SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» RNA sequencing

» Data processing

» Peptides

» Entomology

Comprehensive analysis of the venom gland transcriptome of the spider *Dolomedes fimbriatus*

Sergey A. Kozlov[1], Vassili N. Lazarev[2,3], Elena S. Kostryukova[2,3], Oksana V. Selezneva[2], Elena A. Ospanova[2], Dmitry G. Alexeev[2,3], Vadim M. Govorun[1,2,3] & Eugene V. Grishin[1]

Received: 08 April 2014

Accepted: 09 June 2014

Published: 5 August 2014

A comprehensive transcriptome analysis of an expressed sequence tag (EST) database of the spider *Dolomedes fimbriatus* venom glands using single-residue distribution analysis (SRDA) identified 7,169 unique sequences. Mature chains of 163 different toxin-like polypeptides were predicted on the basis of well-established methodology. The number of protein precursors of these polypeptides was appreciably numerous than the number of mature polypeptides. A total of 451 different polypeptide precursors, translated from 795 unique nucleotide sequences, were deduced. A homology search divided the 163 mature polypeptide sequences into 16 superfamilies and 19 singletons. The number of mature toxins in a superfamily ranged from 2 to 49, whereas the diversity of the original nucleotide sequences was greater (2–261 variants). We observed a predominance of inhibitor cysteine knot toxin-like polypeptides among the cysteine-containing structures in the analyzed transcriptome bank. Uncommon spatial folds were also found.

Design Type(s)	Nucleic Acid Sequencing
Measurement Type(s)	transcription profiling assay
Technology Type(s)	Expressed Sequence Tag
Factor Type(s)	
Sample Characteristic(s)	Dolomedes fimbriatus • venom gland

[1]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences ul. Miklukho-Maklaya, 16/10, Moscow 117997, Russia. [2]Scientific Research Institute of Physical-Chemical Medicine of the Federal Medical and Biological Agency of Russian Federation, 1a, Malaya Pirogovskaya st., Moscow 119435, Russia. [3]Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region 141700, Russia.
Correspondence and requests for materials should be addressed to S.A.K. (email: serg@ibch.ru).

Background & Summary

Natural toxins comprise a vast group of compounds that are diverse in their chemical nature, the most numerous of which are polypeptide molecules. As a rule, peptide toxins are present in animal venoms, and these venoms usually contain a diverse array of bioactive peptides affecting various receptor targets. A number of polypeptide toxins are currently used in fundamental research as biological tools for the study of various receptor systems and for the development of pharmaceutical agents used in medicine[1–5]. Among venomous animals, marine snails and spiders can be considered the leaders in venom component diversity[6–10], with venoms containing up to several hundred distinct components. Taking into account that the number of spider species (~ 44,500) surpasses that of all other known species of venomous animals, spider venoms represent the largest naturally edited library of biologically active molecules, which is predicted to include more than several million compounds[9,11]. It is likely that peptide components with unique specialization to any given ionotropic receptor can be found in spider venoms[12–15].

Peptidomic and proteomic approaches can be practically applied for the identification and characterization of venom components and as high-throughput screening techniques for bioactivity testing[10,16]. Such approaches make it possible to mine rare components in venoms that were not examined by earlier researchers and to determine their structure. In previous proteomic studies of Australian spider venoms, approximately 1,000 individual components for *Hadronyche versuta* and more than 600 for *Atrax robustus* were identified with a combination of chromatographic and off-line mass spectrometry procedures[17,18]. A total of 378 venom polypeptides were found in the venom of the spider *Dolomedes sulfurous* by means of off-line HPLC/MALDI-TOF-MS analysis[19]. The presence of 400 different compounds, many of which overlapped between species, was detected by means of mass spectrometry analysis for three related Brazilian species: *Phoneutria nigriventer*, *Phoneutria reidyi*, and *Phoneutria keyserlingi*[20]. Using a combination of a multidimensional chromatographic approach and tandem mass spectrometry, 286 components were identified in the venom of the spider *Cupiennius salei*[21]. However, the investigation of other spider venoms has not revealed such component variety. A proteomic analysis of the venom of the Chinese tarantula *Haplopelma huwenum* (*Selenocosmia huwena)* indicated the presence of 133 polypeptides[22], and the venom of another Chinese spider *Chilobrachys guangxiensis (Chilobrachys jingzhao),* comprised 120 polypeptides with a molecular weight of 2,000–8,000 Da[10]. The venom of the tarantula *Psalmopoeus cambridgei* included 150 polypeptides[23] and that of the tarantula *Theraphosa leblondi* included 65 polypeptides[24] and was the least diverse venom reported among tarantulas. Furthermore, the Central Asian species *Agelena orientalis*[25] and the South American *Loxosceles intermedia*[26] included 21 and just over 30 components, respectively.

The diversity of polypeptides in a particular venom can be successfully estimated by gene sequencing. Until recently, the most popular procedure involved cDNA library construction generated on the basis of known toxin structures such as ω- HXTX -Hv1a[11], several huwentoxins[10,27], and hainantoxins[28], but today, researchers pay more attention to EST techniques[29–32]. Parallel proteomic studies of the same spider venom serve as reliable proof for the analysis of EST data[18,28,33–35]. It should be noted that transcriptomic analyses generally identify fewer polypeptide components in spider venoms compared to proteomic analyses. Although the detection of more than 200 individual polypeptides is uncommon[29,35], proteomic analyses usually establishes dozens of structures[10,32,33,36]. Therefore, these two approaches produce unequal predictions of polypeptide variety in spider venom. Apparently, this difference is the result of overestimation by the proteomics approach caused by a large number of false positive data.

Methods

Dolomedes fimbriatus (Clerck, 1,757) spiders were collected in the south European region of Russia. *D. fimbriatus*, of the family Pisauridae, is also known as the raft or fishing spider. The raft spider is semi-aquatic: it hunts on the water surface and can submerge itself if necessary. Venom glands were dissected from several specimens and frozen in liquid nitrogen until sample preparation. To obtain a sufficient amount of mRNA from the venom gland, a preliminary (one week) milking procedure was performed to activate massive toxin expression in accordance with a previous study[31].

Total RNA was extracted with an SV Total RNA Isolation System (Promega, USA). The yield and purity were assessed using a Nanodrop ND-1,000 spectrophotometer (Thermo Scientific, USA), with the RNA integrity determined by the RNA Integrity Number (RIN) using a Bioanalyzer 2,100 (Agilent Technologies, USA). The PCR-based cDNA library was created following the instructions for the SMART cDNA library construction kit (Clontech, USA). Competent *E. coli* One ShotTOP10 cells (Invitrogen, USA) were transformed with the cDNA library plasmids to amplify the cDNA. Plasmid DNA was purified with alkaline lysis and sequenced in both directions using an ABI Prism 3730xl automatic DNA sequencer (Sanger technique) with the BigDye Terminator version 3.1 cycle sequencing kit (Applied Biosystems, USA).

Single-residue distribution analysis (SRDA)[37,38] was used to search the polypeptide structures in the crude EST bank. Basic sequence transformation for the search was performed using the key residue Cys and the termination translation symbol -SRDA ('C.'). The deduced proteins were retrieved from a translated database by 9 structural motifs that consider all the major structural features of spider venom polypeptides (Table 1).

Name	Screening line§	% Structures retrieved
motif 1	c#c*cc*c*c#.	8.7
motif 2	c#c*cc*c1c*c1c#.	1.7
motif 3	c1c*c1c#.	0.7
motif 4	c#c*cc*c*c##.	54.5
motif 5	c#c*cc*c1c*c1c##.	3.4
motif 6	c1c*c1c##.	4.2
motif 7	c#c*cc*c*c*c*c	60.8
motif 8	c#c*cc*c1c*c1c*c*c	35.2
motif 9	c1c*c1c*c*c	43.1

Table 1. The structure of the search motifs used. The last column shows the impact of each motif on the number of overall sequences retrieved. §special symbol used: #—any digit (0–9), *—gap in the search line, . —termination translation symbol encoded by the genes' stop codon.

The presence of some features specific to spider toxin precursor proteins was assessed to identify incorrect data with a high degree of reliability. Because toxin-like polypeptides are secretory proteins that are synthesized together with the preceding signal peptide, the deduced protein structure precursors were evaluated for the presence of the correct eukaryotic signal peptide using a Phobius predictor[39]. The second validation criterion was based on the specificity of a maturation process that is typical for cysteine-containing and linear spider venom polypeptides; this criterion is also known as the presence of the processing quadruplet motif (PQM)[31,40]. The last validation criterion was based on the 5' and 3'-read identity of the clones in a first ATG to termination translation codon range. To determine the novelty of the identified structures, BLASTX was used against a non-redundant protein sequence database.

Data Records

The generated database includes 11,712 ESTs from 5,952 individual clones, with almost all of the clones being sequenced from both ends. Raw data were submitted to GenBank of National Center for Biotechnology Information after verification by VecScreen (http://www.ncbi.nlm.nih.gov/tools/vecscreen/) in according to dbEST sequence deposition requirements (Data Citation 1). The data after verification involved only the deduced sequences of secreted venom gland polypeptides. In total, 7,169 translated sequences were identified as possible polypeptide compounds of *D. fimbriatus* spider venom. Rejected sequences corresponded to partially defined polypeptides (fragments) or secreted molecules with enzymatic, regulatory, structural, and other functions that were beyond the goal of the investigation. Only one exception was made, for linear cysteine-free polypeptides called cytotoxins or antimicrobial peptides (AMPs) GO:0003795.

For comprehensive spider venom characterization, we performed step-by-step transcriptome processing that reckons the growth of the deduced polypeptide number to the size of the EST sequenced (Fig. 1a). We considered all clones coding for identical mature sequences as one polypeptide toxin contig to estimate the amount of compounds in the natural venom. A total of 163 trusted sequences with structures confirmed by at least two repeats according to our results could be completely determined, even from 10,000 sequences (curve with closed squares in Fig. 1a). We assumed that 163 different polypeptides represent the diversity level of *D. fimbriatus* venom. The diversity of mistrusted mature sequences encoded in dbEST was estimated at 420 polypeptides (curve with closed circles in Fig. 1a). One third of all sequenced clones were assembled into 6 contigs, each including more than 100 initial clones, though one contig included 1,063 clones (see Fig. 1b). Approximately 50 contigs have moderate representativeness (groups 6–19 and 20–99 in Fig. 1b). The largest number of deduced structures was assembled into 256 singletons and 113 small-sized contigs (2–5 clones). Overall the toxin structure representation was approximately 61%, which was in good agreement with other spider venom ESTs[10,31]. To determine all moderately present sequences (represented by 6–99 clones), it was sufficient to generate a dbEST from 3,500 sequences (Zone 2 in Fig. 1a), whereas it was only necessary to analyze 300 sequences for the detection of the six major toxins (Zone 1 in Fig. 1a).

In total, 163 mature polypeptides were identified in the investigated venom, but many more nucleotide sequence variants were measured by genes. A total of 451 different sequences encoding polypeptide precursors were counted when comparing structural variation in whole-protein precursor sequences. Another 344 sequence variants with synonymous mutations were found after an alignment of the original

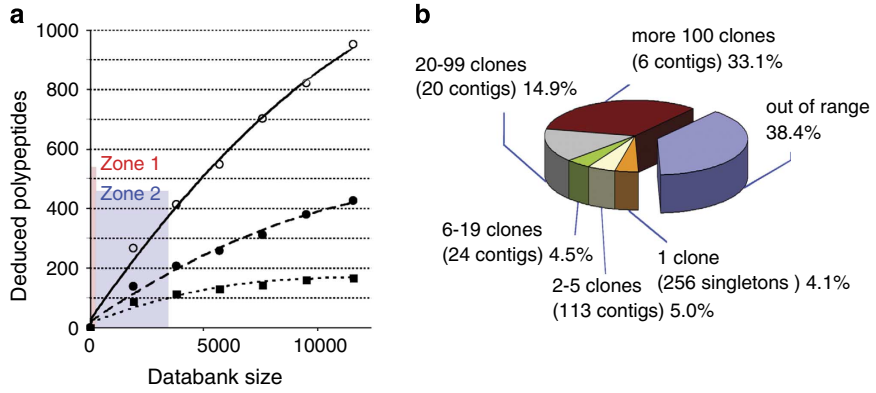

Figure 1. The databank representativeness. (**a**) Relationship of the deduced polypeptide number (ordinate) with the size of the analyzed bank (abscissa). The curve with open circles corresponds to the growth of total toxin-like sequences, the curve with closed circles reflects the growth of validated sequences, and the curve with close squares represents the growth of true sequences. The marked area denotes the bank size that would be sufficient to identify all major compounds (Zone 1) or all moderately distributed compounds (Zone 2). (**b**) The distribution of clones in the dbEST by contig size. The number of unique structures is shown in brackets for each group of contigs. The contigs were generated for exact mature polypeptide sequence deducing. A singleton corresponds to a singular sequence, and the out of range group included all sequences that were not retrieved by the motif search together with the sequences that were not proven by verification.

nucleotide sequences. Thus, the analysis of the studied database revealed the presence of 795 unique nucleotide sequences encoding 451 different polypeptide precursors.

Structural information about the deduced venom polypeptides is available in Supplementary Files (Supplementary File 1 contains all nucleotide sequence variants in FASTA format; Supplementary File 2 contains polypeptide structures expected in crude venom also in FASTA format; Supplementary File 3 includes information about both nucleotide and polypeptide sequences).

Technical Validation

Superfamily features

Because spider venoms commonly contain a large number of homologous sequences with point substitutions[6], it is convenient to designate a group of related sequences as a superfamily. Some peculiarities of the superfamily organization can be illustrated using the first superfamily, consisting of 1,467 ESTs in the total database. After verification, only 644 valid mature sequences were recognized. In the superfamily, the number of errorless nucleotide sequences coding for full-size precursor proteins was 563, with 25 mature venomous polypeptides. For each mature polypeptide, additional transcriptome variants were detected possessing substitutions in the signal and propeptide regions (Table 2). For example, 11 transcripts were found with nonsynonymous substitutions and 16 transcripts with synonymous substitutions that coded for the same mature polypeptide LTDF 01-01. As a result, this peptide is expressed in venom glands from 27 different mRNA sequences and can be described by 11 precursor proteins. For superfamily 01, 78 unique precursors and 136 transcriptome variants were found. The same diversity of nucleotide sequences was observed for other superfamilies and singleton sequences.

The dominance of one nucleotide sequence in each superfamily was observed by inspecting the distribution of unique transcripts. The sequence named as the preferable nucleotide sequence (PNS) was that present in the EST database at a much higher frequency than the other variants. Very rarely, a second transcriptome variant was also moderately represented at a frequency up to 2/3 of that of the PNS. To estimate the superiority of the PNS over other EST variants, we calculated the variability level as the percentage of the total number of sequences encoding each polypeptide represented by other variants. Similar variabilities of approximately 24% were obtained for the well-distributed polypeptides LTDF 01-01, 02, 03, and 04, whereas other members of the superfamily exhibited variability dispersion due to an insufficient amount of data for analysis. We assessed the variability of all polypeptides represented in the database by precursors encoded by more than 5 variants. In Fig. 2a, we divided the obtained results into several groups according to the variant's quantity. The variability value was similar at approximately 20–25% for all groups, and an increase in the number of variants only led to a deviation decrease. Therefore, we can assume the presence of uniform variability machinery for entire spider polypeptides based on one major gene that is available for further mutagenesis. It is obvious that active mutation

Mature polypeptide name	Precursor variants	Transcriptome variants	EST	PNS count	Variability
total in the superfamily	**78**	**136**	**563**		
LTDF 01-01	11	27	151	118	22%
LTDF 01–02	16	31	157	119	24%
LTDF 01–03	15	27	126	95	25%
LTDF 01–04	9	18	74	55	26%
LTDF 01–05	5	6	9	4	56%
LTDF 01–06	1	3	5	3	40%
LTDF 01–07	1	1	4	4	0%
LTDF 01–08	1	2	3	2	33%
LTDF 01–09	1	1	3	3	0%
LTDF 01–10	1	1	2	2	0%
LTDF 01–11	2	2	2	1	50%
LTDF 01–12	2	2	2	1	50%
LTDF 01–13	1	1	2	2	0%
LTDF 01–14	1	1	2	2	0%
LTDF 01–15	1	1	2	2	0%
LTDF 01–16	1	1	2	2	0%
LTDF 01–17	1	1	2	2	0%
LTDF 01–18	1	2	2	1	50%
LTDF 01–19	1	1	1	1	0%
LTDF 01–20	1	1	1	1	0%
LTDF 01–21	1	1	3	3	0%
LTDF 01–22	1	1	2	2	0%
LTDF 01–23	1	2	3	2	33%
LTDF 01–24	1	1	2	2	0%
LTDF 01–25	1	1	1	1	0%

Table 2. Consolidation table of superfamily 01 precursors. The transcriptome variants column indicates the number of unique nucleotide sequences, and the precursor variants column contains the number of unique amino acid sequences. The PNS column presents the number of ESTs equal to the preferable nucleotide sequence for each mature polypeptide. Variability was calculated as the percentage of ESTs not identical to the PNS.

machinery can also produce substitutions in mature chains. These point mutations correlate to the size of a superfamily. We separated the most frequent genes that are transcribed predominantly in PNS inside each superfamily, evaluated their representation in the analyzed EST bank, and calculated the variability of the superfamily as a whole (Fig. 2b).

Most superfamilies demonstrated differences in the number of major genes but were comparable in terms of the variability of minor transcripts. Superfamilies 01 and 03 have the largest number of major genes. These superfamilies each have four major genes with total PNS contents of 387 for superfamilies 01 (see data in Table 2) and 145 (for superfamily 03), corresponding to calculated variabilities of 31 and 40%.

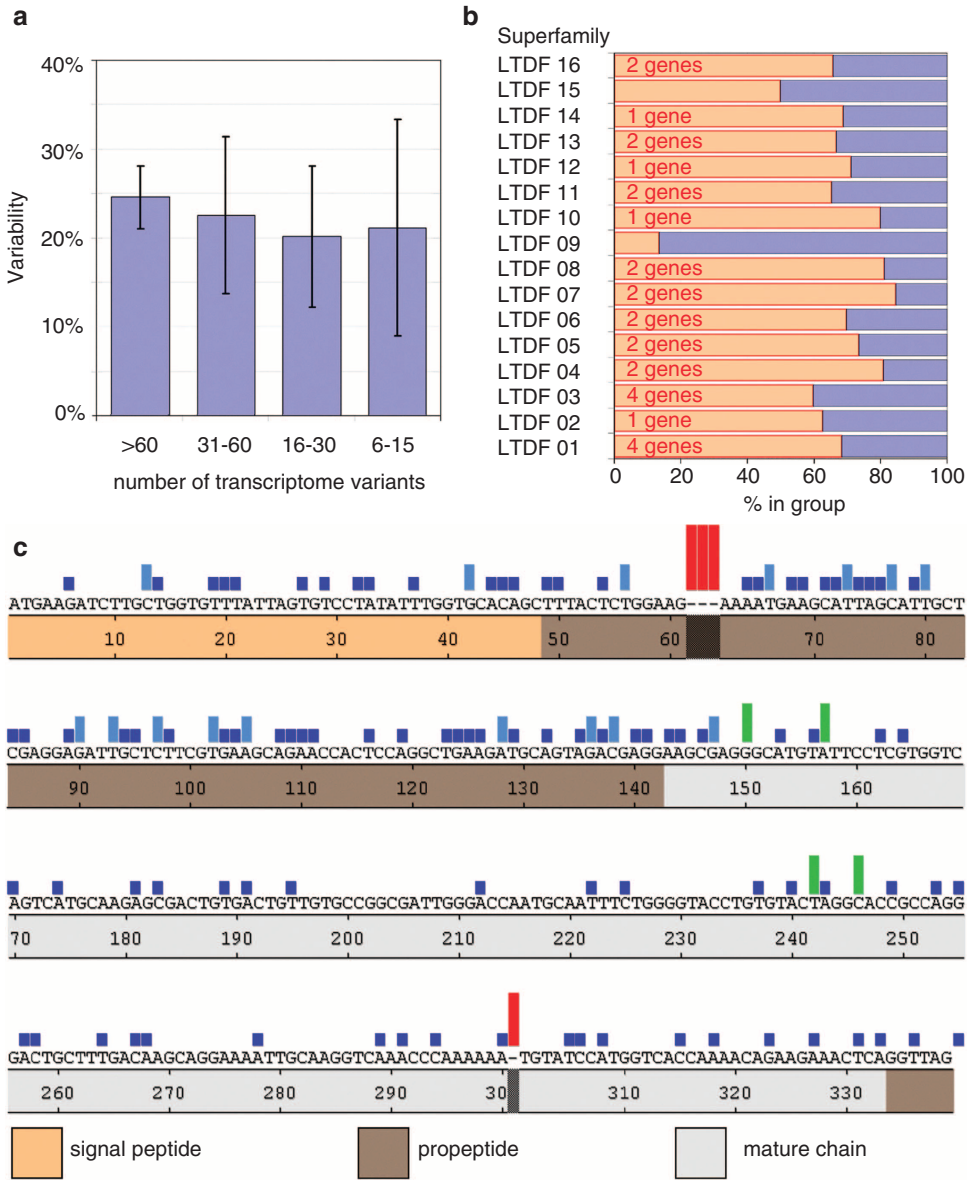

Figure 2. Variability of venomous polypeptides. (**a**) Analysis of the 36 precursor proteins represented in the EST bank by more than 5 transcriptome variants. Standard deviations are shown. (**b**) Variability into 24uperfamilies' for a set of best-represented genes. (**c**) Consensus disagreement for an alignment of nucleotide sequences encoding superfamily 01 protein precursors except the sequences encoding toxins 01–05 and 01–22.

Two superfamilies, 09 and 15, did not include a PNS, and we were unable to identify their major genes. The presence of 1 or 2 major genes was common for the other analyzed superfamilies. The estimated average variability excluding superfamilies 09 and 15 was approximately 29%. To summarize the obtained data, we conclude that there is a core set of 28 major genes in 14 superfamilies that are intensively transcribed in the venom glands of the spider *D. fimbriatus*. There are approximately 2 major genes in each superfamily. The structural abundance of polypeptides is derived from rare transcripts; thus, the total variations inside the superfamily are approximately 30% EST.

We speculate that in contrast to a normal protein's expression, the diversification of major genes into a wide variety of transcripts is an attribute of toxin expression, not only for spiders but for other venomous animals as well. Because the leading roles of gene duplication and diversifying selection have been demonstrated for the formation of functionally variable conotoxins[41], gene duplication is assumed to be the driver of animal toxin diversity[42]. This structural diversity can not be explained by errors of

sequencing and sample preparation, such as dubious data (the curve with open circles in Fig. 1a) have been thoroughly eliminated at the stage of data verification.

A sharp consensus sequence for superfamily 01 was achieved after the elimination of two sequences encoding peptides LTDF 01–05 and LTDF 01–22 (Fig. 2c). The transcripts for these polypeptides were distinct from the other members by 3 insertions of 4, 11, and 4 bp into the mature chain region and by a large unusual 3' region. The remaining members of superfamily 01 exhibited a high level of homology. Strong consensus disagreements were found only for the region encoding the propeptide that had a partial triplet insertion, and for one nucleotide insertion into the mature chain region. This nucleotide insertion of approximately 300 bp leads to a reading frame shift and the production of the longer toxins LTDF 01–06, 07, 21, 23, 24, and 25. For other positions, occasional point mutations were observed. By type, a nucleotide transition occurred 4 times more frequently than a transversion. The more conservative and evolutionarily stable region was located in the area coding for the mature polypeptide, which correlated with previously described variability differences. To date, it has been thought that mature chain sequences should be the predominant mutation sites. In contrast, we found that the most variable region in the analyzed transcripts was the entire propeptide, a large part of the signal peptide, and part of the *N*-terminal sequence of the mature polypeptide (Fig. 2c). Based on the number of substitutions across the full-length protein precursor sequence, a group of polypeptides (LTDFs 01, 02, 06, 12, 13, 20, and 23) with approximately 20 mutations per molecule was clearly distinguished, in contrast to the main group of sequences, which showed 1–8 mutations per molecule.

The analysis of Fig. 2c suggests the probable presence of one or two introns in the major genes of superfamily 01. Such a hypothesis is put forward by us on the basis of a greater variability in the precursor sequence preceding the mature chain, which might indirectly indicate the presence of different splice variants. Alternative splicing is often described for genes coding polypeptide toxins. For cone snails, polypeptide toxins revealed the presence of a quite extended intron inside a propeptide sequence located upstream from the mature chain[43]. Similarly, one or two intronic fragments were found near the end of a signal peptide in scorpion toxin genes[44–46]. In the case of spiders, the situation remains yet undefined, because for some short polypeptide toxins[10,27,47] and for macromolecular latrotoxins[14,48], no introns were found. In contrast, several introns were detected in the genes for long insectotoxins from the venom of the spider *Diguetia canities*[49] and in the genes for sphingomyelinase D from several species of *Loxosceles* and *Sicarius*[50,51].

Polypeptide toxins

All mature sequences were distributed into 16 superfamilies and 19 orphan proteins that did not have homologous polypeptides in the dbEST. The spider *D. fimbriatus* belongs to the superfamily Lycosoidea; thus, the deduced toxin-like polypeptides were named lycotoxins-Df, abbreviated as LTDFs. To distinguish the polypeptides, each was assigned either a number corresponding to a superfamily (01 to 16) or an 'S' for an ungrouped protein together with an ordinal number. The first number was assigned to the most represented sequence and the last number to the rarest one. Because the identity level between superfamily members was rather high, we applied a BLASTX homology search only to the first member in each superfamily. These results are summarized in Table 3. As would be expected, almost all derived polypeptides were found to be homologous to toxins that have previously been identified in spider venoms. However, there were no complete homologies, and all derived polypeptides were found to be novel polypeptide toxin-like molecules. Important levels of homology were found with the cysteine knot toxins predicted from the related spider species *Dolomedes mizhoanus* and to one polypeptide isolated from a natural venom of *Cupiennius salei*[21,32].

In the analyzed transcriptome bank, we found predominant sequences encoding 'inhibitor cysteine knot' (ICK) toxin-like polypeptides (147 out of a total of 163 trusted sequences). The alignment of the discovered toxins among themselves indicated their significant differences from each other in amino acid composition, polypeptide chain length, number of cysteine residues, and distance between cysteines (Fig. 3a). One obvious biological function of ICK toxins is interaction with ion channels.

In addition to ICK toxin-like structures, some polypeptides with other spatial folds were detected in the transcriptome (Fig. 3b). Superfamily 14, with seven members, was larger than many of the ICK polypeptide superfamilies by EST count, but superfamily 15 and LTDF S-18 were rare. The primary feature particular to LTDF 14-01 is the presence in the protein precursor of two propeptides that are removed during maturation. The first propeptide is located between the signal peptide and the heavy protein chain, and a second small one with a length of 7 amino acid residues is found at the *C*-terminus. Such processing was described early for the structural homolog of ω-agatoxin Ia, which was detected in the natural venom as a double-stranded protein[52]. We can suppose that the biological function of polypeptides from superfamily 14 is connected with blocking Ca-channels.

The alignment of the amino acid sequence for another nonstandard protein precursor, LTDF 15-01, indicated moderate homology to a number of known proteins. Unfortunately, no biological function of these homologues has been determined experimentally thus far. Therefore, the functions of the two poorly represented polypeptides from superfamily 15 cannot be predicted. Moderate homology to the spider venom protein PN16C3 (Uniprot ID P84032) was found for LTDF S-18, encoding an extended protein precursor with an estimated molecular weight of greater than 14. The biological function of LTDF 15-01 cannot be predicted.

Superfamily/orphan name	Mature variants	Transcript. variants	Type	%	Homolog	Spider species	ID
LTDF 01	25	136	ICK	79	DMTX-479	*Dolomedes mizhoanus*	S5MFE6
LTDF 02	49	260	ICK	67	DMTX-484	*Dolomedes mizhoanus*	S5MJS8
LTDF 03	13	74	ICK	59	DMTX-174	*Dolomedes mizhoanus*	S5MFE2
LTDF 04	2	15	ICK	68	DMTX-61	*Dolomedes mizhoanus*	S5MYH0
LTDF 05	5	18	ICK	45	PNTx1	*Phoneutria nigriventer*	P17727
LTDF 06	4	16	ICK	54	AgorTX_B7a	*Agelena orientalis*	Q5Y4U3
LTDF 07	2	7	ICK	89	DMTX-41	*Dolomedes mizhoanus*	S5MFI4
LTDF 08	2	7	ICK	45	LSTX-P6	*Lycosa singoriensis*	B6DD57
LTDF 09	9	25	ICK	47	PNTx22C5	*Phoneutria nigriventer*	P84093
LTDF 10	2	3	ICK	40	DMTX-193	*Dolomedes mizhoanus*	S5MK99
LTDF 11	10	52	ICK	81	DMTX-112	*Dolomedes mizhoanus*	S5MFG6
LTDF 12	4	8	ICK	88	DMTX-116	*Dolomedes mizhoanus*	S5N3U8
LTDF 13	3	6	ICK	85	DMTX-142	*Dolomedes mizhoanus*	S5MJV4
LTDF 14	7	58	Cys rich	57	ω-Aga-IA	*Agelenopsis aperta*	P15969
LTDF 15	2	2	Cys rich	43	hypothetical protein	*Latrodectus hesperus*	E7D1U7
LTDF 16	5	43	linear	0			
LTDF S-01	1	6	ICK	42	AgorTX_A4	*Agelena orientalis*	Q5Y4W0
LTDF S-02	1	2	ICK	84	DMTX-12	*Dolomedes mizhoanus*	S5MK94
LTDF S-03	1	1	ICK	87	CSTX-20	*Cupiennius salei*	B3EWT5
LTDF S-04	1	1	ICK	88	DMTX-176	*Dolomedes mizhoanus*	S5MFH9
LTDF S-05	1	1	ICK	0			
LTDF S-06	1	10	ICK	92	DMTX-106	*Dolomedes mizhoanus*	S5MJV8
LTDF S-07	1	10	ICK	67	DMTX-174	*Dolomedes mizhoanus*	S5MFE2
LTDF S-08	1	11	ICK	41	AgorTX_B7a	*Agelena orientalis*	Q5Y4U3
LTDF S-09	1	4	ICK	53	DMTX-29	*Dolomedes mizhoanus*	S5N3V5
LTDF S-10	1	3	ICK	45	Pn3A	*Phoneutria nigriventer*	P81793
LTDF S-11	1	5	ICK	30	hypothetical protein	*Caenorhabditis remanei**	E3NQN5
LTDF S-12	1	1	ICK	87	DMTX-29	*Dolomedes mizhoanus*	S5N3V5
LTDF S-13	1	1	ICK	55	DMTX-104	*Dolomedes mizhoanus*	S5MYC2
LTDF S-14	1	1	ICK	48	DMTX-220	*Dolomedes mizhoanus*	S5N3S4
LTDF S-15	1	2	ICK	43	LSTX-L10	*Lycosa singoriensis*	B6DD28
LTDF S-16	1	2	ICK	45	Tx3–5	*Phoneutria nigriventer*	P8179
LTDF S-17	1	2	ICK	39	LSTX-P6	*Lycosa singoriensis*	B6DD57
LTDF S-18	1	1	Cys rich	40	venom protein PN16C3	*Phoneutria nigriventer*	P84032
LTDF S-19	1	1	linear	0			

Table 3. Polypeptide homologies by BLASTX. The best values of homology correspond to a prevalent mature polypeptide sequence in each superfamily or orphan proteins. The column of mature variants indicates the number of polypeptides in the superfamily, and the transcriptome variants column indicates the diversity of nucleotide sequences in the superfamily. The name of the spider species and database ID for the homologous proteins are included. *polypeptide from nematode.

a

```
LTDF 01-01  ----------ACIPRGQS--CKSDCD----------CCAGDWDQCNFWGTCVQGTARDCFDKQENCKVKPKKCIHGHQNRRNS*
LTDF 11-01  ---------EIYCPKSRNPT-CNLSYKINH--------CCSQAECRMGDVCCVEACGNVCRRGSDTAQGKKFVDGSECTLGREWKSGWF*
LTDF 12-01  ---------SGYCPDFTVKI-CPYKYKIND--------CCAQADCPAGAVCCVQPCGNVCRREYGTPIGEPFVDGTDCKLGHVYPEKWYQKLWHGIS*
LTDF 13-01  -------NELIFCPRPIDEGECPLKSKSNE--------CCSQNDCRSGHLCCTEPCGNVCRVPSDEKMGQMFEEGTKCHLGKEHKHWWSGLFSRSEILKMSGDTIL
LTDF S-01   ----------NCGKTGDD--CEYRYER----------CCDGFRCRYTGRTARLFICL
LTDF S-02   ----------KYCSSSSE---CGEGM-----------CCTGGSYNRHCQRLAEDGTPCQRPNKYDQYSTGCPCKEGLLCSVINYCQKA
LTDF S-03   ---------YTACTKQSD---CAEDE-----------CCLDNLFFKRPYCEKRYGAGQRCSATAIYKEEKDLYYLTCPCVAMYECLGKGSTDENGNTIMKNPKCIMPT
LTDF S-04   IVCPSNYCDKVDCEELTE---CRESNGLRIREKGSFCQCCDICVKVLGEGERCQPEGEFLGVIITSECAKDLVCDYNSRRCTRIGV
LTDF S-05   --------EEEKCLPEGSE--CKATSGPN---------CCGISTRCIVWDTQLPAKHGGGSATWISKCRKYKLGVALEAIGDAFKTKE
LTDF 02-01    ------KCIANGNSCKSDCD---CCYGAYDH--CSCWLWS----------CSCVTGTANDCFNKKSKCNMPNLKCVTRGQNPRNR*
LTDF 03-01    ------SCVGPRQKCESDCD---CCLGDWDT--CSCTLGLFF---------CSCVSGTMRDCMWKDKNCAVKPKKCFMPGQNPRQHKN*
LTDF 04-01    ------SCREPNESCWYPSKP--CCKNRA----CKCGLFFNN---------CECYQPPSEIF*
LTDF 05-01    ---EKKTCIPQGGECDPYKKECQCCGDFTY---CGCYWGVKLFGQ------CKCKYGNIETCYDKYNCDNRSAWTSKPFNQKDCHRG*
LTDF 06-01    ------SCTPIGGDCEPGNDCK-CCNKWAY---CKCTLFNKWG--------CSCVFGDAMVCQRKKEQCRDPQVMDFPKGPCFYSTRKNKRRMT*
LTDF 07-01    -----GYCAENGIKCNDLH----CCSGLV----CKCNSSRSN---------CVCR
LTDF 08-01    -----GLCIRPNMPCGRSVP---CCSNESYNRPCICSLIGTN---------CTCRKTASEILGGWFS
LTDF 09-01    EVTWDTPCIDLGQTCVKGESDCQCCRDNAV---CSCSWLFRSH--------CTCQVFDARASYDTCLIKVNCPNRHKWDRGSKTCPKPCESKHCS*
LTDF 10-01    -----RFCAGLSRTCSKTIK---CCKHGSYERVCYCSTPAGTN--------CRCKKKSSEIVSGWL*
LTDF S-06     -----KTCAGLYKECDAPRLP--CCENRV----CKCSFAFTN---------CKCHKRPSQLIADMFS*
LTDF S-07     ----KRSCTRRGQSCEDDCE---CCADDWDT--CSCTFGLFG---------CSCAHGTARDCYIKAEKCANKPKKCHMPGQNPRN*
LTDF S-08     ------KCVHLGAECSNSCD---CCGDTTY---CNCPLFGLFR--------CSCVPSTSDVCHTKMAECAVKPKSCKVSADHRR*
LTDF S-09     ------GCMKWHGPCNGEDDKS-CCKGK-----CVCWNQRVARGSSRTVTECKCMLW
LTDF S-10     ---QAPACREVNQDCYYPTKP--CCHQRP----CVCSMSGLFSFN------CKCKRF
LTDF S-11     ----KRSCYLKGEECTKDID---CCTMQ-----CLCTTND----------CRCGKRPRPSELRKYFTKRTCGGT
LTDF S-12     ------ACTKWRMSCSGENDNS-CCWPYG----CQCWSQTVSKKSSRKERKCQCRLW
LTDF S-13     ------ACIRQGQTCNNDCD---CCVGDFDT--CSCTLGLFF---------CSCVSGTVNDCTWKEKNCAIKPNKCLLGFLPRQHKN*
LTDF S-14     -----TPCLKTGAECDGSKDDCQCCGAYVE---CKCPFGINWPSVLGP---CKCSMDHLGTCLAKQRCPNKHEWGGGDCKTPKKPKY*
LTDF S-15     ------SCVSPGGVCKDDCD---CCINNNY---CHCPLWGIFGAKS-----CSCIFGDNIVCKERMGKCKRNRPQKCPSTRRSRRYL
LTDF S-16     ------RCLQLREECSGNEKG--CCWPAR----CNCWNQGGGRGSRK----CICALW
LTDF S-17     ---KDSTCIKPPQVCSPSGVG--CCYGSR----CICDAPTTR---------CTCFENYKESLI*
```

b

```
LTDF 14-01              MVKSIVVFAFLIAAVYSFAAEQDEDLFSQDAEVEVVPEEARR---SLPPGSKCDGNAPQPHCQCY
ω-Aga-IA                MMKFVVFLACLFVAAHSFAVEGEEEYF----EAEV-PELE-R-AKALPPGSVCDGNESD--CKCY
LSTX-M3                 MIKYVLISALLVVAVYSFTIEDSEDALLEEAEDELDTEEERR--MALPPGAVCNGHKSD--CQCF

LTDF 14-01              GKWHRCDC------DG-ARCHCKPGTKHTCIQKLKCPNKREWGPYWRSEETGRSPC
ω-Aga-IA                GKWHKCRCPWKWHFTGEGPCTCEKGMKHTCITKLHCPNKAEWGLDWRSEESERSPC
LSTX-M3                 GAKYKCSCPFLWRFRRSAKCHCKKGWAWTAIKKRSCHNRYQWSG

LTDF 15-01              -----MLQTLY--------ISLFFGCTFLLLDTNAYTYIQPLNTESGFCEGPGF-KVAVGEEGYSTDES
hypothetical protein    HEESAILQELQRKMNRYGLLVMCFGLMSMILHTEGYTYMGKVDTSSGFCMMEKLGKIPVGESVYD-DEK
salivary gland peptide  -----MLFPAA-------LVLGLVAASSVWKVSQAEPPFSDVVIVDGKCTYENH-TVANGES-LKWADP
venom protein 7          MSWNNFLVLCP--------SLFFFGTITGNLDFITYVNNRMVPVRNDKCDLDYK-LLDDGEVWYR-TDF

LTDF 15-01              CEKILCH----HGSREVFGCGKVIPPSDPQCHMVRESGHHPDCCL-QVRCT
hypothetical protein    CEKIHCG----QGMYDGMGCGTVSIPESSGCRLEKGTGHYPNCCQ-QPKCD
salivary gland peptide  CQTWECN--VEERSFAVLGCGLVANP--SGCELVRGTGPYPDCCY-RAVCP
venom protein 7          CEKIYCKLVDEFGLTEVKGCVPIT-PVPPNCTIVGRRGTYPDCCVGEVVCN....

Venom_protein PN16C3  ----------------ARPKSDCEKHRESTEKTGTIMKLIPKCKENSDYEELQCYEDSKFCVCYDKKGHAASP
LTDF S-18               MKAVCSVIVFSLCLALALAKTDCEEHRERELKSNTKVKLVPKCDANGDYETLQCFQGSPFCMCWRKDGSHITD

Venom_protein PN16C3  ISTKVKECGCYL-KQKE--RKDSGRESAIIPQCEEDGKWAKKQLWEFNKSCWCVDEKGEQVGKIHHDCDSLKCE
LTDF S-18               PSIKIKTCVCHAHRDKELRRSQTGMVGNFIPSCQESGYYSRKQCHGSTGYCWCADDNGNKVSESVR--GTANC
```

c

```
Pc21g12660  MTSVDELFKKPSTSGSAKRKFEPIQDPNEVYKSAKLDPSGDAISKGKAPMVEDE--IDDGEAGPELPPDFEEDVPDDEEGRFFGGGMERKTAQAMEYIDEQEEEAAPEK....
LTDF 16-01  MNYFKVTCLIFVVLVSTAKSSTGEFENEFEEELETKEIEAELNDDLEEEIARSQALID----------------EEEARFFNGLMLKKT---MEKLKKQ--EAAKRK*

LTDF S-19               ...GEQNSTKRKNNRKLIIRLAPRF-PYIAKQ
hypothetical protein                 ...KLIVDLAPRFSPYIAK...
```

Figure 3. Derived sequences of polypeptides. (**a**) Mature ICK toxin structures. The * symbol after the *C*-terminal amino acid residue indicates an amidation. The key cysteine residues are highlighted in accordance with the two main motifs of spider toxins, PSM (blue) and ESM (pink). (**b**) Alignment of non-ICK toxins. Identical residues are highlighted. Amino acid residues identified in the structure of Ca-channels blocker ω-agatoxin Ia as removable by maturation are underlined. Venom protein-7 from the scorpion *Mesobuthus eupeus* presented a partial structure without 25 *C*-terminal residues. (**c**) Comparison of linear peptide sequences. Similar amino acid residues are highlighted. For LTDF S-19 peptide, only the structure of a mature chain is shown. The * symbol in LTDF 16-01 sequence indicates amidated *C*-terminal residue.

The investigated transcriptome belongs to a Lycosoidea spider, the venoms of which typically contain AMPs, but the amount of deduced linear peptides was negligible. Orphan protein LTDF S-19 and the well-represented superfamily 16 consisted of 5 different sequences were found (Fig. 3c). All identified linear peptides were of small size. A homolog search in the Uniprot database found a certain similarity of the LTDF 16-01 mature chain to filamentous proteins from the fungus *Penicillium chrysogenum*[53]. For the linear peptide LTDF S-19 homologs were detected only in a large number of short fragments from hypothetical protein.

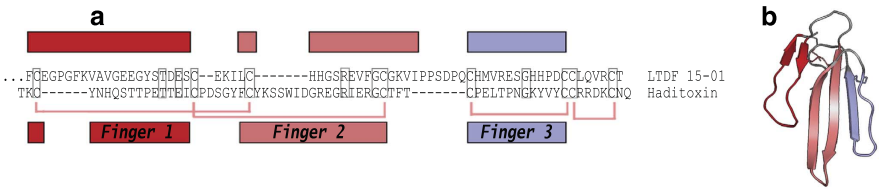

Figure 4. 'Three-finger' neurotoxin. (**a**) Alignment of LTDF 15-01 to king cobra neurotoxin haditoxin. The pattern of cysteine bridges and finger regions is shown in accordance with 3D data[56]. Equal residues are boxed, and the probable finger region is drawn for LTDF 15-01 based only on similarity. For LTDF 15-01, a partial sequence is shown without the 12 *N*-terminal residues. (**b**) 3D structure of haditoxin.

Usage Notes

One peculiarity of a spider venom combinatorial library is the large number of genes with point mutations, which can lead to sequencing procedure limitations. If the reads obtained during the analysis are too short, further attempts at full-length gene reconstruction from several pieces will distort the fine details. As a result, a sequence can be obtained only for the most well-represented transcripts, and the number of coded toxins will be underestimated. The same underestimate of the number of components by a database analysis may occur when contig assembly does have not sufficiently strict parameters. The method based on SRDA and toxin primary structure motifs is more effective for the thorough analysis of combinatorial libraries, as previously confirmed[31,37]. The investigation of nucleotide sequences treated by any enzyme raises the question of how much error was introduced in the sequence array. The technique of double sequencing a clone from forward and reverse primers helps to eliminate such errors, as the probability of a polymerase mistake in the same place several times is negligible. The double sequencing cleared a most part of validated sequences (major reduction from 420 to 163 as shown on Fig. 1a). The BLASTX algorithm can also be considered as a tool for error elimination during EST database analysis. The true homology can be found in other reading frames in the case of a probable non-homologous protein retrieved by the bank screening. We discarded several sequences that thoroughly satisfied the other criteria but showed sufficient homology to a known protein in an alternative reading frame.

The most important point is how the real quantity of polypeptide structures in spider venom can be estimated. The dissimilarity of the compounds identified in the transcriptome from the number of components detected in natural venoms is common and not confined to spiders[54,55]. It has been reported that the most structurally rich spider venoms can contain approximately 500 individual components[18], but recent proteomic studies on the basis of combinations of various separation and detection methods measured approximately 200 components in some spider venoms[21,35]. It is clear that the number of components in venom varies between spider species; moreover, venom composition is observed to vary within species when the venoms of several individuals are compared[25].

A thorough variation search in the transcriptome allowed us to discover important features of spider polypeptide organization. First, there was a strong dominance of ICK folds over other structures. For disulfide-stabilized polypeptides, three alternative folds were found. There were two uncommon spatial folds: the first is similar to ω-agatoxin 1a (superfamily 14) and the second a novel one for toxin LTDF S-18. Another rare fold appears to resemble the well-known 'three-finger' fold of snake neurotoxins (superfamily 15). We assume the possibility of the presence of a toxin-like polypeptide that is most likely folded in the same way as 'three-finger' neurotoxins. In fact, the linear homology of the two polypeptides from superfamily 15 with the snake neurotoxin[56] is quite low, but the arrangement of cysteine residues shows a common peculiarity (Fig. 4). The number of cysteine residues and the distances between the four *C*-terminal cysteines are identical to the three-finger neurotoxin. These differences occur in other parts of the polypeptides: the distance between cysteines 1 and 2 (first finger domain) is longer by 6 amino acids, and the distance between cysteines 3 and 4 (second finger domain) is shorter by 7 amino acids. We assume the possibility of a 'three-finger' fold for spider polypeptides with altered sizes of finger 1 and finger 2 and the same disulfide bond bridging as in snake neurotoxins, but this assumption requires further prove. The presence of the 'three-finger' motif suggests that, originally, venomous terrestrial animals had similar sets of genes for polypeptides with different folds. However, over the course of evolution, ICK polypeptides became predominant in spiders, reaching a large variety of structures, while the development of non-ICK polypeptide diversity was eliminated.

References

1. Brust, A. *et al.* chi-Conopeptide pharmacophore development: toward a novel class of norepinephrine transporter inhibitor (Xen2174) for pain. *J. Med. Chem.* **52,** 6991–7002 (2009).
2. Mamelak, A. N. & Jacoby, D. B. Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601). *Expert Opin. Drug Deliv.* **4,** 175–186 (2007).
3. Wermeling, D. P. & Berger, J. R. Ziconotide infusion for severe chronic pain: case series of patients with neuropathic pain. *Pharmacotherapy* **26,** 395–402 (2006).

4. King, G. F. Venoms as a platform for human drugs: translating toxins into therapeutics. *Expert Opin. Biol. Ther.* **11,** 1469–1484 (2011).
5. Andreev, Y. A., Vassilevski, A. A. & Kozlov, S. A. Molecules to selectively target receptors for treatment of pain and neurogenic inflammation. *Recent Pat. Inflamm. Allergy Drug Discov* **6,** 35–45 (2012).
6. Vassilevski, A. A., Kozlov, S. A. & Grishin, E. V. Molecular diversity of spider venom. *Biochem.* **74,** 1505–1534 (2009).
7. Olivera, B. M. Conus peptides: biodiversity-based discovery and exogenomics. *J. Biol. Chem.* **281,** 31173–31177 (2006).
8. Kaas, Q., Westermann, J. C. & Craik, D. J. Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicon* **55,** 1491–1509 (2010).
9. Escoubas, P. Molecular diversification in spider venoms: a web of combinatorial peptide libraries. *Mol. Divers.* **10,** 545–554 (2006).
10. Chen, J. *et al.* Molecular diversity and evolution of cystine knot toxins of the tarantula Chilobrachys jingzhao. *Cell. Mol. Life Sci.* **65,** 2431–2444 (2008).
11. Sollod, B. L. *et al.* Were arachnids the first to use combinatorial peptide libraries? *Peptides* **26,** 131–139 (2005).
12. Magazanik, L. G. *et al.* Selective presynaptic insectotoxin (Alpha-Latroinsectotoxin) isolated from black-widow spider venom. *Neuroscience* **46,** 181–188 (1992).
13. Fletcher, J. I. *et al.* The structure of a novel insecticidal neurotoxin, omega-atracotoxin-HV1, from the venom of an Australian funnel web spider. *Nat. Struct. Biol.* **4,** 559–566 (1997).
14. Danilevich, V. N., Lukyanov, S. A. & Grishin, E. V. Cloning and structure determination of the α-latrocrustoxin gene from the black widow spider venom. *Russ. J. Bioorganic. Chem.* **25,** 477–486 (1999).
15. Billen, B. *et al.* Unique bell-shaped voltage-dependent modulation of Na+ channel gating by novel insect-selective toxins from the spider Agelena orientalis. *J. Biol. Chem.* **285,** 18545–18554 (2010).
16. Escoubas, P., Quinton, L. & Nicholson, G. M. Venomics: unravelling the complexity of animal venoms with mass spectrometry. *J. Mass. Spectrom.* **43,** 279–295 (2008).
17. Escoubas, P. & King, G. F. Venomics as a drug discovery platform. *Expert Rev. Proteomics* **6,** 221–224 (2009).
18. Escoubas, P., Sollod, B. & King, G. F. Venom landscapes: mining the complexity of spider venoms via a combined cDNA and mass spectrometric approach. *Toxicon.* **47,** 650–663 (2006).
19. Wang, H. *et al.* The venom of the fishing spider Dolomedes sulfurous contains various neurotoxins acting on voltage-activated ion channels in rat dorsal root ganglion neurons. *Toxicon.* **65,** 68–75 (2013).
20. Richardson, M. *et al.* Comparison of the partial proteomes of the venoms of Brazilian spiders of the genus Phoneutria. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **142,** 173–187 (2006).
21. Trachsel, C. *et al.* Multicomponent venom of the spider Cupiennius salei: a bioanalytical investigation applying different strategies. *FEBS J.* **279,** 2683–2694 (2012).
22. Yuan, C. *et al.* Proteomic and peptidomic characterization of the venom from the Chinese bird spider, Ornithoctonus huwena Wang. *J. Proteome Res.* **6,** 2792–2801 (2007).
23. Choi, S. J. *et al.* Isolation and characterization of Psalmopeotoxin I and II: two novel antimalarial peptides from the venom of the tarantula Psalmopoeus cambridgei. *FEBS Lett.* **572,** 109–117 (2004).
24. Legros, C., Celerier, M. L., Henry, M. & Guette, C. Nanospray analysis of the venom of the tarantula Theraphosa leblondi: a powerful method for direct venom mass fingerprinting and toxin sequencing. *Rapid Commun. Mass. Spectrom.* **18,** 1024–1032 (2004).
25. Shlyapnikov, Y. M., Kozlov, S. A., Fedorov, A. A. & Grishin, E. V. A comparison of polypeptide compositions of individual Agelena orientalis spider venoms. *Russ. J. Bioorganic Chem.* **36,** 73–80 (2010).
26. Gremski, L. H. *et al.* A novel expression profile of the Loxosceles intermedia spider venomous gland revealed by transcriptome analysis. *Mol. Biosyst.* **6,** 2403–2416 (2010).
27. Jiang, L. *et al.* Genomic organization and cloning of novel genes encoding toxin-like peptides of three superfamilies from the spider Orinithoctonus huwena. *Peptides* **29,** 1679–1684 (2008).
28. Tang, X. *et al.* Molecular diversification of peptide toxins from the tarantula Haplopelma hainanum (Ornithoctonus hainana) venom based on transcriptomic, peptidomic, and genomic analyses. *J. Proteome Res.* **9,** 2550–2564 (2010).
29. Zhang, Y. *et al.* Transcriptome analysis of the venom glands of the Chinese wolf spider Lycosa singoriensis. *Zool* **113,** 10–18 (2010).
30. Fernandes-Pedrosa Mde, F. *et al.* Transcriptome analysis of Loxosceles laeta (Araneae, Sicariidae) spider venomous gland using expressed sequence tags. *BMC Genomics* **9,** 279 (2008).
31. Kozlov, S. *et al.* A novel strategy for the identification of toxinlike structures in spider venom. *Proteins* **59,** 131–140 (2005).
32. Jiang, L. *et al.* Transcriptome analysis of venom glands from a single fishing spider Dolomedes mizhoanus. *Toxicon* **73,** 23–32 (2013).
33. Jiang, L. *et al.* Venomics of the spider Ornithoctonus huwena based on transcriptomic versus proteomic analysis. *Comp. Biochem. Physiol. Part D Genomics Proteomics* **5,** 81–88 (2010).
34. Diego-Garcia, E., Peigneur, S., Waelkens, E., Debaveye, S. & Tytgat, J. Venom components from Citharischius crawshayi spider (Family Theraphosidae): exploring transcriptome, venomics, and function. *Cell. Mol. Life Sci.* **67,** 2799–2813 (2010).
35. Duan, Z., Cao, R., Jiang, L. & Liang, S. A combined de novo protein sequencing and cDNA library approach to the venomic analysis of Chinese spider Araneus ventricosus. *J. Proteomics* **78,** 416–427 (2013).
36. Jiang, L. *et al.* Molecular diversification based on analysis of expressed sequence tags from the venom glands of the Chinese bird spider Ornithoctonus huwena. *Toxicon* **51,** 1479–1489 (2008).
37. Kozlov, S. & Grishin, E. The mining of toxin-like polypeptides from EST database by single residue distribution analysis. *BMC Genomics* **12,** 88 (2011).
38. Kozlov, S. & Grishin, E. Classification of spider neurotoxins using structural motifs by primary structure features. Single residue distribution analysis and pattern analysis techniques. *Toxicon* **46,** 672–686 (2005).
39. Kall, L., Krogh, A. & Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **35,** W429–W432 (2007).
40. Kozlov, S. A. & Grishin, E. V. The universal algorithm of maturation for secretory and excretory protein precursors. *Toxicon* **49,** 721–726 (2007).
41. Duda, T. F. Jr. & Palumbi, S. R. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus. *Proc. Natl Acad. Sci. USA* **96,** 6820–6823 (1999).
42. Wong, E. S. W. & Belov, K. Venom evolution through gene duplications. *Gene* **496,** 1–7 (2012).
43. Yuan, D. D., Han, Y. H., Wang, C. G. & Chi, C. W. From the identification of gene organization of alpha conotoxins to the cloning of novel toxins. *Toxicon* **49,** 1135–1149 (2007).
44. Zhijian, C. *et al.* Cloning and characterization of a novel calcium channel toxin-like gene BmCa1 from Chinese scorpion Mesobuthus martensii Karsch. *Peptides* **27,** 1235–1240 (2006).
45. Xu, X. *et al.* Genomic sequence analysis and organization of BmKalphaTx11 and BmKalphaTx15 from Buthus martensii Karsch: molecular evolution of alpha-toxin genes. *J. Biochem. Mol. Biol.* **38,** 386–390 (2005).

46. Froy, O. *et al.* Dynamic diversification from a putative common ancestor of scorpion toxins affecting sodium, potassium, and chloride channels. *J. Mol. Evol.* **48,** 187–196 (1999).
47. Qiao, P., Zuo, X. P., Chai, Z. F. & Ji, Y. H. The cDNA and genomic DNA organization of a novel toxin SHT-I from spider Ornithoctonus huwena. *Acta Biochim. Biophys. Sin.* **36,** 656–660 (2004).
48. Danilevich, V. N. & Grishin, E. V. The genes encoding black widow spider neurotoxins are intronless. *Bioorganicheskaya Khimiya* **26,** 933–939 (2000).
49. Krapcho, K. J., Kral, R. M. Jr., Vanwagenen, B. C., Eppler, K. G. & Morgan, T. K. Characterization and cloning of insecticidal peptides from the primitive weaving spider Diguetia canities. *Insect Biochem. Mol. Biol.* **25,** 991–1000 (1995).
50. Binford, G. J., Cordes, M. H. & Wells, M. A. Sphingomyelinase D from venoms of Loxosceles spiders: evolutionary insights from cDNA sequences and gene structure. *Toxicon* **45,** 547–560 (2005).
51. Binford, G. J. *et al.* Molecular evolution, functional variation, and proposed nomenclature of the gene family that includes sphingomyelinase D in sicariid spider venoms. *Mol. Biol. Evol.* **26,** 547–566 (2009).
52. Santos, A. D. *et al.* Heterodimeric structure of the spider toxin omega-agatoxin IA revealed by precursor analysis and mass spectrometry. *J. Biol. Chem.* **267,** 20701–20705 (1992).
53. Van den Berg, M. A. *et al.* Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. *Nat. Biotechnol.* **26,** 1161–1168 (2008).
54. Morgenstern, D. *et al.* The tale of a resting gland: Transcriptome of a replete venom gland from the scorpion Hottentotta judaicus. *Toxicon* **57,** 695–703 (2011).
55. Ma, Y. *et al.* Molecular diversity of toxic components from the scorpion Heterometrus petersii venom revealed by proteomic and transcriptome analysis. *Proteomics* **10,** 2471–2485 (2010).
56. Roy, A. *et al.* Structural and functional characterization of a novel homodimeric three-finger neurotoxin from the venom of Ophiophagus hannah (king cobra). *J. Biol. Chem.* **285,** 8302–8315 (2010).

Data Citation

1. Kozlov, S. *et al.* GenBank JZ520560-JZ530945 (2013).

Acknowledgements

This research work was supported by the Program of Presidium of RAS 'Molecular and Cell Biology' and the Russian Federation president grant for a leading scientific school State support in the Russian Federation No: 1924.2014.4. Eugene Grishin as a principal investigator had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Contributions

S.K., V.L., E.G., and V.G. are responsible for the concept. V.L. constructed the cDNA library. D.A. annotated and submitted the ESTs. EK, OS, and EO sequenced the cDNA library. S.K. carried out database analyses and wrote the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/sdata.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Kozlov, S. A. *et al.* Comprehensive analysis of the venom gland transcriptome of the spider *Dolomedes fimbriatus*. *Sci. Data* 1:140023 doi: 10.1038/sdata.2014.23 (2014).