Marine Genomics 5 (2012) 43–51

Contents lists available at SciVerse ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Novel venom peptides from the cone snail *Conus pulicarius* discovered through next-generation sequencing of its venom duct transcriptome

Arturo O. Lluisma [a,b], Brett A. Milash [c], Barry Moore [d], Baldomero M. Olivera [a,*], Pradip K. Bandyopadhyay [a]

[a] Department of Biology, University of Utah, Salt Lake City, UT, USA
[b] Marine Science Institute, University of the Philippines, Quezon City, Philippines
[c] Bioinformatics Core Facility, University of Utah, Salt Lake City, UT, USA
[d] Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, USA

ARTICLE INFO

Article history:
Received 24 June 2011
Received in revised form 9 September 2011
Accepted 16 September 2011

Keywords:
Conotoxin
Conopeptide
Toxin
Transcriptome

ABSTRACT

The venom peptides (i.e., conotoxins or conopeptides) that species in the genus *Conus* collectively produce are remarkably diverse, estimated to be around 50,000 to 140,000, but the pace of discovery and characterization of these peptides have been rather slow. To date, only a minor fraction have been identified and studied. However, the advent of next-generation DNA sequencing technologies has opened up opportunities for expediting the exploration of this diversity.

The whole transcriptome of a venom duct from the vermivorous marine snail *C. pulicarius* was sequenced using the 454 sequencing platform. Analysis of the data set resulted in the identification of over eighty unique putative conopeptide sequences, the highest number discovered so far from a *Conus* venom duct transcriptome. More importantly, majority of the sequences were potentially novel, many with unexpected structural features, hinting at the vastness of the diversity of *Conus* venom peptides that remains to be explored. The sequences represented at least 14 major superfamilies/types (disulfide- and non-disulfide-rich), indicating the structural and functional diversity of conotoxins in the venom of *C. pulicarius*. In addition, the contryphans were surprisingly more diverse than what is currently known. Comparative analysis of the O-superfamily sequences also revealed insights into the complexity of the processes that drive the evolution and diversification of conotoxins.

1. Introduction

The venom of marine gastropods (members of the genus *Conus*, also known as cone snails) contains a mixture of diverse, small, highly-structured peptides commonly referred to as conotoxins or conopeptides which, when injected by the snail into its target (primarily prey, but could also be their predators and competitors), bind to specific molecular receptors in the envenomated target. This results in the disruption of specific physiological processes in the target and elicits physiological effects such as paralysis. It is estimated that each *Conus* species produces 100–200 different venom peptides, and that there is little or no overlap in the specific kinds of peptides that the different species produce (Olivera, 2002).

Determining the inventory of peptides in the venoms of cone snails is interesting both from a biological and biomedical/biotechnological perspective. Because each species has its own repertoire of peptides that reflect its ecological niche, identification and enumeration of the peptides in the venom may thus provide a "molecular readout" of each species' biotic interactions (Olivera, 2002 and references cited therein). This molecular-level information provides insights on various aspects of the species' biology, ecology, and evolution and facilitates studies on their "exogenome" (Olivera, 2006), including the evolution of the toxins and the molecular mechanisms that generate their diversity.

On the other hand, considering the enormous potential of conotoxins as lead compounds or drugs (Terlau and Olivera, 2004; Olivera and Teichert, 2007), an inventory of peptides in *Conus* venoms (i.e., the "venome") would facilitate a systematic investigation of venome components and of their pharmacological properties and thus would significantly facilitate the identification of drug leads if not development of biomedical applications. Indeed, a number of these peptides are currently in advanced stages of clinical trials while others have become established experimental tools in pharmacological research (Olivera, 2006).

Because the peptides are encoded by genes and are synthesized in a specialized toxin-producing tissue, the venom duct (see Olivera, 2002), the cloning and sequencing of clones from venom duct cDNA libraries (i.e., the transcriptome) have become one of the methods of choice in the discovery of novel venom peptides. Thus, hundreds of *Conus* peptides have been discovered using this "transcriptomics" approach (e.g.,

* Corresponding author at: Dept. of Biology, 257 South 1400 East, University of Utah, Salt Lake City, UT 84112, USA. Tel.: +1 801 581 8370.
E-mail address: olivera@biology.utah.edu (B.M. Olivera).

1874-7787/$ – see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.margen.2011.09.002

Conticello et al., 2001; Garrett et al., 2005; Holford et al., 2009; Peng et al., 2006, 2007; Pi et al., 2006a, b; Liu et al., 2009). The advent of the next generation sequencing technologies (Margulies et al., 2005; Schuster, 2008), however, provides a means for accelerating the transcriptomics-based approach. In particular, shotgun sequencing of transcriptomes can theoretically reveal a complete or near-complete inventory of conotoxin genes expressed in the venom duct.

In this study, we utilized the 454 next generation sequencing platform (Margulies et al., 2005) to carry out whole-transcriptome sequencing of the venom duct of the tropical vermivorous gastropod *Conus pulicarius*. The sequences were then analyzed to identify putative conotoxins in the venom of this species.

2. Materials and methods

2.1. mRNA extraction from C. pulicarius venom duct

The venom duct from *C. pulicarius* was kindly provided by Dr. Jason S. Biggs. The tissue was harvested and stored in RNAlater (Ambion, Austin, Tx) as described in Biggs et al. (2008). Total RNA was isolated using TRIzol Plus RNA purification system (Invitrogen, Carlsbad, CA) according to the manufacturer's recommendation.

2.2. cDNA synthesis and whole-transcriptome shotgun sequencing

cDNA was synthesized from total RNA using the SMART cDNA Library kit (Clontech) following the manufacturer's recommendations, except that, instead of the primers provided in the kit, the following primers were used: (a) modified CDSIII/3′ cDNA Synthesis Primer, 5′ TAG AGA CCG AGG CGG CCG ACA TGT TTT GTT TTT TTT TCT TTT TTT TTT VN-3′, and (b) modified CDSIII/3′ PCR Primer, 5′ TAG AGG CCG AGG CGG CCG ACA TGT TTT GTC TTT TGT TCT GTT TCT TTT VN-3′. The generated cDNA was sequenced using the GS-FLX instrument (Roche, IN, USA) following the manufacturer's instructions (supplied with the system/kits).

2.3. Sequence processing and analysis

The raw sequence reads were processed to remove primer sequences. The primer-trimmed sequences were then assembled on a small cluster of computers running Linux using the Forge-G assembler software (http://www.cebitec.uni-bielefeld.de/forge/wiki/ForgeG, version 20070801) and the LAM/MPI message passing library (http://www.lam-mpi.org/). Forge-G was chosen for its modest memory requirements and its demonstrated ability to assemble 454 sequence data.

The resulting sequences/contigs were then analyzed primarily through comparison with similar sequences in the Swissprot database. Searches for similar sequences were carried out using the BLAST (Basic Local Alignment Search Tool) software (Altschul et al., 1990). Standalone BLAST executables were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) and installed on local desktop computers. A reference database was constructed by adding selected non-redundant conotoxin sequences (downloaded from the Conoserver database http://research1t.imb.uq.edu.au/conoserver/) to a local copy of the UniProtKB/Swiss-Prot Database (release 15.4, downloaded from the UniProt web site, http://www.uniprot.org/downloads). This reference database was formatted using the formatdb software from the downloaded copy of the BLAST executables. Searches for similar sequences in the reference database were made using the blastx option of blastall which was run locally; the output files (in XML format) were processed using custom Python scripts to identify the contigs with hits to conotoxin sequences and to generate files that display the alignment of the sequences of the conotoxin hits with the full sequences of the contigs (translated from three reading frames). These contigs (including those with low scores) were then assigned into categories using the classification of the highest-scoring conotoxins that matched the contigs as a guide to facilitate sequence alignment and comparison. Where necessary, the full precursor sequences of the best-matching conotoxins (the reference sequences) were manually added to the alignment. The alignments were individually inspected to evaluate their quality; the sequence of the reference conotoxins was used as guide to detect frameshifts and to infer the correct translation of the sequences. Sequences that appear to be good conotoxin candidates on the basis of sequence similarity or structural characteristics (i.e., presence and arrangement of multiple Cys residues) were then subjected to multiple sequence alignment (in separate groups according to presumed conotoxin type) and based on this alignment unique peptide sequences which were either full-length or nearly full-length were identified and compiled into a non-redundant list.

To analyze the diversification of the O-superfamily sequences in *Conus*, all O-superfamily sequences (mature peptide region) in the Conoserver database were downloaded and, together with the *C. pulicarius* O-superfamily mature-region sequences generated in the study, were aligned using the software MUSCLE (Edgar, 2004). The resulting alignment was separated into clusters based on overall sequence similarity and length, and the sequence alignment in each cluster was then refined by eye. To generate a cladogram for the species represented in the O-superfamily dataset, 16S rRNA gene sequences for these species were downloaded from GenBank and aligned using the software ClustalW (Larkin et al., 2007). The cladogram was then constructed through Maximum Likelihood analysis as implemented in the software PhyML (Guindon and Gascuel, 2003). The following options were used: Subtree Pruning and Regrafting for the tree topology search algorithm and GTR + Γ + I (discrete gamma model with 4 categories) as the model of nucleotide substitution. Where the option is allowed, the other parameters were set to be optimized by the software.

3. Results and discussion

3.1. Identification of conotoxin sequences from the sequencing reads

Using the 454 Next-Generation DNA sequencing technology, sequencing of the *C. pulicarius* venom duct transcriptome library yielded 359,213 DNA reads and associated quality scores (minimum length: 36, median length: 228, maximum length: 393, total yield: 73,502,057 nucleotides). Primer trimming reduced this data set to 333,478 reads (minimum length: 30, median length: 186, maximum length: 393, total yield: 52,886,072 nucleotides). A total of 81,668 contigs were assembled from these sequence reads. The frequency distributions of the lengths and average read coverage (per bp) of the contigs are shown in Tables 1 and 2. Majority of the contigs (~98%) were less than 300 bp in length, and those that were longer than 500 bp comprised less than 1% of the total. Majority of the contigs (>99%) had relatively low average read coverage (<20), with those having an average read coverage of only 1 accounting for a major proportion (86%) of the total. Contigs with a relatively high average read coverage (>200) accounted for only a minuscule proportion (0.13%) of the total.

Of the 81,668 contigs, 1567 showed high similarity at the amino acid sequence level with conotoxin sequences in our reference database (construction of this reference database is described in the Materials and methods section). After evaluation of the scores and the quality of the match and comparison of the deduced peptide sequences, 82 unique putative conotoxin sequences were identified. Majority of these sequences were full-length but some were truncated at the N-terminus and a few at the C-terminus. A few were also identical with respect to the mature region but were considered unique owing to some divergence at the pre-pro region. A number of other sequences showed some sequence similarity with conotoxins but were either too short (hence could not be reliably identified as conotoxin sequences) or were duplicates of the selected representatives.