Published on Aug 15, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-d3dc-7499

Identifying candidate accessory domains by mining putative venom protein fusions

Hoping to find proteins that alter physiology in useful ways, we screened venom data sets for toxins fused to domains with additional functionality. We identified candidates, but struggled to infer any novel functions, and none seem well-conserved across venomous species.

Contributors (A-Z)

Adair L. Borges, Rachel J. Dutton, Megan L. Hochstrasser, Elizabeth A. McDaniel, Manon Morin, Dennis A. Sun

Version 3 · Mar 31, 2025

Purpose

Animal venoms are complex mixtures of mainly toxin proteins and peptides that can broadly interfere with host physiology. While toxins are the most well-characterized proteins in venoms, there is evidence that molecules facilitating toxin activity are present as well. We decided to search for toxin-like or toxin proteins with extra accessory domains with interesting functions (especially toxin-facilitating functions). Based on evolutionary precedent in bacteria, we thought we might find novel

accessory proteins/domains by searching for uncharacterized domains fused to known toxins.

We developed a computational strategy to screen for potential gene fusion events and identified 1,225 possible candidates across 145 species. The accessory portions of the identified proteins are not well-conserved nor broadly conserved across venomous species. We tried sequence-based analysis (BLASTp and HMM) but ran into issues with annotating sequences of potential accessory domains. Most had only a very general or low-confidence predicted function. We are working to refine functional annotation as a whole, and this project further emphasized the need for new or improved solutions to functional prediction.

While we don't plan to follow up on this work, we're sharing our results in case they may be useful for those studying venom biology or perhaps to the functional annotation community.

- All associated **code** and **metadata** are available in this GitHub repository.
- You can access **data** from this pub on Zenodo, including FASTA files containing representative sequences of the clustered toxin reference database, our custom venom and tick toxin data set, and the accessory sequences of the toxin outliers.

We've put this effort on ice! ☒

#DeadEnd

We didn't find any intriguing "accessory" sequences in our search for longer-than-typical toxins. We couldn't functionally annotate most of our hits and we didn't have a clear path forward to investigate others.

Learn more about the Icebox and the different reasons we ice projects.

Background and goals

Venoms are secretions that an animal produces in a specialized gland that are delivered to a target animal through a wound — they contain molecules that disrupt normal physiology to assist feeding or defense **[1]**. A single venom can contain hundreds of different toxins, and the ability to produce venom has evolved independently more than 100 times across the tree of life **[2]**. There has been a lot of convergent evolution across venom toxin proteins, which target key aspects of host physiology (neurological functions, the cardiovascular system, homeostasis, etc.) **[3] [4]**.

Additional venom proteins are known to facilitate the action of toxins. For instance, hyaluronidases are found in multiple venoms across species. Their hyaluronic acid hydrolysis activity is described as a “spreading agent,” facilitating toxin diffusion through the prey’s skin layers **[5]**. While a great deal of work has focused on venom toxins, little is known about other functional molecules, particularly those with toxin-facilitating functions, in venoms.

For bacterial botulinum neurotoxins produced by *Clostridium* species, non-toxin proteins are known to be essential for the toxins’ activity. The neurotoxin-associated proteins NTNH, HA, and ORFX play important roles to help the botulinum neurotoxin survive the acidic environment of the digestive tract (NTNH) and cross the intestinal barrier (HA). These neurotoxin-associated proteins have emerged from either toxin gene duplication followed by divergence (NTNH) or gene fusion between clostridial toxins and pre-existing HA and ORFX genes, which played other roles in *Clostridium* **[6]**.

By combining old parts to make something new, gene fusion is an efficient mechanism of evolutionary innovation, generating proteins with complex structures and functions. We sought to understand whether similar fusion scenarios between a toxin and another accessory protein have happened in animal venoms, and whether such gene fusions may have evolved convergently in multiple venoms.

There are no set and simple methods to identify gene fusion events and they are rarely investigated at the protein level, but rather at the gene or transcript level. We anticipate that any protein resulting from a fusion event between a toxin and another protein will emerge as a length outlier among its toxin homologs. We screened 145 species’ venom transcriptome public data sets (that provide the protein content information of

venoms), to identify possible venom proteins that result from the association of a toxin domain and an accessory domain as identified length outliers in their toxin category. For such proteins, we further investigate the accessory sequences to determine (i) whether they are found broadly across different venoms, (ii) whether they are associated with multiple toxin types, and (iii) whether we can infer their functions.

While focused on gene fusion involving toxins in venoms, this work provides a general framework for screening evolutionary innovation through gene fusion.

> **SHOW ME THE DATA**: Access FASTA files containing representative sequences of the clustered toxin reference database, our custom venom and tick toxin data set, and the accessory sequences of the toxin outliers on Zenodo (DOI: 10.5281/zenodo.8208984).

The approach

The overarching goal of this work is to determine whether proteins resulting from the fusion of a toxin and another protein with an accessory function are found in venoms and whether any accessory functions are convergent across species.

As we can infer predicted protein sequences from transcriptomes, we started by generating a custom data set of venom proteins from available venom gland or salivary gland transcriptomes (145 species) (Figure 1). In parallel, we clustered the toxin reference database, a curated database that is part of UniProt's animal toxin annotation project (Tox-Prot) [7], and further refer to it as the "Venom proteins and toxins" database. We next compared our custom data set to this database to find sequence-based similarities between our custom data set proteins and the reference toxins. We thereby generated groups of related proteins where each group contains at least one protein from our custom data set and is characterized by a single toxin of the toxin reference database. We then identified proteins emerging as length outliers within each group independently. We continued the analysis by extracting the "non-toxin" sequences (the putative accessory sequences) of each outlier. We further clustered accessory sequences based on their sequence homology. We isolated a representative sequence for each accessory cluster and conducted Pfam annotation in an attempt to identify accessory domains. Finally, we investigated whether these

accessory sequences are broadly shared or specific to venomous species or known toxins.

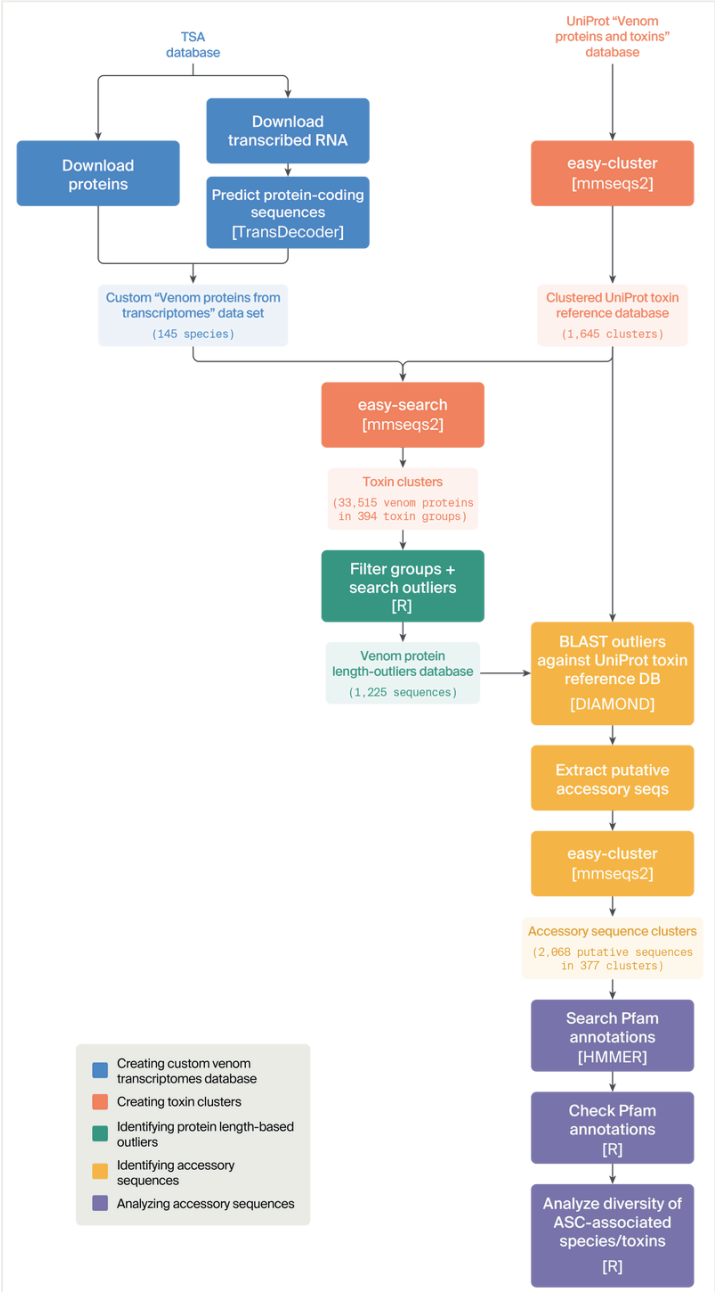

Figure 1

Overview of computational workflow.

Read about our methods in detail below or skip straight to "The results."

More **detailed methodological information** and **code** are available in the GitHub repository (DOI: 10.5281/zenodo.8299305) associated with this work.

Creating the custom venom transcriptome data set

We first downloaded proteins or transcribed RNA data sets from venom glands or salivary glands. We downloaded protein accessions or transcribed RNA files from transcriptome shotgun assemblies (TSA) data sets for transcriptomes that are publicly available and listed them in the metadata file, "SRA_TSA_venom_gland_accessions.csv."

Many species were associated with accessions with already-called predicted proteins. For each of these species, we generated a FASTA file that contains all proteins in the TSA, titled by the species name. Other species only had accessions with transcriptome data without called proteins. After downloading the transcribed sequences for each of these species' individual FASTA files, we used TransDecoder (version 5.7.0) **[8]** to obtain predicted ORFs/proteins.

Overall, we obtained protein sequences for the venom glands of 124 species and from the salivary glands of 21 tick species (Figure 2). We generated Figure 2 using TimeTree **[9]** and the phylogenetic tree viewer software FigTree (version 2018-11-25 - v1.4.4). We pooled all the protein FASTA files into a single FASTA file that represents our custom "venom proteins from transcriptomes" data set, the starting point for the rest of the analysis.

Figure 2

Phylogenetic tree of the species included in this work.

Species have been collapsed into larger groups of species categories. We generated this tree using TimeTree and available information regarding the different species. We further edited the tree to collapse species into groups using FigTree. Because no genetic information was available for some species, this tree is actually missing two groups: crustacean (one species: *Xibalbanus tulumensis*) and bloodworm (one species: *Glycera tridactyla*).

Searching the custom venom proteins against the UniProt venom proteins and toxins reference database

In UniProt, the animal toxin annotation project contains a manually curated database of proteins and toxins from various venoms. We used this curated database as our reference toxin database and refer to it as the “venom proteins and toxins” database in this work. As we expect this database to contain identical or similar homologs, we clustered it using `mmseqs2 easy-cluster` (version 14.7e284) **[10][11][12]**, and kept only

one representative sequence per cluster to remove redundancy. Out of the original 7736 sequences of the “venom proteins and toxins” database, we obtained 1645 clusters and thus 1645 reference sequences. We then used `mmseqs2 createdb` to generate a toxin reference database from these sequences.

We further used `mmseqs2 easy-search` **[10][11][12]** to search this query toxin reference database against the target database: our custom venom proteins data set. During this search, we aligned each venom protein to each representative sequence of the toxin reference database to find possible matches and generate alignment scores.

Ultimately, 33,515 proteins from our custom venom proteins data set got at least one hit in the toxin reference database.

Identifying protein length outliers in toxin groups

As each protein from the custom venom protein data set may have yielded multiple matches with the reference database, we kept only one hit per venom protein, corresponding to the hit with the lowest E-value. Then we defined a “toxin group” as any ensemble of all the venom proteins that hit the same representative toxin from the toxin reference database. Consequently, each toxin group contains at least one venom protein and is characterized by a reference toxin protein. Altogether, we generated 394 groups.

We filtered out any group that contained less than five venom protein sequences, which left us with 236 groups. For each venom protein, we calculated its length ratio compared to the group-associated reference toxin. We further used this length ratio as the metric to identify length outliers within each cluster.

In each cluster, an outlier is defined as any venom protein that meets both of these criteria:

- Soft outlier criteria: Length-ratio greater than Q3 + 1.5*IQ (Q3: 3rd quartile of the group length-ratios distribution, IQ: the interquartile range)
- Minimal ratio criteria: Length-ratio greater than 1.5

We identified 1,225 outlier sequences out of 33,207 sequences, which you can find in the file, "Venomproteins_ticks_toxins_outliers_June2023.csv." We isolated the amino acid sequences of these outliers to identify accessory sequences, as described below.

Extracting and clustering outlier accessory sequences

We next identified and extracted the accessory sequences from the identified outlier sequences. Accessory sequences are defined as any sequence that doesn't match a reference representative sequence from the toxin reference database. We used DIAMOND (version 2.1.6) **[13]** to create DIAMOND individual databases of the 1,225 outlier sequences and of the toxin reference database, and to further perform a protein BLAST (BLASTp) of the outlier sequences against the reference toxin sequences.

Our strategy was to extract putative accessory sequences for every hit obtained for each outlier sequence and generate a multi-FASTA file. For every DIAMOND hit against the toxin reference sequences, we removed the portion of the venom protein that aligned to the reference hit, leaving the non-aligned portion. This process could end up creating different versions of similar or identical putative accessory portions of proteins based on what aligned to each toxin reference hit.

Once identified, we clustered all accessory sequences based on sequence homology, generating accessory sequence clusters from which we extracted a representative sequence and annotated using the whole Pfam HMMs database.

We generated 2,566 accessory sequences from the 1,225 outlier sequences (as we chose to keep any possible accessory sequences per outlier according to the DIAMOND BLAST result). We further clustered these accessory sequences into accessory sequence clusters (ASCs) and extracted one representative sequence for each cluster using `mmseqs2 easy-cluster` (version 14.7e284) **[10][11][12]**. We obtained 371 accessory sequence clusters.

Analyzing accessory sequence clusters

We concluded this work by analyzing the ASCs we generated from our custom venom protein data set. We sought to investigate the potential functions associated with accessory sequences, as well as their distribution and conservation across species and associated toxins.

Filtering out accessory sequence clusters annotated with toxin-associated Pfam

We annotated the representative sequences from the 371 ASCs against the Pfam.hmm database **[14]**. 288 sequences were assigned with one Pfam annotation. We decided to consider any annotation with an E-value greater than 10^{-5} as having no Pfam annotation, as this is a low-confidence annotation.

We identified a list of 481 Pfam annotations that are associated with toxins by combining the lists of Pfams associated with the proteins of the “venom proteins and toxins” database, and the curated list provided in a useful reference paper **[15]**. According to that list, 80 representative sequences of ASCs were annotated with a toxin-associated Pfam. We filtered out these sequences and their associated ASCs for the rest of the analysis.

Investigating the species and toxin diversity within each accessory sequence clusters

For each accessory sequence cluster (ASC), we determined the number of species the clustered accessory sequences originated from as well as the number of different toxins the clustered accessory sequences have been associated with.

> All **code** and **metadata** we generated and used for the pub are available in this GitHub repository.

Additional methods

We used ChatGPT to write some code and add comments to our code.

The results

> **SHOW ME THE DATA**: Access FASTA files containing representative sequences of the clustered toxin reference database, our custom venom and tick toxin data set, and the accessory sequences of the toxin outliers on Zenodo (DOI: 10.5281/zenodo.8208984).

Length-based outlier search identifies venom toxin proteins that may have acquired accessory sequences

Our strategy to identify possible fusion events between toxins and other genes is to screen multiple venom transcriptome data sets for proteins that contain a known toxin component and stand out as length outliers among their homologs.

To start this search, we compiled protein sequences from 124 species' venom glands and from the salivary glands of 21 tick species, collecting them in a custom "venom proteins from transcriptomes" data set. In parallel, we generated a reference toxin database by clustering the "Venom proteins and toxins" database from UniProt's animal toxin annotation project (Tox-Prot) and keeping one single representative toxin sequence per cluster. We used our custom protein data set from venoms to query the reference toxin database. This allowed us to identify the proteins with known toxin sequences in venoms and infer the types of toxin present. We identified 33,515 venom toxin proteins in our custom data set and sorted them into 394 toxin groups based on homology to a unique toxin from the reference database for each group. The size of these groups ranged from one protein (68 groups) to 3,786 proteins (one group). Since we wanted to find group outliers, we omitted any group that contained fewer than five proteins and kept 236 groups.

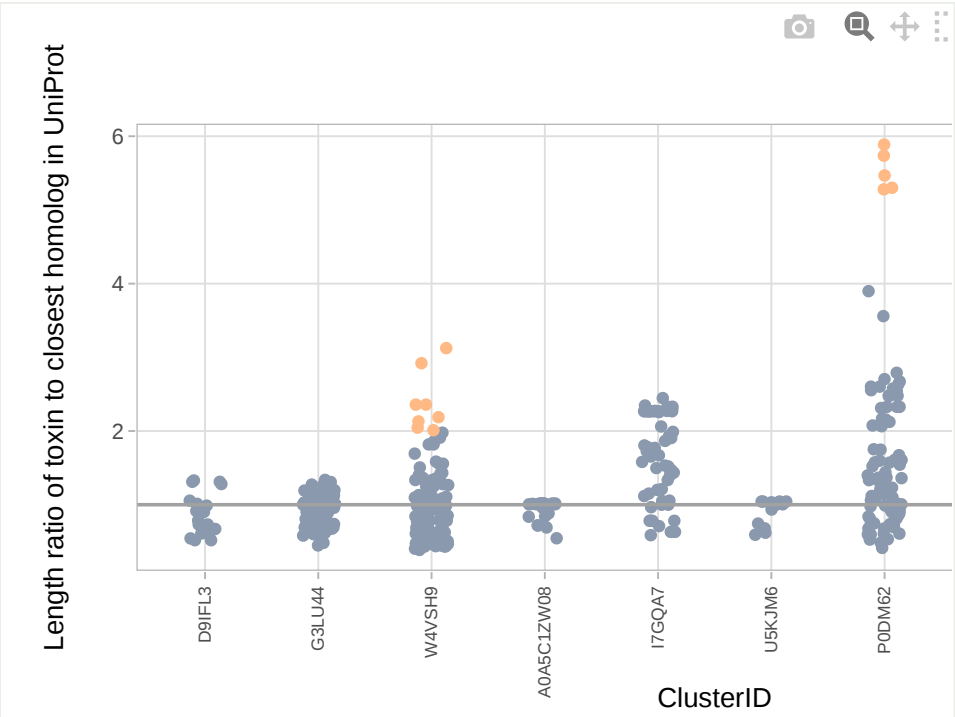

Figure 3

Size ratio distribution of proteins in toxin clusters.

Orange points represent length outliers. Hover over a point to see the protein's ID in our data set. View a static version or open the interactive version in a new tab.

To identify length-based outliers, we considered the length ratio of each protein to its group-associated reference toxin (see "Identifying protein length outliers in toxin groups" for how we define outliers). Figure 3 depicts the size distribution of proteins across 10 different groups, four of which contained outliers. We identified 1,225 outliers that come from 123 different groups, are part of around 30 different categories of toxins, and involve 60 different species (Figure 4). 71% of the outliers were among tick species, with patterns that seem consistent across multiple species, suggesting the possibility that ticks have a higher proportion of gene fusions than other venomous species.

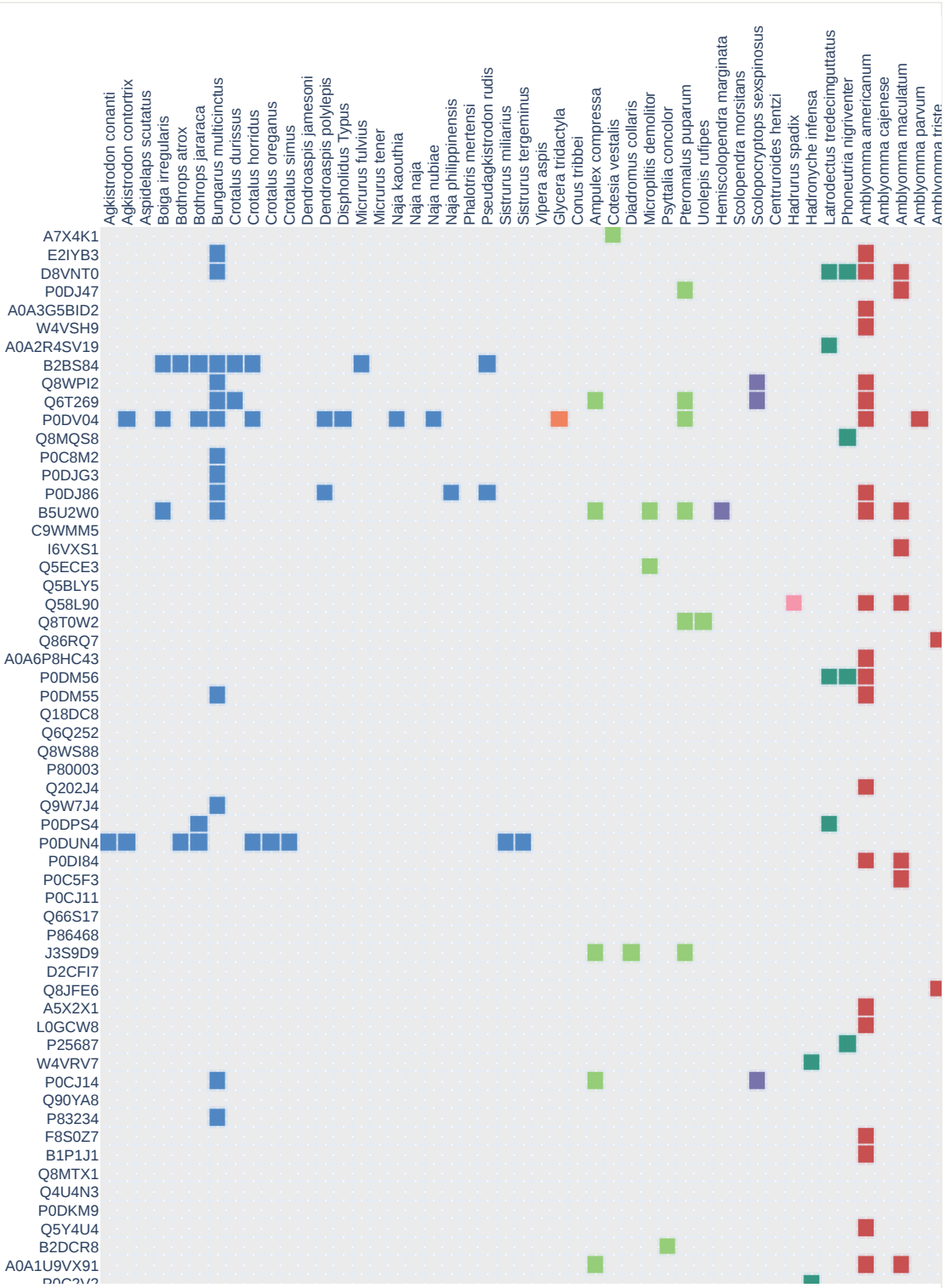

Figure 4

Map of toxin outliers and associated species.

This interactive figure shows the outliers we found in different species (x-axis) within each group where outliers have been found (y-axis). Each group is identified by its associated reference toxin and we've colored data points by taxon. Hover over a point to see each protein sequence's taxon, species,

functional category, and UniProt ID. View a static version or open the interactive version in a new tab.

Altogether, we have been able to identify toxins that are significantly longer than most of their homologs. All these outliers are candidates for further investigation to learn about the non-toxin portion of their sequence, which we refer to as the "accessory sequence."

Pfam annotation of accessory sequences offers limited information about potential function

After BLASTing our outlier sequences against UniProt's curated toxin database, we extracted each outlier's associated accessory sequences (the protein sequence minus the segment or segments that align with a toxin). Consequently, we sometimes obtained multiple accessory sequences for a given outlier. Eventually, we generated 2566 accessory sequences and used sequence homology to cluster them into 371 accessory sequence clusters, further referred to as ASCs (see "Extracting and clustering outlier accessory sequences" for more details). Cluster size ranges from one accessory sequence (110 clusters) to 170 accessory sequences (one cluster) (Figure 5).

Since the goal of this work was to find novel protein functions that modulate host physiology, we next sought to understand what each of our putative toxin accessory domains does in the hope that something useful or intriguing might pop out at us. We annotated one representative accessory sequence for each cluster using HMMER (version 3.3.2) and the whole Pfam database **[14]**. 125 representative sequences obtained a Pfam annotation with an E-value lower than 10^{-5}, our chosen threshold to select for real matches with confidence. Relying on the UniProt-curated "Venom proteins and toxins" database and additional annotations **[15]**, we generated a list of Pfam annotations that are known to be associated with toxins. Using this list, we further identified which accessory sequence clusters are annotated as toxin-associated, non-toxin-associated, or had no Pfam annotation (Figure 5). Overall, 80 ASCs were annotated as toxin-associated, suggesting that our workflow did not fully remove all toxin sequences, either because a single toxin can contain multiple toxin domains and we missed some or that the "toxin" portion of some sequences extends

beyond the area that aligned with the reference sequence. We omitted these ASCs for the rest of the analysis.

291 accessory sequence clusters had either no Pfam annotation or a non-toxin annotation. 45 of them have a non-toxin annotation, and these span 30 different Pfam annotation categories (see these data in "Table-2_Summary_metrics_accessory_sequence_clusters_.csv" on GitHub). Most Pfam annotations are "domains" (collections of related sequence regions that form a distinct structural unit), some are "families" (collections of related sequence regions that may contain one or more domains, but where there is insufficient evidence to support subdivision), and some are repeats.

While multiple annotations point to domains associated with signaling, signal transduction, or protein-protein interactions, these annotations are pretty general and hard to interpret, especially as they indicate functions that can also be found in toxins.

Overall, it is challenging to obtain reliable information about the potential functions carried out by accessory sequences through Pfam annotation. Another solution could be to perform a protein BLAST against the non-redundant protein database **[16]**. This could identify sequence matches between the accessory sequences and annotated non-toxin proteins, informing us of the nature and function of the accessory sequence. However, our putative accessory sequences are portions of proteins that are very likely to be present in the non-redundant database (or homologs from closely related species), so our queries would likely just align with those full-length proteins.

Altogether, current sequence-based annotation had limited success in identifying reliable annotated domains and functions for the accessory sequences, making it challenging to conclude that these toxin length outliers emerge from an actual gene fusion.

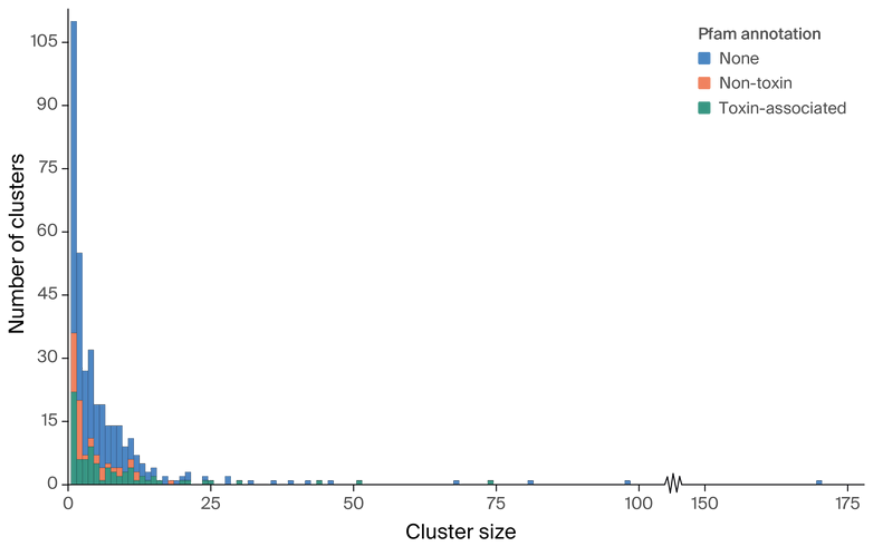

Figure 5

Distribution of putative accessory sequence cluster length.

This histogram shows how many clusters (y-axis) contain a given number of accessory sequences (x-axis).

Cross-species convergence of accessory sequences is limited to closely related species

The functional annotations of the putative accessory sequences didn't contain any strong clues, but we thought we might zero in on exciting functionality by determining which of these sequences are present across venoms from many species and types of toxins (and therefore likely important). Such convergent evolution of accessory sequences would suggest an evolutionary benefit, pointing us to globally important accessory functions in venoms.

For 224 of our 291 clusters, we noticed that the accessory sequence derived from a single outlier sequence (but not from the same outlier for all clusters). This is because we decided to keep all the possible accessory sequences for each outlier based on alignment results, ensuring that we investigated all possible accessory sequences and increasing our chances of identifying something of interest. For some outliers, this led us to include multiple very similar accessory sequences, some differing by only a few

amino acids. Each of these 224 clusters is inherently associated with a single species and a single toxin, and they don't provide any workable information regarding convergence of accessory sequences so we removed them for the rest of the analysis. However, the fact that 77% of the accessory sequence clusters are associated with individual outlier proteins suggests that convergence of accessory sequence is likely uncommon.

We calculated the number of different species represented in each of the remaining 67 clusters. The number of different species identified per cluster ranges from one species (20 clusters) to 10 species (one cluster) (Figure 6 and "Table-1_Accessory_clusters_non_Toxin_Pfam_information.xlsx" on GitHub). 16 clusters are associated with more than three species. When multiple species are present, they are part of the same group. For instance, in 11 clusters with accessory sequences from more than three species, all the species are ticks. In four clusters, the only species represented are snakes, sometimes from the same family (e.g., cluster 61 contains sequences from eight species across four genera of Viperidae snakes).

Overall, we found that accessory sequences that cluster together are usually from the same species or group of species, suggesting some low amount of convergence of accessory sequences across closely related species, and very little or no convergence across a broader range of phylogenetically distant species.

Only one cluster (cluster 19) contained two outliers from species from different groups: the spider species *Latrodectus tredecimguttatus* and the snake species *Bungarus multicinctus*. These outlier sequences are associated with the same toxin, a galactose-specific lectin called nattectin. Protein BLAST of the accessory sequence yields the best matches with other lectin proteins in spiders (macrophage mannose receptor or secretory phospholipase A2 receptor) with best identity matches of 59% and 52%. This suggests that the whole protein is part of the lectin family for both *L. tredecimguttatus* and *B. multicinctus*, including the portion we thought could have been an accessory sequence. Moreover, the transcriptome data we obtained for *B. multicinctus* is the only data set that wasn't restricted to the venom gland and contained protein information from other tissues. As lectins represent a major protein family and they're present in many tissues and not restricted to toxin activity, it is possible that non-toxin lectins from this species have been incorrectly assigned to a toxin group.

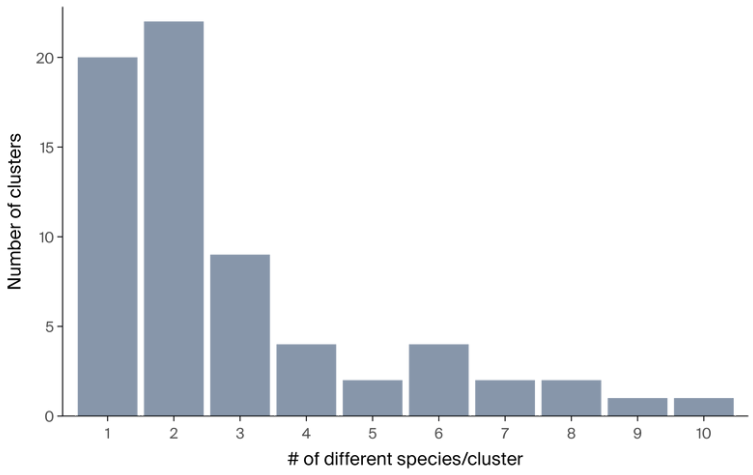

Figure 6

Summary of characteristic metrics of accessory sequence clusters.

Histogram of the number of different species per accessory sequence cluster.

While accessory sequences appear to be species-specific, this doesn't exclude the possibility that they can be shared across different toxins. To test this hypothesis, we looked at each of the 67 ASCs and determined how many different toxins are associated with each accessory sequence (see "Table-1_Accessory_clusters_non_Toxin_Pfam_information.xlsx" on GitHub). Strikingly, most of the accessory sequence clusters are associated with single toxins (61 clusters). The highest number of different toxins found in a cluster is just two (six clusters). For these clusters, the identified toxins are from the same toxin family (clusters 48 and 4: veficolin, cluster 56: protease inhibitor, clusters 32 and 60: thrombin-like enzyme, cluster 59: venom serine protease). Altogether, this undeniably shows the toxin-specificity of accessory sequences.

Overall, by investigating the diversity of species and toxins found in each ASC, we've demonstrated that there is little convergence of toxin-associated accessory sequences across species and toxins.

Key takeaways

In this work, we sought to identify conserved venom proteins with interesting toxin-facilitating functions. Our strategy relied on screening venom transcriptomes to identify toxin proteins potentially fused with extra domains, and then looking for patterns in the distribution or conservation of these accessory domains across a broad range of species.

While our approach identified multiple candidates, analysis of these accessory sequences showed that they are poorly conserved across distant venomous species and highly specific to the toxin they are associated with, refuting the hypothesis that venomous species share fused domains that facilitate the action of toxins.

Our ability to interpret the potential function of these candidate accessory sequences was limited. Currently available sequence-based annotation methods were insufficient for our purposes, highlighting the need for alternative or improved annotation tools.

Finally, ticks stood out as the organisms with the most length outliers, and their accessory sequences appear to be conserved across tick species. This strongly suggests that ticks have evolved specific sets of toxins that are divergent from other venomous species and that could carry out different or modified functions.

Next steps

We found poor conservation of toxin-associated accessory sequences, and couldn't confidently identify specific functions associated with these sequences. Because of current limitations in computationally predicting protein function, we are not pursuing this project further. However, others interested in this subject might consider a couple of directions to take the project one step further.

One obvious direction would be to develop better approaches to characterize the functions associated with identified accessory sequences. An alternative approach to sequence-based annotation that would be particularly relevant in this field would be structural analysis of the toxin and their accessory sequences, as protein structure is crucial for their proper function, especially for proteins that are known to interact with other proteins or molecules. This could reveal whether the accessory portion of the protein has a similar structure to other proteins with known functions. Comparing

structures of length outliers to structures of homologs without extensions could also show how the additional sequence affects the overall structure of the toxin and might hint at how it impacts activity.

Because our data sets were mostly associated with snakes, the most-studied venomous animals, this work is biased toward snake species. Venomous species are incredibly abundant on Earth, so it could be informative to extend this analysis to less studied venoms as more sequencing data becomes available.

Finally, our project assumed that some toxin-facilitating functions would come from additional amino acids fused to toxins that can also exist in a standalone context. Our findings don't refute the possibility that broadly conserved facilitating functions could exist without being cleanly segmented within toxin sequences. Amino acids that confer additional function may be indissociable from the toxin sequence. Identifying whether some toxins have more subtly evolved added functionality would require an in-depth structural analysis of toxins and analysis of protein-protein interactions between toxins and their potential targets.

References

1. Fry BG, Roelants K, Champagne DE, Scheib H, Tyndall JDA, King GF, Nevalainen TJ, Norman JA, Lewis RJ, Norton RS, Renjifo C, de la Vega RCR. (2009). The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms. https://doi.org/10.1146/annurev.genom.9.081307.164356

2. 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. https://doi.org/10.3390/toxins11110666

3. Surm JM, Moran Y. (2021). Insights into how development and life-history dynamics shape the evolution of venom. https://doi.org/10.1186/s13227-020-00171-w

4. Zancolli G, Reijnders M, Waterhouse RM, Robinson-Rechavi M. (2021). Convergent evolution of venom gland transcriptomes across Metazoa.

https://doi.org/10.1101/2021.07.04.451048

5 Girish KS, Jagadeesha DK, Rajeev KB, Kemparaju K. (2002). https://doi.org/10.1023/a:1020651607164

6 Mansfield MJ, Doxey AC. (2018). Genomic insights into the evolution and ecology of botulinum neurotoxins. https://doi.org/10.1093/femspd/fty040

7 Jungo F, Bairoch A. (2005). Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. https://doi.org/10.1016/j.toxicon.2004.10.018

8 Haas, BJ. https://github.com/TransDecoder/TransDecoder

9 Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. https://doi.org/10.1093/molbev/msac174

10 Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988

11 Steinegger M, Söding J. (2018). Clustering huge protein sequence sets in linear time. https://doi.org/10.1038/s41467-018-04964-5

12 Mirdita M, Steinegger M, Söding J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. https://doi.org/10.1093/bioinformatics/bty1057

13 Buchfink B, Reuter K, Drost H-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. https://doi.org/10.1038/s41592-021-01101-x

14 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. (2020). Pfam: The protein families database in 2021. https://doi.org/10.1093/nar/gkaa913

15 Negi SS, Schein CH, Ladics GS, Mirsky H, Chang P, Rascle J-B, Kough J, Sterck L, Papineni S, Jez JM, Pereira Mouriès L, Braun W. (2017). Functional classification of protein toxins as a basis for bioinformatic screening. https://doi.org/10.1038/s41598-017-13957-1

16 Pruitt KD. (2004). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. https://doi.org/10.1093/nar/gki025