Data in Brief 55 (2024) 110629

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of PLA2 family identified from transcriptomic high-throughput sequencing of *Androctonus crassicauda* (Scorpionida: Buthidae) venom gland

Fatemeh Salabi*, Hedieh Jafari

Razi Vaccine and Serum Research Institute, Agricultural Research, Education and Extension Organization (AREEO), Ahvaz, Iran

ARTICLE INFO

Article history:
Received 29 November 2023
Revised 17 May 2024
Accepted 5 June 2024
Available online 12 June 2024

Dataset link: Raw data of female venom glands of Androctonus crassicauda (Original data)
Dataset link: platelet-activating factor acetylhydrolase [Androctonus crassicauda] (Original data)
Dataset link: Raw data of male venom glands of Androctonus crassicauda (Original data)
Dataset link: Raw data of male venom glands of Androctonus crassicauda (Original data)
Dataset link: phospholipase A2 [Androctonus crassicauda] (Original data)

Keywords:
RNA sequencing
Androctonus crassicauda
Venom gland
Transcriptome
PLA2

ABSTRACT

Recently, RNA sequencing has been widely applied to deeply understand the molecular diversity of the venom compounds of various venomous animal species, including scorpions. Among the venomous scorpion species of the Buthidae family, there are many documents of stinging and severe envenoming of victims by the scorpion of *Androctonus crassicauda*. We present here a high-throughput RNA sequencing dataset of the venom glands from five *A. crassicauda* individuals, including male and female scorpions. Furthermore, the assembled data corresponding to annotated PLA2 transcripts are also presented. The dataset in this report is related to our research article entitled: “Whole transcriptome sequencing reveals the activity of the PLA2 family members in *Androctonus crassicauda* (Scorpionida: Buthidae) venom gland” [1]. Here, the venom gland transcriptome analysis of the *A. crassicauda* was performed. The analysis of concatenated clustered transcriptome assembly using TrinityStats.pl showed that *de novo* assembly of 517,799,704 clean read pairs generated 744,804 trinity transcripts representing 563,526 trinity genes. BUSCO score for the concatenated clustered transcriptome assembly against orthologs from Arachnida showed 96.7 % complete, 1.6 % fragmented, 1.7 % missing genes, and 2934 genes. Subsequently, the sequences represented PLA2 annotation

* Corresponding author.
E-mail address: f.salabi@rvsri.ac.ir (F. Salabi).

https://doi.org/10.1016/j.dib.2024.110629

were extracted from the transcriptome dataset using BLAST searches against the local PLA2 database. We found several cDNA sequences representing PLA2 annotations, which based on sequence similarity to previously found PLA2s, we named platelet-activating factor acetylhydrolases, calcium-dependent PLA2s, calcium-independent PLA2s, and secreted PLA2s. The PLA2 data significantly enrich KEGG pathways related to lipid metabolism. This manuscript complements the primary research article by providing additional data on the abundant estimation of PLA2s.

Specifications Table

Subject	Toxicology
Specific subject area	Biotechnology, Transcriptome analysis of the scorpion venom gland
Data format	Raw sequencing data, assembled, and analyzed data of phospholipase A2 (PLA2)
Type of data	Tables, figures, and workflow Chart
Data collection	*Androctonus crassicauda* specimens were collected in summer 2017. The telsons of 5 specimens, three males and two females, were removed three days after venom extraction with the electrical stimulation method, powdered in liquid nitrogen, and the total RNAs of venom glands were extracted with RNeasy Animal Mini Kit (Qiagen, Valencia, CA, USA). The purified total RNAs were used for sequencing the constructed library by Illumina RNA-sequencing (Illumina HiSeq 2000 platform, at Macrogen Co Macrogen, Seoul, Korea) with 150-bp paired-end reads. Raw Illumina read datasets were processed by comprehensive bioinformatics analysis. All reads were trimmed using Trimmomatic v2.10.0 to obtain high-quality sequences, and *de novo* assembly of trimmed reads was performed using Trinity software v2.15.1. The resulting transcripts were then clustered using CD-HIT-EST v4.7. TransDecoder v5.5.0 was then used to predict the Open Reading Frames (ORFs) and the Trinotate v4.0.2 to annotate the venom gland assembled transcriptome. To identify the PLA2s from *A. crassicauda* venom gland, previously identified PLA2s sequenced from scorpions and closely related species were extracted from nucleotide sequence databases of the National Center of Biological Information (NCBI) and they used to build a local PLA2 database for BLAST searches. With Blastx and Blastp, the venom gland assembled transcriptome of *A. crassicauda* was searched against the local PLA2 database to identify and extract the homologous proteins and nucleotide sequences. Expression level calculation of PLA2s was performed using RSEM software. For alternative functional annotation of genes, all de novo transcripts and those representing PLA2 annotation were subjected to the GhostKOALA server.
Data source location	*Androctonus crassicauda* specimens were captured from the Baghmalek of Khuzestan province, southwest of Iran. The identification of specimens based on morphological characteristics was performed at the arthropod laboratory of the Southwest branch of Razi Vaccine and Serum Research Institute, Iran, Ahvaz.
Data accessibility	The bio project, bio samples, and RNA-seq reads are available in National Center for Biotechnology Information database under the accessions: Repository name: Raw data of male's venom glands of *A. crassicauda* Data identification number: PRJNA1040487. Direct link: https://www.ncbi.nlm.nih.gov/sra/PRJNA1040487 BioSample accessions: SAMN38279379, SAMN38279380, SAMN38279381 Direct link to BioSamples: https://www.ncbi.nlm.nih.gov/biosample/38279379 https://www.ncbi.nlm.nih.gov/biosample/38279380 https://www.ncbi.nlm.nih.gov/biosample/38279381

(*continued on next page*)

	Repository name: Raw data of female's venom glands of *A. crassicauda* Data identification number: PRJNA1040746. Direct link: https://www.ncbi.nlm.nih.gov/sra/PRJNA1040746 BioSample accessions: SAMN38273960, SAMN38273961, SAMN38273962, SAMN38273963 Direct link to BioSample s: https://www.ncbi.nlm.nih.gov/biosample/38273960 https://www.ncbi.nlm.nih.gov/biosample/38273961 https://www.ncbi.nlm.nih.gov/biosample/38273962 https://www.ncbi.nlm.nih.gov/biosample/38273963 Repository name for 12 PLA2s transcripts: NCBI Genbank Data identification number: accession numbers: OQ982083, OQ982084, OQ982085, OQ982086, OQ982087, OQ982088, OQ982089, OQ982090, OQ982091, OQ982092, OQ982093, and OQ982094
Related research article	F. Salabi, H, Jafari. Whole transcriptome sequencing reveals the activity of the PLA2 family members in *Androctonus crassicauda* (Scorpionida: Buthidae) venom gland. Faseb journal. 2024 [1].

1. Value of the Data

- The raw reads of *A. crassicauda* venom glands can be helpful to facilitate future studies of genus Androctonus RNA sequencing. They can be used by insect researchers to perform comparative transcriptomic studies of species associated with genus Androctonus or other scorpion species.
- The PLA2 new sequences will serve as references to target scorpions PLA2 sequences. These data provide genomic data for further studying, classification, and exploration of their diversity.
- In this contribution, there is valuable information about the gene expression levels of the PLA2 protein family in terms of FPKM (fragments per kilobase of transcript sequence per million base pairs sequenced), which was calculated using RSEM software.
- Here, we compared the length of the discovered sequences from the venom of *A. crassicauda*. The PLA2 enzymatic activity in the venom of *A. crassicauda* was assessed in our last contribution [1].

2. Background

We present here the RNA sequencing raw reads of male and female *A. crassicauda* venom glands and assembled data corresponding to annotated PLA2 transcripts. The RNA sequencing raw reads represent the first report of *de novo* transcriptome analysis of *A. crassicauda* venom gland, and the annotated PLA2 transcripts provide a first report of the identification, characterization, and classification of different phospholipases A2 (PLA2) and their isomers from the venom glands of *A. crassicauda*. Overall, this study provides a basis for identification of scorpion venom components, development, classification of protein families, and comparison of gene expression of different isoforms of a protein family in future studies.

3. Data Description

Statistical information of data concerning transcripts and unigenes within the transcriptome profile have been shown in Tables 1–3. Table 1 shows the total numbers of raw and cleaned paired reads for each sample (5 libraries), and percentage of mapped reads on the assembled transcripts. The transcriptome sequencing of *A. crassicauda* venom gland generated 518,221,366 paired raw reads (Table 1), out of which 517,799,70vv4 paired clean reads (Table 1) were retained after quality filtering. We first assembled the *A. crassicauda* venom gland transcriptome from raw RNA seq data of five libraries into individual assemblies and concatenated assembly

Table 1
Number of raw and cleaned paired reads per males and females, and percentage of mapped reads on the assembled transcripts.

Samples name	Raw read pairs	Cleaned read pairs	% reads mapped back to transcriptome
Females			
F1	100,214,810	100,209,614	96.29 %
F2	111,463,564	111,458,608	97.10 %
Males			
M1	104,488,530	104,484,814	97.50 %
M2	105,998,054	105,593,736	97.53 %
M3	96,056,408	96,052,932	94.88 %
Total	518,221,366	517,799,704	–

Table 2
Summary statistics of *de novo* clustered transcriptome assembly for *A. crassicauda* venom gland using the combined data of 5 samples.

Total trinity genes	563,526
Total trinity transcripts	744,804
Total assembled bases	350,938,383
GC%	34.10

Table 3
Assessing the quality of the clustered *de novo* transcriptome assembly.

Stats based on all transcript contigs	
Contig N10	2517
Contig N20	1397
Contig N30	905
Contig N40	654
Contig N50	507
Median contig length	322
Average contig	471.18

using Trinity software. Subsequently, the resulting transcripts were clustered using CD-HIT-EST v4.7 [2] with 95 % identity. To evaluate the quality of the final transcriptomes, the filtered unique reads were first mapped back to the final assemblies using Bowtie2 v2.3.4.1. Then the quality of the assemblies was estimated with TrinityStats.pl and the final transcriptome completeness was measured using the BUSCO v5.2.2. The numbers of reads mapped back to transcriptomes resulted in $>$ 95 % mapped reads. The statically analysis of concatenated clustered transcriptome assembly using TrinityStats.pl showed that the *de novo* assembly of all reads resulted in a total assembly of 350,938,383 bp representing 744,804 transcripts with N50 size of 507 bp and corresponding to 563,526 trinity genes (Table 2). The results related to number of contigs, unigenes, and annotated unigenes in databases have shown in Table 2. The evaluation of the quality of the clustered assembly is shown in Table 3. The assembly's completeness of concatenated transcriptome generated from Trinity only (Raw transcriptome) or resulting from the clustering step (Finally transcriptome) of *A. crassicauda* venom gland was measured using BUSCOs [3] against the Arthropoda and Arachnida genes database, which showed high BUSCO completeness scores of $>$96 %. The detailed information is shown in Table 4. The analyses of individual and concatenated transcriptomes of *A. crassicauda* venom gland resulting from the clustering step are also compared in Table 5. These analyses provide interesting information including the number of single, duplicated and fragmented entries form individual and concatenated raw or clustered transcriptomes. It achieves an Arachnida genes number of 2934, but only 1013 Athropoda genes number. In addition, BUSCO analyses also showed that the completeness and the number of duplicated copies were higher in the concatenated assembly than in individual assembly for *A. crassicauda* venom gland datasets. This finding was contrary to the results of single copies. The result showed that the number of duplicated copies of transcriptomes was decreased sig-

Table 4
Quality scores for the raw assembly and the final assembly of *A. crassicauda* venom gland.

Samples name	Cross Arthropoda_odb10		Cross Arachnida_odb10	
	Raw Assembly	Final Assembly	Raw Assembly	Final Assembly
Complete BUSCOs	96.9 %	96.6 %	96.8 %	96.7 %
Single-copy BUSCOs	40.1 %	57.3 %	34.3 %	51.3 %
Duplicated BUSCOs	56.8 %	39.3 %	62.5 %	45.4 %
Fragmented BUSCOs	2.0 %	2.2 %	1.6 %	1.6 %
Missing BUSCOs	1.1 %	1.2 %	1.6 %	1.7 %
Number of genes	1013	1013	2934	2934
Total BUSCOs groups searched	1013	1013	2934	2934

Raw Assembly: The *A. crassicauda* assembly generated from Trinity only.
Final Assembly: The *A. crassicauda* assembly, resulting from the clustering step.

Table 5
BUSCO results from the *de novo* transcriptomes of *A. crassicauda* venom gland with the Arthropoda_odb10 dataset.

Samples name	Individual transcriptome		Concatenated transcriptome	
	Raw Assembly	Final Assembly	Raw Assembly	Final Assembly
Complete BUSCOs	94.3 %	94.0 %	96.9 %	96.6 %
Single-copy BUSCOs	57.3 %	84.9 %	40.1 %	57.3 %
Duplicated BUSCOs	37.0 %	9.1 %	56.8 %	39.3 %
Fragmented BUSCOs	3.6 %	3.8 %	2.0 %	2.2 %
Missing BUSCOs	2.1 %	2.2 %	1.1 %	1.2 %
Number of genes	1013	1013	1013	1013
Total BUSCOs groups searched	1013	1013	1013	1013

nificantly after the clustering step. The *de novo* transcriptomes assemblies may cause the high duplicate matches to the BUSCO sequences, which does not necessarily indicate the unwanted redundancy in the assembly; it may due to assembled from several individuals or the presence of closely related transcripts that represent splicing isoforms. We used Croco v0.1 software [4] to confirm that this redundancy was not due to pervasive cross-species contamination in next-generation sequencing data. The results showed no contamination. The raw data corresponding to this transcriptome were deposited into the NCBI-SRA database under the following BioProjects accession numbers: PRJNA1040487 and PRJNA1040746. Using BLAST searches of the *A. crassicauda* transcriptome against the local PLA2 database, we identified several sequences representing the PLA2 annotation. Then, we classified and named those as platelet-activating factor acetylhydrolases, calcium-dependent PLA2s, calcium-independent PLA2s, and secreted PLA2s. These sequences have been uploaded to the NCBI database with accession numbers OQ982083, OQ982084, OQ982085, OQ982086, OQ982087, OQ982088, OQ982089, OQ982090, OQ982091, OQ982092, OQ982093 and OQ982094.

The characterization of the individual gene functions and reconstruction of KEGG pathways of total assembled sequences and discovered PLA2 sequences was done using the GhostKOALA automatic annotation server (Fig. 1). We found that among the total genes identified in this study, most genes were identified in the metabolic pathway and the least in the pathway related to immune diseases. In contrast, among the PLA2 genes, almost all genes were identified in the lipid metabolism. Based on taxonomic information collected from GhostKOALA server, the transcripts showing similarity to arthropods references.

Furthermore, we characterized the range size for mRNA transcript sequences (Full-length sequences) and CDS sequences of the PLA2s in *A. crassicauda* venom gland transcriptome. The results are shown in Table 6 and Fig. 2. Based on the full-length sequences, the sPLA2G10 with 3033 bp and sPLA2 isoformx2 with 826 bp were the longest and shortest, respectively. Based on CDS sequences or mature protein size ranges, the iPLA2 with open reading frame of 1849 bp and cPLA2 with open reading frame of 468 bp were the longest and shortest sequences, respectively.

The next five figures we provide show the abundant estimation of mRNAs transcripts of PLA2s (Figs. 3-7). Abundances were reported in expected Fragments Per Kilobase of transcript

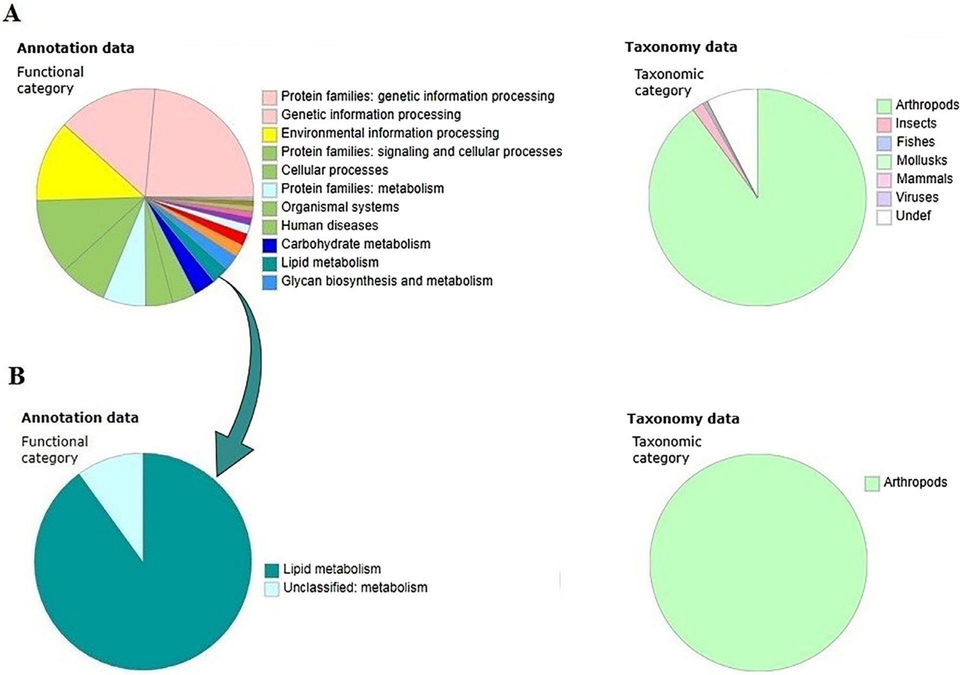

Fig. 1. Output summary of the functional distribution of protein-coding genes and taxonomic information in the GhostKOALA server. **A.** the functional distribution of total genes. **B.** the functional distribution of PLA2 genes.

Table 6
Characterization of the PLA2 family (Full length and mature proteins) [1].

PLA2	Full length (bp)	CDS length (bp)	Naa (mature protein)	Exon No	MW (mature protein)
PAFA	1431	676	224	5	25,233.63
iPLA2	2330	1849	615	8	70,342.69
cPLA2	940	468	155	4	17,747.31
sPLA2-GXIIA	1555	589	169	4	18,621.44
sPLA2G3	1195	969	299	2	35,143.08
sPLA2G10	3033	615	182	5	20,750.79
sPLA2-isox1	870	735	224	4	26,240.03
sPLA2-isox2	826	642	196	4	22,983.03
sPLA2-isox3	1613	693	210	4	24,346.41
sPLA2-isox4	1063	840	263	4	30,497.34

PAFA: platelet-activating factor acetylhydrolase.
iPLA2: calcium-independent phospholipase A2.
cPLA2: cytosolic phospholipase A2.
sPLA2GXIIA: group XIIA secretory phospholipase A2-like.
sPLA2G3: group 3 secretory phospholipase A2.
sPLA2G10: group 10 secretory phospholipase A2.
sPLA2-isox1: secretory phospholipase A2 isoform X1.
sPLA2-isox2: secretory phospholipase A2 isoform X2.
sPLA2-isox3: secretory phospholipase A2 isoform X3.
sPLA2-isox4: secretory phospholipase A2 isoform X4.
Naa: Number of amino acids of mature protein.
MW: Molecular weight.

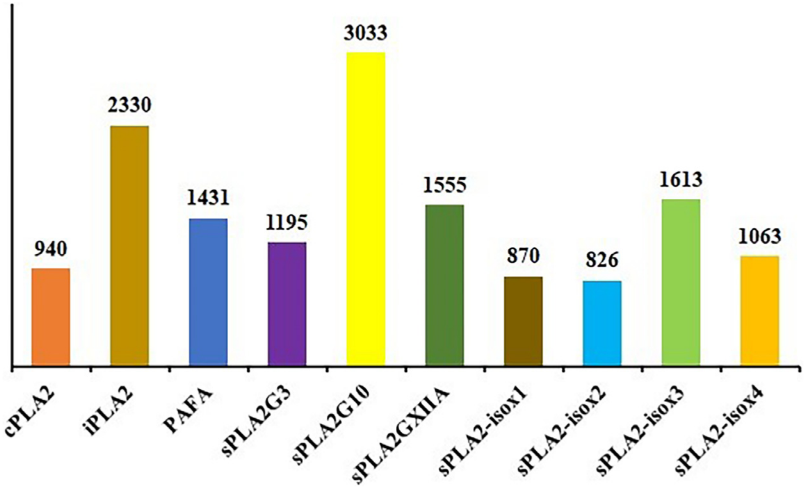

Fig. 2. Length distribution of mRNA transcripts of PLA2 (bp).

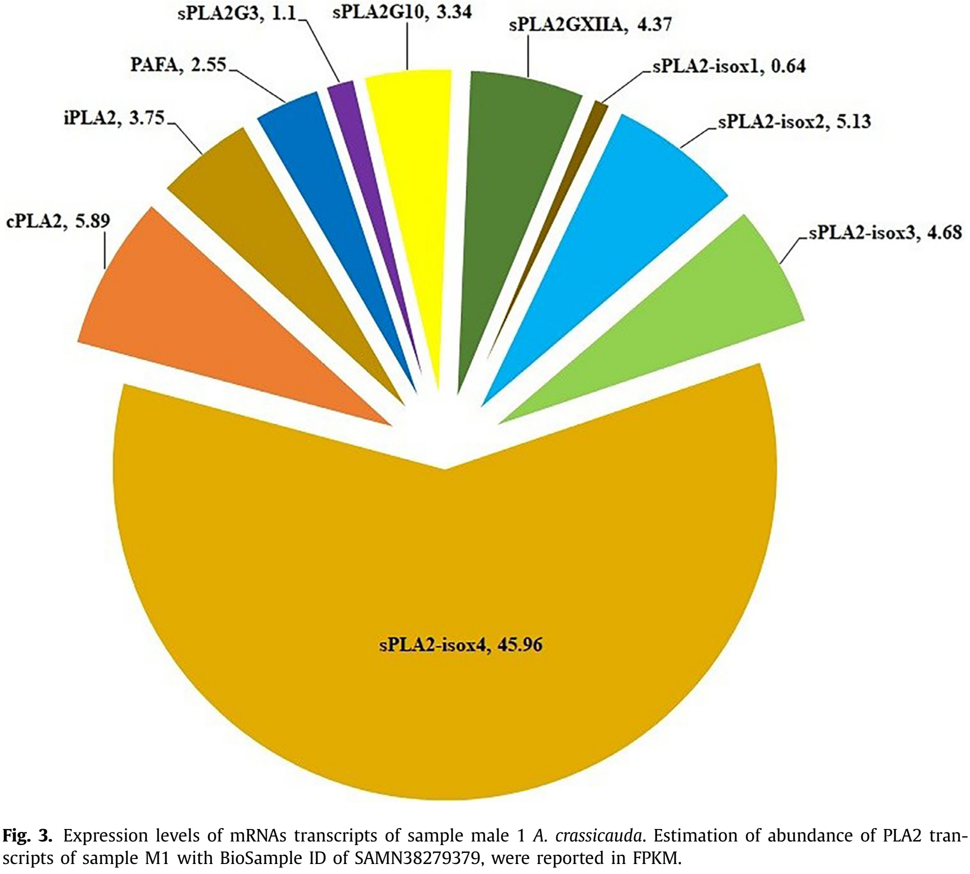

Fig. 3. Expression levels of mRNAs transcripts of sample male 1 *A. crassicauda*. Estimation of abundance of PLA2 transcripts of sample M1 with BioSample ID of SAMN38279379, were reported in FPKM.

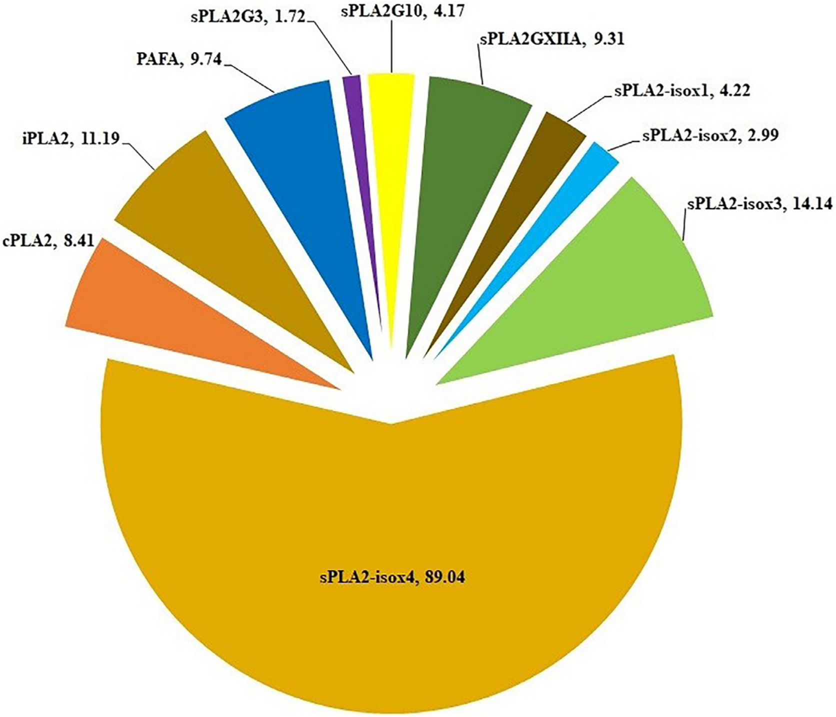

Fig. 4. Expression levels of mRNAs transcripts of sample male 2 *A. crassicauda*. Estimation of abundance of PLA2 transcripts of sample M2 with BioSample ID of SAMN38279380, were reported in FPKM.

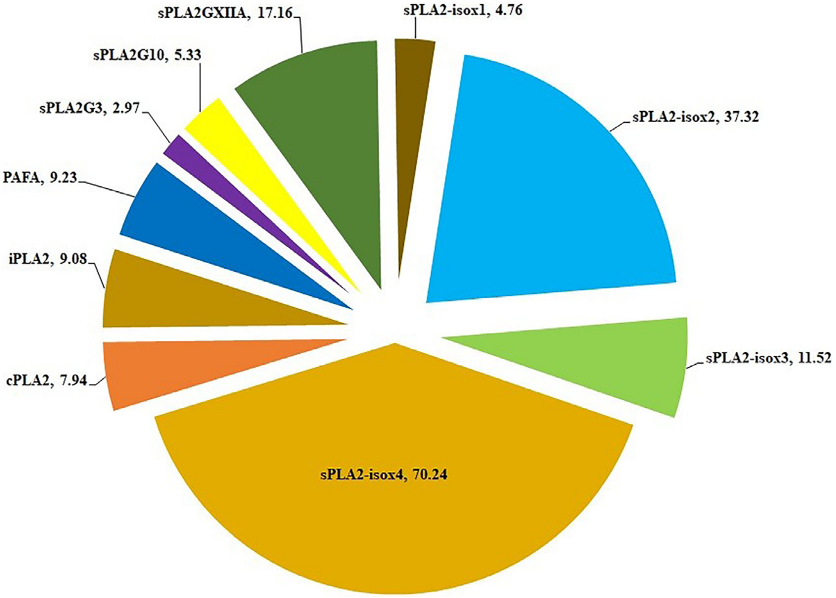

Fig. 5. Expression levels of mRNAs transcripts of sample male 3 *A. crassicauda*. Estimation of the abundance of PLA2 transcripts of sample M3 with BioSample ID of SAMN38279381 was reported in FPKM.

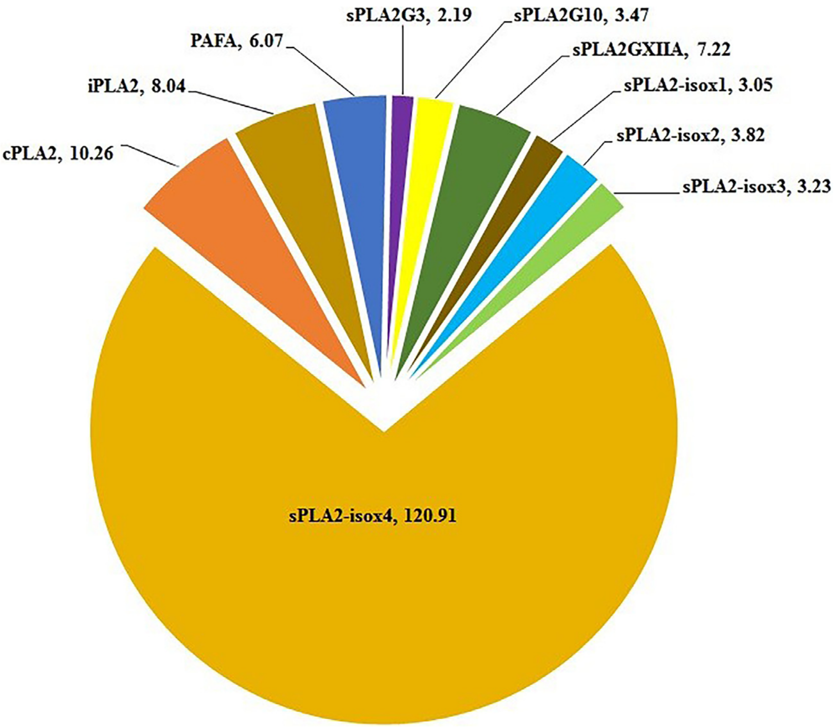

Fig. 6. Expression levels of mRNAs transcripts of sample female 1 *A. crassicauda*. Estimation of abundance of PLA2 transcripts of sample F1 with BioSample IDs of SAMN38273960 and SAMN38273961, were reported in FPKM.

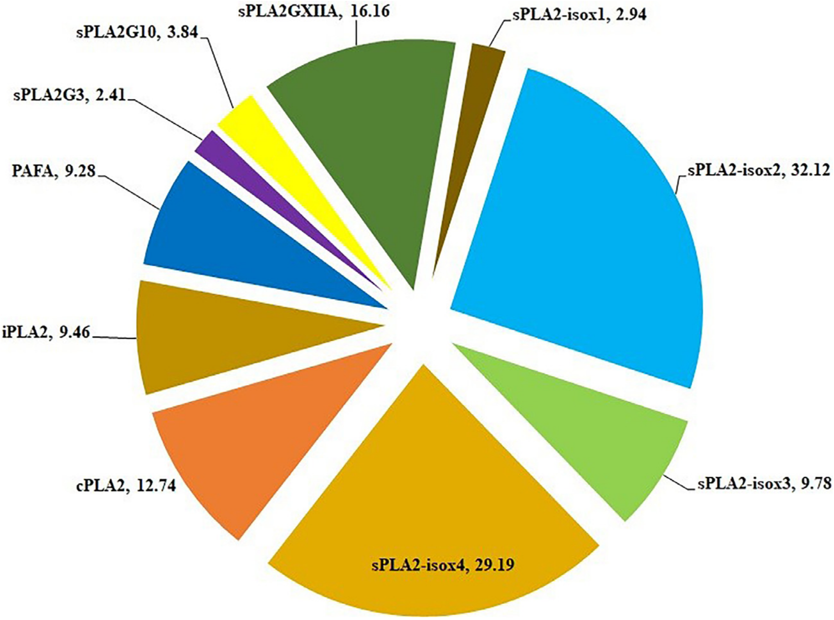

Fig. 7. Expression levels of mRNAs transcripts of sample female 2 *A. crassicauda*. Estimation of abundance of PLA2 transcripts of sample F2 with BioSample IDs of SAMN38273962 and SAMN38273963, were reported in FPKM.

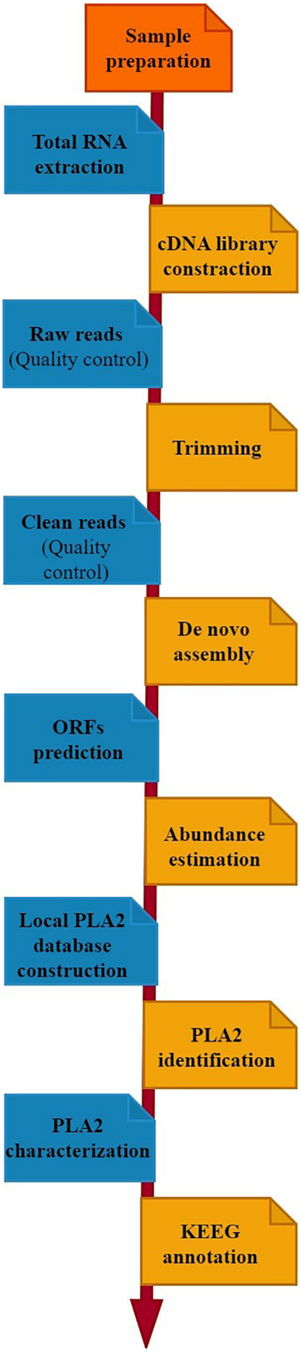

Fig. 8. Workflow of *de novo* assembly, analysis of *A. crassicauda* venom glands transcriptome, and PLA2 identification.

per Million fragments mapped (FPKM) using RSEM tool. By comparing the abundance of all PLA2 genes in almost all samples, we found that sPLA2G3 was the least abundant gene expressed and sPLA2-isox4 was the most abundant gene expressed.

4. Experimental Design, Materials and Methods

The main objective of this work was to explore the presence of PLA2 in the scorpion of *A. crassicauda* venom glands. We develop an experimental and computational pipeline for identifying the PLA2 isoforms from venom gland (Fig. 8). The steps of this pipeline are more explained in our previous works [5-9]. Briefly, the scorpions of *A. crassicauda* were captured from Khuzestan province, Iran. After venom milking, the total RNAs were extracted from the telsons of *A. crassicauda* individuals. The rRNA-depleted libraries were prepared for three independent biological replicates of males and two females. The cDNA libraries were sequenced by the Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA), with 150 bp paired-end reads. After trimming and assessing the quality of raw reads by means of the Trimmomatic v2.10.0 and FastQC v0.11.5 programs, respectively, the clean reads were assembled *de novo* using Trinity software, version 2.15.1. To reduce sequence redundancy, the CD-HIT-EST (v4.7 [2]) program was run for clustering the resulted transcripts. The filtered reads were mapped back to the transcriptome to evaluate individual mapping rates with Bowtie2. Then, in order to test the quality of the assembly, the completeness of the final transcriptome was measured by counting the percentage of orthologues conserved across Arachnida and Arthropoda databases using BUSCO (Benchmarking Universal Single-Copy Orthologs) package (v5.2.2) [3]. Finally, we used Croco v0.1 software [4] to assess the pervasive cross-species contamination in our assemblies. Then, the TransDecoder v5.5.0 was used to predict the Open Reading Frames (ORFs) and potential coding sequences. To quantify the gene expression levels in terms of FPKM in each individual, we used RSEM software. For identification of PLA2 sequences based on BLAST searches, we first created local database using PLA2s from scorpions and closely related species collected from NCBI database. The total assembled reads and extracted PLA2s were enriched by the GhostKOALA server for the alternative functional annotation of genes.

Limitations

Not applicable.

Ethics Statement

No animals were euthanized as part of this study, and all sample collection methods and experimental procedures described herein were rigorously reviewed and approved by the Institutional Animal Care Committee of Razi Vaccine and Serum Research Institute (Permit number IR.RVSRI.REC.1401.017) and AREEO protocols, which comply with Iran guidelines for work with animals.

CRediT Author Statement

Fatemeh Salabi: conceived, designed the experiments, and analyzed the data; **Fatemeh Salabi** and **Hedieh Jafari:** collect and identify the scorpion specimens and wrote the manuscript. All authors have read and approved the manuscript.

Data Availability

Raw data of female venom glands of Androctonus crassicauda (Original data) (MGDS)
platelet-activating factor acetylhydrolase [Androctonus crassicauda] (Original data) (SOL Genomics)
Raw data of male venom glands of Androctonus crassicauda (Original data) (Mendeley Data)
Raw data of male venom glands of Androctonus crassicauda (Original data) (Mendeley Data)
phospholipase A2 [Androctonus crassicauda] (Original data) (SOL Genomics)

Acknowledgments

Authors wish to thank Razi Vaccine and Serum Research Institute of Iran for its financial support.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

References

[1] F. Salabi, Jafari H, Whole transcriptome sequencing reveals the activity of the PLA2 family members in *Androctonus crassicauda* (Scorpionida: buthidae) venom gland, FASEB J. 38 (10) (2024) e23658.
[2] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[3] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (19) (2015) 3210–3212.
[4] P. Simion, et al., A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data, BMC Biol. 16 (2018) 28.
[5] F. Salabi, H. Jafari, Differential venom gland gene expression analysis of juvenile and adult scorpions *Androctonus crassicauda*, BMC Genom. 23 (1) (2022) 636.
[6] F. Salabi, H. Jafari, S. Navidpour, A.S. Sadr, Systematic and computational identification of *Androctonus crassicauda* long non-coding RNAs, Sci. Rep. 11 (1) (2021) 4720.
[7] F. Salabi, H. Jafari, New insights about scorpion venom hyaluronidase; isoforms, expression and phylogeny, Toxin Rev. (2022) 1–16.
[8] M. Baradaran, F. Salabi, Genome-wide identification, structural homology analysis, and evolutionary diversification of the phospholipase D gene family in the venom gland of three scorpion species, BMC Genom. (2023) 730.
[9] M. Nazari, H. roshanfekr, F. Salabi, J. Fayazi, A. Mohamadian, F. Kavosh, Production of the first effective immune equine serum antivenom against iranian honey bees (Apis mellifera meda), Res. Anim. Prod. (2024) 86–94.