A peer-reviewed version of this preprint was published in PeerJ on 31 July 2018.

<u>View the peer-reviewed version</u> (peerj.com/articles/5361), which is the preferred citable publication unless you specifically need to cite this preprint.

Macrander J, Panda J, Janies D, Daly M, Reitzel AM. 2018. Venomix: a simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data. PeerJ 6:e5361 https://doi.org/10.7717/peerj.5361

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/

577 Venomix: A Simple Bioinformatic Pipeline for Identifying and Characterizing Toxin Gene

- 578 Candidates from Transcriptomic Data.
- Jason Macrander^{1,2}, Jyothirmayi Panda³, Daniel Janies^{3,4}, Marymegan Daly², Adam M. Reitzel¹ 1.
- 580 University of North Carolina at Charlotte, Department of Biological Sciences, 9201 University City
- 581 Blvd., 373 Woodward Hall, Charlotte, NC-28223, USA.
- 582 2. The Ohio State University, Department of Evolution, Ecology, and Organismal Biology, 318 W.
- 583 12th Ave. 300 Aronoff Laboratory, Columbus, OH-43215, USA.
- 584 3. University of North Carolina at Charlotte, College of Computing and Informatics, 9201
- 585 University City Blvd., Charlotte, NC-28223, USA.
- 4. University of North Carolina at Charlotte, Department of Bioinformatics and Genomics, 9201
- 587 University City Blvd., Charlotte, NC-28223, USA.
- 588
- 589 Corresponding Author
- 590 Jason Macrander^{1,2}
- 591
- 592 E-mail address: jmacrand@uncc.edu
- 593 Abstract

594 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 595 596 resulted in the area of "venomics": a rapidly growing field using combined transcriptomic and proteomic 597 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 598 599 identification using BLAST, expression level screening, protein sequence alignment, gene tree 600 reconstruction, and characterization of potential toxin function. Here we describe the python package 601 Venomix, which streamlines these processes using commonly used bioinformatic tools along with a

602	public, annotated database comprised of characterized venom proteins. In this study, we use the Venomix
603	pipeline to characterize candidate venom diversity in four phylogenetically distinct organisms, a cone
604	snail (Conidae; Conus sponsalis), a snake (Viperidae; Echis coloratus), an ant (Formicidae; Tetramorium
605	bicarinatum), and a scorpion (Scorpionidae; Urodacus yaschenkoi). Data on these organisms was sampled
606	from public databases and thus different approaches to either transcriptome assembly, toxin identification,
607	or gene expression quantification was used for each. Of the organisms used in our analysis, Venomix
608	recovered numerically more candidate toxin transcripts for three of the four transcriptomes than the
609	original analyses. In four of four organisms we identified new toxin candidates that were not reported in
610	the original analysis. In summary, we show that the Venomix package is a useful tool to identify and
611	characterize the diversity of toxin-like transcripts. Venomix is available at:
612	https://bitbucket.org/JasonMacrander/Venomix/
613	
614	Keywords: Venom, Transcriptome, Python, Transdecoder, SignalP, Protein
615	
616	1. Introduction
617	Throughout the animal kingdom, venom has evolved independently multiple times to be used in
618	prey capture, predatory defense, and intraspecific competition (Casewell et al., 2013). Venoms are toxic
619	cocktails with remarkable diversity in protein action and specificity across animals. The evolutionary and
620	ecological processes shaping this diversity are of major interest (Fry et al., 2009; Wong & Belov, 2012;
621	Casewell, Huttley & Wüster, 2012; Sunagar et al., 2016; Rodríguez de la Vega & Giraud, 2016), with
622	much of this focusing on characterizing protein and RNA composition expressed in the venom gland
623	(Ménez, Stöcklin & Mebs, 2006). As sequencing costs decrease and assembly programs are becoming
624	more efficient, the number of venom-focused studies is increasing at a dramatic rate. For some of the
625	
025	better studied venomous lineages (e.g. Colubroidea), comparative transcriptome and genome sequencing
626	better studied venomous lineages (<i>e.g.</i> Colubroidea), comparative transcriptome and genome sequencing are being used to investigate processes involved with toxin gene recruitment and tissue specific gene

627 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.

634 Bioinformatic tools that use transcriptomic, proteomic, and genomic data sets have emerged for a 635 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 636 (Kaplan, Morpurgo & Linial, 2007; Fan et al., 2011; Wong et al., 2013) and identifiers based on residue 637 638 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 639 640 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 641 642 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, 643 2011). For many of the pipelines, these outputs are then screened using custom query datasets comprised 644 645 of lineage specific toxin genes or the manually curated ToxProt dataset (Jungo et al., 2012), which 646 includes all characterized/annotated animal venom proteins. Following candidate toxin gene 647 identification, downstream analyses often involve predicting open-reading frames using Transdecoder 648 (Haas et al., 2013), in combination with signal region prediction using SignalP (Petersen et al., 2011). 649 These types of data are sometimes complemented with genome and proteome datasets (see Sunagar et al., 650 2016). However, the majority of studies that are exploratory in nature use transcriptomic approaches to 651 describe overall toxin diversity for a variety of poorly studied taxa (Reumont et al., 2014; Macrander,

652 Brugler & Daly, 2015; Barghi et al., 2015; Luna-Ramírez et al., 2015; Macrander, Broe & Daly, 2016; 653 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 654 655 characterize multiple toxin gene families for diverse toxin genes. 656 Here were present Venomix, a bioinformatic pipeline written in the programming languages 657 Python and R that follows generally accepted methods for identifying and characterizing toxin-like genes 658 from transcriptomic datasets. This is a single, easy-to-use bioinformatic pipeline that will screen 659 transcriptomic datasets of diverse taxa for toxin-like genes. Venomix incorporates widely used programs 660 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 661 662 et al., 2011) to predict toxin gene signaling regions, MAFFT (Katoh & Standley, 2013) for protein 663 sequence alignment, and the R package APE (Paradis, Claude & Strimmer, 2004) to construct gene trees. 664 Candidate toxin genes are grouped based on sequence similarity, with each directory corresponding to a 665 specific toxin group based on the ToxProt sequence names (e.g., some variation of conotoxin, Kunitz-type 666 serine protease, phospholipase A2, zinc metalloproteinase). The Venomix pipeline provides the user with 667 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 668 669 among functionally important residues in sequence alignments, and review taxonomic and functional information in combination with tree reconstructions to further evaluate toxin gene candidates. 670 In this study, we use Venomix to characterize the toxin-like diversity from venom gland 671 672 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. 673 Although Venomix alone is not designed to be used to as a definitive validation pipeline for toxin genes, 674 675 it can quickly identify, sort, and characterize transcripts that may be used to further evaluate these venom 676 candidates. By abating time required by these processes, researchers are freed to focus on other aspects of

- 677 toxin gene identification, such functionally characterizing toxin genes from their transcriptome or some of
- 678 the lineage specific tools to better understand venom diversity present in the transcriptome.
- 679 **2.** Materials and Methods
- 680 2.1 Data Acquisition and Transcriptome Assembly
- 681 Raw reads from four different analyses were downloaded from the short read archive (SRA) on
- 682 GenBank (C. sponsalis: SRR260951 (Phuong, Mahardika & Alfaro, 2016); U. yaschenkoi: SRR1557168
- 683 (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
- 685 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
- 690 parameters of its built-in Trimmomatic program to clean up the sequences (Bolger, Lohse & Usadel,
- 691 2014). For each transcriptome, expression values (TPM and FKPM) were calculated in the program
- 692 RSEM (Li & Dewey, 2011) as part of the Trinity package.
- 693 2.2 Analysis Pipeline and Execution

The bioinformatic pipeline for Venomix is outlined in Figure 1. The program requires three inputs 694 695 provided by the user: an assembled transcriptome, gene expression information in the form of a tab 696 delimited output with transcript names in the first column, and tab delimited BLAST output using the 697 ToxProt as query sequences. Following transcriptome assembly and expression level calculations, the 698 final user provided file is created using tBLASTn from NCBI BLAST+ version 2.4.0 (Camacho et al., 699 2009), with the ToxProt dataset as the search query with the final BLAST alignment results shown in a 700 tabular format (-outfmt 6). Query sequences from ToxProt are provided within the Venomix package, 701 however, alternative curations of the ToxProt dataset may be used if the sequence identifiers are not

702 changed. In our analysis, we implemented two BLAST search procedures; the first used a more stringent 703 identification algorithm (E-value 1e-20) and a less stringent identification algorithm (E-value 1e-6). 704 For these tests, Venomix was run on the University of North Carolina at Charlotte 705 COPPERHEAD Research Computing Cluster with 944 Computing Cores, 8 TBs of memory, and 100 706 Gb/s Infiniband connectivity. The implementation of Venomix requires the scripting languages Python 707 2.7 (http://www.python.org/download/releases/2.7/) and R 3.1.1 (https://cran.r-project.org/bin/), in 708 addition to other Biopython packages (Cock et al., 2009) and data from ToxProt and Genbank that are 709 built into the Venomix pipeline (https://bitbucket.org/JasonMacrander/Venomix). We included versions 710 of MAFFT (Katoh & Standley, 2013), NCBI BLAST+ (Camacho et al., 2009), and Transdecoder (Haas et 711 al., 2013) that can be run locally. Although there are two versions of MAFFT (64 bit and 32 bit), the 712 default is the 64-bit, as this is more common for computers used in bioinformatic analyses. Modification 713 to the version of MAFFT in the Venomix pipeline can be done in the support files/Alignment.py file. 714 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 715 716 group directories (Figure 1). The outputs within each of the Toxin Group directories are intended to 717 provide the user with curated information to focus future investigations and analyses. Depending on the 718 next step of the analysis, some of the output files may be used for additional venom related downstream 719 applications or simply a quick reference for the user (see Discussion). Venomix also creates two ancillary 720 products that may be informative to some users: TPM. fasta (only transcripts with TPM values > 1.0) 721 and a large GenBank file with information from ToxProt BLAST hits in a format that may lend itself to 722 quick searches or downstream annotation (Figure 1). The user may choose to rerun Venomix with 723 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. 724 725 **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

728 outputs were compared for both search parameters in terms of the number of toxin groups, number of 729 transcripts, and number of "candidate" transcripts identified by the pipeline. A transcript was considered a 730 "candidate" if the transcript had significantly better e-value associated with a toxin than with a non-toxin 731 protein in Uniprot. Candidate transcripts were translated into their protein sequences using Transdecoder 732 (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 733 734 addition to screening the overall transcriptome analyses, some toxin groups and candidate toxins 735 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) 736 737 (Katoh & Standley, 2013) and visualized in Geneious (Kearse et al., 2012). Toxin gene tree 738 reconstructions were done in Fasttree v2 (Price, Dehal & Arkin, 2010) using maximum likelihood tree 739 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. 740 741 Trinity), in addition to assessing both transcriptomes for overall completeness in BUSCO (Simão et al., 742 2015). Expression values for each transcriptome were calculated using RSEM (Li & Dewey, 2011) rather 743 than raw read counts as originally proposed by Bouzid et al (Bouzid et al., 2014). 744 3. Results

745 **3.1 Transcriptome Assemblies**

Transcriptomes for each species previously assembled in Trinity (Grabherr et al., 2011) resulted 746 747 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 748 encountered during their initial assemblies (Phuong, Mahardika & Alfaro, 2016). The transcriptome for T. 749 750 bicarinatum was originally done using Velvet/Oases (Li & Durbin, 2009), however, we compared this to 751 our Trinity assembly because of its ease of use (Sanders et al., 2018) and frequency in the venom 752 literature (Macrander, Broe & Daly, 2015), in addition to a lower redundancy and chimera rate (Yang & Smith, 2013). 753

754 **3.2 Pipeline Output**

755 In the original published annotations, species-specific transcriptomes of C. sponsalis, E. colaratus 756 and U. yaschenkoi were not subjected to any BLAST searches using ToxProt, but instead were screened 757 using taxonomic specific toxin datasets from venom proteins of closely related species. The Venomix 758 pipeline recovered the majority of these lineage specific toxins and additional transcripts that resemble 759 toxin genes from other taxa. It is worth noting that if there were lineage specific toxins that shared high 760 sequence similarity to other toxins, the toxin group name may be assigned an incorrect lineage 761 classification, yet remain a toxin candidate. For example, analyses of the scorpion U. yaschenkoi resulted 762 in four venom groups with "Snake venom" in the name, however, in these instances, lineage specific 763 toxin names are often members of larger gene families that may not be lineage specific. The number of 764 identified toxin groups varied considerably across species and stringency parameters (Table 2), with the 765 less stringent parameters dramatically increasing the number of toxin groups. Each species had multiple 766 toxin groups that were separated based on sequence similarity and that correspond to large gene families, 767 including astacin-like metalloproteases, conotoxins, phospholipases A2s, Cysteine-rich secretory proteins 768 (CRISPs), Kunitz-type serine protease inhibitors, snaclecs, metalloproteinases, thrombin-like enzymes, 769 and allergens. For each species, there was at least one toxin group that was not retained following the 770 reciprocal BLASTp search after the toxin-like transcripts were translated into their open reading frame 771 (Table 2).

772 **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,

780 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 781 782 recovered more candidate conotoxins than the original study (O1, conodipine, and conophysin 783 superfamilies), however, for the majority of the conotoxin superfamilies identified in the original study 784 (Phuong, Mahardika & Alfaro, 2016) were missing from our analysis (Table 3). This discrepancy is likely 785 due to the different approaches used in the initial transcriptome assemblies, as iterative assemblies used 786 by Phuong et al. (2016) were unable to recover known transcripts using Trinity alone. Beyond the 787 conotoxins, there were many candidate toxin genes found within the Kunitz-type conkuitzin-S1 group (n 788 = 19), which included characterized toxin proteins from the venom Kunitz-type family of sea anemones, 789 cone snails, and snakes.

790 **3.4** Venomix outputs for *E. coloratus*

791 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. 792 793 (2014b). Among the transcripts identified, 45 had a TPM value greater than 100, with 39 of these in the 794 venom gland, four in the scent gland, and two in the skin. The most abundant transcript was actually in 795 the scent gland and identified as a cathelicidin-related peptide (Supplemental Table 1). The majority of 796 the highly expressed transcripts in the venom gland (TPM > 100) corresponded with toxin groups 797 previously identified (Hargreaves et al., 2014b), comprising mostly of C-type lectins, cysteine rich venom 798 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

806	high levels from the latroinsectotoxin gene family, which was previously characterized in spiders
807	(Magazanik et al., 1992). Reciprocal BLAST searches against the entire UNIPROT dataset revealed that
808	two of these were ankyrin rich peptides, and the other a dysferlin-like protein and may not be toxins.
809	3.5 Venomix outputs for T. bicarinatum
810	Of the four datasets included in our analysis, T. bicarinatum was the only transcriptome which
811	originally used the ToxProt dataset for toxin identification. The original transcriptome assembly was done
812	in Velvet/Oases (Li & Durbin, 2009), resulting in very different transcriptome assembly outputs (Bouzid
813	et al., 2014). Despite the alternative approaches, the original assembly resulted in a higher Busco (Simão
814	et al., 2015) score for the Velvet/Oases assembly (95.9%) when compared to our Trinity assembly
815	(92.2%). Interestingly, when considering TPM (rather than raw counts) the number of candidate genes in
816	the venom gland following the approach by Bouzid et al. (2014) was similar to what was originally
817	published (Table 2). Among these 527 candidates, only 44 predicted ORFs from Transdecoder, and only
818	three of these were given a score of 1 or greater in Toxclassifier (Table 2). The BLAST screening,
819	however, resulted in 62 of candidate toxins identified when the E-value threshold was set to 1E-3, but 287
820	when the E-value threshold was set to 10. As E-values were not specified by Bouzid et al. (Bouzid et al.,
821	2014) both are reported here.
822	For the Venomix analysis or our Trinity assemblies, there were 75 and 36 toxin groups for the less
823	stringent (E-value = E-06) and more stringent (E-value = E-20) analyses, respectively. In the less stringent
824	analysis, the largest number candidate toxin genes corresponded to the alpha-latroinsectotoxinLt1a group
825	(N = 280), but overall expression within the ant carcass and venom gland across these transcripts were
826	approximately the same in both the ant carcass and venom gland. Among those more highly expressed in
827	the venom gland, there were six transcripts expressed at TPM values greater than 100 in the venom gland,
828	four corresponding to venom allergen 3 and two to cysteine-rich venom protein Mr30. Upon closer
829	examination, BLAST searches against UNIPROT indicated that all six of these toxins are likely Venom
830	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 833 834 that was previously published by Bouzid et al (2014), we recovered some unexpected results. Although 835 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 836 837 these 33, only 10 were '1000 fold' higher based on expected count values (Supplemental Table 2). 838 Although not all 69 candidate toxin sequences identified by Bouzid et al (Bouzid et al., 2014) were 839 confidently called toxins, all but one was reported as having a higher number of raw read counts in the 840 venom gland transcriptome as opposed to the ant carcass. The reason behind the larger than 50% 841 discrepancy between our analyses and theirs remains unknown, as alternative approaches to quantifying 842 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 843 844 Velvet/Oases assembly with a TPM difference of 100 or greater in the venom gland transcriptome when 845 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 704 846 nucleotides, while the other 8688 transcripts identified in Venomix had an average length of 291 847 848 nucleotides. It is likely that incomplete transcripts for other proteins were recovered here as protein 849 sequences were 407 amino acids (1221 nucleotides) long, which corresponded to the identified transcripts 850 in our analysis. Incomplete transcripts are likely the reason behind this as only 47 of the 8688 transcripts 851 identified in Venomix were translated into their open reading frame.

852

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

857 (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 858 859 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 860 861 2A). These differences may be due to changes in assembly algorithms between the two analyses, as 862 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 863 approximately the same as that identified by Luna-Ramírez et al. (2015). Beyond the exclusive scorpion 864 865 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 866 867 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 868 869 within the delta-latroinsectotoxin-Lt1a group (n = 190), the alpha latroinsectotoxin-Lt1a group (n = 182), 870 and the Neprilysin-1 group (n = 91). Under the more stringent search criteria, these are still the largest 871 toxin groups, however, the most abundant group was Neprilysin-1 group (n = 69), followed by alpha 872 latroinsectotoxin-Lt1a group (n = 51) and delta-latroinsectotoxin-Lt1a group (n = 49). Query toxins which 873 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 874 875 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 876 877 UNIPROT had better E-values than the genes used as queries in Uniprot. Among the neprilysin 878 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 879 880 toxin sequences. Proteins from the less stringent analysis (E-value >1E-6) were used to construct gene 881 trees for neprilysins using proteins of known venomous function along aside other proteins that are 882 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).
- 885
- 886 4. Discussion

887 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 888 889 characterize toxin gene candidates within a transcriptomic dataset. The outputs provided by this pipeline 890 give necessary information for further evaluation of their toxin gene candidates. We recommend using 891 Venomix across multiple BLAST searches with varying E-value thresholds, as the variation among 892 characterized toxin genes and those of the focal taxa may be more accommodating depending on the 893 threshold used. Although Venomix was able to identify more candidate toxin genes in three out of the 894 four datasets tested here, these results require further examination to determine which transcripts are 895 viable toxin gene candidates. Venomix is not meant to be a definitive toxin gene identifier because this 896 determination should not be made by sequence data alone, especially for poorly studied lineages.

897 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 898 ToxProt dataset (Figure 3), whereas the ant venom is poorly characterized among the diverse venomous 899 900 insects found within Hexapoda on ToxProt. Additionally, these datasets represent diverse transcriptome 901 assembly methods, query datasets, and gene expression quantification approaches. The original C. 902 sponsalis assembly had a high number of toxin genes with relatively low variation across gene copies 903 (Phuong, Mahardika & Alfaro, 2016), which likely resulted in many of these being clumped together in 904 our Trinity assembly (Macrander, Broe & Daly, 2015). To get around this issue, Phuong et al (Phuong, 905 Mahardika & Alfaro, 2016) did three assembly iterations involving toxin gene identification and 906 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, 907 908 Auckland, New Zealand). In contrast to this, it is likely the differences we observed for T. bicarinatum

909 using Venomix was due to the alternative transcriptome assembly and gene expression approach (Yang & 910 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, 911 912 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 913 914 and extract candidate toxin sequences. Specifically, our pipeline also will translate transcripts into their 915 predicted protein, screen for signaling regions, assess their similarity through alignment and gene trees, 916 extract expression information, and refer to taxonomic and other information available in the query 917 sequence GenBank entries. This will allow venom biologists to quickly move onto additional downstream 918 identification and characterization of toxin gene diversity using outputs provided by Venomix. 919 Additionally, Venomix is the first pipeline to provide all these outputs in an easy to use search strategy 920 that is flexible, but also repeatable, for all venomous taxa, or non-venomous animal to be used in a tissue 921 specific comparative context (Reumont et al., 2014; Hargreaves et al., 2014b; Reves-Velasco et al., 2015). 922 The pipeline also provides users with easy to navigate directories and organized output files, allowing the 923 user to sort manually or quickly pull information for all toxin groups using simple unix commands (i.e., 924 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 925 testing future hypotheses of venom diversity and tissue specific expression. The E. coloratus 926 927 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 928 929 combination with toxin gene tree reconstruction to determine which of the approximately 16 venom toxin 930 gene families that occur across Toxicofera attribute to the *E. coloratus* venom transcriptome. Of these 931 which are venom candidates in *E. coloratus* a comparison of the Venomix output containing expression 932 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 933 934 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 otherneprilysin toxins in a reciprocal blast hit.

937 Regardless of the bioinformatic approach to identifying toxin genes, one major hurdle using these 938 types of datasets as query sequences is the limited taxonomic diversity present in the ToxProt dataset. 939 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 940 941 reflects the abundance of snake proteins deposited into ToxProt and is in contrast to the paucity of 942 proteins for other, poorly studied venomous lineages (Figure 3). Additionally, Venomix "group" names 943 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. 944 *bicarinatum*, however, the transcripts associated with this appeared to be neurophysins. This was also 945 946 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 947 948 Venom Allergen 3 genes. When investigating venom diversity for poorly studied taxa, caution is 949 warranted in using these gene names because the specific classifiers of the Venomix outputs provide a 950 starting point for toxin gene identification, but does not act as a distinct classification system. 951 In every transcriptome, the machine learning program ToxClassifier failed to recover all of the

952 toxins identified in their respective publications (Table 2). Our downstream analysis of the protein 953 sequences produced by Transdecoder included any candidate toxin with a score > 1. Even though some 954 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 955 956 ToxClassifier and Venomix is that our pipeline is not meant to be a toxin gene identifier. Venomix was 957 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 958 959 building methods are rudimentary and meant to be used for only initial screenings. This allows users to 960 focus their efforts on downstream analyses using complementary proteomics and machine learning to

NOT PEER-REVIEWED

Peer Preprints

961 differentiate between functionally toxic and non-toxic venom components (Gacesa, Barlow & Long,
962 2016) or to complement their transcriptomic data with functional assays of proteins or crude venom
963 extracts.

964 5. Conclusion

965 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 966 967 versatile in that it can accommodate transcriptomic datasets for a variety of species and can quickly 968 identify a large number of toxin gene candidates from venom gland or other tissue specific 969 transcriptomes. Overall, Venomix addresses three shortcomings encountered in similar approaches: (1) it 970 is reproducible, (2) it does not claim to be a toxin gene identifiers as other programs or pipelines do that 971 appear to be less reliable, and (3) it is able to accommodate a wide variety of taxa. Because of its ease of 972 use and ability to quickly identify toxin gene candidates, researchers can move past the tedious and time 973 consuming stages of toxin candidate identification and move onto toxin gene characterization.

974

975 Acknowledgments

976 Venomix is a direct byproduct of the training received at Friday Harbor Marine Labs while 977 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 978 979 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 980 981 enthusiasm for Venomix. The authors would also like to thank Edwin Rice for preliminary input on 982 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 983 984 MD and JM, and NSF OCE 1536530 to AMR. JP was supported by a research assistantship provided by 985 the UNCC Graduate School at UNCC under the guidance of DJ.

986 6. References

987	Barghi N., Concepcion GP., Olivera BM., Lluisma AO. 2015. High conopeptide diversity in Conus
988	tribblei revealed through analysis of venom duct transcriptome using two high-throughput
989	sequencing platforms. <i>Marine Biotechnology</i> 17:81–98. DOI: 10.1007/s10126-014-9595-7.
990	Bhere KV., Haney RA., Ayoub NA., Garb JE. 2014. Gene structure, regulatory control, and evolution of
991	black widow venom latrotoxins. <i>FEBS Letters</i> 588:3891–3897. DOI:
992	10.1016/j.febslet.2014.08.034.
993	Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data.
994	Bioinformatics 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.
995	Bouzid W., Verdenaud M., Klopp C., Ducancel F., Noirot C., Vétillard A. 2014. De Novo sequencing and
996	transcriptome analysis for Tetramorium bicarinatum: a comprehensive venom gland
997	transcriptome analysis from an ant species. BMC Genomics 15:987. DOI: 10.1186/1471-2164-
998	15987.
999	Buczek O., Bulaj G., Olivera BM. 2005. Conotoxins and the posttranslational modification of secreted
1000	gene products. Cellular and Molecular Life Sciences CMLS 62:3067–3079. DOI:
1001	10.1007/s00018-005-5283-0.
1002	Calvete JJ., Fasoli E., Sanz L., Boschetti E., Righetti PG. 2009. Exploring the venom proteome of the
1003	western diamondback rattlesnake, Crotalus atrox, via snake venomics and combinatorial peptide
1004	ligand library approaches. Journal of Proteome Research 8:3055–3067. DOI:
1005	10.1021/pr900249q.
1006	Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+:
1007	architecture and applications. BMC Bioinformatics 10:421. DOI: 10.1186/1471-210510-421.
1008	Casewell NR., Huttley GA., Wüster W. 2012. Dynamic evolution of venom proteins in squamate reptiles.
1009	Nature Communications 3:1066. DOI: 10.1038/ncomms2065.
1010	Casewell NR., Wüster W., Vonk FJ., Harrison RA., Fry BG. 2013. Complex cocktails: the evolutionary
1011	novelty of venoms. Trends in Ecology & Evolution 28:219-229. DOI:
1012	10.1016/j.tree.2012.10.020.
1013	Cock PJA., Antao T., Chang JT., Chapman BA., Cox CJ., Dalke A., Friedberg I., Hamelryck T., Kauff F.,
1014	Wilczynski B., Hoon MJL de. 2009. Biopython: freely available Python tools for computational
1015	molecular biology and bioinformatics. Bioinformatics 25:1422-1423. DOI:
1016	10.1093/bioinformatics/btp163.
1017	Fan Y-X., Song J., Shen H-B., Kong X. 2011. PredCSF: an integrated feature-based approach for
1018	predicting conotoxin superfamily. Protein and Peptide Letters 18:261–267.
1019	Fry BG., Roelants K., Champagne DE., Scheib H., Tyndall JDA., King GF., Nevalainen TJ., Norman JA.,
1020	Lewis RJ., Norton RS., Renjifo C., Vega RCR de la. 2009. The toxicogenomic multiverse:
1021	convergent recruitment of proteins into animal venoms. Annual Review of Genomics and Human
1022	Genetics 10:483-511. DOI: 10.1146/annurev.genom.9.081307.164356.
1023	Gacesa R., Barlow DJ., Long PF. 2016. Machine learning can differentiate venom toxins from other
1024	proteins having non-toxic physiological functions. PeerJ Computer Science 2:e90. DOI:
1025	10.7717/peerj-cs.90.
1026	Garb JE., Hayashi CY. 2013. Molecular evolution of α-latrotoxin, the exceptionally potent vertebrate
1027	neurotoxin in black widow spider venom. Molecular Biology and Evolution 30:999–1014. DOI:
1028	10.1093/molbev/mst011.
1029	Grabherr MG., Haas BJ., Yassour M., Levin JZ., Thompson DA., Amit I., Adiconis X., Fan L.,
	· · · · · · · · · · · · · · · · · · ·

1020	Development D. Zang O. Chan Z. Margali E. Harakan N. Caista A. Dhind N. J. Dalara E.
1030	Raycnowdnury R., Zeng Q., Chen Z., Maucen E., Haconen N., Gnirke A., Knind N., di Palma F.,
1031	binen Bw., Nusbaum C., Lindblad-101 K., Friedman N., Regev A. 2011. Full-tengui
1032	20.644 652 DOL 10.1028/she 1882
1033	29:044-052. DOI: 10.1038/nbt.1883.
1034	Graudins A., Little MJ., Pineda SS., Hains PG., King GF., Broady KW., Nicholson GM. 2012. Cloning
1035	and activity of a novel α -latrotoxin from red-back spider venom. <i>Biochemical Pharmacology</i>
1036	83:170–183. DOI: 10.1016/j.bcp.2011.09.024.
1037	Gupta S., Kapoor P., Chaudhary K., Gautam A., Kumar R., Consortium OSDD., Raghava GPS. 2013. In
1038	silico approach for predicting toxicity of peptides and proteins. <i>PLOS ONE</i> 8:e/395/. DOI:
1039	10.13/1/journal.pone.00/395/.
1040	Haas BJ., Papanicolaou A., Yassour M., Grabherr M., Blood PD., Bowden J., Couger MB., Eccles D., Li
1041	B., Lieber M., MacManes MD., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R.,
1042	William T., Dewey CN., Henschel R., LeDuc RD., Friedman N., Regev A. 2013. De novo
1043	transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity.
1044	Nature protocols 8. DOI: 10.1038/nprot.2013.084.
1045	Hargreaves AD., Swain MT., Hegarty MJ., Logan DW., Mulley JF. 2014a. Restriction and recruitment—
1046	gene duplication and the origin and evolution of snake venom toxins. <i>Genome Biology and</i>
1047	<i>Evolution</i> 6:2088–2095. DOI: 10.1093/gbe/evu166.
1048	Hargreaves AD., Swain MT., Logan DW., Mulley JF. 2014b. Testing the toxicofera: comparative
1049	transcriptomics casts doubt on the single, early evolution of the reptile venom system. <i>Toxicon</i>
1050	92:140–156. DOI: 10.1016/j.tox1con.2014.10.004.
1051	Jungo F., Bougueleret L., Xenarios I., Poux S. 2012. The UniProtKB/Swiss-Prot Tox-Prot program: A
1052	central hub of integrated venom protein data. <i>Toxicon</i> 60:551–557. DOI:
1053	10.1016/j.tox1con.2012.03.010.
1054	Junqueira-de-Azevedo ILM., Bastos CMV., Ho PL., Luna MS., Yamanouye N., Casewell NR. 2015.
1055	Venom-related transcripts from <i>Bothrops jararaca</i> tissues provide novel molecular insights into
1056	the production and evolution of snake venom. <i>Molecular Biology and Evolution</i> 32:754–766.
1057	DOI: 10.1093/molbev/msu337.
1058	Kaas Q., Craik DJ. 2015. Bioinformatics-aided venomics. <i>Toxins</i> 7:2159–2187. DOI:
1059	10.3390/toxins/062159.
1060	Kaas Q., Yu R., Jin A-H., Dutertre S., Craik DJ. 2012. ConoServer: updated content, knowledge, and
1061	discovery tools in the conopeptide database. <i>Nucleic Acids Research</i> 40:D325–D330. DOI:
1062	10.1093/nar/gkr886.
1063	Kaplan N., Morpurgo N., Linial M. 2007. Novel families of toxin-like peptides in insects and mammals:
1064	A computational approach. <i>Journal of Molecular Biology</i> 369:553–566. DOI:
1065	10.1016/j.jmb.2007.02.106.
1066	Katoh K., Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: Improvements
1067	in performance and usability. <i>Molecular Biology and Evolution</i> 30:772–780. DOI:
1068	10.1093/molbev/mst010.
1069	Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A.,
1070	Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012. Geneious
1071	Basic: An integrated and extendable desktop software platform for the organization and analysis
10/2	ot sequence data. <i>Bioinformatics</i> 28:164/–1649. DOI: 10.1093/bioinformatics/bts199.
10/3	Koziov S., Grishin E. 2011. The mining of toxin-like polypeptides from EST database by single residue
1074	distribution analysis. BMC Genomics 12:88. DOI: 10.1186/14/1-2164-12-88.

1075 1076 1077 1078	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 <i>Alatina alata</i> : an emerging model for studying venom, vision and sex. <i>BMC Genomics</i> 17. DOI: 10.1186/s12864-016-2944- 3.
1079 1080	Li B., Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. <i>BMC Bioinformatics</i> 12:323. DOI: 10.1186/1471-2105-12-323.
1081 1082	Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. <i>Bioinformatics</i> 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.
1083	Luna-Ramírez K., Quintero-Hernández V., Juárez-González VR., Possani LD. 2015. Whole transcriptome
1084	of the venom gland from Urodacus yaschenkoi scorpion. PLOS ONE 10:e0127883. DOI:
1085	10.1371/journal.pone.0127883.
1086	Macrander J., Broe M., Daly M. 2015. Multi-copy venom genes hidden in de novo transcriptome
1087	assemblies, a cautionary tale with the snakelocks sea anemone Anemonia sulcata (Pennant, 1977).
1088	Toxicon 108:184–188. DOI: 10.1016/j.toxicon.2015.09.038.
1089	Macrander J., Broe M., Daly M. 2016. Tissue-specific venom composition and differential gene
1090	expression in sea anemones. Genome Biology and Evolution:evw155. DOI: 10.1093/gbe/evw155.
1091	Macrander J., Brugler MR., Daly M. 2015. A RNA-seq approach to identify putative toxins from
1092	acrorhagi in aggressive and non-aggressive Anthopleura elegantissima polyps. BMC Genomics
1093	16:221. DOI: 10.1186/s12864-015-1417-4.
1094	Magazanik LG., Fedorova IM., Kovalevskaya GI., Pashkov VN., Bulgakov OV., Grishin EV. 1992.
1095	Selective presynaptic insectotoxin (a-latroinsectotoxin) isolated from black widow spider venom.
1096	Neuroscience 46:181–188. DOI: 10.1016/0306-4522(92)90017-V.
1097	Ménez A., Stöcklin R., Mebs D. 2006. 'Venomics' or: The venomous systems genome project. Toxicon
1098	47:255–259. DOI: 10.1016/j.toxicon.2005.12.010.
1099	Neumann RS., Kumar S., Shalchian-Tabrizi K. 2014. BLAST output visualization in the new sequencing
1100	era. Briefings in Bioinformatics 15:484–503. DOI: 10.1093/bib/bbt009.
1101	OmPraba G., Chapeaurouge A., Doley R., Devi KR., Padmanaban P., Venkatraman C., Velmurugan D.,
1102	Lin Q., Kini RM. 2010. Identification of a novel family of snake venom proteins veficolins from
1103	Cerberus rynchops using a venom gland transcriptomics and proteomics approach. Journal of
1104	Proteome Research 9:1882–1893. DOI: 10.1021/pr901044x.
1105	Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language.
1106	Bioinformatics 20:289–290. DOI: 10.1093/bioinformatics/btg412.
1107	Petersen TN., Brunak S., von Heijne G., Nielsen H. 2011. SignalP 4.0: discriminating signal peptides
1108	from transmembrane regions. <i>Nature Methods</i> 8:785–786. DOI: 10.1038/nmeth.1701.
1109	Phuong MA., Mahardika GN., Alfaro ME. 2016. Dietary breadth is positively correlated with venom
1110	complexity in cone snails. BMC Genomics 17:401. DOI: 10.1186/s12864-016-2755-6.
1111	Prashanth JR., Lewis RJ. 2015. An efficient transcriptome analysis pipeline to accelerate venom peptide
1112	discovery and characterization. <i>Toxicon</i> 107, Part B:282–289. DOI:
1113	10.1016/j.toxicon.2015.09.012.
1114	Price MN., Dehal PS., Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large
1115	alignments. PLoS ONE 5. DOI: 10.1371/journal.pone.0009490.
1116	Reumont BM von., Blanke A., Richter S., Alvarez F., Bleidorn C., Jenner RA. 2014. The first venomous
1117	crustacean revealed by transcriptomics and functional morphology: remipede venom glands
1118	express a unique toxin cocktail dominated by enzymes and a neurotoxin. <i>Molecular Biology and</i>
1119	Evolution 31:48–58. DOI: 10.1093/molbev/mst199.

1120	Reyes-Velasco J., Card DC., Andrew AL., Shaney KJ., Adams RH., Schield DR., Casewell NR.,
1121	Mackessy SP., Castoe TA. 2015a. Expression of venom gene homologs in diverse python tissues
1122	suggests a new model for the evolution of snake venom. Molecular Biology and Evolution
1123	32:173-183. DOI: 10.1093/molbev/msu294.
1124	Rodríguez de la Vega RC., Giraud T. 2016. Intragenome diversity of gene families encoding toxin-like
1125	proteins in venomous animals. Integrative and Comparative Biology:icw097. DOI:
1126	10.1093/icb/icw097.
1127	Sanders S., Ganote C., Papudeshi B., Mockaitis K., Doak T. 2018. NCGAS makes robust transcriptome
1128	analysis easier with a readily usable workflow following de novo assembly best practices.
1129	Simão FA., Waterhouse RM., Ioannidis P., Kriventseva EV., Zdobnov EM. 2015. BUSCO: assessing
1130	genome assembly and annotation completeness with single-copy orthologs. Bioinformatics
1131	31:3210-3212. DOI: 10.1093/bioinformatics/btv351.
1132	Sunagar K., Morgenstern D., Reitzel AM., Moran Y. 2016. Ecological venomics: How genomics,
1133	transcriptomics and proteomics can shed new light on the ecology and evolution of venom.
1134	Journal of Proteomics 135:62-72. DOI: 10.1016/j.jprot.2015.09.015.
1135	Tan PTJ., Khan AM., Brusic V. 2003. Bioinformatics for venom and toxin sciences. Briefings in
1136	Bioinformatics 4:53-62. DOI: 10.1093/bib/4.1.53.
1137	Terrat Y., Ducancel F. 2013. Are there unequivocal criteria to label a given protein as a toxin? Permissive
1138	versus conservative annotation processes. Genome Biology 14:406. DOI: 10.1186/gb-2013-14-
1139	9406.
1140	Todd EV., Black MA., Gemmell NJ. 2016. The power and promise of RNA-seq in ecology and evolution.
1141	Molecular Ecology 25:1224–1241. DOI: 10.1111/mec.13526.
1142	Undheim EAB., Sunagar K., Herzig V., Kely L., Low DHW., Jackson TNW., Jones A., Kurniawan N.,
1143	King GF., Ali SA., Antunes A., Ruder T., Fry BG. 2013. A Proteomics and transcriptomics
1144	investigation of the venom from the barychelid spider Trittame loki (brush-foot trapdoor). Toxins
1145	5:2488-2503. DOI: 10.3390/toxins5122488.
1146	Vijay N., Poelstra JW., Künstner A., Wolf JBW. 2013. Challenges and strategies in transcriptome
1147	assembly and differential gene expression quantification. A comprehensive in silico assessment of
1148	RNA-seq experiments. Molecular Ecology 22:620-634. DOI: 10.1111/mec.12014.
1149	Vonk FJ., Casewell NR., Henkel CV., Heimberg AM., Jansen HJ., McCleary RJR., Kerkkamp HME.,
1150	Vos RA., Guerreiro I., Calvete JJ., Wüster W., Woods AE., Logan JM., Harrison RA., Castoe
1151	TA., Koning APJ de., Pollock DD., Yandell M., Calderon D., Renjifo C., Currier RB., Salgado D.,
1152	Pla D., Sanz L., Hyder AS., Ribeiro JMC., Arntzen JW., Thillart GEEJM van den., Boetzer M.,
1153	Pirovano W., Dirks RP., Spaink HP., Duboule D., McGlinn E., Kini RM., Richardson MK.
1154	2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom
1155	system. Proceedings of the National Academy of Sciences 110:20651–20656. DOI:
1156	10.1073/pnas.1314702110.
1157	Wong ESW., Belov K. 2012. Venom evolution through gene duplications. Gene 496:1–7. DOI:
1158	10.1016/j.gene.2012.01.009.
1159	Wong ESW., Hardy MC., Wood D., Bailey T., King GF. 2013. SVM-based prediction of propeptide
1160	cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. PLOS ONE
1161	8:e66279. DOI: 10.1371/journal.pone.0066279.
1162	Yang Y., Smith SA. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics.
1163	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

	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					

1 Table 1 . Species specific information for transcriptome as	assemblies.
--	-------------

2 3

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, \dagger = 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.

1 1			1	0						
	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°	780†	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. vaschenkoi	111	11s	210	71(6)	50	277	48(34)	117	689	179(38)

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, \dagger = 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.

2

1

Table 3(on next page)

Table 3. Comparison of Conotoxin Transcripts for C. sponsalis

	 -

1 2

Table 3. Comparison of Conotoxin Transcripts for C. sponsalis

Super Family	Venomix	Phuong et al. (2016)
А	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
М	29	29
MEFRRr	0	3
MKFLLr	0	2
MKISL*	1	1
Ν	0	8
O1	108	95
02	12	28
O3	1	6
Р	0	13
Q	0	1
SF-04	0	1
SF-mi1	0	6
SF-mi2	0	3
Т	3	56
U	0	6
V	0	3
Y	0	3
Total	179	401

3

Table 4(on next page)

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

1

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

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

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.



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 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)

DN100002

DN100003

(tab delimited

gl il 247

- 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.
- sequence.gp ToxProt dataset in GenBank format.

Ancillary Products

TPM.fasta
Transcriptome sequences with a $TPM \ge 1.0$ 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 Gene Tree GenBank SignalP Expression Taxa from Protein sequence alignments information Values for of Toxprot output for Toxprot Unedited Transcripts and toxinfrom transcripts Sequences translated • Protein sequences (ToxProt, like protein ToxProt transcripts Transdecoder output, and Signal sequences Sequences and Toxprot P output) 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.



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.



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