




# In-depth transcriptome reveals the potential biotechnological application of *Bothrops jararaca* venom gland

Leandro de Mattos Pereira<sup>1,2</sup>, Elisa Alves Messias<sup>1</sup>, Bruna Pereira Sorroche<sup>1</sup>, Angela das Neves Oliveira<sup>1</sup>, Lidia Maria Rebolho Batista Arantes<sup>1</sup>, Ana Carolina de Carvalho<sup>1</sup>, Anita Mitico Tanaka-Azevedo<sup>3</sup> , Kathleen Fernandes Grego<sup>3</sup> , André Lopes Carvalho<sup>1</sup>, Matias Eliseo Melendez<sup>1,4,5\*</sup> 

<sup>1</sup>Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, SP, Brazil.

<sup>2</sup>Laboratory of Molecular Microbial Ecology, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil.

<sup>3</sup>Laboratory of Herpetology, Butantan Institute, São Paulo, SP, Brazil.

<sup>4</sup>Pelé Little Prince Research Institute, Curitiba, PR, Brazil.

<sup>5</sup>Little Prince College, Curitiba, PR, Brazil.

## Keywords:

*Bothrops jararaca*

Venom gland

Transcriptome

Biotechnological application

Stonustoxin

Verrucotoxin

## Abstract

**Background:** Lack of complete genomic data of *Bothrops jararaca* impedes molecular biology research focusing on biotechnological applications of venom gland components. Identification of full-length coding regions of genes is crucial for the correct molecular cloning design.

**Methods:** RNA was extracted from the venom gland of one adult female specimen of *Bothrops jararaca*. Deep sequencing of the mRNA library was performed using Illumina NextSeq 500 platform. *De novo* assembly of *B. jararaca* transcriptome was done using Trinity. Annotation was performed using Blast2GO. All predicted proteins after clustering step were blasted against non-redundant protein database of NCBI using BLASTP. Metabolic pathways present in the transcriptome were annotated using the KAAS-KEGG Automatic Annotation Server. Toxins were identified in the *B. jararaca* predicted proteome using BLASTP against all protein sequences obtained from Animal Toxin Annotation Project from Uniprot KB/Swiss-Pro database. Figures and data visualization were performed using ggplot2 package in R language environment.

**Results:** We described the in-depth transcriptome analysis of *B. jararaca* venom gland, in which 76,765 *de novo* assembled isoforms, 96,044 transcribed genes and 41,196 unique proteins were identified. The most abundant transcript was the zinc metalloproteinase-disintegrin-like jararhagin. Moreover, we identified 78 distinct functional classes of proteins, including toxins, inhibitors and tumor suppressors. Other venom proteins identified were the hemolytic lethal factors stonustoxin and verrucotoxin.

**Conclusion:** It is believed that the application of deep sequencing to the analysis of snake venom transcriptomes may represent invaluable insight on their biotechnological potential focusing on candidate molecules.

\* Correspondence: matiasemelendez@gmail.com

<https://doi.org/10.1590/1678-9199-JVATITD-2019-0058>

Received: 24 September 2019; Accepted: 16 September 2020; Published online: 21 October 2020



## Background

Animal venom is composed of a complex and potent mixture of molecules with different physiological activities, ranging from moderate effects, such as allergic reactions and dermatitis [1, 2], to more severe effects like hemorrhage, intravascular coagulation, necrosis, respiratory arrest and death [3–5]. These bioactive compounds are low explored bioresources for the development of new therapeutic drugs for different type of diseases and conditions [6–8].

Venomous snakes Snake venom is a promising source of therapeutic proteins, as these venoms comprise more than 95% of the dry weight of a snake's venom is composed of peptides/proteins [9]. These venoms comprise wide variety of enzymes such as phospholipases A<sub>2</sub>, proteases (metal and serine), L-amino acid oxidases, and esterases, as well as many other non-enzymatic proteins and peptides, which have several biochemical and pharmaceutical properties [10–13].

One venomous snake of particular interest is *Bothrops jararaca*, which is endemic to the tropical/semiotropical forest habitats of southeastern Brazil, northeastern Paraguay, and northern Argentina [14]. In recent years, high throughput technology has been implemented more often in snake venom analyses, allowing better understanding of proteomics and transcriptomics of venom, which has exposed its complexity [15, 16]. High throughput transcriptome analysis allows the identification of complete transcripts expressed in the snake venom gland [17, 18]. Moreover, molecular cloning of biotechnological proteins of *B. jararaca* requires the complete characterization of the full-length coding regions of interesting transcripts. The venom produced by *B. jararaca* has previously used to isolate bradykinin-potentiating factor (BPF), which was the basis for the creation of the antihypertensive agent captopril, and the oral anticoagulant Exanta, also known as ximelagatran [19, 20].

In this study, we generated an RNA-Seq transcriptome, performed *de novo* assembly and annotated the sequences from the venom gland of *Bothrops jararaca*. This work expands the current knowledge of the biotechnological potential of the venom gland of *Bothrops jararaca*. Furthermore, results may also be used for the discovery of novel candidates for cancer treatment, hypertension, inflammatory response, virus infection, and other human diseases, as well as the development of effective treatments of poisoning from *Bothrops jararaca* bite.

## Materials and Methods

### RNA extraction from venom gland

All experiments were performed in accordance with Brazil's National Council for the Control of Animal Experimentation (CONCEA) guidelines and were authorized by the Ethic Committee on Animal Use of the Butantan Institute, under protocol n. 4390280116. Venom glands were removed 3 days after

venom milking, when RNA transcription is at its highest level [21], from a chemically euthanized female adult *B. jararaca* under captivity (Herpetology Laboratory of the Butantan Institute) and immediately stored at -80 °C. Venom glands were then transported in dry ice to the Barretos Cancer Hospital for molecular biology analysis. Tissue samples from the venom glands (40 mg) were then disrupted and homogenized in a Precellys 24 homogenizer (Bertin Technologies) at 4,500 rpm, while being left to cool down on ice between the 3 repeated 20 second-cycles. Total RNA was isolated using RNeasy Mini kit (Qiagen) and quantified by spectrophotometry. In order to avoid cross-contamination during RNA extraction, we performed the RNA extraction in the Molecular Oncology Research Center of the Barretos Cancer Hospital, where no other source of nucleic acid of reptilian origin was ever extracted.

### Transcriptome sequencing and quality analysis

Deep sequencing of the mRNA library from *B. jararaca* was done using the Illumina NextSeq 500 platform (76 bp single-end) and the NextSeq 500/550 Mid Output v2 kit (150 cycles). Libraries were constructed using TruSeq Stranded mRNA LT Sample Prep Kits (Illumina), following the TruSeq Stranded mRNA Sample Prep HS protocol (Illumina). In order to avoid cross-contamination with transcripts of other related species, the venom gland sequenced for this work was the only reptilian sample in the Illumina chip. The quality analysis of the sequenced data was evaluated with the FASTQC software [22]. Reads were filtered with Trimmomatic v0.36 [23] using a 4-base sliding window. Leading or trailing bases with average Phred quality score lower than 20 were removed, along with adapters and reads with a length less than 50 base pairs (bp).

### *De novo* assembly and quality analysis

*De novo* assembly of the *B. jararaca* transcriptome was performed using Trinity [18] with default k-mer size of 25. The statistics of the Trinity Assembly like the Nx statistics (eg. the contig N50 value), total trinity genes, and total of trinity transcripts were obtained with perl script 'TrinityStats.pl' of the Trinity toolkit [24]. Gene open reading frames (ORFs) or protein coding regions within the transcripts were predicted with the TransDecoder program [24]. CD-HIT-EST version 4.6.1 [25] was subsequently used for clustering predicted proteins with 100% of sequence identity and 100% alignment coverage. For analysis of transcript abundance, the sequenced paired-end reads were realigned with the assembled transcripts for quantification by the RSEM (RNA-Seq by Expectation Maximization) software [26] using the Trinity script 'align\_and\_estimate\_abundance.pl' present in the Trinity toolkit [18]. The transcriptome completeness and contiguity of *B. jararaca* was assessed by comparing the assembly transcripts to benchmarking sets of the universal single-copy (BUSCO) of Eukaryota, Metazoa, and Vertebrata using BUSCO v3, based on evolutionarily informed expectations of gene content from

near-universal single-copy orthologs selected with the database OrthoDB v9 [27, 28]. Full-length transcript or near full length transcript of *B. jararaca* was identified using BLASTX (BLAST+ v2.2) by alignment against the predicted proteome of *Anolis carolinensis* (GCA\_000090745.1, GenBank ID), *Ophiophagus Hannah* (GCA\_000516915.1, GenBank ID), and *Python bivittatus* (GCA\_000186305.2, GenBank ID), which were obtained from GENOME (<https://www.ncbi.nlm.nih.gov/genome/>) and UniProtKB/Swiss-Prot release 2019\_04 May-08, 2019 using an E-value cut-off set to  $1 \times 10^{-20}$ .

Full-length transcript or near full-length transcript are defined as transcripts (*query*) similar to proteins already annotated in reference genomes with high quality standards of genome completeness and functional annotation. The alignment between the transcripts (*query*) and the reference sequence protein obtained in the the BLASTX covered across more than 80-90% of the transcript length. The resulting table with full-length transcript or near full-length transcript from BLASTX was obtained with the perl script 'analyze\_BLASTPlus\_topHit\_coverage.pl' from Trinity toolkit [24]. For each BLAST hit in the target database of the protein, the best matching Trinity transcript was selected, and the percent of the BLAST hit's length covered by the Trinity transcript was identified.

### Functional annotation of transcriptome

Annotation was performed using Blast2GO version 3.2 [29]. All predicted proteins after clustering step were blasted against the non-redundant protein database (NR) of NCBI (<ftp://ftp.ncbi.nih.gov/blast/db/>; 29-02-2015) using BLASTP with an E-value cut-off set to  $1 \times 10^{-6}$ . The metabolic pathways present in the transcriptome were annotated using the KAAS - KEGG Automatic Annotation Server [30], which provides functional annotation of genes via BLAST using the method BBH (bi-directional best hit), against the manually curated KEGG GENES database. With the KEGG Orthology groups (KOs) identified via KAAS [30], the complete functional modules of the metabolic pathways present in the *B. jararaca* transcriptome were reconstructed using the tool KEGG Mapper tools [31].

### Toxin identification

The toxins were identified in the *B. jararaca* predicted proteome using BLASTP with an E-value cut-off set to  $1 \times 10^{-6}$  and using the protein sequences obtained from the Animal Toxin Annotation Project (version 31/10/2018) from Uniprot KB/Swiss-Pro database as a reference [32]. The BLASTP annotation results obtained of alignment against the Animal Toxin Annotation Project database and with same annotation description obtained from the BLASTP against the NR database were considered toxins. In addition, all inhibitors or tumor suppressor proteins were identified in the annotation from the BLASTP results against the NR database using the search terms: inhibitors AND tumor suppressor and later these results were manually checked. For

identification of the full length transcript of *B. jararaca* that aligns to the Animal Toxin Annotation Project from Uniprot KB/Swiss-Pro database [32], we processed the BLASTx hits (E-value cut-off set to  $1 \times 10^{-20}$ ) using the 'analyze\_BLASTPlus\_topHit\_coverage.pl' script from the Trinity package (<http://trinityrnaseq.sourceforge.net/>), as described above.

### Phylogenetics analysis of stonustoxin and verrucotoxin

The sequences of stonustoxin and verrucotoxin of *B. jararaca* identified were aligned with BLASTP program (<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>) and homologous sequences recovered with an E-value cut-off above of  $1 \times 10^{-20}$  were used for Phylogenetic analyses reconstruction. Phylogenetic analyses were performed in the platform NGPhylogeny.fr [33] using the method Advanced PhyML + SMS. In this mode, the program of alignment is the MAFFT [34]. The parameters of matrix of distance chosen in the program MA FFT was BLOSUM62. After the alignment, the sequences were curated with BMGE [35] for gaps remotion and to keep the informative sites. Then, the Smart Model Selection in PhyML program assesses and chooses the best evolutionary model for phylogenetic analyzes. For phylogenetic analysis was used the program PhyML [36] with maximum likelihood method and the inference of Branch support in the tree as done with Bootstrap (FBP + TBE) using 1000 bootstrap replicates. All other parameters for program BMGE and Newick tree format Display were kept as default. All steps were done in the program NGPhylogeny.fr (<https://ngphylogeny.fr/>).

### Data visualization

Figures and data visualization were performed using the ggplot2 package in R [37], through the R software [38].

## Results

### Transcriptome sequencing, de novo assembly and full-length transcript analysis

Deep sequencing of the venom gland transcriptome of *B. jararaca* was performed using a NextSeq 500 (Illumina), generating 67,551,639 unpaired reads. Transcriptome raw data quality showed a Phred quality score (per base sequence quality average) higher than 30 (Additional file 1). *De novo* assembly of all reads resulted in a total assembly of 64,853,458 bp representing 76,765 genes (N50 length of 1104 bp) and 96,044 transcripts (N50 length of 1,104 bp), with a mean contig length of 675.25 bp and a GC percentage of 43.56% (Table 1).

The completeness of transcriptome assessment performed by the Benchmarking Universal Single-Copy Orthologs (BUSCO; version 3.0) showed that 85.8% of the 303 core eukaryotic genes and 88.7% of the 978 core metazoan genes were found in our

*B. jararaca* transcriptome assembly (Table 2). Furthermore, the BUSCO analysis with 2,586 core vertebrata genes showed that 1,550 (59.9%) and 543 (21%) of the 2,586 expected vertebrata genes were identified as complete or fragmented, respectively, while 493 (19.1%) genes were considered missing.

The quantity of full-length transcripts, or near full-length transcripts in *B. jararaca*, was determined by the number of predicted sequences of Reptiles and UniprotKB/Swiss-Prot

databases (Table 3). Among the 96,044 transcripts of *B. jararaca*, 2,076 (2.16%) matched near full length with coding sequences of *Anolis carolinensis*, 4,555 (4.74%) with *Ophiophagus hannah*, 6,707 (6.98%) with *Python bivittatus*, and 5,472 (5.69%) with UniprotKB/Swiss-Prot (Table 3 and [Additional file 2](#)). Thus, a total of 6,835 full length or near full-length transcripts were identified using the concatenated predicted proteome of all three reptiles cited above.

**Table 1.** Statistics of Trinity *de novo* assembly.

<b>Global Trinity Stats</b>	
Total trinity 'genes' counts	76 765
Total trinity transcripts counts	96 044
Percent GC (%)	43.56
<b>Stats based on all transcript contigs</b>	
Contig N10	3 774
Contig N20	2 674
Contig N30	2 012
Contig N40	1 512
Contig N50	1 104
Median contig length	362
Average contig (%)	675.25
Total assembled bases	64 853 458
<b>Stats based only on longest isoform per 'gene'</b>	
Contig N10	3 302
Contig N20	2 264
Contig N30	1 642
Contig N40	1 179
Contig N50	828
Median contig length	335
Average contig	585.97
Total assembled bases	44 982 090

**Table 2.** Summary of transcriptome completeness assessment by BUSCO notation.

<b>Eukaryotic genes</b>	
Complete and single-copy BUSCOs	222 (73.3%)
Complete and duplicated BUSCOs	38 (12.5%)
Fragmented BUSCOs	35 (11.6%)
Missing BUSCOs	8 (2.6%)
<b>Total BUSCO groups searched</b>	<b>303 (100%)</b>
<b>Metazoan genes</b>	
Complete and single-copy BUSCOs	683 (69.8%)
Complete and duplicated BUSCOs	185 (18.9%)
Fragmented BUSCOs	82 (8.4%)
Missing BUSCOs	28 (2.9%)
<b>Total BUSCO groups searched</b>	<b>978 (100%)</b>

### Functional annotation

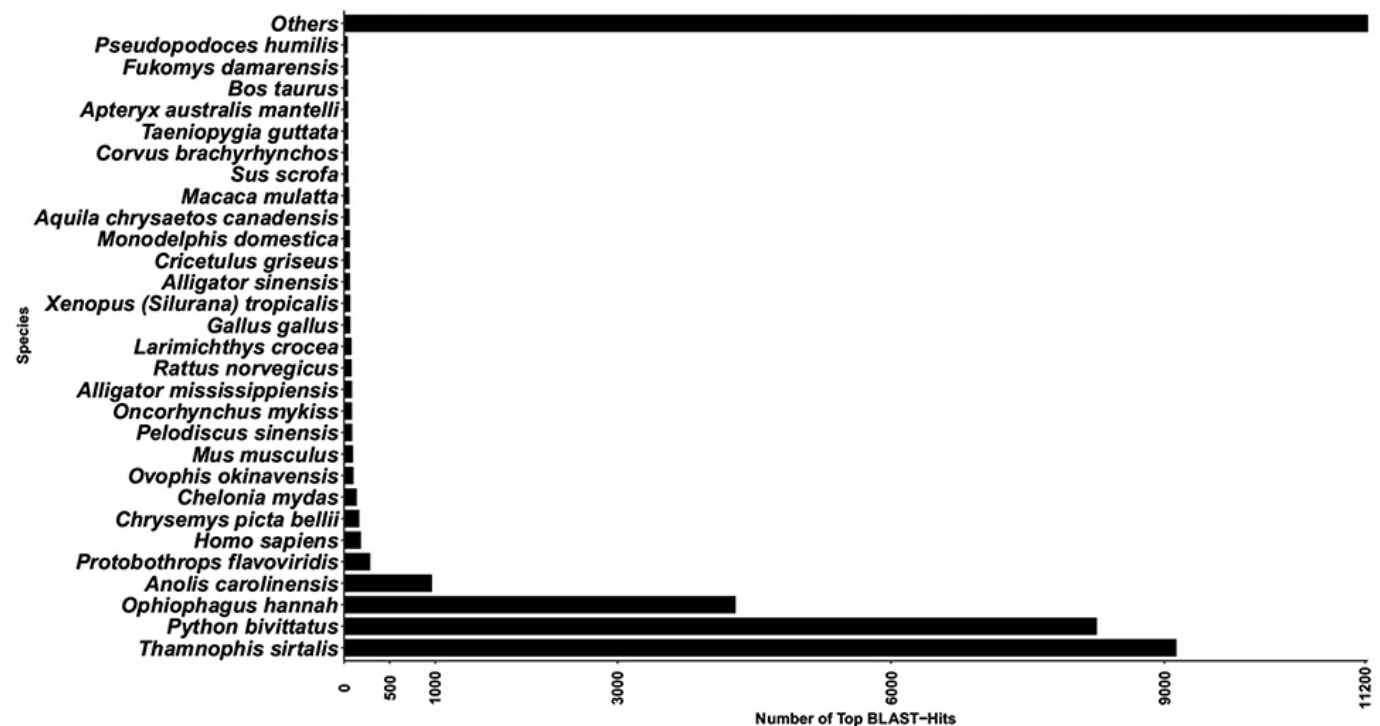
From the 96,044 transcripts identified, 49,345 open reading frames (ORF) were predicted using TransDecoder [10] with *ab initio* model. In this analysis, only predicted ORFs that were at least 60 amino acids long were retained. The clustering steps of predicted proteins in the CD-HIT program (100% amino acid identity) resulted in a set of 41,916 non-redundant sequences. Functional annotations of these coding regions were inferred using BLASTP against the NR database in NCBI. Additional annotation was performed using InterProScan [39] and Gene Ontology [40] using the program Blast2GO [29]. Thus, 78.81%

of unique predicted proteins (33,034) were annotated, while 8,162 (19.47%) transcripts did not have similarity against any proteins in the NR database (Additional file 3). The E-value distribution of hits obtained against NR, the average number of hits per sequence and the High Scoring Segment Pair/Coverage distribution are shown in Additional files 4, Additional file 5 and Additional file 6, respectively. Among the 33,034 annotated sequences, 22,933 (54,7%) best hits were matched in BLAST Top-hits against predicted proteins of the species *Thamnophis sirtalis*, *Python bivittatus*, *Ophiophagus hannah*, *Anolis carolinensis*, and *Protobothrops flavoviridis* (Figure 1). The other 1,858 sequences

**Table 3.** Full-length transcript reconstruction analysis of *B. jararaca* venom gland transcriptome in relation to *Anolis carolinensis*, *Ophiophagus hannah*, *Python bivittatus* and UniprotKB/Swiss-Prot.

<i>Anolis carolinensis</i>			<i>Ophiophagus hannah</i>			<i>Python bivittatus</i>			UniprotKB/Swiss-Prot		
Pct_cov_Hit (%)	PC	PCS	Pct_cov_Hit (%)	PC	PCS	Pct_cov_Hit (%)	PC	PCS	Pct_cov_Hit (%)	PC	PCS
100	1438	1438	100	3069	3069	100	4974	4974	100	3739	3739
90	367	1805	90	834	3903	90	928	5902	90	1034	4773
80	271	2076	80	652	4555	80	805	6707	80	699	5472

Cumulative number of protein of the *A. carolinensis*, *O. hannah*, *Python bivittatus* and UniprotKB-Swiss-Prot databases recovery by BLASTX that aligned by at least one transcript in the assembly *B. jararaca* transcriptome across at 80-100 percentage (%) of coverage. The transcripts identified in *B. jararaca* were annotated as full-length transcripts if they match a protein in the reference proteome database at E-value threshold of  $1e^{-20}$ . Pct\_cov\_Hit: percentage of coverage of top matching hits of reference proteome that align across more than X% (80-100) with a transcript of *B. jararaca*. PC: protein counts of target reference proteome that aligned by at least one transcript of *B. jararaca*. PCS: protein count sum of reference proteins of reference proteome that aligned at X% (80-100) coverage by at least one transcript of *B. jararaca*.



**Figure 1.** Top BLAST hit distribution of predicted proteins from of *Bothrops jararaca* venom gland transcriptome. Recovery by Blast2GO with similarity filter parameter of 55%, E-Value-Hit-Filter:  $10^{-6}$ .



had best hits against other organisms, such as reptiles, humans, mammals, fishes, bacteria, virus, fungi, and others (Additional file 7).

### Gene ontology mapping with Blast2GO

The predicted proteins identified in *B. jararaca* venom gland were mapped into putative functional group-based Gene Ontology (GO) terms assigned using NR derived BLAST hits and InterProScan using Blast2GO, categorized into three ontologies: biological processes (BP), cellular components (CC) and molecular functions (MF) (Figure 2).

### Enzymes and metabolic pathways identification

We identified 4,203 predicted proteins in the KO (orthologous groups) of the signal transduction metabolic pathway (Additional file 8) through the orthologous assignment KAAS – KEGG server [30]. Using 91 modules of the KEGG Reconstruct Module [41] tools, predicted proteins of the Proteasome 20S core particle (M00340), Proteasome 19S regulatory particle (PA700) (M00341), Ski complex (M00392), DNA polymerase delta complex (M00262), SCF-BTRC complex (M00380), SCF-SKP2 complex (M00381), Cul4-DDB1-DDB2 complex (M00385), and ECS complex (M00388) were identified (Additional file 9). A total of 389 metabolic pathways were annotated as having at least 1 predicted protein of *B. jararaca* described in the KEGG among the metabolic pathways identified (Additional file 10).

### Functional domains identified

All predicted proteins obtained from the transcript ORFs were searched for functional domains signatures present in the Smart [42], PFAM [43], and Superfamily [44] thought of the profile hidden Markov models with InterProScan [39]. Overall, a total of 20,605 predicted proteins were categorized into 3,992 domain/family signatures with PFAM (Figure 3).

### Transcript abundance in *B. jararaca*

The top 20 genes and alternatives isoforms codified with higher expression inference of gene abundance in transcripts per million (TPM) determined by RSEM software [20] were represented by zinc metalloproteinase-disintegrin-like jararhagin, natriuretic peptide, metalloprotease, C-type lectin 8a, snake venom serine protease HS114, proteasome 26S subunit, non-ATPase 14, serine endopeptidase, metalloproteinase type II 4, acidic secretory phospholipase A2 sPLA2-II, serine proteinase 20a, metalloprotease BOJUMET II, putative disulfide-isomerase, sphingomyelin phosphodiesterase-like, metalloproteinase type III 8, snake venom serine protease homolog, L-amino-acid oxidase, and snake venom vascular endothelial growth factor toxin (Additional file 11).

### Toxins, inhibitors and tumor suppressors

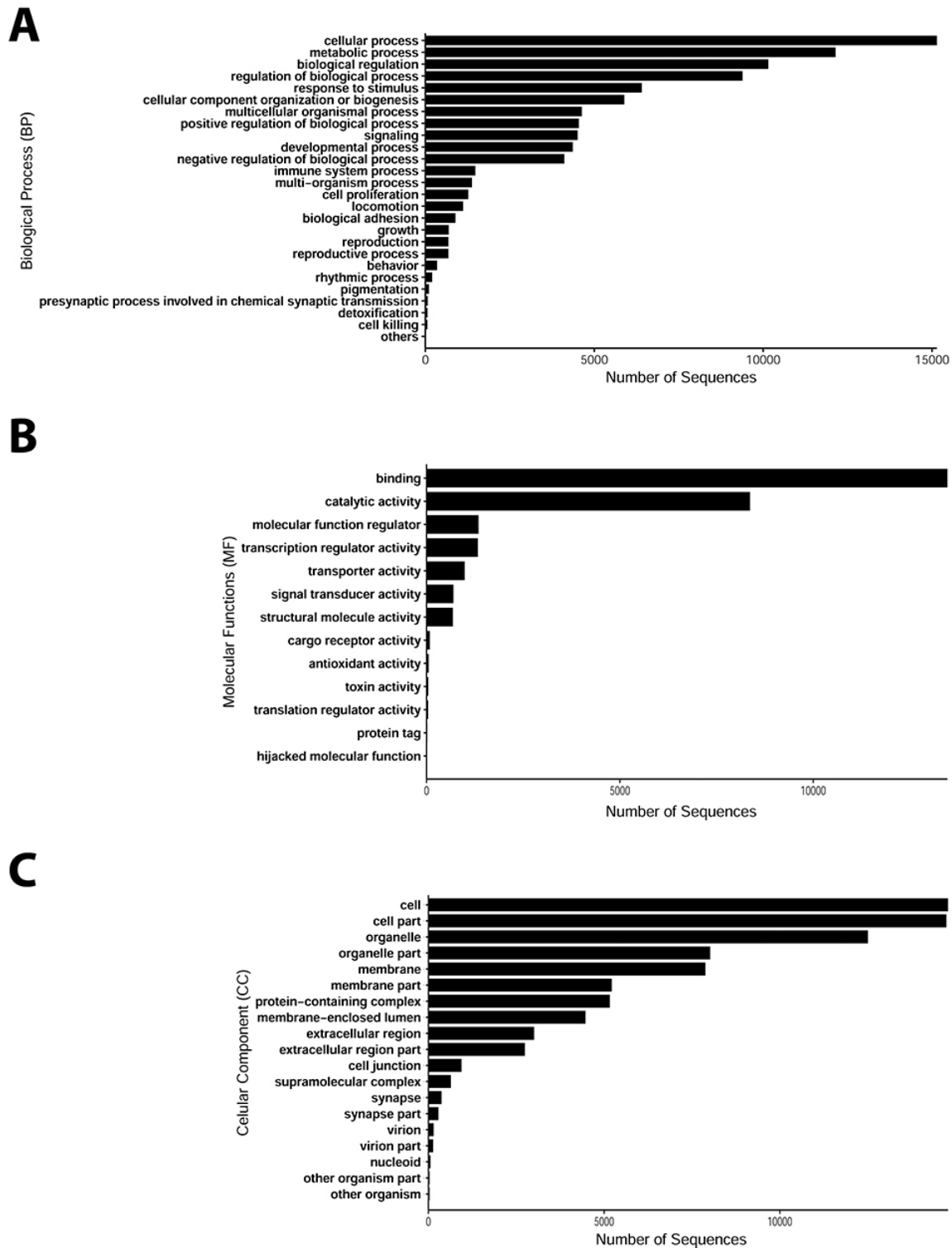
Known animal toxins encoded in the *Bothrops jararaca* transcriptome, using all of the sequences from the Animal Toxin Annotation Project [45] as a reference were also searched. The 831 hits obtained from this approach (Additional file 12) were further validated in our BLASTP annotation against the NCBI non-redundant (NR) database, confirming a same description for a total of 525 coding sequences in both databases (Additional file 13). This set encoded 83 toxin and inhibitors related functional classes (Additional file 14), the more abundant (> 1%) are showed in the Figure 4. Among the less abundant toxins identified, minor toxins, such as venom factor (n = 4), L-amino acid oxidase (n = 4), phospholipase B (n = 4), Mannan-binding lectin serine protease 1 (n = 1), Mannan-binding lectin serine protease 2 (n = 3), verrucotoxin (n = 3), translationally-controlled tumor protein (n = 3), phospholipase A<sub>1</sub> (n = 3), nerve growth factor (n = 2) and stonustoxin subunit alpha (n = 2) were also found (Additional file 14).

Moreover, several inhibitors were also found, such as phosphatase inhibitor 2 (n = 1), kunitz-type protease inhibitor 1 (n = 1), kunitz-type protease inhibitor 2 (n = 1), kunitz-type protease inhibitor 4 (n = 1) and protease inhibitor 3-like (n = 1) (Additional file 14).

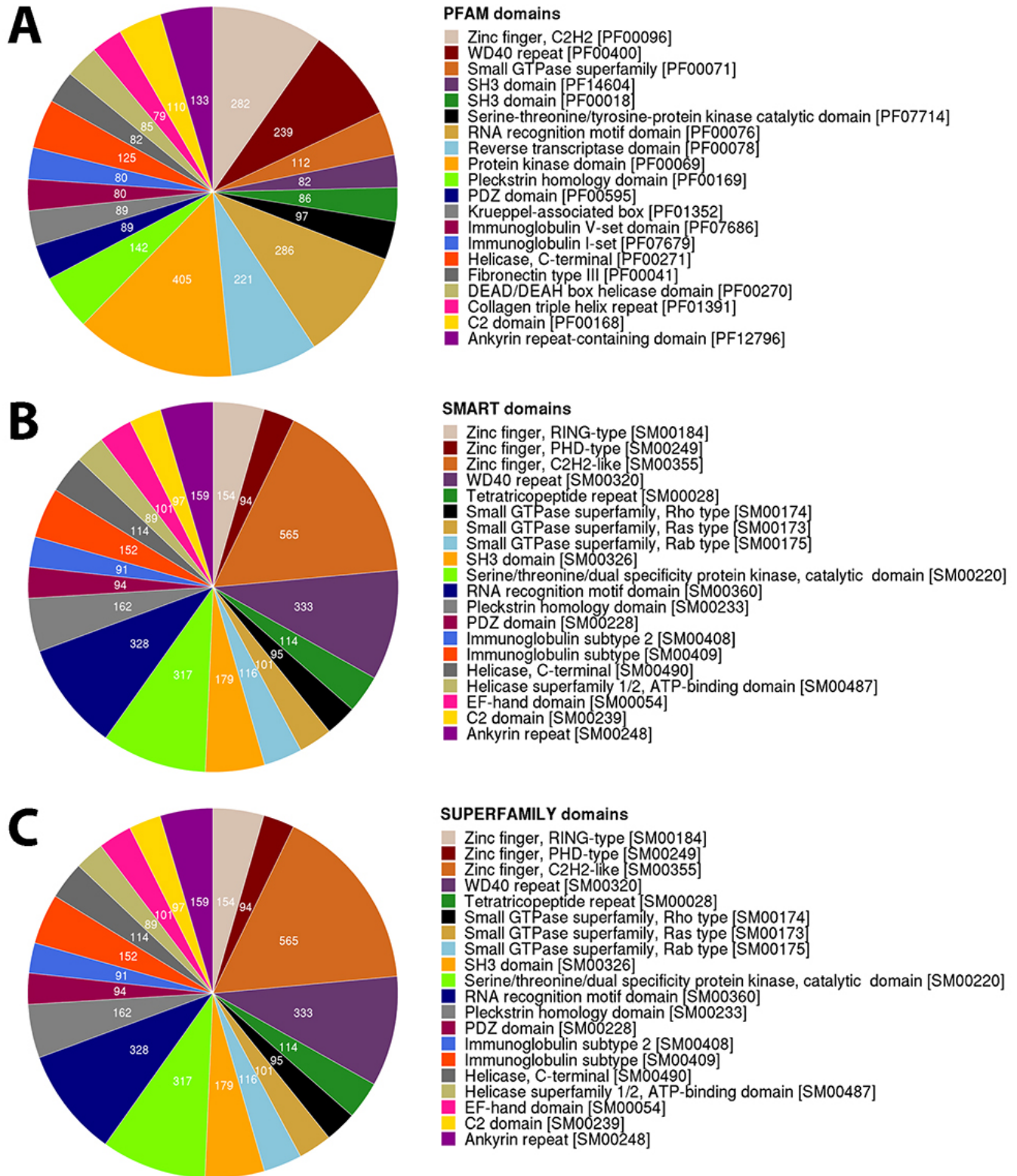
In addition, the 107 full-length transcripts or nearly full-length transcripts of *B. jararaca* encoding toxins were identified by BLASTX recovery against the Animal Toxin Annotation Project (version 31/10/2018) from Uniprot KB/Swiss-Pro database and with the same annotation obtained when NCBI non-redundant database was used as reference. Among these, we annotated transcripts for the vascular endothelial growth factor, acidic phospholipase A<sub>2</sub>, phospholipase B, zinc metalloproteinase-disintegrin-like, venom nerve growth factor, among others (Additional file 15).

### Phylogenetic analysis of stonustoxin and verrucotoxin

We identified two genes (DN34719\_i1, DN56183\_i1) that codified similar (41.1%, using EMBOSS water local alignment program) sequences to stonustoxin subunit alpha; and two genes codifying (DN37544\_i1 and DN1850) similar (41.8%, using EMBOSS water local alignment program) sequences to verrucotoxin subunit beta-like in the annotation recovered by BLASTP against the Nr database (Figure 5 and Additional file 16). Subsequently, the homologous sequences for these proteins were recovery by BLASTP against NR database of NCBI using an E-value cut-off set to 10<sup>-20</sup>. The bootstrapped analysis (1000:100%) of the phylogenetic tree revealed that *Acipenser ruthenus*, *Anabarrilius grahami*, *Salvelinus alpinus*, *Lacerta agilis*, *Anolis carolinensis*, *Bothrops jararaca* (DN37544\_i1, DN56183), *Python bivittatus*, *Bothrops jararaca* (DN34719\_1), *Apteryx rowi*, *Alligator sinensis*, *Trachemys scripta elegans*, *Pelodiscus sinensis*, *Chelonoidis abingdonii*, *Gopherus evgoodei*, *Chelonia mydas*, *Terrapene carolina triunguis*, *Chrysemys picta bellii*

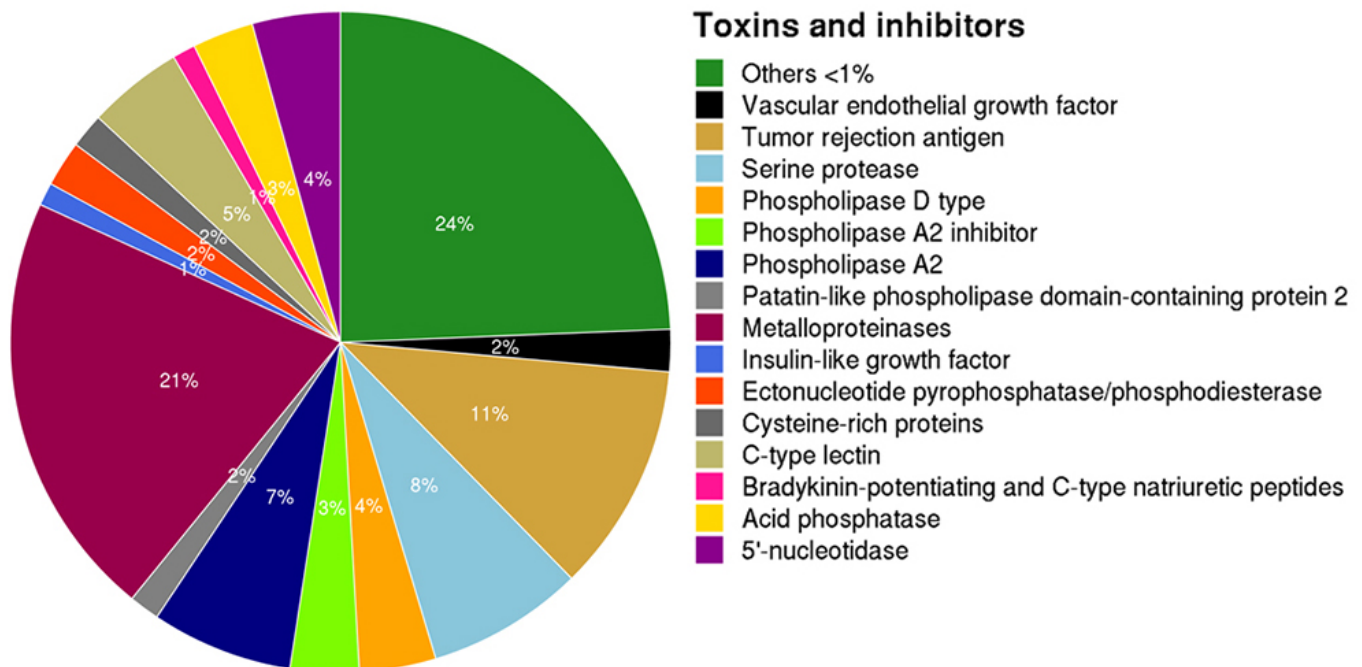


**Figure 2.** Gene Ontology category classification at level 2 and functional distribution of the transcriptome of *B. jararaca* performed by Blast2GO. The predicted proteins were functionally mapped according to the three major classifications of Gene Ontology: **(A)** biological process (BP), **(B)** molecular function (MF) and **(C)** cellular component (CC). They were annotated by setting the following parameters – E-Value-Hit-Filter:  $10^{-6}$  and others parameters default.

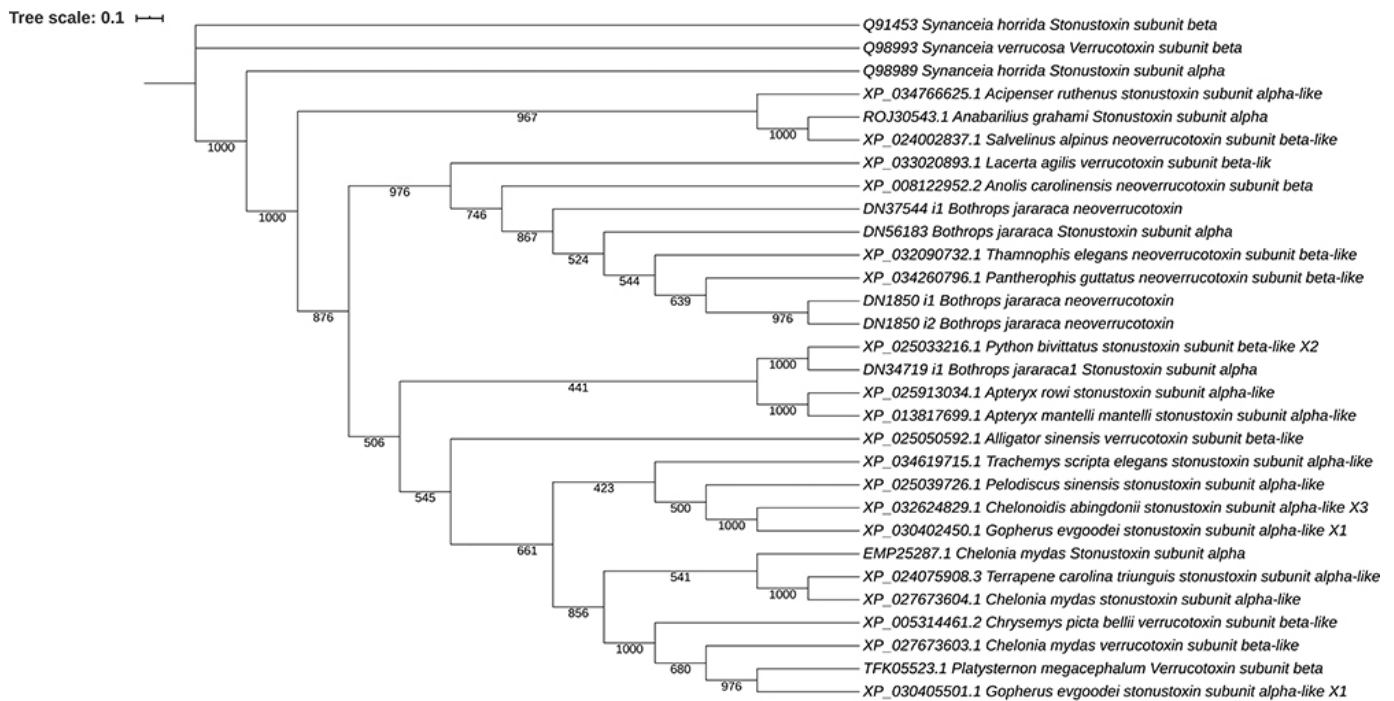


**Figure 3.** Functional domain annotation of predicted proteins of *B. jararaca* transcriptome with InterProScan: **(A)** PFAM, **(B)** SMART and **(C)** SUPERFAMILY databases.





**Figure 4.** Functional class annotation of toxins and accessory family proteins identified in *B. jararaca* venom using Animal Toxin Annotation Project as reference.



**Figure 5.** Phylogenetic tree of stonustoxin and verrucotoxin proteins.

and *Platysternon megalcephalum* have a common origin with stonustoxin subunit alpha Q98989. The stonustoxin of *Synanceia horrida* 13279 have a functional domain SPRY (IPR003877, access identifier in Interproscan), but the sequences of snakes, reptiles and turtles have a functional domain GTP binding (PF01926: access identifier in Pfam) called MMR\_HSR1 (PF01926) and one domain fibronectin type 3 (SM000060:

access identifier in Smart). The *B. jararaca* stonustoxin have one domain MMR\_HSR1 and one domain fibronectin type 3. The sequences annotated as verrucotoxin subunit beta-like in *B. jararaca* have a domain fibronectin type 3. The verrucotoxin of *Synanceia horrida* 13279 also has a domain SPRY as well as stonustoxin subunit beta.

## Discussion

More than 90% of the dry weight of snake venom is represented by peptides and proteins, potentially representing a natural arsenal of biotechnological relevant proteins [13, 46]. Traditional snake venom protein purification involves the extraction of crude venom, followed by purification via physical or chemical methods, such as high-performance liquid chromatography purification, which produces small amounts of purified products, but not all the products are purified. While this procedure is well standardized, the amount of peptides/proteins purified is not enough for pre-clinical or clinical studies. Less abundant snake venom protein classes, well reviewed by Boldrini-França et al. [47]. With the intent to further increase the amount of known proteins, we described here the transcriptome from the *B. jararaca* venom gland and its annotations. Data presented in this study can be used to produce proteins in heterologous expression systems, such as Eukaryotic, bacteria or yeast cell cultures. Recent advances in molecular biology and genomics has made possible the analysis of whole transcriptomes from different tissue sources, being found five publications describing transcriptome analyses of *B. jararaca* [14, 46–51].

An initial study of the transcriptome of an adult specimen of *B. pauloensis* caught in São Paulo State (Brazil) was described by Rodrigues and collaborators in 2012 [52], using an Expressed Sequence Tag (EST) singleton library of 668 EST sequences. Moreover, a transcriptome analysis using deep sequencing approach was recently published by Gonçalves-Machado et al. [14], in which they compared the transcriptomes of two *B. jararaca* populations from the Brazilian Southern (S) and Southeastern (SE) Atlantic rainforest. This transcriptome analysis was performed with 205,449 (SE)/281,569 (S) reads and generated 14,246 (SE)/12,240 (S) contigs, describing 15 (SE)/16 (S) different family proteins [14]. Similarly, Junqueira-de-Azevedo et al. [49] described the transcriptome of the venom gland of an adult *B. jararaca*, generating 116,236 reads. In the present work, we describe the transcriptome analysis of *B. jararaca*, with 67,551,639 unpaired reads, which was assembled using Trinity. This analysis generated 76,765 genes with an N50 length of 828 bp along with 96,044 transcripts with an N50 length of 1,104 bp and a total assembly of 64,853,458 bp with 41,916 unique predicted proteins. To date, an approximate N50 value of 1,431 bp and average contig length of 894 bp was reported for the transcriptome of the snake *Bothrops moojeni* [53], but other works [8, 10, 41–44] did not describe this value.

Our transcriptome of the vertebrata core genes was more complete than any previously sequenced transcriptome of other snakes [54]. However, it is expected that not all genes are expressed in the venom gland. BUSCO recovery tends to be highest when the entire organism and/or multiple developmental stages or the same organism is used to generate the assemblies, compared to those assembled from a select number of tissues [55]. The BUSCO results of our assembly generated here was comparable to the other transcriptomes reported, where recovery varied between 68 and 95% [56–59].

Our predicted proteome had a greater number of predicted proteins with high identity or best hit obtained from BLASTP against the proteomes available for *Thamnophis sirtalis*, *Ophiophagus hannah*, *Python bivittatus*, *Anolis carolinensis*, and *Protobothrops flavoviridis* (Figure 1) (Additional file 7). We also identified and annotated a higher number of predicted proteins than these previously deposited genomes, which could make the current transcriptome analysis a reference or complement for the functional knowledge of the coding proteins present in other snakes or reptiles.

We identified 107 full-length or near full-length transcripts (80–100% coverage length of query in relation to reference toxins) which were defined as toxins or inhibitors (Additional file 13). These transcripts included hemolytic lethal factor stonustoxin [60, 61] and verrucotoxin [62], which were identified for the first time in snakes. We identified MASP1 and MASP2 that are of alternative pathways of complement activation, involved in the activation of factor D [63], complement factor I, complement C1r subcomponent and complement component C7, which possibly participate in the tissue damage in a host bitten by *B. jararaca* and may aid in the venom poisoning. The excessive complement activation and immune activation could exacerbate the severity of the tissue injury [64].

Several predicted toxins were identified for the first time in this transcriptome, such as veficolin-1, ryncolin-1-like, pancreatic alpha-amylase, venom allergen 3-like, stonustoxin subunit alpha and verrucotoxin (Additional file 14). Moreover, different inhibitors and potential tumor suppressors were also found, not yet reported in the venom gland of *B. jararaca*, such as the relA-associated inhibitor, ribonuclease inhibitor, reversion-inducing cysteine-rich protein with Kazal motifs and Insulin-like growth factor-binding protein 6 (Additional file 14). Moreover, up to now, the cobra venom factor and the three-finger like transcripts were never described in venom glands of *B. jararaca* [61, 65, 66].

We also identified two genes coding toxins similar to stonustoxin subunit beta (DN34719\_i1, TPM:105, DN56183, TPM: 17.00) and to verrucotoxin (DN37544\_i1, TPM: 16, DN1850, TPM: 34), which has been studied in venom of fishes of the species *Synanceia verrucosa* [67]. The phylogenetic analysis indicated that *B. jararaca* stonustoxin (DN34719\_c0\_g1) and verrucotoxin (DN1850\_c0\_g1) are related to stonustoxin and verrucotoxin of *Synanceia verrucosa*. Results for stonustoxin indicated that this protein diverged from an ancestral node, with a bootstrap value of 100% significance (Figure 5 and Additional file 15). However, these proteins have different functional domains predicted. Hence, it is not possible to affirm that these homologous sequences have the same function, because neofunctionalization is a common evolutionary process found after speciation in homologous or paralogous sequences.

The *Synanceia verrucosa* inhabits shallow waters of the tropical or subtropical Indo-Pacific regions and are among the most venomous and dangerous fishes in the world. The purified stonefish toxins commonly present potent hemolytic activities due to its ability to form pores in the cell membrane. Also,

these toxins elicit potent hypotension, inhibit neuromuscular function, and induce cardiovascular collapse in humans and native predators [54, 68]. The stonustoxin-like were also found in the transcriptome of the venom gland of annelids [69] and are found in a variety of vertebrates, including the common ostrich, platypus, tasmanian devil, and coelacanth [70].

We identified with KEGG mapper [71] one of the largest mapped/annotated pathways of the components of the human retrovirus response. Subsequently, one of the identified components of the Gene Ontology term [34] identified by Blast2Go in our analysis is the virion part, consisting of sequences coding for several endogenous retrovirus described in snakes such as *Python curtus endogenous retrovirus*, *Python molurus endogenous retrovirus*, Endogenous retrovirus group PABLB member 1, Endogenous retrovirus group K members of families 1, 8, 9, 10, 11, 18, 19 and 25, and human endogenous retroviruses (HERVs) such as HCML-ARV already isolated from blood cells of patient with chronic myeloid leukemia [72].

One important finding was the identification of Syncytin and L1-retrotransposons genes in the venom gland of *B. jararaca*. Syncytin genes are associated with placental evolution in mammals and viviparous animals [73]. Up to now, these coding sequences were never described in transcriptomes of viviparous snakes, mainly *B. jararaca*. Only recently it was found in the *Mabuya* lizards [74].

In addition, we also identified the metabolic pathways essential for survival such as thermogenesis, catabolism and anabolism of carbons, amino acids metabolism, fatty acid metabolism, pathways involved with splicing of RNA, and processing and degradation of proteins (Additional file 11). Among the most frequent signatures of functional domains identified in our work was the kinase domain and serine-threonine/tyrosine-protein kinase catalytic domains, which are known to regulate or activate most cellular pathways. Similarly, the RNA recognition motif was also abundant and acts in the recognition of RNA and proteins known to bind single-stranded RNAs [75]. Other prominent functional domains found within the *B. jararaca* transcriptome were the zinc finger domain, which is a DNA, RNA, protein or lipid binding domain [76]; as well as metalloproteases; WD40 repeats, involved in several functions such regulation to cell cycle control and apoptosis [77]; a reverse transcriptase domain, characteristic of retroviruses involved in replication in the host; and a pleckstrin homology domain, with a role in recruiting proteins to different membranes [78]. These signatures reflected the biologic role of the proteins identified.

The additional analysis of RSEM showed that the most abundant transcripts genes belong to zinc metalloproteinase-disintegrin-like jararhagin, natriuretic peptide metalloprotease, C-type lectin 8a and L-amino acid oxidase. The distribution of these protein types is characteristic of the *Bothrops* genus, whose species produce venoms most notable for local tissue damage such as edema, hemorrhage, and necrosis, which has already been described by other researchers [60].

So far, our work represents the largest transcriptome analysis of *B. jararaca* venom gland. All raw and analyzed data are available for further studies, making possible further exploitation for biotechnological uses.

## Conclusion

Our transcriptome analysis strategy yielded unique insights into the diversity of *B. jararaca* venom gland transcriptome toxins. The present results bring an important contribution to the development of snake venom-derived proteins as potential biotechnological sources, especially for the search and development of candidate molecules.

## Acknowledgments

The authors would like to thank High Performance Computing Lab – LAD/PUCRS for allowing access to run the high-throughput sequences analyses.

## Availability of data and materials

The raw FASTQ files analyzed during the current study are publicly available and were deposited in BioProject under accession number PRJNA549912. All data tables of functional annotation generated during this study are included in this published article (as Additional files). The FASTA files of all ORFs generated during the current study are not publicly available due to further analysis of the datasets in our research, but are available from the corresponding author on reasonable request.

## Funding

This study was supported by the Barretos Cancer Hospital (PAIP – Research Support Incentive Program).

## Competing interests

The authors declare no potential competing interests.

## Authors' contributions

LMP and EAM contributed equally to this work. LMP was responsible for conceptualization, bioinformatics analysis, interpretation of data, drafting of the article and revision of the manuscript. EAM carried out molecular biology experiments and took part in drafting of the article. BPS and ANO participated in molecular biology experiments. AB, LMRBA, ACC and ALC were in charge of the revision of the manuscript. AMTA and KFG were involved in animal care and biological sample preparation. MEM was responsible for the conceptualization and design of the project, search for financial support, analysis, interpretation of data and revision of the manuscript.

## Ethics approval

The present study was approved by the Ethics Committee on Animal Use of the Butantan Institute under protocol n.



4390280116. All experiments were performed in accordance with Brazil's National Council for the Control of Animal Experimentation (CONCEA) guidelines.

### Consent for publication

Not applicable.

### Supplementary material

The following online material is available for this article:

**Additional file 1.** Phred quality score (Sanger encoding, Phred+33 format) obtained with FastQC program.

**Additional file 2.** Full-length transcript identified in *B. jararaca* transcriptome using UniprotKB Swiss database.

**Additional file 3.** Distribution of annotated number of sequences with BLASTP using with BLASTP using Nr database, Interproscan and Gene Ontology through of Blast2GO.

**Additional file 4.** E-value distribution of all predicted proteins obtained from *Bothrops jararaca* transcriptome.

**Additional file 5.** BLAST hit distribution of all predicted proteins obtained from *Bothrops jararaca* transcriptome.

**Additional file 6.** HSP/Seq coverage distribution of all predicted proteins obtained from *Bothrops jararaca* transcriptome.

**Additional file 7.** BLAST hits against predicted proteins.

**Additional file 8.** KO annotation in the predicted proteome of *B. jararaca* with KAAS – KEGG server.

**Additional file 9.** KO annotation in the predicted proteome of *B. jararaca* with KEGG Reconstruct Module.

**Additional file 10.** Metabolic pathways with at least one predicted protein of *B. jararaca* described in the KEGG annotation.

**Additional file 11.** Top 20 more abundant transcripts and their isoforms (based on TPM values) identified in the transcriptome of *B. jararaca*.

**Additional file 12.** Animal toxins encoded in the *Bothrops jararaca* transcriptome identified in the Animal Toxin Annotation Project.

**Additional file 13.** Animal toxins encoded in the *Bothrops jararaca* transcriptome, identified in both Animal Toxin Annotation Project and NCBI non-redundant (NR) database.

**Additional file 14.** Toxins and inhibitors identified in the predicted proteome of *B. jararaca*.

**Additional file 15.** Function of toxins and accessory proteins identified in the predicted proteome of *B. jararaca*.

**Additional file 16.** Alignment of proteins of stonustoxin and verrucotoxin subunit beta.

### References

- Pontes LG, Cavassan NR, Creste CF, Lourenço A Jr, Arcuri HA, Ferreira RS, et al. Crotoxin: a novel allergen to occupational anaphylaxis. *Ann Allergy Asthma Immunol*. 2016 Jun 1;116(6):579-81.e1. doi: 10.1016/j.ana.2016.03.015.
- Utup MS, Jamal MS. Anaphylactic shock following a bite of a wild Kayan slow loris (*Nycticebus kayan*) in rural Sarawak, Malaysian Borneo. *Rural Remote Health*. 2019 Aug 18;19(3):5163. doi: 10.22605/RRH5163.
- Heinen TE, Farias CB, Abujamra AL, Mendonça RZ, Roesler R, Veiga AB. Effects of *Lonomia obliqua* caterpillar venom upon the proliferation and viability of cell lines. *Cytotechnology*. 2014 Jan;66:63-74. doi: 10.1007/s10616-013-9537-7.
- Heinen TE, Veiga AB. Arthropod venoms and cancer. *Toxicon*. 2011 Mar 15;57(4):497-511. doi: 10.1016/j.toxicon.2011.01.002.
- Albuquerque PL, Paiva JH, Martins AM, Meneses GC, Silva GB, Buckley N, et al. Clinical assessment and pathophysiology of *Bothrops* venom-related acute kidney injury: a scoping review. *J Venom Anim Toxins Incl Trop Dis*. 2020 Jul 10;26:e20190076. doi: 10.1590/1678-9199-JVATITD-2019-0076.
- Horta CC, Chatzaki M, Rezende BA, Magalhaes BF, Duarte CG, Felicori LF, et al. Cardiovascular-active venom toxins: an overview. *Curr Med Chem*. 2016 Feb 1;23(6):603-22. doi: 10.2174/0929867323666160126142837.
- Lebbe EK, Tytgat J. In the picture: disulfide-poor conopeptides, a class of pharmacologically interesting compounds. *J Venom Anim Toxins Incl Trop Dis*. 2016 Nov 7;22:30. doi: 10.1186/s40409-016-0083-6.
- Bordon KC, Calogna CT, Fornari-Baldo EC, Pinheiro EL Jr, Cerni FA, Amorim FG, et al. From animal poisons and venoms to medicines: achievements, challenges and perspectives in drug discovery. *Front Pharmacol*. 2020 Jul 24;11:1132. doi: 10.3389/fphar.2020.01132.
- Carregari VC, Rosa-Fernandes L, Baldasso P, Bydlowski SP, Marangoni S, Larsen MR, et al. Snake Venom Extracellular vesicles (SVEVs) reveal wide molecular and functional proteome diversity. *Sci Rep*. 2018 Aug 13;8(1):12067. doi: 10.1038/s41598-018-30578-4.
- Calvete JJ, Juárez P, Sanz L. Snake venomomics: strategy and applications. *J Mass Spectrom*. 2007 Nov;42(11):1405-14. doi: 10.1002/jms.1242.
- Moura AA, Kayano AM, Oliveira GA, Setubal SS, Ribeiro JG, Barros NB, et al. Purification and biochemical characterization of three myotoxins from *Bothrops mattogrossensis* snake venom with toxicity against *Leishmania* and tumor cells. *Biomed Res Int*. 2014 Mar 3;2014:195356. doi: 10.1155/2014/195356.
- Saviola AJ, Burns PD, Mukherjee AK, Mackessy SP. The disintegrin tzabcanin inhibits adhesion and migration in melanoma and lung cancer cells. *Int J Biol Macromol*. 2016 Jul;88:457-64. doi: 10.1016/j.ijbiomac.2016.04.008.
- Warrell DA. Snake bite. *Lancet*. 2010 Jan 2;375(9708):77-88. doi: 10.1016/S0140-6736(09)61754-2. Erratum in: *Lancet*. 2010 Feb 20;375(9715):640.
- Gonçalves-Machado L, Pla D, Sanz L, Jorge RJ, Leitão-De-Araújo M, Alves ML, et al. Combined venomomics, venom gland transcriptomics, bioactivities, and antivenomics of two *Bothrops jararaca* populations from geographic isolated regions within the Brazilian Atlantic rainforest. *J Proteomics*. 2016 Mar 1;135:73-89. doi: 10.1016/j.jprot.2015.04.029.
- Fox JW, Serrano SM. Exploring snake venom proteomes: multifaceted analyses for complex toxin mixtures. *Proteomics*. 2008 Feb 22;8(4):909-20. doi: 10.1002/pmic.200700777.
- Zelanis A, Andrade-Silva D, Rocha MM, Furtado MF, Serrano SM, Junqueira-de-Azevedo IL, et al. A transcriptomic view of the proteome variability of newborn and adult *Bothrops jararaca* snake venoms. *PLoS Negl Trop Dis*. 2012 Mar 13;6(3):e1554. doi: 10.1371/journal.pntd.0001554.
- Margres MJ, Aronow K, Loyacano J, Rokyta DR. The venom-gland transcriptome of the eastern coral snake (*Micrurus fulvius*) reveals high venom complexity in the intragenomic evolution of venoms. *BMC Genomics*. 2013 Aug 2;14:531. doi: 10.1186/1471-2164-14-531.
- Rokyta DR, Lemmon AR, Margres MJ, Aronow K. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics*. 2012 Jul 16;13:312. doi: 10.1186/1471-2164-13-312.



19. Braley LM, Menachery A, Williams GH. Angiotensin II's role in mediating angiotensin I- and tetradecapeptide-induced steroidogenesis by rat glomerulosa cells. *Endocrinology*. 1981 Sep 1;109(3):960-5. doi: 10.1210/endo-109-3-960.
20. Hrebickova L, Nawarskas JJ, Anderson JR. Ximelagatran: a new oral anticoagulant. *Heart Dis*. 2003 Oct 31;5(6):397-408. doi: 10.1097/01.hdx.0000099777.39577.e8.
21. Paine MJ, Desmond HP, Theakston RD, Crampton JM. Gene expression in *Echis carinatus* (carpet viper) venom glands following milking. *Toxicon*. 1992 Apr;30(4):379-86. doi: 10.1016/0041-0101(92)90534-c.
22. Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front Genet*. 2013 Dec 17;4:288. doi: 10.3389/fgene.2013.00288.
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.
24. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013 Jul 11;8:1494-512. doi: 10.1038/nprot.2013.084.
25. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006 Jul 1;22(13):1658-9. doi: 10.1093/bioinformatics/btl158.
26. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011 Aug 4;12:323. doi: 10.1186/1471-2105-12-323.
27. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351.
28. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol Biol*. 2019;1962:227-45. doi: 10.1007/978-1-4939-9173-0\_14.
29. Conesa A, Gotz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast<sub>GO</sub>: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005 Aug 4;21(18):3674-6. doi: 10.1093/bioinformatics/bti610.
30. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2007 Jul 1;35(Suppl 2):W182-5. doi: 10.1093/nar/gkm321.
31. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. 2012 Jan;40(D1):D109-14. doi: 10.1093/nar/gkr988.
32. Jungo F, Bougueleret L, Xenarios I, Poux S. The UniProtKB/Swiss-Prot Tox-Prot program: a central hub of integrated venom protein data. *Toxicon*. 2012 Sep 15;60(4):551-7. doi: 10.1016/j.toxicon.2012.03.010.
33. Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, et al. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res*. 2019 Jul 2;47(W1):W260-5. doi: 10.1093/nar/gkz303.
34. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013 Apr;30(4):772-80. doi: 10.1093/molbev/mst010.
35. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol*. 2010 Jul 13;10:210. doi: 10.1186/1471-2148-10-210.
36. Lefort V, Longueville JE, Gascuel O. SMS: Smart Model Selection in PhyML. *Mol Biol Evol*. 2017 Sep 1;34(9):2422-4. doi: 10.1093/molbev/msx149.
37. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. New York: Springer International Publishing; 2016.
38. R Core Team [Internet]. R: a language and environment for statistical computing [cited 2020 Aug 31]. Available from: <https://www.R-project.org/>.
39. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014 May 1;30(9):1236-40. doi: 10.1093/bioinformatics/btu031.
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000 May;25:25-9. doi: 10.1038/75556.
41. Kanehisa M. Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol Biol*. 2017;1611:135-45. doi:10.1007/978-1-4939-7015-5\_11.
42. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res*. 2000 Jan 1;28(1):231-4. doi: 10.1093/nar/28.1.231.
43. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014 Jan 1;42(D1):D222-30. doi: 10.1093/nar/gkt1223.
44. Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res*. 2007 Jan;35(Suppl 1):D308-13. doi: 10.1093/nar/gkl910.
45. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015 Jan 28;43(D1):D204-12. doi: 10.1093/nar/gku989.
46. Markland FS. Snake venoms and the hemostatic system. *Toxicon*. 1998 Dec;36(12):1749-800. doi: 10.1016/S0041-0101(98)00126-3.
47. Boldrini-França J, Cologna CT, Pucca MB, Bordon KC, Amorim FG, Anjolette FA, et al. Minor snake venom proteins: structure, function and potential applications. *Biochim Biophys Acta Gen Subj*. 2017 Apr;1861(4):824-38. doi: 10.1016/j.bbagen.2016.12.022.
48. Cidade DA, Simão TA, Dávila AM, Wagner G, Junqueira-de-Azevedo IL, Ho PL, et al. *Bothrops jararaca* venom gland transcriptome: analysis of the gene expression pattern. *Toxicon*. 2006 Sep 15;48(4):437-61. doi: 10.1016/j.toxicon.2006.07.008.
49. Junqueira-de-Azevedo IL, Bastos CM, Ho PL, Luna MS, Yamanouye N, Casewell NR. Venom-related transcripts from *Bothrops jararaca* tissues provide novel molecular insights into the production and evolution of snake venom. *Mol Biol Evol*. 2015 Mar;32(3):754-66. doi: 10.1093/molbev/msu337.
50. Kashima S, Roberto PG, Soares AM, Astolfi-Filho S, Pereira JO, Giulati S, et al. Analysis of *Bothrops jararacussu* venomous gland transcriptome focusing on structural and functional aspects: I--gene expression profile of highly expressed phospholipases A2. *Biochimie*. 2004 Mar;86(3):211-9. doi: 10.1016/j.biochi.2004.02.002.
51. Valente RH, Luna MS, Oliveira UC, Nishiyama MY Jr, Junqueira-de-Azevedo IL, Portes JÁ Jr, et al. *Bothrops jararaca* accessory venom gland is an ancillary source of toxins to the snake. *J Proteomics*. 2018 Apr 15;177:137-47. doi: 10.1016/j.jpro.2017.12.009.
52. Rodrigues RS, Boldrini-França J, Fonseca FP, de la Torre P, Henrique-Silva F, Sanz L, et al. Combined snake venomomics and venom gland transcriptomic analysis of *Bothropoides pauloensis*. *J Proteomics*. 2012 May 17;75(9):2707-20. doi: 10.1016/j.jpro.2012.03.028.
53. Amorim FG, Morandi-Filho R, Fujimura PT, Ueira-Vieira C, Sampaio SV. New findings from the first transcriptome of the *Bothrops moojeni* snake venom gland. *Toxicon*. 2017 Dec 15;140:105-17. doi: 10.1016/j.toxicon.2017.10.025.
54. Duan J, Sanggaard KW, Schauser L, Lauridsen SE, Enghild JJ, Schierup MH, et al. Transcriptome analysis of the response of Burmese python to digestion. *Gigascience*. 2017 Aug 1;6(8):1-18. doi: 10.1093/gigascience/gix057.
55. Carruthers M, Yurchenko AA, Augley JJ, Adams CE, Herzyk P, Elmer KR. De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genomics*. 2018 Jan 8;19(1):32. doi: 10.1186/s12864-017-4379-x.
56. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011 May 15;29(7):644-52. doi: 10.1038/nbt.1883.
57. Oliveira AL, Wollesen T, Kristof A, Scherholz M, Redl E, Todt C, et al. Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. *BMC Genomics*. 2016 Nov 10;17(1):905. doi: 10.1186/s12864-016-3080-9.
58. Theissinger K, Falckenhayn C, Blande D, Toljamo A, Gutekunst J, Makkonen J, et al. De Novo assembly and annotation of the freshwater crayfish *Astacus astacus* transcriptome. *Mar Genomics*. 2016 Aug;28:7-10. doi: 10.1016/j.margen.2016.02.006.

59. Haynsen MS, Vatanparast M, Mahadwar G, Zhu D, Moger-Reischer RZ, Doyle JJ, et al. De novo transcriptome assembly of *Pueraria montana* var. *lobata* and *Neustanthus phaseoloides* for the development of eSSR and SNP markers: narrowing the US origin(s) of the invasive kudzu. *BMC Genomics*. 2018 Jun 5;19:439. doi: 10.1186/s12864-018-4798-3.
60. Chen D, Kini RM, Yuen R, Khoo HE. Haemolytic activity of stonustoxin from stonefish (*Synanceja horrida*) venom: pore formation and the role of cationic amino acid residues. *Biochem J*. 1997 Aug 1;325(3):685-91. doi: 10.1042/bj3250685.
61. Low KS, Gwee MC, Yuen R, Gopalakrishnakone P, Khoo HE. Stonustoxin: a highly potent endothelium-dependent vasorelaxant in the rat. *Toxicon*. 1993 Nov;31(11):1471-8. doi: 10.1016/0041-0101(93)90212-2.
62. Ueda A, Suzuki M, Honma T, Nagai H, Nagashima Y, Shiomi K. Purification, properties and cDNA cloning of neoverrucotoxin (neoVTX), a hemolytic lethal factor from the stonefish *Synanceia verrucosa* venom. *Biochim Biophys Acta*. 2006 Nov;1760(11):1713-22. doi: 10.1016/j.bbagen.2006.08.017.
63. Takahashi M, Ishida Y, Iwaki D, Kanno K, Suzuki T, Endo Y, et al. Essential role of mannose-binding lectin-associated serine protease-1 in activation of the complement factor D. *J Exp Med*. 2010 Jan 18;207(1):29-37. doi: 10.1084/jem.20090633.
64. Haihua C, Wei W, Kun H, Yuanli L, Fei L. Cobra Venom Factor-induced complement depletion protects against lung ischemia reperfusion injury through alleviating blood-air barrier damage. *Sci Rep*. 2018 Jul 9;8:10346. doi: 10.1038/s41598-018-28724-z.
65. Cardoso KC, Silva MJ, Costa GG, Torres TT, Del Bem LE, Vidal RO, et al. A transcriptomic analysis of gene expression in the venom gland of the snake *Bothrops alternatus* (urutu). *BMC Genomics*. 2010 Oct 26;11:605. doi: 10.1186/1471-2164-11-605.
66. Hargreaves AD, Swain MT, Logan DW, Mulley JF. Testing the Toxicofera: comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon*. 2014 Dec 15;92:140-56. doi: 10.1016/j.toxicon.2014.10.004.
67. Garnier P, Sauviat MP, Goudey-Perriere F, Perriere C. Cardiotoxicity of verrucotoxin, a protein isolated from the venom of *Synanceia verrucosa*. *Toxicon*. 1997 Jan;35(1):47-55. doi: 10.1016/s0041-0101(96)00075-x.
68. Low KS, Gwee MC, Yuen R, Gopalakrishnakone P, Khoo HE. Stonustoxin: effects on neuromuscular function in vitro and in vivo. *Toxicon*. 1994 May;32(5):573-81. doi: 10.1016/0041-0101(94)90205-4.
69. von Reumont BM, Campbell LI, Richter S, Hering L, Sykes D, Hetmank J, et al. A Polychaete's powerful punch: venom gland transcriptomics of *Glycera* reveals a complex cocktail of toxin homologs. *Genome Biol Evol*. 2014 Sep 5;6(9):2406-23. doi: 10.1093/gbe/evu190.
70. Ellisdon AM, Reboul CF, Panjkar S, Huynh K, Oellig CA, Winter KL, et al. Stonefish toxin defines an ancient branch of the perforin-like superfamily. *Proc Natl Acad Sci U S A*. 2015 Dec 15;112(50):15360-5. doi: 10.1073/pnas.1507622112.
71. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016 Jan 4;44(D1):D457-62. doi: 10.1093/nar/gkv1070.
72. Naveira H, Bello X, Abal-Fabeiro JL, Maside X. Evidence for the persistence of an active endogenous retrovirus (ERVE) in humans. *Genetica*. 2014 Oct;142:451-60. doi: 10.1007/s10709-014-9789-y.
73. Chuong EB. The placenta goes viral: retroviruses control gene expression in pregnancy. *PLoS Biol*. 2018 Oct 9;16(10):e3000028. doi: 10.1371/journal.pbio.3000028.
74. Cornelis G, Funk M, Vernochet C, Leal F, Tarazona OA, Meurice G, et al. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental *Mabuia* lizard. *Proc Natl Acad Sci U S A*. 2017 Dec 19;114(51):E10991-11000. doi: 10.1073/pnas.1714590114.
75. Dreyfuss G, Swanson MS, Piñol-Roma S. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem Sci*. 1988 Mar 1;13(3):86-91. doi: 10.1016/0968-0004(88)90046-1.
76. Klug A. Zinc finger peptides for the regulation of gene expression. *J Mol Biol*. 1999 Oct 22;293(2):215-8. doi: 10.1006/jmbi.1999.3007.
77. Smith TF, Gaitatzes C, Saxena K, Neer EJ. The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci*. 1999 May 1;24(5):181-5. doi: 10.1016/s0968-0004(99)01384-5.
78. Wang DS, Shaw R, Hattori M, Arai H, Inoue K, Shaw G. Binding of pleckstrin homology domains to WD40/beta-transducin repeat containing segments of the protein product of the Lis-1 gene. *Biochem Biophys Res Commun*. 1995 Apr 17;209(2):622-9. doi: 10.1006/bbrc.1995.1545.