Accepted Manuscript

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

Jutty Rajan Prashanth, Richard J. Lewis

PII: S0041-0101(15)30072-6
DOI: 10.1016/j.toxicon.2015.09.012
Reference: TOXCON 5182

To appear in: Toxicon

Received Date: 31 July 2015
Revised Date: 26 August 2015
Accepted Date: 10 September 2015

Please cite this article as: Prashanth, J.R., Lewis, R.J., An efficient transcriptome analysis pipeline to accelerate venom peptide discovery and characterisation, Toxicon (2015), doi: 10.1016/j.toxicon.2015.09.012.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
An efficient transcriptome analysis pipeline to accelerate venom peptide discovery and characterisation

Jutty Rajan Prashanth and Richard J. Lewis.

IMB Centre for Pain Research, The University of Queensland, 306 Carmody Road, St. Lucia, Australia – 4072.
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 *C. geographus* and vermivorous *C. miles* and identified sequences in incorrect superfamilies in the molluscivorus *C. 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.
Introduction

Venoms are among the most common adaptations across the animal kingdom ranging from bees and wasps, snakes, scorpions, spiders and marine animals such as 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 defensive 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 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 defensive 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 Coninsulins 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 *Conus 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.

Materials and Methods

Sequence analysis pipeline
Our pipeline approach is outlined in Figure 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; Kaas et al., 2011) database using Regular Expressions first to sort the sequences. ConoSorter also calculates class and superfamily scores ranging from 0–3 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 >= 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-redundant UniPROT database. Housekeeping proteins such as transporters, structural proteins or contaminants are discarded at this step and identified conotoxin sequences are placed into various superfamilies based on the identity of their signal peptides. Of the third group of sequences that are novel or similar to hypothetical proteins, only sequences that were not singletons in the clustering step i.e. only sequences representing clusters with n>1 are considered for further analysis. The signal, pro- and mature regions of the sequences are then generated using ConoPrec available on the conoserver (Kaas et al., 2011) and sequences with cysteine frameworks typical of a particular superfamily, and signal sequences that are more than 53.3% similar to it are designated as new members of the superfamily. Those sequences with less than 53.3% similarity despite displaying the archetypal conotoxin architecture are designated as
putative new superfamilies pending proteomic validation and characterisation of activity. The threshold of 53.3% was used for sorting sequences into superfamilies based on a comparison of signal sequence identities (Lavergne et al., 2013). Sequences without readily identified known conotoxin characteristics were discarded to reduce false positives although they may represent highly divergent toxin families. All sequences are finally verified manually using the ConoPrec program (Kaas et al., 2011).

To compare our pipeline approach to published results, we performed a reanalysis of three previously published transcriptomes of the worm-hunting Conus miles (Jin et al., 2013), the mollusc-hunting Conus marmoreus (Dutertre et al., 2013) and the fish-hunting Conus geographus (Dutertre et al., 2014)

**Results**

**Reanalysis of the C. marmoreus transcriptome**
Sequencing of the C. marmoreus venom gland transcriptome yielded 179,843 raw reads. In addition, to the original analysis by Dutertre et al (Dutertre et al., 2013), this dataset was previously used for benchmarking the conotoxin-sorting program ConoSorter (Lavergne et al., 2013). Here, quality controlled sequences sorted by ConoSorter produced 59198 transcripts in the regular expression file and 1453 transcripts in the pHMM file. Filtering and removal of duplicates left 315 sequences of which ConoSorter classified 183 sequences into various superfamilies. The remaining 132 sequences were subsequently analysed through the pipeline to detect any previously undetected superfamilies. In all, 192 conotoxin sequences were detected and classified into 17 superfamilies in total. The original study by Dutertre et al. where the transcriptome was manually annotated by BLAST searching reported 105 sequences classified into 13 superfamilies. Importantly, all previously detected superfamilies were also observed in our analysis with the exception of the S-superfamily represented by two transcripts expressed by low levels in the original analysis. Neither sequence were detected by ConoSorter in our reanalysis, suggesting that they were excised
from the dataset at the quality filtering stage presumably due to insufficient sequence read quality.

Lavergne et al used ConoSorter to identify and classify conotoxins from the same C. marmoreus dataset to benchmark the program (Lavergne et al., 2013). They discovered 264 transcripts including 158 novel conotoxins, with 125 sequences belonging to previously discovered superfamilies and 33 novel peptides classified into 13 new gene superfamilies of which 60% were validated using MS/MS evidence. Of these 13 superfamilies, 9 contained sequences with only single reads (H2, M2, N2, Q, R, W, Y2, Y3, and Z), with only the H2, M2 and Y2 superfamilies validated by proteomics (Lavergne et al., 2013). Our reanalysis discovered more 'I4' superfamily sequences and reclassified these into the I2 superfamily. A closer inspection of the reported novel Y2 superfamily sequences revealed that they were also misclassified (Safavi-Hemami et al., 2015) and belonged to the recently described coninsulin class (Figure 2). The H2 superfamily described by Lavergne et al. comprised one transcript with a single read that was identical to the peptide Mr3.8 belonging to the M superfamily with the exception of the first few residues of the reported signal sequence (Lavergne et al., 2013). Here, we retrieved several sequences including a number with high reads (>100), containing this sequence region from the pHMM module of ConoSorter, with some sequences containing an extended region before the reported signal region. Analysis by SignalP detected no signal region from these sequences and they reported a superfamily score of 0 in the Regex module. Taken together, this information suggests that the original H2 superfamily sequence was likely a product of a sequencing read error (Figure 2). Apart from the aforementioned transcripts that were misclassified, we found O4, U, W, and X superfamilies, all containing sequences with at least 2 reads (Figure 3). While the analysis using ConoSorter reported 2 sequences belonging to the I4/I2 superfamily, we found 6 such sequences in our reanalysis (Figure 2). Also, eleven O4 superfamily sequences instead of the eight O4 superfamily sequences originally reported were identified here (Figure 3). A sequence belonging to the recently identified new superfamily one from Conus geographus (NSG1) was identified with 4 reads (Dutertre et al., 2014) (Figure 3). All the major
superfamilies reported in the previous analysis except the S-superfamily was found in our reanalysis. Thus, our pipeline approach has helped discover additional conotoxin transcripts in the transcriptome of *C. marmoreus*, despite two previous analyses of the same dataset, and has allowed the reclassification of a number of sequences.

**Re-analysis of the *C. miles* venom gland transcriptome**

The *C. miles* transcriptome dataset (Jin et al., 2013) yielded 255,829 reads in total that were filtered using the quality threshold of 30 and sorted using ConoSorter producing 69,814 sequences in the Regex file and 1616 sequences in the pHMM file. After filtering based on various parameters, 129 transcripts classified by ConoSorter were separated and 754 unsorted sequences were further analysed to reveal an additional 41 conotoxin sequences, giving a total of 168 conotoxin transcripts with a minimum of two reads classified into superfamilies. The original analysis of the same dataset by Jin et al. identified 662 precursors including single reads belonging to a total of 8 previously identified (D, I2, L, M, O1, O2, P, and T) and 8 new superfamilies named SF-Mi1–SF-Mi8 (Jin et al., 2013) including the P-superfamily supported only by single reads (Jin et al., 2013). In addition to identifying all previously reported superfamilies (except the single read P superfamily), we additionally identified E superfamily and Con-ikot-ikot sequences as well as two novel superfamilies previously identified in the transcriptome of *C. vexillum* (Prashanth et al., manuscript in preparation). The Con-ikot-ikot sequences were expressed at moderate levels (>10 reads) while the other sequences were found at low levels (2–10 reads) (Figure 4). Thus, a total of 18 superfamilies were discovered from *C. miles* using our pipeline approach, compared to the 16 superfamilies reported in the original study.

**Re-analysis of the *C. geographus* transcriptome**

The *C. geographus* dataset yielded 152,752 raw sequence reads that were filtered and sorted initially using ConoSorter producing 38,751 sequences in the Regular Expression file and 1230 sequences in the pHMM file. ConoSorter classified 287 of these sequences into 13 different gene superfamilies. After parameteric
filtering, an additional 221 sequences were analysed. A further 54 conotoxin sequences were discovered and classified, bringing the total to 341 sequences from 22 gene superfamilies. In comparison, the original transcriptomic analysis of the same dataset by Dutertre et al found 20 superfamilies (Dutertre et al., 2014). Here, we discovered sequences belonging to N2, SF-Mi7, U and W superfamilies and sequences misclassified as belonging to the I1 superfamily based were placed into the NSVx5 superfamily (Prashanth et al, Manuscript in preparation) based on signal sequence identity. We also found a novel superfamily named NSG5 (Figure 5). The NSG5 superfamily had one sequence that was expressed at a high level (171 reads), further demonstrating that even highly expressed sequences maybe missed by relying on ConoSorter and/or similarity searching alone. However, two of the superfamilies reported originally, NSG1 and NSG4 were discarded in our reanalysis since they did not meet the superfamily score threshold. While relaxed filtering parameters would help increase the sensitivity of the pipeline and help discover highly divergent sequences such as those belonging to NSG1 and NSG4, this came at the cost of more false positives.

The reanalysis of the three transcriptomes using our method identified a number of unreported sequences and superfamilies from the transcriptome datasets. We compared our method with ConoSorter to show that our method can detect sequences and superfamilies from these datasets that maybe otherwise missed. While ConoSorter performed nearly as well as our pipeline on the C. marmoreus dataset, on which it was originally benchmarked due to these sequences being a part of the Conoserver training dataset, in both the C. geographus and the C. miles datasets, many superfamilies were not classified by ConoSorter. In addition, two of the superfamilies that were erroneously assigned as reported here were identified by ConoSorter, showing that misclassified superfamilies used to train the ConoSorter models could lead to downstream inaccuracies unless verified with similarity-searching as is the case with our method (Table 1). Thus, our method offers an advantage in both accuracy and sensitivity over using ConoSorter only.
Discussion

The adoption of next-generation sequencing technology has accelerated the discovery of new venom peptides and helped inform mechanistic aspects of envenomation (Prashanth et al., 2014). Advances in sequencing throughput have gradually shifted the rate-limiting step in venom analysis from data collection to analysis (Wang et al., 2009). Thus, a robust and efficient pipeline to identify toxin-like sequences from large datasets is required. Sequence annotation at the protein level is reliant on sequence or structural similarities for identification and classification. Since similarity searching is computationally expensive for large datasets dedicated programs like ConoSorter have been developed to analyse such data (Lavergne et al., 2013). However, ConoSorter is restricted to detecting sequences belonging to known superfamilies and cannot discriminate sufficiently between novel toxin sequences and other proteins apart from generating class and superfamily scores. Using only these scores as a guide to delineate new superfamilies can lead to cases where sequences are misclassified as seen for the Coninsulins that were incorrectly classified as new superfamilies in two different studies (Dutertre et al., 2014; Lavergne et al., 2013). In addition, while the majority of conotoxins are small disulfide-rich peptides and thus easily detected, larger sequences such as con-ikot-ikots that have widely varying signal regions (Barghi et al., 2014) maybe missed as seen in our reanalysis of the *C. miles* transcriptome. The pipeline described here integrates ConoSorter with a series of freely available bioinformatics tools in an eliminatory workflow before similarity searching to accelerate toxin discovery and classification. In designing the pipeline, we have strived to make this approach accessible to toxinologists with only basic bioinformatics knowledge by only using GUI driven programs freely available on the web. As a proof-of-concept, we reanalysed three published transcriptomes to draw comparisons between our pipeline and current methods for analysis.

Two aspects of our pipeline design underpin its advantages. Firstly, the sequence in which the various tools are applied improves the efficiency of the pipeline by greatly reducing the number of sequences at each subsequent step through a process of elimination, leaving very few sequences that have to be manually
sorted following the BLAST results. Secondly, the filtering parameters used after the ConoSorter step allowing careful selection of additional sequences for further analysis were developed iteratively using three test datasets. The stringency of these parameters determines the speed, sensitivity, and false discovery rates of the pipeline. Here, we aimed to minimise false positives by only retaining sequences with the archetypical conotoxin architecture, including length, hydrophobicity of the probable signal region, class and superfamily scores generated by the Regular Expression module of ConoSorter (Lavergne et al., 2013). Briefly, Conosorter compares each segment of the sequence (Signal, Pro-, and Mature regions) against various classes and superfamilies, with matches or mismatches in each section getting a score of 1 and 0 respectively and overall scores are calculated by adding the scores of each section of the sequence giving a maximum possible class and superfamily scores of 3 (Lavergne et al., 2013). Thus, these scores act as an index of similarity to the canonical conotoxin architecture. Here, only sequences with a class score \( \geq 2 \) and superfamily \( \geq 1 \) were retained for further analysis, ensuring all sequences shared architectural similarity to other conotoxins. Since sequences not classified by the Regular Expression module are classified by the more sensitive pHMM module in the ConoSorter pipeline (Lavergne et al., 2013), we used a superfamily e-value cut-off score of 0.0001 for sequences in the pHMM module of the program to avoid false identification.

The use of stringent filters can on occasion lead to the elimination of known sequences. For example, two novel superfamilies characterised by a lack of cysteines in the mature region namely NSG1 and NSG4 reported were (Dutertre et al., 2014) eliminated in our analysis because the superfamily scores of the sequences were 0. Despite this, a NSG1 sequence with a slightly different signal sequence was detected in the transcriptome of *C. marmoreus* since it had a superfamily score of 1, demonstrating that even variation in a few residues can alter the match or mismatch output from the Regular Expression module of ConoSorter and thus affect all the downstream steps. In spite of this, only the aforementioned highly divergent sequences were missed with our pipeline approach. Entering the results from more sophisticated approaches such as
those described here into the training dataset used by ConoSorter will continue
to expand its capability to detect such divergent sequences. We also eliminated
transcripts with single reads, which comprised ~ 75% of transcripts, to avoid
sequencing errors. While single reads encoding several novel sequences have
been described, the rate of proteomic validation is dramatically reduced, with
just 3 out of 9 superfamilies expressed by single reads in the transcriptome of C.
marmoreus validated by proteomics (Lavergne et al., 2013). Hence, we discarded
single read sequences from our analysis. Our reanalysis of the transcriptomes
using our pipeline and the aforementioned parameters identified nearly all the
sequences reported in previous studies, clarifying misclassifications and also
discovered novel toxin classes, demonstrating the utility and efficiency of our
pipeline. The method described in this paper also offers clear sensitivity
advantages over using ConoSorter alone, in particular, in those species whose
transcriptomic sequences were not used to train the original dataset (Table 1).

While our reanalyses have been performed on 454 pyrosequencing data, which
has been the most commonly used platform for the analysis of cone snail venom
gland transcriptomes thus far (Prashanth et al., 2014), more recent studies
(Barghi et al., 2015; Lavergne et al., 2015) have begun to use Illumina sequencing
technology (Schirmer et al., 2015), generating shorter reads but more data and
uncovering further sequence diversity (Lavergne et al., 2015). Since ConoSorter
is able to handle these larger datasets after assembly (Lavergne et al., 2015), it
follows that our pipeline is well placed to probe larger datasets. The method can
be broadened to the identification of non-secreted sequences by skipping the
signal peptide requirement and proceeding directly to clustering sequences
using CD-HIT, and can be adapted to transcriptomes of different venomous
animals, once the Regular Expression and pHMM models are trained for their
venom sequences (Lavergne et al., 2015).

**Acknowledgements:** This work was supported by an IMB PhD Scholarship (to
JRP), and an NHMRC Program Grant and Principal Research Fellowship (to RJL).
References


correlates distinct binding orientations with nAChR subtype selectivity. EMBO J. 26, 3858-3867.


Table 1. Comparison of ConoSorter (21/08/15) with our integrated pipeline method

<table>
<thead>
<tr>
<th></th>
<th>Number of Transcripts/Superfamilies</th>
<th>Additional Superfamilies reported here</th>
<th>Sensitivity (Transcripts)</th>
<th>Sensitivity (Superfamily)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ConoSorter Only</td>
<td>Our Pipeline</td>
<td>ConoSorter Only</td>
<td>Our Pipeline</td>
</tr>
<tr>
<td>C. marmoreus</td>
<td>184/18*</td>
<td>192/17</td>
<td>1</td>
<td>96%</td>
</tr>
<tr>
<td>C. miles</td>
<td>129/16</td>
<td>167/18</td>
<td>4</td>
<td>77%</td>
</tr>
<tr>
<td>C. geographus</td>
<td>287/13</td>
<td>341/21</td>
<td>6</td>
<td>84%</td>
</tr>
</tbody>
</table>

* - Two of the 18 superfamilies detected by ConoSorter were misclassifications. The H2 superfamily delineated by Lavergne et al appears to be the product of a sequencing error. Similarly, the I4 superfamily transcripts appear to be a part of the I2 superfamily (Figure 2).

# - The sensitivity percentage was calculated using the total number of superfamilies that included misclassified superfamilies.
Figures

![Flowchart of the process](image)

Figure 1: Cone snail venom gland transcriptome analysis pipeline.
Figure 2: Sequence alignments of misclassified superfamilies from the *C. marmoreus* transcriptome.

Figure 3: Sequence alignments of novel superfamilies discovered in our reanalysis of the *C. marmoreus* transcriptome.
Figure 4: Sequence alignments of novel superfamilies discovered in our reanalysis of the *C. miles* transcriptome in this study.

Figure 5: Sequence alignments of novel superfamilies discovered in our reanalysis of the *C. geographus* transcriptome in this study.
• Transcriptomes are changing our understanding of the chemical diversity of venoms.
• We have developed an optimised analysis pipeline for cone snail transcriptomes.
• Reanalysis of three published cone snail transcriptomes generated improved results with significantly less manual intervention.
• The method can be "trained" to retrieve toxins from other venomous animals.