Peer-reviewed version available at *Toxins* **2018**, *10*, 503; doi:10.3390/toxins10120503

Article

Discovery of novel conotoxin candidates using machine learning

Qing Li [1,2], Maren Watkins[3], Samuel D. Robinson[3], Helena Safavi-Hemami [3,4,*#] and Mark Yandell [1,5,*#]

[1] Eccles Institute of Human Genetics, University of Utah, USA; liqing850104@gmail.com
[1] Eccles Institute of Human Genetics, University of Utah, USA; myandell@genetics.utah.edu
[2] Huntsman Cancer Institute, University of Utah, USA; liqing850104@gmail.com
[3] Department of Biology, University of Utah, USA; maren.watkins@hsc.utah.edu
[3] Department of Biology, University of Utah, USA; helena.safavi@utah.edu
[4] Department of Biochemistry, University of Utah, USA; helena.safavi@utah.edu
[5] USTAR Center for Genetic Discovery, University of Utah; myandell@genetics.utah.edu
* Correspondence: safavihelena@gmail.com; myandell@genetics.utah.edu
# Equal contribution

Abstract: Cone snails (genus *Conus*) are venomous marine snails that inject prey with a lethal cocktail of conotoxins, small, secreted, cysteine-rich peptides. Given the diversity and often high affinity for their molecular targets, consisting of ion channels, receptors or transporters, many conotoxins have become invaluable pharmacological probes, drug leads and therapeutics. Transcriptome sequencing of *Conus* venom glands followed by *de novo* assembly and homology-based toxin identification and annotation is currently the state-of-the-art for discovery of new conotoxins. However, homology-based search techniques, by definition, can only detect novel toxins that are homologous to previously reported conotoxins. To overcome these obstacles for discovery we have created *Conus*Pipe, a machine learning tool that utilizes prominent chemical characters of conotoxins to predict whether a certain transcript in a *Conus* transcriptome, which has no otherwise detectable homologs in current reference databases, is a putative conotoxin. By using *Conus*Pipe on RNASeq data of 10 species, we report 5,230 new putative conotoxin transcripts that have no homologues in current reference databases. 893 of these were identified by at least 3 out of 4 models used. These data significantly expand current publicly available conotoxin datasets and our approach provides a new computational avenue for the discovery of novel toxin families.

Keywords: machine learning; conotoxins; cone snails, venom, drug discovery

Key Contribution: By using ensemble methods, we report 5,230 new candidate conotoxins that have no homologues in current reference databases and provide a unique dataset for future pharmacotherapeutic exploration.

1. Introduction

Predatory marine cone snails (genus *Conus*) have attracted the attention of biologists and pharmacologists for the great neuropharmacological potential of their venom toxins [1-3]. It is estimated that each of the ~750 extant *Conus* species produces ~100-400 distinct venom toxins (conotoxins) with almost no overlap in the toxin repertoire between the ~750 species, not even between sister species [4]. Despite the tremendous diversity and drug discovery potential of *Conus*

venoms, only ~ ~5,000 nucleotide sequences of conotoxin-encoding transcripts have been reported from 100 *Conus* species over the past decades, with most sequences having been discovered in recent years [5,6]. Traditional methods, such as isolation of conotoxins from venom and subsequent Edman- or *de novo* mass spectrometric (MS) sequencing are time-consuming and limited by sample availability. In contrast, high throughput transcriptome sequencing can achieve greater sequencing depth and only requires small amounts of biological sample [7]. Recent studies on the venom gland transcriptomes of several cone snail species, using next generation sequencing technologies (NGS), have discovered ~100-400 conotoxin genes per *Conus* species [4,8-12]. Other authors have reported larger diversities but these are likely to have resulted from inappropriate analyses of NGS datasets, as previously discussed [4,10].

Conotoxins can be classified into different gene superfamilies based on their conserved N-terminal signal sequence [13]. To date, more than 53 conotoxin gene superfamilies have been described for *Conus* [14]. After NGS sequencing and *de novo* transcriptome assembly, candidate conotoxin genes are usually assigned to different superfamilies using BlastX, regular expression-based techniques and profile hidden Markov model (HMMER) analysis against a local reference database of known conotoxins from the Uniprot and/or ConoServer databases. Available tools include ConoPrec, ConoDictor and Conosorter [5,15-17]. However, these approaches can only detect novel toxins that are similar to previously reported sequences. An approach that could overcome the limitations of homology based-searches would thus be highly desirable.

Most conotoxin transcripts can be readily divided into three distinct regions: (1) an N-terminal signal sequence for targeting to the endoplasmic reticulum; (2) an intermediate propeptide region that has been suggested to play a role in secretion, posttranslational modification and folding; and (3) a single copy of the mature toxin region, located at the C terminus [18-20]. We hypothesized that even though conotoxin sequences evolve very rapidly [14], conotoxin retain these three traits even when their sequence similarities become too low to allow detection by alignment-based methods such as Blast and HMMER. If this were true, even highly divergent conotoxins might be identifiable using machine learning methods trained to identify these three traits as features. With this in mind, we implemented three machine learning models for data mining of 12 *Conus* transcriptomes from 10 different species: logistic regression (logit), semi-supervised learning (LabelSpreading) and an artificial neural network (perceptron) [21-25]. The resulting tool for conotoxin discovery is called *Conus*Pipe.

Generalized Linear Models (GLMs) are versatile, powerful, commonly used statistical approaches to model relationship between scalar response variables given several predictor variables/features [21,26]. In particular the logistic regression model computes a weighted sum of the input features (plus a bias term), but instead of outputting the result directly like the linear regression model does, it outputs the logistic of this result [21,25]. This approach often outperforms simple linear regression for binary outcome prediction [26].

Unlike GLMs which completely employ labeled data for training, in semi-supervised learning only some of the training data is labeled. Graph-based methods are then used to make use of additional unlabeled data in order to better capture the shape of the underlying data distribution and generalize the method to apply to new samples [27]. The assumption is that unlabeled data with features that render them neighbors of labeled data are likely to have a common label. In keeping with general practice, we used the K-nearest neighbors method to connect each data point [28]. Semi-supervised learning approaches can perform well when only a small number of labeled data but large amounts of unlabeled data are available.

The Perceptron is one of the simplest Artificial Neural Network ANN architectures, which is composed of a single layer of linear threshold unit (LTU). The LTU computes a weighted sum of its inputs and then applies a step function to that sum and outputs the result [22]. Unlike regression-based approaches, ANNs can capture dependencies within the data, potentially resulting more accurate classification.

The nature of the *Conus* training sets and input data also needs to be considered. For example, the true-positive (labeled) training data is limited in scale and is incomplete, as many conotoxins

presumably remain uncatalogued [29]. Moreover, the input data is very large, with RNA-seq datasets typically exceeding five million reads and tens of thousands of assembled transcripts. Given these features, logistic regression provides a well-established base-line approach, whereas semi-supervised learning (LabelSpreading) provides a means to further leverage unlabeled true positives during training. Finally, the Perceptron model scales well to very large training sets and is widely used for pattern recognition with good results [30-33]. Additionally, since using these machine learning models are based on the hypothesis that conotoxins will retain all three traits in sequence evolution, we added cross-species Blastp to search for similar unknown sequences (if they contain a signal sequence) between different *Conus* species to rescue potential conotoxins that only have one trait (signal sequence) based on the knowledge that the signal peptide sequences of conotoxins from the same superfamilies are highly similar to each other even if they are from different *Conus* species [4].

By employing three different machine learning models plus Blastp, *Conus*Pipe allows users to take an ensemble approach for discovery to maximize the prediction power. All four methods were applied to 12 RNAseq datasets from 10 different species of *Conus*. The pipeline discovered 5,230 new conotoxin candidates that provide a unique dataset for future pharmacotherapeutic exploration.

2. Results

2.1 The ConusPipe toolkit

*Conus*Pipe is implemented in Perl and Python as a complete conotoxin discovery package. It is available at https://github.com/Yandell-Lab/ConusPipe. *Conus*Pipe takes 6-frame-translated peptide sequences from nucleotide sequences which have no hit in current reference database and extracts conotoxin sequence features to train datasets (**Figure 1A**). In addition to 3 machine learning models, cross-species Blastp is used as the 4thmethod to retrieve putative toxin candidates that have a signal sequence but may not have all features used in machine learning (**Figure 1B**).

*Conus*Pipe then generates different combinations (single method, union or overlap) of the four methods to predict candidate conotoxins (**Figure 2**). Users can change the settings in sample.config file to use different cut-off options for transcript per million (tpm) values, blast e-values, signalP D-values and provide paths to input and output fasta files and databases used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2018 doi:10.20944/preprints201809.0587.v1

Peer-reviewed version available at Toxins 2018, 10, 503; doi:10.3390/toxins10120503

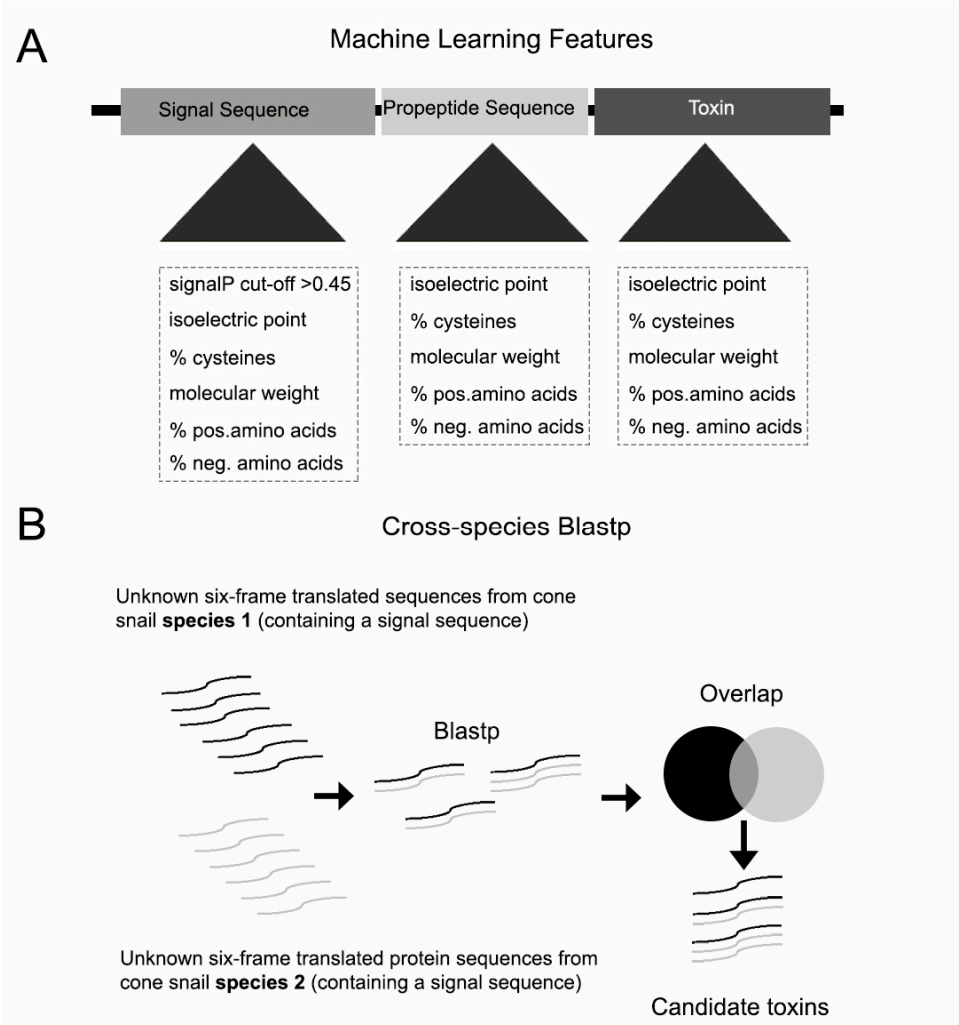


Figure 1. **A.**Overview of feature selection for machine learning models (**B**) and cross-species Blastp methodology used in addition to the machine learning model.

2.1.1 Building and cross-validation of the machine learning models

4,950 known conotoxin sequences (from the ConoServer [5] and Uniprot databases [34] and 52,613 randomly selected non-conotoxin *Conus* transcripts were used to build the machine learning models. In order to assess the performance of the models, 10-fold cross validation was applied to the same dataset [35]. The main measures of performance were sensitivity, specificity and accuracy under different regularization parameters. Sensitivity was defined as the fraction of known conotoxins predicted as conotoxin divided by the number of known conotoxins in the test dataset. Specificity was defined as the fraction of known non-conotoxins predicted as non-conotoxin divided by the number of known non-conotoxins in the test dataset. Accuracy was defined as the fraction of known sequences (conotoxin/non-conotoxin) predicted divided by the total number of sequences in the test dataset. The regularized parameter settings were chosen by plotting accuracy vs parameter settings for each model to make sure the trained model has the best accuracy with minimum overfitting/under fitting. Since the prevalence of conotoxins in the training dataset is only 9.97%, sensitivity is an important performance measure in consideration when choosing regularization parameter settings. The sensitivity, specificity and accuracy for the chosen regularization parameter settings for each

model is shown in **Table 1**. The highest overall testing accuracy and sensitivity were 98.2% and 90.93%, respectively, achieved by the LabelSpreading model.

Table 1. Maximized sensitivity, specificity and accuracy for chosen regularization parameter settings for three machine learning models in 10-fold cross validation.

Machine learning model	Performance measure		
	Sensitivity	Specificity	Accuracy
Logit	82.85%	99.30%	97.78%
LabelSpreading	90.93%	99.07%	98.32%
Perceptron	83.24%	97.65%	96.32%

2.2. Benchmark by identifying known superfamilies

To assess the sensitivity of *Conus*Pipe in identifying conotoxin transcripts, we performed a benchmark analysis for sequences belonging to known conotoxin gene superfamilies [6]. For these analyses, we deleted an entire superfamily from the training set and then tried to re-discover this superfamily as a putative new superfamily using *Conus*Pipe. The specificity of *Conus*Pipe (*i.e.*, the ability to distinguish between known conotoxins and non-conotoxin proteins from various organisms) was assessed by screening hits against the entire UniProtKB/Swiss-Prot database.

We found that the sensitivity varied among different superfamilies and combinations of models used. The highest sensitivity was achieved by the union of the 4 methods (logit, LabelSpreading, perceptron and Blastp, mean sensitivity = 95.7%, SD = 0.11) and the union of 3 methods (logit, perceptron and Blastp or LabelSpreading, perceptron and Blastp, mean sensitivity = 95.7%, SD = 0.11) across all superfamilies. A union of methods means that a conotoxin predicted by one or more methods is a putative positive. A graphical overview of the different groups is provided in **Figure 2**. Results are shown in **Supp. Figure 1** and **Tables 1** and **2.**

In addition to machine learning, using cross-species Blastp to search candidate sequences from different *Conus* species against each other provides better performance for conotoxin superfamilies that contain sequences which don't satisfy the hypothesis of having all 3 traits (signal sequence, propeptide and mature toxin at the C-terminus), such as the SF-mi2, I4, MEFRR, B4 and Prohormone gene families. However, this approach is less powerful for superfamilies that are limited to a small number of species, such as conorfamides, DivMTFLLLLVSV, MEVKM, MTSTL, Teretoxins and Conocaps. The sensitivity for recovering all known conotoxin superfamilies by different combinations of methods is provided in **Supp. Table 1**.

The ability of *Conus*Pipe to distinguish between conotoxins and other proteins from various organisms was assessed by screening the entire UniProtKB/Swiss-Prot database. Using the version released on June 2013 we examined a total of 540,261 protein sequences isolated from diverse organisms. The overall highest specificity was achieved using an overlap of the 4 methods (logit, LabelSpreading, Perceptron and Blastp, specificity = 99.92%) and the overlap of 3 methods (LabelSpreading, Perceptron and Blastp, specificity = 99.92%). These results are shown in **Table 2**.

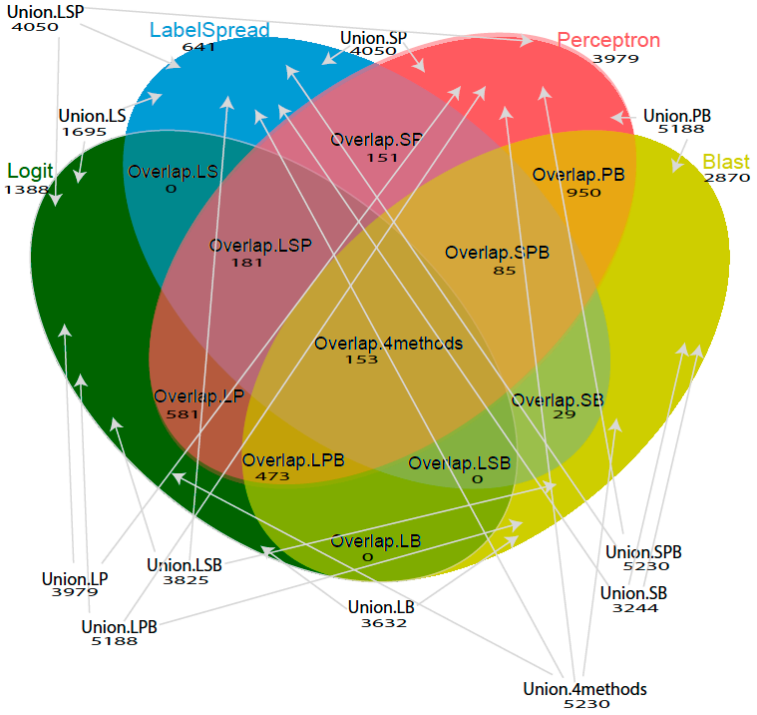


Figure 2. Venn diagram illustrates the different combinations of methodologies used (single method, overlap or union of methods). A union of methods means that a conotoxin is predicted by one or more methods, for example, Union.4methods = predicted by Perceptron or Logit or Label Spreading or Blast. An overlap of methods means that the conotoxin is predicted by all the applied methods, for example, Overlap.4methods = predicted by Perceptron and Logit and Label Spreading and Blast. Abbreviations used: Blast – B, Logit – L, Label Spreading – S, Perceptron - P. Numbers indicate putative novel toxins identified by *Conus*Pipe.

Table 2. The specificity of the individual machine learning methods and their unions/combinations when searching results against the Uniprot/Swissprot non-conotoxin database and mean sensitivity for recovering all known conotoxin superfamilies.

Method	Mean sensitivity	Specificity
Overlap.4methods	34.19%±0.32	99.92%
Overlap.LSP	41.53%±0.35	99.90%
Overlap.LSB	35.15%±0.32	99.87%
Overlap.LPB	76.25%±0.37	99.57%
Overlap.SPB	34.22%±0.32	99.92%
Overlap.LS	43.18%±0.34	99.85%
Overlap.LP	83.61%±0.32	99.53%
Overlap.SP	41.68%±0.35	99.90%
Overlap.LB	79.57%±0.35	98.32%
Overlap.SB	35.52%±0.32	99.86%
Overlap.PB	78.02%±0.36	99.57%
Logit	87.61%±0.26	99.83%
LabelSpreading (SemiS)	43.67%±0.34	99.49%
Perceptron (NeuroNetWork)	85.96%±0.29	94.02%
Blastp	87.10%±0.28	98.19%
Union.4methods	95.73%±0.11	93.89%
Union.LSP	90.31%±0.22	98.15%
Union.LSB	95.25%±0.12	93.89%
Union.LPB	95.73%±0.11	93.90%
Union.SPB	95.73%±0.11	93.97%
Union.LS	88.10%±0.26	98.17%
Union.LP	89.96%±0.23	98.17%
Union.SP	87.95%±0.25	99.41%
Union.LB	95.14%±0.12	93.90%
Union.SB	95.24%±0.12	93.99%
Union.PB	95.05%±0.15	93.98%

2.3. Identification of new conotoxin candidates

To identify new conotoxin candidates, we assembled *Conus* transcripts from RNA-seq datasets derived from venom glands of 10 different *Conus* species using our previously published methods [4]

(see Table 3 for species used in this study). The resulting transcripts were prescreened for conotoxin homology against the Uniprot and Conoserver databases as previously published [4] and described under the methods section. The remaining transcripts were then used as inputs to *Conus*Pipe. New conotoxin candidates were defined as those which were predicted as conotoxins by at least one of the 4 models in *Conus*Pipe, but lacked significant homology to known conotoxins using Blast against the Uniprot database [17].

Since conotoxin gene superfamilies are generally found across multiple *Conus* species (hence, the term superfamily) we considered those sequences that lacked significant homology to known conotoxins but had high homology (Blastp e-value <1e-10) to sequences from at least two other *Conus* species examined here, as members of a new putative superfamily. 5,455 transcripts passed these criteria. In order to validate our predictions, we used NCBI-Blastp to search the 5,455 transcripts against the NCBI non-redundant (NR) protein database (August 2018 version), which includes recently published conotoxin sequences that were not yet available in Uniprot/conoserver at the time of original analysis (and even now) and also includes large numbers of uncharacterized molluscan sequences not available in Uniprot. Out of 5,455 transcripts, 225 had significant Blastp hits (e-values < 1e-4) against the NCBI-NR database. 92 transcripts had hits against other molluscan transcripts. As the majority of conotoxins are not found outside of the genus *Conus* and these transcripts could encode endogenous signaling/housekeeping polypeptides rather than polypeptides used for envenomation, these were removed from our final datasets. 109 sequences were identified as conotoxins. These were also removed from the final machine learning dataset and are provided in **Supp. File 1** ("sequences.with.Blastp.hits.to.ncbi.nr.conotoxins"). Finally, 24 sequences had Blastp hits against non-molluscan species such as fish, tardigrade, sea anemone, worm, plant, bird. These were removed from the final dataset.

A total of 5,230 sequences were left in our final dataset as new conotoxin candidates, 153 of these were identified by all 4 models, 739 of these were identified by 3 models, 1,711 of these were identified by 2 models, 2,627 of these were identified by 1 model (**Figure 1**). All of the 5,230 transcripts are provided in **Supp. File 2** ("total.predicted.MLLabel.rmNrHit.rmImpCo.label.pep"). Using single linkage cluster analysis with a Jaccard Index [36,37] of 0.5 grouped 301 of sequences identified by at least 3 models into 107 clusters that are likely to represent novel conotoxin gene superfamilies (**Supp. File 3,** "Superfamily.cluster.0.5.seq").

The highest number of putative new sequences were identified in *C. striatus*. Blastp against Uniprot/Conoserver identified 99 toxin transcripts in *C. striatus* and 68 of these were also retrieved using the machine learning method (*i.e.*, 21 of these were missed by machine learning, see discussion section below). 1,097 additional putative toxins were subsequently identified by machine learning, 7 of these were confirmed as toxins by Blastp against the NCBI-NR database (**Table 3**).

Table 3. Conotoxin candidates expressed in 12 samples from 10 *Conus* species identified by Blastp and machine learning (ML).

Conus species	No. of conotoxins identified by Blastp against Uniprot/Conoserver database	No. of conotoxins identified by Blastp against Uniprot/Conoserver database also retrieved using ML	No. of conotoxins identified by ML and subsequently identified as conotoxins by Blastp against NCBI	No. of conotoxin candidates identified by ML only
C. magus	119	78	24	980
C. striatus	99	72	32	1065
C. marmoreus	108	75	13	495
C. marmoreus2 [10]	80	55	6	83
C. textile	224	126	34	505
C. gloriamaris	145	82	26	518
C. imperialis	90	67	9	50
C. virgo	159	104	36	701
C. virgo2 [10]	96	65	3	60
C. terebra	154	116	15	374
C. coronatus [10]	286	206	15	72
C. ebraeus [10]	100	77	12	102

2.3. Transcripts identified using 3 or 4 out of 4 methods

In this study, we provide a list of all new conotoxin candidates from the venom gland transcriptomes of several cone snail species. We emphasise that these sequences require further experimental validation to confirm their designation as genuine conotoxins. This is discussed in more detail below. However, we propose that many of the transcripts identified by 3 out of 4 methods (739 sequences) or by a combination of all 4 methods (153 sequences) are likely to represent genuine conotoxin sequences since they were independetly identified using different approaches. Several of these sequences exhibit clear hallmarks of conotoxins (presence of propeptides, found in multiple species, similar but distinct sequences found in different species, multiple cysteines in mature toxin region). An alignment of some of these is shown in Figure 3. Several sequences that were identified by 3 or more models but seem unlikely to represent genuine conotoxins are also shown. These typically exhibit a long series of cysteine repeats, several vicinal cysteines and/or series of methionines (M) in the signal sequence. In future our model could be further refined to exclude such sequences as candidates.

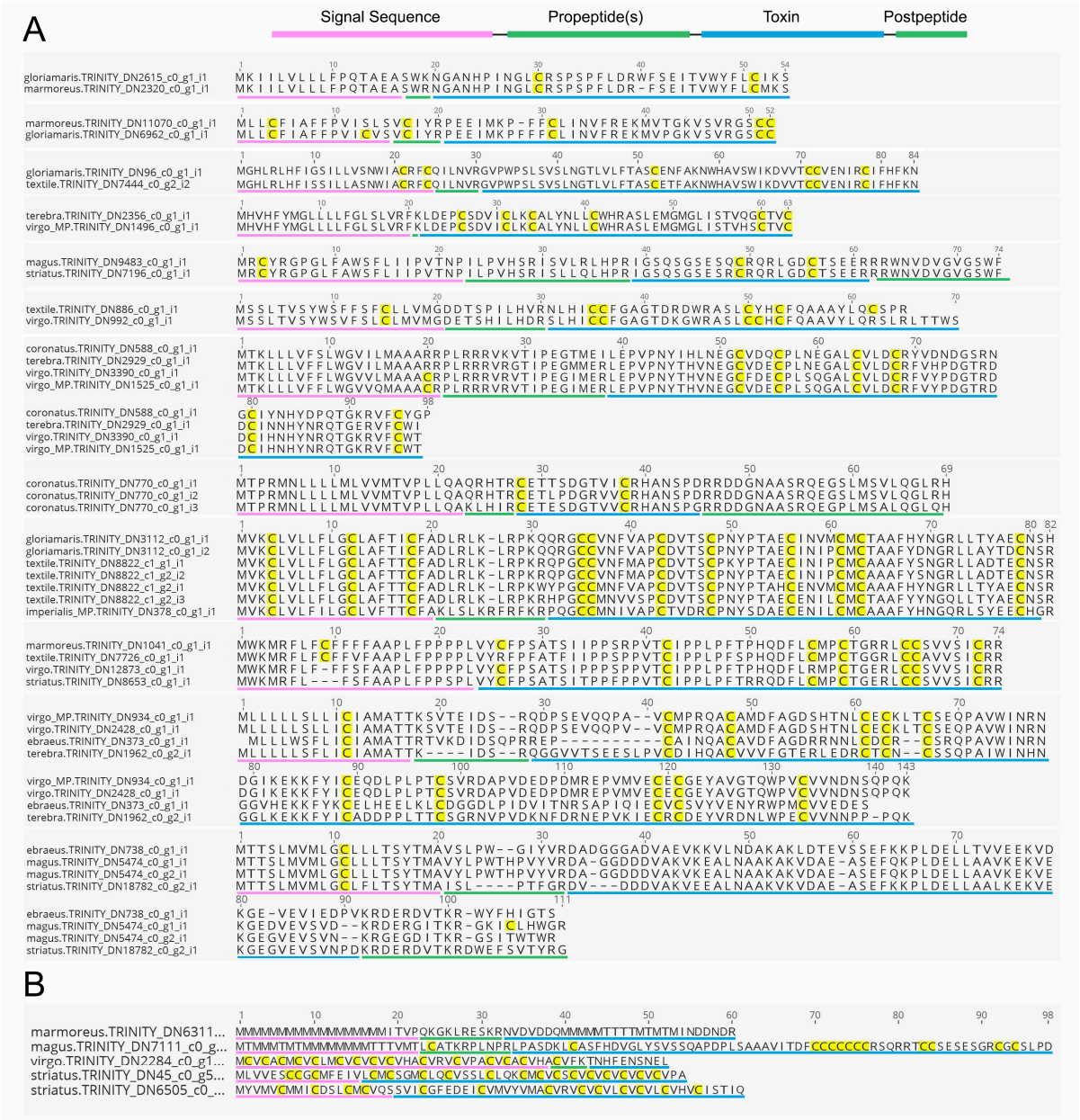


Figure 3. Comparative alignments of selected sequences identified by at least 3 out of 4 methods that are (**A**) likely or (**B**) not likely to represent genuine novel conotoxins. Cysteines are highlighted in yellow, signal sequences, pro- and postpeptides and predicted mature toxins are underlined in purple, green and blue, respectively, as shown on top of panel A. Sequence labels (contigs) correspond to those provided in **Supp. File 2**.

3. Discussion

Previous approaches for toxin discovery have been alignment based, using regular expressions, Blast and/or HMMER to identify new members of known conotoxin superfamilies [5,15-17] . Because conotoxins are hyperdiverse, these approaches are intrinsically limited. In an attempt to cast a wider net for discovery, we have created a machine learning based pipeline, *Conus*Pipe, that utilizes functional characteristics of conotoxins to identify new conotoxin candidates that have no significant sequence homology to conotoxin sequences currently available in reference databases.

By using more than one machine learning model, we expected to see that an ensemble of different models can maximize the prediction power. Indeed, as determined by benchmark analyses, the highest sensitivity is achieved by the union of 3 or more methods and the highest specificity is achieved in the overlap of 3 or more methods.

*Conus*Pipe allows users to choose different combinations of methods according to their requirement on discovery specificity and sensitivity (**Table 1**). In addition to developing machine learning, we demonstrate that using Blast to search candidate sequences from different *Conus* species against each other provides better performance for conotoxin superfamilies that contain sequences that don't satisfy the hypothesis of having all 3 traits. However, this approach is less powerful for superfamilies that are limited to a small number of species and only works if more than one transcriptome is to be analyzed.

We would like to note that, for best performance, *Conus*Pipe should be used in combination with standard Blast-based annotation methodologies as *Conus*Pipe relies on the presence of full-length sequences (containing N-terminal signal sequences) and will not work for truncated contigs. Furthermore, several conotoxin gene superfamilies reported in Uniprot/Conoserver have low SignalP values (D value of < 0.45). These would also be missed by our pipeline. Table 1 provides information on how many conotoxins which have significant homology to sequences in Uniprot/Conoserver database could also be retrieved by *Conus*Pipe (average value for recovery: 68 %, ranging from 56 % for *C. textile* – 77 % for *C. ebraeus*).

We provide a large set of new conotoxin candidates and a bioinformatic pipeline that is freely available and can be applied to any newly sequenced venom gland. No doubt our approach also identifies false positives, particularly when only a single or a combination of two methods is employed. Thus, using our method without further validation is not suitable for defining the venom composition of a species. What it provides is a new tool to identify candidate toxin transcripts that are not able to be detected by homology-based methods. Generating comprehensive databases of all putative toxin candidates expressed in a venom gland will empower current mass spectrometric toxin sequencing approaches [7]. Using mass spectrometry, candidate transcripts can then be verified and subjected to functional characterizations.

4. Conclusions

Using *Conus*Pipe, we identified 5,230 new conotoxin candidates from 1,359,647 transcripts derived from venom gland transcriptomes of 10 *Conus* species. None of these candidate conotoxins has significant homology to any known conotoxin in the Uniprot database, although like known conotoxins, most candidates have an N-terminal signal sequence, a characteristic propeptide spacer region and a single copy of a mature peptide at the C-terminus. Moreover, we have shown that several of these candidates share high homology to newly published conotoxins in the NCBI-NR molluscan database. In conclusion, our approach opens new avenues for the discovery of novel contoxin transcripts from cone snails and other venomous animals with similar venom repertoires.

5. Materials and Methods

5.1. Transcriptome sequencing

Total RNA was isolated from venom glands using *TRIzol® Reagent (*Life Technologies) or the RNeasy kit (Qiagen) following the manufacturers' instructions. RNA integrity, quantity and purity were determined on a 2100 Bioanalyzer (Agilent Technologies). cDNA libraries were prepared and sequenced on an Illumina HiSeq 2000 instrument (Sanger/Illumina 1.9 reads, 101 bp or 125 bp paired-end). Publically available Illumina datasets were used for the venom gland transcriptomes of *C. marmoreus* (specimen 2), *C. virgo* (specimen 2), *C. coronatus* and *C. ebraeus* [10].

Adapter clipping and quality trimming of raw reads were performed using fqtrim software (version 0.9.4, http://ccb.jhu.edu/software/fqtrim/) and PRINSEQ (version 0.20.4 [38]). After processing, sequences shorter than 70 bps and those containing more than 5% ambiguous bases (Ns) were discarded. De novo transcriptome assembly was performed using Trinity version 2.0.5 [39] with

a kmer size for building De Bruijn Graphs of 31, a minimum kmer coverage of 10 and a minimum glue of 10. Assembled transcripts were annotated using Blastx ((NCBI-Blast-2.2.28+, [40]) against conotoxin sequences extracted from the Conoserver [5] and UniProt databases [34].

5.2. Development of Conus*Pipe*

*Conus*Pipe proceeds by first extracting 16 features (see feature explanation below) from known conotoxin sequences (from the ConoServer [6] and Uniprot databases) and non-conotoxin *Conus* transcripts to make the training dataset. Sequences are required to contain a signal sequence as determined by SignalP [41]. Next, the 16 features extracted from *Conus* transcripts of 10 species, which have no homologues in current reference database (the combined ConoServer and UniProtKB database) are used as real test data to run the machine learning methods to predict whether a certain input transcript is conotoxin. All the transcripts predicted by any of the 4 method are output by the pipeline as putative new conotoxins.

5.2.1 Feature Selection and Classifiers

We employed 16 features for the machine learning models: signalP D value for signal sequence; cysteine percentage; molecular weight; percentage of positively/negatively charged amino acids; and isoelectric point for all three regions of the precursor sequence (signal sequence, propeptide and mature toxin). All features are continuous re-normalized to lie between 0 and 1 (**Figure 1A**). The features are mainly chemical characters of amino acids in the three different parts of conotoxin sequence. The motivating hypothesis is that even though conotoxins evolve very rapidly they must still share similar chemical characters in amino acid composition, since they carry out similar functions, e.g. bind transporters and receptors. For example, the three parts of conotoxin sequence - signal sequence, propeptide and mature toxin carry out different functions in conotoxin secretion process in the cell, so their charge distributions are stereotypical and different. The signal sequence is mainly hydrophobic, while the amino acids in propeptide are mainly charged, and the mature toxin is somewhat intermediate as regards charge distribution. To train the models we first ran signalP on a training dataset consisting of known conotoxin/nonconotoxin sequences to get the signalP D value for each known sequence (Petersen et al. 2011). Next, we calculated the cysteine percentage, molecular weight, percentage of positively/negatively charged amino acids and isoelectric point for signal sequence, propeptide and mature toxin, respectively. The pipeline uses the 16 extracted features in training dataset to train the logistic regression model, LabelSpreading model and Perceptron model using the Python scikit learn package (Pedregosa et al. 2011). The accuracy under different regularization parameters of the three models were tested by cross validation with training dataset from known conotoxin and nonconotoxin sequences.

5.2.2. Cross validation

4,950 known conotoxin sequences (from ConoServer and Uniprot/Swissprot) and 5,2613 non-conotoxin *Conus* transcripts with matched sequence length were first split into 10 equal bins, and then sequentially 1 tenth of the data was taken as the test set and the remaining other 9 tenths were used for the training set. The three models were trained under different regularization parameters and evaluated with the test set in 10 iterations. For the logistic regression model, the regularization parameter we tested is slack number C, which is the inverse of regularization strength. . We tested this parameter from 0.001 to 10**10. For the LabelSpreading model, we chose knn as the kernel function, so the regularization parameter we tested is n_neighbors, which was tested from 1 to 15. For the Perceptron model, the regularization parameter we tested is n_iter, the number of passes over the training data, which we tested from 5 to 70. The plots of accuracy versus regularization parameter of different models are shown in **Supp. Figure 2**.

5.2.3. Discovering new putative conotoxins and conotoxin gene families

Paired-end RNAseq data from 10 Conus species were generated by Illumina HiSeq 2000 platform (Table 4). RNAseq reads were assembled using best practice Trinity settings, annotated with Blastx against our reference dataset, and all the *Conus* transcripts which do not have homologous sequences in the current reference databases were selected and 6-frame translated into peptide sequences. These peptide sequences were then used as the input dataset for *Conus*Pipe to drive machine learning models built in previous steps to predict whether they are conotoxins. Cross-species Blastp is also used on all putative peptide sequences that have a signal sequence as an independent method to predict putative conotoxin candidates. The pipeline output all input transcripts which were predicted as conotoxins.

We then used NCBI- Blastp to search all putative toxin transcripts against the NCBI non-redundant (NR) protein database (August 2018 version), which includes recently published conotoxin sequences that were not yet available in Uniprot/conoserver at the time of original analysis. Sequences with significant Blastp hits were excluded from final datasets (e-values < 1e-4).

Then all by all Blastp and single linkage cluster analysis were conducted among the new conotoxins, and the new conotoxins that shared at least 50% hit connections with one another were designated as to be in the same superfamily.

Table 4. RNAseq Data sets from 12 samples in 10 *Conus* species used in the discovery pipeline.

Conus Species	Illumina HiSeq 2000 Number of reads	Read length (nt)
C. magus	85,877,500	101
C. striatus	101,170,402	101
C. marmoreus	53,901,510	125
C. marmoreus2 [10]	50,652,396	101
C. textile	63,365,620	125
C. gloriamaris	28,783,428	125
C. imperialis	30,784,548	125
C. virgo	30,038,902	125
C. virgo2 [10]	31,056,732	125
C. terebra	31,180,460	125
C. coronatus [10]	27,927,952	125
C. ebraeus [10]	19,556,244	125

Author Contributions: "Conceptualization, Q.L.; Methodology, Q.L. and H.S.; Software, Q.L.; Validation, Q.L., M.W., H.S.; Formal Analysis, Q.L.; M.Y., and H.S..; Writing-Original Draft Preparation, Q.L. and H.S.; Writing-Review & Editing, M.Y. and H.S.; Visualization, Q.L. and H.S.; Supervision, H.S. and M.Y.; Funding Acquisition, M.Y and H.S."

Funding: This work was supported in part by a National Institutes of Health Grants 1RO1GM122869 (to H.S.-H.).

Acknowledgments: The authors like to thank the High Throughput Genomics Facility at the Huntsman Cancer Institute in Salt Lake City, Utah.

Conflicts of Interest: The authors declare no conflict of interest.

418 **Appendix**

419 **Supporting Table 1.** The sensitivity for recovering all known conotoxin superfamilies by different combinations
420 of methods.

421 **Supporting File 2.** *Conus*Pipe discovered 109 putative novel conotoxins confirmed by Blastp against the
422 NCBI NR database.

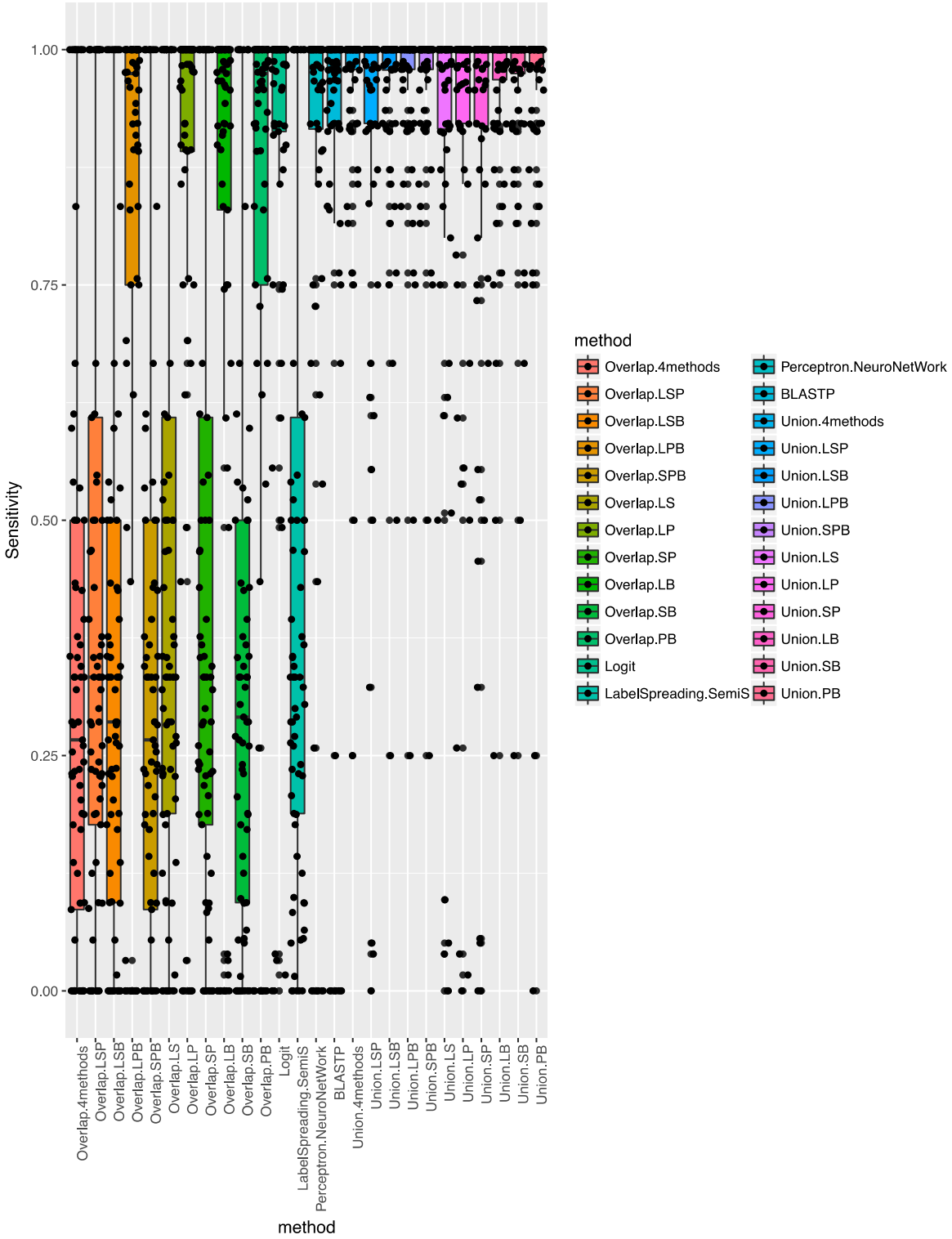
423 **Supporting File 1.** *Conus*Pipe discovered 5,230 putative novel conotoxins with no significant sequence
424 homology in any current reference database (Unipro/Conoserver/NCBI NR).

425 **Supporting File 3.** Cluster analysis of putative new gene superfamilies.

426

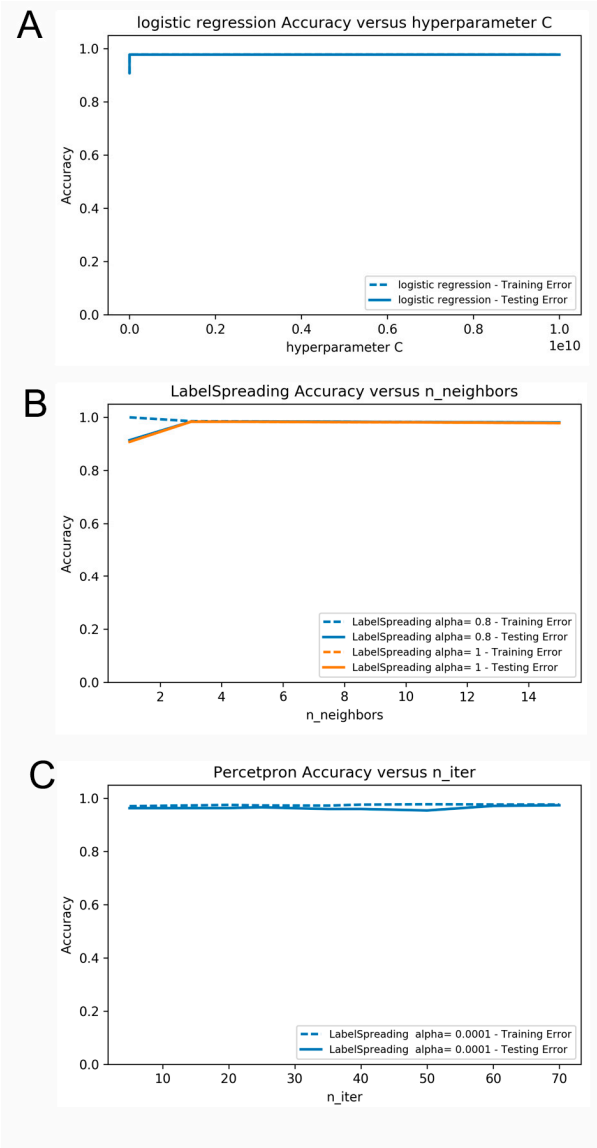
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2018 doi:10.20944/preprints201809.0587.v1

Peer-reviewed version available at Toxins 2018, 10, 503; doi:10.3390/toxins10120503



Supporting Figure 1. Box plot shows that the sensitivity varied among different combinations (single method, overlap or union of methods) of methods used. The union of different methods can achieve higher sensitivity. Union of methods means that the conotoxin is predicted by one method or another, as described in the caption for **Figure 1**.

433



434

435 **Supporting Figure 2**. Plot of accuracy vs regularization parameter settings in each machine learning model. (A)
436 Accuracy vs hyper-parameter C in logistic regression model. No overfitting/under fitting was observed. When
437 choosing C= 10**10, the logit model achieved the best accuracy and sensitivity. (B) Accuracy vs hyper-parameter
438 n_neighbors in LabelSpreading model. No overfitting/under fitting was observed. When choosing n_neighbors=
439 3 and alpha=1, the LabelSpreading model achieved the best accuracy and sensitivity. (C) Accuracy vs hyper-
440 parameter n_iter in Perceptron model. Overfitting was observed between n_iter =12 and n_iter=25, between
441 n_iter =30 and n_iter=50. When choosing n_iter= 5, the Perceptron model achieved the best accuracy and
442 sensitivity.

443

444

445

References

1. Shen, G.S.; Layer, R.T.; McCabe, R.T. Conopeptides: From deadly venoms to novel therapeutics. *Drug Discov Today* **2000**, *5*, 98-106.
2. McIntosh, J.M.; Jones, R.M. Cone venom--from accidental stings to deliberate injection. *Toxicon* **2001**, *39*, 1447-1451.
3. Livett, B.G.; Gayler, K.R.; Khalil, Z. Drugs from the sea: Conopeptides as potential therapeutics. *Curr Med Chem* **2004**, *11*, 1715-1723, doi:Doi 10.2174/0929867043364928.
4. Li, Q.; Barghi, N.; Lu, A.P.; Fedosov, A.E.; Bandyopadhyay, P.K.; Lluisma, A.O.; Concepcion, G.P.; Yandell, M.; Olivera, B.M.; Safavi-Hemami, H. Divergence of the Venom Exogene Repertoire in Two Sister Species of Turriconus. *Genome Biol Evol* **2017**, *9*, 2211-2225, doi:10.1093/gbe/evx157.
5. Kaas, Q.; Westermann, J.C.; Halai, R.; Wang, C.K.; Craik, D.J. ConoServer, a database for conopeptide sequences and structures. *Bioinformatics* **2008**, *24*, 445-446, doi:10.1093/bioinformatics/btm596.
6. Kaas, Q.; Westermann, J.C.; Craik, D.J. Conopeptide characterization and classifications: An analysis using ConoServer. *Toxicon* **2010**, *55*, 1491-1509, doi:10.1016/j.toxicon.2010.03.002.
7. Robinson, S.D.; Undheim, E.A.B.; Ueberheide, B.; King, G.F. Venom peptides as therapeutics: advances, challenges and the future of venom-peptide discovery. *Expert review of proteomics* **2017**, *14*, 931-939, doi:10.1080/14789450.2017.1377613.
8. Robinson, S.D.; Safavi-Hemami, H.; McIntosh, L.D.; Purcell, A.W.; Norton, R.S.; Papenfuss, A.T. Diversity of conotoxin gene superfamilies in the venomous snail, Conus victoriae. *PloS one* **2014**, *9*, e87648, doi:10.1371/journal.pone.0087648.
9. Robinson, S.D.; Li, Q.; Lu, A.; Bandyopadhyay, P.K.; Yandell, M.; Olivera, B.M.; Safavi-Hemami, H. The Venom Repertoire of Conus gloriamaris (Chemnitz, 1777), the Glory of the Sea. *Mar Drugs* **2017**, *15*, doi:10.3390/md15050145.
10. Phuong, M.A.; Mahardika, G.N.; Alfaro, M.E. Dietary breadth is positively correlated with venom complexity in cone snails. *BMC Genomics* **2016**, *17*, 401, doi:10.1186/s12864-016-2755-6.
11. Hu, H.; Bandyopadhyay, P.K.; Olivera, B.M.; Yandell, M. Elucidation of the molecular envenomation strategy of the cone snail *Conus geographus* through transcriptome sequencing of its venom duct. *Bmc Genomics* **2012**, *13*, doi:Artn 284 10.1186/1471-2164-13-284.
12. Barghi, N.; Concepcion, G.P.; Olivera, B.M.; Lluisma, A.O. Comparison of the Venom Peptides and Their Expression in Closely Related Conus Species: Insights into Adaptive Post-speciation Evolution of Conus Exogenomes. *Genome Biol Evol* **2015**, *7*, 1797-1814, doi:10.1093/gbe/evv109.
13. Buczek, O.; Bulaj, G.; Olivera, B.M. Conotoxins and the posttranslational modification of secreted gene products. *Cell Mol Life Sci* **2005**, *62*, 3067-3079, doi:10.1007/s00018-005-5283-0.
14. Olivera, B.M.; Safavi-Hemami, H.; Horvarth, M.P.; Teichert, R.W. Conopeptides, Marine Natural Products from Venoms: Biomedical Applications and Future Research Applications. In *Marine Biomedicine: From Beach to Bedside*, Baker, B.J., Ed. CRC Press 2015.
15. Koua, D.; Brauer, A.; Laht, S.; Kaplinski, L.; Favreau, P.; Remm, M.; Lisacek, F.; Stocklin, R. ConoDictor: a tool for prediction of conopeptide superfamilies. *Nucleic Acids Res* **2012**, *40*, W238-241, doi:10.1093/nar/gks337.
16. Lavergne, V.; Dutertre, S.; Jin, A.H.; Lewis, R.J.; Taft, R.J.; Alewood, P.F. Systematic interrogation of the Conus marmoreus venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene superfamilies. *BMC Genomics* **2013**, *14*, 708, doi:10.1186/1471-2164-14-708.

17. Wheeler, T.J.; Eddy, S.R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **2013**, *29*, 2487-2489, doi:10.1093/bioinformatics/btt403.
18. Bandyopadhyay, P.K.; Colledge, C.J.; Walker, C.S.; Zhou, L.M.; Hillyard, D.R.; Olivera, B.M. Conantokin-G precursor and its role in gamma-carboxylation by a vitamin K-dependent carboxylase from a Conus snail. *J Biol Chem* **1998**, *273*, 5447-5450.
19. Conticello, S.G.; Kowalsman, N.D.; Jacobsen, C.; Yudkovsky, G.; Sato, K.; Elazar, Z.; Petersen, C.M.; Aronheim, A.; Fainzilber, M. The prodomain of a secreted hydrophobic mini-protein facilitates its export from the endoplasmic reticulum by hitchhiking on sorting receptors. *J Biol Chem* **2003**, *278*, 26311-26314, doi:10.1074/jbc.C300141200.
20. Buczek, O.; Olivera, B.M.; Bulaj, G. Propeptide does not act as an intramolecular chaperone but facilitates protein disulfide isomerase-assisted folding of a conotoxin precursor. *Biochemistry-Us* **2004**, *43*, 1093-1101, doi:10.1021/bi0354233.
21. Cox, D.R. The Regression-Analysis of Binary Sequences. *J Roy Stat Soc B* **1958**, *20*, 215-242.
22. Pollack, J.B. Perceptrons - an Introduction to Computational Geometry, Expanded Edition - Minsky,Ml, Papert,Sa. *J Math Psychol* **1989**, *33*, 358-365, doi:Doi 10.1016/0022-2496(89)90015-1.
23. Widrow, B.; Lehr, M.A. 30 Years of Adaptive Neural Networks - Perceptron, Madaline, and Backpropagation. *P Ieee* **1990**, *78*, 1415-1442, doi:Doi 10.1109/5.58323.
24. Delalleau, O.; Bengio, Y.; Le Roux, N. Efficient Non-Parametric Function Induction in Semi-Supervised Learning. In Proceedings of AISTATS; p. 100.
25. Yu, H.F.; Huang, F.L.; Lin, C.J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach Learn* **2011**, *85*, 41-75, doi:10.1007/s10994-010-5221-8.
26. Zhao, L.; Chen, Y.; Schaffner, D.W. Comparison of logistic regression and linear regression in modeling percentage data. *Appl Environ Microbiol* **2001**, *67*, 2129-2135, doi:10.1128/AEM.67.5.2129-2135.2001.
27. Belkin, M.; Niyogi, P. Semi-supervised learning on Riemannian manifolds. *Mach Learn* **2004**, *56*, 209-239, doi:DOI 10.1023/B:MACH.0000033120.25363.1e.
28. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am Stat* **1992**, *46*, 175-185, doi:Doi 10.2307/2685209.
29. Robinson, S.D.; Norton, R.S. Conotoxin gene superfamilies. *Mar Drugs* **2014**, *12*, 6058-6101, doi:10.3390/md12126058.
30. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput* **1989**, *1*, 541-551, doi:DOI 10.1162/neco.1989.1.4.541.
31. Hyvarinen, A.; Koster, U. Complex cell pooling and the statistics of natural images. *Network-Comp Neural* **2007**, *18*, 81-100, doi:10.1080/09548980701418942.
32. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *Ieee T Pattern Anal* **2015**, *37*, 1904-1916, doi:10.1109/Tpami.2015.2389824.
33. Zhou, L.J.; Li, Q.W.; Huo, G.Y.; Zhou, Y. Image Classification Using Biomimetic Pattern Recognition with Convolutional Neural Networks Features. *Comput Intel Neurosc* **2017**, 10.1155/2017/3792805, doi:10.1155/2017/3792805.
34. Consortium, U. UniProt: a hub for protein information. *Nucleic Acids Res* **2015**, *43*, D204-212, doi:10.1093/nar/gku989.
35. Picard, R.R.; Cook, R.D. Cross-Validation of Regression-Models. *J Am Stat Assoc* **1984**, *79*, 575-583, doi:Doi 10.2307/2288403.

36. Tan, P.N.; Steinbach, M.; Kumar, V. *Introduction to data mining*; Pearson Addison Wesley: Boston, 2005; Vol. 1.
37. Jaccard, P. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* **1901**, *37*, 241-272.
38. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863-864, doi:10.1093/bioinformatics/btr026.
39. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q., et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **2011**, *29*, 644-652, doi:10.1038/nbt.1883.
40. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **1990**, *215*, 403-410, doi:10.1016/s0022-2836(05)80360-2.
41. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **2011**, *8*, 785-786, doi:10.1038/nmeth.1701.