Article - Discovery

Bee core venom genes predominantly originated before aculeate stingers evolved

Ivan Koludarov¹ *, Mariana Velasqué², Thomas Timm³, Gánter Lochenit³, Michael Heinzinger⁴, Andreas Vilcinskas¹,⁵, Rosalyn Gloag⁶, Brock A. Harpur⁷, Lars Podsiadlowski⁸, Burkhard Rost⁴, Sebastien Dutertre⁹, Eckart Stolle⁷, and Björn M von Reumont¹,¹⁰*

¹ Justus Liebig University of Gießen, Institute for Insect Biotechnology, Heinrich-Buff-Ring 58, 35392 Gießen
² Genomics & Regulatory Systems Unit, Okinawa Institute of Science & Technology, 1919, Tancha, Okinawa, Japan
³ Protein Analytics, Institute of Biochemistry, Justus Liebig University, Friedrichstrasse 24, 35392 Giessen, Germany
⁴ Department of Informatics, Bioinformatics and Computational Biology, i12, Technical University of Munich, Boltzmannstr. 3, Garching, 85748 Munich, Germany.
⁵ Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Bioresources, Ohlebergs weg 12, 35392 Gießen, Germany
⁶ Rosalyn Gloag - School of Life and Environmental Sciences, The University of Sydney, NSW 2006, Australia
⁷ Brock A. Harpur – Dept of Entomology, Purdue University, West Lafayette, IN. USA
⁸ Leibniz Institute for the Analysis of Biodiversity Change, Zoological Research Museum Alexander Koenig, Centre of Molecular Biodiversity Research, Adenauerallee 160, 53113 Bonn, Germany
⁹ IBMM, Université Montpellier, CNRS, ENSCM, 34095 Montpellier, France
¹⁰ LOEWE Centre for Translational Biodiversity Genomics (TBG), Senckenberganlage 25, 60325 Frankfurt, Germany

*Corresponding authors
BMvR: bmvr@reumont.net
IK: Ivan.Koludarov@agr.uni-giessen.de
Keywords
Hymenoptera venom, bee toxins, solitary bee venom, proteo-transcriptomics, genomics, venom gene evolution, machine learning, melittin, apamin, aculeatoxins

Abstract
Venoms have convergently evolved in all major animal lineages and are ideal candidates to unravel the genomic processes underlying convergent trait evolution. However, few animal groups have been studied in detail, and large-scale comparative genomic analyses to address toxin gene evolution are rare. The hyper-diverse hymenopterans are the most speciose group of venomous animals, but the origin of their toxin genes has been largely overlooked. We combined proteo-transcriptomics with comparative genomics compiling an up-to-date list of core bee venom proteins to investigate the origin of 11 venom genes in 30 hymenopteran genomes including two newly sequenced genomes of stingless bees. We found a more distinct pattern in which toxin genes originated predominantly by single gene co-option, a prevalent mechanism in parasitoid wasps. These are always accompanied by parallel expansion events in other bee and hymenopteran sister gene groups. The short toxic peptides melittin and apamin appear to be unique to bees, a result supported by a novel machine learning approach. Based on the syntenic pattern we propose here Anthophilin1 as a bee-unique gene family that includes apamin and MCDP. It appears that bees co-adapted their venom genes which predominantly already existed before the aculeate ovipositor evolved to a stinger which exclusively injects venom. Our results provide insight into the large-scale evolution of bee venom compounds, and we present here the first study revealing general processes of venom evolution at a comparative genomic level for this mega-diverse Hymenoptera.

Introduction
Molecular processes that drive the evolution of adaptive traits are among the most widely discussed topics in biology (Dunn and Munro 2016; Yin et al. 2016). Venoms, one of the most successful adaptive traits, have arisen by convergent evolution more than 100 times in more than 220,000 species from all major animal groups, covering ~15% of global animal biodiversity (Casewell et al. 2013; Holford et al. 2018; Schendel et al. 2019). Venoms are cocktails of short peptides, enzymes and other proteins that are injected into other organisms for predation, defense or competition, using a specialized morphological structure known as the venom apparatus. The function of venom genes makes them ideal candidates to study the molecular mechanisms that facilitate the evolution of adaptive traits (Drukewitz and von Reumont 2019; Zancolli and Casewell 2020). Advances in comparative genomics and sequencing provide an opportunity to understand those mechanisms at the genomic level (Drukewitz et al. 2019; Zancolli and Casewell 2020; Almeida et al. 2021). Nevertheless, there have been few large comparative studies focusing on the genomic origins of toxin
genes and their weaponization. One reason for this is the predominant interest in venoms for their pharmacological and agrochemical applications or clinical toxinology (Casewell et al. 2013; Holford et al. 2018). Researchers have therefore prioritized groups such as snakes, scorpions, spiders, and cone snails that may not be species-rich compared to insects but are known for powerful venom components of which many cause strong envenomation effects on humans (Holford et al. 2018; Drukewitz and von Reumont 2019). This taxonomic bias hinders a deeper understanding of the origin and evolution of venoms, particularly the vast body of knowledge hidden in the mega-radiations of insects. However, some exceptional insects have been studied in detail recently (for review see Walker et al. 2018; Walker 2020), including the honeybee (Dotimas and Hider 1987; Danneels et al. 2015).

A major dogma in venom evolution is that genes generally gain new functions via duplication followed by neofunctionalization or subfunctionalization (Ohno 1970; Innan and Kondrashov 2010; for the toxicological implications see Jackson and Koludarov 2020). For toxins this process has been illustrated in many cases and species lineages. In toxinology the process of toxin gene origin, referred to as “recruitment”, implies that a gene expression location changes from a venom-unrelated tissue to the venom system associated with the acquisition of toxic functions (Nei and Rooney 2005; Casewell et al. 2013; Drukewitz and von Reumont 2019). Genomic research in cnidarians and snakes, the latter is one of the best-studied taxa in this context, suggests that venoms evolve predominantly via gene duplication followed by neofunctionalization and with a subsequent gene loss in several venom protein families (Wong and Belov 2012; Vonk et al. 2013; Moran and Zakon 2014; Dowell et al. 2018; Koludarov et al. 2020; Almeida et al. 2021). However, recent studies (including several in snakes) have challenged this simplistic paradigm and proposed a more complex relationship between gene duplication and neofunctionalization (Dowell et al. 2016; Dowell et al. 2018; Koludarov et al. 2020). Furthermore, gene duplication might not be the driving force behind venom evolution in some vertebrate taxa. In platypus, for example, single gene co-option was proposed as the major recruitment method for toxin genes (Whittington et al. 2018). The same was proposed for parasitoid wasps (Nasonia spp.) (Sim and Wheeler 2016; Martinson et al. 2017). Indeed, diverse mechanisms may operate in insects: both gene duplication and single gene co-option were shown to shape the evolution of venom genes in robber flies (Drukewitz and von Reumont 2019).

Hymenopterans (sawflies, wasps, ants and bees) are the most species-rich insect group that are of tremendous ecological and economical importance. They also feature the largest number of venomous species, and close encounters between humans and few common species of pollinating bees or hunting wasps regularly lead to accidental and sometimes lethal envenomation. Unsurprisingly, honeybee and bumblebee venoms are among the best-characterized venoms in the animal kingdom. The venoms of the remaining species of the hymenopteran mega-radiation, including bees with their large number of solitary species, remain largely unexplored despite recent proteo-transcriptomic venom descriptions of several (mostly ant and wasp) species.
(Moreno and Giralt 2015, Supplementary Table 1). However, most of the studies deal with single crude fractions or even individual components either due to technical limitations at the time or because of applied research focus (Lima and Brochetto-Braga 2003; Pick 1986; Lee et al. 2016; Santos-Pinto et al. 2018). Consequently, we still know little about the evolutionary origin and dynamics of hymenopteran and in particular bee venoms, even though high-quality genomic resources are now available for many species (Sim and Wheeler 2016; Branstetter et al. 2018; Elsik et al. 2018).

To address this knowledge gap, we combined all relevant verified published toxin data on hymenopterans with new proteo-transcriptomic data of two solitary bees (Xylocopa violacea and Halictus scabiosae) and the honeybee (Apis mellifera) to establish a core venom profile (CVP) for bees (Anthophila). We then used genomic data of 30 species (including two new genomes of stingless bees), representing bees and additional hymenopteran outgroups, for the first taxon-wide comparative genomics analysis to trace the origin and evolution of bee core venom genes. This analysis allowed us to address three key questions: (1) whether bee venoms are dominated by toxin genes that are novel and unique to this clade, (2) whether single gene co-option is the major mechanism of venom gene evolution in bees, as is the case for parasitoid wasps, and (3) whether morphological and ecological adaptations are reflected in the mechanism of venom gene evolution.

Results

Venom proteo-transcriptomics

To address the lack of detailed bee venom profiles other than for honeybee and bumblebee, we generated high-resolution proteome-transcriptome venom data from two phylogenetically distant solitary bees: the great-banded furrow-bee (H. scabiosae) and the violet carpenter bee (X. violacea). We also generated complementary data from the honeybee (A. mellifera). First, we generated tissue-specific short-read transcriptomes for the whole venom system in all three species. Afterwards we processed the data using a multiple-assembler pipeline (Oyster River Pipeline (MacManes 2018)) to predict open reading frames, allowing us to deduce the translated proteins. Next, we identified secreted venom proteins in the crude venom by Orbitrap-based mass spectrometry (MS) using the venom gland transcriptomes as specific databases. The matched protein sequences were annotated (BlastP searches against UniProt, Uniprot-Apocrita and ToxProt) to identify venom peptides and proteins that are supported by proteome and transcriptome data (Supplementary Tables 2–4). We then identified the proteomically supported venom protein families by searching sequences and their annotation against known ToxProt candidates and additionally against described bee specific venom components (Supplementary Table 1). We considered transcripts as venom components if they showed a ToxProt bitscore ≥ to the UniProt bitscore to reduce false-positive matches from possible house-keeping genes. In brief, we identified 36 peptide and protein families as venom components in the crude venoms of A. mellifera, H. scabiosae and X. violacea. All three venoms
predominantly contained low-molecular-weight peptides, in particular melittin, apamin and mast-cell degranulating peptide (MCDP). Larger proteins such as phospholipase A2, venom acid phosphatase, venom dipeptidyl peptidase 4 and venom allergens made up less than 10% of the transcriptome (Figure 1).

Figure 1: The venom composition of *A. mellifera*, *H. scabiosae* and *X. violacea*. Only venom protein transcripts also identified in the proteome are listed here. Transcript expression is shown as a percentage based on scaled transcript per million (TPM) values. The eleven selected venom proteins that we discuss herein further as bee core venom profile are printed in bold.

These results match the few known venom profiles of aculeate wasps and bees, in which small peptides were identified as the dominant components (Lima and Brochetto-Braga 2003; Danneels et al. 2015; Santos-Pinto et al. 2018). Even so, the dominant venom proteins were neither similarly expressed nor necessarily present in all three species. Furthermore, some of the venom components present in the transcriptomes of all three
species were only recovered in one of them by MS analysis (Apidaecin, MCDP, and venom allergen isoforms, see Supplementary Tables 2-4). One major reason for this very heterogeneous picture is possibly that the physiological state of the venom glands in the used bees could not be synchronized. It was neither possible to keep the solitary bees individuals over several days alive in lab conditions nor possible to establish a milking protocol that the individuals survived. We have to point out that we generated the proteo-transcriptomic datasets in this study predominantly to validate missing venom data for solitary bees to compile a more general core profile of bee venom genes. An in-depth comprehensive discussion of the proteo-transcriptomic data will be published elsewhere while we focus here on selected eleven proteo-transcriptomically supported gene families that represent a core venom profile for secreted bee venom components, consistent with the literature on predominant components of injected bee venom (Figure 2).

The core venom profile of bees

From our proteo-transcriptomics dataset we selected a set of toxins that we propose as a core venom profile for bees. It is important to note that we focus here on bees as an anchor taxon and venom proteins that were predominantly characterized as the major injected hymenopteran venom components in bees based on known toxic venom components together with literature-based toxin candidates. (Moreno & Giralt, 2015; Supplementary Table 1). We restricted our further analysis to these five toxins and six auxiliary venom peptide and protein families mostly with known function, but also including two typical aculeate families with yet unknown biological function (Venom allergen 3/5 and Icarapin). We designated these venom components as a preliminary Core Venom Profile (CVP) in bees, see Figure 2. It is important to stress here that we distinguish in the following between toxins with known immediate toxic function such as neurotoxicity (e.g. Apamin) or cytotoxicity (e.g. Mellitin) and venom proteins consistently present in the crude venom but without a direct (or at least known) toxic function.

To unravel the evolutionary history of the CVP proteins, we analysed the corresponding genomic regions by searching for homologs in 30 high-quality bee genomes and hymenopteran outgroups (sawflies, jewel wasp, ants, paper wasps), spanning 300 million years of evolution including representatives of the phytophagous sawflies (Symphyta) as oldest hymenopteran group. We first built an exon database of venom genes and their flanking genes based on the well-annotated A. mellifera reference genome. We then identified orthologs for each exon in other genomes, which were collected into a secondary, extended database. All genomes were searched using this secondary database and the results were manually inspected to establish completeness and microsynteny, which reflects the arrangement and position of flanking exons of genes around venom protein genes. Our results strongly indicate that venom proteins (as defined above) including enzymatic components are present in highly-stable genomic regions (as multi-copy and single-copy genes). They probably originated in the most recent common ancestor of the earliest diverging hymenopteran lineages
sawflies) and lineages leading to parasitic wasps and later to derived aculeate wasps. Exceptions from this finding are the short, single-copy genes encoding toxin peptides known from bees such as apamin/MCDP and melittin, which are bee or honeybee unique.

**Figure 2: The core venom profile (CVP) for bees.** Functionally predominantly characterized/injected bee venom toxins and protein families (light grey) and additional venom components with unclear functions (dark grey) were identified in our proteo-transcriptomics dataset and by literature mining. Green circles indicate proteo-transcriptome matches, whereas blue circles indicate transcriptome-only candidates and grey circles indicate the absence of a match. Question marks indicate undefined biological function. See Supplementary Table 1 for a more detailed overview of known hymenopteran venom components and Supplementary Table 2-4 for all identified venom components in the present study.

**Apamin is restricted to honeybees and is part of the larger bee unique toxin family Anthophilin1**

Apamin, another dominant *A. mellifera* venom component, is encoded by a three-exon gene located next to a very similar three-exon gene encoding MCDP, with which it shares an exon. This tandem duplication is flanked by MOXD1 homolog 2 and TBC1 domain family member 30. Although the two flanking genes are present and identically arranged in the genomes of all the bees we surveyed, we did not detect the full set of apamin or MCDP exons outside of the genus *Apis* (Figure 3). Genomic analysis indicated that apamin and MCDP (from *Apis*) are restricted to the Apini clade (*Apis* spp.).

There are, however, multiple uncharacterised genes in Bombini and some other non-*Apis* bees that share the microsyntenic position and basic intron-exon structure with apamin and encode peptides that share cysteine scaffold and signal peptide structure with apamin. Of these apamin-like genes we identified six copies in the *Dufourea* genome, in *Nomia* and *Megachile* five copies, *B. terrestris* two copies, while *Osmia bicornis*, *Habropoda* and *Megachile* revealed only a single copy.

In addition to that we identified a novel apamin-like gene in *Apis mellifera* located right next to MCDP gene. This gene is potentially a link between *Apis* apamin forms and those of other bees. We identified no similar
genes or exons of apamin in homologous regions from other hymenopterans or in other parts of their genomes.

Figure 3. Gene arrangement and microsynteny for the apamin family (Anthophilin1). Question marks indicate coding sequences with products of unknown functions. Pseudogenes are symbolized by ψ. The arrows reflect gene orientation. We show here only species for which the genomic sequence in the region with apamin genes is contiguous.

Taken together with the apamin-like sequences we discovered in the core venom profile of Xylocopa and Halictus, our results indicate that apamin and mast-cell-degranulating peptide are members of a highly variable bee-unique family of apamin-like peptides that undergoes independent duplication events in different lineages. We propose here to rename this family Anthophilin1, which reflects its uniqueness to several lineages within bees (Anthophila), see Supplementary Figure 1 and Supplementary File 1 for phylogenetic alignment and tree.

Melittin is restricted to the bee lineage

Melittin is a key pain-inducing venom peptide in A. mellifera (Raghuraman and Chattopadhyay 2007; Chen et al. 2016). The syntenic structure of the A. mellifera genome shows that melittin is encoded by a two-exon single-copy gene located between two four-exon genes, one of which encodes vegetative cell wall protein gp1 while the other remains uncharacterized. Melittin-like sequences in other Apis species (A. dorsata, A. cerana and A. florea) feature a similar microsyntenic arrangement (Figure 4). Other bee species also possess
melittin-like sequences (bombolittin, osmin, collectin, lasioglossin, melectin and macropin (Cerovský et al. 2008; Cerovský et al. 2009; Choo, Lee, Yoon, Je, et al. 2010; Stöcklin et al. 2010; Čujová et al. 2014; Monincová et al. 2014; Kawakami et al. 2017). Microsynteny analysis provided evidence that they are orthologous in at least some species from the genera Colletes, Osmia and Bombus (Figure 4).

Figure 4: Gene arrangement and microsynteny around the melittin sequence. All species for which the genome data allowed for a microsyntenic analysis are shown. Vollenhovia emeryi was not included in other genomic analyses due to its relatively low genome quality. However, it is shown because it was the only one of the eight analysed ant species that features a seemingly related gene in the correct position but with a very different mature sequence. Genes labelled with $\Psi$ in ants and wasps bear little similarity with melittin genes, however, they might be sister genes to the melittin group that underwent severe pseudogenisation.

In B. vosnesenski, the melittin gene has undergone a tandem duplication that is thus far unique to this genus. Some Bombus genomes show assembly gaps in this region, preventing the detection of all exons, but recently published genomes of several Bombus species (Sun et al. 2021) show the same sequence and duplication in the microsyntenic region identified in B. vosnesenski (Figure 4). Although tracing the corresponding genomic region in non-bee aculeataans proved to be difficult because of the relative syntenic instability of it, we
successfully located it in ants and wasps with no detectable homology to melittin. However, one particular ant genome — *Vollenhovia emeryi* (excluded from our main genomic analysis due to the relatively low genome contiguity) — had a superficially similar looking peptide in almost the exact location (Figure 4). That peptide has a proline-rich propeptide resembling that of melittin, however, its mature form is very different. We conclude here that our results support the hypothesis that melittin is indeed restricted to bee lineages, however, its ancestral gene might have had sister genes in ancestors of wasps and ants, see Supplementary Figure 2 and Supplementary File 2 for phylogenetic alignment and tree.

In a previous study it was conjectured that melittin is part of a pan-aculeate family of aculeatoxins that occur in ants, wasps and bees (Robinson et al. 2018). The basis for this grouping is the similarity in signal and propeptides among these proteins, mainly their proline pattern. Other suggested aculeatoxin members include wasp mastoparans and eumenine peptides as well as ant myrmecitotoxins and poneratoxins. Although no experimental or statistical evidence (e.g. model based phylogenetic reconstruction or alignment-based distance matrix) that supports this hypothesis has been published yet, it reflects a plausible evolutionary scenario. For that reason, we test this hypothesis here focused on the evolution of bee venom genes but include representative wasp and ant toxic peptides to trace their possible occurrence in bees: vespid wasps’ mastoparan and ants’ poneratoxins. We did not find any genomic location and sequences that match eumenine, vespakinins and myrmecitoxin genes in our bee data and thus could not include them.

However, we located poneratoxin-like sequences near or within the *atrial natriuretic peptide-converting enzyme* in all ant genomes of our dataset. The number of the genes seem to vary between the species. Although the genomic region appears unstable, we located atrial natriuretic peptide-converting enzymes and some of its nearby and proximal neighboring genes in *Vespa* and *Apis*. No genomic feature in the genome of *Vespa* seems to relate to poneratoxins, while *Apis* features an uncharacterized small 2 exon gene in a similar position. Aligning it with poneratoxins from *H. saltator*, *S. invicta* and *O. brunneus* (extracted from genomic regions) revealed its close similarity to the 5' end of prepro mRNA of the longest of *H. saltator* poneratoxins, see Supplementary File 3. This finding indicates that the original gene got truncated and only the 5' part of it was retained in *Apis*.

We restricted our synteny analysis for mastoparan to the main representative species of bees, ants and wasps and substituted *Polistes* spp. with *Vespa mandarinia* because in both *Polistes* assemblies the mastoparan homologs are located on small scaffolds that prevent the tracing of the synteny. We located mastoparan in the genome of *Vespa mandarinia* and used it as an anchor. The gene is located downstream from *LIM/homeobox protein Lhx9-like*. The genomic region appears stable and we were able to locate it in *Apis mellifera* and *Solenopsis invicta*. However, both species only show annotated pseudogenes (no reading frame seems to resolve in a functional product) in this genomic region roughly in the expected position of
mastoparan. Both pseudogenes have three exons and in both cases one of the reading frames encodes for a proline rich region that could have resembled the mastoparan’s propeptide before the pseudogenisation (see relevant part of Additional File 8). Given the short peptide sequences and the difficulties to conduct meaningful phylogenetic analyses we tested the aculeatoxin hypothesis focused on melittin further, applying a novel, sequence independent machine learning approach (see methods section for the detail).

Figure 5: Structure-function space of small peptidic aculeatan toxins as revealed by the machine learning analysis and their genomic position in respect to each other. Combined data of available verified toxin sequences from Robinson et al., and the present study. A.) Sequences with signal peptide included B.) Only mature peptides. Melittin sequences are outlined with a black circle, different ones on the panel A represent different transcripts with the same mature peptide. For the interactive plots see supplementary Files 7 and 8. C.) Schematic of genomic position of the three groups of hymenopteran toxins. Coloured rectangles represent regions of microsynteny: blue for melittin, green for mastoparan and red for poneratoxins. See text for details.

For the machine learning analyses we used all sequences that Robinson et al presented in their study (kindly provided to us by the author) and all melittin-like toxins known from bees included in our study. We created two datasets – with signal peptides (Supplementary File 5, and without signal peptides Supplementary File 6), the latter dataset includes more sequences due to the fact that bee melittins are mostly known from proteomic studies and therefore only their mature sequence is known (Figure 5). The structure-function space occupied by peptides from wasps and bees revealed in both cases a clear distinction to ants, which is in contradiction to aculeatoxins hypothesis. However, our results reveal a close similarity between bee and wasp
peptides, which is even more apparent when signal peptides were removed (which opposes the argumentation of Robinson et al). Our combined results from the microsyntenic analysis and the machine learning approach indeed indicate that melittin is rather independent and unique in bees but convergently similar to genes in wasps and ants.

**Abundant venom proteins are encoded by more widespread single-copy genes**

Phospholipase A2, hyaluronidase and icarapin are among the most abundant bee venom components (Piek 1986; Danneels et al. 2015; de Graaf et al. 2020). Phospholipase and icarapin are encoded by four-exon single-copy genes, whereas the hyaluronidase gene features nine exons. Dipeptidyl peptidase-4 has a strongly conserved single gene, which was present in all hymenopterans in our dataset, probably due to its enzymatic role in the maturation of some toxins. We found that these protein families were highly conserved and ubiquitously present in the genomes of bees, wasps and ants (*Figure 6*, see also Supplementary Files 8-11 and Supplementary Figures 3-6 for phylogenetic alignments and trees). Our results support the hypothesis that these genes were recruited into venom functions without any associated duplication – similar to the co-option of single-copy genes proposed as the main process of *Nasonia* venom protein evolution (Martinson et al. 2017). In comparison, phospholipase A2 genes in viperid snakes have undergone significant multiplication and diversification before recruitment into the venom system (Dowell et al. 2018; Koludarov et al. 2020).

**Some venom proteins form multi-copy gene families with ancient duplication events**

Interestingly, we find larger duplication and diversification events restricted to protein families that comprise enzymatic or larger proteins and not toxin peptides or proteins. Three venom protein classes in the CVP showed copy number variation across the dataset: venom allergens 3/5, venom acid phosphatases, and serine proteases. For these three multi-copy gene families, we found different copy numbers in different species (*Figure 6*). The genes were found in genomic regions that appear evolutionarily stable and allow for the tracing of homologous regions between species by screening for microsynteny.

Among the venom acid phosphatases, for which we identified 10 subfamilies (*Figure 6*), the largest expansion of genes occurred in subfamily 7, which is exclusively found in parasitoid wasps. In contrast, the large gene expansion in subfamily 10 appears to be an ancient pattern that is found in sawflies (9 genes) and parasitoid wasps (13 genes). In all remaining hymenopterans, including bees, only one or occasionally two to three genes are present.

A similar pattern, though on a smaller scale, was observed for subfamily 5 with ant species having 2-4 copies, while all other hymenopterans (with the exception of *Athalia*) have 1. At the same time subfamily 3 seems to have undergone multiple duplication events in some bee species with up to 10 copies in *Ceratina* and *Bombini*, while other species have 1-2 copies or lost all genes (*meliponini*), see Supplementary File 12 and...
Supplementary Figure 7 for phylogenetic alignment and tree).
and 4. In bees subfamily 6 appears to have been lost (Figure 6; see Supplementary File 13 and Supplementary Figure 8 for phylogenetic alignment and tree).

Venom allergens 3/5 have been identified in many hymenopterans (Hoffman 2006; Piek 1986) and we inferred the presence of five subfamilies in our study. Subfamily 5 appears to have undergone greater diversification in sawflies, parasitoid wasps, ants and the solitary bees Ceratina, Osmia. Only a single member of subfamily 5 is present in the solitary bees Habropoda, Colletes and Nomia while it is striking that eusocial wasps and bees of the family Apidae (Apis, Bombus, Melipona, Frieseomelitta and Eufriesea) appear to have lost all subfamily 5 genes. In contrast, subfamily 1 is only present in parasitoid wasps and ants with a single gene in Euglossa.

Other subfamilies generally have a single copy in every species with subfamily 4 occasionally experiencing duplication. In general, the distribution of genes in the venom allergen family is dynamic, but shows some phylogenetic patterns, see Supplementary File 14 and Supplementary Figure 9 for phylogenetic alignment and tree). Two secapin genes were present in most of the genomes we surveyed, but were notably absent in sawflies (indicating an origin in the stem Apocrita) and wasps of the genus Polistes. This class of peptides displayed intriguing N-terminal sequence variation but strong C-terminal conservation (see Supplementary File 15 and Supplementary Figure 10 for phylogenetic alignment and tree). The location of both genes was also strongly conserved, with one always present between exons of the neurexin-1 gene and the other located near the carbonic anhydrase-related protein 10. Our inability to locate both genes in some species may reflect technical issues relating to genome quality and/or the more general challenges associated with the location of small and highly-variable genes.

Discussion

Gene expansions are restricted to few venom protein families in major taxa

Most of the CVP peptides and proteins are encoded as single-copy genes in the bees and other hymenopterans we screened (Figure 6). Our data supports the hypothesis that gene duplications are a less dominant evolutionary mechanism in the evolution of hymenopteran venom components. This hypothesis was already proposed for parasitoid wasps (Nasonia), in which single gene co-option was identified as the major driver of venom protein evolution (Martinson et al. 2017). However, we have to point out that our results indicate a more distinct pattern in which especially the heavier protein and enzyme components represent the few families of venom proteins in which large gene duplications and expansions have occurred in conserved genomic regions. These expansions are further restricted to particular subfamilies and larger hymenopteran clades (Figure 6). The gene duplications and subsequent gene expansions in venom serine proteases, venom allergens and venom acid phosphatases appear to be ‘simple’ events restricted to the expansion of few genes.
This is in stark contrast to other venomous organisms that have been studied more extensively, such as snakes and cone snails, in which venom genes have evolved rapidly by extensive multiplication, expansion and subsequent deletion (Fry et al. 2009; Jackson et al. 2016; Grandal et al. 2021). We have to point out though that this picture is based on our preselected CVP. This includes the most dominant venom components described, however, bee venoms of course contain more venom proteins. Our results indicate that genes of direct toxins and other (assisting) venom proteins evolve differently in bees. This finding should be tested further in the future using extended venom profiles. Complementary activity studies would be important to address the still undefined biological functions of many venom components, which would in turn support the better interpretation of the evolutionary pattern. A good example here would be venom allergens.

Venom allergens (3/5) show a more heterogeneous pattern of gene duplications, especially in subfamily 5. This subfamily has expanded in parasitoid wasps, leafcutter bees (Megachilidae) and carpenter bees (Xylocopinae), but has been lost in other Apidae lineages. We can only speculate about the original and actual biological function of venom allergens in general because until today the only activities tested are related to immune responses in mice and humans linked to allergic reactions (Elieh Ali Komi et al. 2018). No study so far has addressed the possible bioactivity linked to the ancestral and venom variant’s biological function. However, maybe the strong allergenic activity reflects an original function in sawflies that was linked to the modulation of the immune response of plants, which was later adapted to animal hosts.

Genes representing venom acid phosphatase (APH) subfamily 10 have expanded substantially in parasitoid wasps, ants and sawflies, and the 19 genes of venom APH subfamily 7 are unique to parasitoid wasps. In contrast, bees feature only one or occasionally two genes in any APH family. This may signify that ancestral APHs functioned as pre-digestion factors that allowed the offspring of parasitoid wasps to feed more easily on their host (Danneels et al. 2010). In bees, the retained APHs may be adapted to defensive functions, potentially supported by the origin of APH subfamily 8, which is unique to bees.

Venom serine protease (VSP) subfamilies show more phylogenetic variation, with several subfamilies undergoing secondary loss in bees, wasps and ants. Genes representing VSP subfamilies 8 and 9 have expanded in several ants, and partially also in bees without a clear phylogenetic pattern. However, VSP subfamily 5 appears to have expanded in leaf cutter bees, carpenter bees and bumblebees. VSPs are dual function toxins in bees, triggering the phenoloxidase cascade leading to melanization when injected into insects but acting as spreading factors when injected into mammals, similar to snake VSPs with fibrinogen-degrading activity (Choo, Lee, Yoon, Kim, et al. 2010). We hypothesize that the expansion of VSP genes may be linked to this dual function, achieving more effective defense against insects, arthropods and mammals.

**Bee-specific toxin genes encoding for short peptides**
Bees produce apamin and melittin as dominant venom components, but their genomic origin beyond the honeybee lineage has not been investigated before. One major difference of these toxin peptides to the previously discussed venom components is that the genomic region in which they are encoded appear more dynamic and less conservative. This picture is also reflected by taxon restricted gene duplications. The genomic region containing a tandem repeat of apamin and mast cell degranulating peptide in *Apis* could be identified in other bee genomes based on microsynteny and the characteristic cysteine scaffold. Interestingly, we reveal multiple duplication events that are restricted to single bee lineages. Our conclusion based on this pattern is that apamin and mast cell degranulating peptide are members of a so far unrecognized, highly variable bee-unique peptide family, which we named Anthophilin1. The genes of this family seem to diversify independently in different bee lineages. Whether the duplication events are linked to neofunctionalization or co-option (as dominant venom component) remains to be addressed in future studies. These should include more contiguous genomic data of additional bee lineages. More importantly, however, complementary venom proteomes will be needed to understand in more detail the recruitment and diversification processes of members of this family into bee venom.

We identified melittin in a genomic region with conserved synteny in the genera *Apis, Osmia, Ceratina* and *Bombus* (families Megachilidae and Apidae), with a tandem duplication in bumblebees. The synteny confirmed that melittin-like peptides produced by solitary bees are indeed members of the melittin family. Accordingly, melittin is not unique to *Apis* but originated before the divergence of Megachilid and Apid bees. We did not find a syntenic region or sequences similar to melittin in genomes of bees from the families Andrenidae, Halictidae and Colletidae. If melittin evolved in earlier bee lineages and underwent secondary loss in some families remains unclear from our data due to the lack of high-quality genome assemblies for the early-diverging bee lineages. Our data further indicates that the ant *Vollenhovia* features a gene which vaguely could reflect an ancestral version of melittin, however, the mature sequence looks very different. Nevertheless, we have to note that we cannot rule out the possible origin of melittin in earlier aculeatan lineages until a larger taxon sampling for those and earlier bee lineages are available with high-quality proteo-transcriptome-genome data. Nevertheless, our data supports that melittin is co-opted as a single copy gene as one major component in bees. This contrasts with other venomous lineages such as snakes or cnidarians in which major toxin genes are duplicated to become more dominant (Moran et al. 2008; Margres et al. 2017).

In future studies this finding should be tested further by analysing more proteo-transcriptomic venom profiles linked to genomic data.

**Melittin’s origin in the context of the aculeatoxins hypothesis**

Hymenopterans are well known for their short toxic venom peptides. Similar to bees with melittin and apamin as short peptides, ants and wasps likewise produce peptidic toxins known as mastoparans, poneratoxins,
eumenins, myrmecitotoxins and vespakinins. These were proposed as members of a family of aculeatoxins (Robinson 2018). Testing this hypothesis focused on bees we did not identify any of these peptides directly in clear-cut primary sequence-based genes in bees. However, we located a pseudogenised gene in Apis (as bee representative) and Solenopsis (as ant representative) in the syntenic region of mastoparan in V. mandarinia. In wasps no poneratoxin similar genes are identified, but in Apis a truncated gene with a poneratoxin similar 5' end is located in a related syntenic region (see relevant genomic annotations in Additional Files 8). Based on these results we conclude that our data supports rather a scenario in which ancestors of all of the three toxin families (poneratoxins, mastoparans and melittins) were already present in their modern genomic locations at the very early stages of Aculeatan evolution – latest before the split of vespid wasps from bees and ants. Any duplication of ants' or bees' toxins happened after that independently and likely continues in several lineages. Focused on bee venoms, this implies that melittin is a bee unique toxin. However, our findings do not reject the aculeatoxin hypothesis but reflect a possibly more complex picture in which bee toxins could have evolved independently. It is also plausible that the three groups do not share a common ancestor and have evolved relatively similar propeptide and signal peptide. Their superficial similarity can be explained via shared maturation and expression mechanisms rather than ancestral relation – as was previously proposed (Robinson et al. 2018). The enzymatic cleavage of the mature peptides melittin and apamin by DPP4 is ensured by a single recognition motif. DPP4 cleaves proline-alanine dipeptide – which is exactly the defining feature of ‘aculeatoxins’ propeptide. An unconstrained BLAST search reveals similarities between individual toxins and binding sites of some of the enzymes (most notably, mastoparan to ATPase).

However, our findings could also be in line with the aculeatoxins hypothesis. Further studies using more data could infer that all the three peptide groups share a common ancestral gene at the earlier stages of Hymenopteran evolution. At this stage there's no data that would allow to unequivocally support either of the discussed scenarios. To elucidate this further, the ancestral genes need to be identified and (based on more proteome data) connected with extended genomic data of non-Aculeatan hymenopterans. Nevertheless, there are many other scenarios of origin for these short peptides possible, including a relationship to venom allergen domains (poneratoxins share superficial sequence similarity with some of the groups of those) and de novo emergence from non-coding regions. And of course, if melittin is indeed a related gene to either or both mastoparan and poneratoxin (or other small peptides from hymenopteran venoms) it would fall into a general trend of other CVP genes of bees as discussed below.

**Most bee core venom proteins originated in early hymenopterans**

The pattern that we infer reveals an ancient origin for most of the CVP genes in bees (Figure 7). Most subgroups of its major venom protein gene families show a clear-cut orthology with genes already present in
the earliest hymenopteran lineage (sawflies). Female sawflies use their ovipositor to lay eggs in plants but also co-inject proteins that biochemically interfere with the physiology and immune response of plants to ensure the offspring’s survival, thus resembling a primitive venom system (Wang et al. 2016). The composition of these original hymenopteran venoms has not been studied in detail.

Figure 7. Simplified visualization of the core venom profile (CVP) proteins of bees and their representation in outgroup taxa. The numbers of genomes are shown in brackets after the family names. Genes are colour-coded and feature a colour range for duplicates. Duplications are summarized by numbers. Phylogeny and divergence times are shown as previously described (Peters et al. 2017).

Our results suggest that the core venom genes present in bees today were already present in the early Triassic in ancestors of the symphytan lineage, predating the radiation of apocritans starting more than 200 million years ago (Figure 7) (Peters et al. 2017). The restricted waist of apocritans is needed to manoeuvre the ovipositor in such a way that allows its use for predation, parasitism or defense, and only in aculeate
hymenopterans (ants, bees and wasps) is the retractable ovipositor modified into a stinger used exclusively for venom injection. Our data suggest that the genes encoding the predominant bee venom proteins emerged long before the morphological adaptations of a narrow waist and the stinger in aculeates did, which gave this group its common name – the stinging wasps. The core bee venom profile, including known allergens such as phospholipase A2, icarapin and hyaluronidase, was not only already present in sawflies, but is also still present in a group of bees that has secondarily reduced or lost its stinger (stingless bees, Meliponini).

One general hypothesis might explain this pattern, namely that hymenopteran venom is a biochemical machine tailored around production and facilitation of small and highly abundant toxic peptides. Thus, neither means of injection (stinger or ovipositor) nor physiology on the receiving end (plant, developing egg or adult animal) would majorly affect it, since any required adaptive change would most dramatically affect the nature of those small abundant peptides as main interactors. Other enzymes would either be unaffected at all (if they are facilitators of the main peptide) or change slightly to adapt to biochemical particularities of a new ecological reality (e.g. serine proteases), varying in their effectiveness but never rendered useless due to broadly applicable mode of action. The subject of this study, bees, seem to support this by having little variation in their venom genes, but for the Anthophilin1 and Melittin groups.

**Conclusion**

Our comparative analysis provides the first comprehensive insight into evolution and origin of toxin genes in bees. We found that most genes of the predominant bee venom proteins (reflected by the core venom profile) originated at the base of the hymenopteran tree. This means that they already evolved in the last common ancestor of phytophagous sawflies and apocritan Hymenoptera more than 280 million years ago (Figure 7). Only the short peptides melittin and the newly described herein family Anthophilin1 that is constituted by apamin, apamin-like and MCDP-like genes are unique to bees. In contrast to other venomous organisms such as snakes, the origin of bee venoms appears to be the result of a process that is based predominantly on single gene co-option, in agreement with previous findings in parasitoid wasps. Gene duplications occur, but only in certain (not major toxin) protein families and in only a few hymenopteran lineages, reflecting a generally more diverse pattern of gene origin. As a consequence, our results indicate that toxin and venom protein genes probably evolve under different evolutionary processes. Our results (limited to the core venom profile) establish requirements for future studies to provide insight into the evolution of bee and hymenopteran venoms. These should include more high-quality genomes, especially for early bee lineages but also including a more widespread taxon sampling of other hymenopterans. More importantly, extended proteo-transcriptomics venom profile data are essential. Corresponding to genomes this data improves genome annotations and allows to address appropriately venom protein recruitment processes to differentiate more precisely between gene variants expressed in the venom system and non-
venom related genes. Finally, bioassays for many still unknown venom components are needed to identify functional differences linked to the gene evolution and diverse ecology of hymenopteran species.

**Materials and methods**

**Sample collection and preparation.** All specimens of *X. violacea* (15 individuals) and *H. scabiosae* (17 individuals) were collected in June and July 2019/2020 in the alluvial area of the River Wieseck in Giessen, Germany, with permission from the HNLAG Giessen (IV.2 R28). *A. mellifera* specimens (15 individuals) were obtained from the beehive at the Institute for Insect Biotechnology at Justus-Liebig-University Giessen. The whole venom system of each specimen was dissected on ice under sterile conditions. We extracted crude venom from glands and venom reservoirs by squeezing with forceps in sterile ultrapure water (Thermo Fisher Scientific, Waltham, MA, USA) after prewashing twice to minimize hemolymph contamination. The remaining tissue was preserved in RNAlater (Thermo Fisher Scientific) for subsequent RNA sequencing.

**RNA extraction and venom gland transcriptome sequencing.** RNA extraction, library preparation and short-read genome sequencing were outsourced to Macrogen (Seoul, Korea) for *A. mellifera* and *X. violacea* and to Novogene (Cambridge, UK) for *H. scabiosae*. RNA was extracted with Trizol and after a successful RNASeq quality control the cDNA libraries (150bp, paired end reads) were sequenced using a low input protocol (Illumina Truseq) on an Illumina HiSeq2500 (Macrogen) and Illumina NovaSeq (Novogene). For *H. scabiosae* an in-house ultra-low input protocol was used by Novogene due to very low RNA concentration and quantity. All raw data are submitted to NCBI GenBank PRJNA733472 (SRA entries: SRR14690757, SRR14690758, SRR14690759).

**Transcriptome assembly, protein prediction and annotation.** Venom gland transcriptomes were assembled separately using Oyster River Pipeline v2.2.6 (MacManes 2018) with TPM filtration set to 1. The resulting assemblies were processed using Transdecoder (minimum length 20 amino acids) to predict peptides, and Kallisto v0.46 (Bray et al. 2016) to calculate individual transcript abundance, see Additional Files 1-3. The assembled transcripts and their corresponding longest ORFs (Transdecoder output) were used as local BLAST queries against ToxProt and UniProt (the latter limited to insects only) with an e-value cutoff of $1 \times 10^{-3}$. Any highly abundant (TPM > 100) transcripts without significant matches were manually screened using BLAST, InterPro scan and ProteinPredict online suites to determine the closest characterized homolog. For subsequent venom protein identification, we only included transcripts identified in our proteomic dataset representing proteins secreted in the venom system. To compare subsequently all venom proteins in the three datasets we calculated the percentage of scaled TPMs using the package txtimport on R, the script is available via github ([https://github.com/marivelasque/VenomEvolution.git](https://github.com/marivelasque/VenomEvolution.git)), see Supplementary Tables 2-4. Relatively low BUSCO value for *H. scabiosae* transcriptome could be explained by the fact that of the three species its glands are by far the smallest and thus gland-specific transcripts will have higher
abundance compared to housekeeping transcripts, as well as computed score being lower than actual due to BUSCO methodology inflating the proportion of missing genes as elaborated by Simão et al. (Simão et al. 2015). Due to the protection status of bees and subsequent nature conservation regulations we were limited in the number of specimens and could not dramatically increase the sample size to increase tissue quantities.

Table 1: Overview of the newly-generated transcriptome and proteome datasets for the venom glands of A. mellifera, H. scabiosae and X. violacea.

|                     | A. mellifera | H. scabiosae | X. violacea |
|---------------------|--------------|--------------|-------------|
| Illumina reads      | 26,092,431   | 38,605,338   | 16,077,242  |
| Assembled transcripts | 31,282      | 80,388       | 63,006      |
| BUSCO values        | 97.7%        | 41.6%        | 97.4%       |
| Predicted peptides  | 74,609       | 94,492       | 144,998     |
| MS-supported protein matches | 121        | 211          | 187         |

Proteomic analysis of crude venom. All transcriptome assembly-based predicted ORFs were used as specific databases to identify peptides and proteins detected by mass spectrometry. For the tryptic digestion of the crude venom from H. scabiosae, we dissolved 10 µg of protein in 10 µl 10 M urea containing 0.1% ProteasMax (Promega. Madison, WI, USA). Cysteine residues were reduced with 5 mM DTT (30 min at 50 °C) and modified with 10 mM iodoacetamide (30 min at 24 °C). The reaction was quenched with an excess of cysteine and trypsin was added at a protein:enzyme ratio of 40:1 in 100 µl 25 mM ammonium bicarbonate (Sigma-Aldrich, Taufkirchen, Germany). After incubation for 16 h at 37 °C, the reaction was stopped by adding 1% trifluoroacetic acid (TFA). The sample was purified using a C18-ZipTip (Merck-Millipore, Darmstadt, Germany), dried under vacuum and redissolved in 10 µl 0.1% TFA.

LC-ESI-MS analysis was carried out at 35 °C by loading 1 µg of the sample in 0.1% formic acid (Sigma-Aldrich) onto a 50-cm µPAC C18 column (Pharma Fluidics, Gent, Belgium) mounted on an UltiMate 3000RSLCnano (Thermo Fisher Scientific). Peptides were eluted with a linear gradient of 3–44% acetonitrile over 240 min followed by washing with 72% acetonitrile at a constant flow rate of 300 nl/min. They were then infused via an Advion TriVersa NanoMate (Advion BioSciences, New York, NY, USA) into an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific) in positive-ionization mode with a NanoMate spray voltage of 1.6 kV and a source temperature of 275 °C. Using data-dependent acquisition mode, the instrument performed full MS scans every 3 s over a mass range of m/z 375–1500, with the resolution of the Orbitrap set to 120,000. The RF lens was set to 30%, and auto gain control (AGC) was set to standard with a maximum injection time of 50 ms. In each cycle, the most intense ions (charge states 2–7) above a threshold ion count of 50,000 were selected with an isolation window of 1.6 m/z for higher-energy C-trap dissociation at a normalized collision energy of 30%. Fragment ion spectra were acquired in the linear ion trap with the scan rate set to rapid, the mass range to normal and a maximum injection time of 100 ms. After fragmentation, the selected precursor ions were excluded for 15 s for further fragmentation. The raw proteomics data have
been deposited to PRIDE with the dataset identifier PXD026642.

Prior to shotgun proteomics, the *X. violacea* and *A. mellifera* venom samples were denatured, reduced, and alkylated. Briefly, each sample (~50 μg) was dissolved in 89 μl 100 mM triethylammonium bicarbonate (TEABC), and cysteine residues were reduced by adding 1 μl 1 M DTT (30 min at 60 °C) and modified by adding 10 μl 0.5 M iodoacetamide (incubation for 30 min in the dark). We then added 2 μg trypsin (Promega) in 100 mM TEABC and incubated overnight at 30 °C. The peptides were then purified and concentrated using OMIX Tips C<sub>18</sub> reversed-phase resin (Agilent Technologies, Santa Clara, CA, USA). The peptides were dehydrated in a vacuum centrifuge and analysed by NanoLC-MS/MS. Briefly, the samples were resuspended in 20 μl buffer A (0.1% formic acid) and 1 μl was loaded onto an analytical 25 cm reversed-phase column (Acclaim Pepmap 100 C<sub>18</sub>) with a 75 mm inner diameter (Thermo Fisher Scientific) and separated on the Ultimate 3000 RSLC system coupled via a nano-electrospray source to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). Peptides were separated using a 6–40% gradient of buffer B (80% acetonitrile in 0.1% formic acid) over 123 min at a flow rate of 300 nl/min. Using data-dependent acquisition mode, full MS/MS scans (375–1500 m/z) were performed in the Orbitrap mass analyser (Thermo Fisher Scientific) with a 60,000 resolution at 200 m/z. For the full scans, 3 × 106 ions accumulated within a maximum injection time of 60 ms. The 12 most intense ions with charge states ≥ 2 were sequentially isolated to a target value of 1 × 105 with a maximum injection time of 45 ms and were fragmented by higher-energy collisional dissociation in the collision cell (normalized collision energy 28%) and detected in the Orbitrap mass analyser at a resolution of 30,000. PEAKS Studio v8.5 (Bioinformatics Solutions, Waterloo, ON, Canada) was used to match MS/MS spectra from *X. violacea* and *A. mellifera* venom samples against an in-house database resulting from the annotated transcriptome of each species. Carbamidomethylation was set as a fixed modification, and oxidation of methionine as a variable modification, with a maximum of three missed cleavages for trypsin digestion. Parent and fragment mass error tolerances were set at 5 ppm and 0.015 Da, respectively. A false discovery rate (FDR) of 1% and a unique peptide number ≥ 2 were used to filter out inaccurate proteins. A −10lgP value > 120 was used to estimate whether detected proteins were identified by a sufficient number of reliable peptides. In order to identify more relevant sequences, the Spider algorithm (PEAKS Studio) was used to find additional mutations or to correct sequences. This algorithm corrects the sequences stored in transcriptomic databases with de novo sequences based on MS/MS spectra, allowing the detection of post-translational modifications (PTMs) and mutations. The minimum ion intensity for PTMs and mutations was set to 5%, and the ALC score was set to ≥ 90 for de novo sequences, leading to low precursor mass errors. Transcripts supported by proteomic data were manually filtered by excluding non-venom-related proteins and peptides, such as house-keeping and structural genes (*Supplementary Tables 2-4*). Proteome raw data are submitted to PRIDE (PXD029934, PXD029823).
**Genome mining, sequencing and annotation.** For our comparative analysis of venom toxin proteins across the order Hymenoptera, we made use of the following publicly available genome sequences: *Apis cerana* (GCA001442555.1), *Apis dorsata* (GCA000469605.1), *Apis florea* (GCA000184785.2), *Apis mellifera* (GCA003254395.2), *Athalia rosae* (GCA000344095.2), *Bombus terrestris* (GCA000214255.1) *Bombus vosnesenskii* (GCA011952255.1), *Camponotus floridanus* (GCA0003227725.1), *Ceph us cinctus* (GCA000341935.1), *Ceratina calcarata* (GCA001652005.1), *Dufourea novaean galiae* (GCA001272555.1), *Euglossa dilemma* (GCA002201625.1), *Friseomelitta varia* (GCA011392965.1), *Habropoda laboriosa* (GCA001263275.1), *Harpegnathos saltator* (GCA003227715.1), *Linepithema humile* (GCA000217595.1), *Megachile rotundata* (GCA000220905.1), *Melamp ta genalis* (GCA0001865705.1), *Melipona quad rifasciata* (GCA001276565.1), *Nasonia vitripennis* (GCA009193385.2), *Nomia melanderi* (GCA003710045.1), *Odontomachus brunn eus* (GCA010583005.1), *Ooceraea biroi* (GCA003672135.1), *Osmia bicorns* (GCA004153925.1), *Osmia lignaria* (GCA012274295.1), *Polistes canadensis* (GCA_001313835.1), *Polistes domin uis* (GCA001465965.1), *Solenopsis invicta* (GCA_016802725.1), *Vespa mandarinia* (GCA014083535.1), *Vollenhovia emeryi* (GCA000949405.1), *Wasmannia auropunctata* (GCA000956235.1) (Kapheim et al. 2015; Wallberg et al. 2019; Kapheim et al. 2020). The genomes and annotations of the stingless bees *Tetragobula carbonaria* and *Melipona beechii* will be published as part of another study, but have already been uploaded to NCBI.

We also annotated protein-coding genes based on the genome sequence assembly of *C. gigas* (GCA013123115.1, ASM1312311v1,(Brand et al. 2017) because the annotation presented in the publication is not publicly available. Repeats were soft-masked using RepeatMasker annotations (GCA013123115.1_ASM1312311v1_rm.out) with tabtk, bioawk and seqtk (https://github.com/lh3). We used Funannotate v1.8.1 (Palmer and Stajich 2020) and Uniprot (sprot) for homology-based evidence based on protein sequences from 11 related bee species: *B. impatiens*: GCF000188095.2, *B. terrestris*: GCF000214255.1, *A. mellifera*: GCF003254395.2, *M. quad rifasciata*: GCA001276565.1, *E. mexicana*: GCF001483705.1, *F. varia*: GCA011392965.1, *M. rotundata*: GCF000220905.1, *H. laboriosa*: GCF001263275.1, *D. novaean galiae*: GCF001272555.1, *M. genalis*: GCF011865705.1, *N. melanderi*: GCF003710045.1 (Kapheim et al. 2015; Wallberg et al. 2019). Briefly, funannotate used gene predictions from Genemark-ES, Snap v2006-07-28, glimmerHmm v3.0.4, Augustus v.3.3.3, and CodingQuarry v2.0 together with protein alignments in Evidence Modeler v.1.1.1. Too short, gap-spanning or repeat-overlapping gene models were removed (n = 5446) and tRNA genes were detected (n = 168) with tRNAscan-SE v2.0.6. Genes were functionally annotated using PFAM v33.1, the UniProt database v2018_11, EggNog (eggno g_4.5/hmmd b databases: Arthropoda, Insecta, Hymenoptera, Drosophila), MEROPS v12.0, CAZYmes in dbCAN v7.0, BUSCO Hymenoptera models v3.0.2, Hymenoptera odb9, SignalP v4.1, and
InterProScan5 v81.0. The final annotation contained models for 20,016 protein-coding genes and 168 tRNAs, and was estimated to be 87.1% complete (BUSCO4 v4.1.4). The resulting gene annotation files for *C. gigas*, *E. dilemma*, *M. beecheii* and *T. carbonaria* will be made available as Additional Files 4-7 in the Zenodo archive accompanying this manuscript (10.5281/zenodo.5734574).

Genomic analysis. We traced abundant venom gland transcripts that potentially encoded toxins to homologs in the annotated, highly-continuous publicly-available genomes of bees (and wasps, ants, parasitoid wasps and sawflies as outgroup species) using the online BLAST (Altschul et al. 1990) suite against genomic databases. To identify conserved synteny blocks, we first identified the reciprocal best-match paralogs from hymenopteran all-against-all BLASTP comparisons of the venom genes. Based on the matching sequences, we then extracted exons from the candidate venom genes and their flanking genes. We used those to create BLAST databases to survey the selected genomes using tblastx with an e-value cutoff of 0.01. We then applied filters to select venom genes at least 20 kbp in length (to exclude partial genes) with at least two exons. Where gene annotations were insufficient, we manually re-annotated venom genes where possible, following intron boundaries and using known sequences as templates. We extracted the coding sequences of all complete genes for phylogenetic analysis to establish ortholog groups in addition to their microsyntenic patterns. All resulting annotations are available as part of the Additional Materials (Additional File 8).

Phylogenetic analysis. All toxin transcripts together with toxin genes and their outgroup venom-unrelated homologs (e.g. trypsins and chymotrypsins in case of serine proteases) were arranged by gene family and aligned as translated amino acids using MAFFT (Katoh and Standley 2013) (L-INS-I, 1000 iterations). Name convention was established to differentiate between genomic sequences (first two letters of both genus and species name, followed by the last three digits of a bioinformatic scaffold ID, followed – if applicable - by an abbreviation of a pre-existing gene annotation, followed by letters a to z to differentiate between sequences from the same scaffold); proteo-transcriptomic sequences (names kept the same as generated by transcriptome assemblers); homologues from UniProt and SwissProt databases used to provide outgroups and fill the gaps in sequence space (kept as UniProt or SwissProt IDs). Alignments were manually inspected for overt errors (e.g., proper alignment of the cysteine backbone) and used to construct phylogenetic trees in Exabayes (Aberer et al. 2014) (four parallel runs of four chains each, and 1,000,000 generations). Resulting trees are shown in the Supplementary Figures 1-10, with toxin sequences recovered from *Apis, Halictus* or *Xylocopa* venom marked as red arrows and non-toxic physiological sequences marked with grey arrow.

A novel perspective on relations of short peptide: embedding space analysis. Every year, algorithms improve natural language processing (NLP) tasks such as automated translation or question answering, in particular by feeding large text corpora into Deep Learning (DL) based Language Models (LMs) (Devlin et al. 2018; Peters et al. 2018; Raffel et al. 2019). These advances have been transferred to protein sequences.
by learning to predict masked or missing amino acids using large databases of raw protein sequences as input (Alley et al. 2019; Bepler and Berger 2019; Heinzinger et al. 2019; Bepler and Berger 2021; Rives et al. 2021; Elnaggar et al.). Such methods leverage the wealth of information in exponentially growing but unlabelled protein sequence databases by solely relying on sequential patterns found in the input. Processing the information learned by such protein LMs (pLMs), e.g., by feeding a protein sequence as input to the network and constructing vectors thereof from the activation in the network’s last layers, yields a representation of protein sequences referred to as embeddings (e.g., Fig. 1 in Elnaggar et al.). This way, features learned by the pLM can be transferred to any (prediction) task requiring numerical protein representations (transfer learning) which has already been showcased for various aspects ranging from protein structure (Rao et al. 2020; Weißenow et al. 2021) over protein function (M. Littmann et al. 2021). Further, it was shown that distance in embedding space correlates with protein function (Maria Littmann et al. 2021) and can be used as an orthogonal signal for clustering proteins into functional families (Littmann et al.).

Here, we used the pLM ProtT5-XL-UniRef50 (Elnaggar et al.) (in the following ProtT5) to create fixed-length vector representations for each protein sequence (per-protein embeddings) irrespective of its length. Towards this, we first created individual vector representations for each residue in a protein. In order to derive fixed-length vector representations for single proteins (per-protein embedding) irrespective of a protein’s length, we then averaged over all residue embeddings in a protein (Fig. 1 in Elnaggar et al.). The protein Language Model (pLM) ProtT5 was trained solely on unlabelled protein sequences from BFD (Big Fantastic Database; 2.5 billion sequences including meta-genomic sequences) (Steinegger and Söding 2018; Steinegger et al. 2019) and UniRef50. ProtT5 has been built in analogy to the NLP (Natural Language Processing) T5 (Raffel et al. 2019) ultimately learning some of the constraints of protein sequence. As ProtT5 was only trained on unlabelled protein sequences and no supervised training or fine-tuning was performed, there is no risk of information leakage or overfitting to a certain class or label. As a result, every protein was represented as 1024-dimensional per-protein embeddings. Those high-dimensional representations were projected to 3-d using UMAP (n_neighbors=10, min_dist=0.3, random_state=42, n_components=3) and coulored according to their taxonomic group to allow for visual analysis. Embeddings and 3-d plots were created using the bio_embeddings package (Dallago et al. 2021).

**Author contributions**

BMvR and IK conceived the project. Proteo-transcriptomic and genomic data were analysed by IK, MV and BMvR. Mass spectrometry was conducted by S.D, T.T. and G.L. Machine learning analysis was conducted by M.H. and B.R. New genome and annotation data were provided by E.S., R.G., B.A.H. All authors wrote the manuscript.
Acknowledgements

BMvR and IK thank the German Science Foundation (DFG) for funding this work by a grant to BMvR (DFG RE3454/6-1). BMvR and IK thank Frank Förster for fruitful discussions on bioinformatics. BMvR is grateful to Georg Petschenka and Hermann Falkenhahn for helpful insights and discussion on solitary bees and localities in Giessen. AV and BMvR acknowledge generous funding from the Hessian Ministry of Higher Education, Research, and the Arts for the “Animal Venomics” project embedded in the LOEWE Centre for Translational Biodiversity Genomics. The authors thank Dr. Richard M. Twyman for manuscript editing and in particular Timothy Jackson for comments and edits on the final version. We thank three anonymous reviewers from a previous manuscript version for helpful comments.

Supplementary Material (See also: https://zenodo.org/record/5734574)

Supplementary Table 1. List of described hymenopteran venom proteins and their known functions.

Supplementary Table 2. Proteo-transcriptomically identified venom components in *X. violacea*.

Supplementary Table 3. Proteo-transcriptomically identified venom components in *H. scabiosa*.

Supplementary Table 4. Proteo-transcriptomically identified venom components in *A. mellifera*.

Supplementary Figure 1. Phylogenetic tree of anthophilin1 peptides.

Supplementary Figure 2. Phylogenetic tree of melittin peptides.

Supplementary Figure 3. Phylogenetic tree of phospholipase A2 proteins.

Supplementary Figure 4. Phylogenetic tree of hyaluronidase proteins.

Supplementary Figure 5. Phylogenetic tree of icarapin proteins.

Supplementary Figure 6. Phylogenetic tree of dipeptidyl peptidase 4 proteins.

Supplementary Figure 7. Phylogenetic tree of acid phosphatase proteins.

Supplementary Figure 8. Phylogenetic tree of venom serine protease proteins.

Supplementary Figure 9. Phylogenetic tree of venom allergen proteins.

Supplementary Figure 10. Phylogenetic tree of secapin proteins.

Supplementary File 1. Alignment of anthophilin1 peptides.

Supplementary File 2. Alignment of melittin peptides.

Supplementary File 3. Alignment of ant toxins from genomic regions with *Apis* candidate.

Supplementary File 4. Alignment of “aculeatoxins” with signal peptide.

Supplementary File 5. Alignment of only mature regions of “aculeatoxins”.

Supplementary File 6. Machine learning results for “aculeatoxins” with signal peptide.

Supplementary File 7. Machine learning results for mature region of “aculeatoxins”.

Supplementary File 8. Alignment of phospholipase A2 proteins.

Supplementary File 9. Alignment of hyaluronidase proteins.

Supplementary File 10. Alignment of icarapin proteins.

Supplementary File 11. Alignment of dipeptidyl peptidase 4 proteins.
Supplementary File 12. Alignment of venom acid phosphatase proteins.

Supplementary File 13. Alignment of venom serine protease proteins.

Supplementary File 14. Alignment of venom allergen proteins.

Supplementary File 15. Alignment of secapin proteins.

Additional material (ZENODO 10.5281/zenodo.5734574).

Additional File 1. Assembly file of Xylocopa violacea following ORF prediction by Transdecoder.

Additional File 2. Assembly file of Halictus scabiosa following ORF prediction by Transdecoder.

Additional File 3. Assembly file of Colletes gigas following ORF prediction by Transdecoder.

Additional File 4. Genome annotation of Apis mellifera.

Additional File 5. Genome annotation of Euglossa dilemma.

Additional File 6. Genome annotation of Melipona beecheii.

Additional File 7. Genome annotation of Tetragonula carbonaria.

Additional File 8. Gff files of all toxin gene annotations.

References

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

Alley EC, Khimulya G, Biswas S, AlQuaishi M, Church GM. 2019. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods. 16:1315–1322.

Almeida DD, Viala VL, Nachtigall PG, Broe M, Gibbs HL, Serrano SM de T, Moura-da-Silva AM, Ho PL, Nishiyama-Jr MY, Junqueira-de-Azevedo ILM. 2021. Tracking the recruitment and evolution of snake toxins using the evolutionary context provided by the Bothrops jararaca genome. Proc. Natl. Acad. Sci. U. S. A. 118.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Bepler T, Berger B. 2019. Learning protein sequence embeddings using information from structure. arXiv [cs.LG]. Available from: http://arxiv.org/abs/1902.08661

Bepler T, Berger B. 2021. Learning the protein language: Evolution, structure, and function. Cell Syst 12:654–669.e3.

Branstetter MG, Childers AK, Cox-Foster D, Hopper KR, Kapheim KM, Toth AL, Worley KC. 2018. Genomes of the Hymenoptera. Curr Opin Insect Sci 12:654–669.e3.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Erratum: Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34:888.

Casewell NR, Wüster W, Vonk FJ, Harrison RA, Fry BG. 2013. Complex cocktails: the evolutionary novelty of venoms. Trends Ecol. Evol. 28:219–229.

Cerovský V, Budesinský M, Hovorka O, Cvacka J, Voburka Z, Slaninová J, Borovicková L, Fucík V, Bednárová L, Votuba I, et al. 2009. Lasioglossins: three novel antimicrobial peptides from the venom of the eusocial bee Lasioglossum laticeps (Hymenoptera: Halictidae). Chembiochem 10:2089–2099.

Cerovský V, Hovorka O, Cvacka J, Voburka Z, Bednárová L, Borovicková L, Slaninová J, Fucík V. 2008. Melectin: a novel antimicrobial peptide from the venom of the cleptoparasitic bee Melecta albifrons. Chembiochem 9:2815–2821.

Choo YM, Lee KS, Yoon HJ, Je YH, Lee SW, Sohn HD, Jin BR. 2010. Molecular cloning and antimicrobial activity of bombolitin, a component of bumblebee Bombus ignitus venom. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 156:168–173.

Choo YM, Lee KS, Yoon HJ, Kim BY, Sohn MR, Roh JY, Je YH, Kim NJ, Kim I, Woo SD, et al. 2010. Dual
function of a bee venom serine protease: prophenoloxidase-activating factor in arthropods and fibrin(ogen)olytic enzyme in mammals. PLoS One 5:e10393.

Čujošová S, Bednárová L, Slaninová J, Straka J, Čeřovský V. 2014. Interaction of a novel antimicrobial peptide isolated from the venom of solitary bee Colletes daviesanus with phospholipid vesicles and Escherichia coli cells. J. Pept. Sci. 20:885–895.

Dallago C, Schütze K, Heinzinger M, Olenyi T, Littmann M, Lu AX, Yang KK, Min S, Yoon S, Morton JT, et al. 2021. Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. Curr Protoc 1:e113.

Danneels EL, Rivers DB, de Graaf DC. 2010. Venom proteins of the parasitoid wasp Nasonia vitripennis: recent discovery of an untapped pharmacopoea. Toxins. 2:4.

Danneels EL, Van Vaerenbergh M, Debyser G, Devreese B, de Graaf DC. 2015. Honeybee venom proteome profile of queens and winter bees as determined by a mass spectrometric approach. Toxins. 7:11.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [cs.CL] [Internet]. Available from: http://arxiv.org/abs/1810.04805

Dotimas EM, Hider RC. 1987. Honeybee venom. Bee World 68:51–70.

Dowell NL, Giorgianni MW, Griffin S, Kassner VA, Selegue JE, Sanchez EE, Carroll SB. 2018. Extremely Divergent Haplotypes in Two Toxin Gene Complexes Encode Alternative Venom Types within Rattlesnake Species. Curr. Biol. 28:1016–1026.e4.

Dowell NL, Giorgianni MW, Kassner VA, Selegue JE, Sanchez EE, Carroll SB. 2016. The Deep Origin and Recent Loss of Venom Toxin Genes in Rattlesnakes. Curr. Biol. 26:2434–2445.

Drukewitz SH, Bokelmann L, Undheim EA, von Reumont BM. 2019. Toxins from scratch? Diverse, multimodal gene origins in the predatory robber fly Dasypogan diadema indicate a dynamic venom evolution in dipteran insects. GigaScience. 8(7):giz081.

Drukewitz SH, von Reumont BM. 2019. The Significance of Comparative Genomics in Modern Evolutionary Venomics. Frontiers Ecol. Evol. 7:163.

Dunn CW, Munro C. 2016. Comparative genomics and the diversity of life. Zoologica Scripta. 45:5–13.

Elieh Ali Komi D, Shafaghat F, Zwiener RD. 2018. Immunology of bee venom. Clin. Rev. Allergy Immunol. 54:386–396.

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Learning. Available from: http://dx.doi.org/10.1101/2020.07.12.199554

Elsik CG, Tayan A, Unni DR, Burns GW, Hagen DE. 2018. Hymenoptera Genome Database: Using HymenopteraMine to Enhance Genomic Studies of Hymenopteran Insects. Methods Mol. Biol. 1757:513–556.

Fry BG, Rolands K, Champagne DE, Scheib H, Tyndall JDA, King GF, Nevalainen TJ, Norman JA, Lewis RJ, Norton RS, et al. 2009. The toxicogenomic multiverse: convergent recruitment of proteins into animal venom research. Annu. Rev. Genomics Hum. Genet. 10:483–511.

de Graaf DC, Brochetto Braga MR, de Abreu RM, Blank S, Bridts CH, De Clerck LS, Devreese B, Ebo DG, Ferris TJ, Hagendorens MM, et al. 2020. Standard methods for Apis mellifera venom research. J. Apic. Res.:1–31.

Grandal M, Hoggard M, Neely B, Davis WC, Mari F. 2021. Proteogenomic Assessment of Intraspecific Venom Variability: Molecular Adaptations in the Venom Arsenal of Conus purpurascens. Mol. Cell. Proteomics:100100.

Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20:723.

Hoffman DR. 2006. Hymenoptera venom allergens. Clin. Rev. Allergy Immunol. 30:109–128.

Holford M, Daly M, King GF, Norton RS. 2018. Venoms to the rescue. Science 361:842–844.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11:97–108.

Jackson TNW, Koludarov I. 2020. How the Toxin got its Toxicity. Frontiers in Pharmacology. 11: 574925.

Jackson TNW, Koludarov I, Ali SA, Dobson J, Zdenek CN, Dashevsky D, Op den Brouw B, Masci PP, Nouwens A, Josh P, et al. 2016. Rapid Radiations and the Race to Redundancy: An Investigation of the Evolution of Australian Elapid Snake Venoms. Toxins. 8:11.

Kapheim KM, Jones BM, Pan H, Li C, Harpur BA, Kent CF, Zayed A, Ioannidis P, Waterhouse RM, Kingwell C, et al. 2020. Developmental plasticity shapes social traits and selection in a facultatively eusocial bee. Proc. Natl. Acad. Sci. U. S. A. 117:13615–13625.
Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. 2020. Transformer protein language models are unsupervised

Raghuraman H, Chattopadhyay A. 2007. Melittin: a membrane-active peptide with diverse functions. Biosci. Rep. 27:189–223.
structure learners. Available from: https://openreview.net/pdf?id=fylclEqgvgd
Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnicky CL, Ma J, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U. S. A. 118.
Robinson SD, Mueller A, Clayton D, Starobova H, Hamilton BR, Payne RJ, Vetter I, King GF, Undheim EAB. 2018. A comprehensive portrait of the venom of the giant red bull ant, Myrmecia gulosa, reveals a hyperdiverse hymenopteran toxin gene family. Sci Adv 4:eaau4640.
Santos-Pinto JRA dos, dos Santos-Pinto JR, Perez-Riverol A, Lasa AM, Palma MS. 2018. Diversity of peptidic and proteaceous toxins from social Hymenoptera venoms. Toxicon. 148:172–196.
Schendel V, Rash LD, Jenner RA, Undheim EAB. 2019. The Diversity of Venom: The Importance of Behavior and Venom System Morphology in Understanding Its Ecology and Evolution. Toxins. 11:666.
Sim AD, Wheeler D. 2016. The venom gland transcriptome of the parasitoid wasp Nasonia vitripennis highlights the importance of novel genes in venom function. BMC Genomics 17:571.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.
Steinegger M, Mirdita M, Söding J. 2019. Protein-level assembly increases protein sequence recovery from metagenomic samples manifold. Nat. Methods 16:603–606.
Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. Nature Communications. 9:2453.
Stöcklin R, Favreau P, Thai R, Pflugfelder J, Bulet P, Mebs D. 2010. Structural identification by mass spectrometry of a novel antimicrobial peptide from the venom of the solitary bee Osmia rufa (Hymenoptera: Megachilidae). Toxicon 55:20–27.
Sun C, Huang J, Wang Y, Zhao X, Su L, Thomas GWC, Zhao M, Zhang X, Jungreis I, Kellis M, et al. 2021. Genus-Wide Characterization of Bumblebee Genomes Provides Insights into Their Evolution and Variation in Ecological and Behavioral Traits. Mol. Biol. Evol. 38:486–501.
Vonk FJ, Casewell NR, Henkel CV, Heimberg AM, Jansen HJ, McCleary RJR, Kerkkamp HME, Vos RA, Guerreiro I, Calvete JJ, et al. 2013. 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.
Walker AA. 2020. The evolutionary dynamics of venom toxins made by insects and other animals. Biochem. Soc. Trans. 48:1353–1365.
Walker AA, Robinson SD, Yeates DK, Jin J, Baumann K, Dobson J, Fry BG, King GF. 2018. Entomo-venomics: The evolution, biology and biochemistry of insect venoms. Toxicon 154:15–27.
Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics 20:275.
Wang T, Zhao M, Rotgans BA, Ni G, Dean JFD, Nahrgang HF, Cummins SF. 2016. Proteomic analysis of the venom and venom sac of the woodwasp, Sirex noctilio - Towards understanding its biological impact. Journal of Proteomics 146:195–206.
Weißenow K, Heinzinger M, Rost B. 2021. Protein language model embeddings for fast, accurate, alignment-free protein structure prediction. bioRxiv. Available from: www.biorxiv.org/content/10.1101/2021.07.31.454572
Whittington AC, Mason AJ, Rokyta DR. 2018. A Single Mutation Unlocks Cascading Exaptations in the Origin of a Potent Pitviper Neurotoxin. Mol. Biol. Evol. 35:887–898.
Wong ESW, Belov K. 2012. Venom evolution through gene duplications. Gene 496:1–7.
Yin W, Wang Z-J, Li Q-Y, Jian J-M, Zhou Y, Lu B-Z, Jin L-J, Qiu P-X, Zhang P, Zhu W-B, et al. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. Nat. Commun. 7:13107.
Zancolli G, Casewell NR. 2020. Venom Systems as Models for Studying the Origin and Regulation of Evolutionary Novelties. Mol. Biol. Evol. 37:2777–2790.
