Analysis of small non-coding RNAs of poorly annotated species in Galaxy with the help of ortholog information of well annotated species

Jochen Bick, Susanne E. Ulbrich and Stefan Bauersachs
ETH Zurich, Animal Physiology, Institute of Agricultural Sciences, Zurich, Switzerland

Introduction

The analysis of RNA-seq data with a basic analysis pipeline including quality control, filtering, trimming, and adapter clipping followed by mapping to a reference genome or transcriptome is a straightforward task using Galaxy. For processing of smallRNA-Seq data it is necessary to modify this analysis pipeline because the results do not correspond, at least in theory directly to small RNA. An additional common problem is that the number of annotated small non-coding RNAs is very low for certain species including pig and cattle which makes it very difficult to annotate smallRNA data in such species. Our workflow is mainly based on basic Galaxy tools and some own in-house scripts. The idea is to use well annotated related species information such as human to improve the annotation (mapping with BLASTn-short) of each sequence found in smallRNA-Seq results. The collection of BLAST databases contains sequences from miRBASE (precursor and mature microRNAs), sequences from NCBI and Ensembl, mostly non-coding RNAs but also protein-coding transcripts, as well as tRNA and piRNA cluster sequences. Finally, all BLAST results have to be filtered and joined by removing all duplicated hits. The annotated sequences are used for DEG analysis with EdgeR and/or DESeq2.

Data analysis workflow

Generate a count table for EdgeR

Join datasets by identifier column

Table with unique sequences and counts

Count table statistics + filtering

Generate annotation sources used for BLAST

Sequence statistics of smallRNA project EPI10

Annotation sources used for BLAST

Workflow

- Translate RNA to DNA
- Join FASTA files into groups
- Remove duplicates
- Create BLAST databases

---

<table>
<thead>
<tr>
<th>Library 1</th>
<th>Sequence</th>
<th>count1</th>
<th>count2</th>
<th>count3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence_1</td>
<td>32</td>
<td>8</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Sequence_2</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Sequence_3</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Sequence_4</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sequence_5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Library 2</th>
<th>Sequence</th>
<th>count1</th>
<th>count2</th>
<th>count3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence_1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sequence_2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Unique sequences

Raw reads

10.3% annotated

1.7% filtered

~117,000

~14,000

~12,000

~1,200