Tracking the recruitment and evolution of snake toxins using the evolutionary context provided by the *Bothrops jararaca* genome

Diego Dantas Almeida[a,1], Vincent Louis Viala[a,1], Pedro Gabriel Nachtigall[a], Michael Broe[b], H. Lisle Gibbs[b], Solange Maria de Toledo Serrano[a], Ana Maria Moura-da-Silva[c,d], Paulo Lee Ho[e], Milton Yutaka Nishiyama-Jr[a], and Inácio L. M. Junqueira-de-Azevedo[a,2]

[a]Laboratório de Toxinologia Aplicada, Center of Toxins, Immune-Response and Cell Signaling, Instituto Butantan, São Paulo 05503-900, Brazil; [b]Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH 43210; [c]Laboratório de Imunopatologia, Instituto Butantan, São Paulo 05503-900, Brazil; [d]Programa de Pós-Graduação em Medicina Tropical, Universidade do Estado do Amazonas (UEA), Manaus 69040-000, Brazil; and [e]Serviço de Bacteriologia, Divisão BioIndustrial, Instituto Butantan, São Paulo 05503-900, Brazil

Edited by Kerstin Lindblad-Toh, Broad Institute, Cambridge, MA, and approved April 1, 2021 (received for review July 20, 2020)

Venom is a key adaptive innovation in snakes, and how nonvenom genes were co-opted to become part of the toxin arsenal is a significant evolutionary question. While this process has been investigated through the phylogenetic reconstruction of toxin sequences, evidence provided by the genomic context of toxin genes remains less explored. To investigate the process of toxin recruitment, we sequenced the genome of *Bothrops jararaca*, a clinically relevant pitviper. In addition to producing a road map with canonical structures of genes encoding 12 toxin families, we inferred most of the ancestral genes for their loci. We found evidence that 1) snake venom metalloproteinases (SVMPs) and phospholipases A_2 (PLA2) have expanded in genomic proximity to their nonvenomous ancestors; 2) serine proteinases arose by co-opting a local gene that also gave rise to lizard gilatoxins and then expanded; 3) the bradykinin-potentiating peptides originated from a C-type natriuretic peptide gene backbone; and 4) VEGF-F was co-opted from a PGF-like gene and not from VEGF-A. We evaluated two scenarios for the original recruitment of nontoxin genes for snake venom: 1) in locus ancestral gene duplication and 2) in locus ancestral gene direct co-option. The first explains the origins of two important toxins (SVMP and PLA2), while the second explains the emergence of a greater number of venom components. Overall, our results support the idea of a locally assembled venom arsenal in which the most clinically relevant toxin families expanded through posterior gene duplications, regardless of whether they originated by duplication or gene co-option.

snake venom | toxin evolution | genome | gene recruitment | co-option

The evolutionary history of snakes involved striking trait transformations, such as body elongation, limb loss, the development of chemo- and thermoperception, and different sexual reproductive modes and, in some groups, the acquisition of a complex venom apparatus (1). Whereas most of these extreme adaptations are likely controlled by gene systems that are shared in common with other vertebrates, the snake venom system represents a novel key adaptive innovation.

The "advanced snakes" (Caenophidia) developed diverse fang types from the same embryonic origin as their specialized venom glands (VG) (2), which harbor a wide range of bioactive compounds used for predation and defense (3). A large body of knowledge about the evolutionary history of toxin families, selective pressures acting on specific components, and degrees of intra- and interspecific variation in venom was acquired through the sequencing of messenger RNAs (mRNAs) from snake VGs. Thus, many of the hypotheses developed in the last decade about the co-option of proteins involved in physiological functions and the evolutionary origin of venoms, for example, refs. 4 to 7, were largely based on transcript or protein data. However, fundamental questions related to identifying the evolutionary history of this key trait remain. These include the following: 1) "From which preexisting elements did the venom genes arise?" and 2) "How did these ancestral genes transform into toxin genes with unique protein domains?" Recently, large-scale genomic landscapes from venomous snakes became available in the literature (2, 8–14). Gene structures of toxins have been described (2, 15–18), although only a few gene clusters have been studied in a detailed way in viperids (17, 19–21). With these advances, the early origins and the evolutionary routes followed by snake toxins have started to be elucidated and the above questions can now be better addressed with the information provided by the genomic context of toxin genes (22).

Of particular interest, *Bothrops jararaca* (common name, jararaca lancehead) is a representative species of the most diverse and common genus of viperid snakes in South America and provides one of the best-studied models of viperid venom. The venom of *B. jararaca* is diverse in terms of different protein families (23) and different proteoforms (24, 25) present. Many of these components have been characterized (24, 26–31), and early studies on *B. jararaca* toxins helped to establish the basis of the kallikrein-kinin system and led to the development of antihypertensive drugs (32,

Significance

The jararaca lancehead genome provides a comprehensive road map of the genomic context of pitviper toxin genes. Comparisons of these genomic segments across the phylogeny revealed an unexpectedly high number of toxin families that originated via the direct co-option of preexisting nontoxin genes, indicating that the snake toxin arsenal was mostly assembled from local elements of the ancestral genome. These results support a new perspective in venom evolution in which gene duplications in most toxin families occurred after, rather than before, initial toxin recruitment from nontoxin genes, contributing to the evolutionary optimization of snake venoms. They also emphasize the importance of correctly identifying orthologous loci to accurately trace the genomic pathways that lead to the evolutionary origination of new traits.

Author contributions: S.M.d.T.S., A.M.M.-d.-S., P.L.H., and I.L.M.J.-d.-A. designed research; D.D.A., V.L.V., and S.M.d.T.S. performed research; H.L.G. and I.L.M.J.-d.-A. contributed new reagents/analytic tools; D.D.A., V.L.V., P.G.N., M.B., H.L.G., M.Y.N., and I.L.M.J.-d.-A. analyzed data; D.D.A., V.L.V., and I.L.M.J.-d.-A. wrote the paper; and A.M.M.-d.-S. and P.L.H. discussed the results.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]D.D.A. and V.L.V. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: inacio.azevedo@butantan.gov.br.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2015159118/-/DCSupplemental.

Published May 10, 2021.

33). Although the venom of this species has been broadly investigated through transcriptomic and proteomic techniques (7, 34–37), its genomic background has yet to be determined.

Here, we address this need by sequencing the genome of *B. jararaca* and then conducting genome prospecting targeting venom-related genes and scaffolds. By retrieving genes for all major toxin classes known in *Bothrops*, we provide a comprehensive but accessible road map of toxin genes from a Viperidae snake. Moreover, by providing the genomic contexts of these genes relative to homologous loci of other snakes and vertebrates in general, we are able to infer the genes originally located in similar positions in the ancestors and, thus, to deduce the initial steps followed by non-venom genes in becoming part of the snake venom arsenal.

Results and Discussion

B. jararaca Genome Sequencing and Strategies for Targeting Venom Genes. Given the high content of repetitive elements predicted in Viperidae genomes (13, 38, 39), we used four different strategies to optimize the chance of obtaining full-length genes and long genomic segments of interest (Fig. 1): 1) a main hybrid assembly of short and long reads of whole-genome shotgun sequencing (HA-WGS), 2) the screening of toxin genes in independent assemblies of subsets of high quality short reads (SA-WGS), 3) the direct screening of toxin genes within unassembled long reads (LR-WGS), and 4) the high throughput bacterial artificial chromosome (BAC) sequencing and screening for toxin genes (BAC-SeqSc). The last approach was uniquely designed for this work (Fig. 1, *Right*) and is based on the large-scale sequencing of pools of BACs, which are screened for the presence of toxin genes, with the selected BACs then resequenced with high coverage.

Through the k-mer analyses of the short reads, we estimated the genome size of *B. jararaca* to be 2.1 gigabase pairs (Gbp) (*SI Appendix*, Fig. S1), consistent with the size of 2.2 Gbp predicted previously (40). The assembly of the HA-WGS resulted in an N50 contig size of 163.5 kb, for a total contig length of 1.66 Gbp. We evaluated the completeness of the *B. jararaca* genome assembly using BUSCO (Benchmarking Universal Single-Copy Orthologs) datasets (41). From 3,354 ortholog groups searched in BUSCO, 3,096 (92.3%) were identified; from those, 2,775 (82.7% of total) were complete, and 321 (9.6%) genes were "fragmented." The repetitive sequences totaled 285 megabase pairs (Mbp) (17% of the genome). The most common repetitive elements were retroelements (14.6%), among which the long interspersed nuclear element L2/CR1/Rex was the most abundant one (8.8%), as observed in other snakes (38). The scaffolds were deposited in GenBank under Bioproject PRJNA691605 (74). Genome browsing is available at: http://cetics.butantan.gov.br/gb2/gbrowse/bothrops_jararaca/ (75).

In addition to genomic sequences, we generated transcriptomic data for seven different tissues (VG, gut, kidney, stomach, lung, heart, and brain) of the same *B. jararaca* specimen used for the genome assembly. VG reads were also de novo assembled (42) and annotated by BLAST searches against UniProt and previously annotated transcripts of *B. jararaca* (7). We obtained 45 nonredundant full-length transcripts encoding major venom-related proteins (*SI Appendix*, Table S1). The quantitative toxin profile (*SI Appendix*, Fig. S2) was similar to those previously reported from the same or closely related species (23, 34, 43). The major toxin classes observed were SVMP (snake venom metalloproteinase) class P-III and class P-II, followed by C-type lectins (CTL), phospholipase A_2 (PLA2), bradykinin-potentiating peptide and C-type natriuretic peptide precursor (BPP/CNP), and snake venom serine proteinase (SVSP), accounting for 42.7% of the total VG transcription.

The venom proteome of the same specimen was analyzed by in-solution trypsin digestion and liquid chromatography-tandem mass spectrometry (LC-MS/MS). The protein spectra were searched against the proteins predicted from the transcripts as well as reptile protein sequences available at UniProt (Dataset S1). This analysis confirmed the presence in the venom of almost all toxins predicted from the transcriptome (*SI Appendix*, Table S1).

The Venom-Coding Genome of _B. jararaca_. To obtain an overview of the dataset of genes encoding venom components and their genomic context, we prospected for toxin loci within the whole genomic datasets of *B. jararaca* using the VG transcripts as probes. In total, 55 full-length venom genes (considering the presence of all exons and introns within the coding sequence [CDS]) from 12 different toxin families were identified (*SI Appendix*, Tables S1 and S2).

Our data showed that in *B. jararaca,* most toxin families (PLA2, BPP/CNP, CRISP, HYAL, NGF, VEGF-F, NUCL and PLB) are represented by single genes, although multiple unique genes were recognized for other toxins, such as SVMP (P-III and P-II classes), SVSP, and CTL (*SI Appendix*, Table S2). In the case of the CTL family, there are likely more genes in the genome than we were able to retrieve, since multiple transcripts were recognized in the

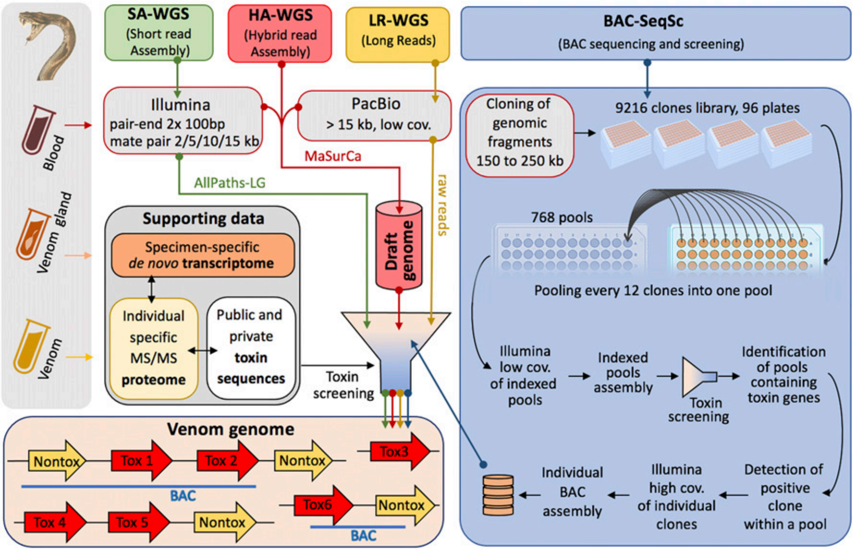

Fig. 1. Schematic diagram of the genomic sequencing strategies used to obtain toxin genes and their flanking regions in *B. jararaca*.

VG (*SI Appendix*, Table S1). On the other hand, no SVMP of the P-I class existed in the specimen sampled in this study, since no such sequence was identified in the genome, the VG transcriptome, or the venom proteome, despite our specific efforts to screen the raw data for this type of metalloproteinase. Since other individuals have been shown to possess P-I class of SVMPs (7, 37, 44), it suggests a polymorphism in the species for this toxin.

Gene size varied greatly, from 1.7 kb (PLA2) to 40.6 kb (PLB). A representative structure from each toxin family is shown in *SI Appendix*, Fig. S3. Within the families containing multiple paralogs, the number and size of the exons were conserved, but the intron sizes varied greatly. Of note is that the first exons of the CRISP gene (exhibiting eight exons) and some SVSP genes (exhibiting five exons) corresponded to noncoding 5′ untranslated region (UTR) sequences. For the NGF gene, only the last exon contained the whole mature protein coding sequence.

By investigating the regions surrounding toxin genes in *B. jararaca* genome, we were able to identify flanking nontoxin genes and thus to recognize gene blocks that could have synteny with the genomes of other snakes, lizards, and nonsquamate Chordata. We then interrogated several Chordata genome sequences to find evidence of synteny with these blocks, reannotating the regions when necessary to assure the correct gene set in the species. The syntenic blocks recognized for each toxin class are summarized in Fig. 2, providing an overview of the whole venom gene landscape. The location of the venom genes between their flanking genes is generally conserved among species, but because of selective pressure and accelerated evolution rates (10, 12), some toxin families vary greatly in terms of copy (paralogs) number and protein primary structure.

For some toxin families (SVMPs and PLA2), it was possible to recognize clusters of venom genes adjacent to a related nonvenom paralog encoding a member of the same protein family that was not expressed in the VGs (red pentagons next to yellow ones in Fig. 2). In the case of PLA2, we inferred the typical gene clustering by considering that the loci sequences described in other Viperidae species contain more PLA2 genes (2, 10, 11, 13, 17, 20, 45), and we assumed the paralog flanking the venom gene as nonvenom because it was not detected in the VG transcriptome nor in the venom proteome of *B. jararaca* (*SI Appendix*, Table S1). However, the presence of only one venom PLA2 gene (and only one transcript) in the *B. jararaca* genome investigated here indicates secondary losses of other PLA2 genes in this specimen and, thus, intraspecific variability at this locus. It is worth mentioning that other transcripts encoding PLA2, including a K-49 PLA2, have been reported in *B. jararaca* VGs (7). Nevertheless, secondary losses of PLA2 genes have been well documented in rattlesnakes (17).

In other venom families (SVSP, HYAL, NGF, VEGF-F, PLB, BPP/CNP, and NUCL), the synteny of the loci indicates that these venom genes are located in the same position occupied by their putative orthologs in nonvenomous species (red pentagons aligned with yellow ones in Fig. 2). HYAL, despite occupying the position of a likely ortholog, is flanked by a gene encoding another hyaluronidase that is not expressed in the VGs, while the others (SVSP, NGF, VEGF-F, PLB, BPP/CNP, and NUCL) do not show any related nonvenom paralogs located nearby at the same locus. For CTL and LAAO, it is not possible to determine whether the corresponding gene positions in nonvenomous species are occupied by orthologs due to a lack of complete locus sequence, and for cysteine-rich secretory protein (CRISP), it is not clear if the adjacent related gene has a role in the venom.

With the above general overview of the venom genes from *B. jararaca* genome and the additional genomic components compiled from public databases and literature, it was possible to explore in more detail the origins and the evolution of specific toxins present in Viperidae venoms. Below, we describe our inferences for four toxin classes of *B. jararaca* venom (SVMP, SVSP, BPP/CNP, and VEGF-F), and we then discuss the general significance of these results for a broad understanding of snake venom evolution in general.

SVMP Genes Show a Discrepancy between Domain and Exon Losses during the P-III to P-II Transition. The HA-WGS strategy of *B. jararaca* genome identified most of the SVMP genes, while the BAC-SeqSc approach recovered seven BAC clones (~150 kb each) providing physical corroboration for some genes (Fig. 3*A*). In total, 20 different (<93% identity) P-III SVMP and seven different (<85% identity) P-II SVMP genes were identified. The BJARBC_30E11N1 scaffold, generated from a single BAC clone, contained one P-II gene (BJARBC_SVMP2_g07) followed by a downstream P-III gene (BJARBC_SVMP3_g03), providing physical evidence of the adjacent positioning of two SVMP classes at the locus. The SA-WGS contig BJARHA_S804283_A also exhibited P-II genes following a string of five P-III genes that are downstream of the flanking ADAM28 gene, which is considered the ancestral gene of all SVMPs (5, 21). Although we were unable to reconstruct the entire locus, the SA-WGS and BAC-SeqSc data together allowed us to infer that *B. jararaca* SVMP genes are organized in a large cluster containing multiple paralogs, starting with the ADAM28 gene (Fig. 3*A*), and this segment is likely flanked by STC1, NEFM, and NEFL genes, as observed in other species. Our results are in accordance with other work that have described a large tandem array of SVMP genes in snakes (11, 13, 14, 20, 21).

It is currently accepted that P-II class SVMPs arose once from a P-III class ancestor at the base of the Viperidae radiation, which may have occurred via gene duplication followed by domain loss (21, 46, 47). A more complete analysis of the exon/intron arrangement of the P-III and the P-II SVMP genes in *B. jararaca* highlighted some details about the initial process leading to the generation of P-II SVMPs. In particular, *B. jararaca* P-III SVMP genes have 17 exons (the CDS starts in exon 1 and ends in exon 17), similar to the ADAM28 gene (up to the Cys-rich domain) but differing from a P-III gene of *Echis ocellatus* (18) in which exons 4, 5, and 6 have merged into a single exon. *B. jararaca* P-II SVMP genes have 15 exons (the CDS starts in exon 1 and ends in exon 14) (Fig. 3*B*). The alignment of a P-III SVMP with a neighboring P-II SVMP gene found in the same BAC showed correspondence of the first 14 exons (Fig. 3*B*). However, we observed that the segment corresponding to part of the disintegrin domain and the entire Cys-rich domain, which was lost upon the P-III to P-II transition, unexpectedly starts in the middle of exon 14, expands to include the entire exon 15 (according to P-III numbering), and ends at the end of exon 16 (long deletion in Fig. 3 *B* and *C*). Therefore, the borders of the domains and the borders of the exons do not exactly match. This is in agreement with the recent observations by Giorgianni and colleagues showing that this deletion is conserved among all P-II SVMP genes in *Crotalus atrox* (21), further corroborating the hypothesis of a single origin of this class of SVMPs (46). However, a simple deletion of entire exons is not sufficient to generate the actual C-terminal region of a P-II SVMP. Instead, an intraexon event would be necessary to complete the deletion of the entire segment.

In fact, we noted another short deletion of 25 bp within exon 14 (short deletion in Fig. 3 *B* and *C*), which caused a frameshift leading to a premature stop codon. Without the acquisition of this stop codon, the simple deletion of the following exons would result in a dysfunctional C-terminal sequence of the protein, likely compromising its structure and function. This short deletion is part of the reason why the disintegrin domains present in P-II precursors are shorter than P-III disintegrin-like domains. The deletion caused the direct removal of eight amino acid residues, and the introduction of a stop codon immediately thereafter prevented the translation of the remaining part of the disintegrin-like domain. The 3′ UTR of a P-II transcript was consequently added, with a small extension of 65 bp (Fig. 3), but the remaining 3′ UTR was mostly unchanged, likely preserving regulatory sites of the mRNAs. Our

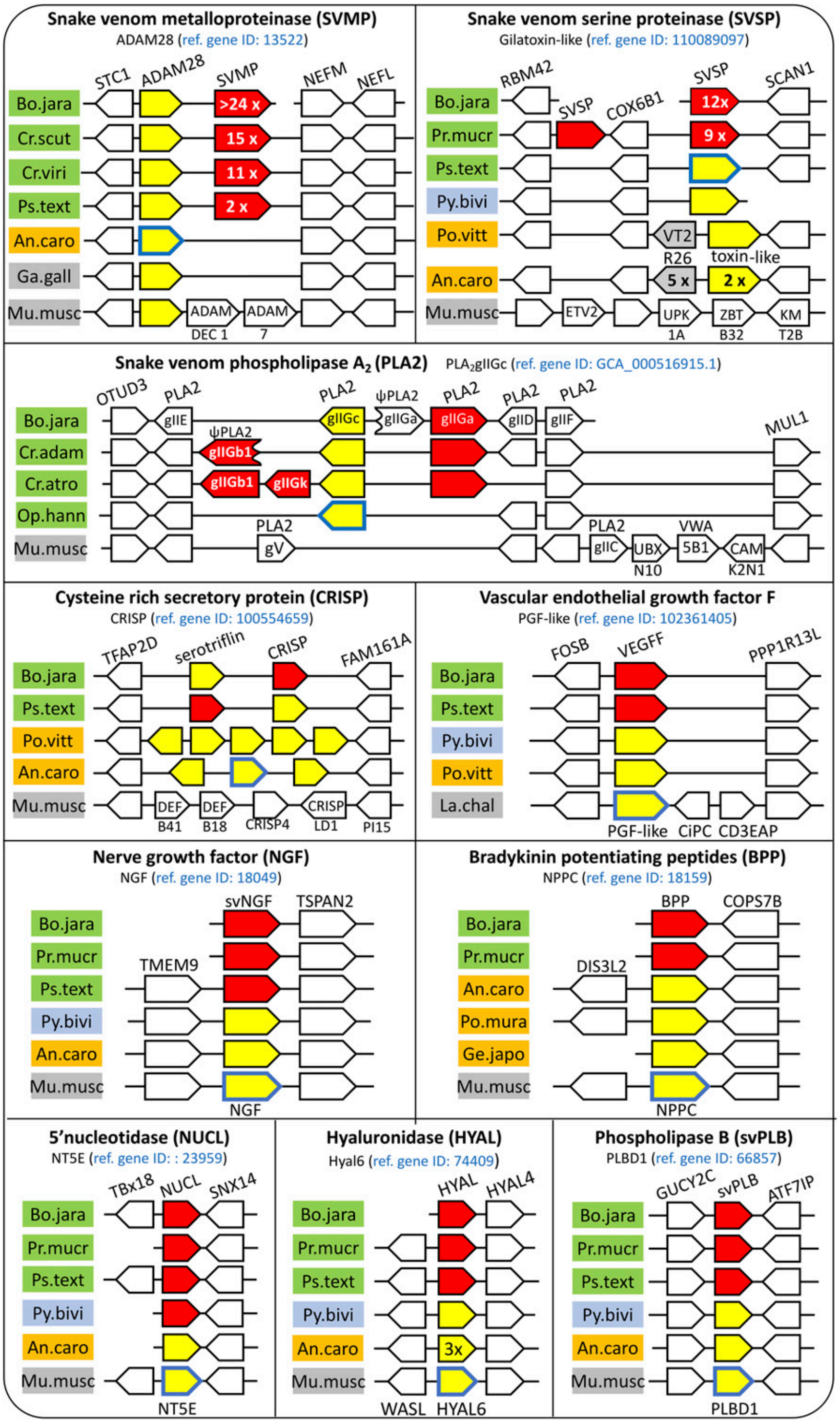

Fig. 2. Schematic architecture of venom gene loci of different toxins showing syntenic blocks among different species. Red pentagon, toxin gene; yellow pentagon, nontoxin ortholog of a toxin gene (or paralog if in the same species); and white pentagon, flanking nontoxin gene. In each box, the name of the ortholog representing the putative ancestral gene for the toxin family is noted bellow the toxin family name, followed in parenthesis by the gene ID of a reference gene (which is noted in blue and outlined in blue in the scheme) from an organism that do not contain the toxic character for this family. Gene names are indicated over the array of orthologs or within pentagons. Some paralogous genes are represented by one pentagon internally marked with the number of paralogs occurring in the species. Relevant pseudogenes are indicated with Ψ. Species were classified according to the following color code: green box, venomous snake; blue box, nonvenomous snake; orange box, nonsnake Squamata; and gray box, none of the above. Species codes and GenBank Genome ID or segment accession number are as follows: Bo.jara, *Bothrops jararaca* (this study); An.caro, *Anolis carolinensis*, ID: 708; Cr.adam, *Crotalus adamanteus*, PLA2 scaffold KX211996; Cr.atro, *Crotalus atrox*, PLA2 scaffold KX211994; Cr.scut, *Crotalus scutulatus*, ADAM28 scaffold MT032003.1; Cr.viri, *Crotalus viridis*, ID: 71654; Ga.gall, *Gallus gallus*, ID: 111; Ge.japo, *Gekko japonicus*, ID: 40475; La.chal, *Latimeria chalumnae*, ID: 3262; Mu.musc, *Mus musculus*, ID: 52; Op.hann, *Ophiophagus hannah*, ID: 10842; Po.mura, *Podarcis muralis*, ID: 8765; Po.vitt, *Pogona vitticeps*, ID: 7589; Pr.mucr, *Protobothrops mucrosquamatus*, ID: 18192; Ps.text, *Pseudonaja textilis*, ID: 72610; and Py.bivi, *Python bivittatus*, ID: 17893. For *B. jararaca*, the scheme is based on the combination of data gathered from the sequences obtained by the different strategies used in this work. CTLs and LAAO were not included in the figure since the *B. jararaca* scaffolds did not provide enough information to define the architecture of their loci.

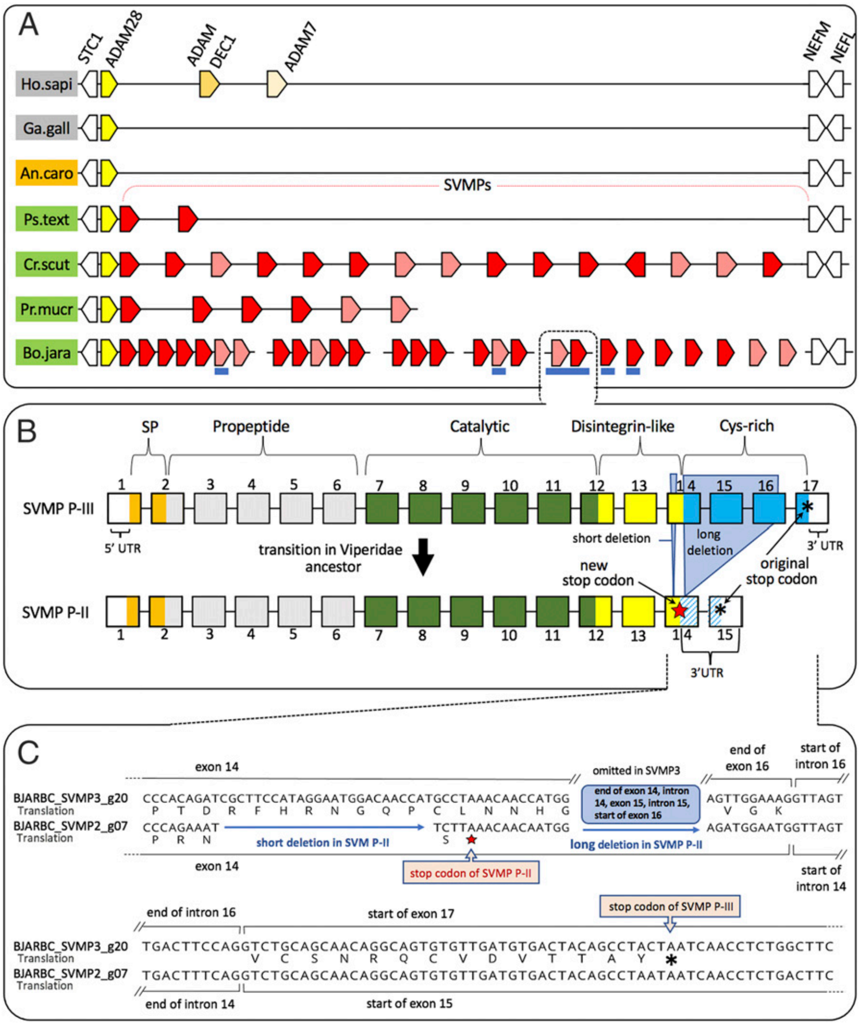

Fig. 3. The SVMP gene structure and arrangement at the locus. (*A*) Architecture of the ADAM28 genomic locus in different vertebrates (not in scale). The putative SVMP ancestral gene ADAM28 (yellow arrow) and flanking genes (STC1, NEFM, and NEFL: white arrows) form a syntenic block among vertebrates. Orange and beige arrows represent ADAM family genes in humans (ADAMDEC1 and ADAM7). Red and pink arrows represent genes from the SVMP classes P-III and P-II, respectively. Solid lines are contiguous sequences, and dotted lines indicate uncertain order or no contiguity. Blue bars represent regions covered by BAC. (*B*) Schematic alignment of SVMP gene structures showing the conservation of exons (squares) between SVMP P-III and P-II. A short and a long deletion at exon 14 of SVMP P-II are marked. These deletions result in the loss of the Cys-rich domain and the shortening of the disintegrin-like sequence through the acquisition of a new stop codon (red star) preceding the original one (black star). (*C*) Details of the nucleotide alignment with the encoded amino acid residues between the two neighboring SVMP genes belonging to the P-III and P-II classes (BJARBC_SVMP3_g20 and BJARBC_SVMP2_g07, respectively) in the region between exons 14 and 17.

results reinforce the idea that the evolution of SVMPs in Viperidae is intimately associated with intron/exon indels (47, 48) and provide a further explanation for the acquisition of the premature stop codon in P-II SVMPs.

SVSP Locus Shows Homology to Gilatoxins. The *B. jararaca* genomic scaffolds containing SVSP genes revealed that these genes are clustered and organized in tandem (Fig. 2). We observed that some of the SVSP genes are preceded by genes of the cytochrome *c* oxidase 6B1 (cox 6B1) subunit or its pseudogenes, and that these cox 6B1 pseudogenes are present in the *Protobothrops mucrosquamatus* SVSP genes (National Center for Biotechnology Information [NCBI] genomic sequence: NW_015386730). This pattern suggests that SVSP duplications in snakes may have involved a genomic segment comprising these two genes and/or that cox 6B1 may have facilitated the SVSP duplication process.

A syntenic locus of the SVSP genes located between the flanking SCAN-domain containing protein-1 and RBM42 genes could be identified across different squamates (Fig. 2). While the locus exhibits essentially the same SVSP expansion in the Viperidae species *P. mucrosquamatus* as in *B. jararaca*, in the Elapidae species *Pseudonaja textilis*, there is no expansion of SVSPs (a single SVSP gene exists, and it exhibits low expression in the VGs). More interestingly, in Toxicofera lizards with sequenced genomes, this locus contains a serine proteinase gene referred to as gilatoxin-like gene. Gilatoxin is a serine proteinase that has been demonstrated to be a major component of the venom from *Heloderma* sp. and other venomous Anguimorpha lizards (49, 50). The identification of a conserved genomic position for the venom serine proteinases of snakes and lizards suggests possible orthology between them. This favors a possible single origin of this specific toxin class in Toxicofera (50) or at least indicates a single ancestral gene source, which could have been recruited one or more times during the evolution of snakes and lizards. Interestingly, the VT2R26 vomeronasal receptor gene has expanded in *Anolis carolinensis* (Fig. 2), suggesting that this locus is prone to gene duplication via unknown mechanisms.

BPPs Were Added to an Ancestral CNP Precursor. The BPP/CNP precursor is one of the most abundant transcripts in *B. jararaca* VGs (*SI Appendix*, Table S1). The BPP/CNP gene, described here, is ~11 kb in length and has two introns, but only one intron interrupts the CDS region (Fig. 4). The coding region of the precursor is encoded by two exons, while a third region comprises most of the 3′ UTR sequence. The first exon contains the signal peptide and the region with the BPP repeats, whereas the second exon contains the spacer and the CNP.

The similarities between parts of the *B. jararaca* BPP/CNP precursor (51) and vertebrate CNP indicate that the first may have originated from the latter. However, the absence of BPPs and the presence of a C-terminal extension in the Elapidae natriuretic peptide precursor, resembling vertebrate brain natriuretic peptide, suggested a nonhomologous origin of these precursors in Viperidae and Elapidae (4). We subsequently found greater similarities of a CNP precursor from Dipsadidae with both Viperidae BPP/CNP and Elapidae natriuretic peptide precursors and hypothesized that all of these precursors were in fact orthologous and derived from a CNP gene (52).

The structure of the BPP/CNP gene identified in this study (Fig. 4) shows the same overall structure as the CNP gene of other vertebrates, and both genes occupy the same position in the locus (Fig. 2). Moreover, intron 1 is positioned just between the last BPP repeat and the beginning of the spacer, indicating that the divergent region (containing the BPPs) is restricted to the end of the first exon. BPP acquisition seems to have occurred as an extension of exon 1 and not via the insertion or shuffling of a new exon, since no such sequence has been identified in other genomes to our knowledge. By reviewing data from other available genomes, we could recognize the same extended first exon in the *P. mucrosquamatus* (Viperidae) BPP gene (Gene ID: 107296050). The corresponding exons in other squamates (*A. carolinensis* and *Gekko japonicus)* and in *Homo sapiens* encode only the signal peptide and a short prodomain. Therefore, BPPs indeed seem to have arisen over the CNP gene, apparently without any gene duplication, since we did not locate any paralog of the "endogenous" CNP gene in any of these genomes. Unfortunately, the lack of annotated genes for venom natriuretic peptide precursors in Elapidae and Dipsadidae prevents a more robust confirmation of the shared origin of venom CNP in these families with the Viperidae BPP/CNP.

VEGF-F Gene Locus Indicates a Non–VEGF-A Origin. Snake venom vascular endothelial growth factor (svVEGF or VEGF-F) (53) is part of the VEGF superfamily of growth factors, which also includes placental growth factors (PGFs). The VEGF-F gene from *Protobothrops* (former *Trimeresurus*) *flavoviridis* was first amplified and sequenced in 2009, as was the endophysiological VEGF-A gene from the same species (16). Although these two genes are very different in size, they supposedly show some conservation of short segments of intronic sequences; therefore, it was hypothesized that VEGF-F could have originated from a duplication of VEGF-A followed by accelerated evolution (16). However, the genomic context of these genes could not be observed in that study.

Here, we identified a single gene encoding VEGF-F in *B. jararaca* (BAC clone BJARBC_02H08Ma1). In this scaffold, the VEGF-F gene is flanked by the downstream genes PPP1R13L and ERCC2 (Fig. 5 *A*, *Left*). Looking for this set of genes in other species, we observed a similar organization in the *P. mucrosquamatus* genome, including an upstream RTN2+FOSB gene block. A similar set of genes was found in more distantly related species of snakes and in other vertebrates. In the snakes *P. textilis* and *Python bivittatus*, the lizard *Pogona vitticeps*, and the amphibian *Microcaecilia unicolor* the "growth factor" gene positioned at this locus is named after VEGF-F–like, probably due to the high similarity of the encoded protein to the snake toxins, whereas in the coelacanth *Latimeria chalumnae,* it is named after "PGF-like" (placental growth factor-like) (Fig. 5*A*). We note that PGF is also a member of the VEGF family of growth factors.

We also identified the *B. jararaca* VEGF-A gene and its flanking genes, based on which we retrieved VEGF-A loci from multiple species (Fig. 5 *A*, *Right*). As observed in Fig. 5*A*, the VEGF-A gene is placed in a completely different genomic context than VEGF-F. Therefore, there are two different loci, each of which is relatively conserved among the Chordata phylogeny, and they represent distinct genes encoding similar proteins belonging to the VEGF family. We performed a phylogenetic reconstruction of VEGF genes retrieved from both loci, as well as with other classes of VEGF from several vertebrates (Fig. 5*B* and *SI Appendix*, Fig. S4). There is a robust grouping of VEGF-F from Viperidae snakes nested within the VEGF-F–like/PGF-like clade of sequences from other Squamates and Chordata.

Given this scenario, we now suggest that the VEGF-F gene (the snake venom VEGF) is, in fact, an ortholog of the gene positioned at the same site in other organisms, sometimes annotated as "PGF-like," and is not a result of a recent (after snake appearance) duplication and neofunctionalization of a VEGF-A (or other VEGF-like) gene positioned elsewhere. The PGF-like gene has likely been positioned in that locus since the time of the ancestral vertebrates, as shown by the conservation of the gene block in *Latimeria*, at the base of Sarcopterygii. The early origin of the whole ortholog group composed of VEGF-F/VEGF-F–like and PGF-like cannot be deduced from our phylogenetic analysis, but the fact that it is not nested within the robustly supported VEGF-A clade indicates the common ancestor of them preceded the appearance of snakes.

Indeed, more detailed observation of the structures of VEGF-A and VEGF-F (Fig. 5*C*) revealed that the two paralogs are very different in size, exon number, and show very low identity in the sequence composition of their noncoding regions. The coding regions of the two paralogs show some similarity (39 to 42% within the same species, Fig. 5*C*), while the conservation within each ortholog is higher, even for distantly related Squamata (e.g., 47% for VEGF-F and 87% form VEGF-A between *B. jararaca* and *Pogona* lizard, Fig. 5*C*). Our hypothesis is that in an ancestral Viperidae, the protein derived from the PGF-like/VEGF-F–like gene was co-opted without gene duplication to be a component of the venom and later it underwent a process of functional specialization. The resulting toxin is similar to the well-characterized VEGF-A, as well as to any VEGF family member, thus providing the suggestion for its name when it was discovered (53).

General Inferences About Toxin Gene Recruitment. The genome of *B. jararaca* described here allowed us to identify syntenies between the gene arrays flanking the toxin genes and the respective genome segments in other organisms, thus enabling us to examine what kinds of genes are present in similar positions across venomous and nonvenomous species. This has permitted to infer if a related nontoxin gene was likely present at the locus in an ancestral snake and to check if this gene is present in extant snakes. The presence in an extant venomous snake of both the toxin gene and its related nontoxin paralog predicted to exist in the ancestral snake is suggestive of an ancestral duplication in the locus, whereas the presence of only the toxin gene at the position where the ancestral snake had a related nontoxin gene will indicate the absence of ancestral duplication in the locus. This locus structure-based approach represents an alternative way of tracking toxin gene origins that has been applied in some cases (2, 11, 17–19, 54, 55) and which is independent of the traditional method based on reconstructing the phylogeny of toxin sequences to infer the ortholog ancestral.

With respect to the whole set of 12 toxin families present in the Viperidae *B. jararaca* (Fig. 2), we identified the two scenarios above for the recruitment of ancestral nontoxin genes. They correspond to the mechanisms pointed out by Vonk et al. (2) based

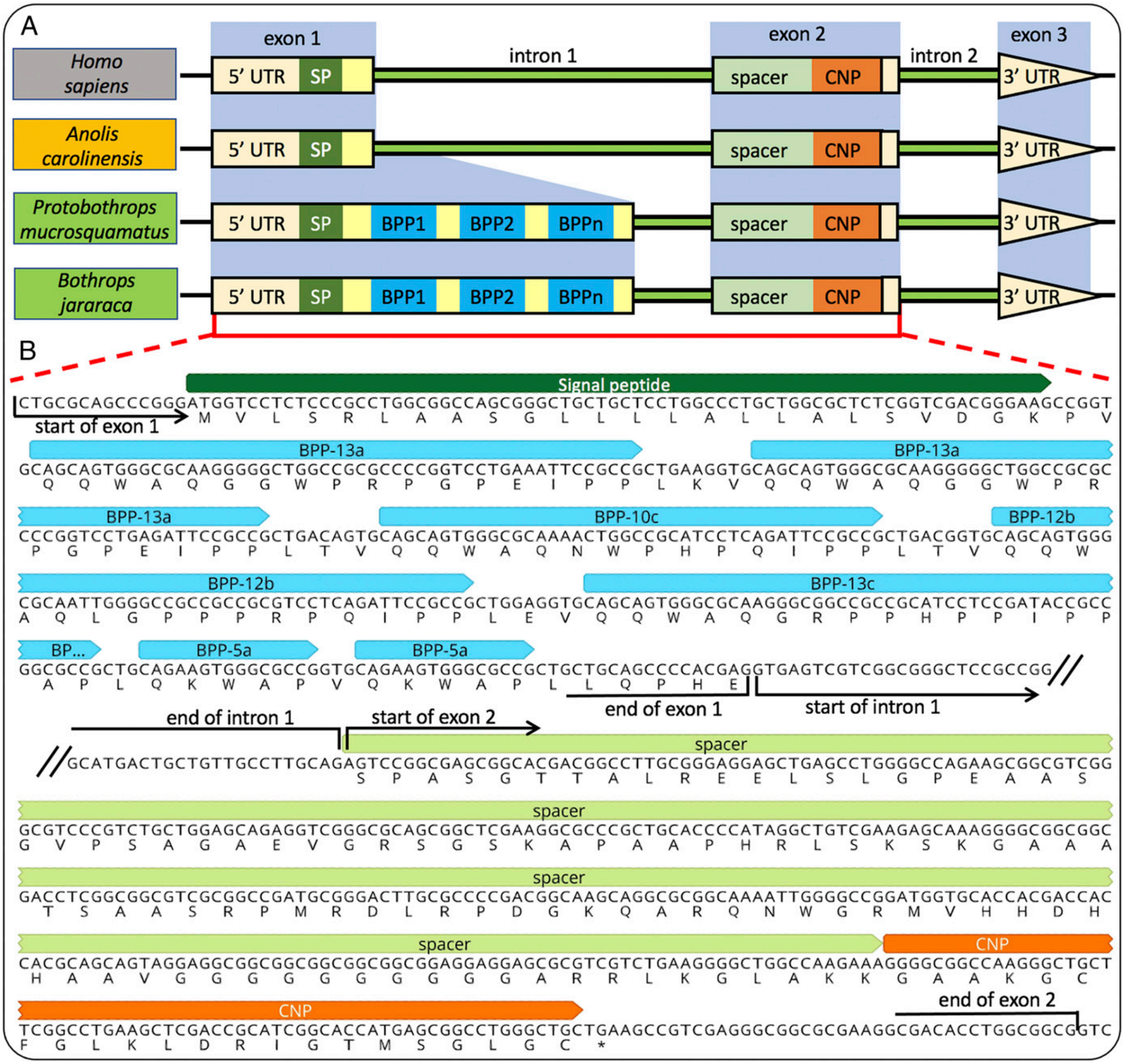
A
exon 1
intron 1
exon 2
intron 2
exon 3
Homo sapiens
Anolis carolinensis
Protobothrops mucrosquamatus
Bothrops jararaca
5′ UTR
SP
BPP1
BPP2
BPPn
spacer
CNP
3′ UTR
B
Signal peptide
CTGCGCAGCCCGGGATGGTCCTCTCCCGCCTGGCGGCCAGCGGGCTGCTGCTCCTGGCCCTGCTGGCGCTCTCGGTCGACGGGAAGCCGGT
M V L S R L A A S G L L L L A L L A L S V D G K P V
start of exon 1
BPP-13a
BPP-13a
GCAGCAGTGGGCGCAAGGGGGCTGGCCGCGCCCCGGTCCTGAAATTCCGCCGCTGAAGGTGCAGCAGTGGGCGCAAGGGGGCTGGCCGCGC
Q Q W A Q G G W P R P G P E I P P L K V Q Q W A Q G G W P R
BPP-13a
BPP-10c
BPP-12b
CCCGGTCCTGAGATTCCGCCGCTGACAGTGCAGCAGTGGGCGCAAAACTGGCCGCATCCTCAGATTCCGCCGCTGACGGTGCAGCAGTGGG
P G P E I P P L T V Q Q W A Q N W P H P Q I P P L T V Q Q W
BPP-12b
BPP-13c
CGCAATTGGGGCCGCCGCCGCGTCCTCAGATTCCGCCGCTGGAGGTGCAGCAGTGGGCGCAAGGGCGGCCGCCGCATCCTCCGATACCGCC
A Q L G P P P R P Q I P P L E V Q Q W A Q G R P P H P P I P P
BP...
BPP-5a
BPP-5a
GGCGCCGCTGCAGAAGTGGGCGCCGGTGCAGAAGTGGGCGCCGCTGCTGCAGCCCCACGAGGTGAGTCGTCGGCGGGCTCCGCCGG//
A P L Q K W A P V Q K W A P L L Q P H E
end of exon 1
start of intron 1
end of intron 1
start of exon 2
spacer
//GCATGACTGCTGTTGCCTTGCAGAGTCCGGCGAGCGGCACGACGGCCTTGCGGGAGGAGCTGAGCCTGGGGCCAGAAGCGGCGTCGG
S P A S G T T A L R E E L S L G P E A A S
spacer
GCGTCCCGTCTGCTGGAGCAGAGGTCGGGCGCAGCGGCTCGAAGGCGCCCGCTGCACCCCATAGGCTGTCGAAGAGCAAAGGGGCGGCGGC
G V P S A G A E V G R S G S K A P A A P H R L S K S K G A A A
spacer
GACCTCGGCGGCGTCGCGGCCGATGCGGGACTTGCGCCCCGACGGCAAGCAGGCGCGGCAAAATTGGGGCCGGATGGTGCACCACGACCAC
T S A A S R P M R D L R P D G K Q A R Q N W G R M V H H D H
spacer
CNP
CACGCAGCAGTAGGAGGCGGCGGCGGCGGCGGCGGAGGAGGAGCGCGTCGTCTGAAGGGGCTGGCCAAGAAAGGGGCGGCCAAGGGCTGCT
H A A V G G G G G G G G G G A R R L K G L A K K G A A K G C
CNP
end of exon 2
TCGGCCTGAAGCTCGACCGCATCGGCACCATGAGCGGCCTGGGCTGCTGAAGCCGTCGAGGGCGGCGCGAAGGCGACACCTGGCGGCGGTC
F G L K L D R I G T M S G L G C *

Fig. 4. BPP/CNP gene structure. (*A*) Schematic alignment of BPP/CNP and CNP genes from different organisms emphasizing the correspondence of introns and exons, the conservation of domain structures, and the extension of exon 1 harboring the BPPs in Viperidae. Species are classified according to the following color code: green box, venomous snake; orange box, nonsnake Squamata; and gray box: none of the above. (*B*) Part of the BPP/CNP gene sequence from *B. jararaca* and its translation, showing that BPPs are restricted to exon 1.

on the genomic organization of three toxin families from the Elapidae *Ophiophagous hannah*, referred to as "duplication of nontoxin genes" and "gene hijacking/modification." These mechanisms relate to more general concepts in gene evolution, referred to as, respectively, "neofunctionalization model" (lato senso) and "gene co-option without duplication" or "moonlighting" (56–60), which have been considered for explaining the recruitment of snake toxins, for example, refs. 2, 3, 6, 11, 54, 61, 62. Our analyses indicate which mechanism likely occurred in nine out of the 12 toxin families present in *B. jararaca* and highlight that these mechanisms occurred mostly locally in the genome of ancestral venomous snake.

Under the duplication of nontoxin gene mechanism, genes have been recruited after the duplication of an ancestral gene existing within the locus, without the direct co-option of the original gene, to become a toxin. This is the case for highly abundant toxins such as SVMP and PLA2, for which a closely related nonvenom paralog is still present flanking the venom genes. In these cases, secondary rounds of copy expansion may have followed the initial duplication, allowing the neofunctionalization of specific toxin paralogs. SVMP is a clear example of this, since its multigene cluster starts just 3′ to the ADAM28 gene (Fig. 3). Likewise, in the case of PLA2, a non-VG–expressed PLA2 IIGc is present within the same locus (Fig. 2), although Jackson and Koludarov (54) considered a potential co-option of PLA2 to venom prior to duplication based on the very low expression of this gene in the VGs of some Crotalinae. Under the gene hijacking/modification mechanism, genes have been recruited by the direct specialization of ancestral genes with nonvenom functions, preexisting within the locus, into

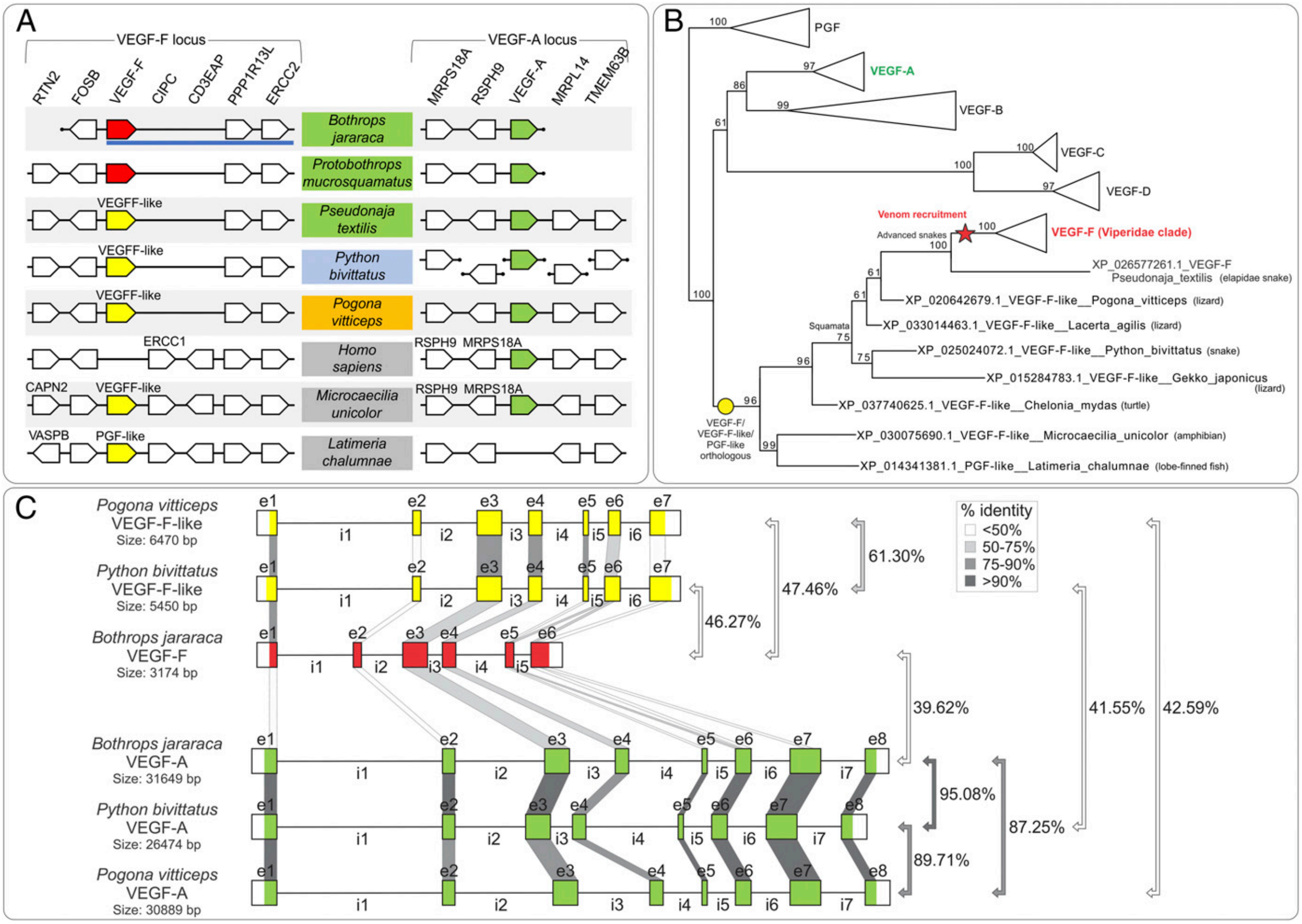

Fig. 5. (*A*) Architecture of the venom VEGF-F gene and nonvenom VEGF-A gene loci in synteny among different organisms. Red pentagon, VEGF-F toxin gene; yellow pentagon, PGF-like or VEGF-F–like genes; green pentagon, VEGF-A gene; and the white pentagon represents adjacent nonrelated genes. Dots at the end of solid lines indicate scaffold ends. Blue bar represents region covered by BAC. Gene representations are not to scale. Species were classified according to the following color code: green box, venomous snake; blue box, nonvenomous snake; orange box, nonsnake Squamata; and gray box, none of the above. (*B*) Summarized phylogenetic tree of the VEGF family of growth factor focusing the origin of VEGF-F (snake venom VEGFs) from the PGF-like/VEGF-F–like ortholog. The complete phylogenetic analysis is shown in *SI Appendix*, Fig. S4. (*C*) Schematic comparison of VEGF-F and VEGF-A genes in three Squamata, pointing out the levels of conservation throughout these genes. Percentage values on the right represent pairwise identity of CDS regions.

toxin genes. This is the case for SVSP, BPP/CNP, and the less abundant ancillary toxins HYAL, NGF, VEGF-F, PLB, and NUCL, whose genes are located at the same position putatively occupied by their ancestral genes and without a nearby closely related nonvenom paralog in the venomous species. VEGF-F is a good example of this process, as we demonstrated that it is placed neither near the VEGF-A gene nor at a random site, but it occupies the genomic position of a preexisting member of the VEGF family (Fig. 5). Our results are in agreement with what has been proposed based on the genomic context for the recruitment of SVMP (2, 11, 20, 21) and PLA2 (17) by ancestral gene duplication and for HYAL and PLB (2) by hijacking/modification, and we provide further support for the gene hijacking/modification mechanism by associating five other families (SVSP, NGF, VEGF-F, BPP/CNP, and NUCL) with this mechanism.

Interestingly, ancestral gene co-options occurring in the locus seem to explain the emergence of most toxin classes, according to our analysis. This is similar to what was observed for the origin of venom genes in parasitoid wasp (55), in which a minor part of the venom genes showed evidence of ancestral gene duplications whereas a greater number likely derived from single-copy ancestral genes. Nevertheless, we should consider that in snakes, most of these directly co-opted genes are minor venom components (except for SVSP and BPP/CNP) and do not represent the most prevalent toxins in Viperidae venoms (58). The ancestral gene duplication, generally considered to be the primary mechanism of toxin recruitment, continues to be supported in our analysis as underlying the recruitment of the highly abundant toxins (PLA2 and SVMP) in these venoms, which are likely the most relevant for venom function.

An intriguing question is how the snake deals with the loss of the functional products of the nonvenom genes directly co-opted to the venom arsenal. A clue to answering this question could be the fact that the majority of toxins arising from ancestral gene co-option belong to families with preexisting paralogs (clustered or spread in the genome) that may assume physiological functions once a member is diverted to venom. However, it seems likely that these preexisting paralogs will not produce proteins with exactly the same function and so any sort of loss will likely result in a shift in physiological function. Another possibility is that the gene products recruited had dual function (both endogenous and toxic), at least at the early stages of their recruitment. The dynamic nature of the toxin recruitment process, balancing gene products toward venom or nonvenom function, was already shown by the phylogenetic positioning of nonvenom proteins within toxin clades

(63) and from the identification of basal levels of toxin expression in nonvenom organs of *B. jararaca* (7). This is suggestive that these animals may have tolerated some toxicity level in their bodies, favoring the possibility of a dual function of co-opted genes, although a significant presence of circulating toxins was not demonstrated in snakes. The exact mechanism allowing the start of the co-option process to venom, especially in the absence of gene duplication, is yet to be elucidated and it certainly has many nuances for each toxin family but ends either with an increase in gene expression in the VGs or a process of gradually restricting the expression to this tissue (64).

For the cases of SVMP and PLA2, however, early duplication events may have facilitated an escape from adaptive conflict (65), in which an ancestral gene is constrained to not specialize in an intrinsic secondary function due to the selective pressure on its primary function. A classic example is the vertebrate δ-crystalline eye protein, which is presumed to have been encoded by a single gene for arginosuccinate lyase in the common archosaur ancestor, then underwent an ancestral duplication followed by the loss of enzymatic activity and subsequent specialization to produce δ1, a crystalline structural protein. However, in the same eye system, α- and βγ-crystallines, thought to be derived from preexisting multiparalogue genes, are believed to have been recruited via a nonduplicative process from a chaperone gene (59). In fact, well-documented cases of gene neofunctionalization before or after ancient duplication are rare (66–68), and the venom genes addressed here in their orthologous context represent a system in which to explore such events in greater detail.

Independent of the initial recruitment process, the most abundant toxin families in Viperidae (SVMP, SVSP, PLA2, and CTL) are those that underwent more expansion at their loci, indicating that most relevant venom pathological effects are more closely associated with secondary expansion of relevant genes into multiple paralogs than with the type of initial recruitment. Since paralogs within each family are not exact copies but divergent genes known to be under accelerated evolution (12), it does not seem likely that pressure driving the accumulation of high levels of proteins in venom was the selective pressure underlying the expansion of these genes. A functional pressure driving the availability of diversified important gene products is more likely, perhaps for the fine-tuning of receptor interactions and prey specificity in different environments.

In conclusion, when the genomic landscape of toxin genes and venom loci in *B. jararaca* is considered in a comparative context with related organisms it demonstrates that the Viperidae venom arsenal was assembled from locally existing elements. More broadly, it illustrates how important it is to consider the genomic background from which innovation arises. We predict that additional venomous snake genomes will be critical for evaluating the generality of the mechanisms proposed for the evolution of this iconic example of a molecular adaptation.

Materials and Methods

Specimen Sampling. An adult female individual of *B. jararaca* from Embu das Artes, São Paulo State, Brazil was used as a source of DNA, RNA, and venom. We followed protocol 1131/13 approved by the Committee of Ethics on the Use of Animals of the Butantan Institute. The specimen is registered in the Herpetological Collection of the Butantan Institute (IBSP84406).

Genome Sequencing and Toxin Gene Locus Identification. The *B. jararaca* genome was assembled de novo using a hybrid approach (HA-WGS), utilizing both Pacific Biosciences long reads at 25× coverage (read N50 9,474 bp) and Illumina 100 bp paired-end data at around 60× coverage, employing MaSuRCA version 3.2.8 (69). The genome size of *B. jararaca* was determined using Jellyfish 2.2.3 (70) and the GenomeScope tool (71). The detailed sequencing and the complementary assemble strategy based on short reads (SA-WGS) are described in *SI Appendix*, *Supplementary Methodology*.

BAC-SeqSc was uniquely designed for this work (Fig. 1, *Right*), detailed in *SI Appendix*, *Supplementary Methodology*. It was based on the sequencing of pools of 12 BACs containing long genomic segments (150 to 250 kbp), screening them for the presence of toxin genes, and resequencing the selected BACs with high coverage.

All scaffolds generated via HA-WGS, SA-WGS, and BAC-SeqSc strategies were screened for segments matching toxin fragments through BLASTn searches using toxin sequences obtained from the de novo transcriptome as well as other *B. jararaca* toxin sequences available in GenBank as queries. Long reads were also screened to identify missing genes, providing additional data to manually link scaffolds and solve gene structures. The scaffolds containing toxin genes were manually annotated with CLC Genomics Workbench version 9 to 11 or Geneious version 10. UTRs and CDSs were annotated for each gene whenever possible, and automatically predicted exon/intron boundaries were manually checked for consistency following the AG-TC rule.

RNA sequencing (RNA-Seq). Total RNA from the VG, gut, kidney, stomach, lung, heart, and brain was extracted with TRIzol (Thermo), and polyA+ RNAs were obtained via magnetic bead purification (DYNAbeads, Life Technologies). The mRNA concentration was estimated with a Quant-iT RiboGreen Kit (Invitrogen). Sequencing libraries were constructed using the TruSeq RNA Sample Prep Kit version 2 and sequenced on HiSeq1500 equipment (Illumina). We used the software Tophat2 (72) and bowtie2 (73) for genome read mapping. Trinity software (42) was used for de novo analysis and guided by the draft genome assembly. Detailed procedures of the RNA-seq analysis are provided in *SI Appendix*, *Supplementary Methodology*.

Venom Proteome Analysis and Toxin Identification. The venom from the specimen described above was analyzed by protein tandem mass spectrometry using a bottom-up shotgun approach. Fresh venom was obtained through milking before VG extraction. Detailed procedures of the proteomic analysis are provided in *SI Appendix*, *Supplementary Methodology*.

Data Availability. Sequences data have been deposited in GenBank (PRJNA691605) (74) and a genome browsing tool is available at:http://cetics.butantan.gov.br/gb2/gbrowse/bothrops_jararaca/ (75).

ACKNOWLEDGMENTS. This work was supported by the São Paulo Research Foundation (Grants 2013/07467-1, 2012/00177-5, and 2016/50127-5 to I.L.M.J.-d.-A.; 2013/07964-0 to D.D.A.; 2015/03509-7 to V.L.V.; and 2018/26520-4 to P.G.N.), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (Grant 304532/20134 to I.L.M.J.-d.-A.), and NSF Grant Division of Environmental Biology 1638872 to H.L.G. We thank Dr. Jesus Ferro, Prof. Dr. Anete Pereira de Souza, and Dr. Danilo Augusto Sforça for helping with BAC library preparation, Dr. João Carlos Setubal and George Willian Condomitti, MSc for assistance with the initial bioinformatic analysis, and the technicians of the Next Generation Sequencing core at the Instituto Butantan.

1. H. W. Greene, M. Fogden, P. Fogden, *Snakes: The Evolution of Mystery in Nature* (University of California Press, 1997).
2. F. J. Vonk *et al.*, The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20651–20656 (2013).
3. N. R. Casewell, W. Wüster, F. J. Vonk, R. A. Harrison, B. G. Fry, Complex cocktails: The evolutionary novelty of venoms. *Trends Ecol. Evol.* **28**, 219–229 (2013).
4. B. G. Fry, From genome to "venome": Molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. *Genome Res.* **15**, 403–420 (2005).
5. N. R. Casewell, On the ancestral recruitment of metalloproteinases into the venom of snakes. *Toxicon* **60**, 449–454 (2012).
6. J. Reyes-Velasco *et al.*, Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. *Mol. Biol. Evol.* **32**, 173–183 (2015).
7. I. L. M. Junqueira-de-Azevedo *et al.*, Venom-related transcripts from Bothrops jararaca tissues provide novel molecular insights into the production and evolution of snake venom. *Mol. Biol. Evol.* **32**, 754–766 (2015).
8. W. Yin *et al.*, Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nat. Commun.* **7**, 13107 (2016).
9. T. A. Castoe *et al.*, The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20645–20650 (2013).Corrected in: *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3194 (2014).
10. S. D. Aird *et al.*, Population genomic analysis of a pitviper reveals microevolutionary forces underlying venom chemistry. *Genome Biol. Evol.* **9**, 2640–2649 (2017).
11. K. Suryamohan *et al.*, The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nat. Genet.* **52**, 106–117 (2020).
12. H. Shibata *et al.*, The habu genome reveals accelerated evolution of venom protein genes. *Sci. Rep.* **8**, 11300 (2018).

13. D. R. Schield *et al*., The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res.* **29**, 590–601 (2019).
14. M. J. Margres *et al*., The Tiger Rattlesnake genome reveals a complex genotype underlying a simple venom phenotype. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2014634118 (2021).
15. N. Itoh *et al*., Organization of the gene for batroxobin, a thrombin-like snake venom enzyme. Homology with the trypsin/kallikrein gene family. *J. Biol. Chem.* **263**, 7628–7631 (1988).
16. Y. Yamazaki *et al*., Snake venom Vascular Endothelial Growth Factors (VEGF-Fs) exclusively vary their structures and functions among species. *J. Biol. Chem.* **284**, 9885–9891 (2009).
17. N. L. Dowell *et al*., The deep origin and recent loss of venom toxin genes in rattlesnakes. *Curr. Biol.* **26**, 2434–2445 (2016).
18. L. Sanz, R. A. Harrison, J. J. Calvete, First draft of the genomic organization of a PIII-SVMP gene. *Toxicon* **60**, 455–469 (2012).
19. K. Yamaguchi *et al*., The finding of a group IIE phospholipase A2 gene in a specified segment of Protobothrops flavoviridis genome and its possible evolutionary relationship to group IIA phospholipase A2 genes. *Toxins (Basel)* **6**, 3471–3487 (2014).
20. N. L. Dowell *et al*., Extremely divergent haplotypes in two toxin gene complexes encode alternative venom types within rattlesnake species. *Curr. Biol.* **28**, 1016–1026.e4 (2018).
21. M. W. Giorgianni *et al*., The origin and diversification of a novel protein family in venomous snakes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10911–10920 (2020).
22. H. M. I. Kerkkamp *et al*., Snake genome sequencing: Results and future prospects. *Toxins (Basel)* **8**, 360 (2016).
23. R. H. Valente *et al*., Bothrops jararaca accessory venom gland is an ancillary source of toxins to the snake. *J. Proteomics* **177**, 137–147 (2018).
24. S. M. T. Serrano, A. K. Oliveira, M. C. Menezes, A. Zelanis, The proteinase-rich proteome of Bothrops jararaca venom. *Toxin Rev.* **33**, 169–184 (2014).
25. C. Augusto-de-Oliveira *et al*., Dynamic rearrangement in snake venom gland proteome: Insights into Bothrops jararaca intraspecific venom variation. *J. Proteome Res.* **15**, 3752–3762 (2016).
26. Y. Fujimura *et al*., Isolation and chemical characterization of two structurally and functionally distinct forms of botrocetin, the platelet coagglutinin isolated from the venom of Bothrops jararaca. *Biochemistry* **30**, 1957–1964 (1991).
27. R. B. Zingali, M. Jandrot-Perrus, M. C. Guillin, C. Bon, Bothrojaracin, a new thrombin inhibitor isolated from Bothrops jararaca venom: Characterization and mechanism of thrombin inhibition. *Biochemistry* **32**, 10794–10802 (1993).
28. A. L. J. Coelho *et al*., Effects of jarastatin, a novel snake venom disintegrin, on neutrophil migration and actin cytoskeleton dynamics. *Exp. Cell Res.* **251**, 379–387 (1999).
29. S. M. T. Serrano *et al*., A novel phospholipase A2, BJ-PLA2, from the venom of the snake Bothrops jararaca: Purification, primary structure analysis, and its characterization as a platelet-aggregation-inhibiting factor. *Arch. Biochem. Biophys.* **367**, 26–32 (1999).
30. D. Ianzer *et al*., Identification of five new bradykinin potentiating peptides (BPPs) from Bothrops jararaca crude venom by using electrospray ionization tandem mass spectrometry after a two-step liquid chromatography. *Peptides* **25**, 1085–1092 (2004).
31. M. J. I. Paine, H. P. Desmond, R. D. G. Theakston, J. M. Crampton, Purification, cloning, and molecular characterization of a high molecular weight hemorrhagic metalloprotease, jararhagin, from Bothrops jararaca venom. Insights into the disintegrin gene family. *J. Biol. Chem.* **267**, 22869–22876 (1992).
32. H. W. Raudonat, M. Rocha e Silva, *Separation of the Bradykinin Releasing Enzyme from the Clotting Factor in Venom from Bothrops Jararaca* (Naunyn-Schmiedeberg's Arch. für Exp. Pathol. und Pharmakologie, 1962).
33. S. H. Ferreira, D. C. Bartelt, L. J. Greene, Isolation of bradykinin-potentiating peptides from Bothrops jararaca venom. *Biochemistry* **9**, 2583–2593 (1970).
34. D. A. P. Cidade *et al*., Bothrops jararaca venom gland transcriptome: Analysis of the gene expression pattern. *Toxicon* **48**, 437–461 (2006).
35. A. Zelanis *et al*., Analysis of the ontogenetic variation in the venom proteome/peptidome of Bothrops jararaca reveals different strategies to deal with prey. *J. Proteome Res.* **9**, 2278–2291 (2010).
36. G. S. Dias *et al*., Individual variability in the venom proteome of juvenile Bothrops jararaca specimens. *J. Proteome Res.* **12**, 4585–4598 (2013).
37. L. Gonçalves-Machado *et al*., Combined venomics, venom gland transcriptomics, bioactivities, and antivenomics of two Bothrops jararaca populations from geographic isolated regions within the Brazilian Atlantic rainforest. *J. Proteomics* **135**, 73–89 (2016).
38. T. A. Castoe *et al*., Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biol. Evol.* **3**, 641–653 (2011).
39. G. I. M. Pasquesi *et al*., Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 2774 (2018).
40. N. B. Atkin, G. Mattinson, W. Beçak, S. Ohno, The comparative DNA content of 19 species of placental mammals, reptiles and birds. *Chromosoma* **17**, 1–10 (1965).
41. 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**, 3210–3212 (2015).
42. M. G. Grabherr *et al*., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
43. D. R. Amazonas *et al*., Molecular mechanisms underlying intraspecific variation in snake venom. *J. Proteomics* **181**, 60–72 (2018).
44. A. Zelanis *et al*., A transcriptomic view of the proteome variability of newborn and adult Bothrops jararaca snake venoms. *PLoS Negl. Trop. Dis.* **6**, e1554 (2012).
45. I. Koludarov *et al*., Reconstructing the evolutionary history of a functionally diverse gene family reveals complexity at the genetic origins of novelty *bioRxiv* [Preprint] (2020). https://doi.org/10.1101/583344. Accessed 10 February 2021.
46. N. R. Casewell, S. C. Wagstaff, R. A. Harrison, C. Renjifo, W. Wüster, Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Mol. Biol. Evol.* **28**, 2637–2649 (2011).
47. L. Sanz, J. J. Calvete, Insights into the evolution of a snake venom multi-gene family from the genomic organization of Echis ocellatus SVMP genes. *Toxins (Basel)* **8**, E216 (2016).
48. R. M. Kini, Accelerated evolution of toxin genes: Exonization and intronization in snake venom disintegrin/metalloprotease genes. *Toxicon* **148**, 16–25 (2018).
49. P. Utaisincharoen, S. P. Mackessy, R. A. Miller, A. T. Tu, Complete primary structure and biochemical properties of gilatoxin, a serine protease with kallikrein-like and angiotensin-degrading activities. *J. Biol. Chem.* **268**, 21975–21983 (1993).
50. B. G. Fry *et al*., Early evolution of the venom system in lizards and snakes. *Nature* **439**, 584–588 (2006).
51. N. Murayama *et al*., Cloning and sequence analysis of a Bothrops jararaca cDNA encoding a precursor of seven bradykinin-potentiating peptides and a C-type natriuretic peptide. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1189–1193 (1997).
52. A. T. C. Ching *et al*., Some aspects of the venom proteome of the Colubridae snake Philodryas olfersii revealed from a Duvernoy's (venom) gland transcriptome. *FEBS Lett.* **580**, 4417–4422 (2006).
53. I. L. Junqueira-de-Azevedo, S. H. Farsky, M. L. Oliveira, P. L. Ho, Molecular cloning and expression of a functional snake venom vascular endothelium growth factor (VEGF) from the Bothrops insularis Pit Viper. *J. Biol. Chem.* **276**, 39836–39842 (2001).
54. T. N. W. Jackson, I. Koludarov, How the toxin got its toxicity. *Front. Pharmacol.* **11**, 574925 (2020).
55. E. O. Martinson, Y. D. Mrinalini, Y. D. Kelkar, C. H. Chang, J. H. Werren, The evolution of venom by co-option of single-copy genes. *Curr. Biol.* **27**, 2007–2013.e8 (2017).
56. S. Ohno, *Evolution by Gene Duplication* (Springer Berlin Heidelberg, 1970).
57. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
58. T. Tasoulis, G. K. Isbister, A review and database of snake venom proteomes. *Toxins (Basel)* **9**, E290 (2017).
59. J. R. True, S. B. Carroll, Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **18**, 53–80 (2002).
60. S. D. Copley, An evolutionary perspective on protein moonlighting. *Biochem. Soc. Trans.* **42**, 1684–1691 (2014).
61. A. D. Hargreaves, M. T. Swain, D. W. Logan, J. F. Mulley, Testing the Toxicofera: Comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon* **92**, 140–156 (2014).
62. J. D. Bayona-Serrano *et al*., Replacement and parallel simplification of nonhomologous proteinases maintain venom phenotypes in rear-fanged snakes. *Mol. Biol. Evol.* **37**, 3563–3575 (2020).
63. N. R. Casewell, G. A. Huttley, W. Wüster, Dynamic evolution of venom proteins in squamate reptiles. *Nat. Commun.* **3**, 1066 (2012).
64. A. D. Hargreaves, M. T. Swain, M. J. Hegarty, D. W. Logan, J. F. Mulley, Restriction and recruitment-gene duplication and the origin and evolution of snake venom toxins. *Genome Biol. Evol.* **6**, 2088–2095 (2014).
65. A. L. Hughes, The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256**, 119–124 (1994).
66. C. T. Hittinger, S. B. Carroll, Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677–681 (2007).
67. D. L. Des Marais, M. D. Rausher, Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**, 762–765 (2008).
68. C. Deng, C.-H. C. Cheng, H. Ye, X. He, L. Chen, Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21593–21598 (2010).
69. A. V. Zimin *et al*., The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
70. G. Marcais, C. Kingsford, *Jellyfish : A Fast K-Mer Counter* (Tutorialis e Manuais, 2012).
71. G. W. Vurture *et al*., GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
72. D. Kim *et al*., TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
73. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
74. I. L. M. Junqueira-de-Azevedo, Genome sequencing and assembly, raw sequence reads, transcriptome or gene expression data deposited in GenBank database under Bioproject accession number PRJNA691605 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA691605. Accessed 12 January 2021.
75. I. L. M. Junqueira-de-Azevedo, Genome Browse for Bothrops jararaca genome http://cetics.butantan.gov.br/gb2/gbrowse/bothrops_jararaca/. Accessed 12 January 2021.

PNAS
www.pnas.org

Supplementary Information for

Tracking the recruitment and evolution of snake toxins using the evolutionary context provided by the *Bothrops jararaca* genome

Diego Dantas Almeida[1#], Vincent Louis Viala[1#], Pedro Gabriel Nachtigall[1], Michael Broe[2], H. Lisle Gibbs[2], Solange Maria de Toledo Serrano[1], Ana Maria Moura da Silva[3], Paulo Lee Ho[4], Milton Yutaka Nishiyama-Jr[1], Inácio L. M. Junqueira de Azevedo[1*]

*Corresponding Author: Inácio Junqueira de Azevedo
inacio.azevedo@butantan.gov.br

This PDF file includes:

Supplementary Methodology
Figures S1 to S4
Tables S1 to S2
Legends for Dataset S1
SI References

Other supplementary materials for this manuscript include the following:

Dataset S1

Supplementary Methodology

Specimen sampling

An adult female individual of *B. jararaca* was used as source of DNA, RNA and venom. The specimen was received by Instituto Butantan from Embu das Artes, São Paulo State, Brazil – Geographic coordinates: -23.654919, -46.864827. We followed the protocol 1131/13 approved by the Committee of Ethics in the Use of Animals from Butantan Institute. The specimen was registered in the Herpetological Collection of the Butantan Institute (IBSP84406). Prior to tissue collection, the venom was extracted, immediately freeze-dried and stored at -80ºC until use for the proteomics analysis. The blood was collected on blood collection tubes and immediately used for DNA extraction as described below. We dissected the venom glands, brain, lungs, stomach, kidney, gut and heart, and immediately frozen and stored at -80ºC until use for the for RNAseq analysis.

Short reads sequencing

We performed DNA extraction from blood using the DNeasy Blood and Tissue kit (Qiagen). One shotgun library was constructed using TruSeq DNA LT Sample Prep Kit (Illumina), following instructions of the manufacturer, with insert sizes about 500 bp long. In addition, five mate-pair libraries (2 kb, 5 kb, 10 kb, 15 kb and a range of 2-15 kb) were constructed using Nextera Mate Pair Library kit (Illumina). Paired-end sequencing was performed on a HiSeq1500 equipment set to read 2 x 100bp (Illumina).

Using Illumina Casava software (v1.8.2), with Illumina quality control QC > Q30, two paired-end fastq files were generated. The raw data reads were trimmed and filtered for PhiX, *Picchia pastoris* and *Escherichia coli* contaminants, using the software bowtie2 version 2.2.3 (1) and by quality, read size (> 40 bp), homopolymer (>90%), low complexity sequences (> 90%) and poly-A/T/N tails and adapters, using the software fastq-mcf version 1.04.662 (2). The long-insert raw mate-pair reads (2 kb, 5 kb, 10 kb, 15 kb and a range of 2-15 kb) were processed and filtered with NextClip tool (v0.8) (Leggett, R.M., et. al., 20014). The sequencing quality of pre-processed reads was checked with FastQC (v0.11.4).

Long reads sequencing

DNA samples from blood and liver of the same individual were extracted by proteinase K digestion from agarose plugs. DNA was quantified by fluorimetry using Quant-iT PicoGreen dsDNA Assay Kit (Thermo) and aliquots of 4 ug of DNA were sent (IBAMA export license No 15BR017067/DF) to Genome Sequencing Shared Resource at Duke University, Durham NC, USA for sequencing service. Three libraries were prepared by selecting fragments above 15kb and sequenced in 60 SMRT cell using P6-C4 chemistry in PACBIO RS II (Pacific Biosciences).

Genome assembly

Before the genome assembly, we estimated the genome size by *k*-mer analyses of the 100bp Illumina pair end reads. The genome size was calculated using the formula: $G = K_{num}/K_{depth}$ (3) where K_{num} is the total counts of *k*-mer and K_{depth} is the *k*-mer depth. We generated a *k*-mer profile with Jellyfish (v2.2.6) (4), which calculates the *k*-mer number and distribution. We then used two different models to generate estimates of genome size. The first method assumes a Poisson distribution for the *k*-mers. When multiple peaks are observed, the peak with lower *k*-mer frequencies is considered as the result of heterozygosity. The second method, which is integrated into the program GenomeScope, uses a mixed negative binomial model, granting more flexibility in genome size estimation (5).

The genome was assembled de novo using a hybrid approach, referred as HA-WGS (hybrid assemble whole genome shotgun) utilizing both Pacific Biosciences long-reads (read N50 9,474 bp for 20X genome coverage) together with Illumina short read and mate-pair data (2 X 100bp for 150X genome coverage), employing MaSuRCA v.3.2.8 (6). For Illumina data, mean fragment length and standard deviation were estimated by mapping a lane of data to a previously assembled *Sistrurus catenatus* genome using bowtie2, filtering reads for pairs properly mapping concordantly one time, and submitting the results to Picard tools CollectInsertSizeMetrics. The MaSuRCA algorithm operates by first creating high fidelity ‘supereads’ from the raw Illumina data, which reduces coverage (typically from 100X to 2–4X), and then using these to map to and tile long-reads, effectively error correcting them in the process. Given the high coverage of our Illumina data, the resulting scaffolds/contigs are expected to be highly accurate. MaSuRCA 3.2.8 produces scaffolds and contigs of identical length. The sequence total was 1666.74 MB, N50 163.55 KB, maximum length 1.85 MB, and this version was deposited in GenBank, as the reference genome for this work.

Since not all expected toxin genes were found in this version, we also performed independent assemblies using only subsets of the high quality short reads (paired-end reads combined with mate-pair reads), referred as SA-WGS (short reads assemble whole genome shotgun).For these assemblies AllPaths-LG assembler version 52488 (7) was used with following parameters: TARGETS=full_eval, THREADS=68, MAX_MEMORY_GB=800, MIN_CONTIG=500, HAPLOIDIFY=True, VAPI_WARN_ONLY=True). The scaffolds and contigs generated were subsequently oriented into larger supercontigs (scaffolds) using SSPACE (8) with the parameters “-k 5 -v 1 -z 100 -a 0.7 -x 0 -m 35 -o 20”. SSPACE aligns paired reads to the assembly using Bowtie2 (1), to create a new scaffold in a hierarchical way using first links obtained from the paired-end libraries to generate intermediate super scaffolds, which were then used as the input for subsequent runs with links from individual mate-pair libraries at increasing in sizes. At each stage, a minimum of three nonredundant links was required to join two contigs. Gaps in scaffolds from partial assemblies were then filled using GapCloser with default parameters from SOAPdenovo2 package (9). Redundancy of final joined scaffolds from partial assemblies was reduced with the Redundans tool version 0.14a (10), with the parameters “--nocleaning -m 600 --identity 0.95 --overlap 0.85 --norearrangements --nogapclosing –noscaffolding”. We used a sequence identity cut-off of 95% and a sequence overlap cut-off of 85%. The sequence identity cut-off chosen prevents loss of repetitive regions and allows for more divergence between haplotypes than expected based on observed levels of heterozygosity in the *B. jararaca* genome.

To assess the integrity of the assembled genome, BUSCO tool (11), the reference for quality and completeness of the genome assembly, was used for searching the set of 3,554 core vertebrate genes.

Automatic gene annotation

To improve the genome annotation, we searched for repetitive and transposable elements in the *B. jararaca* assembled genome. 'Squamata' clade of repeat consensus library based on RepBase version 20.05 (12) was used for RepeatMasker (13) to annotate all the known repetitive elements.

Automatic gene annotation was performed with PASA2 software, which was based on Trinity De novo and Genome guided assemblies from *B. jararaca* RNA-seq transcriptome, allowing for splicing variants, intron size of 15kb, and polyAdenylation sites identification, using the Launch_PASA_pipeline.pl script with the parameters "-d -c alignAssembly.config --MAX_INTRON_LENGTH 15000 -R --CPU 50 --TRANSDECODER -g BjararacaAssemblyGenome.fasta -t transcripts.fasta.clean --ALT_SPLICE --TDN tdn.accs --ALIGNERS blat,gmap -T -u transcripts.fasta". Exon/intron junction and UTR sequences were extracted to create a hint file for *ab initio* gene prediction for AUGUSTUS software (14), using the Scipio software (15) and AUGUSTUS scripts. Gene models were predicted with AUGUSTUS using the hint file for the prediction of protein-coding genes from *B. jararaca*.

Bacterial Artificial Chromosome Sequencing-Screening (BAC-SeqSc)

Blood from the specimen was washed three times with Phosphate-Buffered Saline (PBS) and the cell concentration was adjusted to 5 x 10^7 cells/ml. This cell suspension was equilibrated at 45°C for 10 minutes and mixed (1:1 ratio) with 1% Low-Melting Agarose (InCert® Agarose-Lonza), which was previously equilibrated at 45°C. The mixture was poured into a plug mold and left on ice for solidification. The resulting agarose plugs were submitted to Proteinase K digestion (0.5 M EDTA/1% n-lauroylsarcosine/200 µg/ml Proteinase K) at 50°C for 48 h (the Proteinase K solution was changed after 24h). The plugs were then washed in 50 mL TE pH 8.0 (10 mM Tris/1 mM EDTA) for 1 h under gentle agitation on the rocker at room temperature. The TE buffer pH 8.0 was changed and the plugs were incubated for 1 h at 50°C. The buffer was again replaced by TE pH 8.0/200 µM PMSF and the plugs were incubated for 30 minutes at 50°C twice and then rinsed again with TE buffer pH 8.0 at room temperature for 1 h twice. The plugs were stored in 0.5M EDTA at 4°C or in 70% ethanol at -20°C (in this case they were kept at 4°C overnight and then stored at - 20°C). High molecular mass DNA for BAC library preparation was prepared as described for Long Reads WGS sequencing. The BAC library was constructed following the Luo and Wing protocol (16), with few modifications. The pCUGIBAC1 vector was kindly provided by Dr. Jesus Aparecido Ferro (São Paulo State University, Faculty of Agricultural and Veterinary Sciences of Jaboticabal). Plated colonies of BAC-transformed *Escherichia coli* TransforMax EPI300 (Epicentre) were individually picked up, inoculated in LB medium containing chloramphenicol 12.5 ug/ml and incubated overnight at 37°C. The preparation yielded over 10^4 individual clones, and agarose gel electrophoresis of random prepared samples indicated inserts sizes over 100 kb (average of 150 kb).

For the BAC sequencing, we collected 9216 individual BAC clones and prepared glycerol stocks on 96 well plates. From each plate, we picked up 12 BAC clones (corresponding to 12 wells in a row) and pooled them into a single well of a new 96 well plate, thus each well in the new plate corresponds to a pool of 12 clones. It was possible to create 8 plates for a total of 768 pools of 12 clones (9216 BACs). These pools were inoculated into liquid LB medium for growth and plasmid DNA was extracted by standard alkaline lysis and filtration on MultiScreen MAGVN2250 (Millipore). Illumina sequencing libraries were prepared for each pool using Nextera DNA Sample Preparation kit (Illumina) and receiving a unique sequencing index per pool. The 768 libraries were combined in 4 groups of 192 libraries. Each of the four groups of libraries were sequenced in one lane of an Illumina HiSeq 1500 equipment, in Rapid Run mode (2 x 150bp) to obtain low coverage.

The raw sequences from the lane were demultiplexed considering the index used. The sequences within each pool were assembled with SOAPdenovo program (9) using the parameters –K 75 – k 63, and then the gaps within scaffolds were filled using GapCloser from SOAPDenovo2 package with default parameters. All contigs generated, identified by pools, were *in silico* screened for toxin genes of interest (based on BLASTN of toxin mRNAs from the transcriptome, contigs from SR-WGS and cDNAs from GenBank). Once a target sequence was matched in a pool, the contig could be readily used for downstream analysis and/or the 12 original clones in the pool were screened by PCR to identify the exact BAC clone corresponding to the toxin. In these cases, the primers used for PCR were designed based on previously assembled contigs from BAC pool sequencing. Following identification of PCR positive clones inside a pool and isolation of its plasmid DNA, we re-sequenced the individual DNA in MiSeq (Illumina) in order to obtain a higher coverage of a physically isolated BAC.

This strategy proved to be an efficient way of sequencing long genomic regions of interest with relatively low cost and fewer assembly errors due to the combination of *in silico* selection and the use of physically cloned DNA segments.

*S*creening for venom genes and re-annotation

All scaffolds generated from HA-WGS, SA-WGS, and BAC-SeqSc, and the raw reads from LR-WGS strategy were screened for segments matching toxin fragments through BLASTn searches using toxin sequences obtained from the De novo transcriptome as well as other *B. jararaca* toxin sequences available in GenBank as query. The following gene families were considered: 5'nucleotidase (5NUCL), bradykinin potentiating peptides (BPP), cysteine rich secretory protein (CRISP), C-type lectin (CTL), hyaluronidase (HYALU), L-amino-acid oxidase (LAO), nerve growth factor (NGF), phospholipase A_2 (PLA_2), phospholipase B (PLB), snake venom metalloproteinase (SVMP) (P-I, P-II and P-III classes), snake venom serine proteinase (SVSP), snake venom vascular endothelial growth factor (VEGF-F) and also nonvenom vascular endothelial growth factor (VEGF-A). Parameter varied for each toxin family, respecting sequence conservation in the family, transcript size, completeness of the scaffold, etc. In specific cases (e.g. BPP/CNP) where the screening failed to recovery full length scaffolds, LR-WGS reads were also screened in order to prospect missing genes, providing additional data to manually link scaffolds and solve gene structures. The scaffolds containing toxin genes were manually annotated on CLC Genomics Workbench v.9 to v.11 or Geneious v.10. UTRs and CDS were annotated for each gene whenever possible, and automatically predicted exon/intron boundaries were manually checked for consistency following the AG-TC rule (17). The NCBI GenBank release 240 from October 15, 2020, was used for most gene annotations and genomic comparisons.

Phylogenetic analysis and sequence comparison of VEGFs

We used GenBank (https://www.ncbi.nlm.nih.gov/genbank/) to retrieve the protein sequences of VEGF-A, VEGF-B, VEGF-C, VEGF-D, PGF, PGF-like, VEGF-F-like, and VEGF-F from several vertebrate species. We checked the genomic context of the sequences of VEGF-F, VEGF-F-like, and PGF-like to ensure that they came from a similar locus that is different from the one of VEGF-A, despite the original name annotation provided. To reconstruct the phylogenetic trees, we performed multiple sequence alignment using MAFFT (v7.450; Rozewicki et al., 2019). Then, we used IQTree v1.6.12 (18–21) to search for the Maximum Likelihood tree with ultrafast bootstrap replicates set to 1000 (-bb 1000). The final trees were adjusted using FigTree v1.4.4; (https://github.com/rambaut/figtree/). To check the percent identity among VEGF-A, VEGF-F-like, and VEGF-F, we used their coding sequences and performed alignment using MAFFT.

RNA-seq

We extracted total RNA from all collected organs using the method described by Chomczynski and Sacchi (22), with few modifications. Briefly, the tissues were ground in Polytron PT3100 homogenizer and total RNA was isolated with TRIZOL reagent (Invitrogen). PolyA+ RNAs were obtained by magnetic bead purification (DYNAbeads, LifeTechnologies). The mRNA concentration was estimated by Quant-iT RiboGreen kit (Invitrogen). The quality of total RNAs and mRNAs was evaluated by electrophoresis on picoRNA chip using the Bioanalyzer system (Agilent Technologies). The libraries for Illumina sequencing were constructed using TruSeq RNA Sample Prep Kit v2 starting from 300 ng of mRNA and sequenced on a HiSeq1500 equipment, following manufacturer recommendations.

Raw paired-end reads were preprocessed for quality control. The Trimmomatic software version 0.36 (23) was used to remove adapters and contaminants from UniVec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/), to trim the 5' and 3' ends with mean quality score below 25 (Phred+33), and discard reads shorter than 40 bp after trimming. The software fastq-mcf version 1.04.662 (2) was used for filtering by read size (> 40 bp), homopolymer content (>90%), low complexity sequences (> 90%) and poly-A/T/N tails and adapters. Paired-end reads mapping to PhiX Illumina spike-in were removed using Bowtie 2 version 2.2.5 (1), with the parameter --very-sensitive-local. The processed forward and reverse read files were then paired using Pairfq software (https://github.com/sestaton/Pairfq). After preprocessing, the high quality paired-reads were mapped into the assembled Bothrops jararaca genome, with the TopHat2 program (24) and with the following parameters: --no-mixed, minimum intron size (30pb), number of mismatches per read (3pb), number of gaps per read (3pb), --very-sensitive, maximum insertion size deletion (3bp), maximum paired-reads distance (100pb), maximum standard deviation (30bp), and only concordant uniquely mapped reads (approximately 92% of the mapped reads) were used for further analyses. The venom gland RNA-seq transcriptome was assembled with two Trinity software approaches, de novo assemble and Genome guided mode (performing a *de novo* assembly for reads aligned along the reference draft genome). The de novo and the genome guided assemblies provided complete transcripts for the identification of known and new venom genes, the respective isoforms and variants. Annotation of venom transcripts were performed by BlastX searches against Uniprot (release 07_2019) and *B. jararaca* venom proteins compiled from Genbank.

In order to estimate transcript abundance, we aligned the reads of the venom gland tissue back to the *de novo* assembled transcriptome, and maximum likelihood abundance estimates were obtained using the RSEM method (25). Final relative abundance estimates for each venom gene were calculated as Fragments Per 12*Kilobase of exon per Million fragments mapped to the CDS (Coding sequence) as follow RP12/kb = ((count/12.10^6)*10^6))/(Kb of CDS).

Venom proteomic analysis and toxin identification

The venom from the same specimen described above was analyzed by protein tandem mass spectrometry using a bottom-up shotgun approach. Fresh venom was milked before the venom gland extraction. Trypsin digestion (Sigma-Aldrich, proteomic grade trypsin) of 400 µg of protein was performed in solution into Microcon YM-10 centrifuge filters (Millipore) using Filter-Aided Sample Preparation (FASP) method (26) with minor modifications. The final tryptic peptide solution was acidified to pH ≤ 3 with TFA before desalting with STAGE tip procedure (27). Peptides were eluted with 80% ACN, 0.1% TFA, dried and resuspended on solvent A (0.1% formic acid). The trypsin digestion was performed in duplicate and tryptic peptides was analyzed as three technical replicates each.

LC-MS/MS analysis was performed on an Easy nanoLC system (Thermo) coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo). Five µL of peptide solution was loaded into a 2 cm C18 trap column (Jupiter 10um C18, Phenomenex, 100 µm i.d. × 360 µm o.d.) and separated on a 10-cm long C18 column (Aqua 5 µm C18, Phenomenex, 75 µm i.d. x 360 µm o.d.), packed in-house, by a linear gradient of 5 to 35% B (0.1% formic acid in acetonitrile) in 85 min at a flow rate of 200 nL/min (5–35% B in 85 min; 35–85% B for 10 min and 85-5% B in 2 min). Spray voltage and capillary temperature were set at 2.2

kV and 200 °C and the mass spectrometer was operated in data dependent mode (one full MS scan was acquired in the m/z range of 300–1650 followed by MS/MS acquisition using CID dissociation of the 12 most intense ions per scan (charge state >=2). Orbitrap analyzer MS resolution was 60,000 (at 400 m/z) and MS/MS spectra was acquired by ion trap mass analyzer. The maximum injection time and AGC target were set to 250 ms and 1E6 for full MS, and 100 ms and 5E4 for MS/MS. The minimum signal threshold to trigger fragmentation event, isolation window and normalized collision energy (NCE) were set to, respectively, 5E3 cps, 2 m/z and 35%. A dynamic peak exclusion was applied to avoid the same m/z of being selected for the next 90 s.

Venom protein identification was performed on the search engine PEAKS (version X) using the venom gland transcriptome from the same individual and all “Squamata” sequences available at Uniprot (release 07_2019). A decoy and contaminant database were generated to exclude contaminants and false positive results. Parameters were set as follows: Parent Mass Error Tolerance: 0.5 Da; Fragment Mass Error Tolerance: 0.5 Da; Precursor Mass Search Type: monoisotopic; Enzyme: Trypsin; Max Missed Cleavages: 1; Nonspecific Cleavage: none; Fixed Modifications: Carbamidomethylation: 57.02; Variable Modifications: Oxidation (M): 15.99; and False Discovery Rate (FDR) Estimation: Enabled. MS/MS identification results were filtered for FDR = 0.1%.

Supplementary Figures and Tables

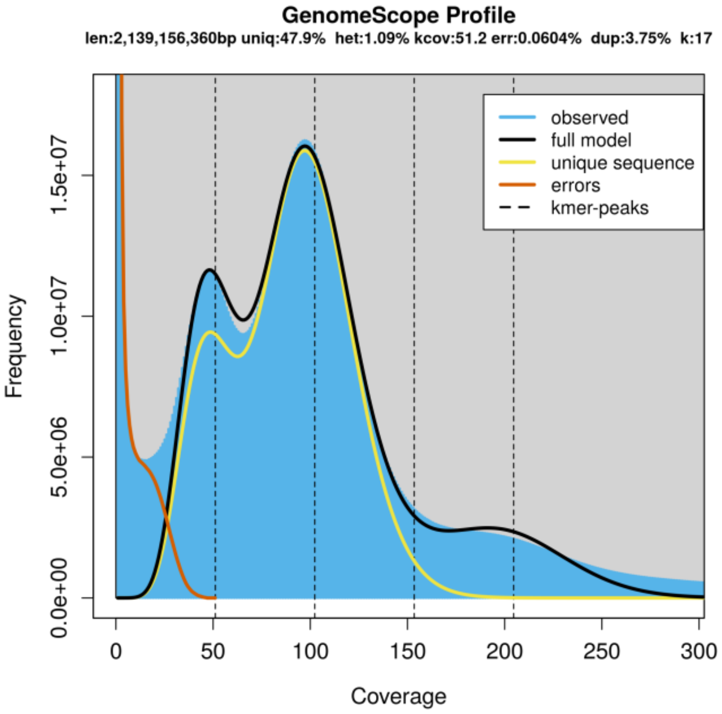

Fig. S1. Estimation of genome size, repeat content, and heterozygosity by GenomeScope software, based on 17-mers in HiSeq Illumina sequence reads (max kmer coverage at 1000). The higher peak at the coverage around 90 is the homozygous portion of the genome, which accounts for the strands of the DNA having identical 17-mers. The smaller peak to the left of the higher one corresponds to the heterozygous portion of the genome, which accounts for the strands of the DNA having different 17-mers. If the genome is highly heterozygous, the height of the smaller peak would be closer to that of the homozygous peak.

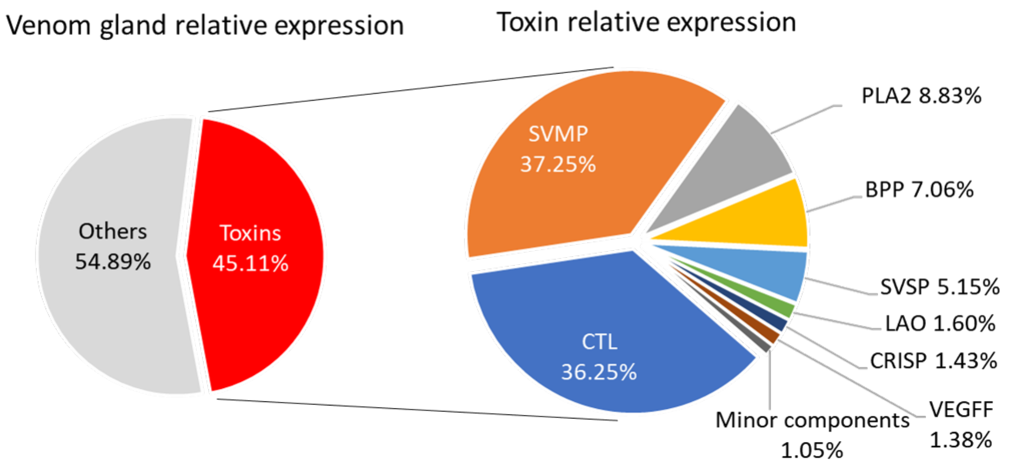

Fig. S2. Relative expression of toxin families in the venom gland transcriptome. Proportinos were calculated based on FPKM values obtained from the mapping of venom gland reads to the de novo assembled transcripts including curated toxin transcripts.

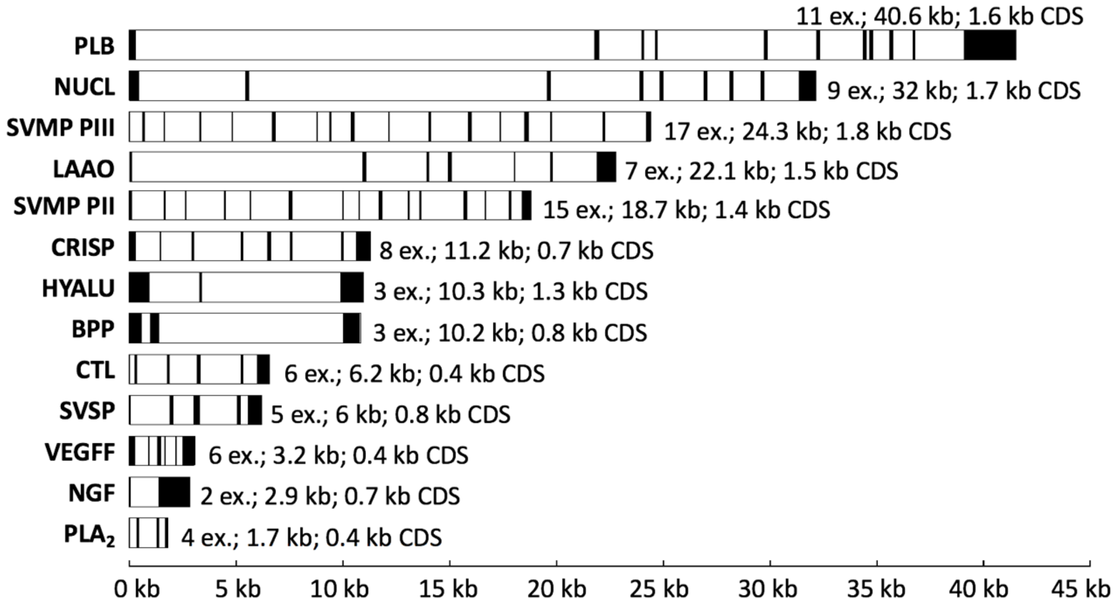
PLB
NUCL
SVMP PIII
LAAO
SVMP PII
CRISP
HYALU
BPP
CTL
SVSP
VEGFF
NGF
PLA$_2$
11 ex.; 40.6 kb; 1.6 kb CDS
9 ex.; 32 kb; 1.7 kb CDS
17 ex.; 24.3 kb; 1.8 kb CDS
7 ex.; 22.1 kb; 1.5 kb CDS
15 ex.; 18.7 kb; 1.4 kb CDS
8 ex.; 11.2 kb; 0.7 kb CDS
3 ex.; 10.3 kb; 1.3 kb CDS
3 ex.; 10.2 kb; 0.8 kb CDS
6 ex.; 6.2 kb; 0.4 kb CDS
5 ex.; 6 kb; 0.8 kb CDS
6 ex.; 3.2 kb; 0.4 kb CDS
2 ex.; 2.9 kb; 0.7 kb CDS
4 ex.; 1.7 kb; 0.4 kb CDS
0 kb
5 kb
10 kb
15 kb
20 kb
25 kb
30 kb
35 kb
40 kb
45 kb

Fig. S3. Intron/exon organization of representative genes of venom proteins. Protein family codes are followed by a graphical representation of genes, in which exons are represented as black boxes and introns as white boxes. Additional information is supplied, such as the number of exons (ex.), gene size and CDS size in kb.

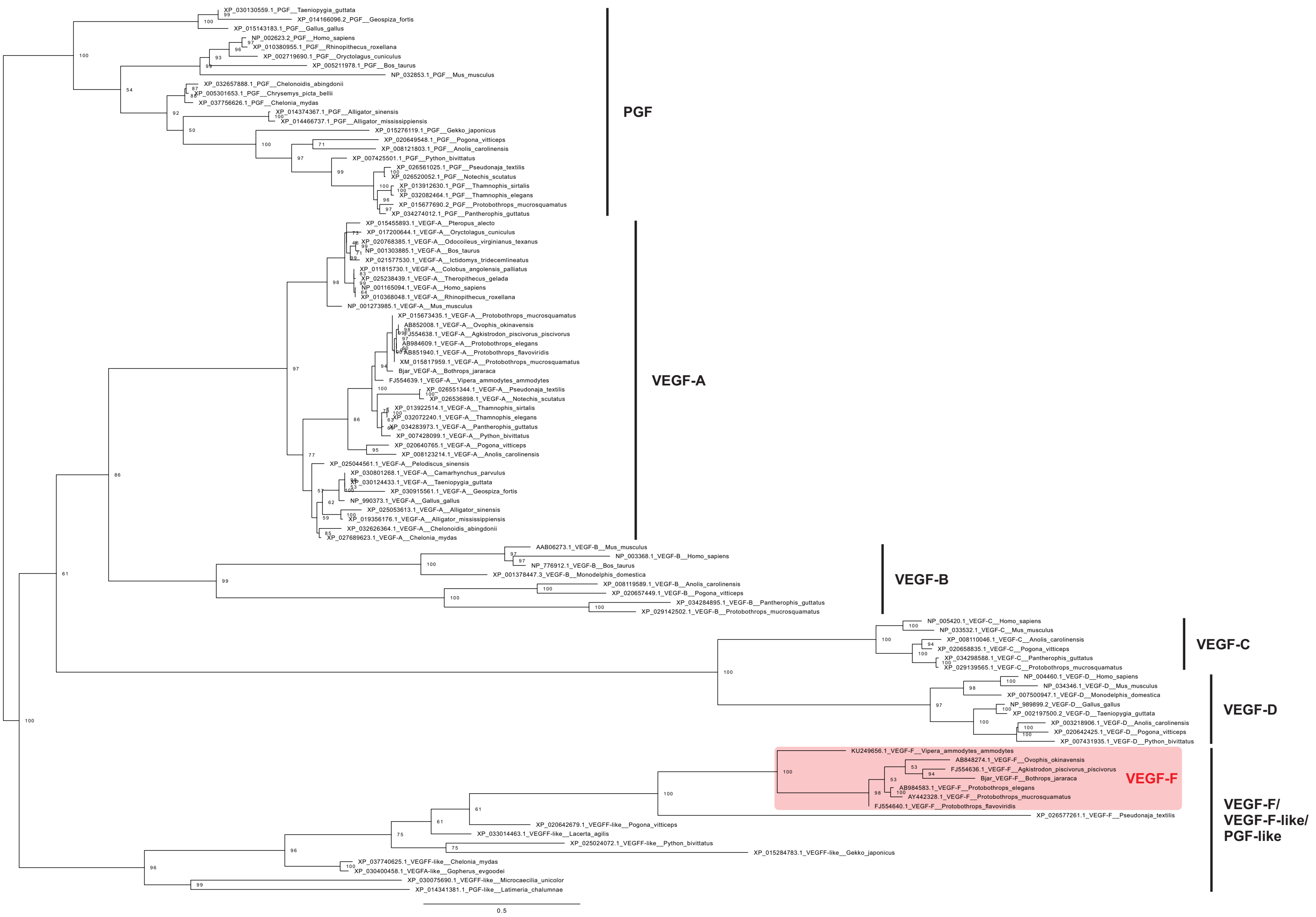
XP_030130559.1_PGF__Taeniopygia_guttata
XP_014166096.2_PGF__Geospiza_fortis
XP_015143183.1_PGF__Gallus_gallus
NP_002623.2_PGF__Homo_sapiens
XP_010380955.1_PGF__Rhinopithecus_roxellana
XP_002719690.1_PGF__Oryctolagus_cuniculus
XP_005211978.1_PGF__Bos_taurus
NP_032853.1_PGF__Mus_musculus
XP_032657888.1_PGF__Chelonoidis_abingdonii
XP_005301653.1_PGF__Chrysemys_picta_bellii
XP_037756626.1_PGF__Chelonia_mydas
XP_014374367.1_PGF__Alligator_sinensis
XP_014466737.1_PGF__Alligator_mississippiensis
XP_015276119.1_PGF__Gekko_japonicus
XP_020649548.1_PGF__Pogona_vitticeps
XP_008121803.1_PGF__Anolis_carolinensis
XP_007425501.1_PGF__Python_bivittatus
XP_026561025.1_PGF__Pseudonaja_textilis
XP_026520052.1_PGF__Notechis_scutatus
XP_013912630.1_PGF__Thamnophis_sirtalis
XP_032082464.1_PGF__Thamnophis_elegans
XP_015677690.2_PGF__Protobothrops_mucrosquamatus
XP_034274012.1_PGF__Pantherophis_guttatus
PGF
XP_015455893.1_VEGF-A__Pteropus_alecto
XP_017200644.1_VEGF-A__Oryctolagus_cuniculus
XP_020768385.1_VEGF-A__Odocoileus_virginianus_texanus
NP_001303885.1_VEGF-A__Bos_taurus
XP_021577530.1_VEGF-A__Ictidomys_tridecemlineatus
XP_011815730.1_VEGF-A__Colobus_angolensis_palliatus
XP_025238439.1_VEGF-A__Theropithecus_gelada
NP_001165094.1_VEGF-A__Homo_sapiens
XP_010368048.1_VEGF-A__Rhinopithecus_roxellana
NP_001273985.1_VEGF-A__Mus_musculus
XP_015673435.1_VEGF-A__Protobothrops_mucrosquamatus
AB852008.1_VEGF-A__Ovophis_okinavensis
FJ554638.1_VEGF-A__Agkistrodon_piscivorus_piscivorus
AB984609.1_VEGF-A__Protobothrops_elegans
AB851940.1_VEGF-A__Protobothrops_flavoviridis
XM_015817959.1_VEGF-A__Protobothrops_mucrosquamatus
Bjar_VEGF-A__Bothrops_jararaca
FJ554639.1_VEGF-A__Vipera_ammodytes_ammodytes
XP_026551344.1_VEGF-A__Pseudonaja_textilis
XP_026536898.1_VEGF-A__Notechis_scutatus
XP_013922514.1_VEGF-A__Thamnophis_sirtalis
XP_032072240.1_VEGF-A__Thamnophis_elegans
XP_034283973.1_VEGF-A__Pantherophis_guttatus
XP_007428099.1_VEGF-A__Python_bivittatus
XP_020640765.1_VEGF-A__Pogona_vitticeps
XP_008123214.1_VEGF-A__Anolis_carolinensis
XP_025044561.1_VEGF-A__Pelodiscus_sinensis
XP_030801268.1_VEGF-A__Camarhynchus_parvulus
XP_030124433.1_VEGF-A__Taeniopygia_guttata
XP_030915561.1_VEGF-A__Geospiza_fortis
NP_990373.1_VEGF-A__Gallus_gallus
XP_025053613.1_VEGF-A__Alligator_sinensis
XP_019356176.1_VEGF-A__Alligator_mississippiensis
XP_032626364.1_VEGF-A__Chelonoidis_abingdonii
XP_027689623.1_VEGF-A__Chelonia_mydas
VEGF-A
AAB06273.1_VEGF-B__Mus_musculus
NP_003368.1_VEGF-B__Homo_sapiens
NP_776912.1_VEGF-B__Bos_taurus
XP_001378447.3_VEGF-B__Monodelphis_domestica
XP_008119589.1_VEGF-B__Anolis_carolinensis
XP_020657449.1_VEGF-B__Pogona_vitticeps
XP_034284895.1_VEGF-B__Pantherophis_guttatus
XP_029142502.1_VEGF-B__Protobothrops_mucrosquamatus
VEGF-B
NP_005420.1_VEGF-C__Homo_sapiens
NP_033532.1_VEGF-C__Mus_musculus
XP_008110046.1_VEGF-C__Anolis_carolinensis
XP_020658835.1_VEGF-C__Pogona_vitticeps
XP_034298588.1_VEGF-C__Pantherophis_guttatus
XP_029139565.1_VEGF-C__Protobothrops_mucrosquamatus
VEGF-C
NP_004460.1_VEGF-D__Homo_sapiens
NP_034346.1_VEGF-D__Mus_musculus
XP_007500947.1_VEGF-D__Monodelphis_domestica
NP_989899.2_VEGF-D__Gallus_gallus
XP_002197500.2_VEGF-D__Taeniopygia_guttata
XP_003218906.1_VEGF-D__Anolis_carolinensis
XP_020642425.1_VEGF-D__Pogona_vitticeps
XP_007431935.1_VEGF-D__Python_bivittatus
VEGF-D
KU249656.1_VEGF-F__Vipera_ammodytes_ammodytes
AB848274.1_VEGF-F__Ovophis_okinavensis
FJ554636.1_VEGF-F__Agkistrodon_piscivorus_piscivorus
Bjar_VEGF-F__Bothrops_jararaca
AB984583.1_VEGF-F__Protobothrops_elegans
AY442328.1_VEGF-F__Protobothrops_mucrosquamatus
FJ554640.1_VEGF-F__Protobothrops_flavoviridis
VEGF-F
XP_026577261.1_VEGF-F__Pseudonaja_textilis
XP_020642679.1_VEGFF-like__Pogona_vitticeps
XP_033014463.1_VEGFF-like__Lacerta_agilis
XP_025024072.1_VEGFF-like__Python_bivittatus
XP_015284783.1_VEGFF-like__Gekko_japonicus
XP_037740625.1_VEGFF-like__Chelonia_mydas
XP_030400458.1_VEGFA-like__Gopherus_evgoodei
XP_030075690.1_VEGFF-like__Microcaecilia_unicolor
XP_014341381.1_PGF-like__Latimeria_chalumnae
VEGF-F/
VEGF-F-like/
PGF-like
0.5

Fig. S4. Maximum likelihood tree of peptide sequences of vascular endothelial growth factor (VEGF) family (VEGF-A to -F), VEGF-F-like, PGF-like, and PGF genes. The venom-specific VEGF-F clade is highlighted in red. The names of each sequence follow the schema: Genbank accession numbers, name of the gene, and species. Numbers in the branches correspond to Bootstrap values.

Table S1: Relationship between *de novo* assembled toxin transcripts, retrieved genes, proteins identified in the venom, and most similar sequence in GenBank.

Transcriptome					Genome	Proteome		NCBI BLASTn hit	
Transcript ID	Expression (FPKM)[a]	VG r.e.[b] (%)	Tox. r.e.[c] (%)	CDS size[d] (bp)	Gene ID	Score 10lgP[e]	# of Spec[f]	Accession[g]	Id.%[h]
SVMP									
BJAR_SVMP3_t01	116817.9	9.79	21.7	1833	BJARHA_SVMP3_g12	376.4	468	AF056025.2	99.4
BJAR_SVMP2_t05	50970.1	4.27	9.47	1437	BJARHA_SVMP2_g05	331.8	387	AF345931.1	99.3
BJAR_SVMP3_t05	8871.2	0.74	1.65	1764	BJARBC_SVMP3_g19	297.5	196	AF450503.1	95.9
BJAR_SVMP3_t03	6653.1	0.56	1.24	1836	BJARHA_SVMP3_g02	313.9	140	AY149647.1	93.5
BJAR_SVMP3_t08	5963.1	0.50	1.11	1806	BJARHA_SVMP3_g15	258.8	69	AF149788.5	99.2
BJAR_SVMP3_t09	2836.6	0.24	0.53	1806	BJARHA_SVMP3_g15	261.2	87	AF149788.5	98.9
BJAR_SVMP3_t10	2498.2	0.21	0.46	1632	No full gene assembled	187.3	39	AF149788.5	95.7
BJAR_SVMP2_t06	2076.4	0.17	0.39	1434	BJARHA_SVMP2_g04	142.8	11	AY736107.1	92.8
BJAR_SVMP3_t11	1166.1	0.10	0.22	1701	No full gene assembled	262.5	57	AF450503.1	90.7
BJAR_SVMP3_t04	786.4	0.07	0.15	1746	BJARBC_SVMP3_g20	220.2	66	EU733641.1	98.6
BJAR_SVMP2_t04	608.3	0.05	0.11	1380	BJARHA_SVMP2_g01	n.d.	n.d.	HQ414107.1	92.2
BJAR_SVMP3_t06	394.5	0.03	0.07	1818	No full gene assembled	325.9	359	AF345931.1	97.4
BJAR_SVMP2_t03	389.2	0.03	0.07	1452	BJARHA_SVMP2_g01	n.d.	n.d.	GQ451438.1	92.5
BJAR_SVMP2_t02	270.5	0.02	0.05	1398	BJARHA_SVMP2_g02	169.9	26	HQ414108.1	95.0
BJAR_SVMP2_t01	206.0	0.02	0.04	1437	BJARHA_SVMP2_g01	169.9	26	HQ414108.1	94.1
CTL									
BJAR_CTL_t06	38284.2	3.21	7.11	459	No full gene assembled	214.8	205	AY962524.1	85.6
BJAR_CTL_t02	37418.7	3.14	6.95	441	BJARLR_CTL_g02	218.2	276	MG132014.1	87.0
BJAR_CTL_t04	28823.9	2.42	5.36	468	No full gene assembled	195.4	102	AY962524.1	98.7
BJAR_CTL_t09	27098.1	2.27	5.03	441	No full gene assembled	221.6	274	MG132014.1	86.6
BJAR_CTL_t11	25419.2	2.13	4.72	441	BJARHA_CTL_g03	284.5	139	AY091761.1	89.5
BJAR_CTL_t03	22901	1.92	4.25	453	No full gene assembled	196.1	66	AY962525.1	98.9
BJAR_CTL_t10	9672.7	0.81	1.80	477	No full gene assembled	196.3	109	HQ414092.1	93.3
BJAR_CTL_t05	4767.5	0.40	0.89	453	No full gene assembled	189.9	60	AY962525.1	94.9
BJAR_CTL_t08	554.1	0.05	0.10	372	No full gene assembled	83.4	6	AY091761.1	90.8
BJAR_CTL_t01	197.1	0.02	0.04	477	BJARHA_CTL_g01	107.6	5	AY522720.1	99.1
BJAR_CTL_t07	38.1	0.00	0.01	456	No full gene assembled	82.7	3	MG132013.1	95.3
PLA_2									
BJAR_PLA2gA_t01	47498.8	3.98	8.83	417	BJARHA_PLA2_g01	249.1	105	AY145836.1	88.7
BPP									
BJAR_BPP_t02	20128.1	1.69	3.74	792	BJARLR_BPP_g01	n.d.	n.d.	AF171670.2	97.6
BJAR_BPP_t01	17895.8	1.50	3.32	798	BJARLR_BPP_g01	n.d.	n.d.	AF171670.2	98.2
SVSP									
BJAR_SVSP_t03	10718.7	0.90	1.99	777	BJARHA_SVSP_g03	231.7	107	XM_015816070	93.6
BJAR_SVSP_t06	5467.3	0.46	1.02	777	BJARHA_SVSP_g06	244.7	155	AB178322.1	99.4
BJAR_SVSP_t05	3099.6	0.26	0.58	777	BJARHA_SVSP_g05	191.7	42	AF490536.1	98.8
BJAR_SVSP_t07	3071.5	0.26	0.57	777	BJARBC_SVSP_g07	238.3	70	DQ247724.1	96.4
BJAR_SVSP_t01	3037.8	0.25	0.56	783	BJARHA_SVSP_g01	282.2	109	AY251282.1	98.6
BJAR_SVSP_t04	1072.1	0.09	0.20	777	BJARHA_SVSP_g04	201.7	47	MF974529.1	95.6
BJAR_SVSP_t08	571.4	0.05	0.11	774	BJARHA_SVSP_g08	200.1	49	MF974466.1	91.1
BJAR_SVSP_t02	474.7	0.04	0.09	774	BJARHA_SVSP_g02	189.0	52	AB031394.1	90.0
BJAR_SVSP_t09	145.1	0.01	0.03	777	BJARHA_SVSP_g09	178.6	44	DQ247724.1	94.3
LAAO									
BJAR_LAO_t01	8589.7	0.72	1.60	1509	BJARHA_LAO_g01	376.1	399	EU870608.1	99.0
CRISP									
BJAR_CRISP_t01	7687.4	0.64	1.43	723	BJARSA_CRISP_g01	376.1	399	MG132022.1	98.1
VEGFF									
BJAR_VEGFF_t01	7424.6	0.62	1.38	441	BJARHA_VEGFF_g01	150.0	42	AY033152.1	99.5
NGF									
BJAR_NGF_t01	2424.5	0.20	0.45	726	BJARSA_NGF_g01	150.0	42	AY007318.1	99.5
PLB									
BJAR_PLB_t01	1788.4	0.15	0.33	1662	BJARSA_PLB_g01	255.2	121	MG132007.1	99.2
NUCL									
BJAR_NUCL_t01	1060.4	0.09	0.20	1776	BJARSA_NUCL_g01	296.4	132	AB985247.1	97.9
HAYLU									
BJAR_HYALU_t01	210.1	0.02	0.04	1350	BJARHA_HYALU_g01	171.9	27	MG132023.1	99.7

Table S2: Information on venom gene scaffolds identified in the *B. jararaca* genome.

Protein family	Nº of genes retrieved	Average gene size	Gene clustering[a]	Sequencing strategy contributing[b]	Scaffold ID
SVMP P-III	20	26.8 kb	M+n	HA-WGS and BAC-SeqSc	BJARHA_S804283_A BJARHA_S279442_A BJARHA_S248196 BJARHA_S123025 BJARHA_S83452 BJARHA_S55045 BJARHA_S53622 BJARHA_S49785 BJARBC_30E11N1 BJARBC_23G12N1
SVMP P-II	7	20.1 kb	M+n	HA-WGS and BAC-SeqSc	BJARHA_S1060317F_A BJARHA_S804283 BJARHA_S248196 BJARHA_S123025 BJARHA_S38175 BJARBC_30E11N1 BJARBC_20E09Ma1 BJARBC_27H01N1 BJARBC_23G12Ma1
CTL	6	7.7 kb	nc	HA-WGS and LR-WGS	BJARHA_S16055 BJARHA_S18032 BJARHA_S26073 BJARHA_S28402 BJARHA_S140991 BJARLR_CR002
PLA_2	1	1.7 kb	S+n	HA-WGS	BJARHA_S32675
BPP/CNP	1	10.9 kb	S]	LR-WGS	BJARLR_CR001
SVSP	12	10.5 kb	M	HA-WGS and BAC-SeqSc	BJARHA_S43225 BJARHA_S75416 BJARHA_S90164F_A BJARHA_S116239 BJARHA_S427266_B BJARBC_20G10Ma1
LAAO	2	28 kb	nc	HA-WGS	BJARHA_S280243
CRISP	1	11.1 kb	nc	SA-WGS	BJARSA_S077418
VEGF-F	1	3.1 kb	S	HA-WGS and BAC-SeqSc	BJARHA_S39975 BJARBC_02H08Ma1
NGF	1	76 kb	S	SA-WGS	BJARSA_S035252
PLB	1	40.6 kb	S	SA-WGS BAC-SeqSc	BJARSA_S078071_A
NUCL	1	32 kb	S	SA-WGS	BJARSA_S080912
HYALU	1	10.3 kb	S	HA-WGS	BJARHA_S173890

a- Gene clustering: M: the segments contain multiple paralogues of toxin genes; S: the segments contain a single venom gene; +n: the segments contain non-venom paralogue(s) (coding for the same protein family but not significatively expressed in the venom gland); nc: not conclusive. b- Strategies: HA-WGS: Hybrid Assembly Whole Shotgun Sequencing, SA-WGS: Short read Assembly Whole Shotgun Sequencing, LR-WGS: Long Reads Whole Shotgun Sequencing, and BAC-SeqSc: BAC Sequencing and Screening.

Dataset S1 (separate file). Protein identification in *B. jararaca* venom by trypsin digestion and LC-MS/MS analysis.

SI References

1. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
2. E. Aronesty, Comparison of Sequencing Utility Programs. *Open Bioinforma. J.* (2013) https:/doi.org/10.2174/1875036201307010001.
3. E. S. Lander, M. S. Waterman, Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* (1988) https:/doi.org/10.1016/0888-7543(88)90007-9.
4. G. Marcais, C. Kingsford, Jellyfish : A fast k-mer counter. *Tutorialis e Manuais* (2012).
5. G. W. Vurture, *et al.*, GenomeScope: Fast reference-free genome profiling from short reads in *Bioinformatics*, (2017) https:/doi.org/10.1093/bioinformatics/btx153.
6. A. V. Zimin, *et al.*, The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
7. J. Butler, *et al.*, ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res.* (2008) https:/doi.org/10.1101/gr.7337908.
8. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* (2011) https:/doi.org/10.1093/bioinformatics/btq683.
9. R. Luo, *et al.*, SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* (2012) https:/doi.org/10.1186/2047-217X-1-18.
10. L. P. Pryszcz, T. Gabaldón, Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* (2016) https:/doi.org/10.1093/nar/gkw294.
11. 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* (2015) https:/doi.org/10.1093/bioinformatics/btv351.
12. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* (2015) https:/doi.org/10.1186/s13100-015-0041-9.
13. M. Tarailo-Graovac, N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* (2009) https:/doi.org/10.1002/0471250953.bi0410s25.
14. M. Stanke, *et al.*, AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
15. O. Keller, F. Odronitz, M. Stanke, M. Kollmar, S. Waack, Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* (2008) https:/doi.org/10.1186/1471-2105-9-278.
16. M. Luo, R. A. Wing, “An Improved Method for Plant BAC Library Construction” in

Plant Functional Genomics, (Humana Press, 2003), pp. 3–20.

17. S. M. Mount, A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**, 459–472 (1982).
18. D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* (2018) https:/doi.org/10.1093/molbev/msx281.
19. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
20. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
21. J. Rozewicki, S. Li, K. M. Amada, D. M. Standley, K. Katoh, MAFFT-DASH: Integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47**, W5–W10 (2019).
22. P. Chomczynski, N. Sacchi, Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
23. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) https:/doi.org/10.1093/bioinformatics/btu170.
24. D. Kim, *et al.*, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* (2013) https:/doi.org/10.1186/gb-2013-14-4-r36.
25. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* (2011) https:/doi.org/10.1186/1471-2105-12-323.
26. J. R. Wiśniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
27. J. Rappsilber, Y. Ishihama, M. Mann, Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics. *Anal. Chem.* **75**, 663–670 (2003).