Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

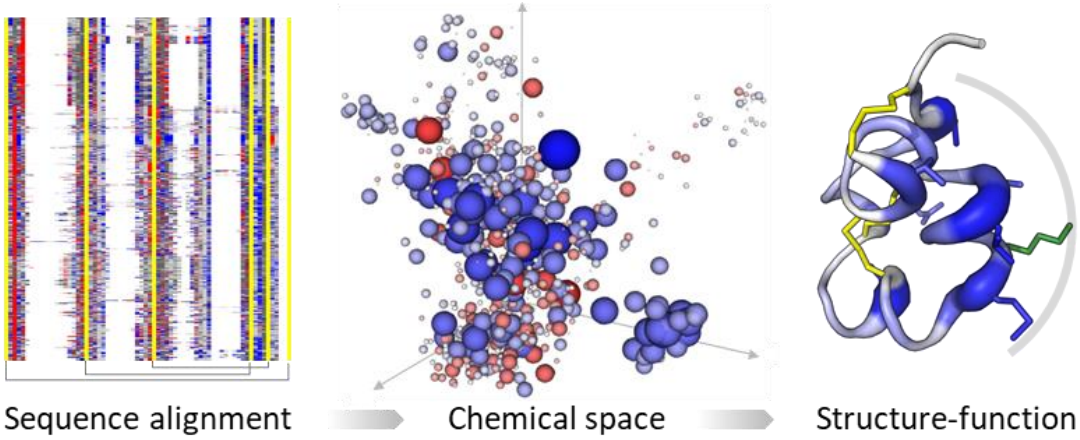
Mapping the chemical and sequence space of the ShKT superfamily

Thomas Shafee[1,2#], Michela L. Mitchell[3,4,5,6], Raymond S. Norton[3]

[1] Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria, 3086, Australia
[2] Department of Animal, Plant, and Soil Science, AgriBio, La Trobe University, Melbourne, Victoria, 3086, Australia
[3] Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria, 3052, Australia
[4] Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia
[5] Marine Invertebrates, Museum Victoria, GPO Box 666, Melbourne, Vic, 3001, Australia
[6] Biodiversity & Geosciences, Queensland Museum, PO Box 3300, South Brisbane, Queensland, 4101, Australia
[#] Correspondence to: T.Shafee@LaTrobe.edu.au
Key words: ShKT; cysteine-rich peptide; disulfide-rich protein; peptide evolution; peptide structure; sequence analysis

Abstract

The ShKT superfamily is widely distributed throughout nature and encompasses a wide range of documented functions and processes, from modulation of potassium channels to involvement in morphogenesis pathways. Cysteine-rich secretory proteins (CRISPs) contain a cysteine-rich domain (CRD) at the C-terminus that is similar in structure to the ShK fold. Despite the structural similarity of the CRD and ShK-like domains, we know little of the sequence-function relationships in these families. Here, for the first time, we examine the evolution of the biophysical properties of sequences within the ShKT superfamily, in relation to function, with a focus on the ShK-like superfamily. ShKT data were sourced from published sequences in the protein family database, in addition to new ShK-like sequences from the Australian speckled anemone (*Oulactis* sp.). Our analysis clearly delineates the ShK-like family from the CRDs of CRISP proteins. The four CRISP subclusters separate out into the main phyla of Mammalia, Insecta and Reptilia. The ShK-like family is in turn composed of seven subclusters, the largest of which contains members from across the eukaryotes, with a continuum of intermediate properties, and smaller sub-clusters containing specialised families such as nematode ShK-like sequences. Several of these ShKT sub-clusters contain no functionally characterised sequences. This chemical space analysis should be useful as a guide to select sequences for functional studies and to gain insight into the evolution of these highly divergent sequences with an ancient conserved fold.



Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

1 Introduction

The *Stichodactyla helianthus* K channel toxin-like superfamily (ShKT) consists of molecules derived from both venomous and non-venomous species (Chang et al., 2018). These molecules have a range of functions, from modulatory actions on potassium channels to roles in morphogenesis and cell differentiation (Gibbs et al., 2008; Prentis et al., 2018; Tsang et al., 2007; Tudor et al., 1996; Yan et al., 2000). ShKT domains are widespread throughout nature, occurring in sea anemones, reptiles, nematodes and mammals (Castañeda et al., 1995; Galea et al., 2014; Gibbs et al., 2006; Minagawa et al., 1998; Nguyen et al., 2013; Sunagar et al., 2012; Tsang et al., 2007). However, we know little of their evolution and the relationship between their chemical properties and function.

The ShKT superfamily contains two main families, the ShK-like proteins and the cysteine-rich domains (CRD) of cysteine-rich secretory proteins (CRISPs), which share a common fold (**Fig. 1A**). ShK-like domains occur as discrete single domains, multiple repeats or in combination with peptidase and other enzyme domains (Castañeda et al., 1995; Finn et al., 2014; Galea et al., 2014) (**Fig. S1**). The CRISPs have a more constrained organisation, with a ShKT domain linked by a flexible hinge to the pathogenesis-related 1 (PR1) domain (Gibbs et al., 2008). Despite their structural similarity, superfamily members have a documented range of functions, processes and taxonomic distribution.

The prototypical superfamily member is the sea anemone toxin ShK, a 35-residue peptide isolated from the sea anemone *Stichodactyla helianthus* (Castañeda et al., 1995). The structure of ShK is characterised by two α-helices (**Fig. 1B**) and three disulfide bonds (Cys^1-Cys^6, Cys^2-Cys^4 and Cys^3-Cys^5) (Tudor et al., 1996). CRISP sequences are larger polypeptides, characterised by 10-16 cysteines and molecular masses of 20-30 kDa, with distinct domains. The CRD region has a similar fold and cysteine connectivity to that of ShK (Error! Reference source not found.**C**) and is therefore thought to be homologous to ShK-like peptides, although their sequence identity is low (Guo et al., 2005).

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

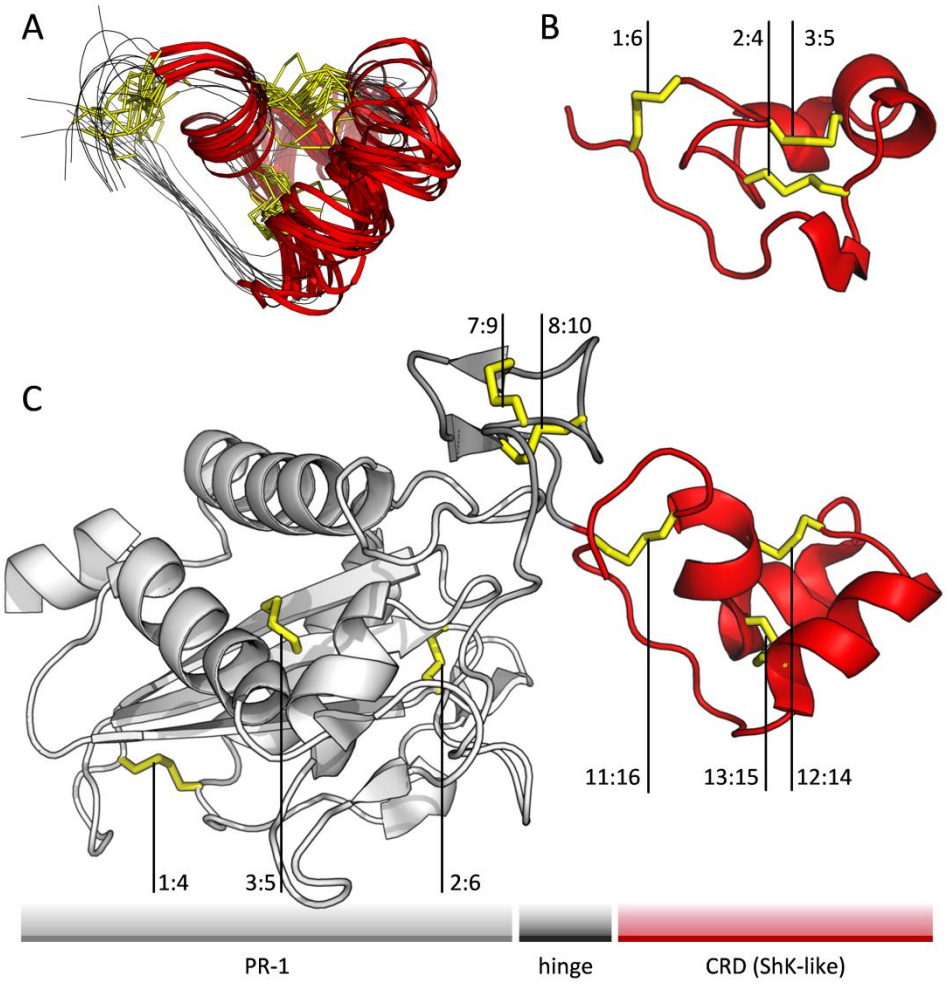


Fig. 1. Comparison of ShKT superfamily structures. A) Multiple structure alignment of 19 ShK-like domains with α-helices in red and disulfide bridges in yellow. B) ShK from the sea anemone *Stichodactyla helianthus* (1ROO; Tudor et al., 1996) C) Triflin CRISP from the snake *Protobothrops flavoviridis* (1WVR; Shikamoto et al., 2005). Highlighted in red are the structurally similar regions of the single domain of ShK and the cysteine-rich domain (CRD) of the CRISP sequence. CRISP pathogenesis-related group 1 domain (PR-1) is shown in white, with the hinge region in grey. Disulfide bridges are shown in yellow with cysteines numbered.

Four main functional and biological processes have been attributed thus far to proteins in the ShKT superfamily: potassium and calcium channel modulation, involvement in morphogenesis pathways and cell regeneration. ShK, the prototypical peptide of the ShKT domain, is a potent K_V1 channel blocker (Castañeda et al., 1995). An analogue of this peptide is in clinical trial for treatment of the autoimmune disease psoriasis (Chandy and Norton, 2017; Chi et al., 2012; Tarcha et al., 2017). Other ShK-like single domain peptides show activity across various K_V channels, for example BgK (from *Bundosoma granulifera*) ($K_V1.1$, 1.2, 1.3) and *Oulactis* sp. (OspTx2a) ($K_V1.2$, 1.6) (Cotton et al., 1997; Punnepalli et al., 2018). ShK-like sequences are also

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

polyfunctional, with ShPI-1 (from *Stichodactyla helianthus*) and APEKTx1 (from *Anthopleura elegantissima*) being both Kunitz-type inhibitors with K_V channel binders (García-Fernández et al., 2016; Peigneur et al., 2011). The precursor domain of NEP3 (from *Nematostella edwardsia*), a ShKT domain repeat protein, was shown to be neurotoxic to fish (Columbus-Shenkar et al., 2018; Moran et al., 2013). The function of the remaining ShKT domains in NEP3 are yet to be determined (Columbus-Shenkar et al., 2018).

The toxin-like domain (TxD) of the matrix metalloprotease MMP-23 has high structural similarity to ShK/BgK and the CRD region of CRISPs (Galea et al., 2014; Nguyen et al., 2013; Rangaraju et al., 2010). Functionally, MMP-23 may play a role in potassium channel trafficking to cell membranes (Galea et al., 2014). The TxD domain is a blocker of several K_V channels (1.1, .3, 1.4, 1.6 and 3.2), although ineffective against others (K_V 1.2, 1.5, 1.7 and KCa3.1) (Galea et al., 2014).

The importance of the ShKT domain in morphogenesis pathways is exemplified by Mab-7, a protein from the roundworm *Caenorhabditis elegans* (Tsang et al., 2007). The ShKT domain in Mab-7 is essential for protein binding, the protein itself being responsible for sensory ray morphogenesis, which is required for contact between the male and hermaphrodite during mating (Tsang et al., 2007).

CRISPs are found predominantly in vertebrates and are composed of four main functional groups. CRISP-1 is located in the reproductive tracts of mammals and CRISP-2 performs an autoantigen function in sperm. Functional studies of the mouse CRD domain of Tpx-1, a CRISP-2 member, regulates the cardiac ryanodine receptor Ca^{2+} signalling in line with its relatedness to the ion-channel modulators BgK and ShK (Gibbs et al., 2006). CRISP-3 is found in the venom of reptiles (snakes and lizards) and CRISP-4 is rodent-specific, where it is implicated in sperm interactions (Gibbs et al., 2008; Turunen et al., 2011). Several homologous CRISP-like sequences have been identified in insects (e.g. mosquitos, beetles and ants), although most of these proteins remain uncharacterised (Holt et al., 2002; Oxley et al., 2014; The Tribolium Genome Sequencing Consortium, 2008).

Cysteine-rich peptide sequences from nature comprise a valuable library of bioactive compounds. Their compact, stable structures permit a large variety of loop sequences to be tolerated, leading to high sequence and function diversity. However, this diversity typically precludes most traditional sequence analyses, such as phylogenetics (Inkpen and Doolittle, 2016; Pearson and Sierk, 2005; Rost, 1999). We therefore built quantitative maps based on sequence/chemical space to define the functional regions explored by the extant ShKs. This technique has been applied successfully to investigate the evolution of two superfamilies of defensins, sequences that are also cysteine-rich and highly sequence-diverse (Mitchell et al., 2019; Shafee et al., 2017; Shafee and Anderson, 2018) as well as for traditional globular proteins (Jackson et al., 2018). Sequences can be explicitly placed within a multidimensional sequence space based on a multiple sequence alignment (MSA) combined with the biophysical properties of their residues (Atchley et al., 2005). This multidimensional space can be summarised in a smaller number of dimensions that capture the important sequence

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

properties. Clusters within such a space indicate sets of co-varying peptide chemical properties (Shafee and Anderson, 2018).

Here, we explore the biophysical determinants of these clusters and their relationship to function. This gives insights into the main groups of ShK-like sequences, and their evolution, while clearly delineating the CRISP CRD family within the ShKT superfamily. This analysis can also be used to guide exploration of the available ShKT sequences by identifying sequences from poorly characterised regions of chemical space.

2 Methods

2.1 Sequence data

Sequences deposited in the protein family database (Pfam - https://pfam.xfam.org/) under the ShKT clan (CL0213), CRISP (PF08562) and ShK-like (PF01549) were downloaded (Finn et al., 2014). Full-length sequences were examined, and the predicted mature peptide retained for parsing through to the alignments and chemical space analysis. Sequences also included two functionally characterised ShK-like sequences isolated from the tentacle transcriptomes of the sea anemone *Oulactis* sp., U-AITX-Oulsp1 and -Oulsp2 (initially designated OspTx2a and b, respectively) (Bankala et al., 2018b; Punnepalli et al., 2018). Additional sequences from the same *Oulactis* sp. transcriptomes (manuscript in preparation), annotated as containing ShK-like domains, are registered in EMBL-ENA (Accession numbers Table S1). ShK-like sequences available and not in Pfam were from *Calliactis polypus* (van der Burg et al., 2016) and the Pacific sea nettle (*Chrysaora fuscescens*) (Ponce et al., 2016). For proteins containing multiple domains, individual ShKT domains were isolated and labelled as sequentially occurring domains, i.e. multi-domain 1, multi-domain 2 and so forth, for sequence alignment. CRISP sequences were truncated and the CRD region retained for alignment.

2.2 Structure alignment

The following ShKT domain structures were aligned using the combinatorial extension method and visualised in PyMol in order to show the level of structure conservation PDB ids: 1BGK, 1RC9, 1ROO, 1WVR, 1XTA, 1XX5, 2GIZ, 2N6B, 2MCR, 2MD0, 2A05, 2CQ7, 2K72, 3MZ8, 2DDA, 2DDB, 6BUC, 6CKD, 6CKF.

2.3 Sequence alignment

Multiple sequence alignments were generated using the CysBar webserver (Shafee et al., 2016) and ClustalΩ (Sievers et al., 2011). Homologous cysteines were barcoded to constrain the alignment. Alternative alignments were checked for consistency using Align Stat (Shafee and

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

Cooke, 2016). The CysBar webserver was also used to calculate the length, charge and hydrophobicity of each sequence. Sequences were aligned in Stockholm format and logo plots generated using Weblogo 3 (Crooks, 2004).

2.4 Chemical/Sequence space analysis

The short sequence length, frequent insertions and deletions, low sequence conservation of peptides and evolutionary plasticity of cysteine-rich proteins (CRPs), can cause conventional phylogenetic methods to suffer from saturation effects (Wake, 1991). We therefore also analysed the full dataset of ShK-like peptides by position-specific biophysical property distance analyses, or sequence-space analyses. The variables used were R-group molecular mass (Dalton), net charge (Coulomb), hydrophobicity (Doolittle index) (Kyte and Doolittle, 1982), disorder propensity (TOP-IDP) (Campen et al., 2008), disulfide potential (binary descriptor; noncys=0, cys=1), and occupancy (binary descriptor; gap=0, nongap=1). The numerical sequence space was projected by principal component analysis (PCA) in [R] (Development Core Team R, 2011). Bayesian clustering performed using Mclust to identify groups of sequences with similar biophysical properties (Fraley and Raftery, 2012). PCA loadings were used to investigate co-varying sequence property sets. The same process was repeated with the column subsets of the MSA to confirm that clustering was not an artefact of the lack of the usually conserved cysteines and glycine.

2.5 Visualisation

Data were visualised with custom [R] scripts based on rgl, ggplots, igraph, and phytools (Csárdi and Nepusz, 2006; Revell, 2012; Wickham, 2009). Structures were visualised with Pymol (DeLano, 2002).

3 Results

3.1 Sequence retrieval and data preparation

A total of 1082 sequences, including single and multiple domains, was subject to analysis after division into their ShK and CRD domain components. The analysis included 50 new sequences from the speckled anemone (*Oulactis* sp.) transcriptome containing 90 ShK-like domains. Many of these sequences contained multiple ShK-like repeat domains, up to nine repeats in one sequence, or were in multi-domain proteins associated with peptidases (e.g. astacin) or protease inhibitors (e.g. Kunitz-type) (**Fig. S1**).

ShK-like domains have extremely divergent sequences (**Fig. 2**), but a highly conserved and constrained structure (**Fig. 1B**) owing to the stability conferred by their disulfide linkages. They

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

especially have frequent insertions and deletions (indels), which change the length of the inter-cysteine loops, common for cysteine-rich protein superfamilies (Shafee et al., 2017) . This poses a problem for traditional sequence analysis methods, which typically struggle with such short, divergent, indel-rich sequences. To adapt to these limitations, multiple sequence alignments can be constrained by homologous cysteines (Shafee et al., 2016). From this alignment, chemical/sequence space analysis can be calculated by arranging sequences based on the biophysical properties of their residues. Within the high sequence variations, this method allows covarying sets of sequence properties to be identified, as well as favoured and disfavoured combinations, and can robustly classify small protein and peptide sequence-function relationships (Dash et al., 2019).

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

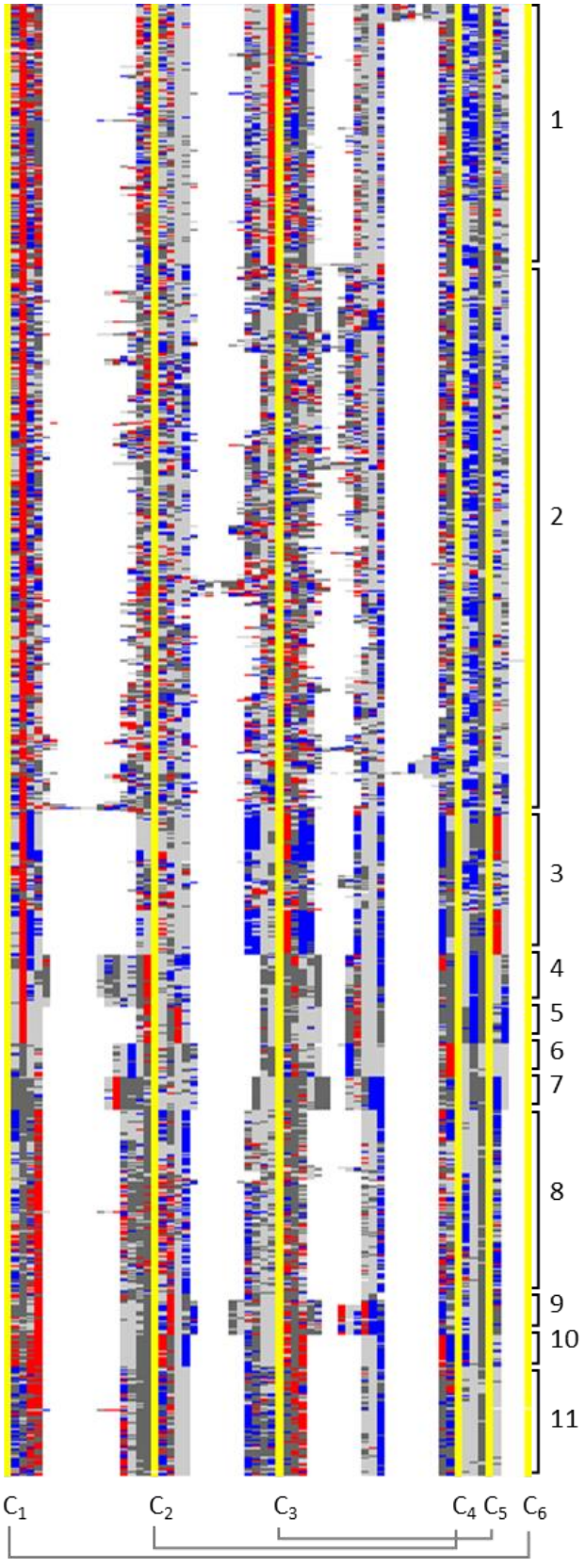


Fig. 2. Sequence variation in ShKT domains. Multiple sequence alignment of 1082 ShK-like domains with cysteines in yellow, cationic residues in blue, anionic residues in red, hydrophobic residues in light grey, polar residues in dark grey, and gaps in white. Numbered groups are described in Fig 3. ShK-like and CRISPs form the main separation in sequence space.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

3.2 ShK-like and CRISPs form the main separation in sequence space

Within the calculated sequence space, the ShK-like and CRISP families can be clearly differentiated when Bayesian model-based clustering is used to determine the optimal number of groups needed to describe the data (**Fig. 3A,B**). The key sequence properties that determine this separation include a longer loop 1 with its anionic charge at a later position in the loop, a smaller residue at the start of loop 2, a slightly shorter loop 3, a smaller and more hydrophobic residue in the middle of loop 4, and shorter loop 5, as defined by the principal component 1 axis (**Table S2**). In addition, axis PC3 separates the nematode and mammalian ShKs as well as the insect CRISPs, indicating some common (convergent) features among these disparate groups, including their more cationic loop 3 and larger residue at the start of loop 5 (**Fig. S2, Table S2**). These features primarily affect the surface surrounding the lysine residue, which plays a key role in those members of the family that block Kv1 channels (**Fig. S3**) (Chi et al., 2012; Kalman et al., 1998; Lanigan et al., 2002).

Each of these two main groups contains several sub-clusters (**Fig. S4**). When these two main groups are analysed separately, the ShK-like cluster contains seven discernable groups (**Fig. 3C,D**) and the CRISPs four (**Fig. 3E,F**), separated by their biophysical properties (**Table S3** and **S4**) and strongly correlating with the main phyla present. These 11 groups remain well-organised when projected back onto the overall sequence space (**Fig. 3G**).

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

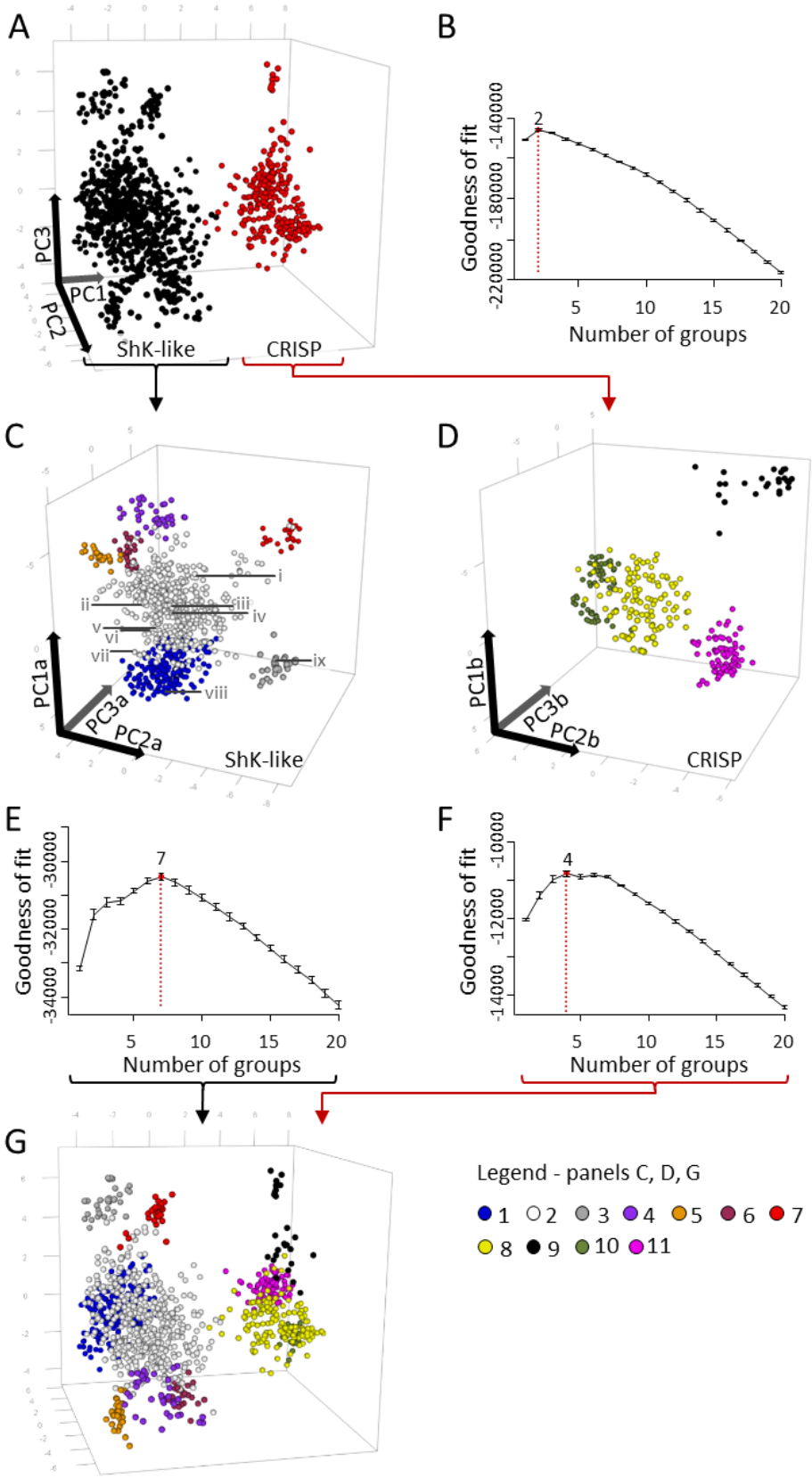


Fig. 3. Sequence space of the ShKT superfamily and its two main groups. A) Overall, the sequence space of ShK-like and CRISPs are clearly delineated into their two families, with groups automatically defined by Bayesian model-based clustering annotated in black and red. B) Estimate of optimum number of groups in the sequence space by Bayesian model-based clustering. C) Sequence space of ShK-like group, with seven coloured groups and example sequences indicated as: i, Aurelin; ii, ShK; iii, AcK; iv, BgK; v, Mad-7; vi, NEP3; vii, BmK; viii, HMP1; ix, MMP23. Group 1 in blue, 2 in white, 3 in grey, 4 in purple, 5 in orange, 6 in maroon, and 7 in red. D) Estimate of optimum number of groups in the ShK-like sequence space by Bayesian model-based clustering E) Sequence space of CRISP CRD's, with four coloured groups. F) Estimate of best number of groups in the CRISP CRD sequence space by Bayesian model-based clustering. G) Overall sequence space of all ShKT domains, arranged as in part A, but coloured by the 11 groups identified in parts C and D. Rotatable versions in supplementary data file 1.

3.3 The ShK-like family contains clear clusters of properties

Within the ShK-like sequence space, clear sub-clusters can be identified, separated along the main axes (principal components, PCs) based on a further set of covarying biophysical properties (**Fig. S2, Table S3**). The majority of well-studied sequences (e.g. ShK, AcK1 and MMP23) fall within Groups 1, 2, and 3, whereas the remaining groups remain largely uncharacterised functionally (**Fig. 3C**). Phyla composition, and sequence function characterised within each of the seven discernible groups of ShK-like and four CRISPs groups are summarised in **Table S5**.

Groups 1, 2 and 4 form a continuum along axis PC1. Groups 1 and 2 contain sequences from a mixture of eukaryotic origins, with plant, algal and planktonic sequences concentrated in Group 1 (**Fig. S5**). Group 4 contains mostly nematode sequences. Most (91%) of the single-domain ShKs also fall in Group 2, as do the most highly charged proteins (mostly cationic, but also the few anionic examples). The single ShK-like domains occur only in cnidarians and nematodes. Group 3 contains the channel-binding domains of vertebrate metalloproteinases like MMP23, strongly separated by axis PC2. Groups 5, 6 and 7 consist of 3-repeat ShKs, none of which have been characterised, with Group 7 being particularly highly separated along axis PC3a. Each three-domain sequence contains one domain from each cluster in the order cluster 6-7-5. The consistency of this order suggests that the radiation of the domain within each cluster occurred subsequent to the domain triplication event, as opposed to the frequent domain gain-and-loss seen in the other multi-domain ShKs.

Many ShK-like sequences are arranged in even longer multi-domain protein complexes or repeats (up to 9 ShK repeats **Fig S5**, found in *Oulactis* sp.). The domains from the longest of these multi-domain sequences are distributed throughout Groups 1 and 2 (**Fig. 4A**). The ShK-like sequences are mostly cationic, with the most highly charged concentrated in Groups 1, 2 and 3, whereas Groups 4 and 7 are largely neutral (**Fig. 4B**).

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

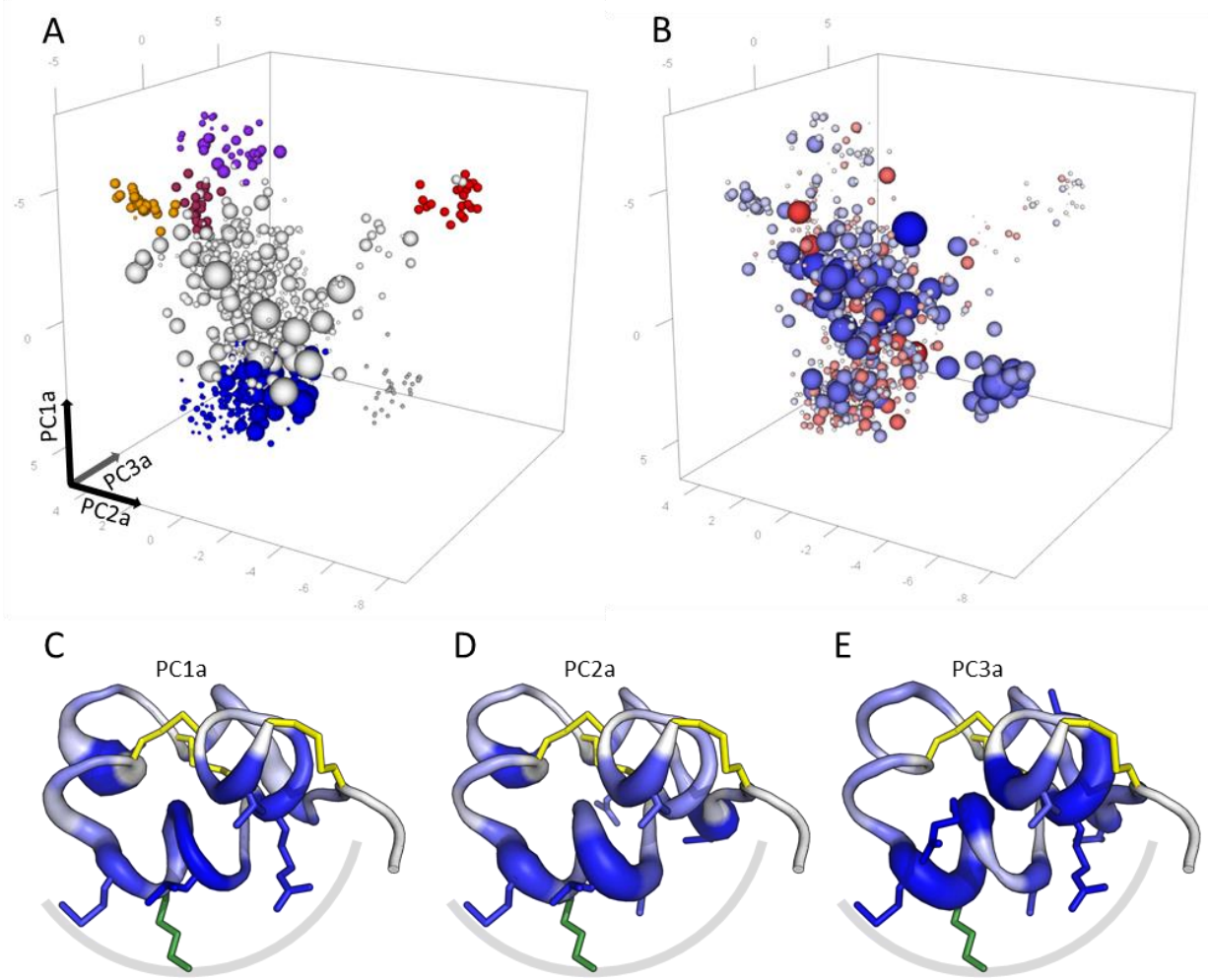


Fig. 4. Details of sequence space of ShK-like group. A) Sequence space of ShK-like group with radius proportional to ShK repeat number (min=1, max = 9). Clusters coloured as in Fig. 3C. B) Net ShK charge. Radius proportional to net charge, coloured with cationic in blue, anionic in red (min = -8, max = 11). Combined loading across all properties for each residue mapped onto structure of ShK (PDB: 1ROO) for each principal component axis C) PC1a, D) PC2a, and E) PC3a. Property loading magnitude indicated by thickness and colour intensity. Residues with average combined loading in top decile shown as sticks. General region containing the majority of highly-loaded residues indicated by grey arc. Inhibitory Lys shown as green stick.

3.4 Clusters are most influenced by residues on the binding surface

The positions of sequences within the space are primarily determined by covarying sets of sequence properties. These are the properties with the highest loading in each principal component axis. For the ShK-like clusters, the residues whose properties most strongly affect position in the sequence space are clustered around the lysine, which is thought to insert into the pore of the K_V channel (**Fig. 4C-E**) (Lanigan et al., 2002; Pennington et al., 1996; Rashid et

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

al., 2013). Residue variation on the opposite face of the protein, in contrast, appears to have little correlation and so has little effect on clustering in the sequence space. Each ShK-like cluster preferences a different set of properties for these residues This pattern is observed for the features convergent along axis PC3, but not for the other axes of the ShKT sequence set or for the CRISP sequences analysed separately (**Fig. S3**).

4 Discussion

The ShKT fold is an ancient scaffold found in cnidarians (~700 *mya*), nematodes, mammals and toxicoferan-reptiles (~65 *mya*) in the form of ShK-like proteins and the CRD region of CRISPs (Chhabra et al., 2014; Fautin, 1998; Gibbs et al., 2008, 2006; Sunagar et al., 2012; Tudor et al., 1996). Within this constrained disulfide structure, sequence divergence is exceptionally high. However, the sequences are clearly not random; otherwise, the sequence space would show a spherical cloud. The clusters define favoured sets of biophysical properties, with voids indicating possible sequences with disfavoured properties. Our analysis supports the two families, ShK-like and the CRISP CRD region, as the main groupings in chemical space, with only a few sequences falling in between these two clusters. We were further able to identify eleven natural clusters within these two families – seven within the ShK-like sequences and four within the CRISPs. In the case of the CRISPs, goodness of fit is almost unchanged upon splitting the largest (yellow) subcluster into three, although the resulting subclusters did not obviously correlate to any taxa or characterised functions.

Although the ShK-like and CRISPs form clearly separated clusters, the middle domain of the 3-repeat nematode proteins and the ShKT domain of the vertebrate matrix metalloproteases share a range of convergent properties with the insect CRISPs, suggesting similar evolutionary solutions to selection pressures in highly separated organisms. These features primarily affect the surface region surrounding the inhibitory lysine, making it markedly more cationic, and likely modulating channel target specificity. Those targets are $K_V1.6$ and $K_V1.3$ in the case of MMP23 (Rangaraju et al., 2010), although it is not known how conserved this is in the other members of that group. Conversely, nothing is known of the targets for the 3-domain nematode group, or insect CRISPs. Vertebrate CRISPs bind a variety of calcium and potassium channels (Gibbs et al., 2008) depending on their role in reproduction (e.g. CRISP-2) or as toxins (e.g. triflin). Future characterisation will determine if these overlapping sequences have any correlation with their function.

Within the ShK-like superfamily there are seven clear groups. Groups 1, 2, and 4 have a continuity of properties spanning the spectrum from floral to faunal ShKs, indicating remarkable tolerance of the scaffold to variation in the inter-cysteine loops. Since ShK-like peptides can also be polyfunctional (Peigneur et al., 2011), it remains to be seen whether sequences with intermediate biophysical properties display multiple and/or intermediate functions at the interfaces of single-function clusters.

Group 1 mostly consists of proteins that include protease domains, such as HMP1 and PMP1 (Pan et al., 1998; Yan et al., 2000), predominantly from plankton, plants and algae, with some

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

nematode and cnidarian members (**Fig. S4**). It also contains a number of uncharacterised 2-ShK-domain proteins.

Group 2 is the largest cluster and represents the most functionally characterised group of all the ShKT clusters and spans all phyla. Three main functions fall within this group. It contains cnidarian potassium channel modulators targeting a range of K_V channels, e.g. ShK and AcK target K_V1.3, and BgK targets K_V1.1, K_V1.2 and K_V1.3 (Castañeda et al., 1995; Cotton et al., 1997). The first domain of NEP3 is included in this group; although its specific target has not yet been identified, it is lethal to fish (Columbus-Shenkar et al., 2018). AsK and Osptx2b also fall in Group 2, even though they do not show any activity against potassium channels (Bankala et al., 2018a, 2018b).

Group 2 also contains sequences with the ShK-like domain at the C-terminus that have been shown to be important for processes including morphogenesis, cell differentiation and digestion. BmK1 from the filial worm (*Brugia malayi*) and Mab-7 are ShK-like domains that are structurally similar. In BmK1 the domain shows a range of immunomodulatory functions against rat and human T cells and in Mab-7 is important for morphogenesis (Chhabra et al., 2014; Tsang et al., 2007). HMP2 from *Hydra vulgaris*, an astacin-metalloproteinase, is implicated in cell differentiation and morphogenesis (Yan et al., 2000). Similar to HMP2, HMP1 contains an astacin and ShK-like cysteine-rich region Tox1 (TH) domain at the C-terminus, as does the jellyfish (*Podocoryne carnea*) protein PMP1, both of which implicated in digestion (Pan et al., 1998; Yan et al., 2000).

Lastly, Group 2 contains those superfamily members that have been characterised as antimicrobial. Aurelin displays antibacterial and antiviral activity respectively (Ovchinnikova et al., 2006). This antimicrobial function may be a native function, or a promiscuous side-activity in the same manner as immunomodulatory functions (Peigneur et al., 2011).

Group 4 contains further uncharacterised nematode sequences with 1-4 ShKT domain repeats. Conversely, ShK-like Groups 3, 5, 6, and 7 have much clearer cluster delineation and are therefore more likely to have distinctly specialised functions. For instance, Group 3 contains MMP-23-like sequences. The grouping of MMP-23 is clearly separate from the ShK toxin-like sequences and CRISPs, reflecting previous findings where these sequences form a phylogenetic outgroup (Galea et al., 2014).

The main differences between these clusters are in the residues that surround the lysine thought to insert into the target channel (K22 in ShK) (Chandy and Norton, 2017; Chi et al., 2012; Kalman et al., 1998; Lanigan et al., 2002). This supports their likely role in determining specificity interactions with their target channels and suggests the properties that may be useful for predicting and manipulating function in these molecules.

This ancient scaffold has widespread taxonomic distribution and supports an array of functions. However, our study highlights that only a fraction of the chemical space explored for the scaffold has been characterised functionally, especially those sequences with ShKT domain repeats. With the proliferation of ShK-like sequences being identified in transcriptome studies, a process is needed to prioritise sequence characterisation efforts. These classifications can

help to inform future sequence-mining efforts that aim to maximise sampling of the superfamily's chemical diversity and bring order to our understanding of their sequence-function relationships.

This work also represents the first use of multidimensional analyses of peptide biophysical properties in a family of toxins other than the defensins. Using the method to identify covarying property sets and visualise sequence clustering is especially useful for the sorts of cysteine-rich peptides often found in venoms, since they are highly sequence-variable yet can be reliably aligned based on conserved cysteine placement. We therefore expect it to be a versatile tool for other toxin superfamilies.

5 Conflict of Interest

The authors declare there are no conflicts of interest.

6 Acknowledgment

The authors acknowledge Dr. Rodrigo A. V. Morales and Edward Airey for their contribution to the preparation of sequences used in the analysis. This project was funded in part by ARC linkage grant LP150100621. M.L.M acknowledges an Australian Government Research Training Program Scholarship, Monash Medicinal Chemistry Faculty Scholarship and Monash University-Museum Victoria Scholarship top-up. R.S.N acknowledges fellowship support from the Australian National Health and Medical Research Council.

7 Supplementary Data

Supplementary data related to this article can be found at (journal website).

8 References

Atchley, W.R., Zhao, J., Fernandes, A.D., Druke, T., 2005. Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. 102, 6395–6400. https://doi.org/10.1073/pnas.0408677102

Bankala, K., MacRaild, C.A., Punnepalli, S., Morales, R.A.V., Peigneur, S., Macrander, J., Yu, H.H., Daly, M., Raghothama, S., Dhawan, V., Chauhan, S., Tytgat, J., Pennington, M.W., Norton, R.S., 2018a. Structure, folding and stability of a minimal homologue from *Anemonia sulcata* of the sea anemone potassium channel blocker ShK. Peptides 99, 169–178. https://doi.org/10.1016/j.peptides.2017.10.001

Bankala, K., Villegas-Moreno, J., Mitchell, M.L., Csoti, A., Peigneur, S., Amero, C., Pennington, M.W., Tytgat, J., Panyi, G., Norton, R.S., 2018b. Synthesis, folding, structure and activity of a predicted peptide from the sea anemone Oulactis sp. with an ShKT fold. Toxicon 150, 50–59. https://doi.org/10.1016/j.toxicon.2018.05.006

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242. https://doi.org/10.1093/nar/28.1.235

Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., Dunker, A.K., 2008. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept. Lett. 15, 956–63. https://doi.org/10.2174/092986608785849164

Castañeda, O., Sotolongo, V., Amor, A.M., Stöcklin, R., Anderson, A.J., Harvey, A.L., Engström, Å., Wernstedt, C., Karlsson, E., 1995. Characterization of a potassium channel toxin from the Caribbean sea anemone *Stichodactyla helianthus*. Toxicon 33, 603–613. https://doi.org/10.1016/0041-0101(95)00013-C

Chandy, K.G., Norton, R.S., 2017. Peptide blockers of $K_V1.3$ channels in T cells as therapeutics for autoimmune disease. Curr. Opin. Chem. Biol. 38, 97–107. https://doi.org/10.1016/j.cbpa.2017.02.015

Chang, S.C., Bajaj, S., Chandy, K.G., 2018. ShK toxin: history, structure and therapeutic applications for autoimmune diseases. WikiJournal Sci. 1. https://doi.org/10.15347/wjs/2018.003

Chhabra, S., Chang, S.C., Nguyen, H.M., Huq, R., Tanner, M.R., Londono, L.M., Estrada, R., Dhawan, V., Chauhan, S., Upadhyay, S.K., Gindin, M., Hotez, P.J., Valenzuela, J.G., Mohanty, B., Swarbrick, J.D., Wulff, H., Iadonato, S.P., Gutman, G.A., Beeton, C., Pennington, M.W., Norton, R.S., Chandy, K.G., 2014. $K_V1.3$ channel-blocking immunomodulatory peptides from parasitic worms: implications for autoimmune diseases. FASEB J. 28, 3952–3964. https://doi.org/10.1096/fj.14-251967

Chi, V., Pennington, M.W., Norton, R.S., Tarcha, E.J., Londono, L.M., Sims-Fahey, B., Upadhyay, S.K., Lakey, J.T., Iadonato, S., Wulff, H., Beeton, C., Chandy, K.G., 2012. Development of a sea anemone toxin as an immunomodulator for therapy of autoimmune diseases. Toxicon 59, 529–546. https://doi.org/10.1016/j.toxicon.2011.07.016S0041-0101(11)00247-9 [pii]

Columbus-Shenkar, Y.Y., Sachkova, M.Y., Macrander, J., Fridrich, A., Modepalli, V., Reitzel, A.M., Sunagar, K., Moran, Y., 2018. Dynamics of venom composition across a complex life cycle. eLife 7, e35014. https://doi.org/10.7554/eLife.35014

Cotton, J., Crest, M., Bouet, F., Alessandri, N., Gola, M., Forest, E., Karlsson, E., Castaneda, O., Harvey, A.L., Vita, C., Menez, A., 1997. A potassium-channel toxin from the sea anemone *Bunodosoma granulifera*, an inhibitor for Kv1 channels - revision of the amino acid sequence, disulfide-bridge assignment, chemical synthesis, and biological activity. Eur. J. Biochem. 244, 192–202. https://doi.org/10.1111/j.1432-1033.1997.00192.x

Crooks, G.E., 2004. WebLogo: A sequence logo generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004

Csárdi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Syst. 1695, 1–9.

Dash, T.S., Shafee, T., Harvey, P.J., Zhang, C., Peigneur, S., Deuis, J.R., Vetter, I., Tytgat, J.,

Anderson, M.A., Craik, D.J., Durek, T., Undheim, E.A.B., 2019. A centipede toxin family defines an ancient class of CSαβ defensins. Structure 27, 315–326.e7. https://doi.org/https://doi.org/10.1016/j.str.2018.10.022

DeLano, W.L., 2002. Pymol: An open-source molecular graphics tool. CCP4 Newsl. Protein Crystallogr. 40, 82–92.

Development Core Team R, 2011. R: A language and environment for statistical computing.

Fautin, D.G., 1998. Cnidaria, in: Skinner, M., Knobil, E., Neill, J.D. (Eds.), Encyclopedia of Reproduction. Academic Press, San Diego, pp. 645–653.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230. https://doi.org/10.1093/nar/gkt1223

Fraley, C., Raftery, A.E., 2012. MCLUST Version 4 for R : Normal Mixture Modeling for Model-Based Clustering , Classification , and Density Estimation. Tech. Rep. - Dep. Stat. Univ. Washingt.

Galea, C.A., Nguyen, H.M., George Chandy, K., Smith, B.J., Norton, R.S., 2014. Domain structure and function of matrix metalloprotease 23 (MMP23): role in potassium channel trafficking. Cell. Mol. Life Sci. 71, 1191–1210. https://doi.org/http://dx.doi.org/10.1007/s00018-013-1431-0

García-Fernández, R., Peigneur, S., Pons, T., Alvarez, C., González, L., Chávez, M., Tytgat, J., 2016. The Kunitz-type protein ShPI-1 inhibits serine proteases and voltage-gated potassium channels. Toxins (Basel). 8, 110. https://doi.org/10.3390/toxins8040110

Gibbs, G.M., Roelants, K., O'Bryan, M.K., 2008. The CAP Superfamily: Cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 Proteins—roles in reproduction, cancer, and immune defense. Endocr. Rev. 29, 865–897. https://doi.org/10.1210/er.2008-0032

Gibbs, G.M., Scanlon, M.J., Swarbrick, J., Curtis, S., Gallant, E., Dulhunty, A.F., O'Bryan, M.K., 2006. The cysteine-rich secretory protein domain of Tpx-1 is related to ion channel toxins and regulates ryanodine receptor Ca^{2+} signaling. J. Biol. Chem. 281, 4156–63. https://doi.org/10.1074/jbc.M506849200

Guo, M., Teng, M., Niu, L., Liu, Q., Huang, Q., Hao, Q., 2005. Crystal structure of the cysteine-rich secretory protein stecrisp reveals that the cysteine-rich domain has a K^+ channel inhibitor-like fold. J. Biol. Chem. 280, 12405–12. https://doi.org/10.1074/jbc.M413566200

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M.C., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., Lai, Z., Kraft, C.L., Abril, J.F., Anthouard, V., Arensburger, P., Atkinson, P.W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G.K., Chrystal, M.A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C.A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R.,

Hillenmeyer, M.E., Hladun, S.L., Hogan, J.R., Hong, Y.S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.-J., Lobo, N.F., Lopez, J.R., Malek, J.A., McIntosh, T.C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S.D., O'Brochta, D.A., Pfannkoch, C., Qi, R., Regier, M.A., Remington, K., Shao, H., Sharakhova, M. V, Sitter, C.D., Shetty, J., Smith, T.J., Strong, R., Sun, J., Thomasova, D., Ton, L.Q., Topalis, P., Tu, Z., Unger, M.F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K.J., Wortman, J.R., Wu, M., Yao, A., Zdobnov, E.M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S.C., Zhimulev, I., Coluzzi, M., della Torre, A., Roth, C.W., Louis, C., Kalush, F., Mural, R.J., Myers, E.W., Adams, M.D., Smith, H.O., Broder, S., Gardner, M.J., Fraser, C.M., Birney, E., Bork, P., Brey, P.T., Venter, J.C., Weissenbach, J., Kafatos, F.C., Collins, F.H., Hoffman, S.L., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298, 129–49. https://doi.org/10.1126/science.1076181

Inkpen, S.A., Doolittle, W.F., 2016. Molecular Phylogenetics and the Perennial Problem of Homology. J. Mol. Evol. 1–9. https://doi.org/10.1007/s00239-016-9766-4

Jackson, M.A., Gilding, E.K., Shafee, T., Harris, K.S., Kaas, Q., Poon, S., Yap, K., Jia, H., Guarino, R., Chan, L.Y., Durek, T., Anderson, M.A., Craik, D.J., 2018. Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases. Nat. Commun. 9, 2411. https://doi.org/10.1038/s41467-018-04669-9

Kalman, K., Pennington, M.W., Lanigan, M.D., Nguyen, A., Rauer, H., Mahnir, V., Paschetto, K., Kem, W.R., Grissmer, S., Gutman, G.A., Christian, E.P., Cahalan, M.D., Norton, R.S., Chandy, K.G., 1998. ShK-Dap22, a potent K_V1.3-specific immunosuppressive polypeptide. J. Biol. Chem. 273, 32697–707. https://doi.org/10.1074/JBC.273.49.32697

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

Lanigan, M.D., Kalman, K., Lefievre, Y., Pennington, M.W., Chandy, K.G., Norton, R.S., 2002. Mutating a critical lysine in ShK toxin alters its binding configuration in the pore−vestibule region of the voltage-gated potassium channel, K_V1.3. Biochemistry 41, 11963–11971. https://doi.org/10.1021/bi026400b

Minagawa, S., Ishida, M., Nagashima, Y., Shiomi, K., 1998. Primary structure of a potassium channel toxin from the sea anemone *Actinia equina* 427, 149–151. https://doi.org/10.1016/S0014-5793(98)00403-7

Mitchell, M.L., Shafee, T.M.A., Papenfuss, A.T., Norton, R.S., 2019. Evolution of cnidarian *trans*-defensins: Sequence, structure and exploration of chemical space. Proteins Struct. Funct. Bioinforma. prot.25679, 1–10. https://doi.org/10.1002/prot.25679

Moran, Y., Praher, D., Schlesinger, A., Ayalon, A., Tal, Y., Technau, U., 2013. Analysis of soluble protein contents from the nematocysts of a model sea anemone sheds light on venom evolution. Mar. Biotechnol. 15, 329–339. https://doi.org/10.1007/s10126-012-9491-y

Nguyen, H.M., Galea, C.A., Schmunk, G., Smith, B.J., Edwards, R.A., Norton, R.S., Chandy, K.G., 2013. Intracellular trafficking of the KV1.3 potassium channel is regulated by the prodomain of a matrix metalloprotease. J. Biol. Chem. 288, 6451–6464.

https://doi.org/10.1074/jbc.M112.421495

Ovchinnikova, T. V, Balandin, S. V, Aleshina, G.M., Tagaev, A.A., Leonova, Y.F., Krasnodembsky, E.D., Men'shenin, A. V, Kokryakov, V.N., 2006. Aurelin, a novel antimicrobial peptide from jellyfish *Aurelia aurita* with structural features of defensins and channel-blocking toxins. Biochem. Biophys. Res. Commun. 348, 514–523. https://doi.org/10.1016/j.bbrc.2006.07.078

Oxley, P.R., Ji, L., Fetter-Pruneda, I., McKenzie, S.K., Li, C., Hu, H., Zhang, G., Kronauer, D.J.C., 2014. The genome of the clonal raider ant *Cerapachys biroi*. Curr. Biol. 24, 451–458. https://doi.org/10.1016/j.cub.2014.01.018

Pan, T., Gröger, H., Schmid, V., Spring, J., 1998. A toxin homology domain in an astacin-like metalloproteinase of the jellyfish *Podocoryne carnea* with a dual role in digestion and development. Dev. Genes Evol. 208, 259–66. https://doi.org/10.1007/s004270050180

Pearson, W.R., Sierk, M.L., 2005. The limits of protein sequence comparison? Curr. Opin. Struct. Biol. 15, 254–260. https://doi.org/10.1016/j.sbi.2005.05.005

Peigneur, S., Billen, B., Derua, R., Waelkens, E., Debaveye, S., Béress, L., Tytgat, J., 2011. A bifunctional sea anemone peptide with Kunitz type protease and potassium channel inhibiting properties. Biochem. Pharmacol. 82, 81–90. https://doi.org/10.1016/j.bcp.2011.03.023

Pennington, M.W., Mahnir, V.M., Khaytin, I., Zaydenberg, I., Byrnes, M.E., Kem, W.R., 1996. An essential binding surface for ShK toxin interaction with rat brain potassium channels. Biochemistry 35, 16407–16411. https://doi.org/10.1021/bi962463g

Ponce, D., Brinkman, D., Potriquet, J., Mulvenna, J., 2016. Tentacle transcriptome and venom proteome of the Pacific Sea Nettle, *Chrysaora fuscescens* (Cnidaria: Scyphozoa). Toxins (Basel). 8, 102. https://doi.org/10.3390/toxins8040102

Prentis, P.J., Pavasovic, A., Norton, R.S., 2018. Sea anemones: Quiet achievers in the field of peptide toxins. Toxins (Basel). 10, 36. https://doi.org/10.3390/toxins10010036

Punnepalli, S., Bankala, K., Peigneur, S., Mitchell, M.L., Rosendo, E., Villegas-Moreno, J., Pennington, M.W., Tytgat, J., Norton, R.S., 2018. Identification, chemical synthesis, structure, and function of a new K_V1 channel blocking peptide from *Oulactis* sp. Pept. Sci. 0, e24073. https://doi.org/10.1002/pep2.24073

Rangaraju, S., Khoo, K.K., Feng, Z.-P., Crossley, G., Nugent, D., Khaytin, I., Chi, V., Pham, C., Calabresi, P., Pennington, M.W., Norton, R.S., Chandy, K.G., 2010. Potassium channel modulation by a toxin domain in matrix metalloprotease 23. J. Biol. Chem. 285, 9124–9136. https://doi.org/10.1074/jbc.M109.071266

Rashid, M.H., Heinzelmann, G., Huq, R., Tajhya, R.B., Chang, S.C., Chhabra, S., Pennington, M.W., Beeton, C., Norton, R.S., Kuyucak, S., 2013. A potent and selective peptide blocker of the Kv1.3 channel: Prediction from free-energy simulations and experimental confirmation. PLoS One 8, e78712. https://doi.org/10.1371/journal.pone.0078712

Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other

things). Methods Ecol. Evol. 3, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng. 12, 85–94.

Shafee, T., Cooke, I., 2016. AlignStat: a web-tool and R package for statistical comparison of alternative multiple sequence alignments. BMC Bioinformatics 17, 434. https://doi.org/10.1186/s12859-016-1300-6

Shafee, T., Lay, F.T., Phan, T.K., Anderson, M.A., Hulett, M.D., 2017. Convergent evolution of defensin sequence, structure and function. Cell. Mol. Life Sci. 74, 663–682. https://doi.org/10.1007/s00018-016-2344-5

Shafee, T., Robinson, A.J., van der Weerden, N., Anderson, M.A., 2016. Structural homology guided alignment of cysteine rich proteins. Springerplus 5, 27. https://doi.org/10.1186/s40064-015-1609-z

Shafee, T.M.A., Anderson, M.A., 2018. A quantitative map of protein sequence space for the cis-defensin superfamily. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty697

Shikamoto, Y., Suto, K., Yamazaki, Y., Morita, T., Mizuno, H., 2005. Crystal structure of a CRISP family Ca^{2+} channel blocker derived from snake venom. J. Mol. Biol. 350, 735–743. https://doi.org/10.1016/j.jmb.2005.05.020

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. https://doi.org/10.1038/msb.2011.75

Sunagar, K., Johnson, W.E., O'Brien, S.J., Vasconcelos, V., Antunes, A., 2012. Evolution of CRISPs associated with Toxicoferan-Reptilian venom and mammalian reproduction. Mol. Biol. Evol. 29, 1807–1822. https://doi.org/10.1093/molbev/mss058

Tarcha, E.J., Olsen, C.M., Probst, P., Peckham, D., Muñoz-Elías, E.J., Kruger, J.G., Iadonato, S.P., 2017. Safety and pharmacodynamics of dalazatide, a $K_V1.3$ channel inhibitor, in the treatment of plaque psoriasis: A randomized phase 1b trial. PLoS One 12, e0180762. https://doi.org/10.1371/journal.pone.0180762

The Tribolium Genome Sequencing Consortium, 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452, 949+.

Tsang, S.W., Nguyen, C.Q., Hall, D.H., Chow, K.L., 2007. mab-7 encodes a novel transmembrane protein that orchestrates sensory ray morphogenesis in *C. elegans*. Dev. Biol. 312, 353–366. https://doi.org/10.1016/j.ydbio.2007.09.037

Tudor, J.E., Pallaghy, P.K., Pennington, M.W., Norton, R.S., 1996. Solution structure of ShK toxin, a novel potassium channel inhibitor from a sea anemone. Nat. Struct. Biol. 3, 317–20.

Turunen, H.T., Sipilä, P., Krutskikh, A., Toivanen, J., Mankonen, H., Hämäläinen, V., Björkgren, I., Huhtaniemi, I., Poutanen, M., 2011. Loss of Cysteine-Rich Secretory Protein 4 (Crisp4) Leads to Deficiency in Sperm-Zona Pellucida Interaction in Mice. Biol. Reprod. 86, Article 12, 1-8. https://doi.org/10.1095/biolreprod.111.092403

van der Burg, C.A., Prentis, P.J., Surm, J.M., Pavasovic, A., 2016. Insights into the innate immunome of actiniarians using a comparative genomic approach. BMC Genomics 17, 850. https://doi.org/10.1186/s12864-016-3204-2

Wake, D.B., 1991. Homoplasy: The result of natural selection, or evidence of design limitations? Am. Nat. 138, 543–567. https://doi.org/10.1086/285234

Wickham, H., 2009. ggplot2: elegant graphics for data analysis. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-98141-3

Yan, L., Fei, K., Zhang, J., Dexter, S., Sarras, M.P., 2000. Identification and characterization of hydra metalloproteinase 2 (HMP2): a meprin-like astacin metalloproteinase that functions in foot morphogenesis. Development 127, 129–41.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

Supplementary information

Mapping the chemical and sequence space of the ShK-like superfamily

9 Supplementary Tables

Table S1. European Nucleotide Archive (ENA) accession numbers for novel transcripts containing one or more ShK-like domain from *Oulactis* sp. tentacle transcriptomes.

Gene name	ENA Acc.	Gene name	ENA Acc.	Gene name	ENA Acc.
U-AITX-Oulsp21	LR535732	U-AITX-Oulsp37	LR535959	U-AITX-Oulsp57	LR535979
U-AITX-Oulsp22	LR535731	U-AITX-Oulsp38	LR535960	U-AITX-Oulsp58	LR535981
U-AITX-Oulsp23a	LR53573–4	U-AITX-Oulsp39	LR535962	U-AITX-Oulsp59	LR535982
U-AITX-Oulsp23b	LR535735	U-AITX-Oulsp40	LR535961	U-AITX-Oulsp60a	LR535983
U-AITX-Oulsp24a	LR535888	U-AITX-Oulsp41	LR535963	U-AITX-Oulsp60b	LR535984
U-AITX-Oulsp24b	LR535889	U-AITX-Oulsp42	LR535964	U-AITX-Oulsp61	LR535986
U-AITX-Oulsp24c	LR535890	U-AITX-Oulsp43	LR535965	U-AITX-Oulsp62	LR535985
U-AITX-Oulsp25a	LR535883–7	U-AITX-Oulsp44	LR535966	U-AITX-Oulsp63	LR535987
U-AITX-Oulsp25b	LR535899	U-AITX-Oulsp45	LR535967	U-AITX-Oulsp64	LR535988
U-AITX-Oulsp26	LR535891	U-AITX-Oulsp46	LR535968	U-AITX-Oulsp65	LR535989
U-AITX-Oulsp27	LR535900	U-AITX-Oulsp47	LR535969	U-AITX-Oulsp66	LR535990
U-AITX-Oulsp28	LR535949	U-AITX-Oulsp48	LR535970	U-AITX-Oulsp67	LR535991
U-AITX-Oulsp29	LR535950	U-AITX-Oulsp49	LR535972	U-AITX-Oulsp68	LR535992
U-AITX-Oulsp30	LR535951	U-AITX-Oulsp50	LR535971	U-AITX-Oulsp69	LR535993
U-AITX-Oulsp31	LR535952	U-AITX-Oulsp51a	LR535974		
U-AITX-Oulsp32a	LR535953	U-AITX-Oulsp51b	LR535975		
U-AITX-Oulsp32b	LR535954	U-AITX-Oulsp52	LR535973		
U-AITX-Oulsp33	LR535955	U-AITX-Oulsp53	LR535978		
U-AITX-Oulsp34	LR535957	U-AITX-Oulsp54	LR535976		
U-AITX-Oulsp35	LR535956	U-AITX-Oulsp55	LR535977		
U-AITX-Oulsp36	LR535958	U-AITX-Oulsp56	LR535980		

Nucleotide sequences lodged in EMBL-ENA (European Nucleotide Archive) under project PRJEB31254. Naming convention follows Oliveira et al. (2012) with modification; U – unknown function, **AI** – Family: Actiniidae, **TX** – toxin, **Oulsp** – *Oulactis* sp, the lowercase letter (e.g. **a**) at the end of the gene number indicates related gene isoforms.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

Table S2. Main biophysical properties for PC axes for all sequences.

consensus	MSA_col	property	PC1_load	consensus	MSA_col	property	PC2_load	consensus	MSA_col	property	PC3_load
A	61	RMW	-0.25	E	35	DISORD	0.30	K	57	CHRG	0.22
D	3	CHRG	0.24	W	23	RMW	0.22	G	64	RMW	0.22
E	35	RMW	-0.23	E	35	HPATH	-0.21	E	35	DISORD	-0.21
W	23	RMW	-0.22	-	40	DISORD	0.21	M	48	HPATH	-0.18
D	5	CHRG	-0.19	W	23	DISORD	-0.20	P	60	RMW	-0.18
G	64	RMW	0.18	G	34	RMW	-0.20	-	40	RMW	0.17
G	34	RMW	0.18	F	47	RMW	0.19	W	23	RMW	0.16
A	61	HPATH	0.18	E	35	CHRG	-0.19	K	49	CHRG	0.16
A	61	CHRG	-0.16	D	5	HPATH	-0.17	W	23	DISORD	-0.16
D	3	RMW	0.15	A	24	RMW	-0.16	-	65	HPATH	0.16
W	23	HPATH	0.15	F	47	DISORD	-0.16	F	47	DISORD	0.16
K	4	CHRG	-0.15	-	65	RMW	-0.14	K	4	RMW	0.15
-	17	DISORD	-0.15	N	39	HPATH	-0.14	F	47	HPATH	-0.14
D	3	DISORD	-0.13	N	38	CHRG	0.13	N	37	CHRG	-0.14
K	49	CHRG	0.11	G	64	RMW	0.12	A	24	RMW	-0.14
G	34	DISORD	-0.11	N	19	CHRG	0.12	M	48	CHRG	0.14
-	17	RMW	0.11	P	60	CHRG	0.09	K	49	HPATH	-0.13
D	5	HPATH	-0.11	K	49	HPATH	0.09	A	61	CHRG	-0.13
K	2	DISORD	0.11	P	60	RMW	0.09	P	60	HPATH	0.13
W	23	DISORD	0.11	A	24	HPATH	0.09	N	19	HPATH	0.13

Top 20 most highly loaded residue properties for the three PC axes (excluding occupancy) for the sequence space analysis performed with full sequence set. MSA_col refers to the MSA column in supplementary data 3. Consensus refers to the most commonly occurring residue in that column. RMW = side chain molecular weight, HPATH = hydropathy, CHRG = charge, DISORD = disorder propensity, CYS = cysteine.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

Table S3. Main biophysical properties for PC axes for ShKT domain sequences.

consensus	MSA_col	property	PC1a_load	consensus	MSA_col	property	PC2a_load	consensus	MSA_col	property	PC3a_load
E	35	DISORD	0.30	E	35	DISORD	0.27	G	64	CHRG	0.24
W	23	RMW	0.24	P	40	RMW	-0.22	P	40	DISORD	-0.17
E	35	CHRG	-0.22	K	4	RMW	-0.21	M	48	HPATH	-0.17
W	23	DISORD	-0.22	W	23	RMW	-0.20	G	64	RMW	0.17
P	40	DISORD	0.22	K	4	HPATH	0.19	P	40	HPATH	0.16
E	35	HPATH	-0.20	K	4	CHRG	-0.19	-	46	RMW	-0.16
N	5	HPATH	-0.18	D	19	HPATH	-0.19	P	60	RMW	-0.16
F	47	RMW	0.17	F	65	HPATH	-0.17	D	3	CHRG	0.15
G	34	RMW	-0.17	W	23	DISORD	0.16	M	48	CHRG	0.15
N	38	CHRG	0.16	K	57	CHRG	-0.16	F	47	DISORD	0.14
A	24	RMW	-0.15	G	64	RMW	-0.16	S	18	DISORD	0.14
F	65	RMW	-0.15	K	49	CHRG	-0.16	K	4	DISORD	-0.14
N	39	HPATH	-0.14	K	49	HPATH	0.15	P	60	HPATH	0.14
D	19	CHRG	0.13	K	49	DISORD	-0.15	K	4	CHRG	-0.14
F	47	DISORD	-0.12	F	65	DISORD	0.14	-	46	HPATH	0.14
K	49	HPATH	0.11	T	37	CHRG	0.14	F	47	HPATH	-0.13
G	64	RMW	0.11	N	39	CHRG	-0.14	K	57	RMW	-0.13
P	60	CHRG	0.10	E	35	RMW	-0.13	F	65	HPATH	0.13
P	60	RMW	0.09	-	32	RMW	-0.13	K	61	CHRG	-0.13
F	65	DISORD	0.09	-	32	CHRG	-0.12	W	23	RMW	0.13

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

Table S4. Main biophysical properties for PC axes for CRISP sequences.

consensus	MSA_col	property	PC1b_load	consensus	MSA_col	property	PC2b_load	consensus	MSA_col	property	PC3b_load
S	34	RMW	-0.31	K	24	HPATH	-0.29	K	58	DISORD	0.27
S	34	DISORD	0.25	K	24	CHRG	0.27	L	16	DISORD	-0.26
L	47	DISORD	-0.25	L	47	HPATH	0.25	K	21	CHRG	0.25
F	64	DISORD	-0.21	S	57	CHRG	-0.23	L	16	HPATH	0.23
F	64	HPATH	0.20	K	24	DISORD	0.23	G	35	RMW	0.22
V	48	HPATH	0.18	F	64	DISORD	0.20	L	47	RMW	0.21
L	47	RMW	0.18	Y	3	DISORD	-0.18	K	2	HPATH	-0.21
G	35	HPATH	-0.16	L	47	RMW	-0.18	K	58	CHRG	0.20
K	58	DISORD	0.14	Q	33	HPATH	0.17	E	4	DISORD	0.19
S	22	RMW	-0.14	K	24	RMW	0.16	K	49	CHRG	0.19
E	4	HPATH	-0.14	G	35	RMW	-0.16	L	23	HPATH	0.18
K	32	CHRG	0.13	Y	17	RMW	-0.15	K	58	HPATH	-0.16
L	16	HPATH	-0.13	L	16	HPATH	0.14	S	57	HPATH	-0.14
V	48	CHRG	-0.13	H	38	RMW	0.14	L	23	DISORD	-0.13
P	60	DISORD	0.13	K	58	CHRG	0.13	K	2	CHRG	0.13
V	48	DISORD	-0.13	Q	33	DISORD	-0.13	S	34	RMW	0.13
Y	17	RMW	-0.12	K	2	DISORD	0.13	E	39	CHRG	-0.13
L	47	CHRG	0.12	F	64	HPATH	-0.13	E	4	RMW	-0.13
S	57	RMW	-0.12	S	34	HPATH	0.13	K	2	DISORD	0.12
K	24	RMW	-0.12	V	48	HPATH	0.13	S	57	DISORD	0.12

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

Table S5. Phyla composition and exemplar function or process for each subcluster in the ShKT superfamily.

Individual sequence data may be found in Supplementary data file 2.

PCA Cluster	Phyla present	CRISP ShKT CRD	Single ShKT domains	ShKT Multi-domain/repeats	Functionality example
1 (blue)	Angiosperm Bacillariophyta Chlorophyta Chordata Cnidaria Nematoda Haptophyta Platyhelminthes Retaria	Absent	0.5% group (2 cnidarian sequences)	Present (99.5% cluster members)	Sequences functionally uncharacterised
2 (white)	Bacillariophyta Basidiomycota Chlorophyta Chordata Cnidaria Nematoda Plankton Platyhelminthes	Absent	Single ShKT domains found only in Cnidaria	Representatives in all phyla	**ShK; Cnidaria**; K_V1.3 blocker (Castañeda et al., 1995) **BgK; Cnidaria**; K_V1.1, 1.2 and 1.3 blocker (Cotton et al., 1997) NEP3:Cnidaria; neurotoxic (Columbus-Shenkar et al., 2018) **Aurelin**; **Cnidaria**; antimicrobial (Ovchinnikova et al., 2006) **BmK, AcK; Nematoda**; K_V1.3 blocker (Chhabra et al., 2014) **Mab-7**; **Nematoda**; morphogenesis
3 (grey)	Chordata	Absent	Absent	Present	**MMP23 family**; cellular function controlling potassium trafficking (Galea et al., 2014)
4 (purple)	Nematoda	Absent	*Ascaris suum**	All *C. elegans*	Sequences functionally uncharacterised
5 (orange)	Nematoda (*C. elegans)*	Absent	Absent	Present	Sequences functionally uncharacterised
6 (maroon)	Nematoda (*C. elegans)*	Absent	Absent	Present	Sequences functionally uncharacterised
7 (red)	Nematoda (*C. elegans)*	Absent	Absent	Present	Sequences functionally uncharacterised
8 (yellow)	Arthropoda Chordata	Insecta: Lepidoptera only	Absent	Absent	**CRISP1**; **Mammalia**; rat epididymis (Cameo and Blaquier, 1976) **CRISP3**; **Mammalia**; salivary gland; androgen control

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

PCA Cluster	Phyla present	CRISP ShKT CRD	Single ShKT domains	ShKT Multi-domain/repeats	Functionality example
					(Haendler et al., 1993) **CRISP4**; **Mammalia;** epididymis (Jalkanen et al., 2005) Arthropoda sequences are functionally uncharacterised
9 (black)	Arthropoda	Insecta: Lepidoptera not present	Absent	Absent	Sequences functionally uncharacterised
10 (green)	Mammalia	Present	Absent	Absent	**CRISP2**; **Primate**; autoantigen epididymal spermatozoa (Hardy et al., 1988) **Tpx-1**; **Rodentia;** inhibits cardiac ryanodine receptor (RyR) 2 (Gibbs et al., 2006)
11 (magenta)	Reptilia: Squamata only	Present	Absent	Absent	**PsTx**; cyclic nucleotide-gated ion channel modulatory – sensory transduction by retinal photoreceptors & olfactory neurons (Brown et al., 1999) **Stecrisp**; functionally uncharacterised (Guo et al., 2005) **Triflin**; Ca+ channel blocker causing artery contraction and muscle smoothing (Yamazaki et al., 2002)

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

10 Supplementary Figures

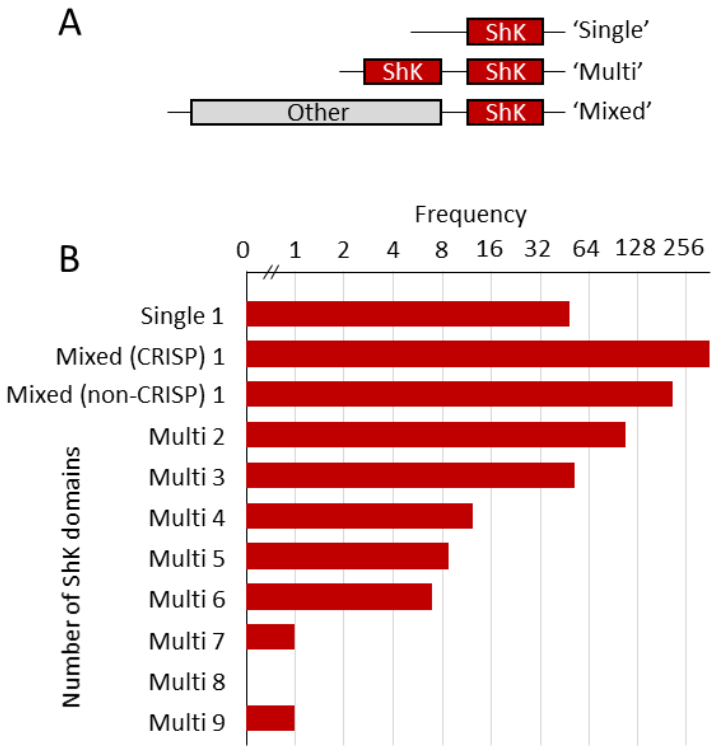


Fig. S1. Domain organisation of ShKT-containing proteins.

A) Schematic of general domain organisations. ShKT domains occur alone, in repeats, or in conjunction with other domains (often proteases). **B)** Number of sequences of each type in dataset (log scale). Mixed-domain proteins are separated between CRISPs (other domain = PR-1) and non-CRISPs (other domain ≠ PR-1).

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

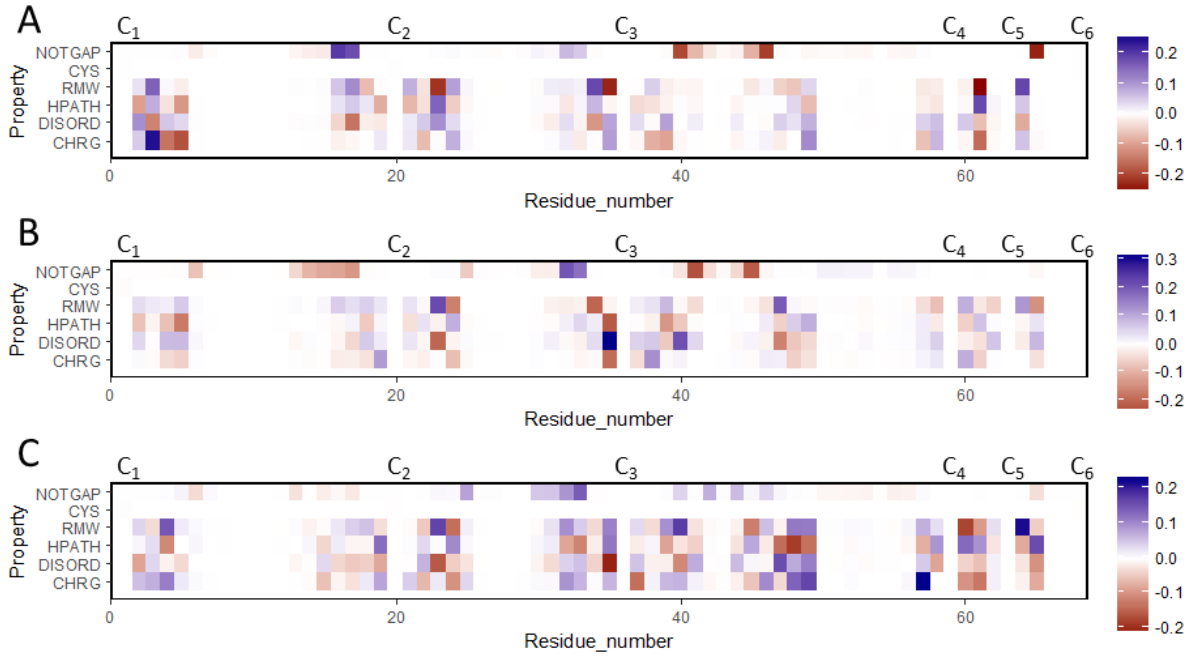


Fig. S2. Sequence properties that influence ShKT placement in sequence space.

Heatmap of biophysical property loadings for **A)** principal component axis 1 **B)** principal component axis 2, **C)** principal component axis 3. Axes are as in Fig 2. Biophysical properties with large positive and negative values strongly influence axis position. Biophysical properties with values near zero do not influence axis position. For the column-averages of these loadings mapped onto the protein structure, see Fig 5. Residue_number refers to the MSA column in supplementary data 3.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

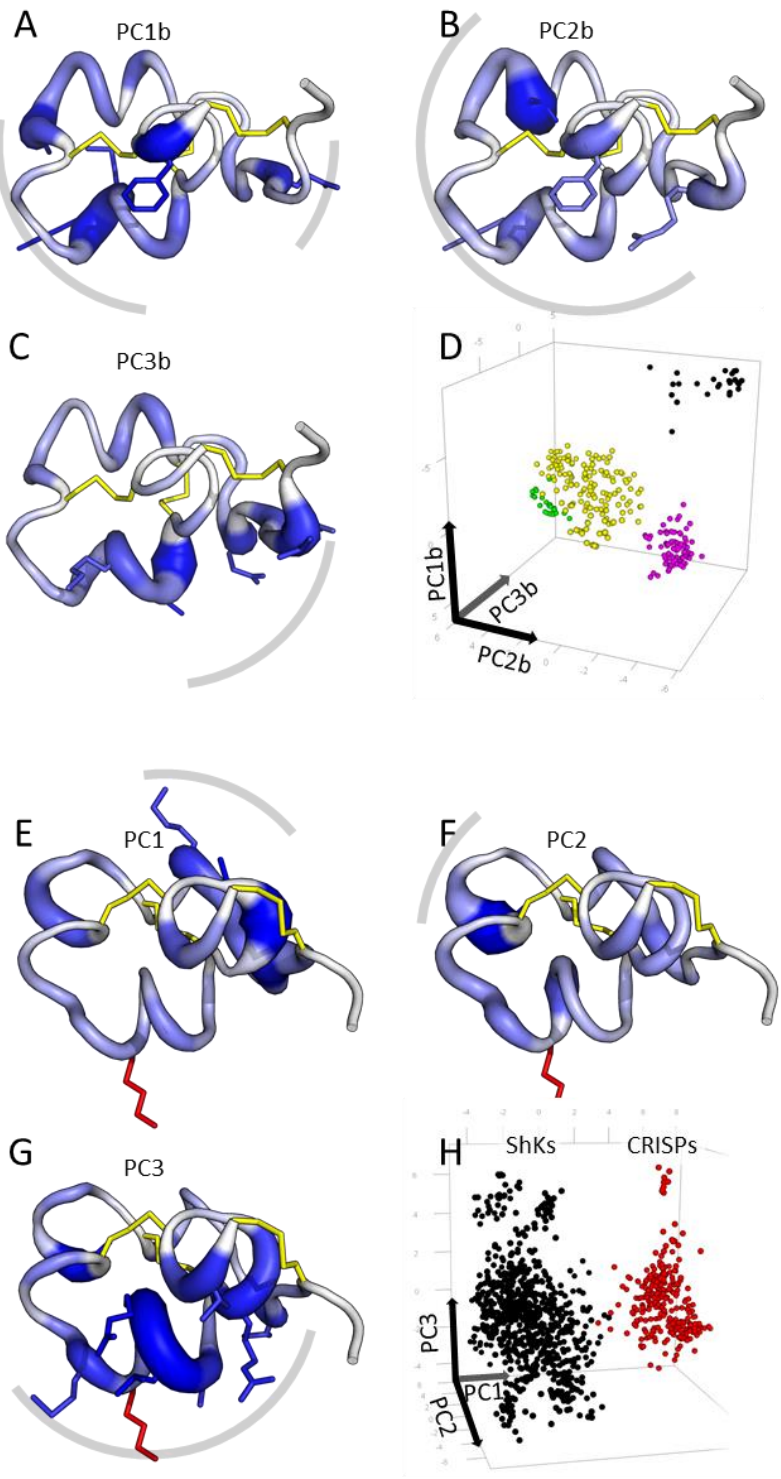


Fig. S3. Locations of residues that affect sequence space coordinates for CRISP cluster and full superfamily.

(A-C) For each principal component axis (PC1-3) combined loading across all properties for each residue mapped onto structure of CRISP CRD domain of triflin (PDB: 1WVR). Property loading magnitude indicated by thickness and colour intensity. Residues with average combined loading in top decile shown as sticks. General region containing the majority of highly-loaded residues indicated by grey arc. Inhibitory lysine shown as red stick. **(D)** Seq space as in Fig. 3E for reference.

(E-G) For each principal component axis (PC1-3) combined loading across all properties for each residue mapped onto structure of ShK (PDB: 1ROO). Property loading magnitude indicated by thickness and colour intensity. Residues with average combined loading in top decile shown as sticks. General region containing the majority of highly-loaded residues indicated by grey arc. Inhibitory lysine shown as red stick. **(H)** Seq space as in Fig. 3A for reference.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

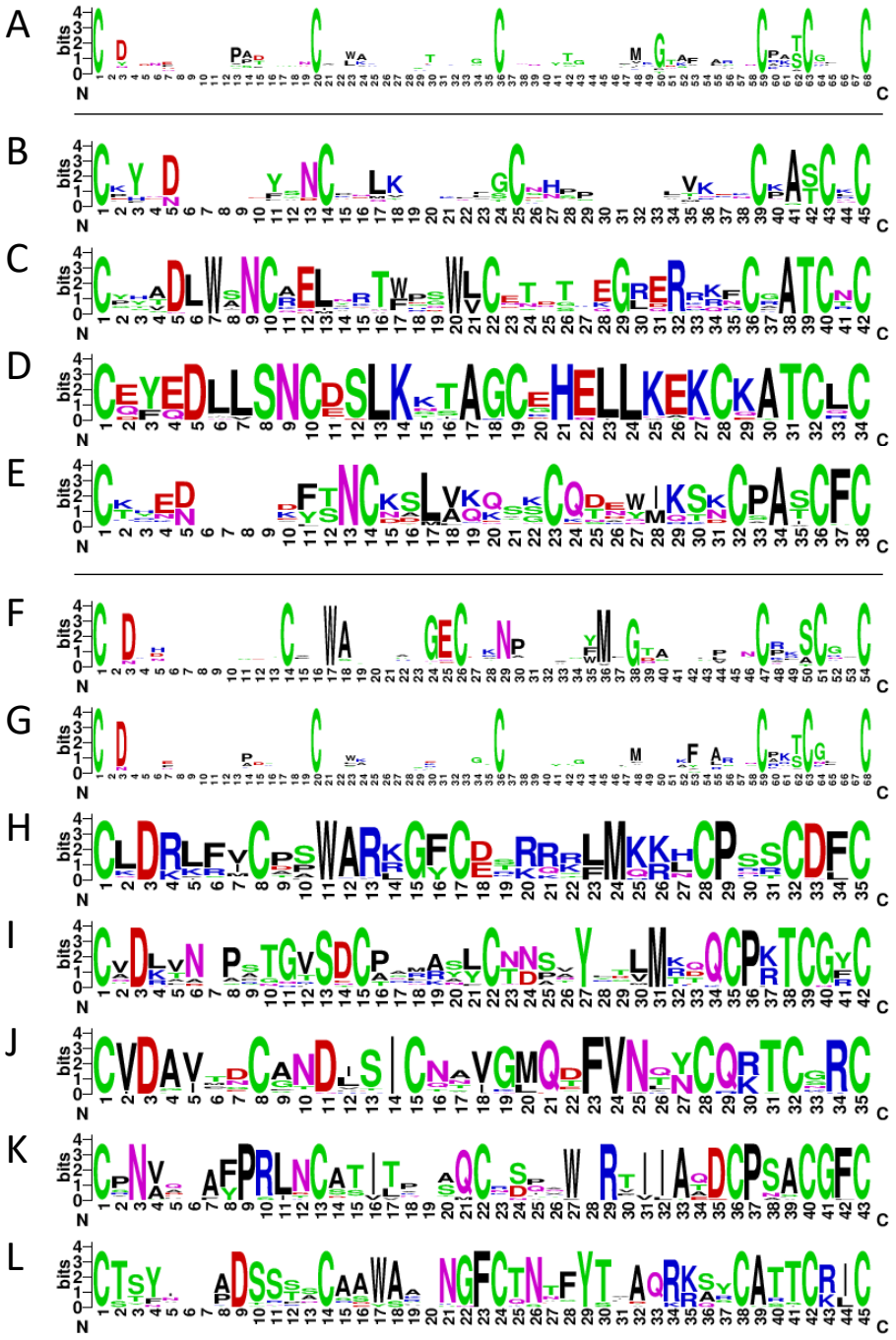


Fig. S4. Sequence logo plots.

A) ShKT superfamily alignment of all ShK-like and the CRISP cysteine-rich domains (CRD) analysed **B-E)** CRISP CRD clusters **F-L)** ShK-like clusters. Alignments are shown in supplementary data 4.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

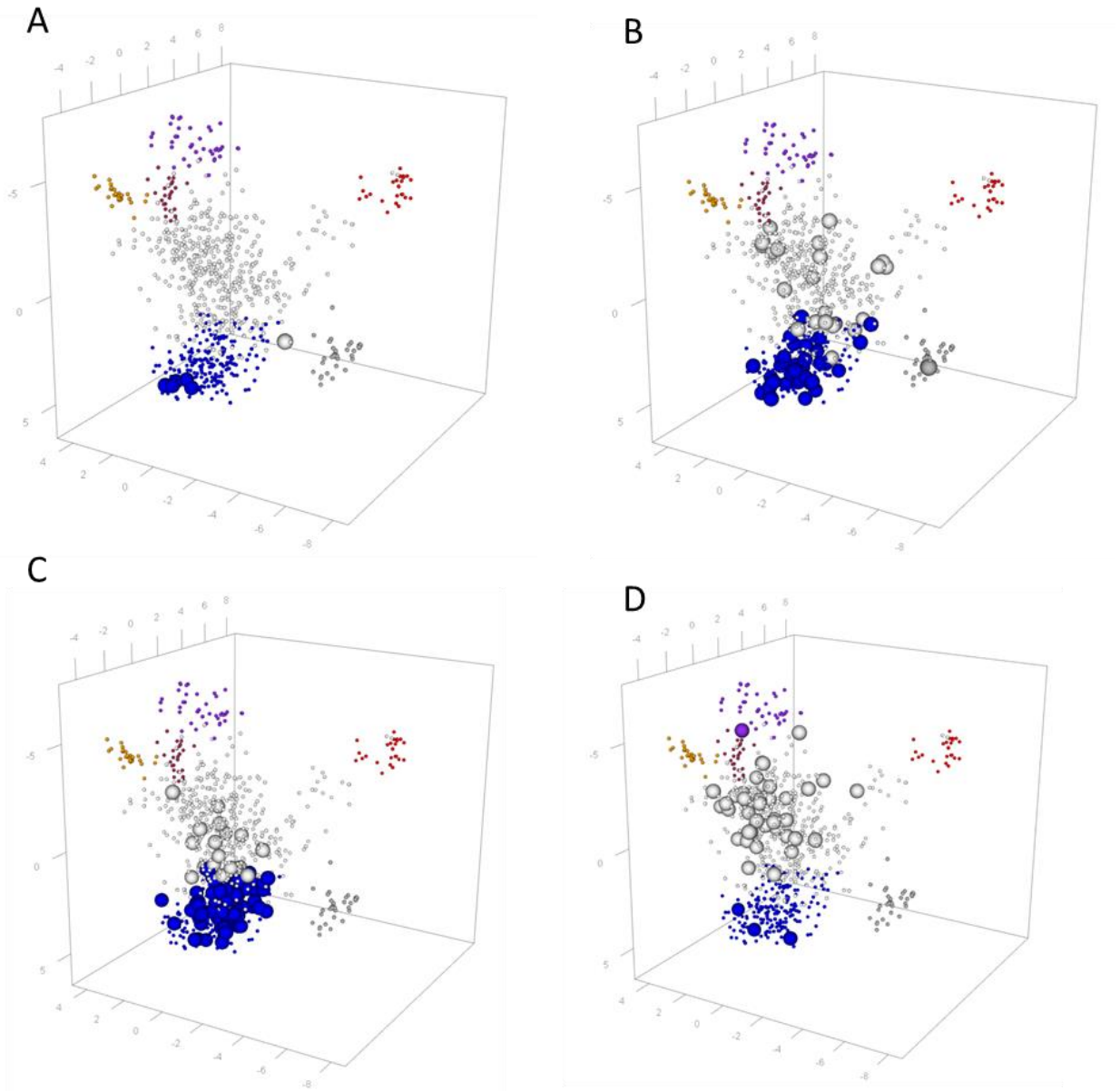


Fig. S5. Example groups within in sequence space.

The sequence space of ShK-like members, with larger spheres indicating; **A)** plant sequences **B)** algal sequences **C)** planktonic sequences **D)** single-domain toxins. Coloured as in Fig. 3.

Shafee, T., Mitchell, M. L., & Norton, R. S. (2019). Mapping the chemical and sequence space of the ShKT superfamily. *Toxicon*, *165*, 95-102. doi:10.1016/j.toxicon.2019.04.008

11 Supplementary Data Files

Data File S1. Rotatable 3D sequence space models.
A) ShKs and CRISPs combined, B) ShKs only, C) CRISPS only.

Data File S2. CRISP and ShK-like labels.
Data annotations for all sequences used in analysis.

Data File S3. Multiple sequence alignment of ShKT domains.
Fasta MSA of all domain sequences used in analysis

Data File S4. Subalignments of each CRISP and ShK-like group.
Fasta MSA of each sequence group from analysis

Data File S5. Rotatable 3D sequence space with network overlay.
Fasta MSA of all sequences used in analysis

12 Supplementary references

Brown, R.L., Haley, T.L., West, K.A., Crabb, J.W., 1999. Pseudechetoxin: A peptide blocker of cyclic nucleotide-gated ion channels. Proc. Natl. Acad. Sci. 96, 754 LP-759.

Cameo, M.S., Blaquier, J.A., 1976. Androgen-controlled specific proteins in rat epididymis. J. Endocrinol. 69, 47–55. https://doi.org/10.1677/joe.0.0690047

Castañeda, O., Sotolongo, V., Amor, A.M., Stöcklin, R., Anderson, A.J., Harvey, A.L., Engström, Å., Wernstedt, C., Karlsson, E., 1995. Characterization of a potassium channel toxin from the Caribbean sea anemone *Stichodactyla helianthus*. Toxicon 33, 603–613. https://doi.org/10.1016/0041-0101(95)00013-C

Chhabra, S., Chang, S.C., Nguyen, H.M., Huq, R., Tanner, M.R., Londono, L.M., Estrada, R., Dhawan, V., Chauhan, S., Upadhyay, S.K., Gindin, M., Hotez, P.J., Valenzuela, J.G., Mohanty, B., Swarbrick, J.D., Wulff, H., Iadonato, S.P., Gutman, G.A., Beeton, C., Pennington, M.W., Norton, R.S., Chandy, K.G., 2014. KV1.3 channel-blocking immunomodulatory peptides from parasitic worms: implications for autoimmune diseases. FASEB J. 28, 3952–3964. https://doi.org/10.1096/fj.14-251967

Columbus-Shenkar, Y.Y., Sachkova, M.Y., Macrander, J., Fridrich, A., Modepalli, V., Reitzel, A.M., Sunagar, K., Moran, Y., 2018. Dynamics of venom composition across a complex life cycle. Elife 7, e35014. https://doi.org/10.7554/eLife.35014

Cotton, J., Crest, M., Bouet, F., Alessandri, N., Gola, M., Forest, E., Karlsson, E., Castaneda, O., Harvey, A.L., Vita, C., Menez, A., 1997. A potassium-channel toxin from the sea anemone *Bunodosoma Granulifera*, an inhibitor for Kv1 channels - revision of the amino acid sequence, disulfide-bridge assignment, chemical synthesis, and biological activity. Eur. J. Biochem. 244, 192–202. https://doi.org/10.1111/j.1432-1033.1997.00192.x

Galea, C.A., Nguyen, H.M., George Chandy, K., Smith, B.J., Norton, R.S., 2014. Domain structure and function of matrix metalloprotease 23 (MMP23): role in potassium channel trafficking. Cell. Mol. Life Sci. 71, 1191–1210. https://doi.org/http://dx.doi.org/10.1007/s00018-013-1431-0

Gibbs, G.M., Scanlon, M.J., Swarbrick, J., Curtis, S., Gallant, E., Dulhunty, A.F., O'Bryan, M.K., 2006. The cysteine-rich secretory protein domain of Tpx-1 is related to ion channel toxins and regulates ryanodine receptor Ca2+ signaling. J. Biol. Chem. 281, 4156–63. https://doi.org/10.1074/jbc.M506849200

Guo, M., Teng, M., Niu, L., Liu, Q., Huang, Q., Hao, Q., 2005. Crystal structure of the cysteine-rich secretory protein stecrisp reveals that the cysteine-rich domain has a K+ channel

inhibitor-like fold. J. Biol. Chem. 280, 12405–12. https://doi.org/10.1074/jbc.M413566200

Haendler, B., Krätzschmar, J., Theuring, F., Schleuning, W.D., 1993. Transcripts for cysteine-rich secretory protein-1 (CRISP-1; DE/AEG) and the novel related CRISP-3 are expressed under androgen control in the mouse salivary gland. Endocrinology 133, 192–198. https://doi.org/10.1210/endo.133.1.8319566

Hardy, D.M., Huang, T.T.F., Driscoll, W.J., Tung, K.S.K., Wild, G.C., 1988. Purification and characterization of the primary acrosomal autoantigen of guinea pig epididymal spermatozoa. Biol. Reprod. 38, 423–437. https://doi.org/10.1095/biolreprod38.2.423

Jalkanen, J., Huhtaniemi, I., Poutanen, M., 2005. Mouse cysteine-rich secretory protein 4 (CRISP4): A member of the crisp family exclusively expressed in the epididymis in an androgen-dependent manner. Biol. Reprod. 72, 1268–1274. https://doi.org/10.1095/biolreprod.104.035758

Oliveira, J.S., Fuentes-Silva, D., King, G.F., 2012. Development of a rational nomenclature for naming peptide and protein toxins from sea anemones. Toxicon 60, 539–550. https://doi.org/10.1016/j.toxicon.2012.05.020

Ovchinnikova, T. V, Balandin, S. V, Aleshina, G.M., Tagaev, A.A., Leonova, Y.F., Krasnodembsky, E.D., Men'shenin, A. V, Kokryakov, V.N., 2006. Aurelin, a novel antimicrobial peptide from jellyfish *Aurelia aurita* with structural features of defensins and channel-blocking toxins. Biochem. Biophys. Res. Commun. 348, 514–523. https://doi.org/10.1016/j.bbrc.2006.07.078

Yamazaki, Y., Koike, H., Sugiyama, Y., Motoyoshi, K., Wada, T., Hishinuma, S., Mita, M., Morita, T., 2002. Cloning and characterization of novel snake venom proteins that block smooth muscle contraction. Eur. J. Biochem. 269, 2708–2715. https://doi.org/10.1046/j.1432-1033.2002.02940.x