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

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. Transdecoder (https://github.com/TransDecoder/TransDecoder/wiki), a companion utility for Trinity [4], 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 the Swissprot database [5]. HMM searches against the PFAM 32.0 database [6] 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 [7], Transmembrane domains with TMHMM 2.0c [8] and rRNAs with RNAmmer 1.2 [9]. Lastly, Trinotate provides annotation from GO [10], eggNOG [11] and KEGG [12] 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 longer than 100 aminoacids, and some proteins of interest for this study are below this threshold. Inhibitor cystine-knots (ICKs), for instance, are generally 30 to 50 residues long [13].

3. After running autoTrinotate, we have also performed a custom Blast searches search against mRNA sequences from *Latrodectus* spp. available at Genbank. All *Latrodectus* mrna sequences were downloaded in November 2019 and cleaned with CD-HIT 4.8.1 [14]. The `-c 1` parameter was used to ensure that only duplicate sequences were removed.

```
cd-hit -c 1 ...
```

Table 1: Number of sequences before and after cleanup

Raw	Cleaned
14067	14013

The difference in sequences is evidenced in Table 1

You can see *Latrodectus curacaviensis* morphology in Figure 1

```
knitr::include_graphics("cura.jpg")
```

Then, ORFs were extracted from the datasets using Transdecoder with the same `-m 30` specification used in Trinotate.

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. 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.
5. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge The UniProt Consortium. Nucleic Acids Research. 2019.
6. 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.
7. 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.
8. 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.
9. 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.
10. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004;32:258D–261. doi:10.1093/nar/gkh036.

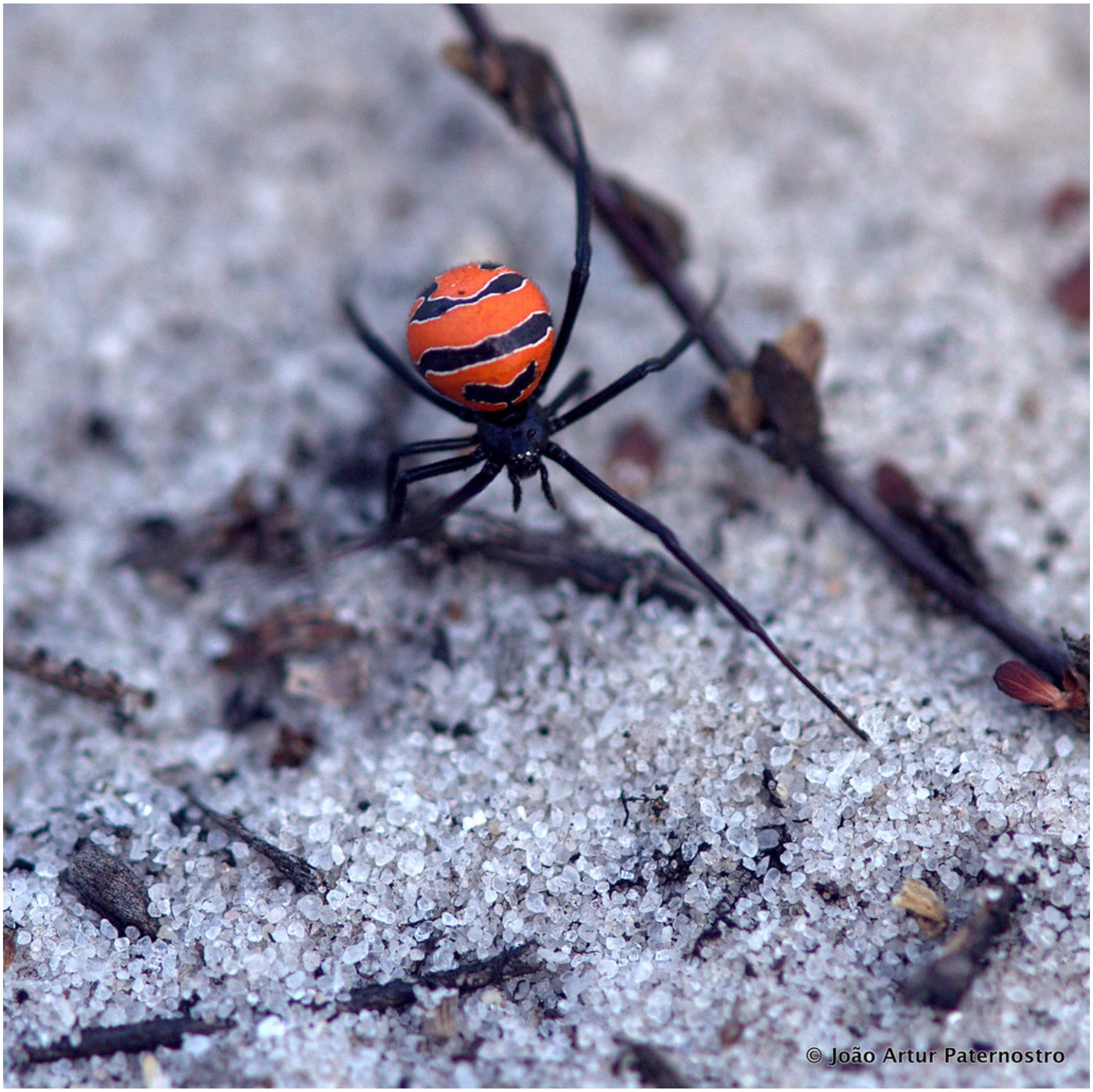

Figure 1: *L. curacaviensis* morphology

11. 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.

12. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28:27–30. doi:10.1093/nar/28.1.27.

13. Postic G, Gracy J, Périn C, Chiche L, Gelly JC. KNOTTIN: The database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. Nucleic Acids Research. 2018;46:D454–8. doi:10.1093/nar/gkx1084.

14. 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.