BIOCHEMISTRY

The authors note that on page E3782, left column, first full paragraph, lines 12–13, “a specialized chemoreceptory organ (called a siphon or osphradium)” should instead appear as “a specialized chemoreceptory organ called an osphradium.”

www.pnas.org/cgi/doi/10.1073/pnas.1520243112
Optimized deep-targeted proteotranscriptomic profiling reveals unexplored Conus toxin diversity and novel cysteine frameworks

Vincent Lavergne\textsuperscript{a}, Ivon Harliwong\textsuperscript{b}, Alun Jones\textsuperscript{a}, David Miller\textsuperscript{b}, Ryan J. Taft\textsuperscript{b}, and Paul F. Alewood\textsuperscript{a,1}

\textsuperscript{a}Division of Chemistry and Structural Biology, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia; and \textsuperscript{b}Division of Genomics and Computational Biology, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved May 27, 2015 (received for review January 27, 2015)

Cone snails are predatory marine gastropods characterized by a sophisticated venom apparatus responsible for the biosynthesis and delivery of complex mixtures of cysteine-rich toxin peptides. These conotoxins fold into small highly structured frameworks, allowing them to potently and selectively interact with heterologous ion channels and receptors. Approximately 2,000 toxins from an estimated number of >70,000 bioactive peptides have been identified in the genus Conus to date. Here, we describe a high-resolution interrogation of the transcriptomes (available at www.ddbj.nig.ac.jp) and proteomes of the diverse compartments of the Conus episcupus venom apparatus. Using biochemical and bioinformatic tools, we found the highest number of conopeptides yet discovered in a single Conus specimen, with 3,305 novel precursor toxin sequences classified into 9 known superfamilies (A, I1, I2, M, O1, O2, S, T, Z), and identified 16 new superfamilies showing unique signal peptide signatures. We were also able to depict the largest population of venom peptides containing the pharmacologically active C-C-CC-C-C inhibitor cystine knot and CC-C-C motifs (168 and 44 toxins, respectively), as well as 208 new conotoxins displaying odd numbers of cysteine residues derived from known conotoxin motifs. Importantly, six novel cysteine-rich frameworks were revealed which may have novel pharmacology. Finally, analyses of codon usage bias and RNA-editing processes of the conotoxin transcripts demonstrate a specific conservation of the cysteine skeleton at the nucleic acid level and provide new insights about the origin of sequence hypervariability in mature toxin regions.

cysteine-rich peptides | conotoxin | transcriptomic | proteomic | bioinformatic

Cone snails are venomous marine gastropods molluscs from the genus Conus (family Conidae), with 706 valid species currently recognized (on April 29, 2015) in the World Register of Marine Species (1). Over the last ∼30 million years, these species have evolved sophisticated predatory and defense strategies, with the elaboration of a highly organized envenomation machinery (2). Their venom apparatus is responsible for the biosynthesis and maturation of short peptide neurotoxins called conotoxins (occasionally referred to as conopeptides) that, once injected in the prey or predator (fish, molluscs, or worms), act as fast-acting paralytics. When the cone snail senses waterborne chemical signals via a specialized chemoreceptor organ (called a siphon or ophthalm), searching behavior begins with the release and extension of the proboscis where, in its lumen, a single dart-like radula tooth loaded from the radular sac (RS) is tightly held by circular muscles and filled with venom (Fig. 1 A and B) (3–5). When the tip of the proboscis comes in contact with the target, the radula is rapidly propelled into the prey and acts like a hypodermic needle to inject the venom (6). This radula tooth then serves as a harpoon to bring the captured prey back to the mouth of the snail (Fig. 1C). The biochemical and cellular mechanisms of toxin synthesis, including their processing and packaging in secretory granules, are poorly described. Nevertheless, epithelial cells bordering the venom duct (VD) are most likely the site of conotoxin production, which may then be released into the duct’s lumen through a holocrine secretion process (Fig. 1B) (7). The muscular venom bulb triggers burst contractions for the circulation of the venom inside the duct up to the pharynx, where conotoxins may undergo sorting and maturation (8). In addition, it has been suggested that certain conotoxins could, to a much lesser extent, be specifically expressed by the salivary gland (SG) (9).

In compensation for their limited mobility, cone snails have developed a vast library of structurally diverse bioactive peptides for prey capture and defense (10). As a result of speciation, a high rate of hypermutations, and a remarkable number of post-translational modifications, little overlap of conopeptides between Conus species has been observed (11, 12), which has led to an estimation of >70,000 pharmacologically active conopeptides although fewer than 1% have been characterized to date (13). The precursor form of conotoxins is composed of three distinct regions: a highly conserved N-terminal endoplasmic reticulum (ER) signal region (used to classify the toxins into gene superfamilies), a central proregion, and a hypervariable mature region, typically between 10 and 35 amino acids long, characterized by conserved cysteine patterns and connectivities (14–16). Mature conotoxins are able to selectively modulate specific subtypes of voltage- or ligand-gated transporters, receptors, and ion channels, expressed in organisms allowing them to potently and selectively interact with heterologous membrane receptors, ion channels, or transporters. In contrast to earlier studies in which the richness and sequence hypervariability of lowly expressed toxins were overlooked, we now describe a comprehensive deep-targeted proteotranscriptomic approach that provides, to our knowledge, the first high-definition snapshot of the toxin arsenal of a venomous animal, Conus episcupus. The thousands of newly identified conotoxins include peptides with cysteine motifs present in FDA-approved molecules or currently undergoing clinical trials. Further highlights include novel cysteine scaffolds likely to unveil unique protein structure and pharmacology, as well as a new category of conotoxins with odd numbers of cysteine residues.

Author contributions: V.L., D.M., R.J.T., and P.F.A. designed research; V.L., I.H., and A.J. performed research; V.L. contributed new reagents/analytic tools; V.L. analyzed data; and V.L., R.J.T., and P.F.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the DNA Data Bank of Japan, www.ddbj.nig.ac.jp/ (accession nos. DRA003531, PRJDB83896, SAMD00029744, DRX030964, DRX034331, SAMD00029745, DRX030965, DRX034332, SAMD00029746, DRX030966, and DRX034333).

Significance

Venomous marine cone snails have evolved complex mixtures of fast-acting paralytic cysteine-rich peptides for prey capture and defense able to modulate specific heterologous membrane receptors, ion channels, or transporters. In contrast to earlier studies in which the richness and sequence hypervariability of lowly expressed toxins were overlooked, we now describe a comprehensive deep-targeted proteotranscriptomic approach that provides, to our knowledge, the first high-definition snapshot of the toxin arsenal of a venomous animal, Conus episcupus. The thousands of newly identified conotoxins include peptides with cysteine motifs present in FDA-approved molecules or currently undergoing clinical trials. Further highlights include novel cysteine scaffolds likely to unveil unique protein structure and pharmacology, as well as a new category of conotoxins with odd numbers of cysteine residues.

Author contributions: V.L., D.M., R.J.T., and P.F.A. designed research; V.L., I.H., and A.J. performed research; V.L. contributed new reagents/analytic tools; V.L. analyzed data; and V.L., R.J.T., and P.F.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the DNA Data Bank of Japan, www.ddbj.nig.ac.jp/ (accession nos. DRA003531, PRJDB83896, SAMD00029744, DRX030964, DRX034331, SAMD00029745, DRX030965, DRX034332, SAMD00029746, DRX030966, and DRX034333).

Significance

Venomous marine cone snails have evolved complex mixtures of fast-acting paralytic cysteine-rich peptides for prey capture and defense able to modulate specific heterologous membrane receptors, ion channels, or transporters. In contrast to earlier studies in which the richness and sequence hypervariability of lowly expressed toxins were overlooked, we now describe a comprehensive deep-targeted proteotranscriptomic approach that provides, to our knowledge, the first high-definition snapshot of the toxin arsenal of a venomous animal, Conus episcupus. The thousands of newly identified conotoxins include peptides with cysteine motifs present in FDA-approved molecules or currently undergoing clinical trials. Further highlights include novel cysteine scaffolds likely to unveil unique protein structure and pharmacology, as well as a new category of conotoxins with odd numbers of cysteine residues.

Author contributions: V.L., D.M., R.J.T., and P.F.A. designed research; V.L., I.H., and A.J. performed research; V.L. contributed new reagents/analytic tools; V.L. analyzed data; and V.L., R.J.T., and P.F.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the DNA Data Bank of Japan, www.ddbj.nig.ac.jp/ (accession nos. DRA003531, PRJDB83896, SAMD00029744, DRX030964, DRX034331, SAMD00029745, DRX030965, DRX034332, SAMD00029746, DRX030966, and DRX034333).

\textsuperscript{1}To whom correspondence should be addressed. Email: p.alewood@imb.uq.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1501334112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1501334112
broadly distributed along the phylogenetic spectrum (10), and are thus considered a rich source of molecular templates with diagnostic and therapeutic interests for the management of human neuro-pathic pain, epilepsy, cardiac infarction, and neurological diseases (10). As described in Table 1, 37 conotoxin cysteine patterns have been reported to date (8 of which have known disulfide bond connectivity) (15). Although cysteine bridges always improve toxin stability and provide resistance to enzymatic degradation, some cysteine frameworks combined to particular loop lengths are more pharmacologically relevant. For instance, ω-conotoxin MVIIA [C(6)C(6)CC(3)C(4)C] (the only FDA-approved venom-derived synthetic peptide, marketed under the name Prialt) (53) and ω-conotoxin CVID [C(6)C(6)CC(3)C(6)C] (phase II clinical trials) (54) both contain the inhibitor cystine knot (ICK) motif where cysteine residues are disposed in a C-C-CC-C-C pattern with a I-IV, II-V, III–VI connectivity. Also, conotoxins such as γ-conotoxin MrLA [C(3)CC(4)C(2)C; I–III, II–IV; phase II] and MrLC [C(3)CC(4)C(2)C; I–III, II–IV; phase II] are important drug leads (55). Despite a weak correlation between gene superfamilies and pharmacological properties, some functional redundancy among members of a same superfamily exists (56). To date, 16 empirical gene superfamilies (designated as A, D, I1, I2, 13, J, L, M, O1, O2, O3, P, S, T, V, Y) have been annotated (57), plus 31 novel superfamilies have been discovered during the past two years (38, 39, 46, 57–60).

Here we describe a deep-targeted pipeline used to analyze the transcriptomes and proteomes of the three main venom apparatus compartments (VD, RS, and SG) of the Bishop’s molluscivorous Conus episcopus. A comprehensive investigation of the cysteine patterns of several thousands of newly identified conotoxin sequences, classified into known and novel gene superfamilies, led to the characterization of numerous peptides containing the ICK and CC-C-C motifs, as well as six novel cysteine scaffolds. We also bring additional insights to explain the hypervariability of mature conotoxin sequences by showing the existence of a specific codon usage bias at the gene level.

Results

RNA Preparation and cDNA Library Sequencing. The lysis of the venom duct, radular sac, and salivary gland of a single C. episcopus specimen provided 401 ng/µL, 314 ng/µL, and 73 ng/µL of total RNA, respectively. The initial qualitative controls of these samples revealed a lack of ribosomal 28S peak along with a strong and sharp 18S band, suggesting that RNA integrity was suitable for library preparation. Lack of 28S rRNA was originally called “the hidden break” by H. Ishikawa (61) and has since been observed in the sea slugs Aplysia (62), insects (63), or nematode parasites (64). To our knowledge, this is the first time the existence of a hidden break has been reported in cone snails. After mRNA isolation and generation of cDNA libraries, we obtained inserts at concentrations of 12,461.6 pM (average of 445 bp; VD), 2,119.9 pM (462 bp; RS), and 803.6 pM (431 bp; SG). Next-generation paired-end sequencing gave rise to average numbers of reads of 20,885,730 (VD), 29,187,419 (RS), and 31,725,853 (SG) (Table 2) (read datasets are freely available at www.ddbj.nig.ac.jp). Filtering of sequences showing an average Phred+33 quality score of >30, and merging of paired-end reads led to a decrease in number of 15.45%, 21.94%, and 36.94% for VD, RS, and SG, respectively.

Tissue-specific sets of concatenated merged and unmerged reads were independently submitted to four de novo assemblers that produced contigs with consistent length ranges and read-vs.-contig mapping rates (Table 2). However, the number of conopeptides identified from the contigs remained very low compared

<table>
<thead>
<tr>
<th>Name</th>
<th>Cysteine pattern</th>
<th>Cysteine connectivity</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>CC-C-C</td>
<td>I-III, II-IV</td>
<td>(17)</td>
</tr>
<tr>
<td>II</td>
<td>CCC-CC-CC-CC-CC</td>
<td>—</td>
<td>(18)</td>
</tr>
<tr>
<td>III</td>
<td>CC-CC-CC-CC-CC</td>
<td>—</td>
<td>(19)</td>
</tr>
<tr>
<td>IV</td>
<td>C-C-C-CC-CC-CC</td>
<td>I-V, II-III, IV-VI</td>
<td>(20)</td>
</tr>
<tr>
<td>V</td>
<td>CC-CC-CC-CC-CC</td>
<td>I-III, IV-IV</td>
<td>(21)</td>
</tr>
<tr>
<td>VI/VII</td>
<td>C-C-C-CC-CC-CC</td>
<td>I-IV, II-V, III-VI</td>
<td>(22)</td>
</tr>
<tr>
<td>VII</td>
<td>C-C-C-CC-C-C-C-C</td>
<td>I-IV, II-V, III-VI</td>
<td>(23)</td>
</tr>
<tr>
<td>IX</td>
<td>C-C-C-CC-CC-CC</td>
<td>I-IV, II-V, III-VI</td>
<td>(24)</td>
</tr>
<tr>
<td>X</td>
<td>CC-CC-[PO]C</td>
<td>I-IV, II-III</td>
<td>(25)</td>
</tr>
<tr>
<td>XI</td>
<td>C-C-C-C-C-C-C-C</td>
<td>I-IV, II-VI, III-VII, V-VIII</td>
<td>(26)</td>
</tr>
<tr>
<td>XII</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(27)</td>
</tr>
<tr>
<td>XIII</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(28)</td>
</tr>
<tr>
<td>XIV</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(29)</td>
</tr>
<tr>
<td>XV</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(30)</td>
</tr>
<tr>
<td>XVI</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(31)</td>
</tr>
<tr>
<td>XVII</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(32)</td>
</tr>
<tr>
<td>XVIII</td>
<td>C-C-C-C-C-C-C</td>
<td>—</td>
<td>(33)</td>
</tr>
<tr>
<td>XIX</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(34)</td>
</tr>
<tr>
<td>XX</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(35)</td>
</tr>
<tr>
<td>XXI</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(36)</td>
</tr>
<tr>
<td>XXII</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(37)</td>
</tr>
<tr>
<td>XXIII</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(38)</td>
</tr>
<tr>
<td>XXIV</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(39)</td>
</tr>
<tr>
<td>XXV</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(40)</td>
</tr>
<tr>
<td>XXVI</td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(41)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(42)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(43)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(44)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(45)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(46)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(47)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(48)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(49)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(50)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(51)</td>
</tr>
<tr>
<td></td>
<td>C-C-C-C-C-C-C-C-C</td>
<td>—</td>
<td>(52)</td>
</tr>
</tbody>
</table>

The name, pattern, and connectivity of cysteine frameworks (“—” when unknown) are reported.

*GenBank accession no. HM003926.
with the direct analysis of the reads (24 score-3 and 6 score-2 contigs displaying signal and proregion cleavage sites were annotated as conotoxins). Also, when using a different assembly approach by pooling together all of the merged and unmerged reads from the three tissues, then by mapping back each tissue-specific set of reads to the contigs generated, fewer toxins were detected (2 score-3 and 3 score-2 toxins previously found with the first strategy). Moreover, their tissue origin couldn’t be retrieved precisely because the number of conotoxins identified tended to be uniformly distributed across the three different compartments.

**Protein Fractionation.** To confirm the presence at the protein level of conotoxin sequences identified in the transcriptomes, we investigated in parallel the proteomes of the venom duct, radular sac and salivary gland. Total protein samples were fractionated by HPLC and 1D-, and 2D-PAGE, giving rise to a total of 300 fractions that have been analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Fig. 2). Reversed-phase HPLC revealed a complex protein mixture in the venom gland of *C. episcopus*, compared with the radular sac and salivary gland samples (Fig. 2 A–D). From a quantitative point of view, the VD sample contains mainly small proteins and peptides (<28 kDa) (Fig. 2 E and F) whereas the major RS and SG components have masses of >28 kDa.

**New Precursor Conopeptides.** ConoSorter was able to identify two of the four full-length conopeptide precursors currently known in *C. episcopus* ([Epl], patent US 6797808/GenBank accession no. AR584835; and [Ep11.1] (65) precursors). The program also identified Pn10.1 and TxMMSK-02 precursors previously isolated from the related molluscivorous species *C. pennaceus* ([AR584835]) and *C. textile* ([Ep11.1] (65)) and confirmed the detection of 26 from *C. episcopus* (66) mature conotoxin regions. Indeed, 84 new precursor sequences from *Conus magnificus* μ-O-MIVIA (67), 26 from *C. episcopus* Ep6.1 (patent US 20020173449), 10 from *C. pennaceus* Pn5.1 precursor conotoxin (66), 7 from *Conus omaria* Om6.5 toxin (also called PnVIB in *C. pennaceus*) (68), and 5 from *C. pennaceus* PnMRC-012 (66) mature conotoxin regions have been deciphered (Fig. S1 and Dataset SL4).
In addition, ConoSorter annotated at the highest score (i.e., the signal, pro, and mature regions simultaneously) 3,303 (99.19%) novel precursor conopeptide sequences in VD, as well as 22 (0.66%) in RS and 5 (0.15%) in SG (≥40 amino acids in length, with >60% hydrophobic residues in their N-terminal region and containing a signal/proregion cleavage site) that were retained for further analysis (nucleotide and amino acid sequences are available in the DNA Data Bank of Japan and UniProtKB, respectively). A majority of VD conopeptides belong to the T (2,356 toxins; 76.78%), O1 (333; 10.08%), and S (199; 6.02%) superfamilies (Fig. 3). We observed that all of the RS and SG precursor conopeptides were also found in the VD, except for 2 lowly expressed RS-specific sequences (Ep21rs is 96.88% identical to Ep3057; Ep22rs shares 98.44% identity with Ep1412) (Table S1). Also, we noticed that these VD conotoxins belonged to the top 0.70% of the most expressed toxin transcripts.

A detailed examination of the similarity between these novel VD precursor sequences revealed 401 “parent” toxins (defined as the longest protein present in a cluster of similar sequences) with a ratio of 1 parent for 7.24 “variant” sequences, when a minimum identity threshold between parent/variant of 96% was applied (Fig. S2). Among multiple cutoffs tested (from 93% to 100%), only this optimal identity limit of 96% produced clusters of sequences all reflecting the characteristics of precursor conotoxins. Indeed, every single cluster contained conotoxins belonging to the same superfamily (as opposed to clusters created at identity thresholds of <96%) and sharing a moderate number of amino acid substitutions in their proregions, as well as high rates of sequence variability in their mature regions (at thresholds of >96% identity, the majority of clusters contained sequences with identical proregion and/or mature regions due to low average numbers of variant per parent toxin: ≤2.74). Moreover, 40 (9.98%) of these parent conotoxin transcripts were retrieved in the VD proteome (at a confidence ≥99%) (Dataset SI B).

Cysteine Motifs in Mature Conotoxins. We identified 1,448 unique cysteine-rich mature toxins (with ≥4 cysteines), among which 1,240 (85.64%) contained an even number of cysteine residues (4 cysteines, 881 sequences; 6 cysteines, 197 sequences; 8 cysteines, 95 sequences; 10 cysteines, 67 sequences) and 208 (14.36%) contained an odd number of cysteine residues (5:145; 7:35; 9:27; and 11:1). Among toxins with an even number of cysteines (104 retrieved at protein level) (Dataset SIC), 9 cone snail cysteine frameworks were represented (Fig. 4D): 44 mature toxins with framework I (CC-C-C), 834 with framework V (CC-CC), 3 with framework XIV (C-C-C-C-C-C), 6 with framework III (CC-C-C-C), 168 with framework VII (C-C-C-C-C-C-C-C), 6 with framework IX (C-C-C-C-C-C-C-C), 77 with framework XI (C-C-C-C-C-C-C-C-C), 15 with framework XXII (C-C-C-C-C-C-C-C-C-C-C-C-C-C), and 66 with framework VIII (C-C-C-C-C-C-C-C-C-C-C-C-C-C).

We then focused specifically on mature toxins containing the VI/VII pattern, which, when complemented by a I–IV, II–V, III–VI cysteine connectivity, forms the well-known ICK fold present in numerous drug leads, including the FDA-approved Ziconotide (53). A total of 166 (98.81%) mature peptides belong to the O1 superfamily, compared with 2 sequences (1.19%) classified in the O2 superfamily [a total of 563 mature conopeptides with framework VI/VII are currently known, among which the majority belong to the O1 (68.56%), O2 (10.66%), and O3 (5.68%) superfamilies]. All these new toxins share the general loop formula (0–16)C(6)C(5–9)CC(2–4)C(3–4)C(0–43) (Fig. 4B, Fig. S3, and Dataset SIC). Interestingly, we observed that several of these new toxins also share high similarity rates with the following: C. magnificus MfVIA μO-conotoxin, a modulator of the pain target Na, 1.8 voltage-gated sodium channels (67) (97% identity with Ep298, Ep299, Ep301, Ep311, Ep523, Ep525, and Ep615); C. pennaceus PnVIB α-conotoxin, which is able to block dihydrotryptophin-insensitive high voltage-activated calcium channels (68) (97% identity with Ep584, Ep587, and Ep589); and...
Although protein sequences containing an odd number of cysteines are usually considered as forming only dinitric structures through the formation of an interchain disulfide bridge, it has been observed that the unpaired cysteine residue can also undergo other types of posttranslational modification, such as ADP ribosylation (72), lipidations (e.g., S-acetylation or S-prenylation) (73), nitrosylation (74), or cysteinylation (75), which provide additional functionality to the molecule. Here, we reveal the presence of 208 new conotoxins containing an odd number of cysteines (5, 7, 9, or 11) that form 21 distinct cysteine patterns (Table S2). Although all these patterns have been observed in larger proteins isolated from other species (75,795 UniProtKB/Swiss-Prot entries), only 5 of them are present in 14 known conotoxins (C-C-C-C-C; C-C-C-C-C; CC-CCC; C-C-C-C-C-C; C-C-C-C-C-C-C). Moreover, 13 of the cysteine patterns described in Table S2 could be derived from known Conus scaffolds containing even numbers of cysteine residues (Fig. S5). New frameworks can have thus been created by insertion of an extra cysteine residue either at the N-terminal end, C-terminal end, or in the core of the sequence, or by deletion of a single cysteine amino acid. Further investigation of these new conotoxin sequences could reveal novel posttranslational modifications at their unpaired cysteine residue, conferring upon them new functional advantages, as recently demonstrated with Conus geographus μO8-conotoxin GVIIIC, which is able to stabilize the sodium channel blockade (75).

**Definition of New Cone Snail Gene Superfamilies.** We have recently reported new gene superfamilies in C. marmoreus (57). Briefly, unclassified signal peptides, isolated from precursor conotoxin transcripts containing proregions and mature Conus regions, were grouped into clusters sharing ≥75.00% identity. Batch pairwise alignments between unclassified signal peptides and ones for which a superfamily has been previously assigned were performed. When the entire cluster of novel signal sequences shared ≥53.30% identity with the signal sequences of any empirical superfamilies, the cluster was considered a putative novel superfamily. Here, the same strategy applied to 1,311 novel signal peptide sequences grouped into 105 clusters led to the identification of 39 precursor sequences classified into 16 categories of toxins that we propose to define as new superfamilies SF-Epi 1–16 (Fig. S6).

**Codon Usage Bias and RNA Editing.** The relative synonymous codon usage (RSCU) of transcripts encoding each group of parent precursor conotoxins and their corresponding isoforms (2,750 sequences distributed into 99 clusters) has been analyzed according to (i) their superfamilies, (ii) their individual signal, proregions and mature regions, and (iii) their cysteine frameworks. Although we were unable to find any correlations between codon usage and superfamilies, we observed that codons encoding cysteines in the mature region are highly conserved, but also specific to the position of the residue in the motif, and specific to the framework considered (Fig. 5A).

We also focused on mutations and potential RNA-editing processes among the clusters of parent and their variant sequences sharing at least 96% identity as described previously (Fig. S8). We observed that point base substitutions occurred more frequently in the mature part of the toxins than in their signal and proregions (except for C to T, and G to T permutations that seemed to be more or equally abundant in the signal peptide) (Fig. S8B). However, the rate of indels is much higher than base substitutions (~3–10 times in number). Interestingly, whereas insertions mainly take place in the mature part of conotoxins, the majority of deletions have been observed in the proregions.
### Table 3. New precursor conopeptide sequences with mature regions containing new cone snail cysteine patterns

<table>
<thead>
<tr>
<th>Cysteine pattern</th>
<th>Name</th>
<th>Precursor sequence</th>
<th>Frequency</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Six cysteines</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| CC-C-CC-C         | Arylsulfatase A (component C)  
(H. sapiens - P15289) (I–V, II–VI, III–IV) | MGASPSLVLALLAAGLAVARPNNVLFADDLGYDGL- 
GCYGHPSSTTNLDQLAAGGGLFDFTYPSVSLCT- 
PSRAALLTGLRVPVMMGYPGVLPVSSRGGLPLLE- 
VTFAEVLAAARGYLTGMAKWHLGVPQAGFLP- 
PHQGFHFLGIPYSHGDQPCQNLTCPAPTPCDG- 
GCDQGLVPILLANSVEAAPPWLPGLEARYMA- 
FAHDLMDAAQDQRDFPLLYASHTYHPQFG- 
QSAERSGRPGFSLMELDDAAVGLTMLAIGDL- 
GLLEELVFTADNPGMRMRSPGCSGLRCCGK- 
GTTYEGGVRPAPLAVPGIAVGTHELASSLD- 
LLPTLALAGAPLNPVLDFDLPLLGLTGGSKPR- 
QLSFYPsyDEVGFAVRGTGKHyHFTQGS- 
AHSQDAPPACHASSSLTAEPPPLYDLSDKPG- 
ENYNLLGQGAVATPELVQALKQOLLKAQLQDALA- 
AVTFGPsQVARGEDPALICCHPGCGTPPRPACC- 
HCPDPHA | — | — |
|                  | Ep214 | MMSKGLVLLTICLLLSSLTAVPLDGQHADQPAER- 
LQGDISELKHPFNVPKRCPPAACACGCG- 
PRCCTV | 1 | M |
| **CC-CC-C-C**     | Heat-stable enterotoxin ST-IA/ST-P  
(E. coli - P01559) (I–IV, II–V, III–VI) | Mkkklmlaifisvlsfpsqsosteldsskkeketalatkcc- 
Dvkkknsekkeksemnntffyclcclnmpacacgy | — | — |
|                  | Ep1802 | MRCLPvfvlllllliaspsvdarpktkdipoqasdq- 
Nakrilqvlksrkncrlqvcqcgqlavelsfhl- 
Wncmikqlckhrnsfvdkhydhvasnyiwtf | 1 | T |
|                  | Ep2291 | MrcLpVFVILLIIIATSPVDARPKTDMPALSFH- 
DdkquirrelQDRNCGIADECGGSEIKENEG- 
CkpcKlsldvkfkgdqtvPfavrRrissngr | 1 | T |
|                  | Ep1646 | MrcLpVFVtllllllaispsvdarpktkdipoqasdq- 
Nakrilqvlkscrkncrlqvcqcgghlwncmikq- 
Lckhrnsfvdkhydhvasnyiwtf | 1 | T |
|                  | Ep2642 | MrcLpVFVIllllllistsvdallktkdmplasfrd- 
DvKrtqltqlnkrgccpyfecccklldrktCicV- 
WlytgIDpdrktGdFQot | 1 | T |
|                  | Ep2653 | MrcLpVFVllllllalstpsvdallktkdmpaslfrd- 
DvKrtqltqlnkrgccpyfeccvqgGdLRcyr- 
LcknnmK | 1 | T |
|                  | Ep2036 | MrcLpVFvllllllaispsvdvrrpkaKddmplasfrd- 
DnplQLqIrrDVTscCPSqPCCRFyGyREMTLDEPT- 
KpcCmyT | 1 | T |
|                  | Ep1629 | MrcLpVFVllllllaispsvdarpktkdipoqasdq- 
DnakrilqvlksrkncrlqvcqcgfEIRKENV- 
Rtdfc | 1 | T |
|                  | Ep1609 | MrcLpVFvllllllaispsvdarpktkdipoqasdq- 
DnakrilqvlksrkncrlqvcqcgfEIRKENV- 
Hadc | 1 | T |
|                  | Ep1092 | MrcLpVFvllllllaispcaLCplkTEGdvPslLFPHD- 
NlkRtrRthlnirecsscdGwCcppAcGPctekvqcLs | 1 | T |
|                  | Ep1100 | MrcLpVFvllllllaispcaLCplkTEGdvPslLFPHD- 
NlkRtrRthlnirecsscdGrcpcAcGstenvHlcP | 1 | T |
|                  | Ep1109 | MrcLpVFvllllllaispcaLCplkTEGdvPslLFPHD- 
NlkRtrRthlnirecsscdGrcpcAcGstenvHlcP | 1 | T |
|                  | Ep1111 | MrcLpVFvllllllaispcaLCplkTEGdvPslLFPHD- 
NlkRtrRthlnirecsscdGwCcppAcGstenvHlcP | 1 | T |
|                  | Ep1112 | MrcLpVFvllllllaispcaLCplkTEGdvPslLFPHD- 
NlkRtrRthlnirecsscdGwCcppAcGstenvHlcP | 2 | T |
|                  | Ep1110 | MrcLpVFvllllllaispcaLCplkTEGdvPslLFPHD- 
NlkRtrRthlnirecsscdGwCcppAcGstenvHlcP | 1 | T |
| **CC-CC-CC**      | Protein YqcK  
(B. subtilis - P45945) | MkyvHyvgvNnvlSlesKinfyEyKfGvKkavKvKtd- 
YakfLettpgLntlnvAdevKgnQnHnFqV- 
DsllelVkhkhrlekFgFareemdtccyAvQ- 
DkFwtdPdpGnewFwFyTksnEvQkDssSc- 
CvTpsdittnscc | — | — |
Eight cysteines

<table>
<thead>
<tr>
<th>Cysteine pattern</th>
<th>Name</th>
<th>Precursor sequence</th>
<th>Frequency</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ep1587</td>
<td>MRCLPV/FVILLLIASSAPSVDARKTVIDIPQASQFDNALKLQVLKESKNCCRLQALASFDNDPLQRLVDTTRCCPSQPCRFG</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Ep2695</td>
<td>MRCLPV/FVILLLIASSAPSDALLTKDDMLASFRDDVRKRTQTLNKRCCQYDFAKRALQTLMDIRECCMGTPGCCPWG</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Eight cysteines

<table>
<thead>
<tr>
<th>Cysteine pattern</th>
<th>Name</th>
<th>Precursor sequence</th>
<th>Frequency</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ep1738</td>
<td>MRCLPV/FVILLLIASSAPSDARKTVIDIPQASQFDNALKLQVLKESKNCCRLWLRPLQTVPGCIEW-KADCSFRTSCWENFEWLSTTRCHLATISLSFHLWNMCIKQLKCHRHY</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Ep2668</td>
<td>MRCLPV/FVILLLIIASSPVDALLTKDDMLASFRDDVRKRTQTLNKRCCQYDFAKRALQTLMDIRECCMGTPGCCPWG</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Cont.

<table>
<thead>
<tr>
<th>Cysteine pattern</th>
<th>Name</th>
<th>Precursor sequence</th>
<th>Frequency</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ep1702</td>
<td>MRCLPV/FVILLLIASSAPSDARKTVIDIPQASQFDNALKLQVLKESKNCCRLQALASFDNDPLQRLVDTTRCCPSQPCRFG</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Ten cysteines

<table>
<thead>
<tr>
<th>Cysteine pattern</th>
<th>Name</th>
<th>Precursor sequence</th>
<th>Frequency</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ep1647</td>
<td>MRCLPV/FVILLLIASSAPSDARKTVIDIPQASQFDNALKLQVLKESKNCCRLQALASFDNDPLQRLVDTTRCCPSQPCRFG</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Cont.

<table>
<thead>
<tr>
<th>Cysteine pattern</th>
<th>Name</th>
<th>Precursor sequence</th>
<th>Frequency</th>
<th>Superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ep1647</td>
<td>MRCLPV/FVILLLIASSAPSDARKTVIDIPQASQFDNALKLQVLKESKNCCRLQALASFDNDPLQRLVDTTRCCPSQPCRFG</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Cont.

Discussion

With their extreme chemical, thermal, and proteolytic stability, small globular cysteine-rich peptides produced by predatory marine cone snails have been considered promising pharmacological alternatives that share the advantages of both small molecules (potential oral delivery, high tissue penetration, cellular internalization, weak immunogenicity) and large protein “biologics,” like antibodies (high affinity and specificity to clinical targets) (76). However, using traditional protein-centric drug discovery approaches has been a tedious and time-consuming task that allows only superficial mining of the huge chemical diversity of natural products and that usually leads to the identification of only a few bioactive peptides per experiment (77).

During the past decade, several studies focusing solely on cone snail venom duct (43, 44, 49, 78–80) or salivary gland (9, 43, 44, 49, 78–80) transcriptomes, and later complemented by proteome profiling (46, 47, 50, 59, 81), have allowed the report of no more than only a hundred (47 on average) full-length precursor conotoxins each. The great majority of these studies used the ROCHE 454 next-generation sequencing platform because it produced low amounts of long reads that were possible to annotate by performing simple homology BLAST searches. However, the sequences produced were often analyzed without applying quality filtering first (or using low thresholds) although single-base call and homopolymer-associated errors are frequent with this platform (82). Moreover, the weak accuracy of global BLAST searches to identify and classify conotoxin transcripts, compared with purpose-built algorithms, favors the discovery only of toxins closely related to known ones and is not suitable for large datasets containing numerous sequence isoforms.

In this article, we used state-of-the-art Illumina 2 x 300 paired-end chemistry and LC/MS/MS proteomic sequencing integrated in a dedicated bioinformatics pipeline that allowed capturing, to our knowledge, the first high-definition snapshot of the toxin arsenal isolated from a single venom apparatus and supported by accurate annotations. We were able to (i) identify 3,303 novel full-length conotoxin precursors belonging to 9 empirical and 16 new gene families, (ii) identify 212 conotoxins containing the pharmacologically active ICK and CC-C-C motifs; (iii) identify six novel cysteine frameworks anticipated to support novel pharmacology; and (iv) highlight the specific conservation of codons encoding the cysteine skeleton of the mature conotoxins.

The high rate of nucleotide substitutions and insertions observed in the intercysteine loops of the mature toxin region, amplified by potential RNA-editing processes, could explain the extensive number of conotoxin isoforms. Indeed, nucleoside modifications such as cytidine (C) to uridine (U) or adenosine (A) to inosine (I) deaminations have been observed in both eukaryotic and prokaryotic tRNAs, rRNAs, microRNAs, and mRNAs (83, 84). Although posttranscriptional editing of mRNAs is far less common than other RNA-processing events, such as alternative splicing, 5′-capping, or 3′-polyadenylation, it could be a source of preference for certain codons to be translated more...
accurately or efficiently, leading to sequence variability and variations of protein expression levels (85). Sequencing the Conus genome might help address these questions and shed light on the organization, expression, and regulation mechanisms of gene-encoding conotoxins.

This exceptional sequence diversity, coupled here with the discovery of new conotoxins with cysteine patterns encountered in other organisms (such as integrin receptor antagonist snake C-type lectins, which have provided lead structures for the design of antimetastatic and antiangiogenic drugs) (86), confirms the extraordinary potential of small Conus peptides to unveil novel pharmacology. Moreover, improvement of de novo assembly programs dedicated to the treatment of datasets with numerous conserved sequences and repeats would open the way to the identification of new classes of longer polypeptides with original modes of action (81). However, de novo transcriptome assembly still remains a challenging task (87–91) requiring dedicated high-depth sequencing strategies and extensive optimization steps (92) especially when performed in nonmodel organisms expressing highly similar transcripts. Indeed, the difference of sequencing depth between Illumina and current mass spectrometers necessitates enriching protein samples to detect low expressed proteins. In addition, bottom-up proteomic technologies can sequence only short fragments of proteins, leading to an enrichment of identical peptides when originating from similar protein isoforms, thus making difficult their precise assignment to their corresponding parent precursor transcripts. This limitation could be alleviated using top-down mass spectrometry where intact protein ions are introduced into a gas phase to be further fragmented and analyzed (94). Although this sequencing approach provides high sequence coverage and usually retains labile PTMs (95), the limited compatibility of the dissociation techniques used (electron capture dissociation or electron transfer dissociation for instance) with front-end separation methods and the difficulty to interpret complex fragmentation spectra generated by large multiply charged precursors limit this application to isolated proteins or simple mixtures (96, 97). We can also mention that a minor fraction of mature conotoxins lacking Arg and/or Lys [345 (10.44%) of the toxins reported here] or showing disadvantageous placement of these amino acids [64 (1.94%)] (98) would not be observable at protein level when using shotgun sequencing. Moreover, peptides not (or too strongly) retained on the LC column,
as well as large peptide fragments containing amino acids that weakly protonate (99) and are able to generate multiply charged ions with m/z value above the mass spectrometer selection threshold, could not be considered for database searching.

Finally, it is noteworthy that the methodology described in this report can be applied to potentially any type of tissue or organisms. The high sensitivity of the sequencing platforms clearly demonstrates the possibility of working with small amounts of starting material, which makes this approach suitable for studying rare samples. Also, the type of sequences to analyze is not restrictive. ConoSorter, the annotation program used here, can be easily modified to study other organisms by incorporating specific search models built from training sets of protein sequences that share conserved or unique primary structure signatures. Thus, this data-mining strategy offers a personalized tool for studying large sets of exome expression products that can be used for fundamental research purposes or applications such as diagnostic or drug discovery.

Materials and Methods

Collection of the Conus specimen, as well as the dissection of its venom duct, radular sac, and salivary glands are described in SI Materials and Methods. mRNA isolation from these compartments followed by the preparation and sequencing of the cDNA libraries with Illumina MiSeq sequencer are described in SI Materials and Methods. The bioinformatic processing of the transcriptome sequencing reads and their de novo assemblies allowing the discovery of new conotoxin sequences, cysteine frameworks, and gene superfamilies, as well as the analysis of codon usage and RNA editing are described in SI Materials and Methods. Finally, protein extraction and fractionation by PAGE and HPLC, followed by MS sequencing showing the existence of conotoxin transcripts at protein level, are detailed in SI Materials and Methods.

ACKNOWLEDGMENTS. We thank Professor Sean M. Grimmel for allowing access to the transcriptome sequencing platform. V.L. acknowledges the provision of a visiting fellowship within the framework of the Vietnam–Australia 2012–2013 joint venture in marine biology, which was supported by the Vietnam National Foundation for Science and Technology Development and the Vietnam Academy of Science and Technology. We thank National Health and Medical Research Council Project Grant 569927.


