METHODOLOGY: Transcriptomic analysis of the black widow spider venom (*Latrodectus curacaviensis*)

Gabriel Alves Vieira

April 2020

Contents

List of Tables

List of Figures

1 Transcriptome sequencing and assembly

The Illumina HiSeq 1500 equipment (Illumina, San Diego, California, United States) was used for RNA sequencing, with ribosomal RNA reduction (rRNA) reagents and large-scale sequencing. After the sequencing platform generated the sequencing images, the CASAVA software (version 1.8.2) was used to separate samples by identifying the barcodes assigned during the cDNA library preparation step. The final step was the conversion from BCL files to fastq format files that contains only nucleotide sequences and their respective qualities, discarding reads that have not been approved by the Q score base quality control (greater than phred 30). The reads were assembled using Trinity software [1] with a minimum contig size of 120 (instead of the default value: 200). Trinity data allowed us to identify the number of isoforms per gene and other features.

2 Bioinformatics analyses

2.1 Transcript annotation

Three different methodologies were used to annotate the contigs generated by Trinity:

1. Alignment using BLASTx to several databases, including: (i) UNIPROT (download on July, the 27 th , 2017 UniprotMetazoaToxin), (ii) eukaryotic COGs, (iii) Arachnoserver [2], and (iv) an subset of UNIPROT including only sequences from the clade Arthropoda. The contigs were compared to these databases BLASTx to provide sequence similarity information. Only the hits an e-value inferior to 1e-5 were considered and we used the best hit against each database to provide a putative annotation for each contig. The contigs were also classified according to a ranking of expression as observed by TPM (transcripts per million transcripts) measures provided by Trinity.

   We also performed a specific ranking containing only the contigs annotated as toxins by Arachnoserver best hits. All mRNA transcripts for the spider genus Latrodectus available at Genbank were downloaded and converted into a BLAST database. Our transcripts were aligned against this database using BLASTn 2.9.0 with an evalue of 10.

2. Submission of sequences to the Trinotate annotation pipeline v3.2.1 [3] using the autoTrinotate perl script. This pipeline has already been succesfully used to annotate venom transcripts from several taxa, such as cnidarians [4, 5], stingrays [6], lizards [7], scorpions [8–10] and spiders [11].

In our runs of the Trinotate pipeline, Transdecoder (https://github.com/TransDecoder/TransDecoder/wiki), a companion utility for Trinity [12], was used to extract and translate ORFs from the assembled contigs. Blast 2.9.0 searches were run using the entire transcripts (BlastX) and ORF aminoacid sequences (BlastP) against all Arthropoda sequences from the Swissprot database [13] (April 2020). HMM searches against the PFAM 32.0 database [14] were performed using Hmmer 3.2.1 (http://hmmer.org/) for protein domain identification. Specific annotation was also performed for signal peptides with signalP 4.0 [15], Transmembrane domains with TMHMM 2.0c [16] and rRNAs with RNAmmer 1.2 [17]. Lastly, Trinotate provides annotation from GO [18], eggNOG [19] and KEGG [20] databases to contigs. Results were integrated into a sqlite database, used to generate a .xls tab-separated report.

Trinotate was run with the default parameters specified in autoTrinotate.pl configuration file, except for Transdecoder, where the `-m` parameter was used to extract ORFs longer than 30 aminoacids from our assembled contigs:

```
TransDecoder.LongOrfs -t nt.fasta -m 30 > protein.fasta
```

```
This custom parameter was necessary because by default Transdecoder will only keep ORFs
```

3. After running autoTrinotate, we have also performed a custom Blast searches search against mRNA sequences from *Latrodectus* spp. available at Genbank. These sequences were downloaded in April 2020 and cleaned to avoid redundancy with CD-HIT 4.8.1 [21]. The `-c 1` parameter was used to ensure that only duplicate sequences were removed, reducing the total number of sequences from 1447 to 1403. This dataset

was used to generate both a nucleotide and aminoacid Blast database. The aminoacid sequences were translated from the mrna dataset using Transdecoder with the same `-m 30` specification used in Trinotate. All our Trinity contigs were aligned to the genbank *Latrodectus* transcripts using Blastn, BlastP and BlastX with an e-value of 10e-5. BlastP and BlastX were incorporated into the Trinotate sqlite database as custom results using the following commands:

```
Trinotate Trinotate-master.sqlite LOAD_custom_blast --outfmt6 Trinity_vs_latrodec_m
--prog blastx --dbtype MRNA_LATRODECTUS_AA
```

```
Trinotate Trinotate-database.sqlite LOAD_custom_blast --outfmt6 Trinity_vs_latrodec
--prog blastp --dbtype MRNA_LATRODECTUS_AA
```

And a new .xls report based on the updated database was generated, including these custom blast hits. `--incl_trans` and `--incl_pep` parameters were used to respectively add nucleotide and aminoacid sequences data to the final Trinotate table:

```
Trinotate Trinotate-database.sqlite report --incl_trans --incl_pep >
trinotate_annotation_report_full.xls
```

A custom python script was used to merge the results of these approaches (BlastX against several databases, Trinotate results and Blast against *Latrodectus* mrna sequences) into a single, massive table.

As a complementary analysis, we have run Trinotate against a custom database with all toxins deposited in Swissprot. Aside from this change in the database for Blast searches, the program was run with the same parameters as in the main annotation process. For both Trinotate runs (annotation against arthropoda and toxins sequences from Swissprot), a support script packaged with Trinotate (trinotate_report_summary.pl) was used to generate result summaries, both in textual and graphical formats.

REFERENCES

1. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;29:644–52.

2. Pineda SS, Chaumeil PA, Kunert A, Kaas Q, Thang MWC, Le L, et al. ArachnoServer 3.0: An online resource for automated discovery, analysis and annotation of spider toxins. Bioinformatics. 2018.

3. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Reports. 2017;18:762–76. doi:10.1016/j.celrep.2016.12.063.

4. Lewis Ames C, Ryan JF, Bely AE, Cartwright P, Collins AG. A new transcriptome and transcriptome profiling of adult and larval tissue in the box jellyfish Alatina alata: An emerging model for studying venom, vision and sex. BMC Genomics. 2016;17. doi:10.1186/s12864-016-2944-3.

5. Mitchell ML, Tonkin-Hill GQ, Morales RAV, Purcell AW, Papenfuss AT, Norton RS. Tentacle Transcriptomes of the Speckled Anemone (Actiniaria: Actiniidae: Oulactis sp.):

Venom-Related Components and Their Domain Structure. Marine Biotechnology. 2020;22:207–19. doi:10.1007/s10126-020-09945-8.
6. Júnior NGDO, Fernandes GDR, Cardoso MH, Costa FF, Cândido EDS, Neto DG, et al. Venom gland transcriptome analyses of two freshwater stingrays (Myliobatiformes: Potamotrygonidae) from Brazil. Scientific Reports. 2016;6. doi:10.1038/srep21935.
7. Lino-López GJ, Valdez-Velázquez LL, Corzo G, Romero-Gutiérrez MT, Jiménez-Vargas JM, Rodríguez-Vázquez A, et al. Venom gland transcriptome from Heloderma horridum horridum by high-throughput sequencing. Toxicon. 2020;180:62–78. doi:10.1016/J.TOXICON.2020.04.003.
8. Romero-Gutierrez T, Peguero-Sanchez E, Cevallos MA, Batista CVF, Ortiz E, Possani LD. A deeper examination of thorellius atrox scorpion venom components with omic techonologies. Toxins. 2017;9. doi:10.3390/toxins9120399.
9. Cid-Uribe JI, Santibáñez-López CE, Meneses EP, Batista CVF, Jiménez-Vargas JM, Ortiz E, et al. The diversity of venom components of the scorpion species Paravaejovis schwenkmeyeri (Scorpiones: Vaejovidae) revealed by transcriptome and proteome analyses. Toxicon. 2018.
10. Valdez-Velázquez LL, Cid-Uribe J, Romero-Gutierrez MT, Olamendi-Portugal T, Jimenez-Vargas JM, Possani LD. Transcriptomic and proteomic analyses of the venom and venom glands of Centruroides hirsutipalpus, a dangerous scorpion from Mexico. Toxicon. 2020.
11. Zobel-Thropp PA, Bulger EA, Cordes MHJ, Binford GJ, Gillespie RG, Brewer MS. Sexually dimorphic venom proteins in long-jawed orb-weaving spiders (Tetragnatha) comprise novel gene families. PeerJ. 2018;2018.
12. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494–512. doi:10.1038/nprot.2013.084.
13. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge The UniProt Consortium. Nucleic Acids Research. 2019.
14. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Research. 2019;47:D427–32. doi:10.1093/nar/gky995.
15. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. 2011;8:785–6. doi:10.1038/nmeth.1701.
16. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology. 2001;305:567–80. doi:10.1006/jmbi.2000.4315.
17. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research. 2007;35:3100–8.
18. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004;32:258D–261. doi:10.1093/nar/gkh036.
19. Jensen LJ, Julien P, Kuhn M, Mering C von, Muller J, Doerks T, et al. eggNOG: Automated construction and annotation of orthologous groups of genes. Nucleic Acids Research. 2008;36 SUPPL. 1. doi:10.1093/nar/gkm796.
20. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28:27–30. doi:10.1093/nar/28.1.27.

21. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.