RNA-Seq on non-model organisms

Jorge Espinosa², Jordi Durban¹, Juan J. Calvete¹
1. Laboratori Venòmica Evolutiva i Trasllacional, Institut Biomedicina de València (CSIC)
2. Departament de Genètica, Universitat de València
email: jdurban@ibv.csic.es

Background

Venoms research has been continuously enhanced by advances in technology. In particular, the emergence of ‘omic’ technologies at the turn of the twenty-first century has revolutionized biological research. The breakthrough experienced by venom research in the last decade is due to the development and application of omics technology to the qualitative and quantitative profiling of the venom gland mRNA expression (venom gland transcriptomics) and the precise identification of the components expressed in the venom (venomics). Both approaches have been fueled by advances on sequencing technologies and improvements on RNA-Seq. However, in the absence of a genomic reference, several approaches have been developed in order to analyze mRNA expression data from Next Generation Sequencing technologies which are not as straightforward as those developed when a genome reference sequence is available.

**de novo RNA-Seq assembly**

**Quality check**

- Reads 1 Paired-end
- Reads 1 Single-end
- Reads 2 Paired-end
- Reads 2 Single-end

Trimming and adapter pre-processing procedure

- Reads with a quality score below 30, adapters, and the first 13 bases were eliminated using FastQC, TrimGalore and PRINSEQ.
- Overlapping short paired-end reads were extended with Fast Length Adjustment of Short reads tool.

**Assembly**

- OASES
- CAP3
- Trinity

**Annotation**

- Repbase
- Repeat masker
- Blast

Homology search annotation process

- Transcript obtained by the assemblers were filtered to eliminate repetitive elements.
- The used Blast to extract the information needed of each hit.
- Then each informative hit was filtered by the snakes and toxin keywords

**Quantification**

- Identified Contigs
- Raw reads
- Bowtie2
- Cd-hit

Quantification of transcript abundance

- Transcript abundance was estimated by mapping paired and unpaired trimmed reads back to annotated encoding transcripts with Bowtie2 and expressed in RPKM
- Molecular protein diversity was assessed by clustering translated sequences with cd-hit

Concluding remarks

Transcriptomics creates many challenges for scientists who are unfamiliar with working with RNA, managing large NGS datasets and executing the required programs, specially if no reference sequence is available. This work provides the methodology used at the Evolutionary and translational venomics lab for the bioinformatic analysis of targeted sequencing of transcripts from snake venom glands. The species-specific obtained transcriptome greatly improves the venom proteomic profiling which takes advantage from a custom protein database.