JOURNAL OF PROTEOMICS 102 (2014) 1–10

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Multifunctional warheads: Diversification of the toxin arsenal of centipedes via novel multidomain transcripts

Eivind A.B. Undheim[a,b], *Kartik Sunagar*[c,d], *Brett R. Hamilton*[e], *Alun Jones*[a], *Deon J. Venter*[e,f], *Bryan G. Fry*[a,b,*], *Glenn F. King*[a,**]

[a]*Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Australia*
[b]*School of Biological Sciences, The University of Queensland, St. Lucia, QLD 4072, Australia*
[c]*Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal*
[d]*CIIMAR/CIMAR–Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Rua dos Bragas 289, P 4050-123 Porto, Portugal*
[e]*OMICS, Pathology Department, Mater Health Services, South Brisbane, QLD 4101, Australia*
[f]*School of Medicine, The University of Queensland, St. Lucia, QLD 4072, Australia*

ARTICLE INFO

Article history:
Received 5 January 2014
Accepted 21 February 2014
Available online 3 March 2014

Keywords:
Centipede venom
Multidomain transcript
Evolution
Posttranslational modification
MALDI imaging

ABSTRACT

Arthropod toxins are almost invariably encoded by transcripts encoding prepropeptides that are posttranslationally processed to yield a single mature toxin. In striking contrast to this paradigm, we used a complementary transcriptomic, proteomic and MALDI-imaging approach to identify four classes of multidomain centipede-toxin transcripts that each encodes multiple mature toxins. These multifunctional warheads comprise either: (1) repeats of linear peptides; (2) linear peptides preceding cysteine-rich peptides; (3) cysteine-rich peptides preceding linear peptides; or (4) repeats of linear peptides preceding cysteine-rich peptides. MALDI imaging of centipede venom glands revealed that these peptides are posttranslationally liberated from the original gene product in the venom gland and not by proteases following venom secretion. These multidomain transcripts exhibit a remarkable conservation of coding sequences, in striking contrast to monodomain toxin transcripts from related centipede species, and we demonstrate that they represent a rare class of predatory toxins that have evolved under strong negative selection. We hypothesize that the peptide toxins liberated from multidomain precursors might have synergistic modes of action, thereby allowing negative selection to dominate as the toxins encoded by the same transcript become increasingly interdependent.

Biological significance
These results have direct implications for understanding the evolution of centipede venoms, and highlight the importance of taking a multidisciplinary approach for the investigation of novel venoms. The potential synergistic actions of the mature peptides are also of relevance to the growing biodiscovery efforts aimed at centipede venom. We also demonstrate the application of

* *Correspondence to:* B.G. Fry, School of Biological Sciences, The University of Queensland, St Lucia, QLD 4072, Australia. Tel.: +61 400 1931 832.
** *Correspondence to:* G.F. King, Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, QLD 4072, Australia. Tel.: +61 7 3346 2025; fax: +61 7 3346 2021.
E-mail addresses: bgfry@uq.edu.au (B.G. Fry), glenn.king@imb.uq.edu.au (G.F. King).

http://dx.doi.org/10.1016/j.jprot.2014.02.024

MALDI imaging in providing a greater understanding of toxin production in venom glands. This is the first MALDI imaging data of any venom gland.

1. Introduction

Proteins earmarked for secretion are typically produced as prepropeptides comprised of a signal peptide and one or two propeptide regions that are posttranslationally excised by endoproteases to yield a single mature protein product [1–3]. There are, however, a number of deviations from this scheme, including transcripts that lack propeptide-encoding regions and transcripts that encode multiple mature proteins [3]. For example, in both vertebrates and invertebrates, neuropeptides and hormones are commonly produced as multifunctional precursors containing a signal peptide and multiple copies of the neuropeptide or hormone separated by propeptide regions [4].

Multifunctional toxin transcripts, however, are exceedingly rare in most venomous taxa, with the reptilian clade Toxicofera being the notable exception. Various strategies leading to multifunctional toxin transcripts have evolved both convergently and divergently on several occasions within Toxicofera. These include the duplication events leading to precursors encoding tandem stretches of sarafotoxins in *Atractaspis* snakes [5], seven newly evolved bradykinin potentiating peptides in the propeptide region of the precursor encoding a C-type natriuretic peptide in the pit viper *Bothrops jararaca* [6], and multiple helokinestatin peptides in the propeptide region of the precursor encoding a B-type natriuretic peptide in Helodermatidae and Anguiidae lizard venoms [7–9]. The venom glands of coleoid cephalopods (cuttlefish, octopus and squid) also produce multifunctional transcripts encoding 3–4 pacifastin peptides that are posttranslationally liberated [10].

In striking contrast to coleoids and toxicoferans, invertebrate venomous animals such as marine cone snails, hymenopterans, sea anemones, scorpions, and spiders strictly adhere to the canonical one gene–one toxin strategy [11,12]. The most widely studied venomous arthropods, namely spiders and scorpions, generate venom diversity via expression of numerous isoforms of each toxin type rather than via multiple posttranslational modifications of a single translated product [12,13]. Their impressive toxin arsenal [14] appears to have evolved through classical gene duplication events followed by explosive diversification driven by positive selection [12,15]. In contrast, transcripts encoding multiple mature toxins are extremely rare in arthropods, and have only been noted for latarcins, linear antimicrobial peptides found in the venom of the spider *Lachesana tarabaevi* [16].

Centipedes may be the oldest extant terrestrial venomous lineage, having arisen more than 400 million years ago (Mya) [17]. Reflecting this ancient divergence, the centipede venom apparatus as well as most centipede toxins described to date bear little resemblance to those of other arthropods [18–20]. However, the centipede toxin transcripts described to date conform to the arthropod paradigm of encoding a prepropeptide containing a single mature toxin domain. In striking contrast, we describe here four different types of multidomain transcripts from the venom gland of four species of scolopendrid centipede and use MALDI imaging to show that these multifunctional "warheads" are activated in the venom gland prior to venom expulsion.

2. Material and methods

2.1. Specimen and venom collection

Ethmostigmus rubripes was purchased from Mini Beast Wildlife (www.minibeastwildlife.com.au), *Scolopendra morsitans* was collected from the Darling Downs region, Queensland, Australia, and *Cormocephalus westwoodi* was collected from the Launceston region, Tasmania, Australia; all were identified according to Koch [21–23]. *Scolopendra alternans* (Haiti) were purchased from La Ferme Tropicale (www.lafermetropicale.com). For venom collection, centipedes were starved for 3 weeks, then anesthetized with CO_2 and venom extracted by electrostimulation (12 V, 1 mA). All species were milked except *S. alternans*. Venom was immediately lyophilised and stored until further use at −80 °C.

2.2. cDNA library construction

Four days after venom depletion by electrostimulation, the venom glands were removed from five anesthetized specimens, flash frozen, and pooled. Total RNA was extracted by using TRIzol (Life Technologies) and enriched for mRNA using a DynaBeads Direct mRNA kit (Life Technologies). mRNA was reverse transcribed, fragmented and ligated to a unique 10-base multiplex identifier (MID) tag and applied to a PicoTitrePlate for simultaneous amplification and sequencing on a Roche 454 GS FLX+ Titanium platform (Australian Genome Research Facility). Automated grouping and analysis of sample-specific MID reads enabled informatic separation of sequences from other transcriptomes on the plates. Low-quality sequences were removed prior to de novo contig assembly using MIRA (version 3.4.0.1). Assembly details (number of reads, average read length, number of contigs and average assembled bases per contig) were: *E.rubripes* 72740, 375, 6980, 1035; *C. westwoodi* 48041, 376, 1706, 544; *S. alternans* 57175, 355, 5044, 612; *S. morsitans* 93436, 356, 6029, 621. Contigs were processed and analyzed using CLC Main Work Bench (ver. 6.2; CLC bio) and the Blast2GO bioinformatic suite [24,25]. To identify putative toxin transcripts, each transcriptome was searched against the Tox-Prot database (http://www.uniprot.org/program/Toxins) to which additional functionally annotated centipede toxin sequences were added [18,20]; the results are shown in Supplementary Fig. 1. Data can be accessed at the National Center for Biotechnology Information under bioprojects PRJNA200639 (*E. rubripes*), PRJNA200641 (*C. westwoodi*), PRJNA200753 (*S. alternans*), and PRJNA200640 (*S. morsitans*).

2.3. LC–MALDI MS

All milked venoms were analyzed using LC–MALDI MS. Crude venom was reconstituted in 0.1% formic acid and 2% acetonitrile (ACN), then particulates were removed by centrifugation. Venom was fractionated by reverse phase-HPLC (RP-HPLC) on an Agilent 1100 series nano-HPLC using a Vydac C18 column (300 μm × 150 mm, 5 μm particle size, 300 Å pore size). RP-HPLC fractions were spotted directly onto a MALDI plate using a Shimadzu Accuspot NSM-1 before batch analysis using a 4700 MALDI TOF/TOF Proteomics Analyser (AB Sciex) in positive reflectron mode; ions of m/z 900–8000 were acquired by accumulating 2500 laser desorptions/spectrum. Samples were analyzed twice, using either α-cyano-4-hydroxycinnamic acid (CHCA) or 1,5-diaminonaphthalene (1,5-DAN) as matrix. For CHCA-spotted samples, high intensity ions with m/z <3500 were manually selected for MS/MS. MS/MS experiments were run twice, with and without nitrogen gas in the collision cell for collision induced dissociation (CID), in both cases using a relative mass precursor window of 200 resolution (full width at half maximum), enabling metastable ion suppression, and accumulating 2000 shots/spectrum.

For generation of sequence tags by in-source decay (ISD) using 1,5-DAN, sample spots corresponding to >95% pure peptide larger than 3 kDa (as determined from initial analysis using CHCA) were re-spotted using 1,5 DAN matrix and immediately analyzed using ISD. Mass ranges were set from m/z 1000 up to 200 m/z higher than the precursor ion; up to 10 acquisition rounds were accumulated to maximize resolution and signal-to-noise ratio (S/N) for fragment ions. ISD spectra were manually interpreted, while CID spectra were automatically searched against the respective transcriptomes translated to all reading frames using Protein Pilot v4.5 (AB Sciex). The searches allowed for biological modifications and amino acid substitutions. Spectra were inspected manually to eliminate false positives, excluding spectra with low S/N, erroneous modification assignments, and confidence values below 99%.

2.4. Shotgun-LC–ESI MS/MS

For shotgun sequencing of milked venoms, peptides were reduced and alkylated [26] then 5 μg was incubated overnight with 0.02 mg/mL trypsin in 40 mM ammonium bicarbonate, pH 8.0. The reduced/alkylated sample (0.75 μg) was then processed by LC–MS/MS, using a, Agilent Zorbax C18 column (2.1 × 100 mm, 1.8 μm particle size, 300 Å pore size) at a flow of 400 μl/min and a gradient of 1–40% solvent B (90% ACN, 0.1% formic acid) in 0.1% formic acid over 15 min on a Nexera UHPLC (Shimadzu) coupled with an AB Sciex Triple TOF 5600 mass spectrometer equipped with a Turbo V ion source heated to 450 °C. MS^2 spectra were acquired at 20 scans/s, with a cycle time of 2.3 s, and optimized for high resolution. Precursor ions with m/z 300–1800, a charge of +2 to +5, and an intensity of at least 120 counts/s were selected; a unit mass precursor ion inclusion window of ±0.7 Da was used, and isotopes within 2 Da were excluded for MS^2. MS^2 spectra were searched against translated transcriptomes from the corresponding centipede species using Proteinpilot v4.5 (ABSciex) as described above but with specification of the alkylation method (iodoethanol) and tryptic digestion. Spectra were inspected manually to eliminate false positives, excluding spectra with low S/N, erroneous modification assignments, and confidence values below 95% unless justifiable by the presence of a-ions after comparison with the theoretical MS/MS product ion spectrum. Proteins from each venom identified using this approach were BLAST searched against the NCBI non-redundant protein sequence database and the results are summarized in Supplementary Fig. 2.

2.5. MALDI imaging

A modification of published protocols [27,28] was used for MALDI imaging of venom glands from *E. rubripes*. The fixation process was optimized to improve the success of venom-gland sectioning while maintaining access to endogenous proteins/peptides. Briefly, centipedes were starved for 3 weeks then anesthetized with CO_2 before forcipules were cut off. Two longitudinal incisions were made along the cuticle to facilitate penetration of fixative, then the forcipules were left in 50% RCL2/ethanol at room temperature overnight. Venom glands were then dissected out, dehydrated sequentially using 70%, 90% and 100% ethanol (3 × 15 min at each concentration), cleared in xylene for 30 min, and embedded in paraffin wax before sectioning at 7 μm thickness. Sections were de-paraffinized by careful washing with xylene, and optically imaged prior to applying CHCA (7 mg/ml in 50% ACN, 0.2% trifluoroacetic acid) using a Bruker ImagePrep automated matrix sprayer. FlexControl 2.1 (Bruker) was used to operate an UltraFlex III TOF–TOF mass spectrometer (Bruker) in either linear positive or reflectron positive mode. For both modes of operation, a small laser size was chosen to give a spatial resolution of 50 μm, and matrix ion suppression was enabled up to m/z 980. MALDI-imaging experiments were performed using FlexImaging 2.1 (Bruker), and 200 laser shots per raster point were acquired. FlexImaging was subsequently used to visualize data in 2D ion intensity maps, producing an averaged spectrum based upon the normalized individual spectra acquired during the experiment. This approach enabled the distribution of individual ions to be related to the venom-gland Sections.

2.6. Molecular evolution analyses

We evaluated the influence of natural selection using maximum-likelihood models [29,30] implemented in the CODEML program of the PAML package [31]. We compared likelihood values for three pairs of models with different assumed ω distributions: M0 (constant ω across all sites) versus M3 (allows ω to vary across sites within n discrete categories, with $n \geq 3$); M1a (a model of neutral evolution) where all sites are assumed to be under negative ($\omega < 1$) or neutral selection ($\omega = 1$) versus M2a (a model of positive selection) which in addition to the site classes mentioned for M1a, assumes a third category of sites; sites with $\omega > 1$ (positive selection) and M7 (β) versus M8 (β and ω), and models that mirror the evolutionary constraints of M1 and M2 but assume that ω values are drawn from a β distribution [32]. Only if the alternative models (M3, M2a and M8: allow sites with $\omega > 1$) yield a better fit in a Likelihood Ratio Test (LRT) relative to their null models (M0, M1a and M7: do not allow sites $\omega > 1$), are their results considered significant. LRT is estimated as twice the difference in maximum likelihood values between

nested models and compared to the χ^2 distribution with the appropriate degree of freedom (i.e., the difference in the number of parameters between the two models). The Bayes empirical Bayes (BEB) approach [33] was used to identify amino acids under positive selection by calculating the posterior probability that a particular amino acid belongs to a given selection class (neutral, conserved or highly variable). Sites with posterior probability ≥95% of belonging to the 'ω > 1 class' were inferred to be positively selected.

Single Likelihood Ancestor Counting (SLAC), Fixed-Effects Likelihood (FEL), REL, and Fast, Unconstrained Bayesian AppRoximation (FUBAR) [34,35] implemented in HyPhy [36] were employed to support the aforementioned analyses and to detect sites evolving under the influence of positive and negative selection. MEME [37] was also used to detect episodically diversifying sites. To clearly depict the proportion of sites under different regimes of selection, an evolutionary fingerprint analysis was carried out using the evolutionary selection distance (ESD) algorithm implemented in datamonkey [38]. All sequence alignments are available in SI Files 2–4.

3. Results

3.1. Identification of multidomain toxin transcripts in centipedes

Due to their large size and clinical importance, we selected four representative centipede species from the family Scolopendridae that facilitated comparisons at the subfamily, genus, and species level: *E. rubripes* (Scolopendridae; Otostigminae), *C. westwoodii* (Scolopendridae; Scolopendrinae), and *S. alternans* and *S. morsitans* (Scolopendrinae). The split of Otostigminae from Scolipendrinae about 300 Mya enabled approximate dating of evolutionary events. Venoms were characterized by next-generation sequencing of transcriptomes from regenerating venom glands and, where venom was available, complemented by both bottom-up and top-down proteomics in order to identify mature venom components. Peptides are named according to the rational nomenclature proposed for venom peptides [39], with scoloptoxin (SLPTX) indicating that toxins derive from a scolopendrid centipede [20]. Decimals denote the relative position of each encoded peptide on a transcript, with lower numbers indicating proximity to the 5′ end.

Linear and disulfide-rich peptides were identified by searching MS/MS and ISD-MS spectra, respectively, of high intensity precursor ions from LC–MALDI analysis of *E. rubripes* venom and matching sequences against the corresponding venom-gland transcriptome (Fig. 1). Remarkably, this strategy revealed three types of multidomain transcripts, each encoding a single prepropeptide product that is posttranslationally cleaved to yield several mature peptides (Table 1): Type I transcripts encode a linear peptide preceding a cysteine-rich peptide (U-SLPTX-Er1.1 and U-SLPTX-Er1.2; Fig. 1); Type II transcripts encode a cysteine-rich peptide preceding a linear peptide (U-SLPTX-Er4.1 and U-SLPTX-Er4.2; Fig. 2); and Type III transcripts encode repeats of linear peptides (U-SLPTX-Er5.1 and U-SLPTX-Er5.2; Fig. 3). These transcripts are the first examples of multidomain toxin precursors in centipedes as well as the first examples of multidomain transcripts encoding cysteine-rich toxins in any venomous arthropod.

3.2. Taxonomic distribution of multidomain centipede-toxin transcripts

A BLAST search of each *E. rubripes* multidomain transcript against venom-gland transcriptomes from *C. westwoodi*, *S. morsitans*, and *S. alternans* uncovered similar transcripts (Figs. 1–3). Thus, multidomain transcripts appear to be a common strategy for diversifying the venom arsenal of centipedes. However, the various transcript classes were not universally conserved across all species.

Intriguingly, the cysteine-rich domains of Er1.2 and Er2.1 from *E. rubripes* were homologous to sequences recovered from *Scolopendra*, but the transcripts encoding these latter toxins lacked the linear domains found in the *E. rubripes* transcript. In the case of Er1, only one ortholog was found in *S. morsitans* (Fig. 1). The sequence corresponding to the linear peptide in Er1.1 is almost entirely absent in the *S. morsitans* precursor, with the notable exception of the C-terminal "ER" cleavage site, and the "AR" cleavage site used to release Er1.1 is missing from the propeptide region. As a result, the *S. morsitans* precursor encodes only a single, cysteine-rich toxin. Hence, as shown in Fig. 4, Type I transcripts most likely evolved after the divergence of Otostigminae from Scolopendrinae around 300 Mya [40] via extension of the propeptide region and addition of an N-terminal "AR" processing site. The newly formed linear toxin (Er1.1) is presumably liberated from the cysteine-rich domain (Er1.2) via the same mechanism that releases the ancestral cysteine-rich peptide in *S. morsitans* from the N-terminal propeptide.

In the case of Er2, single-domain orthologs were found in both *S. morsitans* and *Scolopendra subspinipes dehaani* (Fig. 2). In *E. rubripes* Er2, the cysteine-rich domain (Er2.1) is followed by a dibasic "KR" cleavage site that immediately precedes the linear domain (Er2.2). In contrast to Er1 from *E. rubripes*, this additional domain does not appear to be an extension or functionalization of a propeptide domain as the ortholog recovered from *S. morsitans* lacks a C-terminal processing signal while the orthologs from *S. subspinipes dehaani* lack C-terminal propeptide regions entirely. Hence, as for Type I, Type II transcripts appear to have evolved after the divergence of Otostigminae from Scolopendrinae (Fig. 4), but in this case via addition of a new domain at the C-terminus as well as a C-terminal "KR" processing site.

A BLAST search revealed orthologs of *E. rubripes* Er5 in venom-gland transcriptomes from *C. westwoodi*, *S. morsitans*, and *S. alternans* (Fig. 3). The C-terminal sequences of Er5.1 and Er5.2 are identical, and they contain four of the first five residues of a sequence motif ("RLWRNWE") that is repeated several times in the Er5-like transcripts recovered from all species. However, these latter transcripts lack domains corresponding to the linear peptides Er5.1 and Er5.2. Since transcripts with and without Er5.1 and Er5.2 domains were recovered from *E. rubripes*, it is likely that Type III transcripts were recruited after the split between the two scolopendrid subfamilies Otostigminae and Scolopendrinae about 300 Mya (Fig. 4).

Notably, the RLWRNWE repeats are flanked by N-terminal dibasic sites ("KR" or "RR") and C-terminal sequences ("NW") that correspond to the C-terminal processing sites for Er5.1 and Er5.2. Hence it is conceivable that these transcripts are

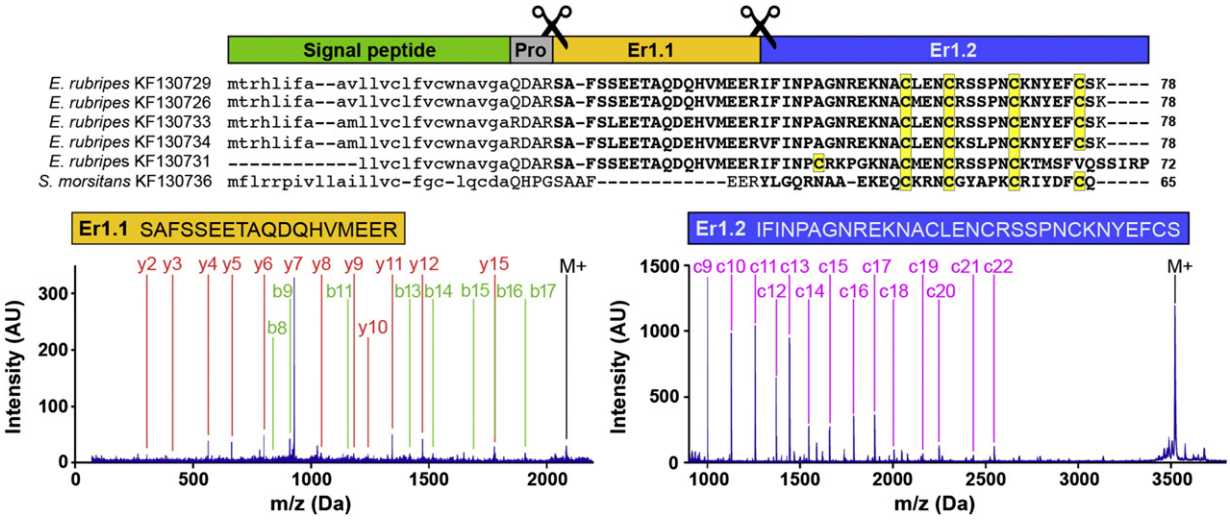

Fig. 1 – Type I multidomain centipede-toxin transcripts. Translated sequence alignments of Er1 precursors from *Ethmostigmus rubripes* and an ortholog from *Scolopendra morsitans*. Signal peptides are in lowercase, cysteines are highlighted in yellow, and mature peptides are shown in bold. A schematic representation of the signal peptide, propeptide and mature peptide domains is displayed above the alignment along with cleavage sites. Mass spectra used for identification of mature Er1.1 and Er1.2 in the secreted venom are displayed below the alignment and are labeled according to fragment-ion type.

processed to yield multiple short "LWRN" peptides that were too small to detect in this study.

An additional cysteine-rich domain was found in the RLWRNWE-repeat-containing transcript from *C. westwoodi*, which we named U_1-SLPTX-Cw1a (henceforth Cw1a) (Table 1; Fig. 3). This peptide was also identified in trypsinized venom from the same species, supporting its existence in secreted venom (Supplementary Fig. 3). Hence, the *C. westwoodi* transcript corresponds to a fourth class of multidomain centipede-toxin transcript (Type IV) in which multiple linear peptide domains precede a cysteine-rich toxin domain. The intriguing cysteine-rich domain encoded by this transcript is presumably a more recent recruitment that occurred after the split between *Scolopendra* and *Cormocephalus* approximately 250 Mya [40] (Fig. 4).

BLAST searches revealed that none of the mature toxins produced from the centipede multidomain transcripts are homologous to peptides or proteins with known function. However, the *E. rubripes* cysteine-rich mature toxin Er2.1 (Fig. 2) is homologous to putative neurotoxins with no assigned function from *S. subspinipes dehaani* and *Scolopendra subspinipes mutilans* [18,20]. As for the transcript recovered from *S. morsitans*, the toxin precursors encoding Er2.1-like sequences in *S. s. dehaani* and *S. s. mutilans* lack the C-terminal linear peptide domain found in *E. rubripes*. Transcripts containing LWRNWE repeats were also found in the venom gland of *S. s. dehaani*, but they lack domains

Table 1 – Sequences of mature peptide toxins encoded by centipede multidomain transcripts that were recovered by mass spectrometry sequencing of crude venom.

Toxin name	Species	Sequence of mature secreted peptide	NCBI accession
Er1.1a	*E. rubripes*	SAFSSEETAQDQHVMEER	KF130724, KF130725, KF130726, KF130727, KF130728, KF130729, KF130730, KF130731, KF130732, KF130738
Er1.1b	*E. rubripes*	SAFSLEETAQDQHVMEER	KF130733, KF130734, KF130735
Er1.2a	*E. rubripes*	IFINPAGNREKNACLENCRSSPNCKNYEFCS	KF130724, KF130727, KF130728, KF130729
Er1.2b	*E. rubripes*	IFINPAGNREKNACMENCRSSPNCKNYEFCS	KF130725, KF130726, KF130730
Er1.2c	*E. rubripes*	IFINPAGNREKNACLENCRSSPNCENYEFCS	KF130733
Er1.2d	*E. rubripes*	VFINPAGNREKNACLENCKSLPNCKNYEFCS	KF130734, KF130735
Er1.2e	*E. rubripes*	IFINPCRKPGKNACMENCRSSPNCKTMSFVQSSIRP	KF130731
Er4.1a	*E. rubripes*	CPSDCKCDYRSSQCRPANDDVHPNVCIDHYCVVMNLA	KF130739, KF130740, KF130741, KF130742, KF130743, KF130744, KF130745, KF130746, KF130747, KF130748, KF130749, KF130750
Er4.2a	*E. rubripes*	EQRPELSPGAWDDSS	KF130739, KF130740, KF130741, KF130742, KF130743, KF130746, KF130747, KF130749, KF130748
Er4.2b	*E. rubripes*	EQRPELSPGALDDSS	KF130744, KF130745, KF130750
Er5.1a	*E. rubripes*	QVANEDDGEKAKELWRN	KF130754, KF130755
Er5.2a	*E. rubripes*	QVADLNDEQETQRDKRLWRN	KF130754, KF130755
Cw1a	*C. westwoodi*	LWRNEDQEVACTTKCSCSDNEIFSKVDHELTTSE TKRVPCCC	KF130762

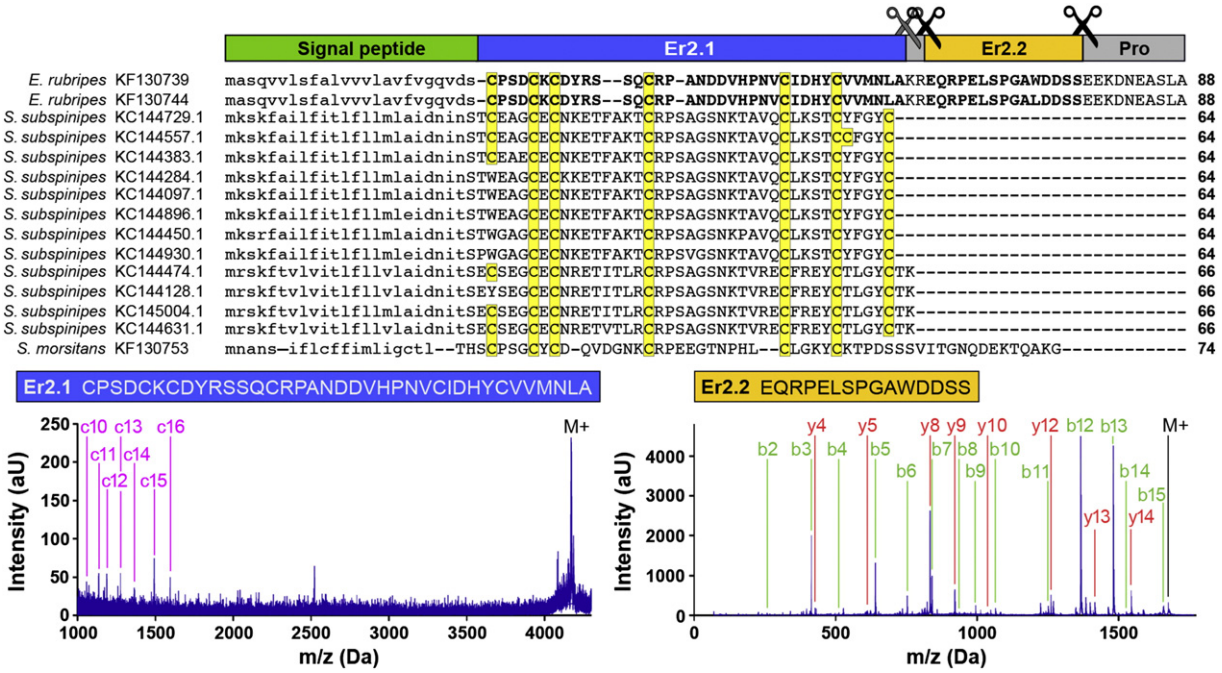

Fig. 2 – Type II multidomain centipede-toxin transcripts. Translated sequence alignments of Er2 precursor sequences from *Ethmostigmus rubripes* and homologous sequences from *Scolopendra morsitans* and *S. subspinipes dehaani* [18]. The sequences represent complete prepropeptides, where signal peptides are in lowercase, cysteines are highlighted in yellow, and mature peptides are shown in bold. A schematic representation of signal peptide, propeptide and mature peptide domains is shown above the alignment along with cleavage sites. Mass spectra used for identification of mature Er2.1 and Er2.2 in the secreted venom are displayed below the alignment and are labeled according to fragment-ion type.

corresponding to Er5.1 and Er5.2. Significant BLAST hits are summarized in Supplementary Table 1. There are no known structural or functional domains in any of the centipede-toxin multidomain transcripts aside from the signal peptides identified by InterPro.

3.3. *MALDI imaging*

We used MALDI imaging to determine whether prepropeptides encoding multiple mature toxins are processed in the centipede venom gland or if the mature toxins are only liberated from the precursor upon venom secretion. For all three types of multidomain transcripts identified in *E. rubripes*, the fully processed mature toxins were identified in the venom gland (Fig. 5). Thus, the mature toxins are clearly liberated from the multidomain precursors in the venom gland prior to venom expulsion and not by venom proteases upon secretion. This conclusion was reinforced by LC–MALDI-TOF analysis of the secreted venom, which revealed no ions with molecular masses corresponding to any of the unprocessed multidomain precursors.

3.4. *Molecular evolution of centipede multidomain toxin transcripts*

Venom proteins involved in prey capture are often subject to extensive positive selection and diversification as a result of an ongoing predator–prey arms race [41–43]. Co-expressed domains with divergent functions often exhibit different rates of evolution, perhaps due to differential evolvability of their cognate molecular targets in prey [44]. Different rates or regimes of natural selection can thus be an indication of the different roles that toxins play in the venom. We therefore investigated the molecular evolution of the centipede multidomain toxin transcripts and their monodomain homologues with a view to gaining insight into the likely importance of the encoded mature toxins in prey envenomation.

Molecular evolutionary assessments using various maximum-likelihood methodologies failed to detect variations in the coding sequences of centipede toxins (Supplementary Table 2). Site-model 8, which computes the non-synonymous to synonymous rate ratio (ω) across all sites in the alignment, indicated that each of the centipede multidomain toxin-transcripts evolved under the significant influence of negative selection. However, site-specific assessments are known to be influenced by sequence divergence and they also assume that the strength of selection remains constant across all lineages over time, which is not always biologically justified [37]. Centipedes, which may well be the oldest extant terrestrial venomous lineage, have relatively short generation times and hence accumulate tremendous sequence divergence. Therefore, to address the shortcomings of site-specific assessments, we employed the Mixed Effects Model of Evolution (MEME) which is known to reliably capture the molecular footprints of both episodic and pervasive diversifying selection [37]. MEME identified very few sites in all but one Er2 toxin transcript from *E. rubripes* as evolving under episodic diversifying selection, further supporting the

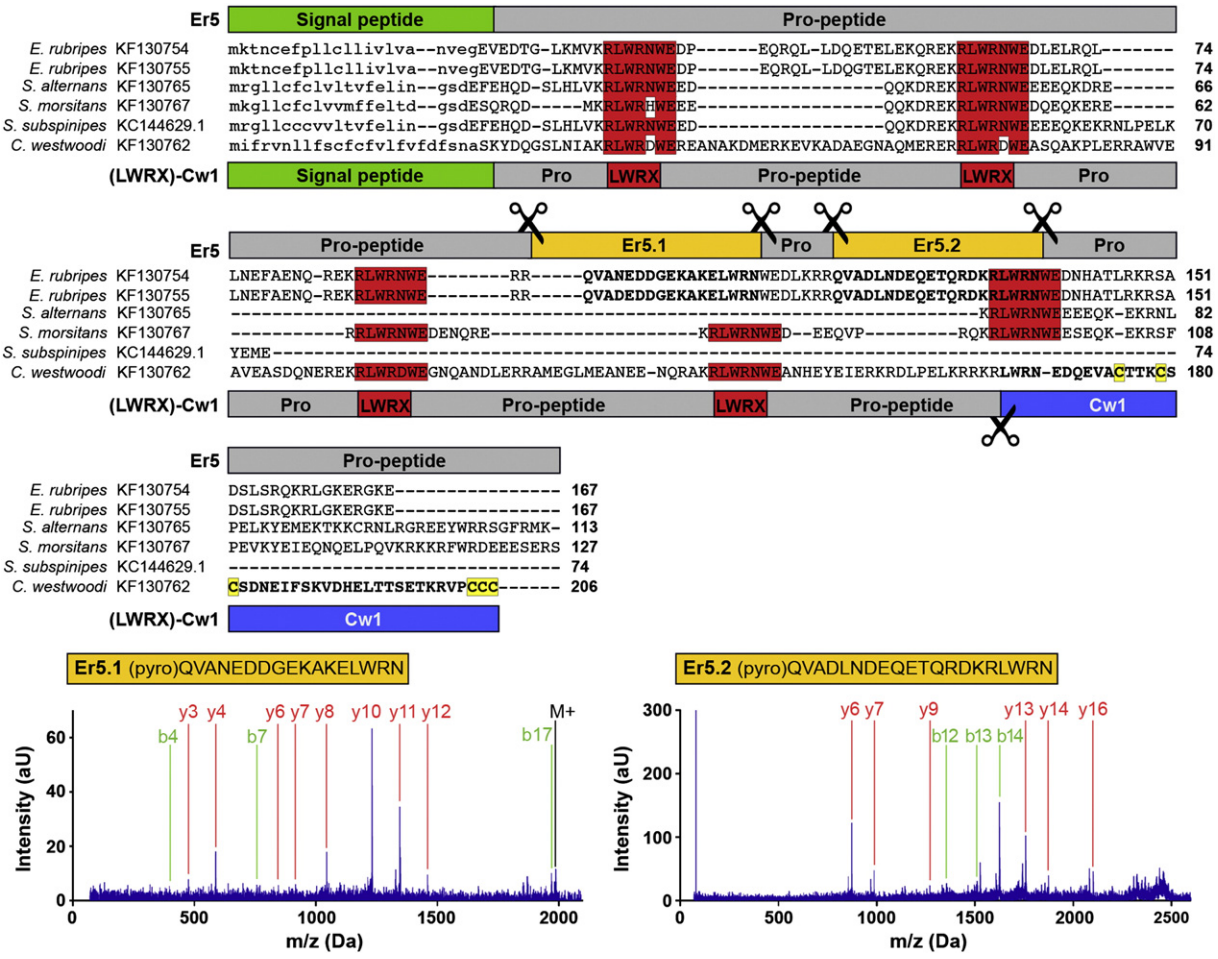

Fig. 3 – Type III and IV multidomain centipede-toxin transcripts. Trimmed translated sequence alignments of Er5 precursors from *Ethmostigmus rubripes* and homologous sequences from *Cormocephalus westwoodi*, *Scolopendra morsitans* and *S. subspinipes dehaani* [18]. The sequences represent complete prepropeptides, where signal peptides are in lowercase, cysteines are highlighted in yellow, RLWRNWE repeats are highlighted in red, and mature peptides are shown in bold. A schematic representation of the signal peptide, propeptide, and mature peptide domains of the Type III Er5 precursor, along with the location of protease cleavage sites, is shown above the alignment. A similar schematic representation of the coding structure of the Cw1a Type IV precursor is shown below the alignment, including the putative "LWRN" mature peptide domains. Mass spectra used for identification of mature Er5.1 and Er5.2 in the secreted venom are displayed below the alignment and are labeled according to fragment-ion type.

conclusion that centipede multidomain toxin transcripts have evolved under strong negative selection.

To complement these analyses, we also used an unconstrained Bayesian approximation method [35] to identify sites evolving under adaptive and purifying selection pressures. This approach identified only a few positively selected sites in the Er2.1 orthologs from *S. subspinipes*. In support of this, evolutionary fingerprint analyses revealed that while a large proportion of sites in most centipede multidomain toxin transcripts follow a regime of negative selection, a smaller proportion of sites in Er2.1 orthologs from *S. subspinipes* evolve under the influence of positive Darwinian selection (Supplementary Fig. 4). Branch-site random effect likelihood (REL) model analyses identified several lineages in all but three Er2.1 centipede-toxin transcripts as being subjected to episodic bursts of adaptive selection pressures (Supplementary Fig. 5), suggesting that even these highly conserved toxin transcripts fine tune their ability to target constantly evolving molecular receptors in prey from time to time.

Thus, various selection assessments conclusively highlight that most centipede multidomain toxin transcripts recovered in this study lack coding sequence variations. They therefore represent a rare class of predatory toxins that have evolved under the constraints of negative selection [43]. In contrast, the cysteine-rich Er2.1 monodomain orthologs from *S. subspinipes* accumulate greater amounts of variation, suggesting that they evolve under positive selection.

4. Discussion

Toxicoferan reptiles and coleoid cephalopods produce several types of multidomain toxin transcripts that contribute to venom complexity by posttranslational generation of functional variants

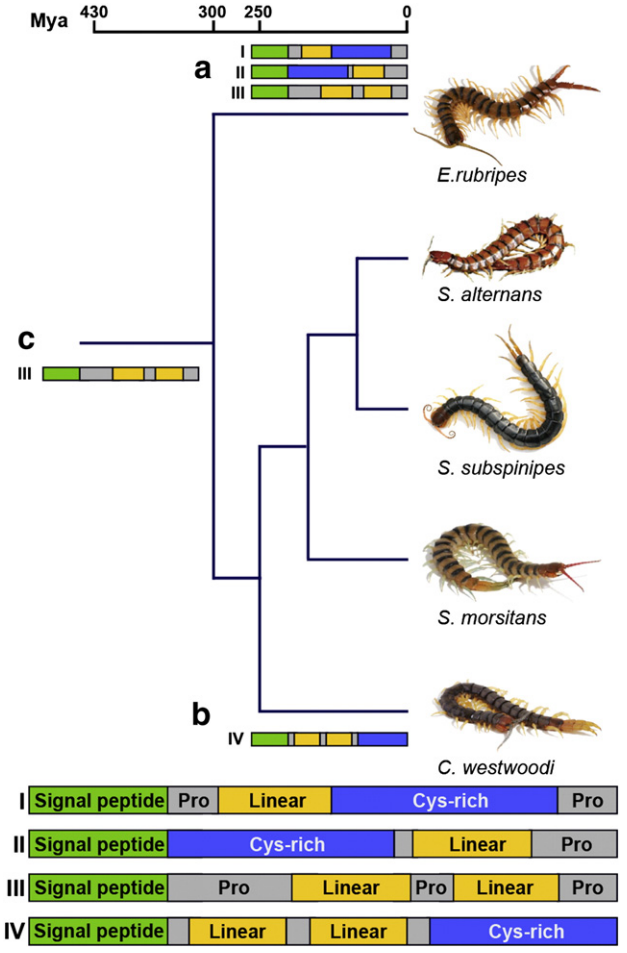

Fig. 4 – Graphical summary of the four types of multidomain centipede-toxin transcripts, with their earliest estimated recruitment indicated on a representational phylogenetic tree. (a) Type 1 (Er1), Type 2 (Er2) and Type 3 (Er5) transcripts in *E. rubripes*; (b) Type 4 transcript (Cw1a) in *C. westwoodii*; and (c) the possible early recruitment of LWRNWE-repeat-containing transcripts in an early scolopendrid ancestor. Toxin precursors on the cladogram are color-coded according to the schematic below the cladogram, with signal peptides, propeptides, mature linear peptides ("Linear"), and mature cysteine-rich toxins ("Cys-rich") colored green, gray, yellow and blue, respectively.

as well as entirely new toxin families [5,7,10,44]. In contrast, no such strategies have been reported for diversifying the venom arsenal of arthropods despite extensive characterization of several venomous taxa. Here we describe four types of multidomain toxin transcripts produced in the venom gland of centipedes. These transcripts encode either a linear peptide preceding a cysteine-rich toxin (Type I), a cysteine-rich toxin preceding a linear peptide (Type II), repeats of linear peptides (Type III), or repeats of linear peptides preceding a cysteine-rich toxin (Type IV).

Type I and II transcripts appear to have evolved by independent recruitment of linear peptide domains into genes encoding cysteine-rich toxins. Type I transcripts typified by Er1 (Fig. 1) most likely arose by extension and functionalization of an N-terminal propeptide region, whereas Type II transcripts typified by Er2 (Fig. 2) appear to have arisen through acquisition of an entirely new linear peptide domain via alternative splicing or loss of a stop codon. The Type III transcripts exemplified by Er5 (Fig. 3) and Type IV transcripts exemplified by the Cw1a-encoding transcript from *C. westwoodi* (Fig. 3) represent an interesting divergent use of the same family of LWRNWE-repeats. Type III transcripts appear to have arisen by duplication and divergence of an LWRNWE-containing linear peptide domain, whereas an additional C-terminal cysteine-rich domain has been appended to the LWRNWE-repeats in the Type IV transcripts (Fig. 4).

Centipede venom glands are essentially composite gland structures composed of multiple secretory units, each comprised of one or more secretory cells that release toxins into an extracellular storage body that is connected to the central chitinous duct through a one-way valve [19]. MALDI imaging revealed that each multidomain toxin precursor is processed in the venom gland to liberate the mature toxins prior to entering the venom duct. To our knowledge, this represents the first application of MALDI imaging to examine the distribution of peptide toxins in the venom gland of any animal. Due to the limited resolution of our MALDI imaging system (~50 μm), we were unable to determine whether precursor processing occurs in the secretory cell granules before release into the secretory body. However, the MALDI imaging data eliminate post-secretory proteolytic cleavage as a means of peptide maturation or as a source of experimental artifacts.

Most animal toxins evolve under the significant influence of positive selection, driven by an ongoing predator–prey arms race, and this is also the case for functional domains on multidomain transcripts [44,45]. However, in striking contrast to all previously described multidomain toxin transcripts, we found that all of the centipede toxins encoded by such transcripts are under the significant influence of negative selection with a distinct lack of coding sequence variation. Although some predatory toxins that act non-specifically on membrane lipids (e.g., snake-venom cytolytic three-finger toxins) evolve under the influence of negative selection, the lack of sequence variation in predatory toxins is considered a rare evolutionary phenomenon. In contrast, predatory toxins that attack more plastic molecular targets in prey are likely to evolve under the influence of positive selection and accumulate variations under an arms race scenario. Interestingly, our biochemical assessments revealed that mature toxins encoded by centipede multidomain transcripts are not cytotoxic. Furthermore, the relatively high abundance of these peptides in the secreted venom suggests that they play an important function in their respective venoms. Hence, it is likely that these centipede toxins attack molecular targets in the prey that have vital functions and consequently evolve under strong evolutionary constraints.

The strong influence of negative selection on centipede toxins encoded by multidomain transcripts could also be explained by toxins encoded on the same transcript having synergistic modes of action. The ability of a single precursor to liberate peptides that attack multiple targets simultaneously would eliminate the necessity to accumulate variations rapidly,

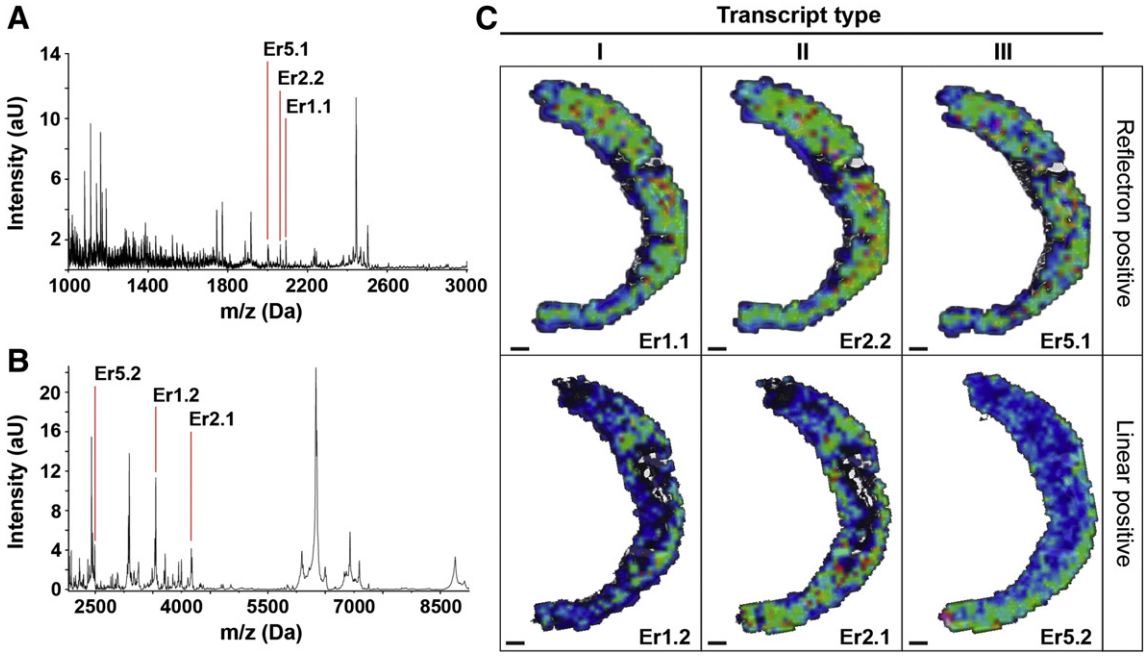

Fig. 5 – MALDI imaging mass spectrometry of *Ethmostigmus rubripes* venom gland. Distribution of Er1.1 and Er1.2, Er2.1 and Er2.2, and Er5.1 and Er5.2 in a central transverse section of an *E. rubripes* venom gland as obtained by MALDI imaging. Each spectrum is an average of all normalized spectra obtained across an entire section in reflectron (a) or linear (b) positive ion mode. Extracted ions corresponding to each mature peptide domain are displayed as heat maps according to their relative intensity in each normalized spectrum acquired on the tissue section (c). The scale bar in each of the MALDI images in panel (c) corresponds to 200 μm.

thereby allowing negative selection to become dominant as the synergistic modes of actions became increasingly interdependent. In support of this hypothesis, the single cysteine-rich toxin encoded by the *monodomain* Er2 transcript in *S. subspinipes dehaani* has accumulated much greater coding sequence variation than the corresponding cysteine-rich domain in the homologous *multidomain* Er2 precursor in *E. rubripes*. Moreover, selection assessments detected several sites in the *S. subspinipes dehaani* Er2 toxin evolving under the influence of positive selection. Multifunctional transcripts may therefore represent a strategy to reduce redundancy within each toxin class, thereby facilitating the expression of higher levels of functionally important toxin isoforms while ensuring that the complexity of the centipede's toxin arsenal is maintained in order to prevent the accumulation of venom resistance in prey.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2014.02.024.

Acknowledgments

This study was supported by the Australian Research Council (Discovery Grant DP1095728 to GFK), The University of Queensland (International Postgraduate Research Scholarship, UQ Centennial Scholarship, and UQ Advantage Top-Up Scholarship to EABU), the Norwegian State Education Loans Fund (EABU), and the Fundação para a Ciência e a Tecnologia (PhD scholarship via SFRH/BD/61959/2009 to KS).

REFERENCES

[1] Gierasch LM. Signal sequences. Biochemistry 1989;28:923–30.
[2] Duckert P, Brunak S, Blom N. Prediction of proprotein convertase cleavage sites. Protein Eng Des Sel 2004;17:107–12.
[3] Kozlov SA, Grishin EV. The universal algorithm of maturation for secretory and excretory protein precursors. Toxicon 2007;49:721–6.
[4] Rholam M, Fahy C. Processing of peptide and hormone precursors at the dibasic cleavage sites. Cell Mol Life Sci 2009;66:2075–91.
[5] Ducancel F, Matre V, Dupont C, Lajeunesse E, Wollberg Z, Bdolah A, et al. Cloning and sequence analysis of cDNAs encoding precursors of sarafotoxins. Evidence for an unusual "rosary-type" organization. J Biol Chem 1993;268:3052–5.
[6] Murayama N, Hayashi MA, Ohi H, Ferreira LA, Hermann VV, Saito H, et al. Cloning and sequence analysis of a *Bothrops jararaca* cDNA encoding a precursor of seven bradykinin-potentiating peptides and a C-type natriuretic peptide. Proc Natl Acad Sci U S A 1997;94:1189–93.
[7] Fry BG, Roelants K, Winter K, Hodgson WC, Griesman L, Kwok HF, et al. Novel venom proteins produced by differential domain-expression strategies in beaded lizards and Gila monsters (genus *Heloderma*). Mol Biol Evol 2010;27:395–407.
[8] Fry BG, Winter K, Norman JA, Roelants K, Nabuurs RJ, van Osch MJ, et al. Functional and structural diversification of the Anguimorpha lizard venom system. Mol Cell Proteomics 2010;9:2369–90.
[9] Koludarov I, Sunagar K, Undheim EA, Jackson TN, Ruder T, Whitehead D, et al. Structural and molecular diversification

of the Anguimorpha lizard mandibular venom gland system in the arboreal species *Abronia graminea*. J Mol Evol 2012;75:168–83.
[10] Ruder T, Sunagar K, Undheim EAB, Ali SA, Wai T-C, Low DHW, et al. Molecular phylogeny and evolution of the proteins encoded by coleoid (cuttlefish, octopus, and squid) posterior venom glands. J Mol Evol 2013;76:192–204.
[11] Olivera BM, Hillyard DR, Marsh M, Yoshikami D. Combinatorial peptide libraries in drug design: lessons from venomous cone snails. Trends Biotechnol 1995;13:422–6.
[12] Sollod BL, Wilson D, Zhaxybayeva O, Gogarten JP, Drinkwater R, King GF. Were arachnids the first to use combinatorial peptide libraries? Peptides 2005;26:131–9.
[13] Pineda SS, Wilson D, Mattick JS, King GF. The lethal toxin from Australian funnel-web spiders is encoded by an intronless gene. PLoS One 2012;7:e43699.
[14] King GF, Hardy MC. Spider-venom peptides: structure, pharmacology, and potential for control of insect pests. Annu Rev Entomol 2013;58:475–96.
[15] Zhu S, Bosmans F, Tytgat J. Adaptive evolution of scorpion sodium channel toxins. J Mol Evol 2004;58:145–53.
[16] Kozlov SA, Vassilevski AA, Feofanov AV, Surovoy AY, Karpunin DV, Grishin EV. Latarcins, antimicrobial and cytolytic peptides from the venom of the spider *Lachesana tarabaevi* (Zodariidae) that exemplify biomolecular diversity. J Biol Chem 2006;281:20983–92.
[17] Shear WA, Edgecombe GD. The geological record and phylogeny of the Myriapoda. Arthropod Struct Dev 2009;39:174–90.
[18] Liu Z-C, Zhang R, Zhao F, Chen Z-M, Liu H-W, Wang Y-J, et al. Venomic and transcriptomic analysis of centipede *Scolopendra subspinipes dehaani*. J Proteome Res 2012;11:6197–212.
[19] Undheim EAB, King GF. On the venom system of centipedes (Chilopoda), a neglected group of venomous animals. Toxicon 2011;57:512–24.
[20] Yang S, Liu Z, Xiao Y, Li Y, Rong M, Liang S, et al. Chemical punch packed in venoms makes centipedes excellent predators. Mol Cell Proteomics 2012;11:640–50.
[21] Koch LE. A taxonomic study of the centipede genus *Ethmostigmus Pocock* (Chilopoda: Scolopendridae: Otostigminae) in Australia. Aust J Zool 1983;31:835–49.
[22] Koch LE. Revision of the Australian centipedes of the genus *Cormocephalus Newport* (Chilopoda: Scolopendridae: Scolopendrinae). Aust J Zool 1983;31:799–833.
[23] Koch LE. Morphological characters of Australian scolopendrid centipedes, and the taxonomy and distribution of *Scolopendra morsitans* L. (Chilopoda: Scolopendridae: Scolopendrinae). Aust J Zool 1983;31:79–91.
[24] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674–6.
[25] Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 2008;36:3420–35.
[26] Hale JE, Butler JP, Gelfanova V, You JS, Knierman MD. A simplified procedure for the reduction and alkylation of cysteine residues in proteins prior to proteolytic digestion and mass spectral analysis. Anal Biochem 2004;333:174–81.
[27] Yarnold JE, Hamilton BR, Welsh DT, Pool GF, Venter DJ, Carroll AR. High resolution spatial mapping of brominated pyrrole-2-aminoimidazole alkaloids distributions in the marine sponge *Stylissa flabellata* via MALDI-mass spectrometry imaging. Mol Biosyst 2012;8:2249–59.
[28] Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Anal Chem 1997;69:4751–60.
[29] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 1994;11:725–36.
[30] Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998;15:568–73.
[31] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–91.
[32] Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 1998;148:929–36.
[33] Yang Z, Wong WSW, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 2005;22:1107–18.
[34] Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005;22:1208–22.
[35] Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol 2013;30:1196–205.
[36] Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics 2005;21:676–9.
[37] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 2012;8:e1002764.
[38] Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD. Evolutionary fingerprinting of genes. Mol Biol Evol 2010;27:520–36.
[39] King GF, Gentz MC, Escoubas P, Nicholson GM. A rational nomenclature for naming peptide toxins from spiders and other venomous animals. Toxicon 2008;52:264–76.
[40] Murienne J, Edgecombe GD, Giribet G. Including secondary structure, fossils and molecular dating in the centipede tree of life. Mol Phylogenet Evol 2010;57:301–13.
[41] Fry BG, Wüster W, Kini RM, Brusic V, Khan A, Venkataraman D, et al. Molecular evolution and phylogeny of elapid snake venom three-finger toxins. J Mol Evol 2003;57:110–29.
[42] Fry BG. From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res 2005;15:403–20.
[43] Casewell NR, Wuster W, Vonk FJ, Harrison RA, Fry BG. Complex cocktails: the evolutionary novelty of venoms. Trends Ecol Evol 2013;28:219–29.
[44] Brust A, Sunagar K, Undheim EA, Vetter I, Yang D, Casewell NR, et al. Differential evolution and neofunctionalization of snake venom metalloprotease domains. Mol Cell Proteomics 2013;12:651–63.
[45] Sunagar K, Johnson WE, O"Brien SJ, Vasconcelos V, Antunes A. Evolution of CRISPs associated with toxicoferan-reptilian venom and mammalian reproduction. Mol Biol Evol 2012;29:1807–22.