Optimized deep-targeted proteotranscriptomic profiling reveals unexplored *Conus* toxin diversity and novel cysteine frameworks

Vincent Lavergne[a], Ivon Harliwong[b], Alun Jones[a], David Miller[b], Ryan J. Taft[b], and Paul F. Alewood[a,1]

[a]Division of Chemistry and Structural Biology, Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia; and [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 episcopatus* 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 hypervariablity in mature toxin regions.

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

Cone snails are venomous marine gastropod 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 chemoreceptory organ (called a siphon or osphradium), 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. 1*C*). 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. 1*B*) (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

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 episcopatus*. 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, PRJDB3896, SAMD00029744, DRX030964, DRR034331, SAMD00029745, DRX030965, DRR034332, SAMD00029746, DRX030966, and DRR034333).

[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.

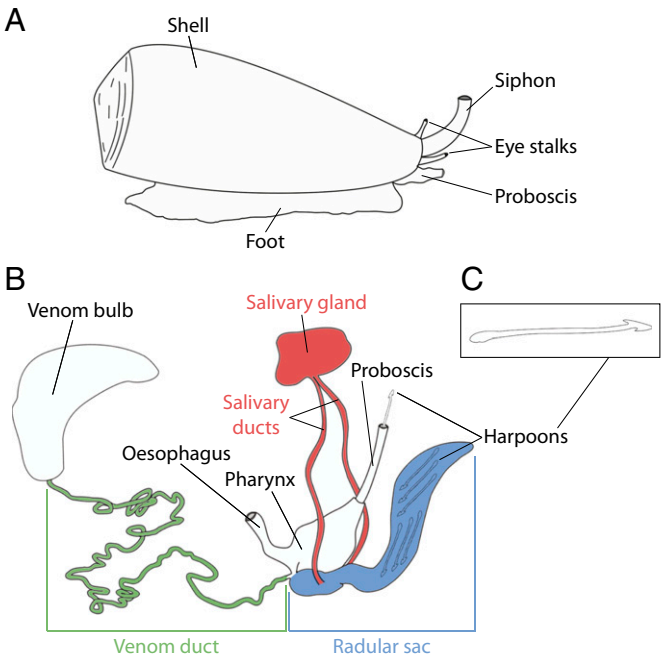

Fig. 1. Macroscopic anatomy of a cone snail (*A*), its venom apparatus (*B*), and a radula tooth (*C*).

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 neuropathic 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 χ-conotoxin MrIA [(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, I3, 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 episcopatus*. 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. episcopatus* 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 1. Cysteine frameworks of mature conotoxins

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

The name, pattern, and connectivity of cysteine frameworks ("—" when unknown) are reported.
*GenBank accession no. HM003926.

Table 2. Description of the raw, merged, and assembled reads from the venom duct, radular sac, and salivary gland

	Sample								
	Read R1	Read R2	Merged R1/R2	Unmerged R1	Unmerged R2	CLC contigs	SOAP contigs	Oases contigs	Trinity contigs
Venom duct									
Total Sequences	20,890,920	20,880,539	17,659,352	2,864,734	511,428	132,719	46,926	30,354	114,771
Length Interval	35–301	35–301	35–592	25–301	37–301	100–7,392	102–5,385	101–29,853	101–5,747
Avg Length	209	211	215	256	281	341	408	761	370
N50	244	251	230	289	301	418	424	1,038	478
N90	301	301	438	301	301	1,379	851	3,060	1,466
%GC	37.44	38.03	37.60	37.68	36.99	38.89	38.41	38.63	38.96
%N	0	0	0	0	0	0	0	0.53	0
Reads to contigs mapping	—	—	—	—	—	90.83%	76.28%	69.61%	87.35%
Known toxins	—	—	6	6	2	1	2	2	1
New toxins (score 3)	—	—	2,061	2,629	1,117	28	20	16	9
New toxins (score 2)	—	—	822	1,323	158	6	1	3	2
Radular sac									
Total sequences	29,442,459	28,932,379	22,782,581	6,166,372	1,324,521	138,284	150,900	269,328	129,509
Length interval	35–301	35–301	35–592	35–301	35–301	100–5,386	100–5,385	100–21,487	101–5,386
Avg length	206	209	204	262	288	258	225	504	267
N50	248	293	215	286	301	301	246	830	315
N90	301	301	441	301	301	979	680	2,091	1,046
%GC	32.90	33.47	32.60	33.11	33.95	36.67	36.69	34.27	36.92
%N	0	0	0	0	0	0	0	0	0
Reads to contigs mapping	—	—	—	—	—	91.93%	89.88%	45.47%	96.48%
Known toxins	—	—	1	1	0	1	1	1	1
New toxins (score 3)	—	—	10	13	6	8	3	7	3
New toxins (score 2)	—	—	1	1	0	2	2	2	0
Salivary gland									
Total sequences	32,701,009	30,750,697	20,005,743	2,454,298	2,174,305	147,817	166,152	276,069	141,353
Length interval	35–301	35–301	35–592	51–301	45–301	100–5,386	100–3,646	100–28,906	101–6,374
Avg length	200	205	164	299	299	249	208	472	260
N50	239	300	165	300	301	305	215	755	325
N90	301	301	358	301	301	928	668	2,044	1,071
%GC	33.08	34.58	32.84	33.33	34.21	36.30	36.40	34.79	36.50
%N	0	0	0	0	0	0	0	0	0
Reads to contigs mapping	—	—	—	—	—	90.70%	87.84%	40.61%	95.71%
Known toxins	—	—	1	1	0	1	0	1	0
New toxins (score 3)	—	—	0	0	0	3	2	3	4
New toxins (score 2)	—	—	0	0	0	0	0	0	0

The length interval, average length, N50 and N90 values, and %GC, as well as %N, of the different sets of sequences, are reported in the table. The overall rates of merged reads aligned back to the contigs have been calculated after using four different de novo assemblers for each compartment of the venom apparatus. The number of sequences annotated as known and new conotoxins (≥40 amino acids long; containing ≥60% hydrophobic residues in their N-ter region) and classified into *Conus* gene superfamilies [based on conopeptide sequence signatures in both signal/pro/mature regions (score 3) or pro/mature regions only (score 2)] are also listed (calculations not performed on read datasets are represented by "—").

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. episcopatus*, 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. episcopatus* [EpI, 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 *Conus pennaceus* and *Conus textile*, respectively (66). We were also able to detect 132 novel precursor forms of known mature toxin regions. Indeed, 84 new precursor sequences from *Conus magnificus* μO-MfVIA (67), 26 from *C. episcopatus* 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* PnMRCL-012 (66) mature conotoxin regions have been deciphered (Fig. S1 and Dataset S1*A*).

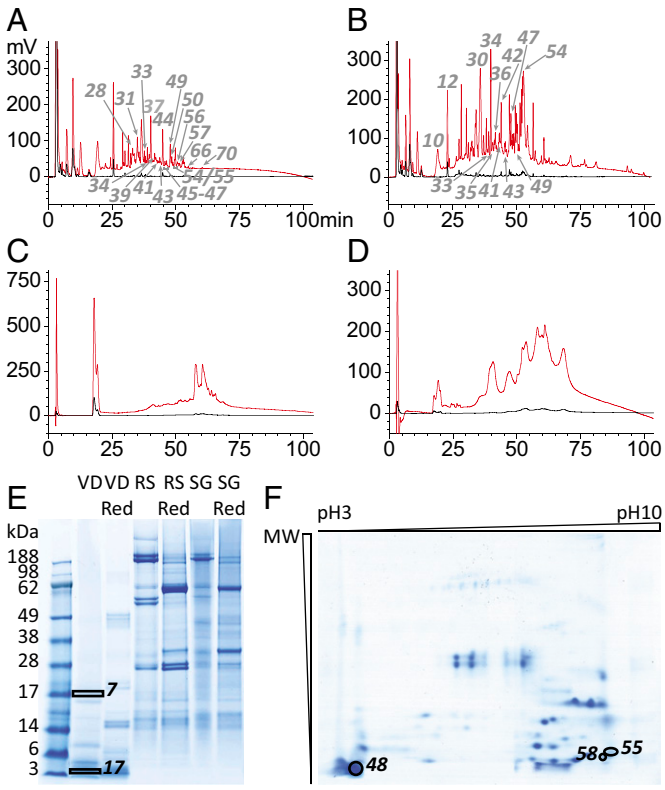

Fig. 2. Fractionation methods used to purify protein samples. Reverse-phase HPLC traces (214 nm) of VD [<30 kDa (*A*); >30 kDa (*B*)], as well as RS (*C*) and SG (*D*) protein extracts are shown. Raw and reduced protein extracts have also been separated by 1D-PAGE and revealed with Coomassie blue (*E*). Total VD proteins have been separated by 2D-PAGE and revealed with Coomassie blue (*F*). The fraction (HPLC) and spot (gels) numbers that refer to MS fragments (Dataset S1) are also mentioned.

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 new 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 S1*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 S1*C*), 9 cone snail cysteine frameworks were represented (Fig. 4*A*): 44 mature toxins with framework I (CC-C-C), 834 with framework V (CC-CC), 3 with framework XIV (C-C-C-C), 6 with framework III (CC-C-C-CC), 168 with framework VI/VII (C-C-CC-C-C), 6 with framework IX (C-C-C-C-C-C), 77 with framework XI (C-C-CC-CC-C-C), 15 with framework XXII (C-C-C-C-C-C-C-C-C), and 66 with framework VIII (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. 4*B*, Fig. S3, and Dataset S1*C*). 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_v 1.8 voltage-gated sodium channels (67) (97% identity with Ep298, Ep299, Ep301, Ep311, Ep323, Ep325, and Ep615); *C. pennaceus* PnVIB ω-conotoxin, which is able to block dihydropyridine-insensitive high voltage-activated calcium channels (68) (97% identity with Ep584, Ep587, and Ep589); and

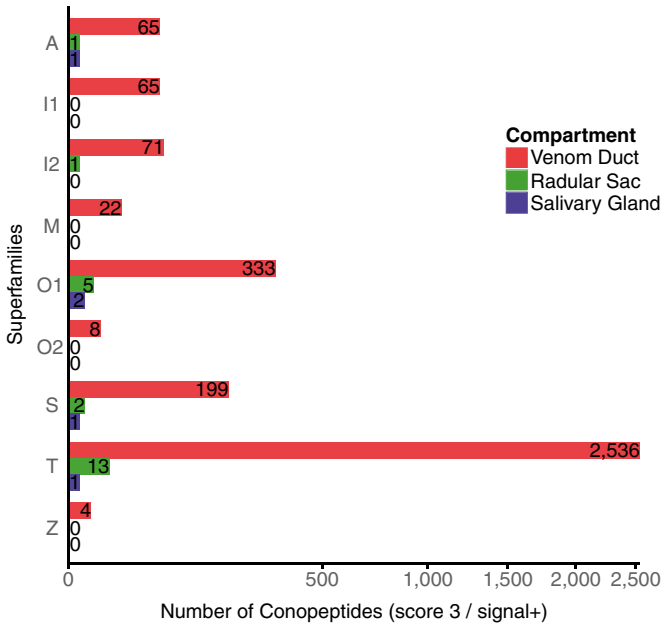

Fig. 3. Number of new precursor conopeptides per gene superfamily and compartment. These toxin precursors all contain a signal peptide and have been classified with ConoSorter at the highest score (matching *Conus* signal, pro, and mature signatures without conflicts).

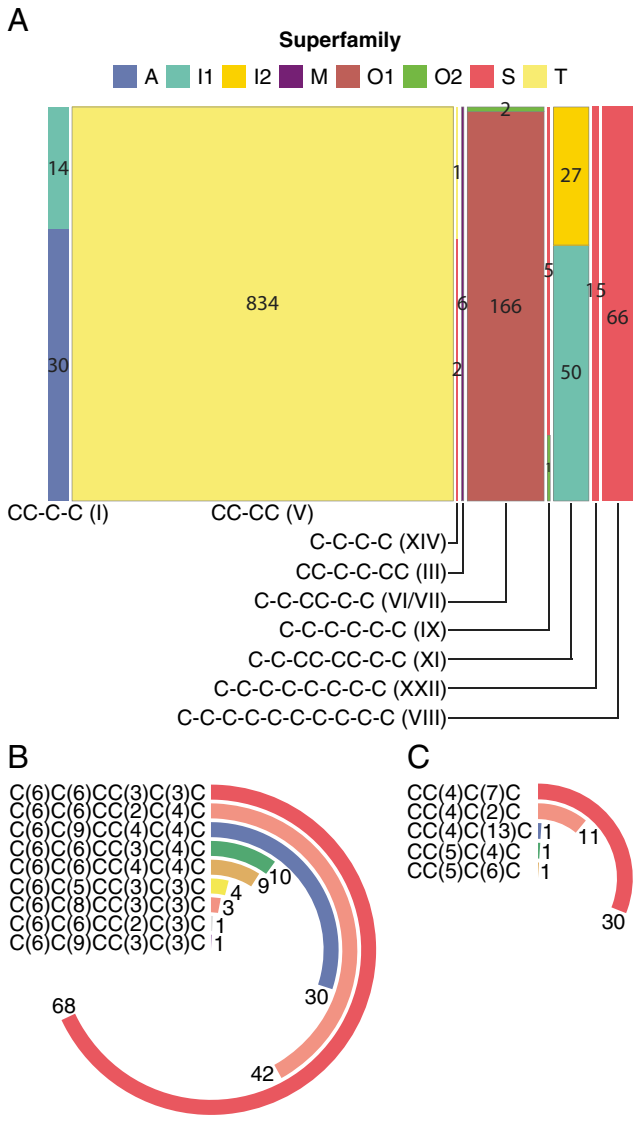

Fig. 4. Distribution of mature toxins per superfamily and known cysteine pattern (*A*). (*Bottom*) The number of mature toxins containing cysteine framework VI/VII (C-C-CC-C-C) and I (CC-C-C) classified per loop formula in (*B*) and (*C*), respectively.

Conus amadis Am2766 δ-conotoxin, targeting the rNa$_v$ 1.2a sodium channel (69) (59% with Ep386).

We also investigated mature conotoxins containing the cysteine framework I, which is found in the norepinephrine transporter inhibitor *Conus marmoreus* χ-conotoxin MrIA [(3)CC(4)C(2)C (phase IIb clinical trial)] (70). Among them, 30 (68.18%) and 14 (31.81%) belong to the A and T superfamilies, respectively [309 mature conotoxins with framework I are known to date, the most populated groups being the A (54.05%), M (3.24%), and T (1.29%) superfamilies]. The general loop formula for these toxins is (1–24)CC(4–5)C(2–13)C(0–35) (Fig. 4*C*, Fig. S4, and Dataset S1*C*). We also observed that the highest similarity rate (73%) is shared between the sequence Ep3055 and the patented TxId conotoxin isolated from *C. textile* (71).

In addition, we found 6 new cone snail cysteine motifs in 21 lowly expressed mature conotoxins [7 (33.33%) have been identified by MS) (Table 3 and Dataset S1*D*). After scanning the UniProtKB/Swiss-Prot database (541,954 entries), we observed that 4 of these patterns (CC-C-CC-C, CC-CC-C-C, CC-CC-CC, and CC-CC-C-C-C-C) are present in 558 mature proteins isolated from 292 different species. A fifth motif (CC-C-C-C-C-C-C) has been found in 19 snake heteromeric C-type lectins. Intriguingly, we also identified a novel 10-cysteine pattern that has never been observed in any organism: CC-CC-CC-CC-C-C.

Although protein sequences containing an odd number of cysteines are usually considered as forming only dimeric 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-acylation 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-CC-C-C; CC-CCC; C-C-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* μO§-conotoxin GVIIJ, 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 signatures 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. 5*A*).

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. 5*B*). 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. 5*B*). 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

Cysteine pattern	Name	Precursor sequence	Frequency	Superfamily
Six cysteines				
CC-C-CC-C	**Arylsulfatase A (component C) (*H. sapiens* - P15289) (I–V, II–VI, III–IV)**	*MGAPRSLLLALAAGLAVA*RPPNIVLIFADDLGYGDL-GCYGHPSSTTPNLDQLAAGGLRFTDFYVPVSLCT-PSRAALLTGRLPVRMGMYPGVLVPSSRGGLPLEE-VTVAEVLAARGYLTGMAGKWHLGVGPEGAFLP-PHQGFHRFLGIPYSHDQGPCQNLTCFPPATPCDG-GCDQGLVPIPLLANLSVEAQPPWLPGLEARYMA-FAHDLMADAQRQDRPFFLYYASHHTHYPQFSG-QSFAERSGRGPFGDSLMELDAAVGTLMTAIGDL-GLLEETLVIFTADNGPETMRMSRGGCSGLLRCGK-GTTYEGGVREPALAFWPGHIAPGVTHELASSLD-LLPTLAALAGAPLPNVTLDGFDLSPLLLGTGKSPR-QSLFFYPSYPDEVRGVFAVRTGKYKAHFFTQGS-AHSDTTADPACHASSSLTAHEPPLLYDLSKDPG-ENYNLLGGVAG**ATPEVLQALKQLQLLKAQLDA-AVTFGPSQVARGEDPALQICCHPGCTPRPACC-HCPDPHA**	—	—
	Ep214	*MMSKLGVLLTICLLLFSLTA*VPLDGDQHADQPAER-LQGDILSEKHPLFNPVKR**CCPAAACAMGCC-PFICGTV**	1	M
CC-CC-C-C	**Heat-stable enterotoxin ST-IA/ST-P (*E. coli* - P01559) (I–IV, II–V, III–VI)**	*MKKLMLAIFISVLSFPSFSQ*STESLDSSKEKITLETKKC-DVVKNNSEKKSENMN**NTFYCCELCCNPACAGCY**	—	—
	Ep1802	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQASFQD-NAKRILQVLKSKR**NCCRLQVCCGLQAAVSLSFHL-WNCMIKQLKCHRNFSVDKHYDHVASNYIIWTF**	1	T
	Ep2291	*MRCLPVFVILLLLIASTPNVDA***RPKTKDDMPLASFH-DDAKRILQILQDRNGCCIAGDCCGGSEIKENEFG-CKPCKLSLDVKFGKQTVPFARV**RRISNGR	1	T
	Ep1646	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQASFQD-NAKRILQVLESKR**NCCRLQVCCGFHLWNCMIKQ-LKCHRNFSVDKHYDHVASNYIIWTF**	1	T
	Ep2642	*MRCLPVFVILLLLIASTPSVDA*LLKTKDDMPLASFRD-DVKRT**LQTLLNKRFCCPYFECCKLLDERLKTCICV-WLYTGIPDNRKTGDPFQT**	1	T
	Ep2653	*MRCLPVFVILLLLIASTPSVDA*LLKTKDDMPLASFRD-DVKRT**LQTLLNKRFCCPYFECCVVGGDQLCYRG-LIKCIMN**K	1	T
	Ep2036	*MRCLPVFVILLLLIASAPS*VDVRPKAK**DDMPLASFH-DNPLQIRLVDTSCCPSQPCCRFGYREMTLDETPT-KCPCMYT**	1	T
	Ep1629	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQASFQ-DNAKRILQVLESKR**NCCRLQVCCGCFEVKENV-RTDFC**	1	T
	Ep1609	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQASFQ-DNAKRILQVLESKR**NCCRLQVCCGCFEIKENV-HADC**G	1	T
	Ep1092	*MRCLPVFIVLLLLIASAPCLDA*LPKTEGDVPLSSFHD-NLKRT**RRTHLNIRECCPDGWCCPAGCPTEKVQLCS**	1	T
	Ep1100	*MRCLPVFIVLLLLIASAPCLDA*LPKTEGDVPLSSFHD-NLKRT**RRTHLNIRECCSDGRCCPAGCSTENVHLCP**	1	T
	Ep1109	*MRCLPVFIVLLLLIASAPCLDA*LPKTEGDVPLSSFHD-NLKRT**RRTHLNIRECCSDGWCCPAGCLTENVHLCP**	1	T
	Ep1111	*MRCLPVFIVLLLLIASAPCLDA*LPKTEGDVPLSSFHD-NLKRT**RRTHLNIRECCSDGWCCPAGCSTENEHLCP**	1	T
	Ep1112	*MRCLPVFIVLLLLIASAPCLDA*LPKTEGDVPLSSFHD-NLKRT**RRTHLNIRECCSDGWCCPAGCSTENVHLCP**	2	T
	Ep1110	*MRCLPVFIVLLLLIASAPCLDA*LPKTEGDVPLSSFHD-NLKRT**RRTHLNIRECCSDGWCCPAGCSTEHVHVCP**	1	T
CC-CC-CC	**Protein YqcK (*B. subtilis* - P45945)**	MKYVHVGVNVVSLEKSINFYEKVFGVKAVKVKTD-YAKFLLETPGLNFTLNVADEVKGNQVNHFGFQV-DSLEEVLKHKKRLEKEGFFAREEMDTTCCYAVQ-DKFWITDPDGNEWEFFYTKSNSEVQKQDSSSC-CVTPPSDITTNSCC	—	—

Table 3. Cont.

Cysteine pattern	Name	Precursor sequence	Frequency	Superfamily
	Ep1587	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQASFQ-DNAKRILQVLESKR**NCCRLQALASFHDNPLQI-RLVDTRCCPSQPCCRF**G	1	T
	Ep2695	*MRCLPVFVILLLLIASTPSVDA*LLKTKDDMPLASFR-DDVKR**TLQTLLNKRFCCQYFDAKRALQTLMD-IRECCMGTPGCCPW**G	1	T
Eight cysteines				
CC-C-C-C-C-C-C	**Snaclec 4 (C-type lectin-like 4) (*D. siamensis* - Q4PRC9) (II–III, IV–VIII, V–inter, VI–VII)**	*MGRFISISFGLLVVFLSLSGTEA***AFCCPSGWSAYD-QNCYKVFTEEMNWADAEKFCTEQKKGSHLV-SLHSREEEKFVVNLISENLEYPATWIGLGNMW-KDCRMEWSDRGNVKYKALAEESYCLIMITHE-KVWKSMTCNFIAPVVCKF**	—	—
	Ep1738	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQASFQ-DNAKRILQVLESKR**NCCRLWLRPLQTVPGCEIW-KADCSFRTCSWNFEWSLTTRCHLQATISLSFHL-WNCMIKQLKCHRHY**	1	T
	Ep2668	*MRCLPVFVILLLLIASTPSVDA*LLKTKDDMPLASFR-DDVKR**TLQTLLNKRFCCPYFECWKADCSFRTC-SWNFEWSLTTRCHLQATISLSFHLWNCMIKQ-LKCHRHY**	1	T
CC-CC-C-C-C-C	**SPRY domain-containing protein 7 (*H. sapiens* - Q5W111)**	MATSVLCCLRCCRDGGTGHIPLKEMPAVQLDTQ-HMGTDVVIVKNGRRICGTGGCLASAPLHQNK-SYFEFKIQSTGIWGIGVATQKVNLNQIPLGRD-MHSLVMRNDGALYHNNEEKNRLPANSLPQE-GDVVGITYDHVELNVYLNGKNMHCPASGIR-GTVYPVVYVDDSAILDCQFSEFYHTPPPGFEK-ILFEQQIF	—	—
	Ep1702	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQA-SFQDNAKRILQVLESKR**NCCRLQVCCGSTQ-NWYGPGESDCLIKTKHCDGHHSVLTQCDFCPVL**	1	T
Ten cysteines				
CC-CC-CC-CC-C-C	Ep1647	*MRCLPVFVILLLLIASAPSVDA*RPKTKDDIPQAS-FQDNAKRILQVLESKR**NCCRLQVCCGFQGN-RLCCVSLPTHTQTVHFCCILNTTCFTTCDSSQ**	1	T

Cysteine patterns found in proteins from non-*Conus* organisms (name, species, UniProtKB accession number and cysteine connectivity in bold) are illustrated with a representative sequence (signal peptide in italic; mature region in bold; data not applicable to the reference proteins are represented by "—").

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 × 300 paired-end chemistry and LC-MS/MS protein 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 superfamilies, as well as displaying 9 *Conus* cysteine frameworks; (*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

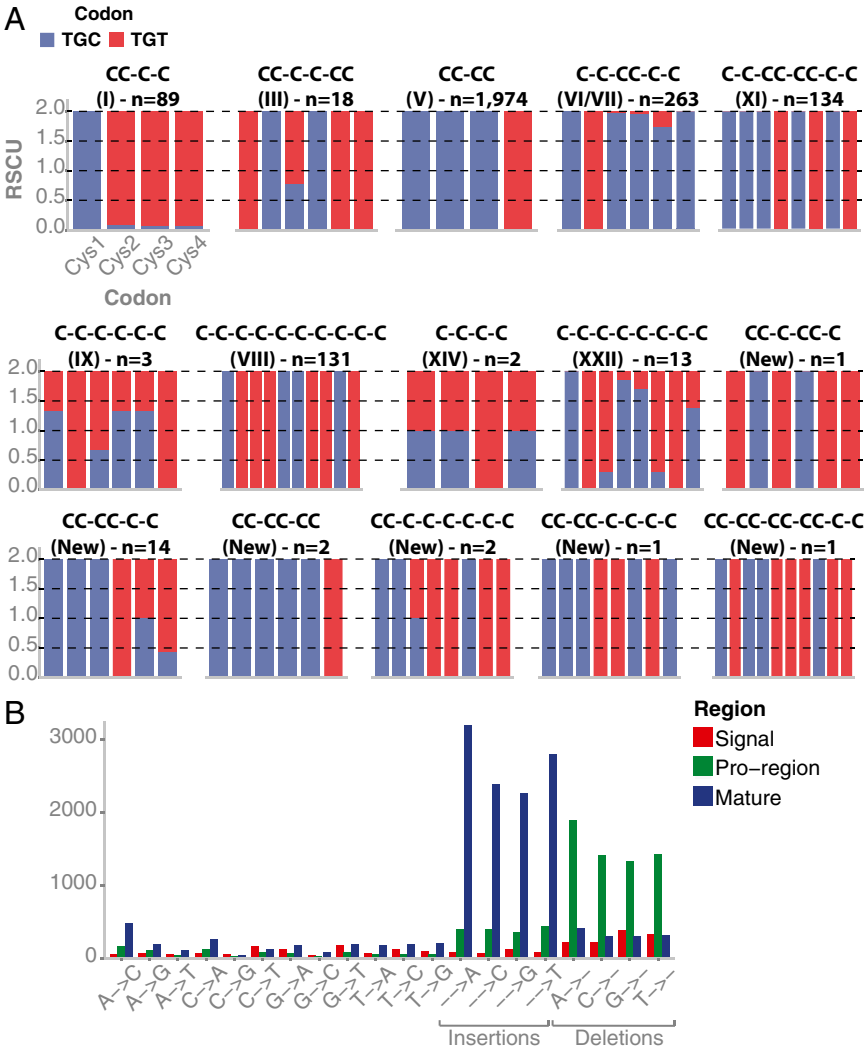

Fig. 5. Relative synonymous codon usage (RSCU) of cone snail cysteine frameworks (*A*) and profiles of point nucleotide substitutions or indels in the signal (red), proregion (green), and mature (blue) conotoxin regions (*B*).

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, most modern de novo assemblers based on de Bruijn graphs (93) still lack the efficiency to treat repetitive sequence regions, often leading to the abortion of the graph resolution after a few cycles and production of chimera sequences.

This chemical diversity could also be expanded with better identification of toxin posttranslational modifications (PTMs) and the amelioration of transcriptome/proteome mapping. Although the pipeline described here has allowed matching more peptide fragments with conotoxin transcripts than the only two comparable *Conus* studies [144 peptides mapped to 3,303 full-length precursor transcripts (4.4%); Dutertre et al. (59), 43 vs. 75–57%; Jin et al. (46), 29 vs. 48–60%], the task still remains delicate. 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 modulated 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. Grimmond for allowing access to the transcriptome sequencing platform. V.L. acknowledges the provision of an Institute for Molecular Bioscience Postgraduate Award and support from National Health and Medical Research Council Program Grant 569927.

1. Bouchet P, Gofas S (2014) *Conus* Linnaeus, 1758. World Register of Marine Species. Available at www.marinespecies.org. Accessed April 29, 2015.
2. Duda TF, Jr, Kohn AJ (2005) Species-level phylogeography and evolutionary history of the hyperdiverse marine gastropod genus Conus. *Mol Phylogenet Evol* 34(2): 257–272.
3. Freeman SE, Turner RJ, Silva SR (1974) The venom and venom apparatus of the marine gastropod Conus striatus Linne. *Toxicon* 12(6):587–592.
4. Kohn AJ (1956) Piscivorous gastropods of the genus Conus. *Proc Natl Acad Sci USA* 42(3):168–171.
5. Spengler HA, Kohn AJ (1995) Comparative external morphology of the Conus osphradium (Mollusca: Gastropoda). *J Zool* 235(3):439–453.
6. Schulz JR, Norton AG, Gilly WF (2004) The projectile tooth of a fish-hunting cone snail: Conus catus injects venom into fish prey using a high-speed ballistic mechanism. *Biol Bull* 207(2):77–79.
7. Marshall J, et al. (2002) Anatomical correlates of venom production in Conus californicus. *Biol Bull* 203(1):27–41.
8. Safavi-Hemami H, Young ND, Williamson NA, Purcell AW (2010) Proteomic interrogation of venom delivery in marine cone snails: Novel insights into the role of the venom bulb. *J Proteome Res* 9(11):5610–5619.
9. Biggs JS, Olivera BM, Kantor YI (2008) Alpha-conopeptides specifically expressed in the salivary gland of Conus pulicarius. *Toxicon* 52(1):101–105.
10. Lewis RJ, Dutertre S, Vetter I, Christie MJ (2012) Conus venom peptide pharmacology. *Pharmacol Rev* 64(2):259–298.
11. Terlau H, Olivera BM (2004) Conus venoms: A rich source of novel ion channel-targeted peptides. *Physiol Rev* 84(1):41–68.
12. Craig AG, Bandyopadhyay P, Olivera BM (1999) Post-translationally modified neuropeptides from Conus venoms. *Eur J Biochem* 264(2):271–275.
13. Olivera BM (2006) Conus peptides: Biodiversity-based discovery and exogenomics. *J Biol Chem* 281(42):31173–31177.
14. Espiritu DJ, et al. (2001) Venomous cone snails: Molecular phylogeny and the generation of toxin diversity. *Toxicon* 39(12):1899–1916.
15. Kaas Q, Yu R, Jin AH, Dutertre S, Craik DJ (2012) ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res* 40(Database issue):D325–D330.
16. Schroeder CI, Craik DJ (2012) Therapeutic potential of conopeptides. *Future Med Chem* 4(10):1243–1255.
17. Gray WR, Luque A, Olivera BM, Barrett J, Cruz LJ (1981) Peptide toxins from Conus geographus venom. *J Biol Chem* 256(10):4734–4740.
18. Ramilo CA, et al. (1992) Novel alpha- and omega-conotoxins from Conus striatus venom. *Biochemistry* 31(41):9919–9926.
19. Sato S, Nakamura H, Ohizumi Y, Kobayashi J, Hirata Y (1983) The amino acid sequences of homologous hydroxyproline-containing myotoxins from the marine snail Conus geographus venom. *FEBS Lett* 155(2):277–280.
20. Fainzilber M, et al. (1995) A new cysteine framework in sodium channel blocking conotoxins. *Biochemistry* 34(27):8649–8656.
21. Walker CS, et al. (1999) The T-superfamily of conotoxins. *J Biol Chem* 274(43):30664–30671.
22. Olivera BM, McIntosh JM, Cruz LJ, Luque FA, Gray WR (1984) Purification and sequence of a presynaptic peptide toxin from Conus geographus venom. *Biochemistry* 23(22):5087–5090.
23. England LJ, et al. (1998) Inactivation of a serotonin-gated ion channel by a polypeptide toxin from marine snails. *Science* 281(5376):575–578.
24. Lirazan MB, et al. (2000) The spasmodic peptide defines a new conotoxin superfamily. *Biochemistry* 39(7):1583–1588.
25. Balaji RA, et al. (2000) Lambda-conotoxins, a new family of conotoxins with unique disulfide pattern and protein folding: Isolation and characterization from the venom of Conus marmoreus. *J Biol Chem* 275(50):39516–39522.
26. Jimenez EC, et al. (2003) Novel excitatory Conus peptides define a new conotoxin superfamily. *J Neurochem* 85(3):610–621.
27. Brown MA, et al. (2005) Precursors of novel Gla-containing conotoxins contain a carboxy-terminal recognition site that directs gamma-carboxylation. *Biochemistry* 44(25):9150–9159.
28. Aguilar MB, et al. (2005) A novel conotoxin from Conus delessertii with post-translationally modified lysine residues. *Biochemistry* 44(33):11130–11136.
29. Möller C, et al. (2005) A novel conotoxin framework with a helix-loop-helix (Cs alpha/alpha) fold. *Biochemistry* 44(49):15986–15996.
30. Peng C, Liu L, Shao X, Chi C, Wang C (2008) Identification of a novel class of conotoxins defined as V-conotoxins with a unique cysteine pattern and signal peptide sequence. *Peptides* 29(6):985–991.
31. Pi C, et al. (2006) Diversity and evolution of conotoxins based on gene expression profiling of Conus litteratus. *Genomics* 88(6):809–819.
32. Yuan DD, et al. (2008) Isolation and cloning of a conotoxin with a novel cysteine pattern from Conus caracteristicus. *Peptides* 29(9):1521–1525.
33. Chen JS, Fan CX, Hu KP, Wei KH, Zhong MN (1999) Studies on conotoxins of Conus betulinus. *J Nat Toxins* 8(3):341–349.
34. Chen P, Garrett JE, Watkins M, Olivera BM (2008) Purification and characterization of a novel excitatory peptide from Conus distans venom that defines a novel gene superfamily of conotoxins. *Toxicon* 52(1):139–145.
35. Loughnan ML, Nicke A, Lawrence N, Lewis RJ (2009) Novel alpha D-conopeptides and their precursors identified by cDNA cloning define the D-conotoxin superfamily. *Biochemistry* 48(17):3717–3729.
36. Möller C, Marí F (2011) 9.3 KDa components of the injected venom of Conus purpurascens define a new five-disulfide conotoxin framework. *Biopolymers* 96(2):158–165.
37. Elliger CA, et al. (2011) Diversity of conotoxin types from Conus californicus reflects a diversity of prey types and a novel evolutionary history. *Toxicon* 57(2):311–322.
38. Ye M, et al. (2012) A helical conotoxin from Conus imperialis has a novel cysteine framework and defines a new superfamily. *J Biol Chem* 287(18):14973–14983.
39. Luo S, et al. (2013) A novel inhibitor of $\alpha 9\alpha 10$ nicotinic acetylcholine receptors from Conus vexillum delineates a new conotoxin superfamily. *PLoS ONE* 8(1):e54648.
40. Aguilar MB, et al. (2013) A novel arrangement of Cys residues in a paralytic peptide of Conus cancellatus (jr. syn.: Conus austini), a worm-hunting snail from the Gulf of Mexico. *Peptides* 41:38–44.
41. Bernáldez J, et al. (2013) A Conus regularis conotoxin with a novel eight-cysteine framework inhibits CaV2.2 channels and displays an anti-nociceptive activity. *Mar Drugs* 11(4):1188–1202.
42. Liu Z, et al. (2012) Diversity and evolution of conotoxins in Conus virgo, Conus eburneus, Conus imperialis and Conus marmoreus from the South China Sea. *Toxicon* 60(6):982–989.
43. Lluisma AO, Milash BA, Moore B, Olivera BM, Bandyopadhyay PK (2012) Novel venom peptides from the cone snail Conus pulicarius discovered through next-generation sequencing of its venom duct transcriptome. *Mar Genomics* 5:43–51.
44. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2012) Elucidation of the molecular envenomation strategy of the cone snail Conus geographus through transcriptome sequencing of its venom duct. *BMC Genomics* 13:284.
45. Biggs JS, et al. (2010) Evolution of Conus peptide toxins: Analysis of Conus californicus Reeve, 1844. *Mol Phylogenet Evol* 56(1):1–12.
46. Jin AH, et al. (2013) Transcriptomic messiness in the venom duct of Conus miles contributes to conotoxin diversity. *Mol Cell Proteomics* 12(12):3824–3833.
47. Safavi-Hemami H, et al. (2014) Combined proteomic and transcriptomic interrogation of the venom gland of Conus geographus uncovers novel components and functional compartmentalization. *Mol Cell Proteomics* 13(4):938–953.
48. Zhou M, et al. (2013) Characterizing the evolution and functions of the M-superfamily conotoxins. *Toxicon* 76:150–159.
49. Terrat Y, et al. (2012) High-resolution picture of a venom gland transcriptome: Case study with the marine snail Conus consors. *Toxicon* 59(1):34–46.
50. Violette A, et al. (2012) Recruitment of glycosyl hydrolase proteins in a cone snail venomous arsenal: Further insights into biomolecular features of Conus venoms. *Mar Drugs* 10(2):258–280.
51. Walker CS, et al. (2009) A novel Conus snail polypeptide causes excitotoxicity by blocking desensitization of AMPA receptors. *Curr Biol* 19(11):900–908.
52. Lirazan M, Jimenez EC, Grey Craig A, Olivera BM, Cruz LJ (2002) Conophysin-R, a Conus radiatus venom peptide belonging to the neurophysin family. *Toxicon* 40(7): 901–908.

53. Wermeling DP (2005) Ziconotide, an intrathecally administered N-type calcium channel antagonist for the treatment of chronic pain. *Pharmacotherapy* 25(8):1084–1094.
54. Kolosov A, Aurini L, Williams ED, Cooke I, Goodchild CS (2011) Intravenous injection of leconotide, an omega conotoxin: Synergistic antihyperalgesic effects with morphine in a rat model of bone cancer pain. *Pain Med* 12(6):923–941.
55. Brust A, et al. (2009) chi-Conopeptide pharmacophore development: Toward a novel class of norepinephrine transporter inhibitor (Xen2174) for pain. *J Med Chem* 52(22): 6991–7002.
56. Kaas Q, Westermann JC, Craik DJ (2010) Conopeptide characterization and classifications: An analysis using ConoServer. *Toxicon* 55(8):1491–1509.
57. Lavergne V, et al. (2013) Systematic interrogation of the Conus marmoreus venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene superfamilies. *BMC Genomics* 14(1):708.
58. Aguilar MB, et al. (2013) Precursor De13.1 from Conus delessertii defines the novel G gene superfamily. *Peptides* 41:17–20.
59. Dutertre S, et al. (2013) Deep venomics reveals the mechanism for expanded peptide diversity in cone snail venom. *Mol Cell Proteomics* 12(2):312–329.
60. Puillandre N, Koua D, Favreau P, Olivera BM, Stöcklin R (2012) Molecular phylogeny, classification and evolution of conopeptides. *J Mol Evol* 74(5-6):297–309.
61. Ishikawa H (1977) Evolution of ribosomal RNA. *Comp Biochem Physiol B* 58(1):1–7.
62. Hernández AI, et al. (2009) Poly-(ADP-ribose) polymerase-1 is necessary for long-term facilitation in Aplysia. *J Neurosci* 29(30):9553–9562.
63. Fujiwara H, Ishikawa H (1986) Molecular mechanism of introduction of the hidden break into the 28S rRNA of insects: Implication based on structural studies. *Nucleic Acids Res* 14(16):6393–6401.
64. Zarlenga DS, Dame JB (1992) The identification and characterization of a break within the large subunit ribosomal RNA of Trichinella spiralis: Comparison of gap sequences within the genus. *Mol Biochem Parasitol* 51(2):281–289.
65. Buczek O, et al. (2005) Characterization of D-amino-acid-containing excitatory conotoxins and redefinition of the I-conotoxin superfamily. *FEBS J* 272(16):4178–4188.
66. Conticello SG, et al. (2001) Mechanisms for evolving hypervariability: The case of conopeptides. *Mol Biol Evol* 18(2):120–131.
67. Vetter I, et al. (2012) Isolation, characterization and total regioselective synthesis of the novel μO-conotoxin MfVIA from Conus magnificus that targets voltage-gated sodium channels. *Biochem Pharmacol* 84(4):540–548.
68. Kits KS, et al. (1996) Novel omega-conotoxins block dihydropyridine-insensitive high voltage-activated calcium channels in molluscan neurons. *J Neurochem* 67(5): 2155–2163.
69. Sudarslal S, et al. (2003) Sodium channel modulating activity in a delta-conotoxin from an Indian marine snail. *FEBS Lett* 553(1-2):209–212.
70. Lewis RJ (2012) Discovery and development of the χ-conopeptide class of analgesic peptides. *Toxicon* 59(4):524–528.
71. Bhatia S, et al. (2012) Constrained de novo sequencing of conotoxins. *J Proteome Res* 11(8):4191–4200.
72. McDonald LJ, Moss J (1994) Enzymatic and nonenzymatic ADP-ribosylation of cysteine. *Mol Cell Biochem* 138(1-2):221–226.
73. Tate EW, Kalesh KA, Lanyon-Hogg T, Storck EM, Thinon E (2015) Global profiling of protein lipidation using chemical proteomic technologies. *Curr Opin Chem Biol* 24:48–57.
74. Qin Y, Dey A, Daaka Y (2013) Protein s-nitrosylation measurement. *Methods Enzymol* 522:409–425.
75. Gajewiak J, et al. (2014) A disulfide tether stabilizes the block of sodium channels by the conotoxin μO§-GVIIJ. *Proc Natl Acad Sci USA* 111(7):2758–2763.
76. Fosgerau K, Hoffmann T (2015) Peptide therapeutics: Current status and future directions. *Drug Discov Today* 20(1):122–128.
77. Vetter I, et al. (2011) Venomics: A new paradigm for natural products-based drug discovery. *Amino Acids* 40(1):15–28.
78. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2011) Characterization of the Conus bullatus genome and its venom-duct transcriptome. *BMC Genomics* 12:60.
79. Remigio EA, Duda TF, Jr (2008) Evolution of ecological specialization and venom of a predatory marine gastropod. *Mol Ecol* 17(4):1156–1162.
80. Robinson SD, et al. (2014) Diversity of conotoxin gene superfamilies in the venomous snail, Conus victoriae. *PLoS ONE* 9(2):e87648.
81. Violette A, et al. (2012) Large-scale discovery of conopeptides and conoproteins in the injectable venom of a fish-hunting cone snail using a combined proteomic and transcriptomic approach. *J Proteomics* 75(17):5215–5225.
82. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* 7(2):e30087.
83. Brennicke A, Marchfelder A, Binder S (1999) RNA editing. *FEMS Microbiol Rev* 23(3): 297–316.
84. Su AA, Randau L (2011) A-to-I and C-to-U editing within transfer RNAs. *Biochemistry (Mosc)* 76(8):932–937.
85. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* 107(8):3645–3650.
86. Arlinghaus FT, Eble JA (2012) C-type lectin-like proteins from snake venoms. *Toxicon* 60(4):512–519.
87. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics* 12:317.
88. Góngora-Castillo E, Buell CR (2013) Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Nat Prod Rep* 30(4):490–500.
89. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nat Rev Genet* 12(10):671–682.
90. Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet* 13(1):36–46.
91. Vijay N, Poelstra JW, Künstner A, Wolf JB (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification: A comprehensive in silico assessment of RNA-seq experiments. *Mol Ecol* 22(3):620–634.
92. Zhao QY, et al. (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: A comparative study. *BMC Bioinformatics* 12(Suppl 14):S2.
93. Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29(11):987–991.
94. Chait BT (2006) Mass spectrometry: Bottom-up or top-down? *Science* 314(5796): 65–66.
95. Ueberheide BM, Fenyö D, Alewood PF, Chait BT (2009) Rapid sensitive analysis of cysteine rich peptide venom components. *Proc Natl Acad Sci USA* 106(17):6910–6915.
96. Wehr T (2006) Top-down versus bottom-up approaches in proteomics. *LCGC North America* 24(9):1004–1011.
97. Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* 11:49–79.
98. Keil B (1992) *Specificity of Proteolysis* (Springer, Berlin).
99. Harrison AG (1997) The gas-phase basicities and proton affinities of amino acids and peptides. *Mass Spectrom Rev* 16(4):201–217.
100. Andrews S (2011) *FastQC: A Quality Control Tool for High Throughput Sequence Data* (Babraham Bioinformatics, Cambridge, UK).
101. Lohse M, et al. (2012) RobiNA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res* 40(Web Server Issue):W622–W627.
102. Magoč T, Salzberg SL (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957–2963.
103. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics* 13:31.
104. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652.
105. Haas BJ, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8): 1494–1512.
106. Xie Y, et al. (2014) SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30(12):1660–1666.
107. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28(8):1086–1092.
108. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
109. Zamora-Bustillos R, Aguilar MB, Falcón A, Heimer de la Cotera EP (2009) Identification, by RT-PCR, of four novel T-1-superfamily conotoxins from the vermivorous snail Conus spurius from the Gulf of Mexico. *Peptides* 30(8):1396–1404.
110. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Methods* 8(10):785–786.
111. Duckert P, Brunak S, Blom N (2004) Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* 17(1):107–112.
112. Fan X, Nagle GT (1996) Molecular cloning of Aplysia neuronal cDNAs that encode carboxypeptidases related to mammalian prohormone processing enzymes. *DNA Cell Biol* 15(11):937–945.
113. Fan X, Spijker S, Akalal DB, Nagle GT (2000) Neuropeptide amidation: Cloning of a bifunctional alpha-amidating enzyme from Aplysia. *Brain Res Mol Brain Res* 82(1-2): 25–34.
114. Hale JE, Butler JP, Gelfanova V, You JS, Knierman MD (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(1):174–181.
115. McInerney JO (1998) GCUA: General codon usage analysis. *Bioinformatics* 14(4):372–373.