein sequence alignment, and the R package APE (Paradis, Claude & Strimmer, 2004) to construct gene trees. Candidate toxin genes are grouped based on sequence similarity, with each directory corresponding to a specific toxin group based on the ToxProt sequence names (e.g., some variation of conotoxin, Kunitz-type serine protease, phospholipase A2, zinc metalloproteinase). The Venomix pipeline provides the user with several output files that can be used to characterize the potential function of these candidate toxins, compare relevant expression level values across toxin-gene candidates, evaluate amino acid conservation among functionally important residues in sequence alignments, and review taxonomic and functional information in combination with tree reconstructions to further evaluate toxin gene candidates.

In this study, we use Venomix to characterize the toxin-like diversity from venom gland transcriptomes for a cone snail (*Conus sponsalis*), a snake (*Echis coloratus*), an ant (*Tetramorium bicarinatum*), and a scorpion (*Urodacus yaschenkoi*), using the Venomix bioinformatic pipeline. Although Venomix alone is not designed to be used to as a definitive validation pipeline for toxin genes, it can quickly identify, sort, and characterize transcripts that may be used to further evaluate these venom candidates. By abating time required by these processes, researchers are freed to focus on other aspects of toxin gene identification, such functionally characterizing toxin genes from their transcriptome or some of the lineage specific tools to better understand venom diversity present in the transcriptome.

2. Materials and Methods

2.1 Data Acquisition and Transcriptome Assembly

Raw reads from four different analyses were downloaded from the short read archive (SRA) on GenBank (*C. sponsalis*: SRR260951 (Phuong, Mahardika & Alfaro, 2016); *U. yaschenkoi*: SRR1557168 (Luna-Ramírez et al., 2015); *E. colaratus*: ERR216311 – ERR216312 (Hargreaves et al., 2014b); *T. bicarinatum*: SRR1106144 - SRR1106145 (Bouzid et al., 2014)). The previously published transcriptome level analysis for *U. yaschenkoi* and *T. bicarinatum* were restricted to just characterizing the venom gland transcriptome in their respective species (Bouzid et al., 2014; Luna-Ramírez et al., 2015), while the *C. sponsalis* and *E. coloratus* venom gland transcriptomes were investigated in a comparative framework alongside closely related taxa (Hargreaves et al., 2014b; Phuong, Mahardika & Alfaro, 2016). All four transcriptomes were assembled in Trinity (Grabherr et al., 2011; Haas et al., 2013), using default parameters of its built-in Trimmomatic program to clean up the sequences (Bolger, Lohse & Usadel, 2014). For each transcriptome, expression values (TPM and FKPM) were calculated in the program RSEM (Li & Dewey, 2011) as part of the Trinity package.

2.2 Analysis Pipeline and Execution

The bioinformatic pipeline for Venomix is outlined in Figure 1. The program requires three inputs provided by the user: an assembled transcriptome, gene expression information in the form of a tab delimited output with transcript names in the first column, and tab delimited BLAST output using the ToxProt as query sequences. Following transcriptome assembly and expression level calculations, the final user provided file is created using tBLASTn from NCBI BLAST+ version 2.4.0 (Camacho et al., 2009), with the ToxProt dataset as the search query with the final BLAST alignment results shown in a tabular format (`-outfmt 6`). Query sequences from ToxProt are provided within the Venomix package, however, alternative curations of the ToxProt dataset may be used if the sequence identifiers are not changed. In our analysis, we implemented two BLAST search procedures; the first used a more stringent identification algorithm (E-value 1e-20) and a less stringent identification algorithm (E-value 1e-6).

For these tests, Venomix was run on the University of North Carolina at Charlotte COPPERHEAD Research Computing Cluster with 944 Computing Cores, 8 TBs of memory, and 100 Gb/s Infiniband connectivity. The implementation of Venomix requires the scripting languages Python 2.7 (http://www.python.org/download/releases/2.7/) and R 3.1.1 (https://cran.r-project.org/bin/), in addition to other Biopython packages (Cock et al., 2009) and data from ToxProt and Genbank that are built into the Venomix pipeline (https://bitbucket.org/JasonMacrander/Venomix). We included versions of MAFFT (Katoh & Standley, 2013), NCBI BLAST+ (Camacho et al., 2009), and Transdecoder (Haas et al., 2013) that can be run locally. Although there are two versions of MAFFT (64 bit and 32 bit), the default is the 64-bit, as this is more common for computers used in bioinformatic analyses. Modification to the version of MAFFT in the Venomix pipeline can be done in the support_files/Alignment.py file. Once the user specifies the input files (Transcriptome, Expression file, and BLAST output), the Venomix pipeline automatically produces several potentially useful and informative files within each of the toxin group directories (Figure 1). The outputs within each of the Toxin Group directories are intended to provide the user with curated information to focus future investigations and analyses. Depending on the next step of the analysis, some of the output files may be used for additional venom related downstream applications or simply a quick reference for the user (see Discussion). Venomix also creates two ancillary products that may be informative to some users: `TPM.fasta` (only transcripts with TPM values > 1.0) and a large GenBank file with information from ToxProt BLAST hits in a format that may lend itself to quick searches or downstream annotation (Figure 1). The user may choose to rerun Venomix with `TPM.fasta` instead of their assembled transcriptome if they would like to characterize only transcripts with a TPM >1.0, but it is not recommended when looking for rare or extremely diverse toxin genes.

2.2 Venomix Evaluation

For each assembled transcriptome, we identified candidate toxin genes using the Venomix pipeline using a stringent (E-value 1e-20) and less stringent (E-value 1e-6) search strategy in BLAST. Venomix outputs were compared for both search parameters in terms of the number of toxin groups, number of transcripts, and number of "candidate" transcripts identified by the pipeline. A transcript was considered a "candidate" if the transcript had significantly better e-value associated with a toxin than with a non-toxin protein in Uniprot. Candidate transcripts were translated into their protein sequences using Transdecoder (Haas et al., 2013) and further evaluated in ToxClassifier (Gacesa, Barlow & Long, 2016). If a protein sequence received a score of > 1according to ToxClassifier, it was retained as a toxin candidate. In addition to screening the overall transcriptome analyses, some toxin groups and candidate toxins identified in our analysis were subjected to additional screening beyond what is included in the Venomix pipeline. Sequence alignments for candidate transcripts shown below were done using MAFFT (v.1.3.3) (Katoh & Standley, 2013) and visualized in Geneious (Kearse et al., 2012). Toxin gene tree reconstructions were done in Fasttree v2 (Price, Dehal & Arkin, 2010) using maximum likelihood tree reconstruction methods and bootstrap supports calculated over 1000 replicates. For the Bouzid et al. (2014) dataset, Venomix was used to compare alternative assembly approaches (Oases/Velvet vs. Trinity), in addition to assessing both transcriptomes for overall completeness in BUSCO (Simão et al., 2015). Expression values for each transcriptome were calculated using RSEM (Li & Dewey, 2011) rather than raw read counts as originally proposed by Bouzid et al (Bouzid et al., 2014).

3. Results

3.1 Transcriptome Assemblies

Transcriptomes for each species previously assembled in Trinity (Grabherr et al., 2011) resulted in a similar assembly parameter in Venomix (Table 1), with the only notable difference being in the number of transcripts for *C. sponsalis,* which may be due to repetitiveness and sequence complexity encountered during their initial assemblies (Phuong, Mahardika & Alfaro, 2016). The transcriptome for *T. bicarinatum* was originally done using Velvet/Oases (Li & Durbin, 2009), however, we compared this to our Trinity assembly because of its ease of use (Sanders et al., 2018) and frequency in the venom literature (Macrander, Broe & Daly, 2015), in addition to a lower redundancy and chimera rate (Yang & Smith, 2013).

3.2 Pipeline Output

In the original published annotations, species-specific transcriptomes of *C. sponsalis*, *E. colaratus* and *U. yaschenkoi* were not subjected to any BLAST searches using ToxProt, but instead were screened using taxonomic specific toxin datasets from venom proteins of closely related species. The Venomix pipeline recovered the majority of these lineage specific toxins and additional transcripts that resemble toxin genes from other taxa. It is worth noting that if there were lineage specific toxins that shared high sequence similarity to other toxins, the toxin group name may be assigned an incorrect lineage classification, yet remain a toxin candidate. For example, analyses of the scorpion *U. yaschenkoi* resulted in four venom groups with “Snake venom” in the name, however, in these instances, lineage specific toxin names are often members of larger gene families that may not be lineage specific. The number of identified toxin groups varied considerably across species and stringency parameters (Table 2), with the less stringent parameters dramatically increasing the number of toxin groups. Each species had multiple toxin groups that were separated based on sequence similarity and that correspond to large gene families, including astacin-like metalloproteases, conotoxins, phospholipases A2s, Cysteine-rich secretory proteins (CRISPs), Kunitz-type serine protease inhibitors, snaclecs, metalloproteinases, thrombin-like enzymes, and allergens. For each species, there was at least one toxin group that was not retained following the reciprocal BLASTp search after the toxin-like transcripts were translated into their open reading frame (Table 2).

3.3 Venomix outputs for *C. sponsalis*

Collectively, conotoxins represent some of the best- studied toxin genes across the genus *Conus*, comprising of multiple gene families with cysteine rich proteins (Buczek, Bulaj & Olivera, 2005; Kaas et al., 2012). Our less stringent analysis identified 76 toxin groups comprising of 246 toxin gene candidates based on our preliminary low stringency BLAST survey; 20 of these groups cluster with various conotoxins (Table 2). In total, there were 179 of the 246 toxin gene candidates from the 20 conotoxin groups. The largest number of candidate toxin genes were associated with the conotoxin O1 superfamily (n = 105), which was also the most abundant conotoxin identified in the original transcriptome (Phuong, Mahardika & Alfaro, 2016) (Table 3). Among the remaining conotoxins, superfamily M was the second largest (n = 29), followed by conoodipine and conophysin. There were three instances where Venomix recovered more candidate conotoxins than the original study (O1, conodipine, and conophysin superfamilies), however, for the majority of the conotoxin superfamilies identified in the original study (Phuong, Mahardika & Alfaro, 2016) were missing from our analysis (Table 3). This discrepancy is likely due to the different approaches used in the initial transcriptome assemblies, as iterative assemblies used by Phuong et al. (2016) were unable to recover known transcripts using Trinity alone. Beyond the conotoxins, there were many candidate toxin genes found within the Kunitz-type conkuitzin-S1 group (n = 19), which included characterized toxin proteins from the venom Kunitz-type family of sea anemones, cone snails, and snakes.

3.4 Venomix outputs for *E. coloratus*

In total, in our low stringency search (E-value = 1E-6) Venomix identified 132 "Toxin Groups" for *E. coloratus*, most of which can be combined and placed within groups outlined by Hargreaves et al. (2014b). Among the transcripts identified, 45 had a TPM value greater than 100, with 39 of these in the venom gland, four in the scent gland, and two in the skin. The most abundant transcript was actually in the scent gland and identified as a cathelicidin-related peptide (Supplemental Table 1). The majority of the highly expressed transcripts in the venom gland (TPM > 100) corresponded with toxin groups previously identified (Hargreaves et al., 2014b), comprising mostly of C-type lectins, cysteine rich venom proteins, disintegrins, metalloproteinases, and several others.

In addition to these venom candidates, we found one cystatin highly expressed in the venom gland, although it was also highly expressed in other tissues and not thought to be a toxic component of the *E. coloratus* venom (Hargreaves et al., 2014b). We also identified a single peroxiredoxin, which may play a role in the structural/functional diversification of toxins (Calvete et al., 2009). Additionally, Venomix recovered a single ficolin, which is involved in platelet aggregation and/or coagulation (OmPraba et al., 2010), a prothrombin-like activator, which may really be a complement factor B-like protein based on reciprocal BLAST search and have no known function, and three transcripts expressed at high levels from the latroinsectotoxin gene family, which was previously characterized in spiders (Magazanik et al., 1992). Reciprocal BLAST searches against the entire UNIPROT dataset revealed that two of these were ankyrin rich peptides, and the other a dysferlin–like protein and may not be toxins.

3.5 Venomix outputs for *T. bicarinatum*

Of the four datasets included in our analysis, *T. bicarinatum* was the only transcriptome which originally used the ToxProt dataset for toxin identification. The original transcriptome assembly was done in Velvet/Oases (Li & Durbin, 2009), resulting in very different transcriptome assembly outputs (Bouzid et al., 2014). Despite the alternative approaches, the original assembly resulted in a higher Busco (Simão et al., 2015) score for the Velvet/Oases assembly (95.9%) when compared to our Trinity assembly (92.2%). Interestingly, when considering TPM (rather than raw counts) the number of candidate genes in the venom gland following the approach by Bouzid et al. (2014) was similar to what was originally published (Table 2). Among these 527 candidates, only 44 predicted ORFs from Transdecoder, and only three of these were given a score of 1 or greater in Toxclassifier (Table 2). The BLAST screening, however, resulted in 62 of candidate toxins identified when the E-value threshold was set to 1E-3, but 287 when the E-value threshold was set to 10. As E-values were not specified by Bouzid et al. (Bouzid et al., 2014) both are reported here.

For the Venomix analysis or our Trinity assemblies, there were 75 and 36 toxin groups for the less stringent (E-value = E-06) and more stringent (E-value = E-20) analyses, respectively. In the less stringent analysis, the largest number candidate toxin genes corresponded to the alpha-latroinsectotoxinLt1a group (N = 280), but overall expression within the ant carcass and venom gland across these transcripts were approximately the same in both the ant carcass and venom gland. Among those more highly expressed in the venom gland, there were six transcripts expressed at TPM values greater than 100 in the venom gland, four corresponding to venom allergen 3 and two to cysteine-rich venom protein Mr30. Upon closer examination, BLAST searches against UNIPROT indicated that all six of these toxins are likely Venom Allergen 3 toxins, indicating that these are likely the most abundant venom toxins in the *T. bicarinatum* transcriptome, with these six toxins making up ~ 92% of the cumulative gene expression across the transcripts identified as potential toxins in Venomix.

When we compared the Venomix outputs for our Trinity assembly to the Velvet/Oases assembly that was previously published by Bouzid et al (2014), we recovered some unexpected results. Although we used RSEM instead of BWA, of the original 69 candidate toxin sequences recovered from their analysis only 33 had higher TPM expression in the venom gland than in the ant carcass. Additionally, of these 33, only 10 were ‘1000 fold’ higher based on expected count values (Supplemental Table 2). Although not all 69 candidate toxin sequences identified by Bouzid et al (Bouzid et al., 2014) were confidently called toxins, all but one was reported as having a higher number of raw read counts in the venom gland transcriptome as opposed to the ant carcass. The reason behind the larger than 50% discrepancy between our analyses and theirs remains unknown, as alternative approaches to quantifying gene expression should exhibit some proportional correlations across the two sample types. Among the toxin-like sequences identified in Venomix, Thrombin-like enzymes recovered 8688 transcripts from the Velvet/Oases assembly with a TPM difference of 100 or greater in the venom gland transcriptome when compared to the ant carcass. After further examination, 33 of the original 69 transcripts were also grouped in this Thrombin-like group (Supplemental Table 2). These transcripts had an average length of 704 nucleotides, while the other 8688 transcripts identified in Venomix had an average length of 291 nucleotides. It is likely that incomplete transcripts for other proteins were recovered here as protein sequences were 407 amino acids (1221 nucleotides) long, which corresponded to the identified transcripts in our analysis. Incomplete transcripts are likely the reason behind this as only 47 of the 8688 transcripts identified in Venomix were translated into their open reading frame.

3.6 Venomix outputs for *U. yaschenkoi*

The original analysis for this species used scorpion-specific toxins as query sequences and identified 210 transcripts representing 111 unique scorpion toxins, venom gland enzymes, and antimicrobial peptides (Luna-Ramírez et al., 2015). By expanding the query sequences with the ToxProt dataset, we recovered 117 toxin groups representing 689 unique transcripts for the less stringent search

(Table 2). Within the Toxprot dataset, there were only 10 of the 117 toxin groups with scorpion derived query sequences. Of these 10, the "Toxin-like protein 14" were the same number as the original investigation. One notable difference, however, is that Venomix recovered the complete protein sequence, whereas the original investigation by Luna-Ramírez et al. (Luna-Ramírez et al., 2015) did not (Figure 2A). These differences may be due to changes in assembly algorithms between the two analyses, as Venomix has no input on the initial assembly parameters. When using exclusively scorpion venom proteins from ToxProt as query sequences, the number of candidate toxins identified by Venomix was approximately the same as that identified by Luna-Ramírez et al. (2015). Beyond the exclusive scorpion query sequences in ToxProt, Luna-Ramírez et al. (2015) also identified several phospholipases and potassium channel toxins, which were also recovered in our analysis, although a more thorough characterization is needed for these candidate toxins as they may also be non-venomous in nature.

The most abundant toxin-like transcripts within the less stringent search criteria were found within the delta-latroinsectotoxin-Lt1a group (n = 190), the alpha latroinsectotoxin-Lt1a group (n = 182), and the Neprilysin-1 group (n = 91). Under the more stringent search criteria, these are still the largest toxin groups, however, the most abundant group was Neprilysin-1 group (n = 69), followed by alpha latroinsectotoxin-Lt1a group (n = 51) and delta-latroinsectotoxin-Lt1a group (n = 49). Query toxins which are used to form these toxin groups were previously identified in spiders (Graudins et al., 2012; Garb & Hayashi, 2013; Undheim et al., 2013; Bhere et al., 2014), and not included in the original transcriptome analysis (Luna-Ramírez et al., 2015). Preliminary screening based on comparative toxin groups indicated that the latroinsectotoxin groups identified in Venomix may not be toxins, as other non-toxin proteins in UNIPROT had better E-values than the genes used as queries in Uniprot. Among the neprilysin candidates, however, seven had better E-values from matching within the ToxProt dataset. Maximum likelihood gene tree reconstructions were used as post-processing steps to further screen these potential toxin sequences. Proteins from the less stringent analysis (E-value >1E-6) were used to construct gene trees for neprilysins using proteins of known venomous function along aside other proteins that are nonvenomous in origin and from the same gene family. The subsequent toxin gene tree for the search revealed that candidate toxins from the Neprilysin-1 group formed a well-supported cluster with neprilysin toxins from other scorpions at high expression levels (Figure 2B).

4. Discussion

Venomix presents a less cumbersome, non-taxon specific alternative to any other pipeline currently being implemented in venom research. The pipeline allows the user to quickly identify and characterize toxin gene candidates within a transcriptomic dataset. The outputs provided by this pipeline give necessary information for further evaluation of their toxin gene candidates. We recommend using Venomix across multiple BLAST searches with varying E-value thresholds, as the variation among characterized toxin genes and those of the focal taxa may be more accommodating depending on the threshold used. Although Venomix was able to identify more candidate toxin genes in three out of the four datasets tested here, these results require further examination to determine which transcripts are viable toxin gene candidates. Venomix is not meant to be a definitive toxin gene identifier because this determination should not be made by sequence data alone, especially for poorly studied lineages.

We chose four very different taxa to highlight some of the benefits and limitations of Venomix. Of the taxa used in this study, three of them are from taxonomic groups with ample representation in the ToxProt dataset (Figure 3), whereas the ant venom is poorly characterized among the diverse venomous insects found within Hexapoda on ToxProt. Additionally, these datasets represent diverse transcriptome assembly methods, query datasets, and gene expression quantification approaches. The original *C. sponsalis* assembly had a high number of toxin genes with relatively low variation across gene copies (Phuong, Mahardika & Alfaro, 2016), which likely resulted in many of these being clumped together in our Trinity assembly (Macrander, Broe & Daly, 2015). To get around this issue, Phuong et al (Phuong, Mahardika & Alfaro, 2016) did three assembly iterations involving toxin gene identification and subsequent mapping, in addition to downstream analysis incorporating the Assembly by Reduced Complexity pipeline (https://github.com/ibest/ARC) and manual alignments in Geneious (Biomatters, Auckland, New Zealand). In contrast to this, it is likely the differences we observed for *T. bicarinatum* using Venomix was due to the alternative transcriptome assembly and gene expression approach (Yang & Smith, 2013; Vijay et al., 2013; Todd, Black & Gemmell, 2016). Additionally, limiting query sequences to only venoms of that lineage – which was done with the *C. sponsalis*, *E. coloratus*, and *U. yaschenkoi,* but not for *T. bicarinatum* – likely limited the number of toxin candidates being identified.

The Venomix pipeline was designed to sidestep much of the rigorous analysis used to identify and extract candidate toxin sequences. Specifically, our pipeline also will translate transcripts into their predicted protein, screen for signaling regions, assess their similarity through alignment and gene trees, extract expression information, and refer to taxonomic and other information available in the query sequence GenBank entries. This will allow venom biologists to quickly move onto additional downstream identification and characterization of toxin gene diversity using outputs provided by Venomix. Additionally, Venomix is the first pipeline to provide all these outputs in an easy to use search strategy that is flexible, but also repeatable, for all venomous taxa, or non-venomous animal to be used in a tissue specific comparative context (Reumont et al., 2014; Hargreaves et al., 2014b; Reyes-Velasco et al., 2015). The pipeline also provides users with easy to navigate directories and organized output files, allowing the user to sort manually or quickly pull information for all toxin groups using simple unix commands (i.e., grep) as the files within each toxin group directory have the same name.

Venomix can facilitate the process of determining what constitutes a venom protein and aid in testing future hypotheses of venom diversity and tissue specific expression. The *E. coloratus* transcriptome used in our analysis was part of a broader study, to test the early evolution of venom in reptiles, the Toxicofera hypothesis (Hargreaves et al., 2014b). They used tissue specific expression in combination with toxin gene tree reconstruction to determine which of the approximately 16 venom toxin gene families that occur across Toxicofera attribute to the *E. coloratus* venom transcriptome. Of these which are venom candidates in *E. coloratus* a comparison of the Venomix output containing expression information would identify toxin candidates with ease (Table 4). Conversely, there are also transcripts highly expressed in the venom gland that are likely not venomous (Terrat & Ducancel, 2013). This was made evident in the *U. yaschenkoi* analysis, as several transcripts within the latroinsectotoxins cluster were actually neprilysins in high abundance, but transcripts resembling neprilysins matched to other neprilysin toxins in a reciprocal blast hit.

Regardless of the bioinformatic approach to identifying toxin genes, one major hurdle using these types of datasets as query sequences is the limited taxonomic diversity present in the ToxProt dataset. Although the transcriptome for *U. yaschenkoi* was larger and had a longer N50 than that of *E. coloratus* (Table 1), there were more toxin-like transcripts identified in the *E. coloratus* transcriptome. This likely reflects the abundance of snake proteins deposited into ToxProt and is in contrast to the paucity of proteins for other, poorly studied venomous lineages (Figure 3). Additionally, Venomix "group" names should be examined closely because some candidate toxin genes were labeled with lineage-specific proteins. For example, our analysis recovered a group called conophysin (a cone snail toxin) for *T. bicarinatum*, however, the transcripts associated with this appeared to be neurophysins. This was also observed when Venomix failed to group Venom Allergen 3 and Cysteine-rich venom protein Mr30 groups together for *T. bicarinatum* even though it was apparent that the most highly expressed were all Venom Allergen 3 genes. When investigating venom diversity for poorly studied taxa, caution is warranted in using these gene names because the specific classifiers of the Venomix outputs provide a starting point for toxin gene identification, but does not act as a distinct classification system.

In every transcriptome, the machine learning program ToxClassifier failed to recover all of the toxins identified in their respective publications (Table 2). Our downstream analysis of the protein sequences produced by Transdecoder included any candidate toxin with a score > 1. Even though some datasets were close (Table 2), ToxClassifier considers a "potential toxin" (> 4), meaning that the number of "toxins" for *C. sponsalis* drops to 243 and *E. coloratus* to 7. Despite this, one major contrast between ToxClassifier and Venomix is that our pipeline is not meant to be a toxin gene identifier. Venomix was designed to be useful for preliminary searches for users new to the command line, or provide a platform that is adaptable for those that are well versed in the command line. The incorporated alignment and tree building methods are rudimentary and meant to be used for only initial screenings. This allows users to focus their efforts on downstream analyses using complementary proteomics and machine learning to differentiate between functionally toxic and non-toxic venom components (Gacesa, Barlow & Long, 2016) or to complement their transcriptomic data with functional assays of proteins or crude venom extracts.

5. Conclusion

The advent of next-generation sequencing has allowed for a large influx of comparative transcriptomic datasets to identify toxin gene candidates in a variety of taxa. Our Venomix pipeline is a versatile in that it can accommodate transcriptomic datasets for a variety of species and can quickly identify a large number of toxin gene candidates from venom gland or other tissue specific transcriptomes. Overall, Venomix addresses three shortcomings encountered in similar approaches: (1) it is reproducible, (2) it does not claim to be a toxin gene identifiers as other programs or pipelines do that appear to be less reliable, and (3) it is able to accommodate a wide variety of taxa. Because of its ease of use and ability to quickly identify toxin gene candidates, researchers can move past the tedious and time consuming stages of toxin candidate identification and move onto toxin gene characterization.

Acknowledgments

Venomix is a direct byproduct of the training received at Friday Harbor Marine Labs while attending the Practical Computing for Biologists workshop with Steve Haddock and Casey Dunn. The authors thank them and other members of the workshop (Aurturo Alvarez-Aguilar, Jimmy Bernot, Bill Browne, Anela Choy, Zander Fodor, Michelle Gather, Joel Kingslover, Jasmine Mah, Adelaide Rhodes, Liz Scheimer, Emily Warschefsky, Linda Wordeman, and Sara Wyckoff) for their continued support and enthusiasm for Venomix. The authors would also like to thank Edwin Rice for preliminary input on python scripts, Daft Punk composing good music to allow for coding, and Tyler Carrier for his PeerJ enthusiasm. Support for this project was provided by NSF DEB 1257796 to MD, NSF DEB 1536530 to MD and JM, and NSF OCE 1536530 to AMR. JP was supported by a research assistantship provided by the UNCC Graduate School at UNCC under the guidance of DJ.

6. References

Barghi N., Concepcion GP., Olivera BM., Lluisma AO. 2015. High conopeptide diversity in *Conus tribblei* revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. *Marine Biotechnology* 17:81–98. DOI: 10.1007/s10126-014-9595-7.

Bhere KV., Haney RA., Ayoub NA., Garb JE. 2014. Gene structure, regulatory control, and evolution of black widow venom latrotoxins. *FEBS Letters* 588:3891–3897. DOI: 10.1016/j.febslet.2014.08.034.

Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

Bouzid W., Verdenaud M., Klopp C., Ducancel F., Noirot C., Vétillard A. 2014. De Novo sequencing and transcriptome analysis for Tetramorium bicarinatum: a comprehensive venom gland transcriptome analysis from an ant species. *BMC Genomics* 15:987. DOI: 10.1186/1471-2164-15987.

Buczek O., Bulaj G., Olivera BM. 2005. Conotoxins and the posttranslational modification of secreted gene products. *Cellular and Molecular Life Sciences CMLS* 62:3067–3079. DOI: 10.1007/s00018-005-5283-0.

Calvete JJ., Fasoli E., Sanz L., Boschetti E., Righetti PG. 2009. Exploring the venom proteome of the western diamondback rattlesnake, *Crotalus atrox*, via snake venomics and combinatorial peptide ligand library approaches. *Journal of Proteome Research* 8:3055–3067. DOI: 10.1021/pr900249q.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. DOI: 10.1186/1471-210510-421.

Casewell NR., Huttley GA., Wüster W. 2012. Dynamic evolution of venom proteins in squamate reptiles. *Nature Communications* 3:1066. DOI: 10.1038/ncomms2065.

Casewell NR., Wüster W., Vonk FJ., Harrison RA., Fry BG. 2013. Complex cocktails: the evolutionary novelty of venoms. *Trends in Ecology & Evolution* 28:219–229. DOI: 10.1016/j.tree.2012.10.020.

Cock PJA., Antao T., Chang JT., Chapman BA., Cox CJ., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., Hoon MJL de. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423. DOI: 10.1093/bioinformatics/btp163.

Fan Y-X., Song J., Shen H-B., Kong X. 2011. PredCSF: an integrated feature-based approach for predicting conotoxin superfamily. *Protein and Peptide Letters* 18:261–267.

Fry BG., Roelants K., Champagne DE., Scheib H., Tyndall JDA., King GF., Nevalainen TJ., Norman JA., Lewis RJ., Norton RS., Renjifo C., Vega RCR de la. 2009. The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annual Review of Genomics and Human Genetics* 10:483–511. DOI: 10.1146/annurev.genom.9.081307.164356.

Gacesa R., Barlow DJ., Long PF. 2016. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Computer Science* 2:e90. DOI: 10.7717/peerj-cs.90.

Garb JE., Hayashi CY. 2013. Molecular evolution of α-latrotoxin, the exceptionally potent vertebrate neurotoxin in black widow spider venom. *Molecular Biology and Evolution* 30:999–1014. DOI: 10.1093/molbev/mst011.

Grabherr MG., Haas BJ., Yassour M., Levin JZ., Thompson DA., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren BW., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644–652. DOI: 10.1038/nbt.1883.

Graudins A., Little MJ., Pineda SS., Hains PG., King GF., Broady KW., Nicholson GM. 2012. Cloning and activity of a novel α-latrotoxin from red-back spider venom. *Biochemical Pharmacology* 83:170–183. DOI: 10.1016/j.bcp.2011.09.024.

Gupta S., Kapoor P., Chaudhary K., Gautam A., Kumar R., Consortium OSDD., Raghava GPS. 2013. In silico approach for predicting toxicity of peptides and proteins. *PLOS ONE* 8:e73957. DOI: 10.1371/journal.pone.0073957.

Haas BJ., Papanicolaou A., Yassour M., Grabherr M., Blood PD., Bowden J., Couger MB., Eccles D., Li B., Lieber M., MacManes MD., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R., William T., Dewey CN., Henschel R., LeDuc RD., Friedman N., Regev A. 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature protocols* 8. DOI: 10.1038/nprot.2013.084.

Hargreaves AD., Swain MT., Hegarty MJ., Logan DW., Mulley JF. 2014a. Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. *Genome Biology and Evolution* 6:2088–2095. DOI: 10.1093/gbe/evu166.

Hargreaves AD., Swain MT., Logan DW., Mulley JF. 2014b. Testing the toxicofera: comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon* 92:140–156. DOI: 10.1016/j.toxicon.2014.10.004.

Jungo F., Bougueleret L., Xenarios I., Poux S. 2012. The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. *Toxicon* 60:551–557. DOI: 10.1016/j.toxicon.2012.03.010.

Junqueira-de-Azevedo ILM., Bastos CMV., Ho PL., Luna MS., Yamanouye N., Casewell NR. 2015. Venom-related transcripts from *Bothrops jararaca* tissues provide novel molecular insights into the production and evolution of snake venom. *Molecular Biology and Evolution* 32:754–766. DOI: 10.1093/molbev/msu337.

Kaas Q., Craik DJ. 2015. Bioinformatics-aided venomics. *Toxins* 7:2159–2187. DOI: 10.3390/toxins7062159.

Kaas Q., Yu R., Jin A-H., Dutertre S., Craik DJ. 2012. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Research* 40:D325–D330. DOI: 10.1093/nar/gkr886.

Kaplan N., Morpurgo N., Linial M. 2007. Novel families of toxin-like peptides in insects and mammals: A computational approach. *Journal of Molecular Biology* 369:553–566. DOI: 10.1016/j.jmb.2007.02.106.

Katoh K., Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780. DOI: 10.1093/molbev/mst010.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649. DOI: 10.1093/bioinformatics/bts199.

Kozlov S., Grishin E. 2011. The mining of toxin-like polypeptides from EST database by single residue distribution analysis. *BMC Genomics* 12:88. DOI: 10.1186/1471-2164-12-88.

Lewis Ames C., Ryan JF., Bely AE., Cartwright P., Collins AG. 2016. A new transcriptome and transcriptome profiling of adult and larval tissue in the box jellyfish *Alatina alata*: an emerging model for studying venom, vision and sex. *BMC Genomics* 17. DOI: 10.1186/s12864-016-2944-3.

Li B., Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. DOI: 10.1186/1471-2105-12-323.

Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.

Luna-Ramírez K., Quintero-Hernández V., Juárez-González VR., Possani LD. 2015. Whole transcriptome of the venom gland from *Urodacus yaschenkoi* scorpion. *PLOS ONE* 10:e0127883. DOI: 10.1371/journal.pone.0127883.

Macrander J., Broe M., Daly M. 2015. Multi-copy venom genes hidden in de novo transcriptome assemblies, a cautionary tale with the snakelocks sea anemone *Anemonia sulcata* (Pennant, 1977). *Toxicon* 108:184–188. DOI: 10.1016/j.toxicon.2015.09.038.

Macrander J., Broe M., Daly M. 2016. Tissue-specific venom composition and differential gene expression in sea anemones. *Genome Biology and Evolution*:evw155. DOI: 10.1093/gbe/evw155.

Macrander J., Brugler MR., Daly M. 2015. A RNA-seq approach to identify putative toxins from acrorhagi in aggressive and non-aggressive *Anthopleura elegantissima* polyps. *BMC Genomics* 16:221. DOI: 10.1186/s12864-015-1417-4.

Magazanik LG., Fedorova IM., Kovalevskaya GI., Pashkov VN., Bulgakov OV., Grishin EV. 1992. Selective presynaptic insectotoxin (α-latroinsectotoxin) isolated from black widow spider venom. *Neuroscience* 46:181–188. DOI: 10.1016/0306-4522(92)90017-V.

Ménez A., Stöcklin R., Mebs D. 2006. ‘Venomics’ or: The venomous systems genome project. *Toxicon* 47:255–259. DOI: 10.1016/j.toxicon.2005.12.010.

Neumann RS., Kumar S., Shalchian-Tabrizi K. 2014. BLAST output visualization in the new sequencing era. *Briefings in Bioinformatics* 15:484–503. DOI: 10.1093/bib/bbt009.

OmPraba G., Chapeaurouge A., Doley R., Devi KR., Padmanaban P., Venkatraman C., Velmurugan D., Lin Q., Kini RM. 2010. Identification of a novel family of snake venom proteins veficolins from *Cerberus rynchops* using a venom gland transcriptomics and proteomics approach. *Journal of Proteome Research* 9:1882–1893. DOI: 10.1021/pr901044x.

Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290. DOI: 10.1093/bioinformatics/btg412.

Petersen TN., Brunak S., von Heijne G., Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8:785–786. DOI: 10.1038/nmeth.1701.

Phuong MA., Mahardika GN., Alfaro ME. 2016. Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics* 17:401. DOI: 10.1186/s12864-016-2755-6.

Prashanth JR., Lewis RJ. 2015. An efficient transcriptome analysis pipeline to accelerate venom peptide discovery and characterization. *Toxicon* 107, Part B:282–289. DOI: 10.1016/j.toxicon.2015.09.012.

Price MN., Dehal PS., Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5. DOI: 10.1371/journal.pone.0009490.

Reumont BM von., Blanke A., Richter S., Alvarez F., Bleidorn C., Jenner RA. 2014. The first venomous crustacean revealed by transcriptomics and functional morphology: remipede venom glands express a unique toxin cocktail dominated by enzymes and a neurotoxin. *Molecular Biology and Evolution* 31:48–58. DOI: 10.1093/molbev/mst199.

Reyes-Velasco J., Card DC., Andrew AL., Shaney KJ., Adams RH., Schield DR., Casewell NR., Mackessy SP., Castoe TA. 2015a. Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. *Molecular Biology and Evolution* 32:173–183. DOI: 10.1093/molbev/msu294.

Rodríguez de la Vega RC., Giraud T. 2016. Intragenome diversity of gene families encoding toxin-like proteins in venomous animals. *Integrative and Comparative Biology*:icw097. DOI: 10.1093/icb/icw097.

Sanders S., Ganote C., Papudeshi B., Mockaitis K., Doak T. 2018. NCGAS makes robust transcriptome analysis easier with a readily usable workflow following de novo assembly best practices.

Simão FA., Waterhouse RM., Ioannidis P., Kriventseva EV., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212. DOI: 10.1093/bioinformatics/btv351.

Sunagar K., Morgenstern D., Reitzel AM., Moran Y. 2016. Ecological venomics: How genomics, transcriptomics and proteomics can shed new light on the ecology and evolution of venom. *Journal of Proteomics* 135:62–72. DOI: 10.1016/j.jprot.2015.09.015.

Tan PTJ., Khan AM., Brusic V. 2003. Bioinformatics for venom and toxin sciences. *Briefings in Bioinformatics* 4:53–62. DOI: 10.1093/bib/4.1.53.

Terrat Y., Ducancel F. 2013. Are there unequivocal criteria to label a given protein as a toxin? Permissive versus conservative annotation processes. *Genome Biology* 14:406. DOI: 10.1186/gb-2013-14-9406.

Todd EV., Black MA., Gemmell NJ. 2016. The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology* 25:1224–1241. DOI: 10.1111/mec.13526.

Undheim EAB., Sunagar K., Herzig V., Kely L., Low DHW., Jackson TNW., Jones A., Kurniawan N., King GF., Ali SA., Antunes A., Ruder T., Fry BG. 2013. A Proteomics and transcriptomics investigation of the venom from the barychelid spider *Trittame loki* (brush-foot trapdoor). *Toxins* 5:2488–2503. DOI: 10.3390/toxins5122488.

Vijay N., Poelstra JW., Künstner A., Wolf JBW. 2013. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology* 22:620–634. DOI: 10.1111/mec.12014.

Vonk FJ., Casewell NR., Henkel CV., Heimberg AM., Jansen HJ., McCleary RJR., Kerkkamp HME., Vos RA., Guerreiro I., Calvete JJ., Wüster W., Woods AE., Logan JM., Harrison RA., Castoe TA., Koning APJ de., Pollock DD., Yandell M., Calderon D., Renjifo C., Currier RB., Salgado D., Pla D., Sanz L., Hyder AS., Ribeiro JMC., Arntzen JW., Thillart GEEJM van den., Boetzer M., Pirovano W., Dirks RP., Spaink HP., Duboule D., McGlinn E., Kini RM., Richardson MK. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences* 110:20651–20656. DOI: 10.1073/pnas.1314702110.

Wong ESW., Belov K. 2012. Venom evolution through gene duplications. *Gene* 496:1–7. DOI: 10.1016/j.gene.2012.01.009.

Wong ESW., Hardy MC., Wood D., Bailey T., King GF. 2013. SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. *PLOS ONE* 8:e66279. DOI: 10.1371/journal.pone.0066279.

Yang Y., Smith SA. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* 14:328. DOI: 10.1186/1471-2164-14-328.

Table 1(on next page)

Table 1. Species specific information for transcriptome assemblies.

*indicates alternative assembly approach, so comparisons were not possible; PE= Paired End; bp = Base Pair

Table 1. Species specific information for transcriptome assemblies.

	Sequencing	Raw Reads	Transcripts	N50
C. sponsalis	PE-100bp	26,419,249 (-8.9%)	53,349 (+22.0%)	546 (-4.7%)
*E. coloratus**	PE-100bp	68,011,342	173,198	1,603
*T. bicarinatum**	PE-100bp	424,743,516	200,106	881
U. yaschenkoi	PE-100bp	82,746,144 (-1.0%)	170,984 (-29. 9%)	1,248 (+8.7%)

*indicates alternative assembly approach, so comparisons were not possible; PE= Paired End; bp = Base Pair

Table 2(on next page)

Table 2. Species-specific Venomix outputs following different search strategies

N = number of candidate toxins identified in original study. Groups = number of toxins types identified based on sequence similarity, c = conotoxins only (Phuong, Mahardika & Alfaro, 2016), s = scorpion toxins only (Luna-Ramírez et al., 2015); Transcripts = total number of unique transcripts evaluated, † = includes duplicates as cumulative after three iterations in Trinity [see 33]. β = >100 TPM difference upregulated in the venom gland compared the ant carcass. E = number of candidates based on different E-values 10/1E-3 thresholds. Evaluated = number of unique transcripts retained after using BLAST screening, parenthesis indicates number of transcripts identified using a Toxclassifier score of 1 or greater.

Table 2. Species-specific Venomix outputs following different search strategies

	Original Publication				Stringent (*e-value 1e-20*)			Less Stringent (*e-value 1e-6*)		
	N	**Groups**	**Transcripts**	**Evaluated**	**Groups**	**Transcripts**	**Evaluated**	**Groups**	**Transcripts**	**Evaluated**
C. sponsalis	***401***	35^{c}	$780^{\dagger}$	393(363)	22	61	44(13)	75	293	246(45)
E. coloratus	***34***	8	82	62(35)	72	339	147(116)	130	775	202(143)
T. bicarinatum	***69***	32	527^{β}	$287/62^{E}(14)$	36	289	95(14)	75	761	201(36)
U. yaschenkoi	***111***	11^{s}	210	71(6)	50	277	48(34)	117	689	179(38)

N = number of candidate toxins identified in original study. Groups = number of toxins types identified based on sequence similarity, c = conotoxins only (Phuong, Mahardika & Alfaro, 2016), s = scorpion toxins only (Luna-Ramírez et al., 2015); Transcripts = total number of unique transcripts evaluated, † = includes duplicates as cumulative after three iterations in Trinity [see 33]. β = >100 TPM difference upregulated in the venom gland compared the ant carcass. E = number of candidates based on different E-values 10/1E-3 thresholds. Evaluated = number of unique transcripts retained after using BLAST screening, parenthesis indicates number of transcripts identified using a Toxclassifier score of 1 or greater.

Table 3(on next page)

Table 3. Comparison of Conotoxin Transcripts for *C. sponsalis*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26733v1 | CC BY 4.0 Open Access | rec: 19 Mar 2018, publ: 19 Mar 2018

Table 3. Comparison of Conotoxin Transcripts for *C. sponsalis*

Super Family	**Venomix**	**Phuong et al.** (2016)
A	0	3
B1	2	2
B4r	0	9
con-ikot-ikot	0	12
conkunitzin	0	7
conodipine	10	3
conophysin	7	3
D	0	4
Divergent_MKFPLLFISL	0	2
E	0	3
F	0	2
G-like	1	11
I1	2	2
I2	0	3
I3	4	3
I4	0	3
J	0	3
L	1	1
M	29	29
MEFRRr	0	3
MKFLLr	0	2
MKISL*	1	1
N	0	8
O1	108	95
O2	12	28
O3	1	6
P	0	13
Q	0	1
SF-04	0	1
SF-mi1	0	6
SF-mi2	0	3
T	3	56
U	0	6
V	0	3
Y	0	3
Total	**179**	**401**

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26733v1 | CC BY 4.0 Open Access | rec: 19 Mar 2018, publ: 19 Mar 2018

Table 4(on next page)

Table 4. Number of previously predicted toxin compared to those derived from Venomix

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26733v1 | CC BY 4.0 Open Access | rec: 19 Mar 2018, publ: 19 Mar 2018

Table 4. Number of previously predicted toxin compared to those derived from Venomix

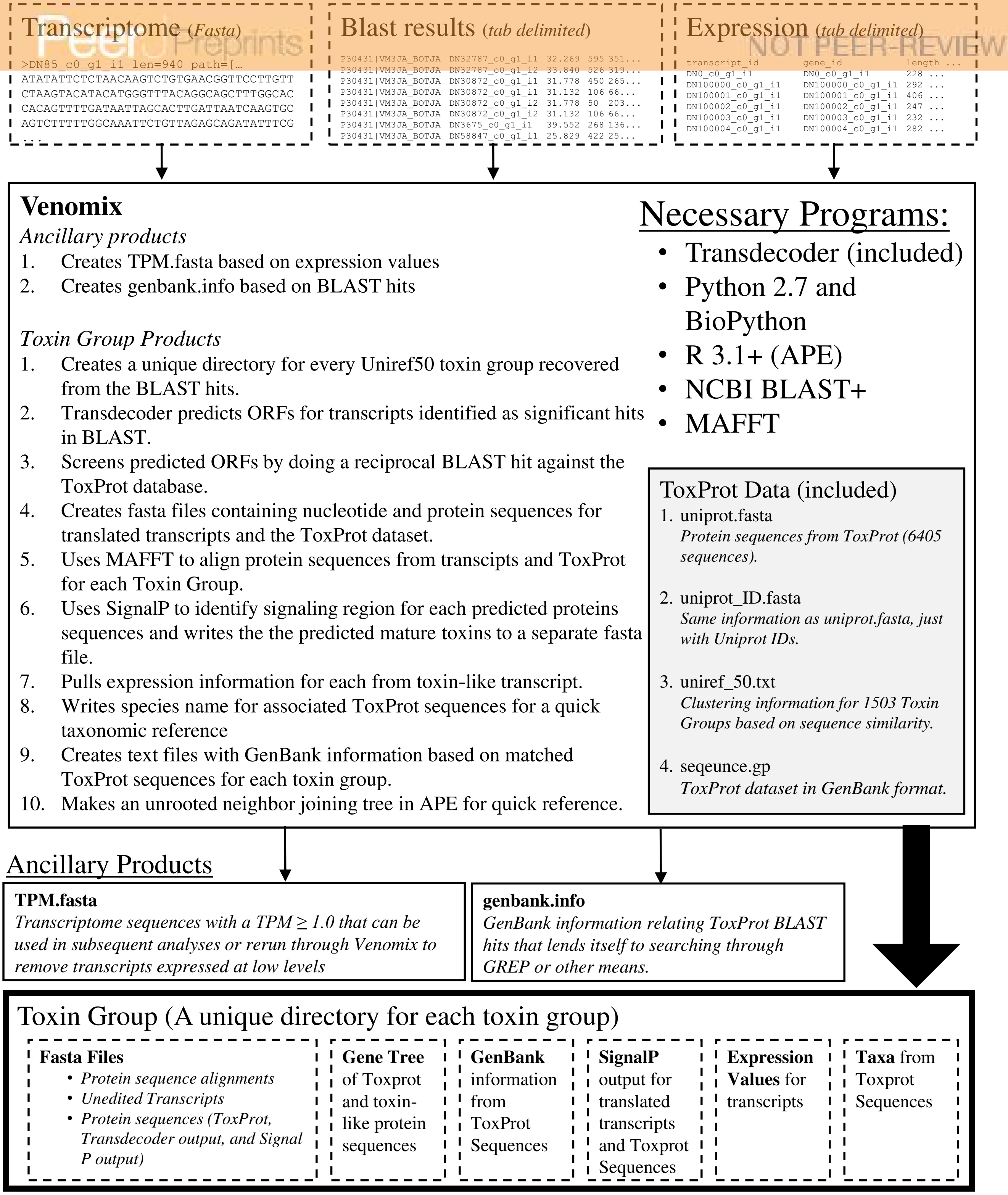
Toxin Family	**Venomix**	**Hargreaves et al *(2014b)***
Snake Venom Metalloproteases	36	13
C-type lectin	49	8
Serine protesase	10	6
Phospholipase A2	6	3
Cysteine-rich secretory proteins	3	1
L-amino acid oxidase	2	1
Vascular Endothelial Growth Factors	1	1
Crotamine	0	1

1

Figure 1(on next page)

Venomix Pipeline Outline

Outline showing the stepwise progression of the Venomix pipeline, including the necessary inputs (dashed lined boxes above), ancillary products, and files included for every Toxin Group directory.



Transcriptome (*Fasta*)

```
>DN85_c0_g1_i1 len=940 path=[…
ATATATTCTCTAACAAGTCTGTGAACGGTTCCTTGTT
CTAAGTACATACATGGGTTTACAGGCAGCTTTGGCAC
CACAGTTTTGATAATTAGCACTTGATTAATCAAGTGC
AGTCTTTTTGGCAAATTCTGTTAGAGCAGATATTTCG
...
```

Blast results (*tab delimited*)

```
P30431|VM3JA_BOTJA  DN32787_c0_g1_i1  32.269  595 351...
P30431|VM3JA_BOTJA  DN32787_c0_g1_i2  33.840  526 319...
P30431|VM3JA_BOTJA  DN30872_c0_g1_i1  31.778  450 265...
P30431|VM3JA_BOTJA  DN30872_c0_g1_i1  31.132  106 66...
P30431|VM3JA_BOTJA  DN30872_c0_g1_i2  31.778  50  203...
P30431|VM3JA_BOTJA  DN30872_c0_g1_i2  31.132  106 66...
P30431|VM3JA_BOTJA  DN3675_c0_g1_i1   39.552  268 136...
P30431|VM3JA_BOTJA  DN58847_c0_g1_i1  25.829  422 25...
```

Expression (*tab delimited*)

```
transcript_id        gene_id              length ...
DN0_c0_g1_i1         DN0_c0_g1_i1         228 ...
DN100000_c0_g1_i1    DN100000_c0_g1_i1    292 ...
DN100001_c0_g1_i1    DN100001_c0_g1_i1    406 ...
DN100002_c0_g1_i1    DN100002_c0_g1_i1    247 ...
DN100003_c0_g1_i1    DN100003_c0_g1_i1    232 ...
DN100004_c0_g1_i1    DN100004_c0_g1_i1    282 ...
```

Venomix

Ancillary products

1. Creates TPM.fasta based on expression values
2. Creates genbank.info based on BLAST hits

Toxin Group Products

1. Creates a unique directory for every Uniref50 toxin group recovered from the BLAST hits.
2. Transdecoder predicts ORFs for transcripts identified as significant hits in BLAST.
3. Screens predicted ORFs by doing a reciprocal BLAST hit against the ToxProt database.
4. Creates fasta files containing nucleotide and protein sequences for translated transcripts and the ToxProt dataset.
5. Uses MAFFT to align protein sequences from transcipts and ToxProt for each Toxin Group.
6. Uses SignalP to identify signaling region for each predicted proteins sequences and writes the the predicted mature toxins to a separate fasta file.
7. Pulls expression information for each from toxin-like transcript.
8. Writes species name for associated ToxProt sequences for a quick taxonomic reference
9. Creates text files with GenBank information based on matched ToxProt sequences for each toxin group.
10. Makes an unrooted neighbor joining tree in APE for quick reference.

Necessary Programs:

- Transdecoder (included)
- Python 2.7 and BioPython
- R 3.1+ (APE)
- NCBI BLAST+
- MAFFT

ToxProt Data (included)

1. uniprot.fasta
   Protein sequences from ToxProt (6405 sequences).
2. uniprot_ID.fasta
   Same information as uniprot.fasta, just with Uniprot IDs.
3. uniref_50.txt
   Clustering information for 1503 Toxin Groups based on sequence similarity.
4. seqeunce.gp
   ToxProt dataset in GenBank format.

Ancillary Products

TPM.fasta
Transcriptome sequences with a TPM ≥ 1.0 that can be used in subsequent analyses or rerun through Venomix to remove transcripts expressed at low levels

genbank.info
GenBank information relating ToxProt BLAST hits that lends itself to searching through GREP or other means.

Toxin Group (A unique directory for each toxin group)

Fasta Files

- *Protein sequence alignments*
- *Unedited Transcripts*
- *Protein sequences (ToxProt, Transdecoder output, and Signal P output)*

Gene Tree of Toxprot and toxin-like protein sequences

GenBank information from ToxProt Sequences

SignalP output for translated transcripts and Toxprot Sequences

Expression Values for transcripts

Taxa from Toxprot Sequences

Figure 2(on next page)

Candidate toxins from *U. yaschenkoi.*

(A) Candidate toxins from *U. yaschenkoi* highlighting alignment difference in the candidate Toxin-like protein 14 sequencing in both analyses. Conserved residues across the alignment are highlighted. (B) Maximum likelihood neprilysin gene tree highlighting the abundance and diversity of candidate neprilysin toxins and non-venomous neprilysin genes. Branches associated with transcripts from *U. yaschenkoi* are highlighted in orange throughout the tree. Venomous taxa emphasized with bold font.

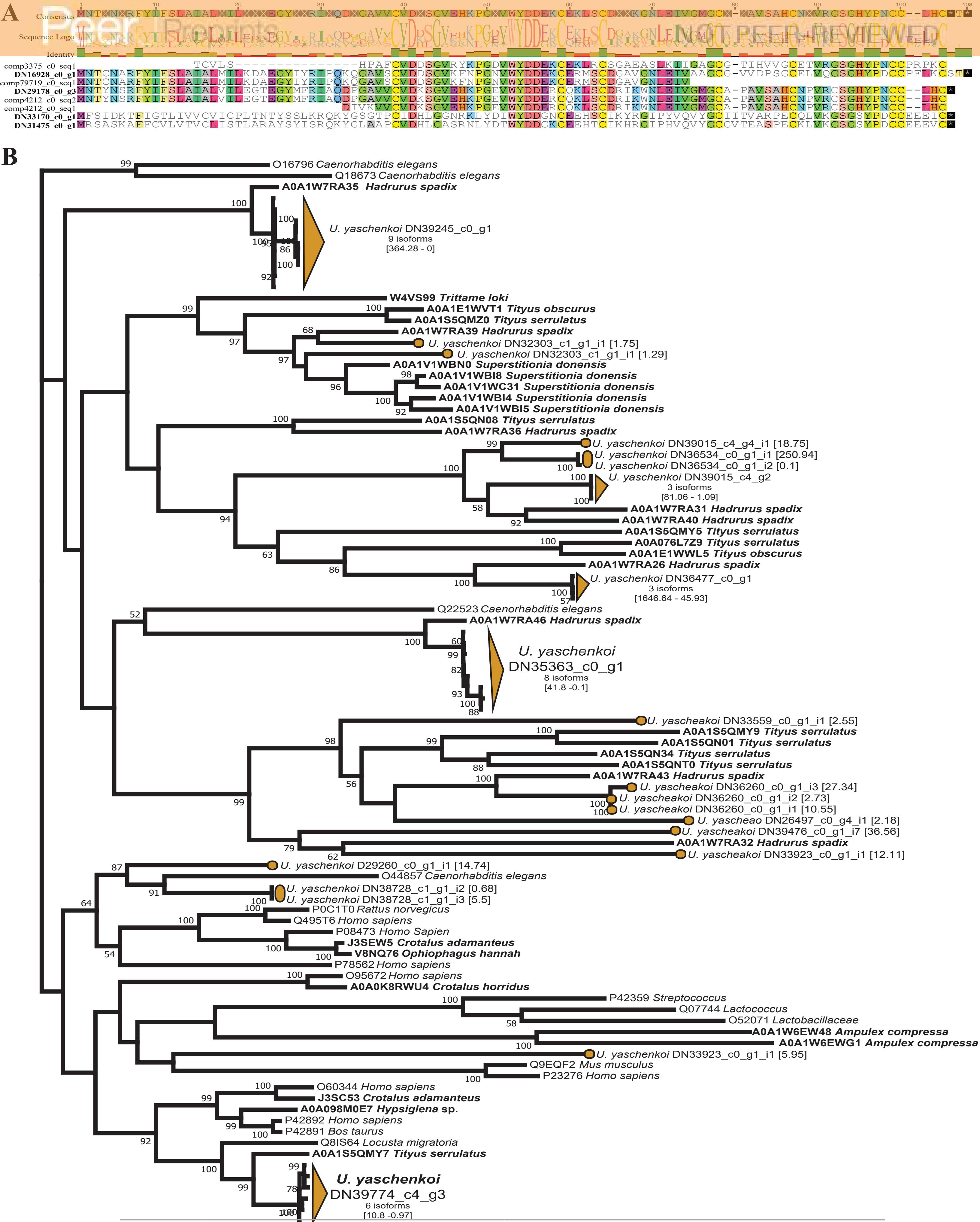
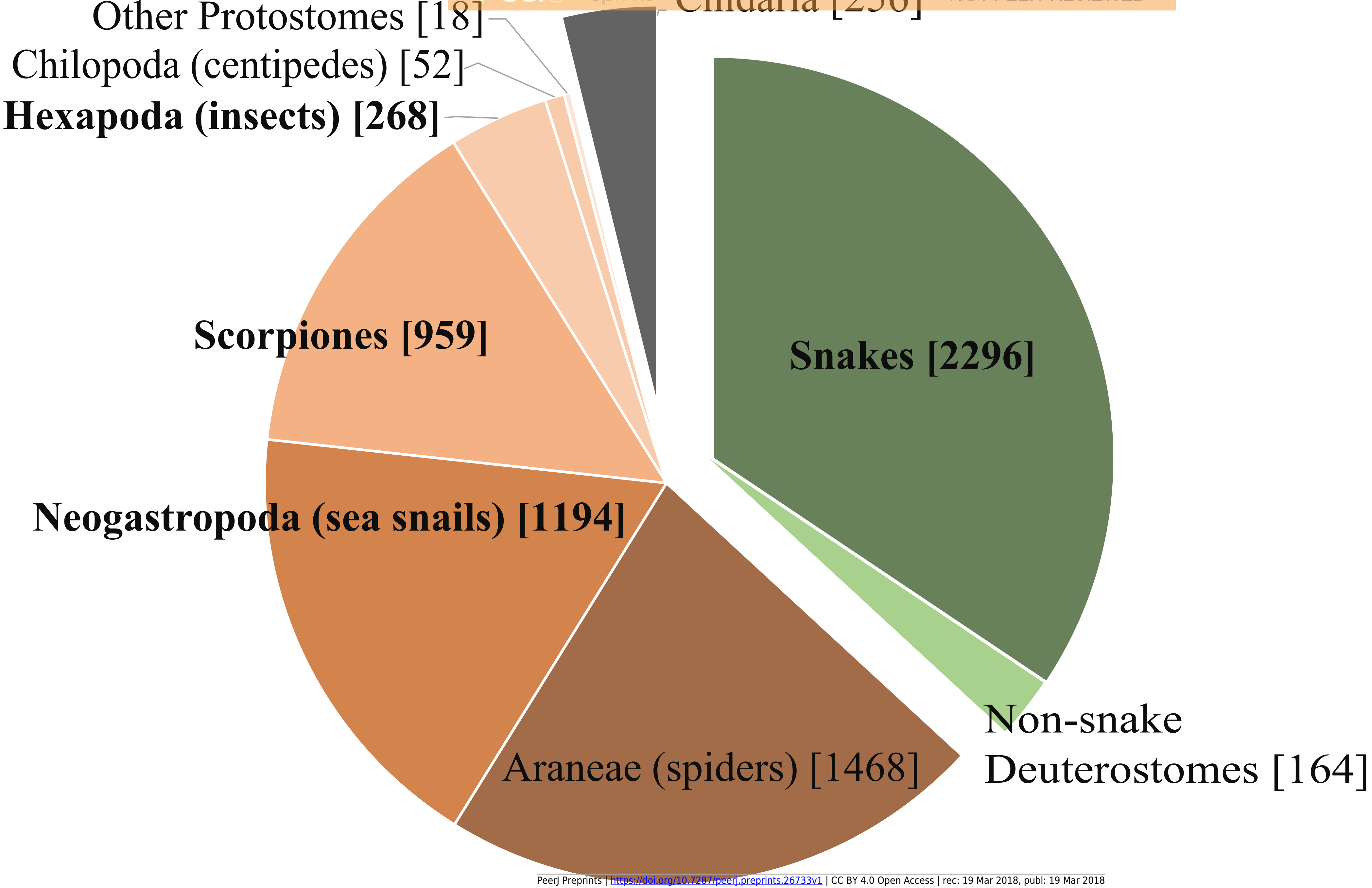

A
Consensus
Sequence Logo
Identity
1
10
20
30
40
50
60
70
80
90
100
108
MNTXNXRFYIFSLAIALXILXXXEGYXXRIXQDXGAVVCVDXSGVEHKPGXVWYDDEKCEKLSCDXXKGNLEIVGMGCX-XAVSAHCNXVRGSGHYPNCC--LHC*T*
comp3375_c0_seq1 TCVLS---------------HPAFCVDDSGVRYKPGDVWYDDEKCEKLRCSGAEASLKIIGAGCG-TIHVVGCETVRGSGHYPNCCPRPKC
DN16928_c0_g1 MNTCNARFYIFSLAIALMILKDAEGYIYRIPQKQGAVSCVDDSGVKFNPGNVWYDDEKCERMSCDGAVGNLEIVAAGCG-VVDPSGCELVQGSGHYPDCCPFLKCST*
comp79719_c0_seq1 MNTCNARFYIFSLAIALMILKDAEGYIYRIPQKQGAVSCVDDSGVKFNPGNVWYDDEKCERMSCDGAVGNLEIV
DN29178_c0_g3 MNTYNSRFYIFSLAIALVILEGTEGYMFRIAQDPGAVVCVDKSGVEHKPGEVWYDDERCQKLSCDRIKWNLEIVGMGCA-PAVSAHCNPVRCSGHYPNCC--LHC*
comp4212_c0_seq2 MNTYNSRFYIFSLAIALVILEGTEGYMFRIAQDPGAVVCVDKSGVEHKPGEVWYDDERCQKLSCDRIKWNLEIVGMGCA-PAVSAHCNPVRCSGHYPNCC--LHC
comp4212_c0_seq1 DIVKVVCVDKSGVEHKPGEVWYDDERCQKLSCDRIKWNLEIVGMGCA-PAVSAHCNPVRCSGHYPNCC--LHC
DN33170_c0_g1 MFSIDKTFIGTLIVVCVICPLTNTYSSLKRQKYGSGTPCIDHLGGNRKLYDIWYDDGNCEEHSCIKYRGIPYVQVYGCIITVARPECQLVKGSGSYPDCCEEEIC*
DN31475_c0_g1 MRSASKAFFCVLVTVCLISTLARAYSYISRQKYGLAAPCVDHLGASRNLYDTWYDDGKCEEHTCIKHRGIPHVQVYGCGVTEASPECKLVKGSGSYPDCCEEEVC*
B
99 O16796 Caenorhabditis elegans
Q18673 Caenorhabditis elegans
A0A1W7RA35 Hadrurus spadix
100
100
100 95 100
86
100
92
U. yaschenkoi DN39245_c0_g1
9 isoforms
[364.28 - 0]
99 W4VS99 Trittame loki
100 A0A1E1WVT1 Tityus obscurus
A0A1S5QMZ0 Tityus serrulatus
97
68 A0A1W7RA39 Hadrurus spadix
U. yaschenkoi DN32303_c1_g1_i1 [1.75]
97
U. yaschenkoi DN32303_c1_g1_i1 [1.29]
A0A1V1WBN0 Superstitionia donensis
98 A0A1V1WBI8 Superstitionia donensis
96
A0A1V1WC31 Superstitionia donensis
100
A0A1V1WBI4 Superstitionia donensis
92 A0A1V1WBI5 Superstitionia donensis
100 A0A1S5QN08 Tityus serrulatus
A0A1W7RA36 Hadrurus spadix
99 U. yaschenkoi DN39015_c4_g4_i1 [18.75]
U. yaschenkoi DN36534_c0_g1_i1 [250.94]
100
100 U. yaschenkoi DN36534_c0_g1_i2 [0.1]
100 U. yaschenkoi DN39015_c4_g2
3 isoforms
[81.06 - 1.09]
100
58 A0A1W7RA31 Hadrurus spadix
94
92 A0A1W7RA40 Hadrurus spadix
A0A1S5QMY5 Tityus serrulatus
100 A0A076L7Z9 Tityus serrulatus
63
A0A1E1WWL5 Tityus obscurus
86 A0A1W7RA26 Hadrurus spadix
100 U. yaschenkoi DN36477_c0_g1
100
57
3 isoforms
[1646.64 - 45.93]
52 Q22523 Caenorhabditis elegans
A0A1W7RA46 Hadrurus spadix
100
60
99
82
93
100
88
U. yaschenkoi
DN35363_c0_g1
8 isoforms
[41.8 -0.1]
U. yascheakoi DN33559_c0_g1_i1 [2.55]
100 A0A1S5QMY9 Tityus serrulatus
98
99 A0A1S5QN01 Tityus serrulatus
A0A1S5QN34 Tityus serrulatus
88 A0A1S5QNT0 Tityus serrulatus
100 A0A1W7RA43 Hadrurus spadix
56
U. yascheakoi DN36260_c0_g1_i3 [27.34]
99
100 U. yascheakoi DN36260_c0_g1_i2 [2.73]
100 U. yascheakoi DN36260_c0_g1_i1 [10.55]
U. yascheao DN26497_c0_g4_i1 [2.18]
U. yascheakoi DN39476_c0_g1_i7 [36.56]
79
A0A1W7RA32 Hadrurus spadix
62
U. yascheakoi DN33923_c0_g1_i1 [12.11]
87 U. yaschenkoi D29260_c0_g1_i1 [14.74]
O44857 Caenorhabditis elegans
91
U. yaschenkoi DN38728_c1_g1_i2 [0.68]
100 U. yaschenkoi DN38728_c1_g1_i3 [5.5]
64
100 P0C1T0 Rattus norvegicus
100 Q495T6 Homo sapiens
P08473 Homo Sapien
100 J3SEW5 Crotalus adamanteus
54
100 V8NQ76 Ophiophagus hannah
P78562 Homo sapiens
100 O95672 Homo sapiens
A0A0K8RWU4 Crotalus horridus
P42359 Streptococcus
100 Q07744 Lactococcus
58 O52071 Lactobacillaceae
A0A1W6EW48 Ampulex compressa
100 A0A1W6EWG1 Ampulex compressa
U. yaschenkoi DN33923_c0_g1_i1 [5.95]
Q9EQF2 Mus musculus
100 P23276 Homo sapiens
100 O60344 Homo sapiens
99
J3SC53 Crotalus adamanteus
A0A098M0E7 Hypsiglena sp.
99 P42892 Homo sapiens
100 P42891 Bos taurus
92
Q8IS64 Locusta migratoria
A0A1S5QMY7 Tityus serrulatus
100
99
99
78
100
U. yaschenkoi
DN39774_c4_g3
6 isoforms
[10.8 -0.97]
0.2

Figure 3(on next page)

Taxonomic distribution of venom and toxin proteins in the ToxProt dataset.

Taxonomic distribution of venom and toxin proteins in the ToxProt dataset. Deuterostomes are highlighted in green, protostomes in brown, and cnidarians in grey.


Other Protostomes [18]
Chilopoda (centipedes) [52]
Hexapoda (insects) [268]
Cnidaria [256]
Scorpiones [959]
Snakes [2296]
Neogastropoda (sea snails) [1194]
Araneae (spiders) [1468]
Non-snake Deuterostomes [164]