Toxicon 107 (2015) 282–289

Contents lists available at ScienceDirect

Toxicon

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

An efficient transcriptome analysis pipeline to accelerate venom peptide discovery and characterisation

Jutty Rajan Prashanth, Richard J. Lewis*

IMB Centre for Pain Research, The University of Queensland, 306 Carmody Road, St. Lucia, 4072, Australia

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form
26 August 2015
Accepted 10 September 2015
Available online 14 September 2015

Keywords:
Venomics
Proteomics
Transcriptomics
Conotoxins
Venom peptides

ABSTRACT

Transcriptome sequencing is now widely adopted as an efficient means to study the chemical diversity of venoms. To improve the efficiency of analysis of these large datasets, we have optimised an analysis pipeline for cone snail venom gland transcriptomes. The pipeline combines ConoSorter with sequence architecture-based elimination and similarity searching using BLAST to improve the accuracy of sequence identification and classification, while reducing requirements for manual intervention. As a proof-of-concept, we used this approach reanalysed three previously published cone snail transcriptomes from diverse dietary groups. Our pipeline method generated similar results to the published studies with significantly less manual intervention. We additionally found undiscovered sequences in the piscovorous *Conus geographus* and vermivorous *Conus miles* and identified sequences in incorrect superfamilies in the molluscivorus *Conus marmoreus* and *C. geographus* transcriptomes. Our results indicate that this method can improve toxin detection without extending analysis time. While this method was evaluated on cone snail transcriptomes it can be easily optimised to retrieve toxins from other venomous animals.

1. Introduction

Venoms are among the most common adaptations across the animal kingdom ranging from bees and wasps, snakes, scorpions, spiders and marine animals such sea anemones, jellyfish and cone snails for both prey capture and defence (Casewell et al., 2013). Venoms induce a range of effects including cardiotoxicity, myotoxicity, and neurotoxicity with potency and specificity leading to the widespread interest in them as possible therapeutics (King, 2011). Several molecules such as the blockbuster ace-inhibitor Captopril, originally isolated from the venom of the snake *Bothrops jararaca* and the intrathecal analgesic Prialt, originally isolated from the venom of the cone snail *Conus* magus, showcase the therapeutic potential of venoms (King, 2011). Toxins have also been used to probe receptor–ligand interactions at their respective molecular targets. For example, the crystal structure of ASIC1a bound to psalmotoxin-1 (Baconguis and Gouaux, 2012; Dawson et al., 2012) isolated from the spider *Araneae theraphosidae* (Escoubas et al., 2000) was used to map the toxin-binding domain and understand activation mechanisms of ASICs (Baconguis and Gouaux, 2012). The co-crystallisation of ASIC1a with MitTx, a pain causing Texas coral snake toxin revealed the open state conformation of the channel (Baconguis et al., 2014). Similarly, a number of α-conotoxins including TxIA (Dutertre et al., 2007) and PnIA (Celie et al., 2005), as well as snake toxins such as α-cobratoxin (Bourne et al., 2005), have been used to study binding interactions of nAChRs via its molluscan glial surrogate protein, AChBP (van Dijk et al., 2001).

Venoms also act as models for evaluating the role of natural selection on predator–prey interactions facilitated by the rapid rates of evolution of toxin genes and the expression of individual toxins by single genes (Casewell et al., 2013). While many venom systems are thought to have evolved primarily for predation, marine cone snails produce distinct predatory and defencive venoms, thus allowing the study of their evolutionary response to different ecological pressures (Casewell et al., 2013; Dutertre et al., 2014). As a result of these diverse evolutionary pressures, venoms continue to provide novel tools for studying receptor function, with a significant number having been evaluated for their therapeutic potential (Casewell et al., 2013).

Venoms invariably consist of complex mixtures of peptides and proteins acting in a synergistic manner. To isolate individual peptides, venoms were traditionally first separated by assay-guided fractionation before assaying in animal models. However, this

* Corresponding author.
E-mail address: r.lewis@imb.uq.edu.au (R.J. Lewis).

http://dx.doi.org/10.1016/j.toxicon.2015.09.012
0041-0101/

method requires large quantities of venom, and is time and resource intensive (Prashanth et al., 2012). Recent advances in transcriptomic and proteomic approaches, and the development of complementary bioinformatics tools have established 'venomics' as an accelerated method for studying venoms, with several seminal discoveries reported using this approach (Pineda et al., 2014; Prashanth et al., 2014; Zelanis and Tashima, 2014). In addition to novel toxin discoveries (Jin et al., 2014; Viala et al., 2015), venomics has helped uncover the mechanisms governing toxin diversification (Dutertre et al., 2013; Jin et al., 2013), distinct defencive and predatory venom gland specialisation in Conidae (Dutertre et al., 2014), and the morphological constraints driving the evolution of centipede venoms (Undheim et al., 2015). In the absence of reference genomes for many venomous animals, transcriptome sequencing of venom glands has come to underpin the venomics approach and has enabled novel toxin discovery at an unprecedented level from snakes (Durban et al., 2011), spiders (Pineda et al., 2014), scorpions (Rendón-Anaya et al., 2015), cone snails (Prashanth et al., 2014), and even relatively poorly characterised animals such as ants (Bouzid et al., 2013).

With the reduced cost of 454-Pyrosequencing and Illumina, sequencing the venom gland transcriptome has become an affordable and relatively quick way to fingerprint the venom profile of animals (Liu et al., 2012). This approach can also uncover rare peptides that maybe missed by traditional assay-guided fractionation (Prashanth et al., 2012). In particular, transcriptome sequencing has been used extensively to study of cone snail venoms because a single read of the 454-Pyrosequencing platform can cover the entire conotoxin precursor cDNA (~300 bp) thus circumventing the issue of assembly and this sequencing platform has been used to uncover the venome of various Conidae (Prashanth et al., 2014). Recent technological advances have increased sequence read lengths generated by the Illumina platform allowing better quality assemblies, which combined with the much greater sequencing depth provided by the platform (Schirmer et al., 2015) has already started to be used to sequence venom gland transcriptomes producing much larger datasets (Barghi et al., 2015; Lavergne et al., 2015).

For such transcriptomic datasets, data analysis involves identifying and classifying putative venom peptides. Sequence annotation typically uses homology searching using BLAST to either nucleotide or protein sequence databases with programs like BLAST2GO (Conesa et al., 2005) used to perform process level annotation (Stein, 2001). However, the sheer volume of data generated in next-generations sequencing experiments renders such an approach computationally restrictive or very time-consuming. Stand-alone programs such as ConoSorter that translate cDNA reads into six reading frames and identify coding sequences of conotoxins using a combination of regular expressions and profile hidden Markov models (pHMM) have partially overcome this issue (Lavergne et al., 2013). Though this program can handle large datasets, an overreliance on such programs can miss novel toxin sequences that frequently possess novel cysteine scaffolds. It can also lead to incorrect annotations, such as the Con-insulins from *Conus geographus* being misidentified as a novel conotoxin gene superfamily (Safavi-Hemami et al., 2015).

To improve transcriptomic data analysis, we have optimised a sequence annotation pipeline designed to efficiently identify conotoxin-like sequences from large datasets using freely available bioinformatics tools. As a proof of concept, we present a reanalysis of three published cone snail venom gland transcriptomes from *Conus marmoreus* (Dutertre et al., 2013; Lavergne et al., 2013), which was used for the original benchmarking of ConoSorter, *Conus miles* (Jin et al., 2013), and *C. geographus* (Dutertre et al., 2014). With the exception of two highly divergent superfamilies reported from *C. geographus*, and the S-superfamily sequences from *C. marmoreus* that were reported at low levels in the original analysis, we quickly discovered all previously reported superfamilies represented by at least two reads in our reanalysis. In addition, we discovered several superfamilies that were missed previously, including putative new superfamilies, and reclassified some misclassified sequences. Thus, our pipeline approach has demonstrated utility and efficiency for the analysis of large venom gland transcriptomes from Conidae. Although this method was designed to identify conotoxins from next generation data sets due to the availability of standalone programs such as ConoSorter and large volumes of next-generation sequencing data (Prashanth et al., 2014), it is adaptable to the study of other venomous animals such as snakes or spiders.

2. Materials and methods

2.1. Sequence analysis pipeline

Our pipeline approach is outlined in Fig. 1. Specifically, raw data from sequencing experiments is either assembled (Illumina) using assemblers such as SOAPdenovo (Xie et al., 2014) or Trinity (Grabherr et al., 2011) or filtered based on the raw read quality score (454-pyrosequencing) using programs such as QTrim (Shrestha et al., 2014) or NGS QC Toolkit (Patel and Jain, 2012). In our pipeline, a stringent quality control score of 30 is used to remove low quality reads. Quality controlled data is then sorted initially using ConoSorter, which translates raw cDNA sequences into six reading frames and extracts sequences from the first start codon in each read to the first subsequent stop codon. Extracted sequences are then searched against a training dataset comprised of sequences from the Conoserver (Kaas et al., 2008, 2012) database using Regular Expressions first to sort the sequences. ConoSorter also calculates class and superfamily scores ranging from 0 to 3 based on the similarity of the predicted signal-, pro- and mature regions of the sequences to known toxin classes and superfamilies with a score of 3 indicating matches for each region and 0 indicating no matches. The total class and superfamily scores for each sequence are calculated by adding the scores of each region of the sequence. The sequences are then classified into their respective superfamilies based on these similarities. Sequences that could not be sorted into known superfamilies by Regular Expressions are then subjected to a pHMM-based scan against profiles generated from the conotoxin training dataset. The pHMM module returns e-values for each matched section indicating the quality of the match (Lavergne et al., 2013).

Sequences that were unequivocally identified by ConoSorter are then separated, while the remaining unclassified sequences are further analysed in the pipeline. The sequences from the regular expression file are filtered based on number of reads ($n \geq 2$), sequence length (Sequence length > 50 amino acids), hydrophobicity of the signal region (Hydrophobicity > 50), class score (Score ≥ 2), superfamily score (Score ≥ 1), with sequences containing unrecognised amino acids removed. For sequences in the pHMM, an e-value cut-off (superfamily e-value < 0.0001) was implemented to prevent false identification of sequences as conotoxins in place of the class and superfamily scores. The other filtering parameters applied to sequences in the regular expression files are then applied to those in the pHMM file. Filtered sequences from each file are pooled and any duplicates removed.

To classify sequences into superfamilies, signal regions from filtered sequences are extracted using SignalP and sequences lacking signal regions discarded. Sequences are then clustered based on their signal sequences using the program CD-HIT using a signal peptide identity threshold of 75%. Representative sequences from each cluster are then annotated using BLASTp against the non-