Briefings in Bioinformatics, 00(00), 2021, 1–16

doi: 10.1093/bib/bbab095
Problem Solving Protocol

ToxCodAn: a new toxin annotator and guide to venom gland transcriptomics

Pedro G. Nachtigall†, Rhett M. Rautsaw†, Schyler A. Ellsworth, Andrew J. Mason, Darin R. Rokyta, Christopher L. Parkinson and Inácio L.M. Junqueira-de-Azevedo

Corresponding author: Pedro G. Nachtigall, Laboratório de Toxinologia Aplicada, CeTICS, Instituto Butantan, São Paulo, SP 05503-900, Brazil. Tel: +5514981744000; E-mail: pedronachtigall@gmail.com
†These authors contributed equally to this work.

Abstract

Motivation: Next-generation sequencing has become exceedingly common and has transformed our ability to explore nonmodel systems. In particular, transcriptomics has facilitated the study of venom and evolution of toxins in venomous lineages; however, many challenges remain. Primarily, annotation of toxins in the transcriptome is a laborious and time-consuming task. Current annotation software often fails to predict the correct coding sequence and overestimates the number of toxins present in the transcriptome. Here, we present ToxCodAn, a python script designed to perform precise annotation of snake venom gland transcriptomes. We test ToxCodAn with a set of previously curated transcriptomes and compare the results to other annotators. In addition, we provide a guide for venom gland transcriptomics to facilitate future research and use *Bothrops alternatus* as a case study for ToxCodAn and our guide. **Results:** Our analysis reveals that ToxCodAn provides precise annotation of toxins present in the transcriptome of venom glands of snakes. Comparison with other annotators demonstrates that ToxCodAn has better performance with regard to run time (> 20x faster), coding sequence prediction (> 3x more accurate) and the number of toxins predicted (generating > 4x less false positives). In this sense, ToxCodAn is a valuable resource for toxin annotation. The ToxCodAn framework can be expanded in the future to work with other venomous lineages and detect novel toxins. **Supplementary Data:** Supplementary data are available online at https://academic.oup.com/bib.

Pedro G. Nachtigall is a biologist and received his PhD in Genetics at UNESP (Universidade Estadual Paulista, Brazil) in the year 2017. Currently, he is a postdoctoral researcher at Instituto Butantan, São Paulo, Brazil. His research focuses on comparative genomics, transcriptomics and evolutionary biology.
Rhett M. Rautsaw is currently a PhD student in biological sciences at Clemson University. He received his MS from the University of Central Florida in 2017. His research focuses on the integration of ecology, evolution and genetics to understand trait evolution.
Schyler A. Ellsworth is currently a PhD candidate in ecology and evolution at Florida State University. His research focuses on characterizing venom complexity and investigating the evolution of venom resistance.
Andrew J. Mason is an evolutionary biologist and received his PhD in Biological Sciences from Clemson University in 2020. Currently, he is a postdoctoral researcher at The Ohio State University. His research focuses on comparative genomics and evolutionary biology.
Darin R. Rokyta is an evolutionary biologist and received his PhD in Bioinformatics and Computational Biology from the University of Idaho in 2006. Currently, he is a professor at Florida State University. His research focuses on adaptation and molecular evolution in complex protein systems such as viruses and venoms.
Christopher L. Parkinson is an evolutionary biologist and received his PhD in Environmental Biology from the University of Louisville in 1996. Currently, he is a professor at Clemson University. His research focuses on phylogenetics, biogeography and the evolution of venom in snakes.
Inácio L.M. Junqueira-de-Azevedo is a biologist and received his PhD in Genetics and Evolutionary Biology at São Paulo University, Brazil, in 2002. Currently, he is head of the Applied Toxinology Laboratory at Instituto Butantan, São Paulo, Brazil. His research focuses on venoms, genomics and protein family evolution.

Submitted: 5 January 2021; **Received (in revised form):** 15 February 2021

Introduction

Next-generation sequencing (NGS) and the 'omics' era has provided a cost-effective way to collect vast amounts of molecular data, transforming the field of evolutionary biology and genetics [1]. In particular, transcriptomics (RNA-Seq) has revolutionized our ability to quantify the expression of genes, identify differentially expressed genes and simultaneously explore sequence variation within transcripts [2]. However, the impact of NGS technologies is felt greatest in nonmodel systems where high-throughput sequencing has begun to close the gap with model systems [3]. For example, in venomous lineages such as snakes, NGS and advances in proteomics have given rise to the field of 'venomics', which seeks to inventory and explore the evolutionary history of toxins in venomous lineages as well as ensure effective production of antivenoms [4–8].

Venom is a complex cocktail of proteins and peptides that has evolved independently in many animal lineages and is used for prey capture and predator defense [6, 9, 10]. The proteins and peptides in venom are a result of many genes working in concert to produce a toxic function; however, unlike many other polygenic traits, venom is the result of a relatively direct pathway from transcription to translation [11–13]. Therefore, venom is easily characterized and nearly congruent at the transcriptomic and proteomic levels [11–13]. Transcriptomics has become very common for characterizing venom composition because it requires less starting material than standard proteomic methods, facilitating studies on small and less traditionally recognized venomous taxa such as rear-fanged snakes [7, 14–16]. This further facilitates the discovery of novel toxins and potential therapeutic compounds [15]. However, venom gland transcriptomics still faces several challenges.

First, assembling the complete transcriptome is nontrivial because traditional 'de Bruijn' graph assemblers can result in errors such as chimeric transcripts or fail to assemble common long toxin transcripts (e.g. snake venom metalloproteinases; SVMPs). This has been reviewed by Holding *et al.* [17] and can be resolved through the use of multiple different assemblers followed by a consolidation of their outputs to reduce redundancy. The second challenge is the identification and annotation of toxin transcripts. That is, identification of the true coding sequence (CDS) within an assembled contig and naming the CDS appropriately. Some computational tools have been designed to perform automated annotation of transcriptomes using CDS prediction, sequence similarity and domain composition (reviewed in [18]).

Among the designed tools for automatic annotation, Trinotate [19] is the most widely used. Trinotate performs a general transcriptome annotation (i.e. not toxin specific) by using TransDecoder [20] to identify the longest open reading frame (ORF) followed by BLAST [21] searches against several protein databases (proteinDBs), additional searches for conserved domains and signalP (SP) prediction. Another tool that uses a similar approach to Trinotate is Dammit [22], which was also designed to perform general transcriptome annotation of the assembly. Although Trinotate and Dammit are similar in their use of TransDecoder, they have unique annotation pipelines with Dammit using Conditional Reciprocal Best LAST [23] to learn an appropriate *e*-value cutoff for different transcript lengths. However, because toxins are frequently derived from duplicated genes that evolve rapidly, resulting in numerous paralogous transcripts coding for similar products, toxin annotation can be challenging. Recently, Macrander *et al.* [8] published Venomix, which is a computational tool designed specifically to perform automated toxin annotation of transcriptome assemblies from vertebrate and invertebrate venomous species. Venomix similarly relies on TransDecoder, but performs its BLAST searches against the ToxProt database [24] followed by SP prediction, protein sequence alignment and gene tree construction.

Despite the availability of such tools to perform automated transcriptome annotation, nearly all automatic annotation pipelines identify the longest ORF in a contig, which is often incorrect due to misidentification of the correct start codon [25]. In addition, these automated pipelines overestimate the number of toxins present in the transcriptome or genome [26–29]. Therefore, these pipelines require manual checks, which can be laborious and time consuming. Instead of using these pipelines, many studies have performed complete manual annotation of the venom gland transcriptome [30], which is similarly strenuous.

Here, we present ToxCodAn, a computational pipeline able to confidently characterize the venom components from a snake transcriptome assembly. We focus on snake toxin annotation; however, future releases of ToxCodAn could be expanded with appropriate training data to work with any venomous lineage or updated upon the discovery of novel toxins. ToxCodAn demonstrates better performance for correctly identifying and annotating full toxin CDS regions with faster run times than traditional and other toxin annotators. Alongside ToxCodAn, we provide a guide to venom gland transcriptomics, which walks through our recommended bioinformatics pipeline from raw data to expression quantification to facilitate future research. To demonstrate the utility of ToxCodAn and our guide, we analyze the venom gland transcriptome of the Urutu Lancehead (*Bothrops alternatus*) as a case study.

Methods

Algorithm implementation

ToxCodAn is implemented in Python (v3.5 or higher) and uses third-party tools to perform the automated analysis (Figure 1). The pipeline consists of a prefilter step, which performs a BLAST (v2.9 or higher) search against a toxin database (toxinDB). This database contains curated toxin protein sequences from Uniprot (https://www.uniprot.org/) and protein sequences from published [31–34] and unpublished transcriptome assemblies (see Training Sets section). The prefilter step generates two sets of sequences, the NonToxins, which have no hits against the toxinDB, and the Putative Toxins (PT), which have hits against the toxinDB. After the prefilter step, ToxCodAn performs a CDS prediction on the PT using CodAn (v1.0; [25]) with generalized Hidden Markov Models (gHMMs) designed specifically for different snake toxin families (see Training Sets section). Then, ToxCodAn screens all of the predicted CDSs; if a transcript has a CDS predicted by one or more toxin models, it will keep the prediction with the highest probability score as the true CDS by following the CodAn application [25].

After the prediction step, ToxCodAn performs a toxin filter step by searching the predicted CDSs against the toxinDB using BLAST to filter and annotate the toxins present in the predicted CDSs. This step generates two sets of sequences: the PT sequences, which have no CDS predicted or the predicted CDS has no hits against the toxinDB, and the Toxin (T) sequences, which have a predicted CDS and hits against the toxinDB. For PT, the CDS is predicted following three steps: (1) BLAST search against the toxinDB with low stringency; (2) iterate through all hits to detect which hits begin with a start codon (ATG) and

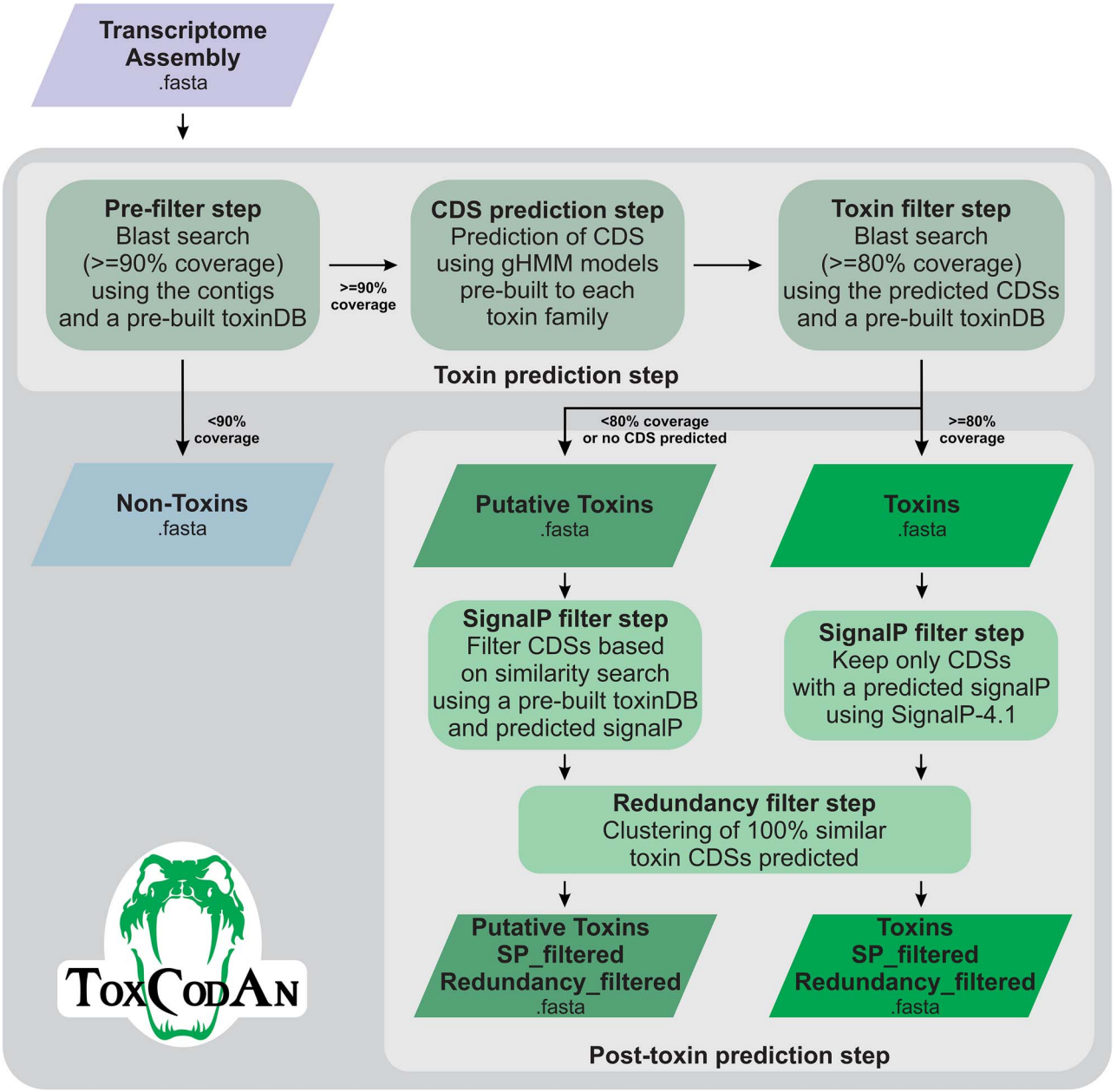

Figure 1. Flowchart of the ToxCodAn pipeline. The prefilter step performs a Blast search on the *de novo* transcriptome assembly. If there is no hit, then the contig gets placed into a NonToxins file, which can annotate using a different database. Contigs with hits move onto CDS prediction with CodAn and toxin family-specific generalized Hidden Markov Models. A final filter step separates toxins into Toxins and Putative Toxins, which can be further filtered.

have successful prediction of a signal peptide; and (3) retrieve the sequence from the position of the start codon to the position of the stop codon or to the end of the sequence. Finally, ToxCodAn performs signal peptide prediction by using SignalP (v4.1; setting the parameters -u 0.34 -U 0.34; [35]) to filter toxins without signal peptides and further removes redundancy by clustering the Toxins and PT sequences with 100% identity in size and sequence by using an in-house Python script. The full ToxCodAn pipeline can be seen in Figure 1.

Training sets

To perform identification of the CDS region for toxins in the transcriptome assembly, ToxCodAn uses gHMM specific to each toxin family present in snake venom. For this purpose, we designed the gHMMs using training sets containing curated sequences of toxin genes from previous published transcriptomic data [31, 34] and unpublished transcriptomic data graciously donated by collaborators. Unpublished transcriptomic data consisted of toxin transcripts that were manually curated, checked and cleaned following Hofmann *et al.* [30] in preparation for independent publications. In total, the training data contained 19 337 full-length transcripts with annotated CDS regions from 56 different species belonging to the Viperidae, Elapidae, Colubridae and Dipsadidae clades. The training set of each toxin family contains a mix of sequences from these snake species. We estimated the parameters on each toxin family model by using the ToPS (Toolkit for Probabilistic models of Sequences) framework [36].

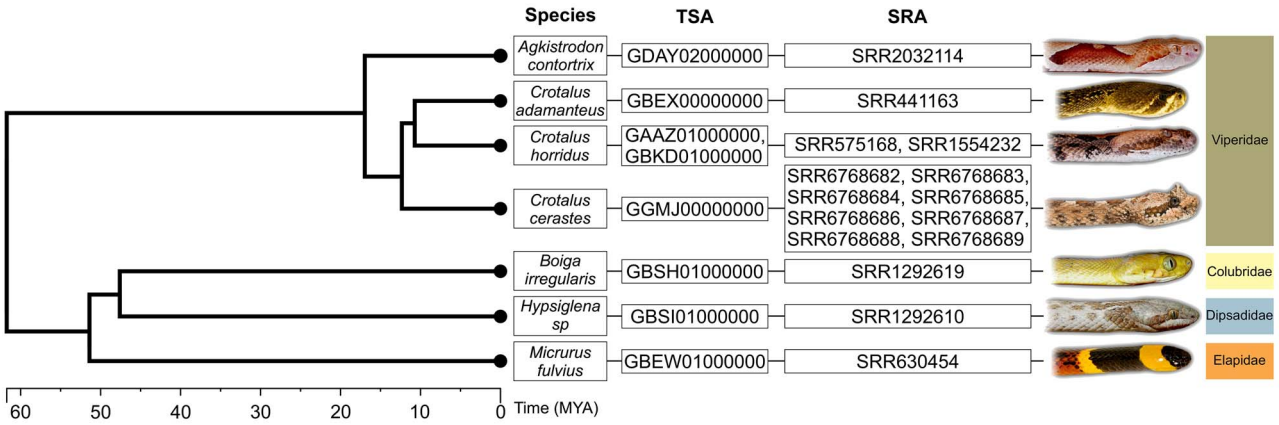

Figure 2. Diagram showing the phylogenetic relationships among the species with their TSA and SRA datasets used to design the testing sets and perform the comparative analysis. The divergence time among the species was estimated using TimeTree [53]. Photo credits: Michael Hogan and Travis Fisher.

All sequences present in the training sets were excluded from the testing sets.

Testing sets

In order to evaluate the performances of ToxCodAn and other annotators, we designed two sets of tests to perform comparative analysis. First, we designed a test set using the Transcriptome Shotgun Assemblies (TSA) from National Center for Biotechnology Information (NCBI) Genbank as a controlled scenario where all transcripts to be annotated are purposefully full length, correctly assembled and manually curated for reference. With the TSA test set, we are able to comparatively measure the accuracy of each pipeline/tool in retrieving the expected annotation of known toxin transcripts.

Second, we designed a test set using *de novo* assemblies from the NCBI Genbank Sequence Read Archive (SRA) data associated with each TSA to simulate a real-world scenario where the datasets are composed of full, partial, misassembled and chimeric transcripts. We detailed the workflow of the testing sets in the Supplementary Figure 1.

TSA testing set

For the TSA comparative analysis, we downloaded seven datasets from the TSA database (https://www.ncbi.nlm.nih.gov/genbank/tsa/) representing six snake species including *Agkistrodon contortrix*, *Crotalus adamanteus*, *Crotalus horridus*, *Boiga irregularis*, *Hypsiglena sp.* and *Micrurus fulvius* (Figure 2). These six species belong to the families Viperidae, Colubridae, Dipsadidae and Elapidae. The TSA datasets were annotated as stated in their original manuscripts [13, 14, 30, 37–39]. These datasets, which contained full-length assembled contigs with known toxin CDS within, were used to assess if all toxins could be annotated, given their known presence in the input assembly. Specifically, the curated toxins data were used in a comparative analysis to check the performance of annotators on reliably identifying toxins in the assemblies.

De novo testing set

For the *de novo* comparative analysis, we downloaded the raw data associated with each of the seven TSA datasets from the SRA database (https://www.ncbi.nlm.nih.gov/sra; Figure 2). We also included an additional dataset for *Crotalus cerastes* [30]. In total, we had 15 datasets from seven snake species belonging to the families Viperidae, Colubridae, Dipsadidae and Elapidae. Following trimming and assembly, these datasets were used to assess if all toxins could be annotated, given the unknown presence in the input dataset due to potential misassembly. In addition, the *de novo* datasets provided more realistic examples with regard to the number of input contigs and run times.

For these datasets, we first trimmed adapters and low-quality reads using Trim Galore! (v0.4.4; https://github.com/FelixKrueger/TrimGalore). Reads were removed if they had Phred scores < 5 and a length < 75 bp. Next, paired-end reads were merged using PEAR (v0.9.10; [40]). *De novo* assembly followed the recommendations of Holding *et al.* [17]. Briefly, we used three assemblers: Trinity (default; [20]), Extender (overlap parameter set to 120; [41]) and NGen (using the default parameters; Lasergene DNAStar software package; Madison, WI: https://www.dnastar.com/t-nextgen-seqman-ngen.aspx). The assemblies generated for each dataset were concatenated and 100% similar transcripts were clustered using cd-hit [42] to generate the final testing set.

Comparative analysis

We compared the annotation performance of ToxCodAn to that of Trinotate (v3.2.0; [19]), Dammit (v1.2; [22]) and Venomix (v0.7; [8]) using the default parameters for each annotator. We included all possible databases (–full) for Dammit and estimated expression using RSEM [43] with Bowtie2 [44] prior to running Venomix. For each annotator, we only kept predictions considered 'complete' by Transdecoder and clustered results at 100% identity with cd-hit in order to remove redundancy. Furthermore, given that Trinotate and Dammit are general annotators (i.e. not designed specifically to annotate toxins), we filtered their output annotations by applying a similar BLAST search to that used in the 'prefilter step' of the ToxCodAn pipeline in order to standardize comparisons.

To compare the results of all annotators, we performed a BLAST search of the curated toxins for each species (i.e. TSA-curated toxins) against the annotations obtained by each software. The BLAST search parameters were set to 99% coverage and identity. Before performing BLAST searches, curated toxins were clustered at 99% using cd-hit to reduce redundancy of repeated transcripts and group allelic variation at a single locus. In addition, curated toxins were checked for chimeric sequences (CKs), which may be present due to misassembly. Previous studies often did not perform these steps; therefore, doing so

ensured that searches were only performed for clean transcripts. To check for chimeras, we used a custom script (ChimeraKiller; https://github.com/masonaj157/ChimeraKiller). Briefly, reads were mapped to the annotated transcriptome and transcripts with zero coverage at any position were removed. Chimeric transcripts were then reported by searching for a difference > 75% (*-d* 0.75) in the average length of reads on either side of a given site based on the average read size. Using the BLAST search results, we calculated the recall—or the percent of curated toxins recovered—and derived statistics from a confusion matrix including false-positive rate, precision, sensitivity and *F*1-score. We considered that True positives are the toxin transcripts assembled in the datasets that were predicted as toxins, False positives are nontoxin transcripts assembled in the datasets that were predicted as toxins, False negatives are toxin transcripts assembled in datasets that were predicted as nontoxins and the True negatives are nontoxin transcripts assembled in datasets that were predicted as nontoxins.

If a toxin transcript fails to assemble, it cannot be annotated. Therefore, to ensure toxins were properly assembled in the *de novo* test sets, we performed an identical BLAST search on the transcriptome assembly prior to annotation. In the case that a curated toxin was not detected in the assembly, we performed an additional analysis to check for partial transcripts of the nonassembled toxins. This was done to assess why certain toxins were not available for annotation. To do this, we performed a BLAST search with 20% coverage and 95% identity to classify the toxin in the assembly into three main categories: (i) complete transcripts, where the full toxin CDS is in the middle of the contig; (ii) partial transcripts, where the border of the toxin CDS aligns with the border of the contig or the full contig is found within the toxin CDS; and (iii) nonpartial, where the hit was partial between the CDS and the contig.

CDS size analysis

The other annotators tested in the comparative analysis use TransDecoder to identify the longest ORF in an assembled contig. This is often incorrect due to misidentification of the correct start codon [25]. Therefore, to check if the CDS predicted and annotated by each tool matched the curated toxin, we performed an analysis comparing the size of the true/curated toxin CDS with the predicted CDS.

Running time analysis

To assess the run time of ToxCodAn compared with the other annotators, we used the *M. fulvius de novo* dataset, which contains 146 077 assembled contigs. We performed these analyses on the Clemson University Palmetto Supercomputing Cluster specifying 16 CPUs and 62 GB of memory.

Guide to venom gland transcriptomics

Alongside ToxCodAn, we designed a guide to venom gland transcriptomics with our recommended bioinformatics pipeline from raw data to expression quantification. Specifically, we provide the command-line code and links to useful resources for basic bioinformatics, data trimming, merging paired-reads, assembly, annotation, cleaning and quantification. We also provide an R script containing useful functions for plotting expression results. Our guide is available in Markdown format on our ToxCodAn GitHub repository (https://github.com/pedronachtigall/ToxCodAn/tree/master/Guide) and in an archived PDF found in Supplementary File 1.

Bothrops alternatus case study

To provide a case study for ToxCodAn and our guide to venom gland transcriptomics, we characterized the venom gland transcriptome for two Urutu Lanceheads (Viperidae: *B. alternatus*) from two distinct regions of Brazil. *Bothrops alternatus* is a large pit viper with an average size of 754.5 mm and geographical distribution ranging from Northern Argentina to the South/Central Brazil, Paraguay and Uruguay [45]. *Bothrops alternatus* is considered a dietary specialist, feeding almost exclusively on mammals [46, 47]. The venom of this species has not been well studied, especially compared with other species of *Bothrops*, but previous reports suggest high proteolytic and hemotoxic effects. In 2010, Cardoso *et al*. [48] published the first transcriptome analysis using Expressed Sequence Tags (ESTs). SVMP made up 81% of toxin expression followed by bradykinin-potentiating peptides (BPP, 8%), phospholipases (PLA2, 5%), snake venom serine proteases (SVSP, 2%) and c-type lectins (CTL, 1%). A follow-up study in 2014 using ESTs from a single individual similarly identified high expression of SVMPs (59%), but much higher expression of CTLs (16%) [49]. Based on these findings, we expect to find a high expression of proteolytic toxins in the venom of *B. alternatus*.

Here, we briefly describe each step of data analysis, and all computational steps and code can be found in our guide (https://github.com/pedronachtigall/ToxCodAn/tree/master/Guide; Supplementary File 1).

Sampling

One specimen (SB0060CVR) was collected in September 2017 in Mato Grosso do Sul state, Brazil. The second specimen (SB0022CVR) was collected in September 2016 in Rio Grande do Sul state, Brazil. For both snakes, venom was collected by allowing the snake to bite a sterile cup and venom glands were excised for transcriptomics after 4 days when transcription is maximized [50]. The specimens were euthanized with a single-step sodium pentobarbital (100 mg/kg) injection following standard approved American Veterinary Medical Association (AVMA) guidelines. Venom glands were transferred to RNAlater (Ambion) and stored at -80°C. The snakes were handled and collected under Protocol Number 4479020217 from Ethic Committee on Animal Use of the Butantan Institute (CEUAIB).

RNA extraction and sequencing protocol

Total RNA from the venom gland was extracted using Trizol Reagent (Invitrogen), following the manufacturer's protocol. The RNA concentration and contamination level were measured by ultraviolet absorbance using NanoDrop 1000 (Thermo Scientific), and RNA integrity was assessed with the equipment Agilent 2100 Bioanalyzer (Agilent Technologies).

Next, the messenger RNA (mRNA) was purified from the total RNA by using the Dynabeads® mRNA DIRECT kit (Ambion) and used to prepare independent cDNA libraries for each venom gland from each snake. The complementary DNA (cDNA) libraries were prepared following the protocol for TruSeqTM RNA Sample Preparation Kits v2 (Illumina), and sequenced using the HiSeq1500 platform (Illumina), generating strand-specific paired-end reads.

Transcriptome assembly, annotation and quantification

Illumina adapters and low-quality reads were trimmed using Trim Galore! (v0.4.4; https://github.com/FelixKrueger/TrimGalore), removing reads with Phred scores < 5 and a length < 75

bp. Paired-end reads were merged using PEAR (Paired-End reAd mergeR) [40] and *de novo* assembly followed Holding *et al.* [17]. Specifically, we used three assemblers: Trinity (default; [20]), Extender (overlap parameter set to 120; [41]) and NGen (using the default parameters; Lasergene DNAStar software package; Madison, WI: https://www.dnastar.com/t-nextgen-seqman-ngen.aspx). All assemblies were combined and clustered with a 100% identity threshold using cd-hit [42] to remove redundancy.

After *de novo* transcriptome assembly, we performed toxin annotation using ToxCodAn with default parameters. We combined the resulting toxin (redundancy filtered) and PT (SP filtered) CDSs into a single file. Next, we eliminated chimeric transcripts resulting from the assembly process using the custom ChimeraKiller script described earlier. Removing CKs further reduces the number of annotated transcripts (false positives) and prevents accidentally keeping spurious transcripts in the final transcriptome, resulting in a cleaner transcriptome. Finally, we clustered the cleaned toxin CDSs with 99% similarity using cd-hit to reduce redundancy of repeated transcripts and group allelic variation at a single locus.

To annotate nontoxin transcripts, we used the nontoxin contigs from ToxCodAn and performed CDS prediction using CodAn [25] with the full model designed for vertebrates. Predicted CDSs were then BLAST searched against two proteinDBs: (1) a custom proteinDB (https://github.com/pedronachtigall/ToxCodAn) and (2) the Swissprot database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz). The proteinDB contains peptide sequences of the snakes proteins with reviewed status at Uniprot, peptide sequences annotated in the genomes of species belonging to Toxicofera clade available at Ensembl (i.e. *Anolis carolinensis*, *Pogona vitticeps* and *Notechis scutatus*), and the peptide sequences of curated snake venom gland transcriptomes available at TSA (i.e. *A. contortrix*, *C. adamanteus*, *C. cerastes*, *C. horridus*, *B. irregularis*, *Hypsiglena sp.* and *M. fulvius*). Then, we performed an HMM search in the CDSs not annotated in the BLAST search step by using hmmsearch (HMMER 3.2.1; http://hmmer.org/) with the HMM (Hidden Markov Model) models from BUSCO [51] and hmmscan (HMMER 3.2.1; http://hmmer.org/) with the HMM models from Pfam (https://pfam.xfam.org/). Although we ran nontoxin annotation separately, this framework has been incorporated into ToxCodAn, and nontoxin annotation can be performed using the provided databases and models (by using the '-n' option). Importantly, nontoxin annotation is not performed directly by ToxCodAn, which only has gene models specific to toxin families, we instead rely on CodAn to make CDS predictions (see (Nachtigall et al. 2020) [25] for review of CodAn performance). All annotated nontoxin transcripts were checked for chimeras to generate the final non-toxins set.

Then, the final non-toxins were combined with the final ToxCodAn-annotated toxins and clustered with 99% similarity using cd-hit. To create a *B. alternatus* consensus transcriptome, we combined our two individual's transcriptomes together and clustered transcripts at 98% similarity using cd-hit. Finally, we performed transcript quantification using RSEM with Bowtie2 [43, 44].

We also performed a comparative analysis with the ESTs sequenced by Cardoso *et al.* ([48]; accession numbers from GW575430 to GW583300 in GenBank). We BLAST searched the EST sequences against our *B. alternatus* consensus transcriptome using a percentage identity threshold of 98% (-perc_identity 98) and only considered the best hit for each EST sequence. We did not include the ESTs sequenced by de Paula *et al.* [49], due to the unavailability of these sequences.

Results

ToxCodAn was designed to reliably identify the venom components within snake venom gland transcriptome assemblies. ToxCodAn uses gHMMs designed for different toxin families to successfully characterize the set of toxins highly and/or lowly expressed. We compared ToxCodAn's performance against that of Trinotate, Dammit and Venomix in two testing sets: the TSA test, containing seven datasets from six snake species, and the *de novo* test, containing 15 datasets from seven snake species. Also, the case study demonstrated the applicability of ToxCodAn and the guide on performing Toxin and Nontoxin annotation of two novel venom gland transcriptome assemblies from *B. alternatus* species.

TSA comparative analysis

The TSA comparative analysis revealed that ToxCodAn has the best performance with an average toxin detection of 96% (Figure 3). Trinotate and Dammit had the next best performance with an average toxin detection of 83 and 68%, respectively (Figure 3). In addition to having better performance in toxin detection, ToxCodAn made comparatively modest number of predictions (i.e. generated fewer false positives; Figure 3). On average, ToxCodAn made only 80 predictions compared with 177 and 167 made by Trinotate and Dammit (Figure 3). Venomix made fewer predictions (53), but also recovered less toxins with an average toxin detection of only 50% (Figure 3). Following filtering with SignalP and removing CKs, ToxCodAn displays little change in the toxin detection, but further reduces the number of predictions made (Figure 3). Results for each individual dataset can be found in Supplementary Figure 2. Overall, on average ToxCodAn had a lower false-positive rate at ~1%, whereas Trinotate, Dammit and Venomix had much higher false-positive rates (11, 8 and 5%, respectively; Figure 4). ToxCodAn also had higher precision, sensitivity and F1-scores (Figure 4).

The few toxins not identified by ToxCodAn (and other annotators) were mainly lowly expressed toxins or transcripts with questionable function in the venom and only recognized as potential or PTs (Figure 5; Supplementary File 2). These include toxins such as Ficolin, Waprin and Hyaluronidase. However, unlike Trinotate and the other annotators, ToxCodAn consistently annotates the most highly expressed transcripts and those considered biologically relevant and critical to venom composition (Figure 5; Supplementary File 2).

De novo comparative analysis

The *de novo* comparative analysis similarly revealed that ToxCodAn has the best performance with an average toxin detection of 96% compared with that of Trinotate (86%), Dammit (83%) and Venomix (69%) (Figure 3). Again, ToxCodAn made a comparatively modest number of predictions at ~200 compared with Trinotate (432), Dammit (281) and Venomix (225) (Figure 3). Following filtering with SignalP and removing CKs, ToxCodAn displays little change in toxin detection, but further reduces the number of predictions made (Figure 3). Results for each individual dataset can be found in Supplementary Figure 3. False-positive rates for the *de novo* analysis were much lower than that of the TSA analysis due to the large number of transcripts present and filtered from the dataset by each annotator (Figure 4). However, ToxCodAn had a much lower false-positive rate at 0.1% compared with Trinotate (0.7%),

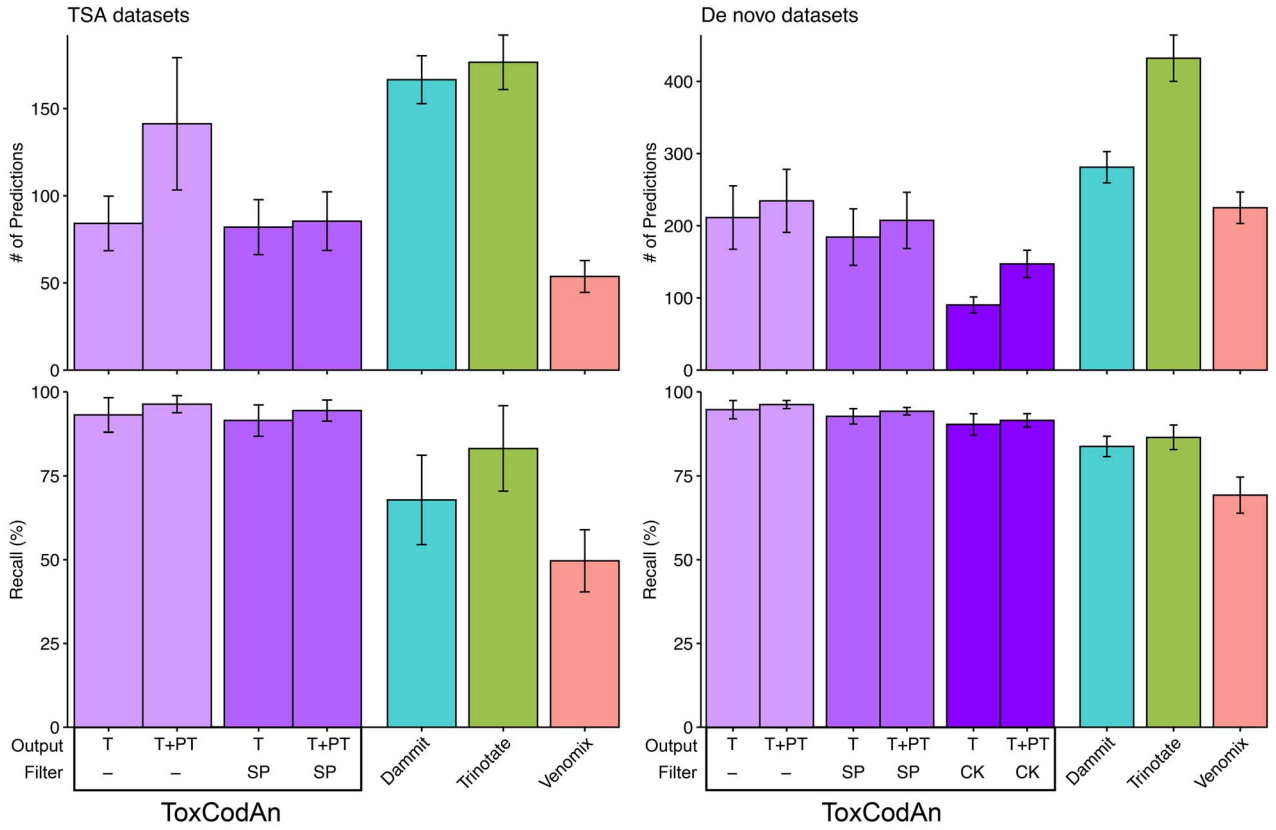

Figure 3. Barplots with 95% confidence intervals for the average number of predictions made (top) and average toxin recall (bottom) by each annotator across the six TSA datasets (left) and 15 *de novo* datasets (right). Here, recall is defined as the percent of curated toxins recovered. Bar colors correspond to different annotators or ToxCodAn recommended filtering protocols. T, Toxins; PT, Putative Toxins; SP,SignalP filtered; CK, ChimeraKiller filtered.

Dammit (0.4%) and Venomix (0.5%) (Figure 4). In the *de novo* analysis, ToxCodAn also had higher precision, sensitivity and *F*1-scores (Figure 4).

Similar to the TSA dataset, many of the toxins which were not annotated by ToxCodAn or other annotators were low-expression toxins or toxins with questionable toxic function (Figure 6; Supplementary File 3). However, Trinotate, Dammit and Venomix often fail to annotate even highly expressed transcripts.

Some toxins were not annotated by any software because they were not detected in the assembly (Figure 6; Supplementary File 3). To determine why certain transcripts were not detected, we searched for partial transcripts. We found that although some toxins failed to assemble at all, most were partially assembled (Figure 6).

CDS size analysis

We found that ToxCodAn also has better performance in correctly identifying the full CDS of the toxins in all datasets analyzed (Figure 7 and Supplementaryl File 4). In particular, although all annotators have a median at or near zero, representing prediction of the correct start codon position, ToxCodAn had a much smaller variance with a percentage of 97.3 and 93.01% of correctly identifying the start codon position in TSA and *de novo* analysis, respectively (Figure 7). In comparison, Trinotate identified the correct start codon position in 55.82% of TSA predictions and in 48.07% *de novo* predictions. Dammit was able to correctly predict 87.20 and 82.99% in TSA and *de novo* datasets, respectively, whereas Venomix identified the correct start codon position in 86.32 and 83.04% of its predictions in the TSA and *de novo* datasets, respectively. Overall, the other annotators consistently annotate ORFs that are much longer than the curated toxin's CDS, demonstrating the bias these annotators suffer related to the CDS prediction tools used in each pipeline. In response, these annotators require additional manual curation of the CDS in most of its predictions, which is not required by ToxCodAn (Figure 7 and Supplementary File 4).

Running time analysis

We found that ToxCodAn is substantially faster than the other annotators, particularly when considering required setup times for the other annotators, which includes running RSEM for Venomix and setting up databases for Trinotate and Dammit (Table 1). The total amount of time needed for ToxCodAn was approximately 23 min, whereas Venomix, Trinotate and Dammit needed > 8 h, 12 h and 35 h to complete, respectively.

Importantly, we did not record installation/troubleshooting times, which is substantial for some annotators. In addition, it is important to note that Trinotate, Dammit and Venomix require extensive setup prior to analysis. ToxCodAn is freely available on GitHub and simple to install and use immediately. The only setup required is unzipping the folder containing the toxin gHMMs. Furthermore, we tested all annotators with a large amount of computational resources (16 CPUs; 62 GB memory) because certain steps in Venomix and Dammit required these

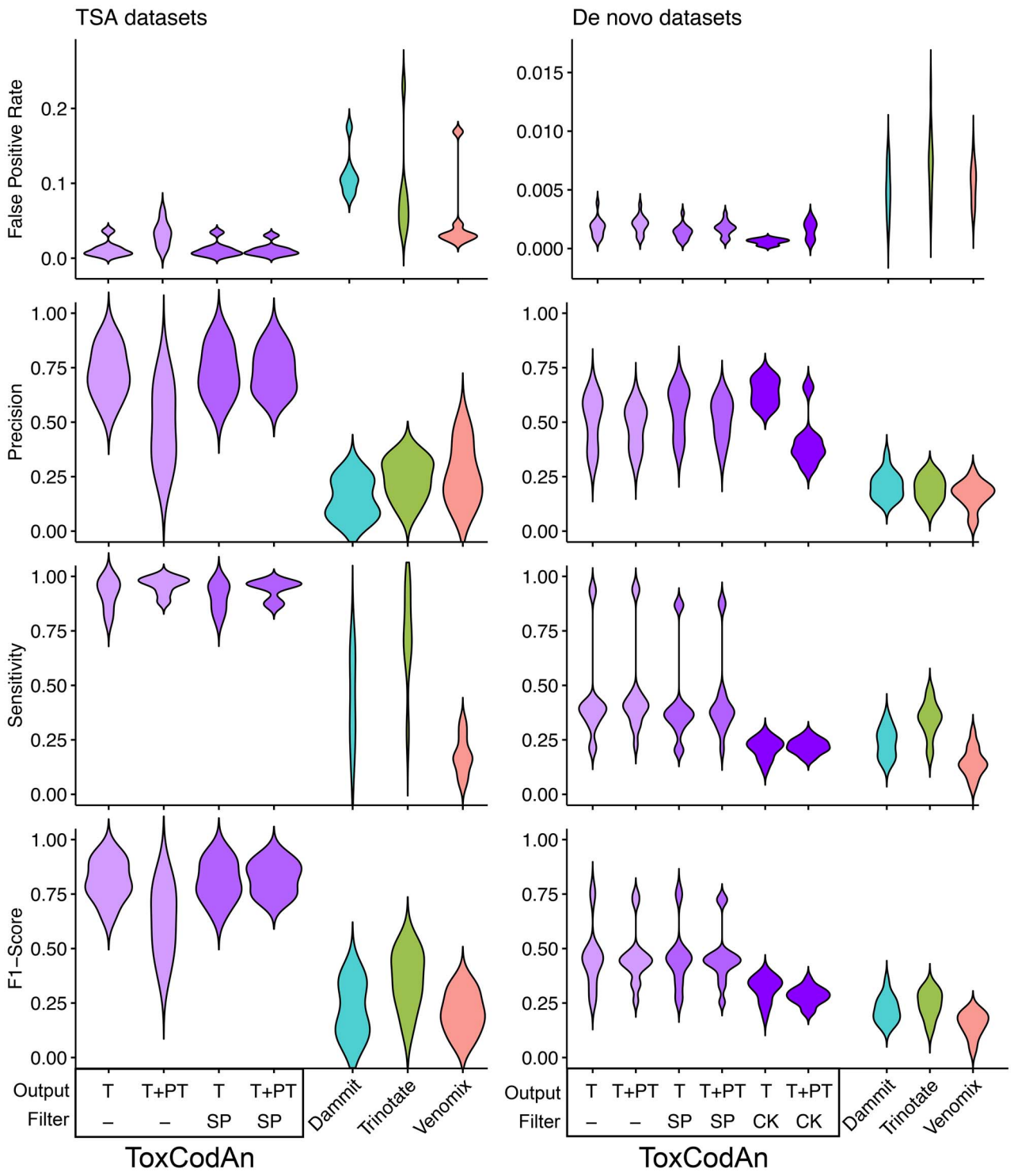

Figure 4. Violin plot of the average false-positive rate, precision, sensitivity and F1-Score for each annotator across the six TSA datasets (left) and 15 *de novo* datasets (right). Violin colors correspond to different annotators or ToxCodAn recommended filtering protocols. T, Toxins; PT, Putative Toxins; SP, SignalP filtered; CK, ChimeraKiller filtered.

resources to successfully complete. In fact, Dammit required additional memory to complete (89 GB). However, running ToxCodAn with only 8 CPU and 32 GB of memory, which is standard on many computers, resulted in only a 1 h run time, which is still faster than the other annotators ran with substantially more resources. Moreover, if the user has a low CPU and memory available for use, ToxCodAn can still process a high amount of data, which indicates that ToxCodAn is a tool that can be used on any personal computer with UNIX operational system or takes advantage of supercomputers.

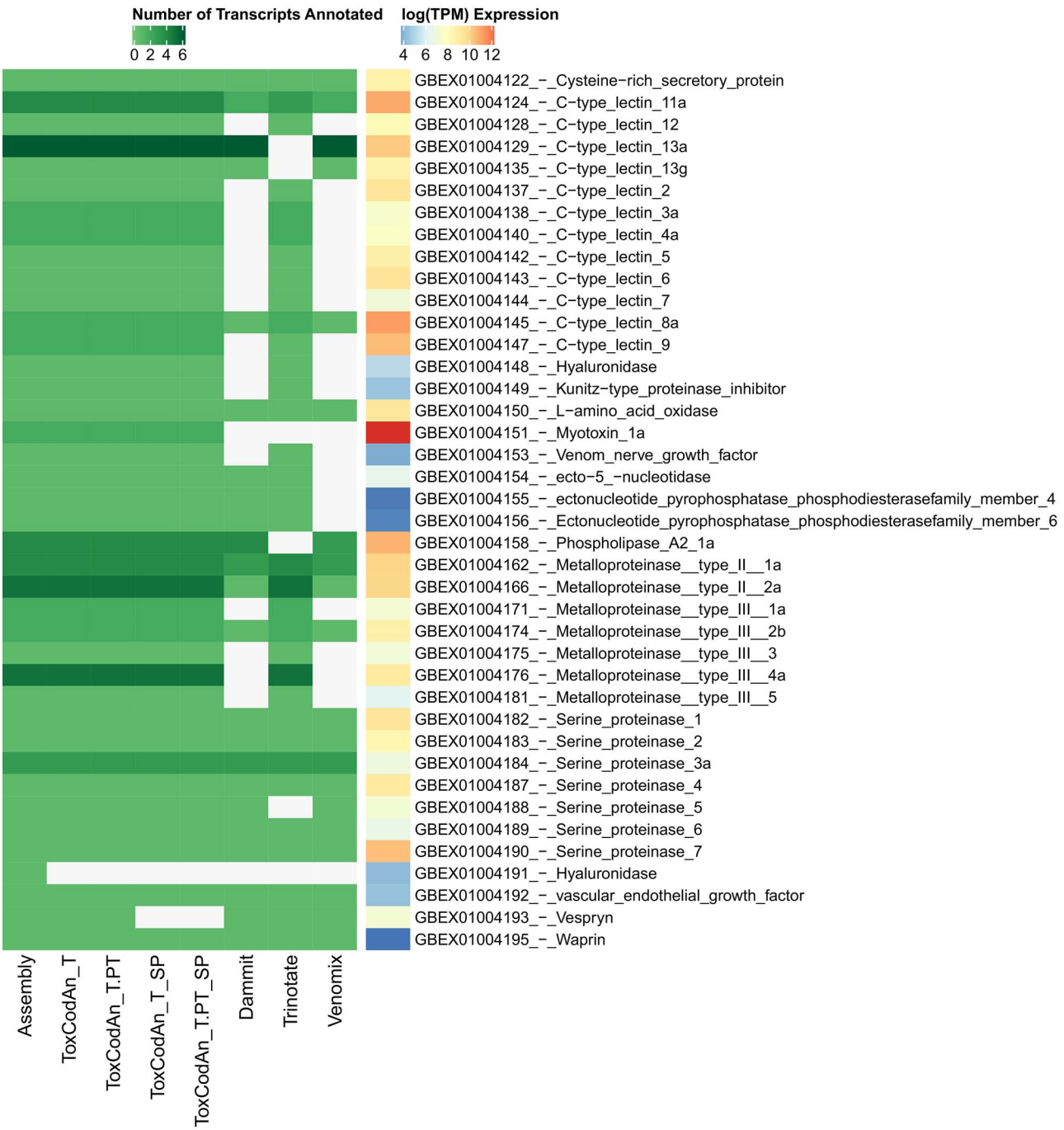

Figure 5. Example of heatmap showing toxin transcript expression level of the *C. adamanteus* TSA dataset. The rows represent the toxin transcripts and the columns represent the assembly, annotators and expression level. T, Toxins; PT, Putative Toxins; SP, SignalP filtered.

Table 1. Run times for each annotator on the *M. fulvius de novo* dataset (n = 146 077 contigs) *Notes* : Runs were performed on the Clemson University Palmetto Supercomputing Cluster with 16 CPU and 62 GB of memory. Despite abundance of resources, Dammit still required additional memory to complete (89 GB). Installation difficulty and time is not included, the setup times for Trinotate, Dammit and Venomix refer to downloading and setting up databases post-installation or running RSEM (required for Venomix).

Annotator	Setup time	Run time	Total time	CPU setup time	CPU run time	CPU total time
ToxCodAn	00:00:00	00:23:43	00:23:43	00:00:00	04:21:24	04:21:24
Venomix	08:38:13	00:11:19	08:49:32	109:19:13	00:12:27	109:31:40
Dammit	03:33:29	31:39:12	35:12:41	02:00:06	39:11:21	41:11:27
Trinotate	12:02:09	00:18:20	12:20:29	159:11:26	00:06:06	159:17:32

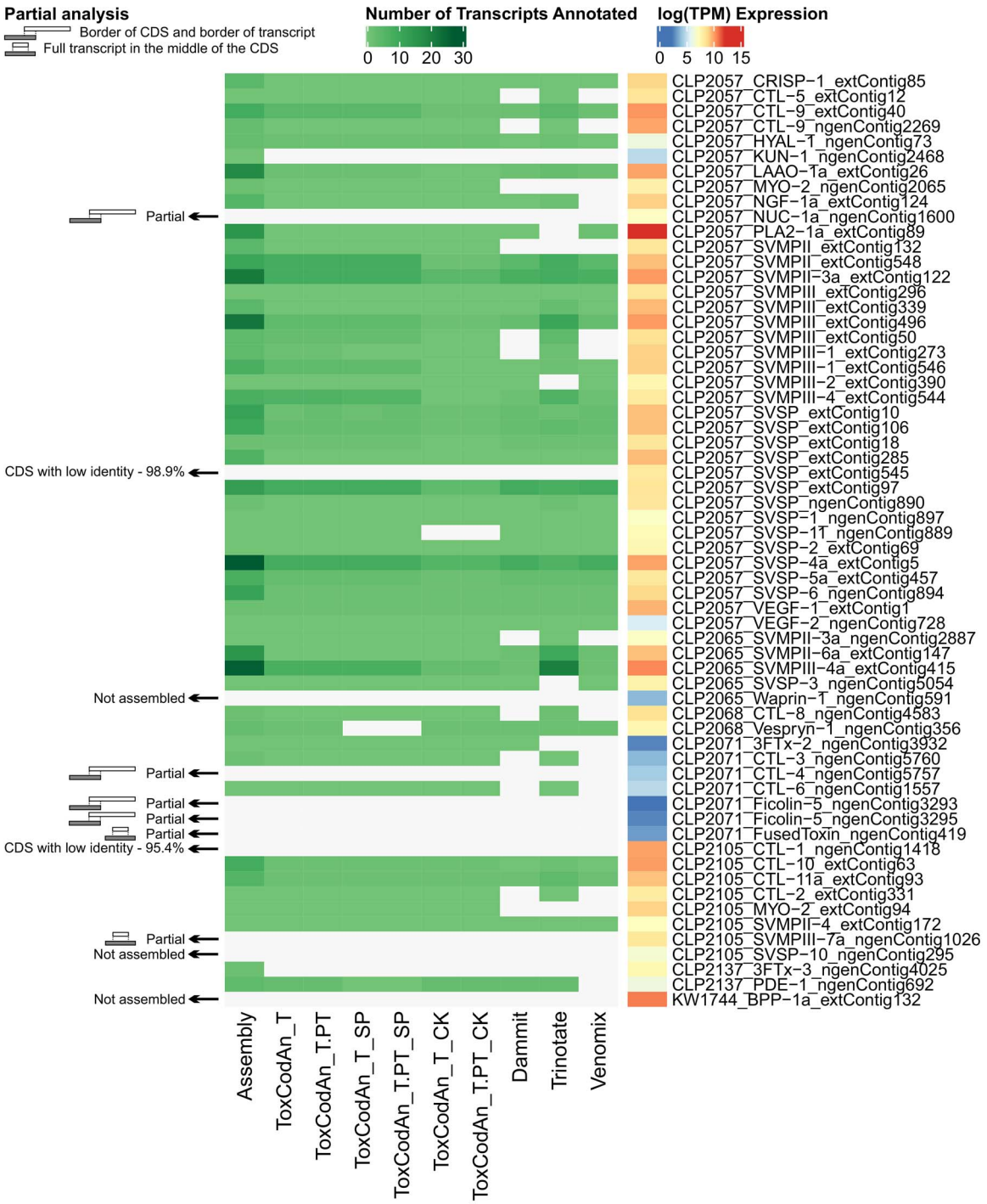

Figure 6. Example of heatmap showing toxin transcript expression level and the partial analysis of the *C. cerastes* Ccera-SRR6768684 *de novo* dataset. The rows represent the toxin transcripts and the columns represent the assembly, annotators and expression level. T, Toxins; PT, Putative Toxins; SP, Signal-P filtered; CK, ChimeraKiller filtered.

Bothrops alternatus case study

Using ToxCodAn and following our guide (https://github.com/pedronachtigall/ToxCodAn/tree/master/Guide; Supplementary File 1), we identified 80 toxin transcripts from 23 gene families for the *B. alternatus* species (Figure 8). Toxins represented 59–71% of total expression, and 64 of the toxin transcripts, on average, had expression levels greater than the transcriptome-wide average (transcript per million > 146; Supplementary Table S1 in Supplementary File 5). We found that the *B. alternatus* samples similarly contain CTLs (22 transcripts), PLA2s (2), SVMPs (16), BPPs (3) and SVSPs (12) as the major components to the venom. Minor components of the venom include vascular

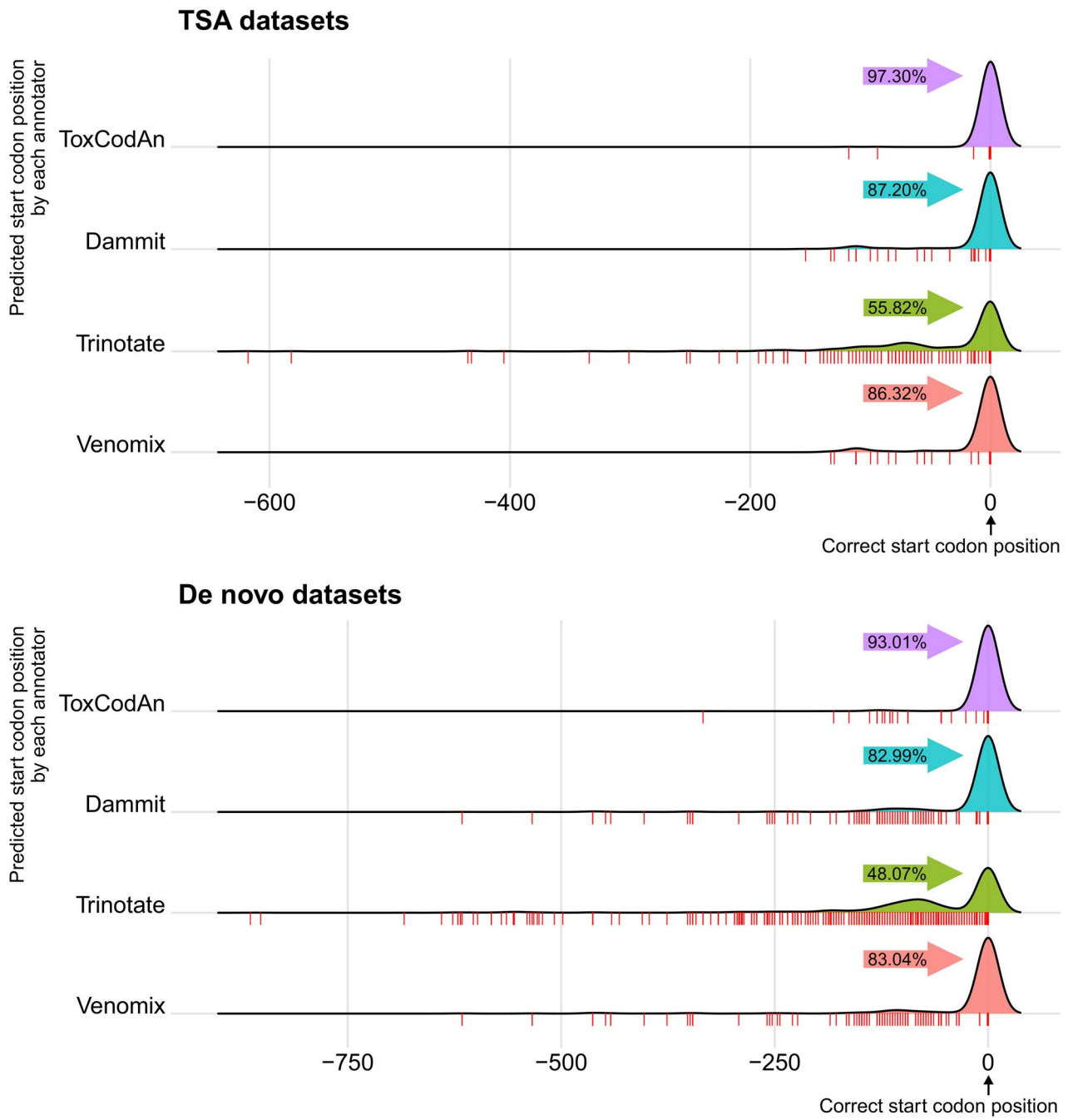

Figure 7. Predicted codon start positions of toxins of each annotator across all six TSA datasets (top) and 15 *de novo* datasets (bottom). Zero represents the correct start codon position, whereas the plot field and red traces show the start codon position identified by each annotator. The percentage of correct start codon position identified by each annotator is indicated in the colored arrows.

endothelial growth factors, a cystine-rich secretory protein, L-amino acid oxidases and 13 other toxin families were annotated but expressed at very low levels (Figure 8). For example, a three-finger toxin was annotated, but is very lowly expressed. Although three-finger toxin may be present in the genome and detected in the transcriptomes of some vipers, it is not expected to be highly expressed [29, 52]; therefore, these 13 toxin families might not be relevant venom components of *B. alternatus*.

Unlike the reports from Cardoso *et al.* [48] and de Paula *et al.* [49], we found that CTLs, rather than SVMPs, represent the most highly expressed toxin family, followed by PLA2s or SVMPs, depending on individual or sampling locality (Figure 8, Supplementary Table S2 in Supplementary File 5). This may suggest differential expression between sampled locations in *B. alternatus*. Particularly, PLA2s are much less represented in the specimen from Southern Brazil than in the specimen from Center-West region. This difference was also observed between

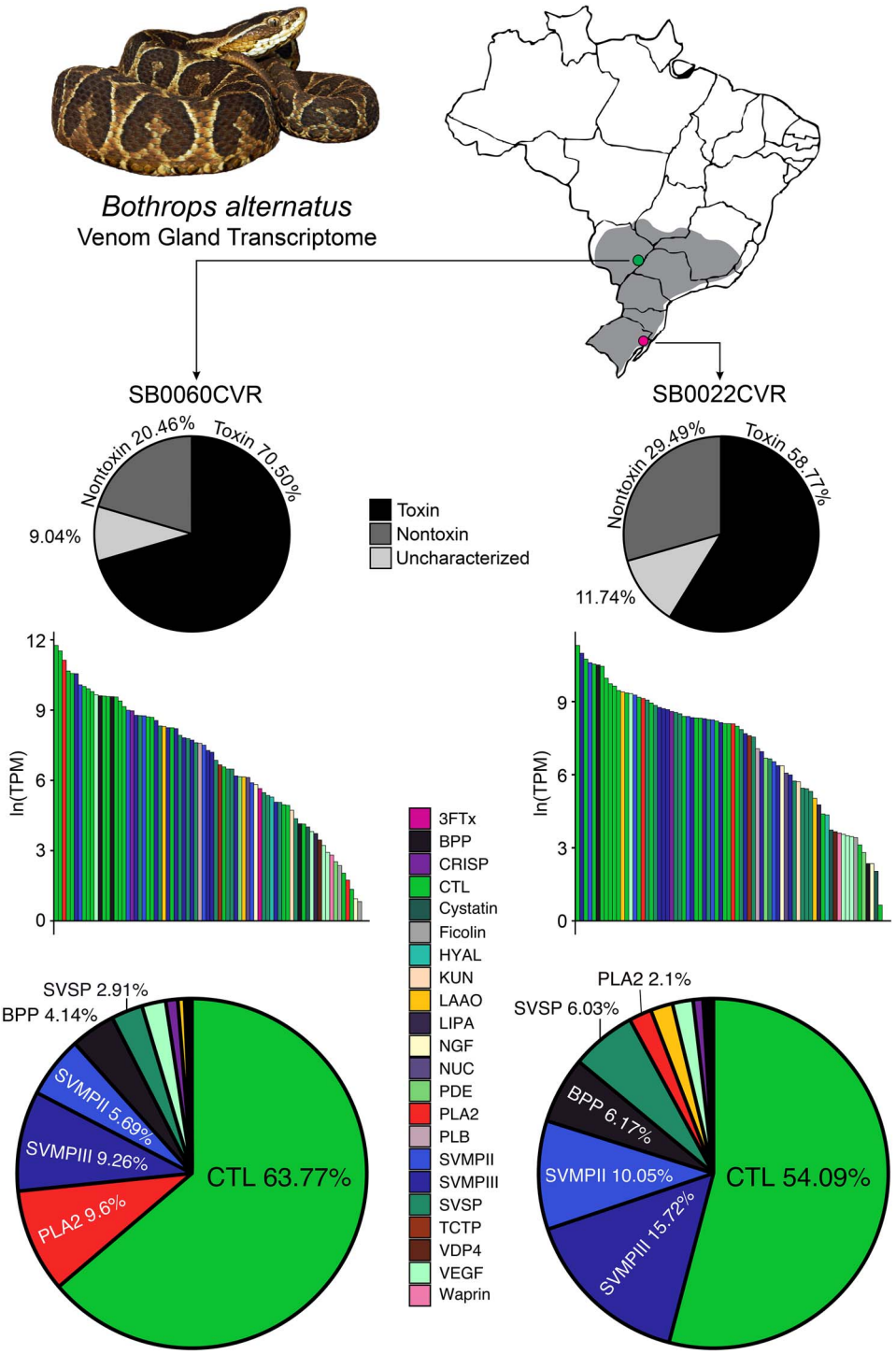

Figure 8. Venom components identified in the *B. alternatus* samples from two distinct regions of Brazil. In the map of Brazil at the top, the gray indicates the distribution range of the species. Photo credits: Arthur D. Abegg. 3FTx, three-finger toxin; BPP, bradykinin-potentiating peptides; CRISP, cysteine-rich secretory proteins; CTL, C-type lectins; HYAL, hyaluronidase; KUN, Kunitz-type proteinase inhibitor; LAAO,L-amino acid oxidase; LIPA, lipase; NGF, nerve growth factor; NUC, Ecto 5' nucleotidase; PDE, phosphodiesterase; PLA2, phospholipase A2; PLB, phospholipase B; SVMP, snake venom metalloproteinase; SVSP, snake venom serine protease; TCTP, translationally controlled tumor protein; VDP4, venom dipeptidylpeptidase 4; VEGF-F, vascular endothelial growth factor.

the specimens from São Paulo state used by de Paula *et al.* [49], which have low PLA2 content, and the specimen used by Cardoso *et al.* [48], from an unknown location, which showed higher PLA2 content.

Moreover, we noticed that the SVMPIII expression in Cardoso's sample, which is a mix of three individuals, was mainly related to two SVMPIII transcripts (~62%). These SVMPIII transcripts are lowly expressed in the two samples used in the

present study (Supplementary Table S3 in Supplementary File 5). On the other hand, the SVMPIII expression observed in the samples used in the present study was represented mainly by one transcript that is lowly expressed in Cardoso's sample. Looking at the SVMPII differential expression profile, we noticed a similar pattern among samples (Supplementary Tables S2 and S3 in Supplementary File 5). In this sense, the differences in metalloproteinases regards the SVMPIII expression. The CTLs were lowly expressed in Cardoso's sample and its expression was distributed among all CTL transcripts, whereas the high expression of CTLs in the samples from the present study were mainly related to five CTL transcripts (Supplementary Tables S2 and S3 in Supplementary File 5). We did not detect several transcripts annotated by our pipeline in the ESTs data, likely because the coverage obtained by EST approach does not allow to capture all set of transcripts expressed in a sample within a similar depth obtained by using RNA-seq experiments.

Discussion

Transcriptomics (RNA-seq) has transformed evolutionary biology and genetics, particularly for nonmodel systems like venomous snakes. However, bioinformatics processing of venom gland transcriptomes can be challenging. Therefore, we provide a convenient computational tool, ToxCodAn, and a guide to venom gland transcriptomics to facilitate research exploring snake venom composition. We demonstrate that ToxCodAn has high precision and sensitivity in toxin recovery from a transcriptome assembly, particularly when compared with other transcriptome annotators (Figure 4). Overall, ToxCodAn can quickly predict most toxins in the transcriptome, displays high accuracy predicting the appropriately sized CDS and generates few false positives; thereby, minimizing overestimation of the number of toxins present in the genome [27–29]. Furthermore, our guide for venom gland transcriptomics provides a useful resource and pipeline to follow for processing venom gland transcriptomic data, which we demonstrate with *B. alternatus*.

ToxCodAn can be easily installed on any UNIX-like operating system and is fast, taking less than an hour to perform confident toxin identification and annotation of 146 077 contigs with 8 CPU and 32GB memory. These resources are available on most modern desktop and laptop computers, demonstrating the applicability of ToxCodAn for projects of any size, regardless of available computational resources. Furthermore, ToxCodAn produces highly accurate annotations with a comparably modest number of false positives (Figures 3and 4). This reduces the time needed for manual checks of the predicted toxins and demonstrates the utility of ToxCodAn even among more established annotation programs.

One of the biggest issues in the annotation procedure is predicting the correct or appropriately sized CDS. CDS predictors generally have good accuracy in identifying the stop codon, but not the start codon [25]. ToxCodAn can accurately identify the full CDS of toxins due to the use of CodAn (Figure 7 and Supplementary File 4) [25]. The other annotators tested use TransDecoder for CDS prediction, which is a self-training algorithm that uses the longest ORFs found in each contig as the training set. TransDecoder is known to have a lower rate of correct identification of the start codon than other CDS prediction tools [25]. Therefore, although Trinotate had good annotation success, it also had a high error rate for identifying the correct start codon. Dammit and Venomix presented better accuracy in identifying the correct start codon because they perform correction steps based on BLAST searches, which helps improve their performance.

ToxCodAn's improved performance compared with the other annotators may be related to two main features: (1) the CDS prediction step and (2) the database used in the search steps. First, ToxCodAn performs the CDS prediction using CodAn [25], which eliminates noncoding transcripts while allowing ToxCodAn to leverage prebuilt CDS models that were designed specifically to detect toxins with high accuracy. The other annotators rely on TransDecoder [20], which is a self-training algorithm that estimates a model specific to that set of sequences by detecting the longest ORF of each transcript. This characteristic of the TransDecoder pipeline led to a prediction biased towards the annotation of longer ORFs. Also, venom gland transcriptomes include toxin and nontoxin transcripts, which may present different features for CDS identification and may disrupt the capability of TransDecoder to accurately estimate toxin CDSs. In this sense, the use of models designed specifically for toxin genes improves the quality of predictions obtained and, consequently, the final annotation. The second reason ToxCodAn may exhibit improved performance is because of its comprehensive toxinDB. ToxCodAn uses a database consisting of 29 757 well-annotated toxins from several snake species. Trinotate and Dammit are general annotators, which use databases including higher percentages of nontoxin genes. Venomix uses a database containing toxin sequences from the ToxProt database [24], which contains 6349 reviewed proteins from 735 vertebrate and invertebrate species available on Uniprot (accessed in February 2021). In this sense, the use of a database highly populated with well-annotated toxin sequences may help improve the annotation performance exhibited by ToxCodAn.

The only toxins that ToxCodAn failed to predict were lowly expressed or transcripts questionable in toxic function (Figures 5 and 6). Nonetheless, with appropriate training data, future releases of ToxCodAn's gHMMs can be improved or expanded to better capture these low-expression toxins and/or to capture other toxins from other venomous taxa (e.g. invertebrate toxins). In addition, although ToxCodAn relies on existing databases and is unable to capture novel toxins, the guide to venom gland transcriptomics is designed to annotate nontoxin transcripts and even keep uncharacterized proteins in the final transcriptome, which may help to discover novel toxins, especially when combined with high-sensitivity proteomic analyses. Specifically, estimating the expression of the final transcriptome, including uncharacterized proteins, may uncover previously unidentified components of the venom that can be further explored. This is particularly important given the potential utility of toxins as therapeutic drug components [15].

Interestingly, in our *de novo* analysis, we identified some toxins that were not annotated by any software and not detected in the assembly (Figure 6; Supplementary File 3). After checking for partial assembly of these toxins, we found that in most cases the toxins were partially assembled, while a few toxins were not assembled at all. For instance, in *C. cerastes* (SRR6768684; Figure 6), despite being moderately expressed NUC-1a and SVMPIII-7a are only partially assembled. NUC-1a only assembled one side of the CDS and SVMPIII-7a only assembled in the center of the CDS, failing to assemble either side. In this sense, ToxCodAn (and the other annotators) performed appropriately, but caution should be taken by researchers to ensure proper assembly. Annotation completeness is necessarily dependent on the completeness of the input assembly. Snake venom toxins have been shown to be difficult to comprehensively assemble,

potentially requiring combinations of *de novo* assemblies with different software and/or parameter variation (see [17]).

Following our guide for venom gland transcriptomics and using ToxCodAn to annotate our *B. alternatus* venom gland transcriptomes, we demonstrate the applicability of these resources for other researchers. The guide to venom gland transcriptomics is an optimized bioinformatics pipeline based on nearly a decade of research that will facilitate accurate assembly, annotation, cleaning and quantification of venom gland transcriptomes for other researchers. The guide can be found in Markdown format on GitHub or archived as a PDF in Supplementary File 1. ToxCodAn and the guide can be used on any snake species and is able to identify both highly and lowly expressed toxins.

With our ToxCodAn-annotated transcriptome, we found high expression of CTLs in *B. alternatus*, unlike that of previous studies, which identified SVMPs to be the most highly expressed toxins [48, 49]. Nonetheless, de Paula *et al.* [49] found that CTLs were the second most highly expressed toxins. Overall, we see substantial differences in the expression of CTLs, PLA2s and SVMPs between our samples from Mato Grosso do Sul state and Rio Grande do Sul state, and the samples represented in de Paula *et al.* and Cardoso *et al.* studies, which are from São Paulo state and unknown location, respectively [48, 49]. Together, these results likely indicate differential expression or local adaptation affecting the venom composition of *B. alternatus* individuals. Our guide to venom gland transcriptomics, alongside ToxCodAn, are valuable resources for future research in the 'venomics' field. By streamlining the toxin annotation process, researchers will have more time to focus on downstream proteomic or functional analyses to better understand venom diversity, functional divergence and adaptation among species and populations.

Overall, ToxCodAn has several advantages for toxin annotation compared with other approaches. Our results revealed that ToxCodAn is suitable for use on any project focused on toxin annotation of snakes. Although ToxCodAn currently only has models designed specifically to snake species, we are working to expand the toxin models to other venomous animals (e.g. scorpions, spiders, cone snails, cnidarians, insects, etc.) and make them available in a near future.

Conclusion

Here, we describe ToxCodAn, a toxin annotator aimed to ease the laborious task of manual annotation of a *de novo* transcriptome assembly. We demonstrate that ToxCodAn performs better than other annotation tools and can be applied on data generated from any snake species. Furthermore, we provide a guide to venom gland transcriptomics, a resource to walk researcher's through venom gland transcriptomics, particularly snakes, but with the goal of expanding this framework for other venomous lineages.

Key Points

- We present ToxCodAn, a computational tool that performs confident toxin annotation on transcriptome assemblies.
- We provide a guide to venom gland transcriptomics to facilitate future research in the field of venomics.
- A comprehensive analysis using data from seven-snake species revealed that ToxCodAn is able to quickly and accurately annotate toxins.
- ToxCodAn presents a higher performance when compared with other annotation tools.
- Using ToxCodAn, we defined the toxin repertoire of *B. alternatus* specimens from a new geographic region

Data availability

The RNA-seq data from venom gland of *B. alternatus* are available at SRA under the accession numbers SRR13153637 (BioSample SAMN16930330) and SRR13153633 (BioSample SAMN16930333). ToxCodAn and the guide to venom gland transcriptomics are freely available at https://github.com/pedronachtigall/ToxCodAn.

Authors' Contributions

P.G.N., R.M.R., C.L.P. and I.L.M.J.A. conceived and designed the experiments. P.G.N. wrote the ToxCodAn script and trained the gHMM models. R.M.R. designed the training and testing sets. S.E. wrote BlastNamer script. P.G.N. and R.M.R. performed all experiments and bioinformatic analyses. I.L.M.J.A. contributed to the biological data. P.G.N., R.M.R., A.J.M., D.R.R., C.L.P. and I.L.M.J.A. validated the data. P.G.N. and R.M.R. wrote the manuscript. P.G.N., R.M.R., S.E., A.J.M., D.R.R., C.L.P. and I.L.M.J.A. critically edited the final manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors thank Jason L. Strickland, Erin E. Stiers, Erich P. Hofmann, Juan D. Bayona-Serrano and Luciana A. Freitas-de-Sousa for providing training data for ToxCodAn. The authors thank Michael Hogan for photographs of *A. contortrix*, *C. adamanteus*, *C. horridus*, *B. irregularis*, *Hypsiglena sp.* and *M. fulvius*, Travis Fisher for the photograph of *C. cerastes*, and Arthur D. Abegg for the photograph of *B. alternatus*. Finally, the authors thank Clemson University for generously providing computational resources on the Palmetto Cluster.

Funding

Fundação de Amparo à Pesquisa no Estado de São Paulo (FAPESP) (Processes Numbers: 2016/50127-5, 2018/26520-4 to I.L.M.J.A. and P.G.N.); the National Science Foundation (DEB 1145987 and DEB 1638902 to D.R.R. and DEB 1822417 and DEB 1638879 to C.L.P.).

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51. doi: 10.1038/nrg.2016.49.
2. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet* 2019;**20**(11):631–56. doi: 10.1038/s41576-019-0150-2.

3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**(1):57–63. doi: 10.1038/nrg2484.
4. Calvete JJ, Sanz L, Angulo Y, *et al*. Venoms, venomics, antivenomics. *FEBS Lett* 2009;**583**(11):1736–43. doi: 10.1016/j.febslet.2009.03.029.
5. Calvete JJ. Snake venomics: from the inventory of toxins to biology. *Toxicon* 2013;**75**:44–62. doi: 10.1016/j.toxicon.2013.03.020.
6. Casewell NR, Wüster W, Vonk FJ, *et al*. Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol Evol* 2013;**28**(4):219–29. doi: 10.1016/j.tree.2012.10.020.
7. Junqueira-de Azevedo IL, Campos PF, Ching AT, *et al*. Colubrid venom composition: an -omics perspective. *Toxins* 2016;**8**(8):1–24. doi: 10.3390/toxins8080230.
8. Macrander J, Panda J, Janies D, *et al*. Venomix: a simple bioinformatic pipeline for identifying and characterizing toxin gene candidates from transcriptomic data. *PeerJ* 2018;**6**:e5361. doi: 10.7717/peerj.5361.
9. Fry BG, Vidal N, Norman JA, *et al*. Early evolution of the venom system in lizards and snakes. *Nature* 2006;**439**(7076):584–8.
10. Zancolli G, Casewell NR. Venom systems as models for studying the origin and regulation of evolutionary novelties. *Mol Biol Evol* 2020;**37**(10):2777–90. doi: 10.1093/molbev/msaa133.
11. Casewell NR, Wagstaff SC, Wüster W, *et al*. Medically important differences in snake venom composition are dictated by distinct postgenomic mechanisms. *Proc Natl Acad Sci* 2014;**111**(25):9205–10. doi: 10.1073/pnas.1405484111.
12. Margres MJ, McGivern JJ, Wray KP, *et al*. Linking the transcriptome and proteome to characterize the venom of the Eastern Diamondback Rattlesnake (*Crotalus adamanteus*). *J Proteomics* 2014;**96**:145–58. doi: 10.1016/j.jprot.2013.11.001.
13. Rokyta DR, Margres MJ, Calvin K. Post-transcriptional mechanisms contribute little to phenotypic variation in snake venoms. *G3* 2015a;**5**(11):2375–82. doi: 10.1534/g3.115.020578.
14. McGivern JJ, Wray KP, Margres MJ, *et al*. RNA-seq and high-definition mass spectrometry reveal the complex and divergent venoms of two rear-fanged colubrid snakes. *BMC Genomics* 2014;**15**(1):1061. doi: 10.1186/1471-2164-15-1061.
15. Saviola AJ, Peichoto ME, Mackessy SP. Rear-fanged snake venoms: an untapped source of novel compounds and potential drug leads. *Toxin Rev* 2014;**33**(4):185–201. doi: 10.3109/15569543.2014.942040.
16. Modahl CM, Mackessy SP. Venoms of rear-fanged snakes: new proteins and novel activities. *Front Ecol Evol* 2019;**7**:1–18. doi: 10.3389/fevo.2019.00279.
17. Holding ML, Margres MJ, Mason AJ, *et al*. Evaluating the performance of de novo assembly methods for venom-gland transcriptomics. *Toxins* 2018;**10**(6):249. doi: 10.3390/toxins10060249.
18. Moreton J, Izquierdo A, Emes RD. Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. *Front Genet* 2016;**6**:361. doi: 10.3389/fgene.2015.00361.
19. Bryant DM, Johnson K, DiTommaso T, *et al*. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep* 2017;**18**(3):762–76. doi: 10.1016/j.celrep.2016.12.063.
20. Haas BJ, Papanicolaou A, Yassour M, *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;**8**(8):1494. doi: 10.1038/nprot.2013.084.
21. Camacho C, Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:1–9. doi: 10.1186/1471-2105-10-421.
22. Scott C. dammit, 2018. http://dib-lab.github.io/dammit/.
23. Aubry S, Kelly S, Kümpers BM, *et al*. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet* 2014;**10**(6):e1004365. doi: 10.1371/journal.pgen.1004365.
24. Jungo F, Bougueleret L, Xenarios I, *et al*. The UniProtKB/Swiss-Prot Tox-Prot program: a central hub of integrated venom protein data. *Toxicon* 2012;**60**(4):551–7. doi: 10.1016/j.toxicon.2012.03.010.
25. Nachtigall PG, Kashiwabara AY, Durham AM. CodAn: predictive models for precise identification of coding regions in eukaryotic transcripts. *Brief Bioinform* 2020;**00**(0):1–11. doi: 10.1093/bib/bbaa045.
26. Smith JJ, Undheim EA. True lies: using proteomics to assess the accuracy of transcriptome-based venomics in centipedes uncovers false positives and reveals startling intraspecific variation in Scolopendra subspinipes. *Toxins* 2018;**10**(3):96. doi: 10.3390/toxins10030096.
27. Schield DR, Card DC, Hales NR, *et al*. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* 2019;**29**(4):590–601. doi: 10.1101/gr.240952.118.
28. Suryamohan K, Krishnankutty SP, Guillory J, *et al*. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nat Genet* 2020;**52**(1:106–117. doi: 10.1038/s41588-019-0559-8.
29. Margres MJ, Rautsaw RM, Strickland JL, *et al*. The Tiger Rattlesnake genome reveals a complex genotype underlying a simple venom phenotype. *Proc Natl Acad Sci* 2021;**118**(4):e2014634118. doi: 10.1073/pnas.2014634118.
30. Hofmann EP, Rautsaw RM, Strickland JL, *et al*. Comparative venom-gland transcriptomics and venom proteomics of four Sidewinder Rattlesnake (*Crotalus cerastes*) lineages reveal little differential expression despite individual variation. *Sci Rep* 2018;**8**(1):1–15. doi: 10.1038/s41598-018-33943-5.
31. Strickland JL, Mason AJ, Rokyta DR, *et al*. Phenotypic variation in Mojave Rattlesnake (*Crotalus scutulatus*) venom is driven by four toxin families. *Toxins* 2018;**10**(4):1–23. doi: 10.3390/toxins10040135.
32. Bayona-Serrano JD, Viala VL, Rautsaw RM, *et al*. Replacement and parallel simplification of nonhomologous proteinases maintain venom phenotypes in rear-fanged snakes. *Mol Biol Evol* 2020;**37**(12):3563–75. doi: 10.1093/molbev/msaa192.
33. Freitas-de Sousa LA, Nachtigall PG, Portes-Junior JA, *et al*. Size matters: an evaluation of the molecular basis of ontogenetic modifications in the composition of Bothrops jararacussu snake venom. *Toxins* 2020;**12**(12):791. doi: 10.3390/toxins12120791.
34. Mason AJ, Margres MJ, Strickland JL, *et al*. Trait differentiation and modular toxin expression in palm-pitvipers. *BMC Genomics* 2020;**21**(1):147. doi: 10.1186/s12864-020-6545-9.
35. Petersen TN, Brunak S, Von Heijne G, *et al*. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;**8**(10):785. doi: 10.1038/nmeth.1701.
36. Kashiwabara AY, Bonadio Í, Onuchic V, *et al*. Tops: a framework to manipulate probabilistic models of sequence data. *PLoS Comput Biol* 2013;**9**(10):e1003234. doi: 10.1371/journal.pcbi.1003234.
37. Rokyta DR, Wray KP, Margres MJ. The genesis of an exceptionally lethal venom in the Timber Rattlesnake

(*Crotalus horridus*) revealed through comparative venom-gland transcriptomics. *BMC Genomics* 2013;**14**(1):394. doi: 10.1186/1471-2164-14-394.
38. Rokyta DR, Wray KP, McGivern JJ, *et al*. The transcriptomic and proteomic basis for the evolution of a novel venom phenotype within the Timber Rattlesnake (*Crotalus horridus*). *Toxicon* 2015b;**98**:34–48. doi: 10.1016/j.toxicon.2015.02.015.
39. Margres MJ, McGivern JJ, Seavy M, *et al*. Contrasting modes and tempos of venom expression evolution in two snake species. *Genetics* 2015;**199**(1):165–76. doi: 10.1534/genetics.114.172437.
40. Zhang J, Kobert K, Flouri T, *et al*. PEAR: a fast and accurate Illumina paired-end reAd mergeR. *Bioinformatics* 2014;**30**(5):614–20. doi: 10.1093/bioinformatics/btt593.
41. Rokyta DR, Lemmon AR, Margres MJ, *et al*. The venom-gland transcriptome of the Eastern Diamondback Rattlesnake (*Crotalus adamanteus*). *BMC Genomics* 2012;**13**(1):312. doi: 10.1186/1471-2164-13-312.
42. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23):3150–2. doi: 10.1093/bioinformatics/bts565.
43. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**(1):323. doi: 10.1186/1471-2105-12-323.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9. doi: 10.1038/nmeth.1923.
45. Nogueira CC, Argôlo AJ, Arzamendia V, *et al*. Atlas of Brazilian snakes: verified point-locality maps to mitigate the Wallacean shortfall in a megadiverse snake fauna. *South Am J Herpetol* 2019;**14**(sp1):1–274. doi: 10.2994/SAJH-D-19-00120.1.
46. Andrade DV, Abe AS. Relationship of venom ontogeny and diet in *Bothrops*. *Herpetologica* 1999;**55**(2):200–4.
47. Zanella N, Cechin SZ. Influência dos fatores abióticos e da disponibilidade de presas sobre comunidade de serpentes do Planalto Médio do Rio Grande Do Sul. *Iheringia Série Zoologia* 2009;**99**(1):111–4. doi: 10.1590/s0073-47212009000100016.
48. Cardoso KC, Da Silva MJ, Costa GG, *et al*. A transcriptomic analysis of gene expression in the venom gland of the snake *Bothrops alternatus* (Urutu). *BMC Genomics* 2010;**11**(1):605. doi: 10.1186/1471-2164-11-605.
49. de Paula FFP, Ribeiro JU, Santos LM, *et al*. Molecular characterization of metalloproteases from *Bothrops alternatus* snake venom. *Comp Biochem Physiol Part D Genomics Proteomics* 2014;**12**:74–83. doi: 10.1016/j.cbd.2014.09.001.
50. Rotenberg D, Bamberger E, Kochva E. Studies on ribonucleic acid synthesis in the venom glands of *Vipera palaestinae* (Ophidia, Reptilia). *Biochem J* 1971;**121**(4):609–12. doi: 10.1042/bj1210609.
51. Waterhouse RM, Seppey M, Simão FA, *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 2018;**35**(3):543–8. doi: 10.1093/molbev/msx319.
52. Junqueira-de Azevedo IL, Ching AT, Carvalho E, *et al*. Lachesis muta (Viperidae) cDNAs reveal diverging pit viper molecules and scaffolds typical of cobra (Elapidae) venoms: implications for snake toxin repertoire evolution. *Genetics* 2006;**173**(2):877–89.
53. Kumar S, Stecher G, Suleski M, *et al*. Timetree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* 2017;**34**(7):1812–9. doi: 10.1093/molbev/msx116.