Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

A centipede toxin family defines an ancient class of CSαβ defensins

Thomas S. Dash[#1], Thomas Shafee[#3], Peta J. Harvey[1], Chuchu Zhang[4], Steve Peigneur[5], Jennifer R. Deuis[1], Irina Vetter[1,6], Jan Tytgat[5], Marilyn Anderson[3], David J. Craik[1], Thomas Durek*[1], Eivind A. B. Undheim*[2]

[1] *Institute for Molecular Bioscience,* [2]*Centre for Advanced Imaging, and* [6]*School of Pharmacy, The University of Queensland, St Lucia, Queensland, 4072, Australia*

[3] *La Trobe Institute for Molecular Science, La Trobe University, Victoria, 3083, Australia*

[4] *Department of Physiology, University of California, San Francisco, CA 94143, USA*

[5] *Toxicology and Pharmacology, University of Leuven, Leuven, 3000, Belgium*

[#] These authors contributed equally

* Address for correspondence: e.undheim@uq.edu.au and t.durek@imb.uq.edu.au

Summary

Disulfide rich peptides (DRPs) play diverse physiological roles and have emerged as attractive sources of pharmacological tools and drug leads. Here we describe the 3D structure of a centipede venom peptide, U-SLPTX$_{15}$-Sm2a, whose family defines a unique class of one of the most widespread DRP folds known, the cystine-stabilised α/β fold (CSαβ). This class, which we have named the two-disulfide CSαβ fold (2ds-CSαβ), contains only two internal disulfide bonds as opposed to at least three in all other confirmed CSαβ peptides, and constitutes one of the major neurotoxic peptide families in centipede venoms. We show the 2ds-CSαβ is widely distributed outside centipedes, and is likely an ancient fold predating the split between prokaryotes and eukaryotes. Our results highlight the usefulness of 3D structures as evolutionary tools and provides some of the first insights into the ancient evolutionary history of any DRP fold.

Introduction

Peptides derived from animal venoms, generally referred to as toxins, have evolved over millions of years to target ion channels, receptors and other membrane proteins in the organisms they are injected into, often with great potency. The majority of these peptides contain a large proportion of cysteine residues that form intramolecular disulfide bonds, which confer a high degree of structural rigidity and stability. This unique combination of stability and pharmacology has attracted substantial interest due to the use of toxins as molecular tools and potential for development into therapeutics or other high-value products (Kalia et al., 2015; King, 2011). Despite the increased attention to their potential uses, however, we know very little about the evolutionary history of most toxins and thus also the traits that underlie their exceptional biophysical and pharmacological properties (Undheim et al., 2016b).

Although the utilization of venom peptides is widespread throughout the animal kingdom, each venomous lineage has evolved its own intricate venom cocktail for primary prey capture and defence. However, it is now clear that the functional and pharmacological diversity of venom peptides is achieved by a comparatively small number of conserved structural folds (Fry et al., 2009; Undheim et al., 2016b). These 'privileged scaffolds' are thought to have functionally diversified from common housekeeping peptides through gene duplication and mutation events while maintaining the structure-stabilizing cysteine pattern and connectivity. Examples of such scaffolds include the inhibitory cystine knot (ICK) motif found in many spider toxins, the three-finger fold common in elapid snake toxins, and the cysteine-stabilized αβ motif of the '*cis*-defensin' superfamily (henceforth CSαβ) frequently encountered in scorpion toxins. Strikingly, these three scaffolds together make up almost 50 % of all structurally characterized disulfide constrained peptides (1–10 kDa), with CSαβ peptides comprising a marginally greater proportion than ICK peptides (19% *vs* 18.5%) (Undheim et al., 2016b). CSαβ peptides are also taxonomically widespread outside venoms, with representatives identified from both opisthokont (including animals and fungi) and plant lineages, where they often serve important roles in the defence against pathogens ('defensins') (Shafee et al., 2017).

As the name suggests, the CSαβ fold is characterized by an α-helix that is anchored to a β-sheet via two disulfide bonds. These disulfides are formed between cystines that are spatially adjacent in the helix (i, i+4) and spatially adjacent in one of the β-strands (i, i+2), resulting in a characteristic cysteine-signature motif of 'CXXXC–X_n–CXC', where X denotes any amino acid residue. All confirmed CSαβ peptides also contain at least one additional stabilizing disulfide bond connecting the N-terminal cap region of the α-helix with another β-strand that is N-terminal to that containing the two canonical cysteines (Shafee et al., 2017). CSαβ peptides are thought to share a common evolutionary origin (Shafee et al., 2016), with a three-disulfide

CSαβ (3ds-CSαβ) fold possibly representing the ancestral eukaryotic form (Shafee et al., 2017). Based on the discovery of a set of putative myxobacterial CSαβ amino acid sequences it has also been proposed that this form originated as a two-disulfide CSαβ (2ds-CSαβ) fold containing only the core disulfides (Zhu, 2007). However, it has yet to be confirmed whether these putative amino acid sequences do assume a CSαβ fold and whether 2ds-CSαβ peptides are found in lineages other than Myxobacteria.

At approximately 450 million years old, centipedes are among the most ancient extant venomous terrestrial lineages (Undheim and King, 2011). However, despite their ancient evolutionary history and notoriously painful stings, very little is known about their venoms: of the approximately 3,500 described species, only nine species have had their venom proteomes analysed in detail (Liu et al., 2012; Rates et al., 2007; Rong et al., 2015; Undheim et al., 2014; Yang et al., 2012). Moreover, most peptide toxins show no amino acid sequence similarity to any known peptides, indicating that they are a rich source of bioactive and structurally novel peptides (Undheim et al., 2015a). One of the main neurotoxic peptide families in venoms from the giant centipedes of the genus *Scolopendra* is the SLPTX15 family (Liu et al., 2012; Rong et al., 2015; Undheim et al., 2015a; Undheim et al., 2016a; Undheim et al., 2014). Curiously, the main defining characteristic of this family is four cysteines with the consensus amino acid sequence 'CXXXC-X_n-CXC' (n typically 20–25 amino acids), with some members also having an additional Cys on each end of this motif. The unusual, 2ds-CSαβ-like cysteine pattern and overall low amino acid sequence similarity to any know venom-derived toxins prompted us to determine the NMR structure of one of these peptides, U-SLPTX_{15}-Sm2a (GenBank accession GASH01000152; hereafter referred to as Sm2).

Here we show using 3D structural comparisons, structure-guided alignments, phylogenetics and multi-dimensional clustering methods that the centipede SLPTX15 family indeed represents a weaponized 2ds-CSαβ-type peptide fold. Interrogation of a taxonomically comprehensive amino acid sequence dataset revealed that the SLPTX15 2ds-CSαβ fold is not unique to centipede venoms, but is in fact a widespread but previously unrecognized family of the CSαβ fold. Our results show that the 2ds-CSαβ and classic CSαβ folds are both likely ancient forms of the CSαβ fold that predate the evolution of eukaryotes, and suggest that they may share a single evolutionary origin. These findings provide some of the first insight into the ancient evolutionary history of DRPs and highlight the value of 3D structures as evolutionary tools.

Results

Isolation and primary structure of Sm2

Sm2 was originally isolated in high purity from the venom of *Scolopendra morsitans*, where it is one of the major peptide components (SI Fig. S1A). However, the yield from milked venom was insufficient for detailed structural or pharmacological studies and we therefore synthesized Sm2 chemically. The disulfide connectivity of the four cysteine residues was initially examined using enzymatic digests of intact, oxidized native toxin and co-elution of native and synthetic Sm2. Briefly, trypsin digestion of native oxidized Sm2 did not produce peptide segments indicative of a CysI–CysII, CysIII–CysIV cysteine connectivity, but also did not allow us to distinguish between the two remaining CysI–CysIII, CysII–CysIV or CysI–CysIV, CysII–CysIII cysteine connectivities. These disulfide isomers were therefore both chemically synthesized using a directed disulfide formation approach via orthogonal protection of cysteine residues (Acm/Trt), which afforded the desired Sm2 disulfide isomers in good yield. Of the two isomers obtained, only the one having a CysI–CysIII, CysII–CysIV connectivity co-eluted with venom-derived Sm2 and yielded well dispersed NMR spectra, indicating that this synthetic isomer represents the disulfide connectivity of native Sm2 (SI Fig. S1D). Finally, to exclude the possibility of any isobaric posttranslational modifications (such as amino acid L/D isomerization), synthetic and venom-derived Sm2 were digested with trypsin and the resulting peptide fragments analyzed by LC-MS. Comparison of the tryptic peptides derived from the synthetic and native material indicated identical fragment patterns, strongly suggesting that the structures of both samples are identical (SI Fig. S1E).

Sm2 adopts a unique form of the CSαβ fold

Because of the unusual cysteine-pattern of Sm2 and its lack of amino acid sequence similarity to any other toxin families, we decided to determine its 3D solution structure by NMR. The backbone amide proton and nitrogen resonances of Sm2 are well dispersed (SI Fig. S2A, B), indicating a well-defined structure. Structure calculations were based upon 577 distance restraints, 21 hydrogen bond pairs, and a total of 103 dihedral angle restraints (Table 1). Both proline residues (Pro6 and Pro14) were determined to adopt the trans conformation due to strong Hδ(i)Pro-Hα(i-1) NOE correlations and the ^{13}C shifts of the Cβ and Cγ proline resonances (Wuthrich, 1986). The resulting assembly of the calculated 20 lowest energy structures overlaid well, with an RMSD for the backbone atoms of 0.69 ± 0.16 Å. The mean overall MolProbity (Chen et al., 2010) score of 1.6 indicated very good structural quality, with 97% of residues falling within the most favoured regions of the Ramachandran plot. Promotif (Hutchinson and Thornton, 1996) was used to confirm that the compact globular fold of Sm2 consisted of a triple stranded anti-parallel β-sheet (β1: Glu2 – Arg8; β2: Ile35 – Lys42; β3: Tyr45 – Ile51) tethered to

Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

an α-helix (Glu16 – Ala28) by the two disulfide bonds. Each of the three loops connecting these elements of secondary structure was well-defined. A type VIII β-turn (Lys11 – Pro14) was identified in the loop connecting the first strand to the α-helix; a type I' β-turn (Ala28 – Asp31) followed by a type I β-turn (Gln32 – Ile35) forms the loop between the helix and the β2 strand; and a β-hairpin turn incorporating a type II β-turn forms the short loop between the β2 and β3 strands. Both disulfide bonds are defined as left-hand spiral.

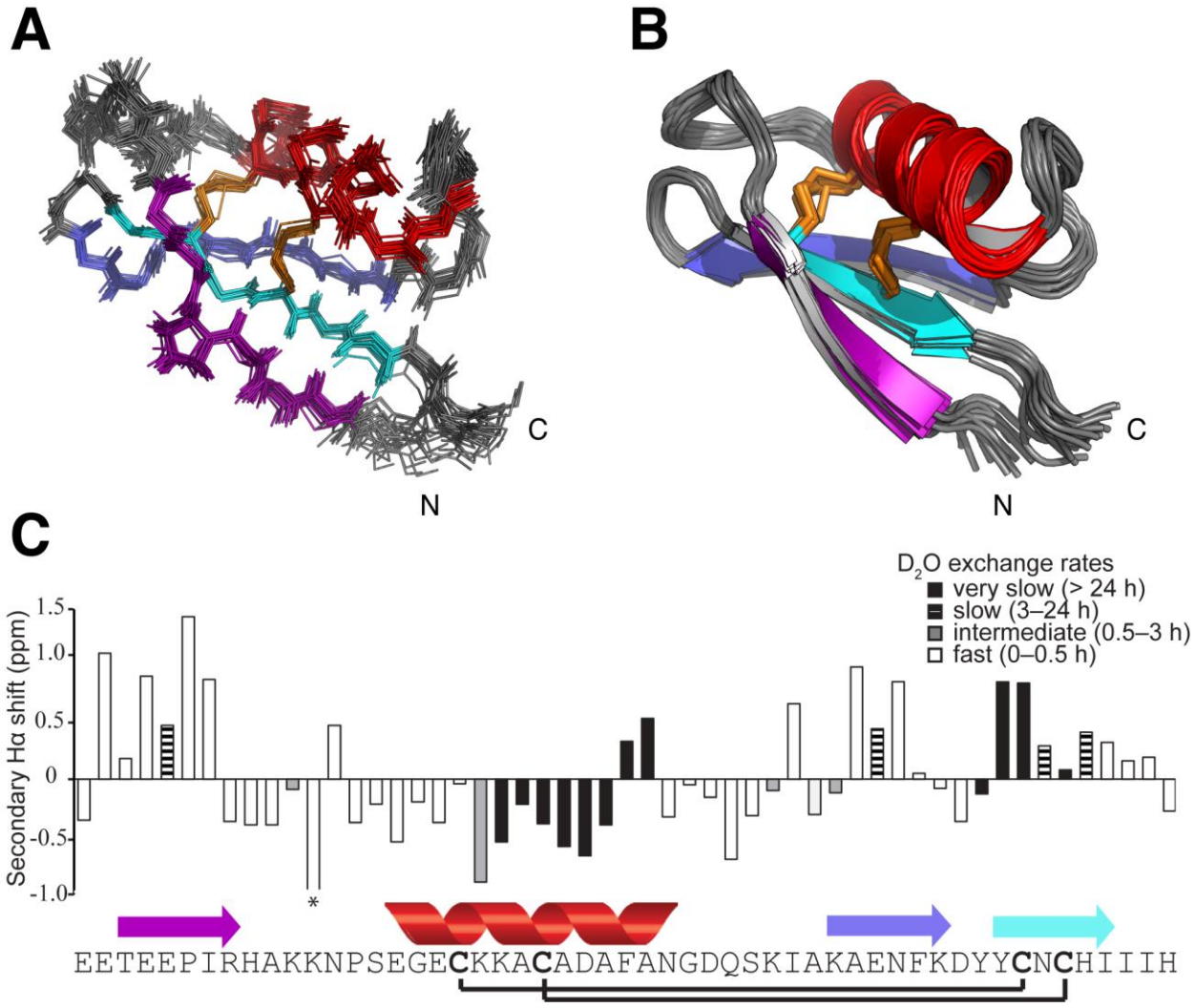

Figure 1. NMR structure of Sm2. Overlaid 20 lowest energy structures of Sm2 displayed as (A) line or (B) ribbon form, and (C) secondary Hα chemical shifts and H/D exchange kinetics. Disulfide bonds are shown in orange while elements of secondary structure are colour coded. See also Figure S1 and Table 1.

NMR analysis revealed that the structure of Sm2 indeed assumes a CSαβ fold frequently encountered among scorpion toxins and defensins from invertebrates, fungi and plants (Shafee et al., 2016; Shafee et al., 2017) (Fig. 1). However, comparison of Sm2 with these classical CSαβ peptides revealed the absence of an otherwise highly conserved third disulfide bond (Fig. 2).

Instead, the 3D structure of Sm2 is similar to another recently published centipede peptide, SsTx from the venom of *Scolopendra subspinipes mutilans* (Luo et a., 2018). Despite sharing only 32% sequence identity (26% excluding cysteines), the two 3D structures overlay with an RMSD of 3.1 Å along their peptide backbones and share near identical disulfide orientations (SI Fig. S2C). Whilst this does not seem a particularly tight fit, the elements of secondary structure overlay well despite Sm2a having a slightly longer helix and longer beta strands.

Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

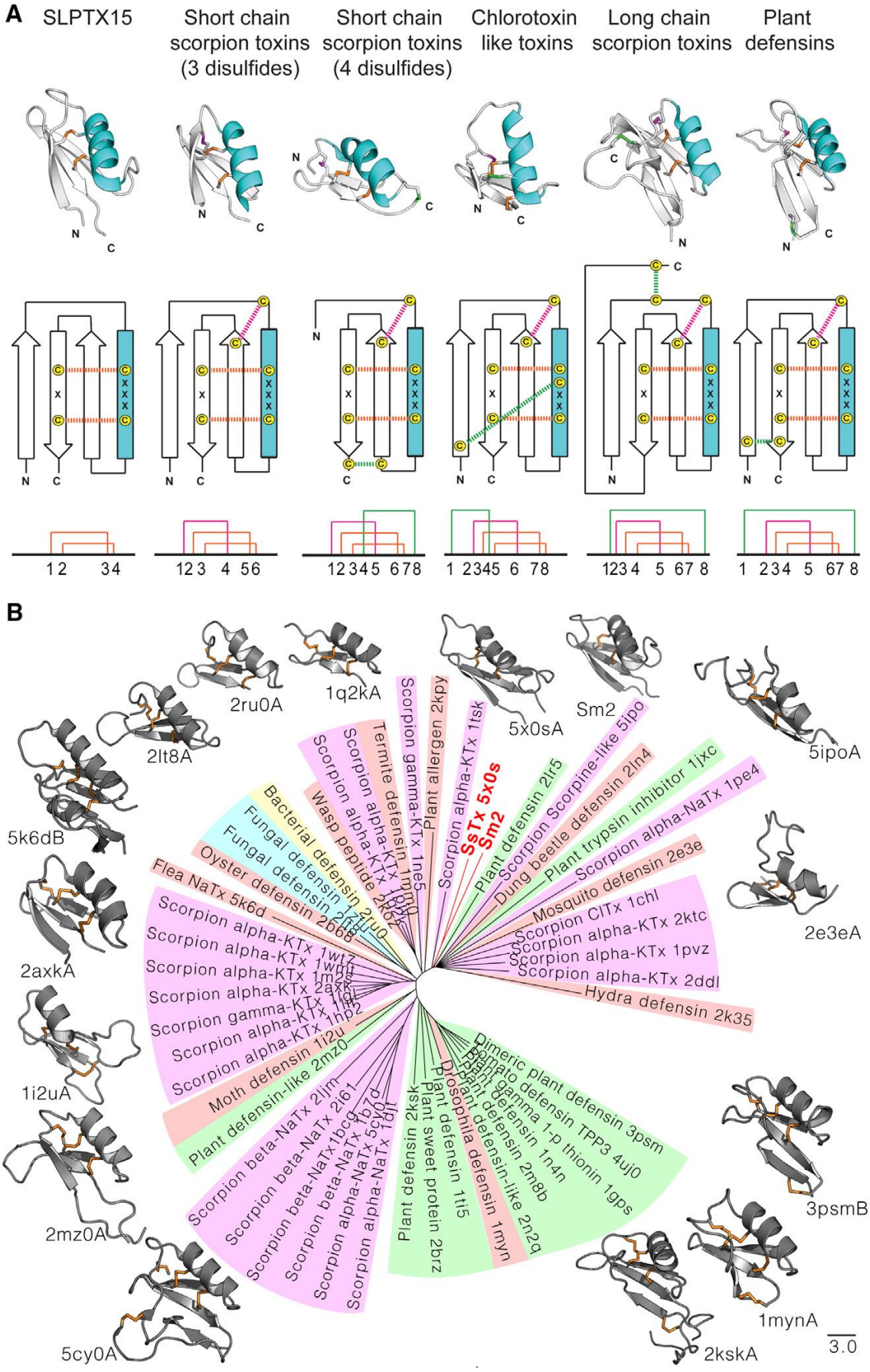

Figure 2. Sm2 represents a unique form of the CSαβ motif. (A) Structural variations of the CSαβ motif. Sm2 and five prominent CSαβ subfamilies are shown in 3D ribbon representation with the corresponding secondary structure topologies and disulfide connectivities shown below. The representative subfamily structures shown are: agitoxin 2 (PDB 1AGT, short chain scorpion toxins), HSTX1 (PDB 1QUZ, short chain scorpion toxins with 4 disulfides), chlorotoxin (PDB 1CHL, chlorotoxin-like peptides), LqhaIT (PDB 2ASC, long chain scorpion toxins), and NaD1 (PDB 4AAZ, plant defensins). The prototypical CSαβ motif is characterized by three conserved disulfide bonds that tether the α-helix to a β-sheet. Additional disulfides in variable positions are common in some of the subfamilies. In contrast, centipede toxin Sm2 represents a minimalists CSαβ that is stabilized by only two disulfide bonds. (B) Dendrogram of structural comparisons of structural homologues of Sm2 calculated using the DALI server all-against-all tool. Representative 3D structures are shown for each clade, while PDB and chain identifiers are provided for each structure at the tip label of each branch. Scorpion toxins are highlighted in pink, defensin peptides in red, fungal peptides in cyan, plant peptides in green, and bacterial peptides in yellow. KTx and NaTx indicate modulation of potassium and sodium channels, respectively. See also Figure S2 and Table S1.

However, some of this variation may also be due to the lower quality of the structure of SsTx as judged by MolProbity (All-atom clashscores of 46 for SsTx vs 9.1 for Sm2, and 10% vs 0.3% Ramachandran outliers for SsTx and Sm2, respectively). Although SsTx was not assigned to a centipede toxin family or a peptide fold, its preproprotein shares 89% amino acid sequence identity to that of a SLPTX15 peptide (NCBI accession KC145039) from the venom of *S. dehaani* (Liu et al., 2012; Undheim et al., 2014). Thus, our results demonstrate that SsTx is another member of SLPTX15 and that this toxin family is characterised by an unusual, and likely diverse form of the CSαβ fold.

The high quality of the Sm2 3D structure also provides insight into the properties of this unusual CSαβ fold. Electrostatic interactions are abundant throughout the peptide and appear to stabilize the overall fold, resulting in a high thermal stability despite having only two disulfides for 53 residues (SI Fig. S2D). Three such interactions are most notable. The first interaction involves the string of four glutamates at the N-terminus which aligns proximally with the C-terminal His49 and His53 of the neighbouring β-strand. Additionally, this also places the N and C-termini next to each other and a salt bridge between the termini can be inferred in 12 of the 20 structures. The second electrostatic interaction is between Lys12 and Asp43 which are located in loop 1 and 3 respectively (SI Fig. S2E). This interaction is between two amino acid sequence-distant secondary structure elements and can therefore effectively be thought of as a disulfide equivalent. The third interaction is a text-book example of α-helix stabilization by side chain interactions. Within the helix, Glu16 is i-4 to Lys20 and Lys21 is i-4 to Asp25. For each pair of residues noted, this places their side chains on the same side of the helix and in close proximity to form two salt bridges which likely contribute to helix stability. In typical globular fold fashion, Sm2 has a hydrophobic core built around its two disulfide bonds. The side chains of some hydrophobic residues (Ile7, Pro14, Phe27, Ile35, Ala38 and Ile50) protrude from the backbone into the core shielding the disulfide bonds.

In order to look for structural homologues of Sm2, we used the Dali server (Holm and Laakso, 2016) to search the 3D structure of Sm2 against all structures of proteins contained within the Protein Data Bank (PDB) that are composed of more than 30 amino acids. This search yielded 107 structures from 74 unique proteins with significant structural similarity (z-score > 2; SI Table S1) that were distributed across a wide taxonomic range of organisms. These hits also included multiple false positives (z-score < 3.7, rank > 45), as judged by a lack of any disulfides corresponding to those characteristic of the CSαβ fold, illustrating the susceptibility to false positives of this approach when used in isolation. The taxa from which CSαβ motifs were identified included both eukaryotes and prokaryotes, illustrating the extremely widespread and ancient nature of this peptide fold. All-against-all structural comparisons of Sm2 with its structural homologues followed by pairwise distance-based dendrogram construction revealed that Sm2 does not just have a unique disulfide pattern, but that the 2ds proteins adopt a unique variant of the overall CSαβ fold (Fig. 2B).

2ds-CSαβ: a widespread form of the classic CSαβ motif

Although homology of Sm2 and SsTx with the other CSαβ proteins is likely given their significant structural similarity, resolving their evolutionary history within the superfamily is challenging. We first investigated whether the centipede 2ds-CSαβ represents a one-off structural adaptation following weaponization from a non-toxin ancestor, as has occurred with the HAND toxins of the SLPTX03 toxin family (Undheim et al., 2015b), or whether it is a more broadly occurring variant of the classical CSαβ motif. BLAST and HMMER searches of UniProtKB (The UniProt Consortium, 2017) and GenBank (NCBI Resource Coordinators, 2017), supplemented by *de novo* assembled transcriptome datasets for an additional 16 myriapod species downloaded from the NCBI sequence reads archive (SRA; SI Table S2), revealed that the latter is indeed the case: the 2ds-CSαβ motif is not only found in the venomous centipedes, but also in several non-venomous myriapods. Moreover, phylogenetic analyses revealed that all myriapod 2ds-CSαβ peptides form a well-supported clade with respect to the remaining CSαβ peptides (Fig. S3). This clade contained 2ds-CSαβ peptides from both millipedes (Diplopoda) and centipedes (Chilopoda), which strongly suggests the 2ds-CSαβ fold did not evolve independently in centipedes and millipedes. Instead, our results suggest the 2ds-CSαβ fold evolved before the split of the Myriapoda some 500 million years ago (Fernández et al., 2016).

We further examined the distribution of the 2ds-CSαβ fold within Arthropoda. Although there are several well-studied arthropod species, the distribution of sequences in public databases is highly skewed to these few model species, meaning the overall taxonomic sampling of the Arthropoda in public databases is relatively poor. We therefore obtained and assembled an additional 114 arthropod transcriptome datasets from the SRA, which we used to supplement the publicly available and myriapod SRA sequences. In addition, we assembled four nematode

transcriptomes to provide ecdysozoan outgroup candidates (SI Table S2). The search of this dataset revealed that the 2ds-CSαβ fold is widespread in Arthropoda, and probably also Ecdysozoa: We found putative 2ds-CSαβ peptides in almost all major clades, including Pancrustacea, Myriapoda and Arachnida, as well as both the nematode groups Chromadorea and Enoplea (Fig. 3A).

Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

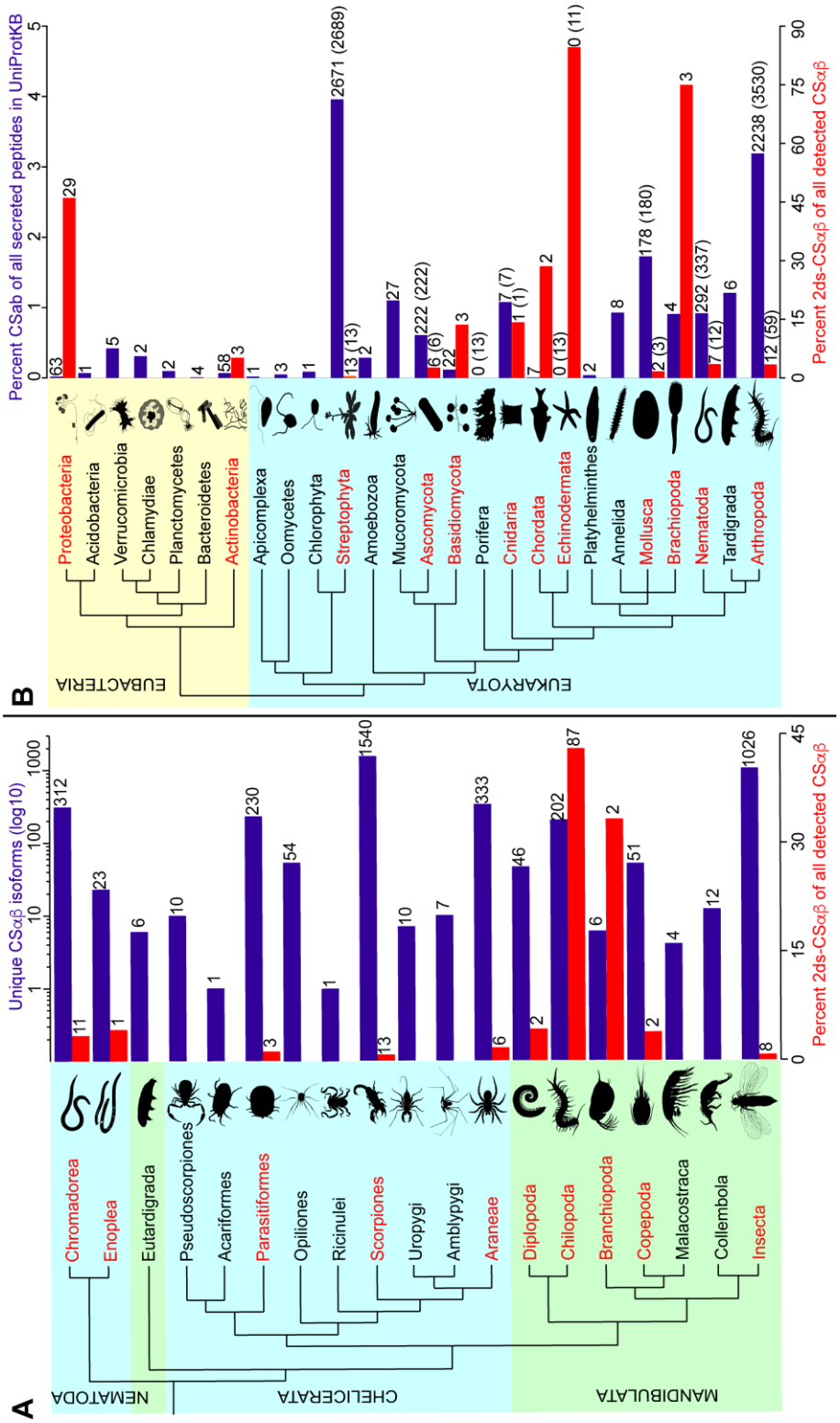

Figure 3. Distribution and diversity of the classic and minimalistic CSαβ folds.

(A) Total number of CSαβ sequences detected (blue bars; upper axis, log10 scale) and the % proportion of these that exhibited the 2ds-CSαβ cysteine pattern (orange bars; lower axis, percent), plotted against a commonly accepted hypothesis on their source organismal phylogeny (Fernández et al., 2016; Regier et al., 2010; Sharma et al., 2014). Total numbers of CSαβ and 2ds-CSαβ are shown at the tip of each bar. (B) Number of CSαβ sequences detected in UniProt relative to the total number of secreted peptides in UniProt (blue bars; upper axis, percent) and the proportion of all CSαβ sequences that exhibited the 2ds-CSαβ cysteine pattern (orange bars, lower axis, percent), plotted for each phylum against a commonly accepted hypothesis on their organismal phylogeny (Dunn et al., 2014). Numbers of CSαβ and 2ds-CSαβ detected in UniProtKB are shown at the tip of each bar, with total numbers detected in the full public and SRA databases shown in parentheses where applicable. Images are sourced from Phylopic (www.phylopic.org, see acknowledgements for image credits). See also Figure S3 and Table S2.

Although there were several taxa in which we did not detect putative 2ds-CSαβ peptides, the apparent lack of 2ds-CSαβ could be an artefact resulting from a combination of a number of

organismal factors and varying sequencing depth. For example, we found seven unique 2ds-CSαβ peptides in the whole-body transcriptome of *Scolopendra subspinipes* (Chilopoda) that had been infected with *Escherichia coli*, compared to three in the corresponding transcriptome of non-infected specimens (Yoo et al., 2014). While the non-infected specimens contained no unique amino acid sequences, four amino acid sequences were unique to the infected specimens, and three of these showed almost 10-fold higher expression levels compared to the shared 2ds-CSαβ transcripts (228.19 – 545.96 fragments per kilobase per million (FPKM) *vs* 11.99 – 68.35 FPKM). The total estimated expression level for all 2ds-CSαβ transcripts was also about 10-fold higher in the infected compared to the non-infected specimens (1294 *vs* 142 FPKM). This result suggests that centipede 2ds-CSαβ peptides may perform an immune-related function as defensin peptides, much like what has been described for many peptides with the classic CSαβ motif (Koehbach 2017; Vriens et al., 2014; Wu et al., 2014). Thus, the wide distribution of the 2ds-CSαβ fold within Arthropoda and Nematoda suggests it also predates the origin of the Ecdysozoa.

To further probe just how taxonomically widespread the 2ds-CSαβ fold may be, we searched all secreted peptides contained within the GenBank non-redundant (nr) and UniProtKB databases, as well as another 21 non-ecdysozoan transcriptome datasets from the SRA (SI Table S2). In addition, we also examined all amino acid sequences annotated as having a CSαβ fold by the InterPro database (IPR003614) (Finn et al., 2017). Using this strategy, we detected a total of 179 unique putative 2ds-CSαβ peptides distributed among 12 of the 25 examined eukaryote and bacterial phyla (Fig. 3B). This included a large relative diversity in Proteobacteria (29 of all 63 CSαβ amino acid sequences were 2ds-CSαβ) where amino acid sequences of putative 2ds-CSαβ peptides have previously been reported but did not have their 3D structures experimentally confirmed (Zhu, 2007), as well as sequences in phyla where CSαβ have previously not been reported, such as echinoderms (Fig. 3B). Although we detected 2ds-CSαβ amino acid sequences in a wide range of organisms, classic CSαβ sequences were found in all examined phyla and almost always at a greater diversity compared to the 2ds-CSαβ sequences. A notable exception are the echinoderms, where 11 of the 13 detected putative CSαβ peptides were 2ds-CSαβ.

Single or multiple origins of the 2ds-CSαβ motif?

While the lack of putative 2ds-CSαβ peptides in about half the examined phyla could be explained in part by variable expression levels or sequencing depth as discussed above, the disjunctive phyletic distribution could also be explained by the evolution of the 2ds-CSαβ motif on multiple independent occasions. Unfortunately, the extreme degree of amino acid sequence divergence and proportion of sequence gaps contained within the full dataset means sequence-based phylogenetic analyses are not well-suited for this task, as can be seen in our unresolved Bayesian phylogenetic reconstruction (SI Fig. S3C). Moreover, the 'CXXXC-X_n-CXC' motif is

insufficient on its own to establish likely homology of 2ds-CSαβ peptides. In addition to similar patterns being present in other unrelated structural motifs, such as some forms of the ICK fold (Nadezhdin et al., 2017), randomized sequence generation based on cysteine frequencies observed across the CSαβ superfamily (*P*(Cys)=0.16) (Sewell and Durbin, 1995; Shafee et al., 2016) showed that this cysteine motif arises in 39% of random 50-aa long cysteine-rich sequences.

However, although the amino acid sequences are too diverse for accurate phylogenetic analysis, evolutionary relatedness would mean they are not random. To test hypotheses on the evolutionary origin of the 2ds-CSαβ structural motif, we therefore used multi-dimensional clustering analyses to see if the 2ds-CSαβ amino acid sequences formed a distinct clade (single origin) or were nested within other CSαβ clades (multiple origins). This alternative approach clusters proteins based on the biophysical properties of their sequences to arrange them within a quantitative 'amino acid sequence space'. Such sequence space methods are effective at identifying co-varying sets of sequence properties, and clustering distantly related and diverse sequence (Jackson et al., 2018; Shafee and Anderson, 2018). Bayesian clustering of the amino acid sequence space map identified ten clear clusters (Fig S4A–B), with dimensions 1-4 segregating alpha toxins, animal antimicrobial proteins, and plant antimicrobial proteins into separate clusters. The 2ds-CSαβ sequences fall within a single cluster with 100% sensitivity and 89% selectivity (Fig S4E), which is most clearly separated in dimensions 5-7 (Fig. 4A, Fig S4B–D). This clustering indicates that there is a set of biophysical properties that unifies these peptide sequences (SI Table S3), despite their amino acid sequences typically sharing less than 15 % identity. The clustering was still observed when the analysis was repeated, omitting the cystine and glycine columns that are conserved in the other CSαβ proteins (Fig. 4B), indicating a concerted set of sequence properties in addition to their cysteine motif. The analysis also gave as good or better clustering when ignoring the varied-length regions outside of the 'CXXXC-X_n-CXC' motif (Fig. 4B). Within the 2ds-CSαβ sequences, the prokaryotic sequences largely group together, while the fungal, plant and animal sequences are mixed through the remaining groups (Fig. S4F).

As reflected in our amino acid sequence space clustering analyses, 2ds-CSαβ amino acid sequences share several properties that are distinct from the rest of the CSαβ superfamily (Fig. 4D). For instance, 2ds-CSαβ amino acid sequences have a longer average core inter-cysteine loop, and in particular the loop preceding the absent 4th canonical CSαβ cysteine (SI Fig. S4G–I). They have a distinct surface charge distribution with higher charge at the surface-exposed N-terminal of the α-helix, and reduced charge at its C-terminal cap region. In the 2ds-CSαβs, the 'CXXXC' motif tends towards basic residues followed by either a large hydrophobic or large charged residue, whereas in other sequences these are generally small or polar neutral

Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

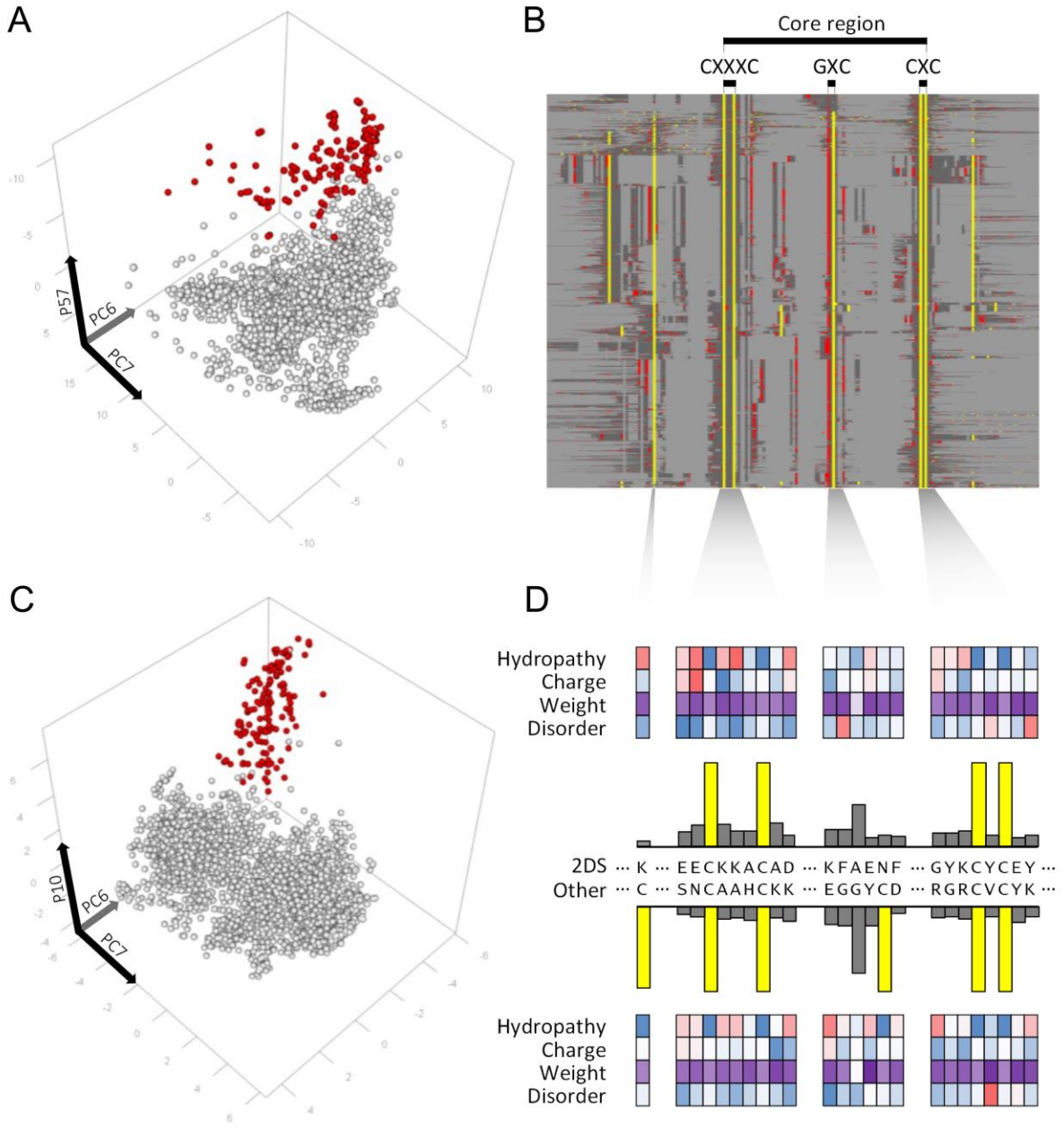

Figure 4. Sequence space analysis of CSαβ peptides. (A) Sequence space of full alignment with 2ds-CSαβ sequences highlighted in red. (B) Alignment overview of all sequences (supplementary data 3). Cysteine in yellow, glycine in red, other residues in grey, gaps in light grey. (C) Sequence space of just the core sequence region (CXXXC-X_n-CXC) and disregarding cysteines. (D) Comparison of sequence conservation for MSA columns with >60 % occupancy (2ds-CSαβ above, others below), along with average values for representative biophysical properties. Positive values in blue, negative values in red. See also Figures S4–S5 and Tables S3–S4.

residues. Similarly, the residue within the 'CXC' motif is far more likely to be hydrophilic in 2ds-CSαβs than in the other sequences. Lastly, the 2ds-CSαβ structure also relaxes a constraint that conserves a glycine in a 'GXC' motif in other CSαβ folds. In most structures with additional disulfides, the α-helix is more tightly packed against the second β-strand such that only a

glycine fits. However, in the 2ds-CSαβ structure, this constraint is slightly relaxed and the homologous residue in the 2ds-CSαβs amino acid sequences can be A/G/V/S/T (SI Fig. S4J–L). Thus, the 2ds-CSαβ amino acid sequences share a set of unifying biophysical properties in addition to their cysteine pattern, leading to their separation in the sequence space (SI Fig. S4C-E). The prokaryotic 2ds sequences do group together (SI Fig. S4F) but still fall within the 2ds cluster. Taken together, our findings suggest the 2ds-CSαβ motif is likely monophyletic with respect to the classic CSαβ fold, placing its origin in a common cellular ancestor of the eukaryotes and prokaryotes.

Activity of Sm2

The widespread function of CSαβ peptides as defensins raises the question as to whether this may also be one of the primary roles of 2ds-CSαβ peptides. Although their general and ancestral function remains highly speculative, the increased expression of 2ds-CSαβ peptides in infected compared to healthy *S. subspinipes* suggests that they may play a role in defence against pathogens in arthropods. However, at 32 µg/mL Sm2 showed no significant antimicrobial activity against five bacteria and two fungi, including four Gram-negative bacteria (*Escherichia coli, Pseudomonas aeruginosa, Acinetobacter baumannii* and *Klebsiella pneumoniae*), one Gram-positive bacterium (*Staphylococcus aureus*), and the fungi *Candida albicans* and *Cryptococcus neoformans* (for details see SI Table S4).

The diversity and neurotoxic activities of 2ds-CSαβ peptides in centipede venoms (Liu et al., 2012; Undheim et al., 2016a) also suggest that this structural scaffold is amenable to evolution of new functions namely the modulation of eukaryotic ion channels and receptors as is the case for other CSαβ variants. However, screening the synthetic Sm2 against sixteen K_V subtypes (K_V1.1–K_V1.6, K_V2.1, K_V3.1, K_V4.2, K_V7.1-7.2, K_V7.4-7.5, K_V10.1, K_V11.1 and *Shaker* IR), nine Na_V subtypes (Na_V1.1–Na_V1.8, BgNa_V1 and VdNa_V1), two Ca_V subtypes (Ca_V1, Ca_V2.2), and two nicotinic acetylcholine receptor subtypes (α7 nAChR, α3β2/α3β4 nAChR) did not return any detectable activity at a concentration of 10 µM (SI Fig. S5A,B). Expanding this screen to trigeminal ganglia (up to 500 µM Sm2) (SI Fig. S5C–E) and insecticidal assays (1 µmol/g in *Acheta domesticus,* data not shown) also returned no hits. Finally, intraplantar injection of Sm2a (10 µM) in mice caused no spontaneous pain behaviors, as evidenced by the absence of licking, flinching or licking of the injected hind paw (data not shown), and had no effect on mechanical thresholds (PWF: control, 2.6 ± 0.6 g; Sm2a, 2.7 ± 0.3 g) or heat thresholds (PWT: control, 50.2 ± 1.6 °C; Sm1a, 51.1 ± 0.9 °C) (SI Fig. S5F). The activity of Sm2 thus remains unknown.

Discussion

Centipedes have emerged as an attractive source of novel bioactive peptides due to the unusually high structural diversity of their venom peptide toxins, and interesting pharmacologies of the toxins that have been functionally characterized to date (Undheim et al., 2016a). Species in the family Scolopendridae are particularly interesting from a discovery perspective, as their venoms contain by far the greatest diversity of peptides (Undheim et al., 2015a). SLPTX15 is the second largest toxin family in the venoms of this genus, and accounts for 38 out of the 241 known venom peptides reported from *Scolopendra*, second only to SLPTX11 with 43 representatives. Our results reveal that SLPTX15 is one of two independently recruited toxin families that belong to the CSαβ superfamily (Undheim et al., 2014). Unlike SLPTX15, the other CSαβ toxin family (SLPTX02) contains the classic 3ds-CSαβ cysteine pattern that is shared by the vast majority of CSαβ peptides and characterizes the majority of scorpion toxins.

Unlike scorpion toxins or SLPTX15, however, the SLPTX02 family does not seem to have undergone any substantial functional or structural radiation, and is instead currently only known from 6 transcripts encoding 5 nearly identical peptides found in the venom of *Ethmostigmus rubripes* (also family Scolopendridae)(Undheim et al., 2014). Thus it is the 2ds-CSαβs that exhibit the greatest structural and functional radiation in centipede venoms: in *Scolopendra* alone, SLPTX15 has diverged into four distinct clades containing peptides with reported activity against voltage gated calcium (Ca_V), potassium (K_V) and sodium (Na_V) channels (Undheim et al., 2014). The diversification of the CSαβ with the least number of disulfides contrasts the perceived importance of disulfide bonds in providing cysteine-rich peptides with the evolutionary plasticity required to successfully partake in a positive selection-driven evolutionary arms race (Undheim et al., 2016b), and highlights the often overlooked importance of other structural features such as the electrostatic interactions that appear to be important for stabilizing the 2ds-CSαβ fold.

An immediate question that arises from the observation that a 2ds-CSαβ fold accounts for one of the most diverse toxin families in giant centipede venoms is whether this reduced number of disulfides is the result of a secondary loss following weaponization of a classic CSαβ peptide gene. Although rare, a loss of otherwise conserved disulfides in peptide toxins probably resulted in the evolution of the disulfide directed hairpin fold found in scorpion venoms from an ICK ancestor (Undheim et al., 2016b). However, this scenario is not the case for SLPTX15, which clearly evolved from a non-toxin 2ds-CSαβ ancestor (Fig. 3, SI Fig. S3). Rather, the 2ds-CSαβ fold found in centipede venoms appears to be just one representative of a widespread and ancient peptide family whose members share a distinct set of biophysical properties in addition to the completely conserved core cysteine pattern 'CXXXC-X_n-CXC'.

Interestingly, the cysteine pattern of the 2ds-CSαβ fold is also the most conserved trait of the CSαβ superfamily (Shafee et al., 2016). This superfamily includes host defence peptides, ion-channel binding toxins, self/non-self-recognition molecules, and enzyme inhibitors, and consists of a large number of known disulfide variations that together define an evolutionary group called the *cis*-defensin superfamily (Shafee et al., 2016). While all confirmed CSαβ structural variants characterized prior to SPLTX15 contain at least three disulfides, the direction of the third non-αβ-stabilizing disulfide varies, with a small proportion of amino acid sequences forming this disulfide with a cysteine near the C- rather than N-terminus, meaning that only five of the cysteines that form these three disulfides are universally conserved among 3ds-CSαβ (Fig. 5b). Thus, the most conserved feature that unifies the CSαβ superfamily is the pair of disulfides that link the terminal β-strand and α-helix and are formed between the cysteines in the 'CXXXC-X_n-CXC' motif. To our knowledge, different disulfide connectivities in peptides carrying this motif have only been reported for Maurotoxin, κ-BUTX-Tt2b and Ts16 (Blanc et al., 1997; Kharrat et al., 1997; Saucedo et al., 2012).

Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

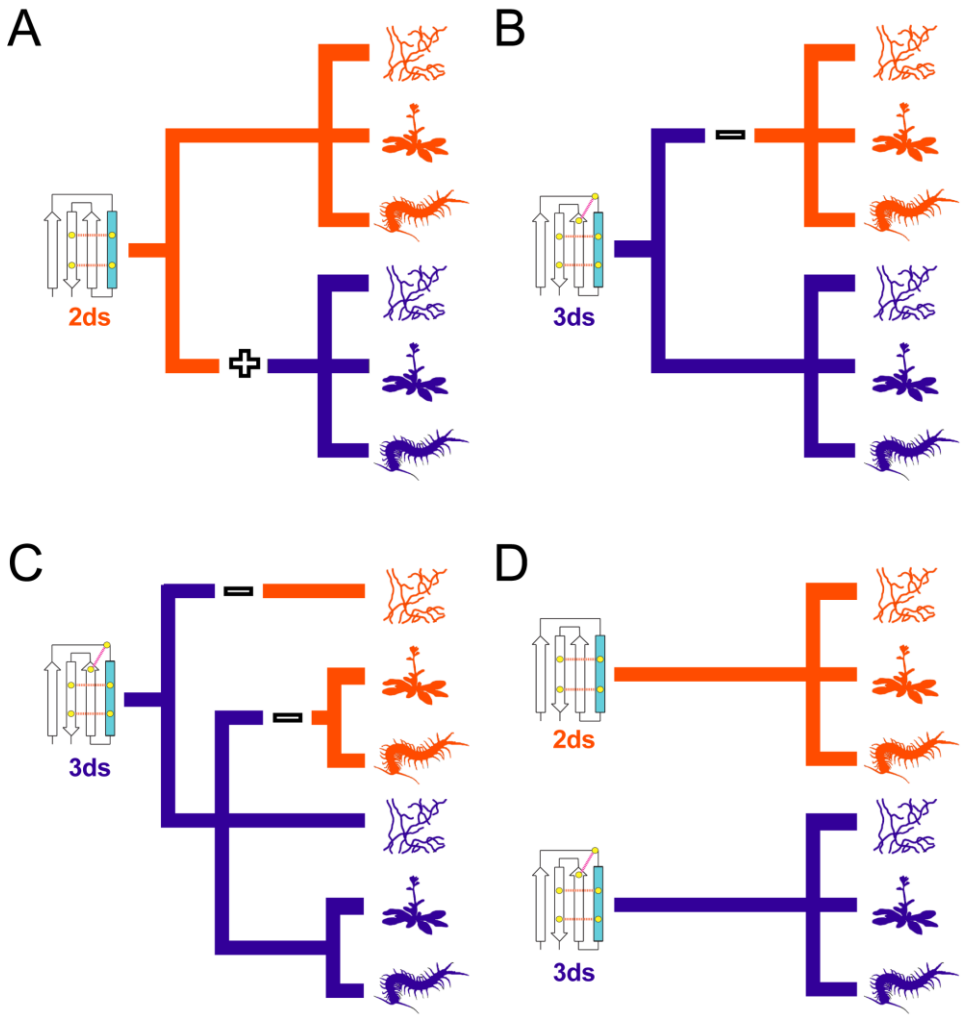

Figure 5. Possible scenarios for the evolution of the CSαβ superfamily. Four evolutionary scenarios can explain the phyletic distribution of the two main forms of CSαβ peptides. The two scaffolds evolved from a common (A) 2ds-CSαβ or (B) CSαβ ancestor, (C) the 2ds-CSαβ scaffold evolved convergently in prokaryotes and eukaryotes from a CSαβ ancestor, or (D) the 2ds-CSαβ and CSαβ arose independently in a common ancestor of prokaryotes and eukaryotes. Although we are unable to confidently distinguish between scenarios (A), (B), or (D), our data suggest scenario (C) is unlikely, and (A) more parsimonious than (B) given other evolutionary trends with the CSαβ superfamily.

Our structural studies on native Sm2 also support previous findings demonstrating that artificial deletion of extra disulfide bonds outside of this motif in a CSαβ peptides generally does not compromise the integrity of its 3D fold (Carrega et al., 2005; Sabatier et al., 1996). Thus, the minimal cysteine motif is by itself sufficient to generate a stable CSαβ structure (Sabatier et al., 1996). Although the 2ds-CSαβ architecture has been proposed previously (Mouhat et al., 2004; Tamaoki et al., 1998), our study demonstrates that this minimized fold is widespread and thereby redefines the basic canonical cysteine signature for one of the most widespread disulfide-rich peptide structural folds found in nature (Bontems et al., 1991).

The shared canonical cysteine signature between the classic and 2ds-CSαβ folds also begs the question as to their evolutionary relationship, i.e. whether the 2ds-CSαβ fold represents an ancestral, derived, or convergent fold with respect to the classic CSαβ fold (Fig. 5). Convergent evolution from *de novo* gene birth is always a possible origin for a such a small fold. For example, the CSαβ superfamily shares many convergent features with the trans-defensin superfamily (Shafee et al., 2017). Similarly, the 'CXXXC-X_n-CXC' motif has arisen in different orientations in several unrelated folds (Tamaoki et al., 1998). However, in each of these cases, analogy was clearly distinguishable by their overall 3D structure (Shafee et al., 2016). The distinct clustering of the 2ds-CSαβ and CSαβ amino acid sequences properties also makes convergent evolution unlikely, although we cannot rule out extremely strong pressure for convergence to these additional properties due to common constraints. These findings, combined with the level of structural similarity between Sm2 and the other CSαβ folds of the cis-defensin superfamily, instead suggest that the 2ds-CSαβ and CSαβ are two ancient and potentially related peptide folds, whose evolutionary history would predate the origin of eukaryotes.

Conclusions

Despite their importance, the evolutionary histories of most DRP folds remain elusive due to the structural resilience to mutations that is offered by the internal disulfide bonds. A combination of 3D structural data, structure guided alignments, phylogenetics, and multi-dimensional clustering methods can provide a solution to this issue. Using this approach, we show that one of the predominant peptide toxins of giant centipede venoms represents a widespread but previously unrecognised natural form of one of the most ubiquitous DRP folds known, the CSαβ fold. Our data suggest this new 2ds-CSαβ fold originated prior to the evolution of eukaryotes, and possibly shares a common ancestor with the rest of the of CSαβ peptides. Our results thus highlight the usefulness of 3D structures as evolutionary tools and provides some of the first insights into the ancient evolutionary history of any DRP fold.

Acknowledgements

This work was supported by the Australian Research Council (DECRA Fellowship grant number DE160101142 and Discovery Project grant number DP160104025 to E.A.B.U.). J.T. was supported by grant CELSA/17/047 – BOF/ISP. We thank Prof. David Julius, University of California, San Francisco, USA, for supporting the contributions of C.Z. Antimicrobial screening was performed by CO-ADD (The Community for Antimicrobial Drug Discovery), funded by the Wellcome Trust (UK) and The University of Queensland (Australia). Phylopic image credits: Frank Förster (Enoplea), Gareth Monger (Pseudoscorpiones, Opiliones), Birgit Lang

(Collembola), Matt Crook (Actinobacteria, Acidobacteria, Bacteroidetes, Chlamydiae, Planctomycetes, Proteobacteria, Verrucomicrobia).

Author contributions

Conceptualization, T.D. and E.A.B.U.; Discovery and purification, C.Z. and E.A.B.U; Sequence determination, E.A.B.U.; Synthesis, T.S.D. and T.D.; Nocifencive assays, J.R.D. and I.V.; Pharmacology, S.P., C.Z., I.V. and J.T.; NMR structure determination, T.S.D., P.J.H., D.J.C. and T.D.; Molecular evolution, T.S., M.A., J.T. and E.A.B.U.; Writing – Original Draft, T.S.D., T.D., and E.A.B.U; Writing – Review & Editing, T.S.D., T.S., J.T., M.A., I.V., P.J.H., T.D., and E.A.B.U.; Funding Acquisition - E.A.B.U., D.J.C. and J.T.

Declaration of interests

The authors declare no competing interests.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Ariki, N.K., Munoz, L.E., Armitage, E.L., Goodstein, F.R., George, K.G., Smith, V.L., Vetter, I., Herzig, V., King, G.F., and Loening, N.M. (2016). Characterization of three venom peptides from the spitting spider *Scytodes thoracica*. PloS one *11*, e0156291.

Blanc, E., Sabatier, J.M., Kharrat, R., Meunier, S., el Ayeb, M., Van Rietschoten, J., and Darbon, H. (1997). Solution structure of maurotoxin, a scorpion toxin from *Scorpio maurus*, with high affinity for voltage-gated potassium channels. Proteins *29*, 321–333.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bontems, F., Roumestand, C., Gilquin, B., Menez, A., and Toma, F. (1991). Refined structure of charybdotoxin: Common motifs in scorpion toxins and insect defensins. Science *254*, 1521–1523.

Brunger, A.T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics *10*, 421.

Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., and Dunker, A.K. (2008). TOP-IDP-scale a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept. Lett. *15*, 956–963.

Carrega, L., Mosbah, A., Ferrat, G., Beeton, C., Andreotti, N., Mansuelle, P., Darbon, H., De Waard, M., and Sabatier, J.M. (2005). The impact of the fourth disulfide bridge in scorpion toxins of the alpha-KTx6 subfamily. Proteins *61*, 1010–1023.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: All-atom structure validation for macromolecular crystallography. Acta Crystallogr. D *66*, 12–21.

Cheneval, O., Schroeder, C.I., Durek, T., Walsh, P., Huang, Y.H., Liras, S., Price, D.A., and Craik, D.J. (2014). Fmoc-based synthesis of disulfide-rich cyclic peptides. J. Org. Chem. *79*, 5538–5544.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B.*, et al.* (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Cock, P.J.A., Grüning, B.A., Paszkiewicz, K., and Pritchard, L. (2013). Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. PeerJ *1*, e167.

Deuis, J.R., and Vetter, I. (2016). The thermal probe test: A novel behavioral assay to quantify thermal paw withdrawal thresholds in mice. Temperature *3*, 199–207.

Dunn, C.W., Giribet, G., Edgecombe, G.D., and Hejnol, A. (2014). Animal Phylogeny and Its Evolutionary Implications. Annu. Rev. Ecol. Evol. Syst. *45*, 371–395.

Fernández, R., Edgecombe, G.D., and Giribet, G. (2016). Exploring phylogenetic relationships within myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. Syst. Biol. *65*, 871–889.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M.*, et al.* (2017). InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res. *45*, D190–d199.

Fraley C, R.A. (2012). MCLUST Version 4 for R : Normal mixture modeling for model-based clustering, classification, and density estimation (Department of Statistics University of Washington), pp. 1–57.

Fry, B.G., Roelants, K., Champagne, D.E., Scheib, H., Tyndall, J.D., King, G.F., Nevalainen, T.J., Norman, J.A., Lewis, R.J., Norton, R.S.*, et al.* (2009). The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. Annu. Rev. Genomics Hum. Genet. *10*, 483–511.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.*, et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Guntert, P. (2004). Automated NMR structure calculation with CYANA. Methods Mol. Biol. *278*, 353–378.

Hale, J.E., Butler, J.P., Gelfanova, V., You, J.S., and Knierman, M.D. (2004). A simplified procedure for the reduction and alkylation of cysteine residues in proteins prior to proteolytic digestion and mass spectral analysis. Anal. Biochem *333*, 174–181.

Han, M.V., and Zmasek, C.M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics *10*, 356.

Holm, L., and Laakso, L.M. (2016). Dali server update. Nucleic Acids Res *44*, W351-355.

Hutchinson, E.G., and Thornton, J.M. (1996). PROMOTIF — A program to identify and analyze structural motifs in proteins. Protein Sci *5*, 212–220.

Jackson, M.A., Gilding, E.K., Shafee, T., Harris, K.S., Kaas, Q., Poon, S., Yap, K., Jia, H., Guarino, R., Chan, L.Y.*, et al.* (2018). Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases. Nat. Comm. *9*, 2411.

Kalia, J., Milescu, M., Salvatierra, J., Wagner, J., Klint, J.K., King, G.F., Olivera, B.M., and Bosmans, F. (2015). From foe to friend: Using animal toxins to investigate ion channel function. J. Mol. Biol. *427*, 158–175.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kharrat, R., Mansuelle, P., Sampieri, F., Crest, M., Oughideni, R., Van Rietschoten, J., Martin-Eauclaire, M.F., Rochat, H., and El Ayeb, M. (1997). Maurotoxin, a four disulfide bridge toxin from *Scorpio maurus* venom: purification, structure and action on potassium channels. FEBS Let. *406*, 284–290.

King, G.F. (2011). Venoms as a platform for human drugs: Translating toxins into therapeutics. Expert Opin. Biol. Ther. *11*, 1469–1484.

Koehbach, J. (2017) Structure-activity relationships of insect defensins. Front. Chem. *5*, 45

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database, C. (2011). The Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21.

Li, W., and Godzik, A. (2006). CD-HIT: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Liu, Z.C., Zhang, R., Zhao, F., Chen, Z.M., Liu, H.W., Wang, Y.J., Jiang, P., Zhang, Y., Wu, Y., Ding, J.P.*, et al.* (2012). Venomic and transcriptomic analysis of centipede *Scolopendra subspinipes dehaani*. J. Proteome Res. *11*, 6197–6212.

Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188–1195.

Mouhat, S., Jouirou, B., Mosbah, A., De Waard, M., and Sabatier, J.M. (2004). Diversity of folds in animal toxins acting on ion channels. Biochem. J. *378*, 717–726.

Nadezhdin, K.D., Romanovskaia, D.D., Sachkova, M.Y., Oparin, P.B., Kovalchuk, S.I., Grishin, E.V., Arseniev, A.S., and Vassilevski, A.A. (2017). Modular toxin from the lynx spider *Oxyopes takobius*: Structure of spiderine domains in solution and membrane-mimicking environment. Protein Sci. *26*, 611–616.

NCBI Resource Coordinators (2017). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *45*, D12–d17.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M.*, et al.* (2005). RECOORD: A recalculated coordinate database of 500+proteins from the PDB using restraints from the BioMagResBank. Proteins *59*, 662–672.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. *10*, 1–6.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nat. Methods *8*, 785–786.

R Development Core Team (2011). R: A language and environment for statistical computing, pp. 2.11.11.

Rates, B., Bemquerer, M.P., Richardson, M., Borges, M.H., Morales, R.A.V., De Lima, M.E., and Pimenta, A.M.C. (2007). Venomic analyses of *Scolopendra viridicornis nigra* and *Scolopendra angulata* (Centipede, Scolopendromorpha): Shedding light on venoms from a neglected group. Toxicon *49*, 810–826.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., and Cunningham, C.W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature *463*, 1079–1083.

Rong, M.Q., Yang, S.L., Wen, B., Mo, G.X., Kang, D., Liu, J., Lin, Z.L., Jiang, W.B., Li, B.W., Du, C.Q.*, et al.* (2015). Peptidomics combined with cDNA library unravel the diversity of centipede venom. J. Proteomics *114*, 28–37.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539–542.

Sabatier, J.M., Lecomte, C., Mabrouk, K., Darbon, H., Oughideni, R., Canarelli, S., Rochat, H., Martin-Eauclaire, M.F., and van Rietschoten, J. (1996). Synthesis and characterization of leiurotoxin I analogs lacking one disulfide bridge: evidence that disulfide pairing 3-21 is not required for full toxin activity. Biochemistry *35*, 10641–10647.

Saucedo, A.L., Flores-Solis, D., Rodríguez de la Vega, R.C., Ramírez-Cordero, B., Hernández-López, R., Cano-Sánchez, P., Noriega R.N., García-Valdés, J., Coronas-Valderrama, F., de Roodt, A., Brieba, L.G., Possani, L.D., del Río-Portilla, F. (2012) New tricks of an old pattern: structural versatility of scorpion toxins with common cysteine spacing. J. Biol. Chem. *6*, 12321–12330.

Sewell, R.F., and Durbin, R. (1995). Method for calculation of probability of matching a bounded regular expression in a random data string. J. Comput. Biol. *2*, 25–31.

Shafee, T., and Anderson, M.A. (2018). A quantitative map of protein sequence space for the cis-defensin superfamily. Bioinformatics *In press (accepted 07/08/2018)*. doi:10.1093/bioinformatics/bty697/5068591

Shafee, T., and Cooke, I. (2016). AlignStat: A web-tool and R package for statistical comparison of alternative multiple sequence alignments. BMC Bioinformatics *17*, 434.

Shafee, T.M., Lay, F.T., Hulett, M.D., and Anderson, M.A. (2016). The defensins consist of two independent, convergent protein superfamilies. Mol. Biol. Evol. *33*, 2345–2356.

Shafee, T.M., Lay, F.T., Phan, T.K., Anderson, M.A., and Hulett, M.D. (2017). Convergent evolution of defensin sequence, structure and function. Cell. Mol. Life Sci. *74*, 663–682.

Sharma, P.P., Kaluziak, S.T., Pérez-Porro, A.R., González, V.L., Hormiga, G., Wheeler, W.C., and Giribet, G. (2014). Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. Mol. Biol. Evol. *31*, 2963–2984.

Shen, Y., and Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J. Biomol. NMR *56*, 227–241.

Tamaoki, H., Miura, R., Kusunoki, M., Kyogoku, Y., Kobayashi, Y., and Moroder, L. (1998). Folding motifs induced and stabilized by distinct cystine frameworks. Protein Eng. *11*, 649–659.

The UniProt Consortium (2017). UniProt: The universal protein knowledgebase. Nucleic Acids Res. *45*, D158–d169.

Undheim, E.A.B., Fry, B., and King, G.F. (2015a). Centipede venom: Recent discoveries and current state of knowledge. Toxins *7*, 679–704.

Undheim, E.A.B., Hamilton, B.R., Kurniawan, N.D., Bowlay, G., Cribb, B.W., Merritt, D.J., Fry, B.G., King, G.F., and Venter, D.J. (2015b). Production and packaging of a biological arsenal:

Evolution of centipede venoms under morphological constraint. Proc. Natl. Acad. Sci. USA *112*, 4026–4031.

Undheim, E.A.B., Jenner, R.A., and King, G.F. (2016a). Centipede venoms as a source of drug leads. Expert Opin. Drug Disc. *11*, 1139–1149.

Undheim, E.A.B., Jones, A., Clauser, K.R., Holland, J.W., Pineda, S.S., King, G.F., and Fry, B.G. (2014). Clawing through evolution: Toxin diversification and convergence in the ancient lineage Chilopoda (centipedes). Mol. Biol. Evol. *31*, 2124–2148.

Undheim, E.A.B., and King, G.F. (2011). On the venom system of centipedes (Chilopoda), a neglected group of venomous animals. Toxicon *57*, 512–524.

Undheim, E.A.B., Mobli, M., and King, G.F. (2016b). Toxin structures as evolutionary tools: Using conserved 3D folds to study the evolution of rapidly evolving peptides. Bioessays *38*, 539–548.

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, P., Ulrich, E.L., Markley, J.L., Ionides, J., and Laue, E.D. (2005). The CCPN data model for NMR spectroscopy: Development of a software pipeline. Proteins *59*, 687–696.

Vriens, K., Cammue, B.P.A., Thevissen, K. (2014) Antifungal plant defensins: Mechanisms of action and production. Molecules *19*, 12280–12303.

Wu, Y. Gao, B., Zhu, S. (2014) Fungal defensins, an emerging source of anti-infective drugs. Chin. Sci. Bull. *59*, 931.

Wuthrich, K. (1986). NMR of proteins and nucleic acids (New York: New York : John Wiley & Sons).

Yang, S., Liu, Z.H., Xiao, Y., Li, Y., Rong, M.Q., Liang, S.P., Zhang, Z.Y., Yu, H.N., King, G.F., and Lai, R. (2012). Chemical punch packed in venoms makes centipedes excellent predators. Mol. Cell. Proteomics *11*, 640–650.

Yoo, W.G., Lee, J.H., Shin, Y., Shim, J.-Y., Jung, M., Kang, B.-C., Oh, J., Seong, J., Lee, H.K., Kong, H.S.*, et al.* (2014). Antimicrobial peptides in the centipede *Scolopendra subspinipes mutilans*. Funct. Integr. Genomics *14*, 275–283.

Zhu, S. (2007). Evidence for myxobacterial origin of eukaryotic defensins. Immunogenetics *59*, 949–954.

Dash, T. S., Shafee, T., Harvey, P. J., Zhang, C., Peigneur, S., Deuis, J. R., ... & Durek, T. (2019). A centipede toxin family defines an ancient class of CSαβ defensins. *Structure*, *27*(2), 315-326. doi:10.1016/j.str.2018.10.022

Main table and legend

Table 1. Statistical analysis of Sm2 NMR structures [a]

Experimental restraints	
total no. distance restraints	577
intraresidue	153
sequential	179
medium range, *i-j*<5	63
long range, *i-j*≥5	182
hydrogen bond restraints	42
dihedral angle restraints: phi	43
dihedral angle restraints: psi	38
dihedral angle restraints: chi1	22
Deviations from idealized geometry	
bond lengths (Å)	0.010 ± 0.000
bond angles (deg)	1.049 ± 0.037
impropers (deg)	1.28 ± 0.09
NOE (Å)	0.012 ± 0.001
cDih (deg)	0.068 ± 0.046
Mean energies (kcal/mol)	
overall	-2295 ± 51
bonds	20.8 ± 1.0
angles	61.8 ± 5.3
improper	23.1 ± 2.8
van Der Waals	-258.9 ± 6.0
NOE	0.08 ± 0.01
cDih	0.08 ±0.10
electrostatic	-2388 ± 55
Violations	
NOE violations exceeding 0.2 Å	0
Dihedral violations exceeding 2.0 Å	0
RMS deviation from mean structure, Å	
backbone atoms	0.69 ± 0.16
all heavy atoms	1.42 ± 0.21
Stereochemical quality[b]	
Residues in most favoured Ramachandran region, %	97.5 ± 1.8
Ramachandran outliers, %	0.3 ± 0.8
Unfavourable sidechain rotamers, %	0.0 ± 0.0
Clashscore, all atoms	9.1 ± 2.5
Overall MolProbity score	1.6 ± 0.1

[a]All statistics are given as mean ± SD.

[b]According to MolProbity

STAR Methods

Contact for reagent and resource sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eivind A. B. Undheim (e.undheim@uq.edu.au).

Experimental model and subject details

Source of venom

Scolopendra morsitans were collected on the property of Dr Marco Inserra on Bartels Road, Kogan, 4406 Queensland, Australia (27°08'54.9"S 150°36'20.1"E). Venom from both sexes was pooled.

In vitro pharmacology

The 12 K_Vs (K_V1.1–K_V1.6, K_V2.1, K_V3.1, K_V4.2, K_V10.1, K_V11.1 and *Shaker* IR) and 8 Na_Vs (Na_V1.1–Na_V1.6, Na_V1.8, BgNa$_V$1 and VdNa$_V$1) included in our *in vitro* activity screen were exogenously expressed in oocytes harvested from adult female *Xenopus laevis*.

Activity against α7 nAChR, α3β2/α3β4 nAChR (henceforth referred to as α3-containing; α3*), Ca_V2.2, Ca_V1.3, and Na_V1.7 responses, Ca^{2+} imaging assays were performed on SH-SY5Y human neuroblastoma cells. SH-SY5Y cells were cultured in RPMI medium (ThermoFisher Scientific, Scoresby, Australia) supplemented with 15% foetal bovine serum and l-glutamine and passaged every 3–5 days using 0.25% trypsin/EDTA (ThermoFisher Scientific). For fluorescent Ca^{2+} assays, cells were plated on 384-well black-walled imaging plates (Corning) at a density of 30,000 cells/well and cultured for 48 h at 37 °C in a 5 % CO_2 incubator.

Activity against trigeminal ganglia (TG) by calcium imaging and electrophysiology, using TG dissected from newborn (P0–P3) C57BL/6 mice and cultured for >12 h before calcium imaging or electrophysiological recording.

In vivo pharmacology

In vivo pharmacological screening was done using adult male C57BL/6J mice aged 8 weeks.

Antimicrobial screening

Antimicrobial screening was done by The Community for Antimicrobial Drug Discovery (CO-ADD; www.co-add.org) using the following organisms (strains): Escherichia coli (ATCC 25922), Klebsiella pneumoniae (ATCC 700603), Acinetobacter baumannii (ATCC 19606), Pseudomonas aeruginosa (ATCC 27853), Staphylococcus aureus (ATCC 43300), Candida albicans (ATCC 90028), Cryptococcus neoforms (ATCC 208821). Bacterial strains were cultured in Luria broth (LB) (In Vitro Technologies, USB75852), at 37 °C overnight. A sample of culture was then diluted 40-fold in fresh Muller Hinton broth (MHB) (Bacto laboratories, 211443) and incubated at 37 °C for 1.5-2 h prior to assays. Fungal strains were strains were cultured for 3 days on Yeast Extract-Peptone Dextrose (YPD) (Sigma-Aldrich, Y1500) agar at 30 °C.

Method details

Isolation of Sm2

Venom was obtained from *S. morsitans* by electrostimulation after starvation for three weeks. Specimens were anesthetized with CO_2 before restrained to a clean cardboard cylinder with a double set of rubber bands, and venom extracted by electrostimulation (12 V, 1 mA) after the centipede recovered. Venom was immediately lyophilised and stored until further use at −80 °C. 500 µg venom was diluted in solvent A (0.01% trifluoroacetic acid (TFA) (v/v)) and fractionated by reverse-phase HPLC (rpHPLC) using a Gemini (Phenomenex, CA, USA) stable-bond C18 column (4.6 x 250 mm, 3 µm particle size, 110 Å pore size) with a gradient of 5–50 % solvent B (acetonitrile (ACN):H_2O:TFA; 90:10:0.043 (v/v/v)) in solvent A at a flow rate of 1mL/min over 60 min. The fraction containing Sm2 was isolated by rpHPLC using a Phenomenex Onyx monolithic C18 column (3.0 x 100 mm, 130 Å pore size) with a gradient of 10–30 % solvent B in solvent A at a flow rate of 3 ml/min over 20 min.

The peptides contained within the first rpHPLC fraction containing Sm2 were identified by LC-MS/MS analysis of reduced and alkylated (Hale et al., 2004) and trypsin digested material on an AB SCIEX 5600 triple-quadrupole TOF mass spectrometer (AB SCIEX, USA), and searching MS/MS spectra against transcriptomic sequence data using Protein Pilot v5.0 (AB SCIEX, Framingham, MA, USA). Reduction and alkylation of peptides was carried out in gas phase by drying down 1.5 ml Eppendorf tubes containing a 10% aliquot of each fraction in a vacuum centrifuge, placing 10 µL of reduction alkylation reagent in the cap (50% (vol/vol) 0.1M ammonium carbonate, 48.75% ACN (vol/vol), 1% iodoethanol (vol/vol), and 0.25% triethylphosphine (vol/vol)(final pH 10)), and incubating upside down at 37 °C in the dark for 60 minutes. The reduced and alkylated samples were digested by incubating with 30 µg/µL trypsin overnight at 37 °C in 10% ACN (vol/vol), 50 mM ammonium bicarbonate, pH 8, at a final substrate to enzyme ratio of approximately 100:1. The digested sample was made to a final concentration of 1 % formic acid (FA)(vol/vol) and analysed on an AB Sciex 5600 TripleTOF equipped with a Turbo-V source heated to 550 °C. Tryptic peptides were fractionated on a Shimadzu (Kyoto, Japan) Nexera UHPLC with an Agilent Zorbax stable-bond C18 column (Agilent, Santa Clara, CA, USA) (2.1 x 100 mm, 1.8 µm particle size, 300 Å pore size), using a flow rate of 180 µL/min and a gradient of 1–40% solvent B (90% ACN (vol/vol), 0.1% FA (vol/vol)) in 0.1% FA over 30 min. MS1 spectra were acquired at 300–1800 m/z with an accumulation time of 250 ms and selecting the 20 most intense ions for MS2. MS2 scans were acquired at 80–1400 m/z with an accumulation time of 100 ms and optimized for high resolution. Precursor ions with a charge of +2 to +5 and an intensity of at least 120 counts/s were selected, with a unit mass precursor ion inclusion window of 0.7 Da and excluding isotopes within 2 Da for MS/MS.

To identify peptides, Protein Pilot v5.0 (AB SCIEX, Framingham, MA, USA) was used to search the resulting MS/MS spectra against the already available transcriptome data (NCBI TSA accession numbers GASH01), while not allowing for amino acid substitutions. False positives were identified using decoy-based false discovery rates (FDR) as estimated by Protein Pilot and only protein identifications with a corresponding local FDR of <0.5% were considered significant. The molecular weight of the isolated Sm2 was confirmed by MALDI-TOF-MS on an AB Sciex 4700 MALDI TOF/TOF (USA) operated in positive reflectron mode, using α-cyano-4-hydroxycinnamic acid (CHCA) as matrix (7 mg/ml in 60 % ACN (vol/vol), 1 % FA (vol/vol)). Ions of m/z 900–8000 were acquired by accumulating 2500 laser desorptions/spectrum.

Chemical synthesis of Sm2

The 53 amino acid residues Sm2 polypeptide (EETEEPIRHAKKNPSEGECKKACADAFANG-DQSKIAKAENFKDYYCNCHIIIH) was synthesized by automated Fmoc SPPS using optimized protocols (Cheneval et al., 2014). The peptide was assembled on 2-chlorotrityl chloride resin using the following side chain protecting groups: Asp(tBu), Glu(tBu), His(Trt), Lys(Boc), Asn(Trt), Gln(Trt), Arg(Pbf), Ser(tBu), Thr(tBu) and Tyr(tBu). The native I-III, II-IV disulfide connectivity was achieved via a directed disulfide formation approach by protecting Cys19 (I) and Cys47 (III) as Cys(Acm), while acid-labile Cys(Trt) was used for Cys23 (II) and Cys49 (IV). Resin cleavage and side-chain deprotection were carried out by suspending the dried peptide-resin in cleavage cocktail (TFA:triisopropylsilane:H_2O; 95:2.5:2.5). After stirring for 1.5 h at room temperature, the majority of TFA was evaporated under vacuum and the peptide was precipitated with ice-cold diethyl ether. The peptide was dissolved in 50 % ACN/water containing 0.05 % TFA and lyophilized. Crude peptides were dissolved in a 10 % (v/v) ACN-water mixture containing 0.05 % (v/v) TFA, before being purified by preparative rpHPLC. The column was equilibrated with 10% of solvent B (ACN:H_2O:TFA; 89.95:10:0.05) in solvent A (H_2O:TFA; 99.95:0.05). Peptides were eluted using linear gradients of solvent B in solvent A, and fractions were collected across the expected elution time. Peptide purity and identity were assessed by ESI-MS on Shimadzu 2020 LC mass spectrometer and by analytical scale uHPLC on a Shimadzu Nexera system equipped with an Agilent Zorbax C18 column (1.8 µm, 2.1 x 100 mm). Fractions containing the desired product were pooled, lyophilized and stored at –20 ºC (observed mass 6138.6 Da; calculated mass 6138.7 Da [average isotope composition]).

Reduced Acm-protected Sm2 (20.3 mg, 1 equ.) was dissolved in 40 mL of a mixture of AcOH/H_2O (9:1 (v/v)). To this mixture 10 eq. I_2 (0.5 M I_2 in methanol) was added and the solution was incubated at room temperature for 20 min to form the Cys23–Cys49 disulfide linkage. After this time, water (40 mL) was added and the mixture was incubated at 40 °C for 50 min to remove Acm protecting groups and form the Cys19 and Cys47 disulfide bond. The remaining I_2 was quenched with ascorbic acid. Sm2 was then purified using HPLC on a RP

column (Phenomenex Gemini C18, 250 x 10 mm, 5 µM particle size, 110 Å pore size) using a gradient elution profile (10–40 % solvent B over 60 min, 3 ml/min, 40 °C). Fractions containing Sm2 were combined and lyophilized to give the final product. Peptide identity, purity, correct folding was assessed by MALDI-TOF MS as described above, as well as by co-elution with native Sm2a using high-resolution LC-MS (SI Figure S2). MALDI-TOF MS yielded 5989.4 M+H for the native toxin, 5989.5 M+H for the synthetic toxin, compared to a calculated monoisotopic 5989.8 M+H. For comparison by LC-MS, native and synthetic material were analysed using an Agilent Zorbax stable bond C18 column (2.1 × 100 mm, 1.8 µm particle size, 300 Å pore size) at a flow of 180 µl/min and a gradient of 1–40% solvent B (ACN:H_2O:formic acid (FA); 90:10:0.1) in 0.1 % FA over 30 min on a Shimadzu Nexera UHPLC coupled with an AB SCIEX 5600 mass spectrometer equipped with a Turbo V ion source heated to 450°C (observed mass 5988.5; calculated mass 5988.8 Da (monoisotopic mass)).

NMR structure determination

NMR samples were prepared by dissolving 1mg of lyophilized synthetic Sm2 in a 90% H_2O/ 10 % D_2O mixture (500 µL) containing 20 mM sodium phosphate, pH 5.9. NMR measurements were performed on a Bruker Avance 600 MHz spectrometer at 298K equipped with a cryogenically cooled probe. 2-D TOCSY (80 ms mixing time), NOESY (200 ms mixing time), E-COSY, and natural abundance ^{15}N and ^{13}C HSQC were used to sequentially assign backbone and side chain protons and heteroatoms. Variable temperature experiments were performed by recording six TOCSY spectra at temperatures ranging from 283–308 K. Slowly exchanging amide protons were identified by incubating Sm2 in 100% D_2O over 24 h. Solvent suppression was achieved using excitation sculpting. Spectra were referenced to water at 4.77 ppm. All spectra were processed using Topspin v2.1 and assigned using CcpNMR Analysis v2.4.1 (Vranken et al., 2005).

The NOESY was successfully assigned using a combination of the sequential assignment protocol and HSQC experiments (Wuthrich, 1986). Initial structure calculations were performed using CYANA with distance restraints derived from NOESY spectra. Disulfide restraints were introduced along with backbone ϕ and ψ dihedral angle constraints generated using TALOS-N (Shen and Bax, 2013). X_1 restraints were derived from E-COSY coupling constants and NOE intensities. Hydrogen bond restraints were introduced as indicated by slow D_2O exchange and for those backbone amide protons whose chemical shift was not temperature sensitive. A final set of structures was refined by CNS (Brunger, 2007) using torsion angle dynamics, refinement and energy minimization in explicit solvent, and protocols as developed for the RECOORD database (Nederveen et al., 2005). The final set of 20 lowest energy structures was chosen based upon stereochemical quality as assessed using MolProbity (Chen et al., 2010). Chemical shifts have been deposited in the BMRB database (BMRB code 30372). Structural coordinates have been deposited in the RCSB PDB (PDB accession 6BL9).

Evolutionary analyses

In order to identify structural homologues of Sm2, its 3D structure was searched against all publicly available structures in the PDB using the Dali server (Holm and Laakso, 2016). The resulting hits were filtered based on structural similarity, removing structures with amino acid sequences >50% identity using CD-HIT v4.6.5 (Li and Godzik, 2006) (0.5 identity, word size 3). Structures lacking any of the disulfides characteristic of any CSαβ fold were treated as definite false positives and removed. We then conducted an iterative, lenient BLAST search with blastp within Blast + (Camacho et al., 2009) to search the amino acid sequences of all DALI hits against all amino acid sequences in the PDB using a e-value threshold of 1. These were filtered by CD-HIT and manual inspection as per above, combined with the filtered DALI search hit, again filtered to remove sequence redundancy > 50%, and analysed by an all-against-all structural comparison and pairwise distance-based dendrogram construction available through the DALI server (http://ekhidna2.biocenter.helsinki.fi/dali/).

In addition, the sequences of Sm2 and all its structural homologues were searched against all secreted sequences with 40–150 amino acids within the UniProtKB database (The UniProt Consortium, 2017) (downloaded 01 June 2017) using blastp within Blast + (Camacho et al., 2009), as well as the GenBank nr and TSA databases (NCBI Resource Coordinators, 2017) using PSI-BLAST (Altschul et al., 1997)(Supplemental Data S1). The sequence datasets were combined, duplicate sequences removed using CD-HIT (0.95 identity, word size 5), filtered based on presence of a predicted signal peptide by SignalP v4.1 (Nielsen et al., 1997; Petersen et al., 2011), inspected manually to remove any false positives, and aligned using MAFFT v7.304 (Katoh and Standley, 2013). The resulting multiple sequence alignment (MSA) was then used to generate HMMER profiles for the combined CSαβ sequence dataset, including the two-disulfide hits, and subsequently used to re-interrogate the above databases by HMMER searches with HMMER v3.1b (hmmer.org). Redundant amino acid sequences and false positives were then removed by CD-HIT and manual inspection.

In addition to interrogating public sequence databases, we constructed a taxonomically comprehensive eukaryotic sequence database from Illumina-sequenced transcriptome datasets in the NCBI SRA(Leinonen et al., 2011) (see SI Table 3). SRA data were downloaded and converted to fastq format using the fastq dump tool in the SRA toolkit v2.4.1, and trimmed with Trimmomatic v035 (Bolger et al., 2014) using a quality cut-off of 25 and adjusting the remaining trimming parameters according to the experimental sequencing parameters of each sample. For each sample, the paired trimmed reads were assembled using Trinity v2.0.6 (Grabherr et al., 2011) using default parameters, and all coding sequences (CDS) longer than 60 amino acids predicted using the Galaxy tool ‘Get open reading frames (ORFs) or coding sequences (CDSs)’ (Cock et al., 2009; Cock et al., 2013). The amino acid sequences were filtered based on

sequence similarity using CD-HIT as above, yielding a database consisting of 48,903,419 unique amino acid sequences encoded by 18,881,531 contigs from a total of 155 transcriptomes from 146 species. The database of high-significance sequence pairs (HSP's) obtained by BLAST and HMMER searches of UniProtKB and GenBank was then used to identify putative CSαβ homologues by BLAST and HMMER as described above. HSPs from both databases were combined, duplicate amino acid sequences removed by CD-HIT, and false positives removed by manual inspection.

Identified CSαβ and putative CSαβ homologues were aligned using MAFFT, and refined by comparison with structure-based sequence alignments generated in PyMol v1.6.0. MSAs were further refined by locally realigning structurally non-conserved regions using the MAFFT regional alignment tool v0.2, CysBar (Shafee and Cooke, 2016) and AlignStat (Shafee and Cooke, 2016) (Supplemental Data S1). The evolution of the CSαβ fold was then attempted reconstructed using maximum likelihood and Bayesian phylogenetic analyses of both the full dataset and only myriapod sequences. For phylogenetic analysis by maximum likelihood we used IQ-Tree v1.5.5 (Nguyen et al., 2015). Evolutionary models were estimated using ModelFinder (Kalyaanamoorthy et al., 2017), while support values were estimated by ultrafast bootstrap using 10,000 iterations (Minh et al., 2013). For Bayesian inference we used MrBayes v3.2 (Ronquist et al., 2012), setting MrBayes to estimate the most appropriate model by setting parameter models to mixed and rates to follow a gamma distribution. Phylogenetic trees were visualized in Archaeopteryx v0.9916 (Han and Zmasek, 2009).

Sequence clustering analyses

Due to the short amino acid sequences, frequent insertions and deletions, low sequence conservation of peptides combined with the evolutionary plasticity of disulfide rich peptides, conventional phylogenetic methods suffer from saturation effects (SI Fig. S3C). We therefore also analysed the CSαβ peptide dataset by quantitative position-specific biophysical sequence-space analyses (Shafee and Anderson, 2018). Full [R] codes are available at the repository https://github.com/TS404/SeqSpace. The MSAs were converted to vectors of biophysical properties describing each position in the MSA in [R] (R Development Core Team, 2011). Net charge in Coulombs, disorder propensity as in TOP-IDP (Campen et al., 2008), hydrophobicity as in the Doolittle index (Kyle and Doolittle, 1982), molecular weight of R group in Daltons, and disulfide potential and column occupancy as binary descriptors. These properties encompass the main differences between the naturally occurring amino acids. Disulfide potential is included in this case since disulfides are particularly important to defensin structures. MSA column occupancy accounts for different sequence lengths. Values were normalised within each property. Gaps were given the average value of their column for each property (other than occupancy) such that they had no effect on subsequent multidimensional scaling. Each

sequence is therefore represented by a row of the resulting matrix which contains its MSA-position-specific biophysical information.

The numerical representation of sequence space was projected by Principal Component Analysis (PCA) via prcomp in [R] (R Development Core Team, 2011). This summarised key covarying property sets and allowed the highly multidimensional data to be summarised in far fewer dimensions. Bayesian clustering was performed using Mclust to identify groups of amino acid sequences with similar biophysical properties (Fraley C, 2012). Briefly, this algorithm calculates the models the distribution of data points as a set of spheroid clusters with varied sizes, elongation and orientation. The optimal number of clusters is chosen based on goodness of fit (Bayesian Information Criterion). Adding clusters to the model improves the models fit to the data until an optimal number of clusters is reached, after which additional clusters fail to improve the model's fit. The first 40 PCs were used for clustering since they summarised the most important 30% of the contained information. Of the 10 clusters identified in this way, one contained all 164 2DS sequences (as well as 21 other sequences). PCA loadings were used to investigate covarying sequence property sets. Components 1-4 separated plant and animal sequences, as well as several families with known neurotoxic or antimicrobial functions, but are not the focus of this work. Components that were strongly differentiated between 2DS and other sequences were identified by mean sequence displacement in each axis (Jackson et al., 2018). The analysis was repeated whilst omitting specific MSA columns as described in the main text to confirm that clustering was not an artefact of the lack of the usually conserved cysteines and glycine. The transformed sequence space matrix and classification likelihoods used for the sequence space analyses are available in Supplemental Data S3.

Pharmacology

Sm2 is among the most abundant peptides in the venom of *S. morsitans*, suggesting it serves a role as a toxin. The activity profile of Sm2 was screened by electrophysiology at a concentration of 10 µM against 12 K_Vs (K_V1.1–K_V1.6, K_V2.1, K_V3.1, K_V4.2, K_V10.1, K_V11.1 and *Shaker* IR) and 8 NaVs (Na_V1.1–Na_V1.6, Na_V1.8, BgNa_V1 and VdNa_V1) exogenously expressed receptors in *Xenopus laevis* oocytes, with each experiment conducted in at least triplicates. K_V1.1–K_V1.6 and *Shaker* IR currents were evoked by 500 ms depolarizations to 0 mV followed by a 500 ms pulse to –50 mV, from a holding potential of –90 mV. K_V10.1 currents were evoked by 2 s depolarizing pulses to 0 mV from a holding potential of –90 mV. Current traces of hERG channels were elicited by applying a +40 mV prepulse for 2 s followed by a step to –120 mV for 2 s. K_V2.1, K_V3.1 and K_V4.2 currents were elicited by 500 ms pulses to +20mV from a holding potential of –90 mV. Sodium current traces were, from a holding potential of -90 mV, evoked by 100 ms depolarizations to V_{max} (the voltage corresponding to maximal sodium current in control conditions).

To assess the effect of Sm2 on α3*, $Ca_V2.2$, $Ca_V1.3$, and $Na_V1.7$ responses, previously described fluorescent Ca^{2+} imaging assays on SH-SY5Y human neuroblastoma cells were performed using the FLIPR^{TETRA} fluorescence plate reader (Molecular Devices, Sunnyvale, CA)(Ariki et al., 2016). SH-SY5Y cells were cultured in RPMI medium (ThermoFisher Scientific, Scoresby, Australia) supplemented with 15% foetal bovine serum and l-glutamine and passaged every 3–5 days using 0.25% trypsin/EDTA (ThermoFisher Scientific). For fluorescent Ca^{2+} assays, cells were plated on 384-well black-walled imaging plates (Corning) at a density of 30,000 cells/well and cultured for 48 h at 37 °C in a 5 % CO_2 incubator. In brief, SH-SY5Y cells were loaded with Calcium 4 no-wash dye (Molecular Devices) diluted in physiological salt solution (composition in mM: NaCl 140, d-glucose 11.5, KCl 5.9, $MgCl_2$ 1.4, NaH_2PO_4 1.2, $NaHCO_3$ 5, $CaCl_2$ 1.8, HEPES 10) for 30 min at 37 °C. Fluorescence responses (excitation 470–495 nm; emission 515–575 nm) to addition of Sm2 (10 µM) were measured every 1 s for 5 min, followed by addition of the following agonists: choline (30 µM) in the presence of 10 µM PNU-120596 to activate α7 nAChR; nicotine to activate α3β2/α3β4 nAChR; KCl (90 mM) and $CaCl_2$ (5 mM) to activate $Ca_V1.3$; KCl (90 mM) and $CaCl_2$ (5 mM) in the presence of nifedipine (10 µM) to activate $Ca_V2.2$; and veratridine (4 µM) in the presence of OD1 (30 nM; UniProt Accession P84646) to activate $Na_V1.7$.

Sm2 was also screened against trigeminal ganglia by calcium imaging and electrophysiology. For calcium imaging, trigeminal ganglia (TG) were dissected from newborn (P0–P3) C57BL/6 mice and cultured for >12 h before calcium imaging or electrophysiological recording. Primary cells were plated onto cover slips coated with poly-L-lysine (Sigma) and laminin (Invitrogen, 10 µg/ml). Sm2 was buffered in isotonic solution (140 mM NaCl, 5 mM KCl, 2 mM CaCl2, 2 mM $MgCl_2$, 10 mM glucose, 10mM HEPES, pH 7.4), while neurons were previously loaded with Fura-2-AM (Molecular Probes) for >1 h. Response to high extracellular potassium (150 mM KCl, 10 mM HEPES, pH 7.4) was used to identify neurons.

For electrophysiological screening against TG, whole-cell recordings were conducted in an extracellular solution contained 150 mM NaCl, 2.8 mM KCl, 1 mM $MgSO_4$, 2 mM $CaCl_2$, 10 mM HEPES, 290–300 mOsmol/kg, pH7.4. Pipette solution contained 130 mM K-gluconate, 15 mM KCl, 4 mM NaCl, 0.5 mM $CaCl_2$, 1 mM EGTA, 10 mM HEPES, 280–290 mOsmol/kg, pH 7.2. Extracellular solution was perfused with or without toxins/drugs using a SmartSquirt Micro-Perfusion system (AutoMate). All recordings were performed using fire-polished glass electrodes with a resistance of 2–5 MΩ at room temperature (20–22 °C). Signals were amplified using an Axopatch 200B amplifier, digitized with a Digidata 1440A and recorded using pCLAMP 10.2 software (Molecular Devices, Sunnyvale, CA, USA). For all TG neurons, the holding potential was –80mV, while tetraethylammonium (TEA, Sigma) and 4-aminopyridine (4-AP, Tocris) were used as controls.

Finally, to assess any *in vivo* effects on nocifencive behaviour, we tested Sm2 for effects on thermal and mechanical pain in mice. Ethical approval for *in vivo* experiments in animals was obtained from The University of Queensland animal ethics committee. Experiments involving animals were conducted in accordance with the Animal Care and Protection Regulation Qld (2012), the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes, 8th edition (2013) and the International Association for the Study of Pain Guidelines for the Use of Animals in Research. Adult male C57BL/6J mice aged 8 weeks were administered an intraplantar injection of Sm2a (10 µM) diluted in PBS/0.1 % BSA or vehicle in a volume of 40 µL under light isoflurane (3 %) anaesthesia. Mice were then immediately placed into individual mouse runs for at least 5 min prior to behavioural assessment. Mechanical thresholds were assessed using an electronic von Frey apparatus (MouseMet Electronic von Frey, TopCat Metrology) as previously described (Deuis and Vetter, 2016). Thermal thresholds were assessed using the thermal probe test (MouseMet Thermal, TopCat Metrology) as previously described (Deuis and Vetter, 2016). The experimenter was blinded to the injection type (Sm2a or vehicle control) each individual mouse received.

Antimicrobial activity

Antimicrobial assays were done by The Community for Antimicrobial Drug Discovery (CO-ADD; www.co-add.org) according to their standardised methods. For screening of antimicrobial activity, Sm2 was first prepared in DMSO (< 1 %) and H_2O to a final testing concentration of 5.3 µM in a 384-well, non-binding surface plate. Experiments were performed twice (n=2, on different plates). All bacteria were cultured in Cation-adjusted Mueller Hinton broth at 37 °C overnight, diluted 40-fold in fresh broth and incubated at 37 °C for 1.5–3 h. The resultant mid-log phase cultures were added to each well of the compound-containing plates, giving a cell density of 5 x 10^5 CFU/mL (measured by absorbance at 600 nm [OD_{600}]) in 50 µL, and incubated at 37 °C for 18 h without shaking. Fungal strains were cultured for 3 days on Yeast Extract-Peptone Dextrose agar at 30 °C. A yeast suspension of 1 x 106 to 5 x 106 CFU/mL (determined by OD_{530}) was prepared from five colonies, diluted, added to each well of the compound-containing plates giving a final cell density of 2.5 x 10^3 CFU/mL in 50 µL, and incubated at 35 °C for 24 h without shaking.

Quantification and statistical analysis

In vitro pharmacology

For fluorescent Ca^{2+} imaging assays on SH-SY5Y human neuroblastoma cells, raw fluorescence readings were converted to response over baseline using the analysis tool SCREENWORKS 3.1.1.4 (Molecular Devices) and were expressed relative to the maximum increase in fluorescence of control responses. All calcium imaging responses in TG were digitized and analysed using MetaFluor software (Molecular Device).

Antimicrobial activity

Antimicrobial assays were done by The Community for Antimicrobial Drug Discovery (CO-ADD; www.co-add.org). Inhibition of bacterial growth was determined measuring absorbance at 600 nm (OD600), using a Tecan M1000 Pro monochromator plate reader. The percentage of growth inhibition was calculated for each well, using the negative control (media only) and positive control (bacteria without inhibitors) on the same plate as references. Growth inhibition of *Candida albicans* was determined measuring absorbance at 530 nm (OD530), while the growth inhibition of *Cryptococcus neoformans* was determined measuring the difference in absorbance between 600 and 570 nm (OD600-570), after the addition of resazurin (0.001% final concentration) and incubation at 35 °C for additional 2 h. The absorbance was measured using a Biotek Synergy HTX plate reader.

The percentage of growth inhibition was calculated for each well, using the negative control (media only) and positive control (bacteria without inhibitors) on the same plate as references. Percentage growth inhibition of an individual sample is calculated based on negative controls (media only) and positive controls (bacterial/fungal media without inhibitors), with an expected variation in growth rates for all bacteria and fungi of 10%. Any significant variation (or outliers/hits) is identified by the modified Z-Score, and actives are selected by a combination of inhibition value and Z-Score. The Z-Score is calculated based on the sample population using a modified Z-Score method which accounts for possible skewed sample population. The modified method uses median and median average deviation (MAD) instead of mean and standard deviation, and a scaling factor: $M(i) = 0.6745 * (x(i) - \text{median}(x))/\text{MAD})$. $M(i)$ values of $> |2.5|$ (absolute) are considered significant.

Data and software availability

Structural coordinates of the Sm2 NMR solution structure are available through the RCSB Protein Data Bank (www.rcsb.org/pdb/; accession 6BL9) while chemical shifts have been deposited in the Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu; accession 30372). All CSαβ peptides identified from UniProtKB, all CSαβ peptides with less than 95 % sequence identity, all CSαβ peptides detected in the combined UniProtKB, NCBI nr, TSA and SRA dataset with less than 95 % sequence identity, and the full SRA assembly dataset generated during the current study are available from the corresponding author (EABU) on request.

Supplemental information titles and legends

Figure S1. Isolation, synthesis and disulfide connectivity of Sm2, relates to Figure 1. (A) Reverse phase HPLC (rpHPLC) chromatogram of the venom of *Scolopendra morsitans* fractionated using a Phenomenex Gemini stable-bond C18 column, with the peak containing Sm2 marked with an asterisk. (B) Fraction containing Sm2 refractionated by rpHPLC using a monolithic Phenomenex Onyx column, with the peak corresponding to Sm2 marked with an asterisk. **(**C) HPLC-MS of the crude 53 residue Sm2 (Cys19, 47 as Cys(Acm)) after cleavage from the solid support. The ESI-MS spectrum of the principal peak (*) is shown on the right-hand side. Observed mass 6138.6 Da; calculated mass 6138.7 Da (average isotope composition). (D) uHPLC co-elution experiments (left) and fingerprint region of ^{1}H-NMR spectra (right) of venom-derived Sm2 and synthetic analogs. The chromatogram shows fully reduced and side-chain deprotected Sm2 (blue), Sm2 isolated from *Scolopendra morsitans* venom (red) and the two products obtained after regioselective disulfide formation (black). ^{1}H-NMR spectra show the two synthetic Sm2 disulfide isomers: I-IV, II-III disulfide isomer (top, red); I-III, II-IV isomer (bottom, black). Collectively these data suggest that native Sm2 has a I-III, II-IV cysteine connectivity. (E) LC-MS of intact (top left) or reduced and alkylated tryptic peptides of native and synthetic Sm2, each showing co-elution by overlaid extracted ion chromatograms of peptides indicated in blue with the corresponding m/z and overlaid mass spectrum as insets.

Figure S2. 15N-HSQC, similarity to SsTx, thermal stability and stabilising electrostatic interactions of Sm2, relates to Figure 2. (A) Natural abundance ^{15}N-HSQC of synthetic Sm2 (2 mg/mL in 20 mM sodium phosphate, pH 5.9, 10 % D_2O). **(B)** 1H-fingerprint region (same conditions). **(C)** Primary and 3D structure alignments of Sm2 and SsTx (PDB 5X0S), with overlaid structures of Sm2 and SsTx shown in grey and cyan, respectively. Corresponding secondary structure features are colour coded and displayed above the amino acid sequence alignment, cysteines are shown in bold, and identical sequences are highlighted in grey. Sm2 and SsTx share 32% amino acid sequence identity (26% not counting cysteines), and their 3D structures overlay with a peptide backbone RMSD of 3.1 Å (lowest RMSD was achieved using “super” alignment function in PyMol). **(D)** Temperature-induced unfolding of Sm2 monitored by CD spectroscopy at 222 nm. Unfolding (red, 20 °C → 95 °C) and re-folding (blue, 95 °C → 20 °C) transition curves indicate a Tm > 60 °C. **(E)** Structural comparison of *Nicotiana alata* Defensin 1 (NaD1; PDB 4AAZ) and Sm2, with disulfides shown as orange tubes and corresponding stabilizing features circled in red. The positioning of the N and C-termini next to each other results in the formation of a salt bridge between the termini that can be inferred in 12/20 structures. The second electrostatic interaction is between Lys12 and Asp43 which are located in loop 1 and 3 respectively. This interaction could be significant as it is between two sequence-

distant secondary structure elements and therefore can effectively be thought of as a disulfide equivalent.

Figure S3. Phylogenetic reconstruction of CSαβ amino acid sequences, relates to Figure 3. Maximum likelihood phylogenetic reconstruction of CSαβ sequences identified from myriapods displayed as **(A)** midpoint rooted and **(B)** unrooted trees with bootstrap values given as node support. 2ds-CSαβ sequences are highlighted in red and all other CSαβ sequences in blue, while sequence names from non-venomous species are shown in bold blue font. The collapsed clade containing SLPTX15 toxin sequences is highlighted by the red rectangle and shown in bold red font. **(C)** Attempted phylogenetic reconstruction of all identified CSαβ-peptides, calculated under the Blosum model of amino acid substitution as selected by MrBayes v3.2 (posterior probability 1), with gamma-distributed rates across sites. The tree shown is a majority rule consensus tree displayed as midpoint rooted and with splits of posterior probability less than 0.5 collapsed. No higher-order clades were resolved, reflecting saturated phylogenetic signal.

Figure S4. Sequence properties of classic CSαβ and 2ds-CSαβ amino acid sequences, related to Figure 4. Sequence space coloured by the automatically identified groups **(A)** PCs 1-3, and **(B)** PCs 5-7 (black cluster contains all 164 2ds sequences and 21 other sequences). **(C)** Mean axis displacement of 2DS sequences used to identify main PCs of interest. **(D)** Histogram and boxplot of PC7 axis position for 2DS (red) and other (grey) sequences. **(E)** Bayesian probability for each sequence being assigned to the cluster coloured black in parts I and J, for the 164 2ds sequences assigned to the cluster, for the 21 other sequences assigned to the cluster, and for the remaining sequences not assigned to the cluster. **(F)** Sequence space and clustering of just the 2ds-CSαβ sequences in alone, with proteobacterial sequences coloured purple. **(G-I)** Inter-cystine loop lengths of 2ds (red) and all other (black) CSαβ-peptides for **(G)** the full CXXXC-X_n-CXC region, **(H)** the region before the usually-conserved central cysteine, and **(I)** the region after the usually-conserved central cysteine. The position of the central cysteine was estimated from alignment, outlined in main manuscript Fig. 4b. **(J)** Relative proportions of different residues in the first position of the 'GxC' motif in 2ds and all other CSαβ sequences. The position of the first residue in the 'GxC' motif shown as spheres in **(K)** Sm2 and **(L)** NaD1 (PDB 4AAZ).

Figure S5. Activity screening of Sm2, related to Figure 4. (A) Electrophysiology experiments with exogenously expressed receptors in Xenopus oocytes. Activity profile of Sm2 on several K_V and Na_V channel isoforms. Representative whole-cell current traces in control and toxin conditions are shown. The dotted line indicates the zero-current level. The asterisk (*) marks steady-state current traces after application of 10 µM toxin. Traces shown are representative traces of at least 3 independent experiments ($n \geq 3$). $K_V1.1$–$K_V1.6$ and Shaker IR currents were evoked by 500 ms depolarizations to 0 mV followed by a 500 ms pulse to –50 mV, from a

holding potential of –90 mV. K_V10.1 currents were evoked by 2 sec depolarizing pulses to 0 mV from a holding potential of –90 mV. Current traces of hERG channels were elicited by applying a +40 mV prepulse for 2 sec followed by a step to –120 mV for 2 sec. K_V2.1, K_V3.1 and K_V4.2 currents were elicited by 500 msec pulses to +20mV from a holding potential of –90 mV. Sodium current traces were, from a holding potential of –90 mV, evoked by 100 msec depolarizations to Vmax (the voltage corresponding to maximal sodium current in control conditions). **(B)** Sm2 (10 µM) had no effect on α7 nAChR, α3β2/α3β4 nAChR, Ca_V2.2, Ca_V1.3, and Na_V1.7 responses assessed using fluorescent Ca^{2+} imaging in SH-SY5Y human neuroblastoma cells. Data are presented as mean ± SEM, n = 3–12 per group. **(C)** Calcium imaging traces and representative images of 500 µM synthetic Sm2 fraction on cultured mouse TG neurons. Subsequent response to high (150 mM) extracellular K^+ reveals all neurons in the field. Each trace represents a responding individual neuron and the thick black trace is average of all individual traces. **(D, E)** Whole-cell electrophysiology recording and quantification of Sm2 effect on cultured TG neurons. **(F)** Assessing the effect of Sm2 on pain behaviours in mice. Intraplantar injection of Sm2 (10 µM) had no significant effect on mechanical thresholds compared to vehicle control assessed using electronic von Frey test, and no significant effect on heat thresholds compared to vehicle control assessed using the thermal probe test. Statistical significance was determined using t-test, * $P < 0.05$ compared to control. Data are presented as mean ± SEM, n = 4 per group.

Table S1. DALI search output table, related to Figure 2. Structures with amino acid sequences that share less than 50% sequence identity and have been used in Figure 2 are highlighted in bold, while clear false positives that lack any disulfides corresponding to those characteristic of the CSαβ fold are shaded grey.

Table S2. Transcriptome datasets assembled from the NCBI SRA database along with the number of CSαβ and 2ds-CSαβ identified in each assembly, related to Figure 3.

Table S3. Biophysical properties that define PC axes of 2ds-CSαβ amino acid sequences, related to Figure 4. Top half shows top 20 most highly loaded residue properties for the three PC axes that most strongly separate of 2ds-CSαβ sequences (excluding occupancy). Bottom half shows top 20 most highly loaded residue properties for the three PC axes when MSA columns outside of the core region, and the missing cysteine is ignored that most strongly separate of 2ds-CSαβ sequences (excluding occupancy). MSA column refers to the MSA in supplementary data. 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.

Table S4. Antimicrobial activity of Sm2, related to Figure 4. For screening of antimicrobial activity, Sm2 was first prepared in DMSO (< 1 %) and H_2O to a final testing concentration of 5.3 µM in a 384-well, non-binding surface plate. Experiments were performed twice (n=2, on different plates). All bacteria were cultured in Cation-adjusted Mueller Hinton broth at 37 °C overnight, diluted 40-fold in fresh broth and incubated at 37 °C for 1.5–3 h. The resultant mid-log phase cultures were added to each well of the compound-containing plates, giving a cell density of 5 x 10^5 CFU/mL (measured by absorbance at 600 nm [OD_{600}]) in 50 µL, and incubated at 37 °C for 18 h without shaking. Inhibition of bacterial growth was determined measuring OD_{600}, using a Tecan M1000 Pro monochromator plate reader. Fungal strains were cultured for 3 days on Yeast Extract-Peptone Dextrose agar at 30 °C. A yeast suspension of 1 x 10^6 to 5 x 10^6 CFU/mL (determined by OD_{530}) was prepared from five colonies, diluted, added to each well of the compound-containing plates giving a final cell density of 2.5 x 10^3 CFU/mL in 50 µL, and incubated at 35 °C for 24 h without shaking. Growth inhibition of *C. albicans* and *C. neoformans* was determined by measuring OD_{530} or the difference in $OD_{600–570}$ after addition of resazurin (0.001% final concentration) and incubation at 35 °C for additional 2 h, respectively. Absorbance was measured using a Biotek Synergy HTX plate reader. The percentage of growth inhibition was calculated for each well, using a negative control (media only) and positive control (bacteria or fungi without inhibitors). The significance of the inhibition values was determined by modified Z-scores, calculated using the median and mean average deviation (MAD) of the samples (no controls) on the same plate.

Supplemental Data1. CSαβ sequences in public domain. All 2ds- and other CSαβ amino acid sequences identified in UniProt and NCBI databases. Available at https://doi.org/10.14264/uql.2018.613.

Supplemental Data S2. All CSαβ MSA. Multiple sequence alignment of all non-redundant (0.95 %) CSαβ amino acid sequences identified. Available at https://doi.org/10.14264/uql.2018.613.

Supplemental Data S3. Sequence space data. Transformed sequence space matrix and classification likelihoods used for sequence space analyses. Available at https://doi.org/10.14264/uql.2018.613.