Toxicon 59 (2012) 34–46

Contents lists available at SciVerse ScienceDirect

Toxicon

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

High-resolution picture of a venom gland transcriptome: Case study with the marine snail *Conus consors*

Yves Terrat [a,*], Daniel Biass [c], Sébastien Dutertre [c], Philippe Favreau [c], Maido Remm [d], Reto Stöcklin [d], David Piquemal [b], Frédéric Ducancel [a]

[a] *Laboratoire d'Ingénierie des Anticorps pour la Santé (LIAS), CEA/DSV/iBiTec-s/SPI, CEA Saclay, Bt. 152, 91191 Gif-sur-Yvette Cedex, France*
[b] *Skuldtech, 134, rue du Curat, Bat. Amarante, 34090 Montpellier, France*
[c] *Atheris Laboratories, CH-1233 Bernex, Geneva, Switzerland*
[d] *Estonian Biocentre, Riia str. 23, 51010 Tartu, Estonia*

ARTICLE INFO

Article history:
Received 9 June 2011
Received in revised form 26 September 2011
Accepted 4 October 2011
Available online 6 November 2011

Keywords:
Toxin
Conopeptide
Cone snail
Conus consors
Pyrosequencing
Transcriptome

ABSTRACT

Although cone snail venoms have been intensively investigated in the past few decades, little is known about the whole conopeptide and protein content in venom ducts, especially at the transcriptomic level. If most of the previous studies focusing on a limited number of sequences have contributed to a better understanding of conopeptide superfamilies, they did not give access to a complete panorama of a whole venom duct. Additionally, rare transcripts were usually not identified due to sampling effect. This work presents the data and analysis of a large number of sequences obtained from high throughput 454 sequencing technology using venom ducts of *Conus consors*, an Indo-Pacific living piscivorous cone snail. A total of 213,561 Expressed Sequence Tags (ESTs) with an average read length of 218 base pairs (bp) have been obtained. These reads were assembled into 65,536 contiguous DNA sequences (contigs) then into 5039 clusters. The data revealed 11 conopeptide superfamilies representing a total of 53 new isoforms (full length or nearly full-length sequences). Considerable isoform diversity and major differences in transcription level could be noted between superfamilies. A, O and M superfamilies are the most diverse. The A family isoforms account for more than 70% of the conopeptide cocktail (considering all ESTs before clustering step). In addition to traditional superfamilies and families, minor transcripts including both cysteine free and cysteine-rich peptides could be detected, some of them figuring new clades of conopeptides. Finally, several sets of transcripts corresponding to proteins commonly recruited in venom function could be identified for the first time in cone snail venom duct. This work provides one of the first large-scale EST project for a cone snail venom duct using next-generation sequencing, allowing a detailed overview of the venom duct transcripts. This leads to an expanded definition of the overall cone snail venom duct transcriptomic activity, which goes beyond the cysteine-rich conopeptides. For instance, this study enabled to detect proteins involved in common post-translational maturation and folding, and to reveal compounds classically involved in hemolysis and mechanical penetration of the venom into the prey. Further comparison with proteomic and genomic data will lead to a better understanding of conopeptides diversity and the underlying mechanisms involved in conopeptide evolution.

1. Introduction

Cone snails of the genus *Conus* are predatory marine gastropods that utilize venom to capture prey. *Conus* venoms mainly consist of small disulfide-rich peptides commonly

Abbreviations: BLAST, Basic Local Alignment Search Tool; bp, base pair; CCG, Common Cellular Gene; EST, Expressed Sequence Tag; GO, Gene Ontology; LTR, Long Terminal Repeat; TE, Transposable Element; TRG, Toxin Related Gene; XB, Xenobiotic.

* Corresponding author.
E-mail address: yves.terrat@yahoo.fr (Y. Terrat).

0041-0101/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.toxicon.2011.10.001

referred to as conotoxins or conopeptides. Each of the near 700 *Conus* species synthesizes its own characteristic repertoire of toxic peptides. It has been estimated that the toxin repertoire of cone snails comprises more than 100,000 different bioactive compounds with various neurological targets (Terlau and Olivera, 2004). The venom cocktail of cone snails is further extended through the addition of post-translational modifications that increase toxin potency (Pisarewicz et al., 2005; Lopez-Vera et al., 2008) and assist the stabilization of the three-dimensional molecular structure (Craig et al., 2001). These components target a variety of ion channels, transporters and receptors besides the interest in their natural functions in venoms. They are therefore of much interest as drug leads (Favreau and Stöcklin, 2009). Even though cone snail venoms have been assiduously studied during the past few decades, a comprehensive study of the transcriptomic content of a cone snail venom duct is a new step in this field. Indeed, earlier studies aiming at specific conopeptide superfamilies and using only a restrained number of sequences did not lead to a complete overview of a whole venom composition. Furthermore, sampling effects have limited the identification of rare transcripts.

A number of research groups have sequenced portions of cone snail venom duct transcriptomes to identify the genes contributing to venoms. Most of these previous studies have relied on cloning of cDNA libraries and Sanger sequencing, generating important, but ultimately limited, data. More than half of the expressed sequence tags (ESTs) from these studies have, in most cases, been found to code for toxin genes (Pi et al., 2006a,b) and a large proportion of the remaining ESTs to code for genes involved in transcription, translation, cell regulation and metabolism. The application of next-generation sequencing, though not without its own drawbacks, should alleviate issues of low coverage and provide a more complete characterization of the genes contributing to cone snail venoms. The current transcriptomic analysis is integrated in the CONCO project (http://www.conco.eu/). This project focuses on the discovery and the development of new therapeutically relevant molecules issued from the venomous marine cone snail species *Conus consors*. Through the deep and exhaustive investigation of the animal biodiversity, of its genome, its venom gland transcriptome and its venom proteome, CONCO aims at exploiting in a sustainable way the great richness offered by these animals to discover the drugs of tomorrow (Dutertre et al., 2010; Kauferstein et al., 2011). The present manuscript presents analysis of data generated by high throughput sequencing. Not only this work focuses on characterizing the most abundant toxin-encoding transcripts in the venom-gland transcriptome, but it provides also analyses of the remaining part of the transcriptome, presenting new insights on the venom cocktail used by the cone snail to capture his prey.

2. Material and methods

2.1. Species collection and RNA extraction

Specimens of *C. consors* were collected in the Chesterfield Islands (New Caledonia) in the frame of the CONFIELD scientific expedition in June 2007. Venom ducts were dissected from three living snails, and were stored in RNA-later stabilization reagent following manufacturer's instructions (Qiagen). Total RNA was extracted with TRIzol® reagent using standard protocols (Invitrogen). The cDNA library was not subjected to a normalization step prior to sequencing.

2.2. 454 Pyrosequencing and contig assembly

Approximately 1 μg of the adaptator-ligated cDNA population was sheared by nebulization and DNA sequencing was performed following protocols for the genome Sequencer GS FLX System (Roche diagnostic). Sequence reads were processed in order to exclude low quality and poly A+ tracts using Trace2dbEST (Parkinson and Blaxter, 2004). Subsequently, assembly was undertaken with SeqMan pro (DNASTAR, USA) using high stringency clustering parameters (100% identity between reads with 40 nucleotides sequence overlap). The contigs were then assembled into clusters using Cd-hit software (Huang et al., 2010) with a 80% identity threshold. Each cluster is represented by its longest sequence for further similarity-based annotation.

2.3. Annotation of ESTs sequences

Bioinformatic processing was carried out using a combination of public softwares and home-made scripts using PERL programming language. The EST work-flow is summarized in Fig. 1. First, representative sequences of each cluster were BLAST-annotated (Altschul et al., 1990) against UniProt/Swiss-Prot (v56.2), Uniprot/TrEMBL (v39.2) (UniProt Consortium, 2011), in-house toxin database (named ToxRelDB) and Repbase repository dedicated to transposable elements (Jurka et al., 2005). Most toxin-related sequences from cone snails are conopeptides but numerous proteins have been convergently recruited into the venoms of various animals (Fry et al., 2009). For this reason, toxin sequences from a wide range of taxonomic groups are integrated in our regularly updated home-made toxin database to increase new toxin detection. A combination of BLASTn and BLASTx algorithms was used for contig annotation. This later was considered successful when the best match displayed an expected e-value $\leq 1.10^{-3}$. After the first round of data treatment, sequences were parsed using a home-made PERL script to classify representative sequences in 6 categories: Common cellular genes (CCGs), Conopeptides, Toxin-related genes (TRGs), Xenobiotics (XBs), Transposable elements (TEs) and Unknown sequences (UKs).

Sequences related to a given cluster and encoding for a putative conopeptide family were submitted to a home-made program for isoforms identification. To increase sensitivity of BLAST searches to detect conopeptides, a second BLAST step was performed using as queries the first dataset of *C. consors* sequences identified so far.

Since many venom genes have been shown to have evolved from genes coding for normal cellular products, it is therefore difficult to reject an expressed gene as a venom component just because it has some similarity with a gene coding for a protein involved in a "conventional" cellular process. Due to this issue, an estimation of the transcription level (see further) for the different contigs was performed