Looking for Secretory Calcium-Binding PhosphoProtein (SCPP) transcripts in jaw transcripts

Protocol for RNA-seq data cleaning, assembly, annotation and detection of candidate sequences

XI LIU,1,2 Erwan CORRE, Christophe CARON* and Jean-Yves SIRE*
1 Évolution Paris-Saclay, UMR7338, Institut de Biologie Paris-Saclay, Université Paris-Saclay, Marne la Vallée, 78140 Le Chesnay, France. 2 Département de Biologie, Université Paris-Saclay, 91405 Orsay, France. 
*Corresponding author (E-mail: christophe.caron@ib-pca.fr)

Introduction

Skeleton mineralization (bones and teeth) was a major event in the emergence of vertebrates. The Secretory Calcium-Binding PhosphoProtein (SCPP) family represents crucial molecules, which control mineralization of skeletal tissues. A total of 25 jaw transcripts were sequenced from species representative of the main lineages of non-mammalian vertebrates and one from the opossum Monodelphis domestica, in order to elucidate the origin and evolution of SCPPs (ANR project JAWS).

In this study, we established a protocol of data cleaning, assembly, annotation and analysis for the opossum’s transcriptome (for which a reference genome is available) to serve as the reference protocol for further analyses of the other transcripts.

Key-words: RNA-seq, assembly, annotation, jaw mineralization, SCPP.

Methods

Step 1: Preprocessing and assembly

![Illumina HiSeq 2500: 50 bp paired-end sequencing](Image)

- Adapters removing & quality trimming
- Reference genome assembly
- Assemble evaluation

Assembly metrics, reads remapping to transcripts, CEGMA\(^1\), DETONATE\(^8\)

Step 2: Annotations

- Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.
- Blast2PM: a tool for annotating proteins on the basis of their sequence similarity to a set of pre-annotated proteins.
- Ortho-SEED: a tool for finding orthologous groups of proteins.

Step 3: Contaminant detection

- Silva: a tool for identifying and classifying rRNA and tRNA sequences.
- Trinotate swiss/uniref: a tool for annotating proteins using non-redundant protein databases.
- NCBI Taxonomy: a tool for identifying and classifying rRNA and tRNA sequences.

Step 4: SCPPs identification

- Candidate gene: POMC, AMBN/ABM1, AMTN/AMTN, ISG15/ISG15, SCPP1/SCPP1
- ACID-IDS: OGPS/DMIPS/IPSIM/SCPP1

Conclusion & Perspectives

- Automatic methods are very efficient, but checking by an expert is necessary in some cases.
- For the future study, we will detect assembly chimera, test other assemblers de novo and find new characteristics for SCPPs identification.