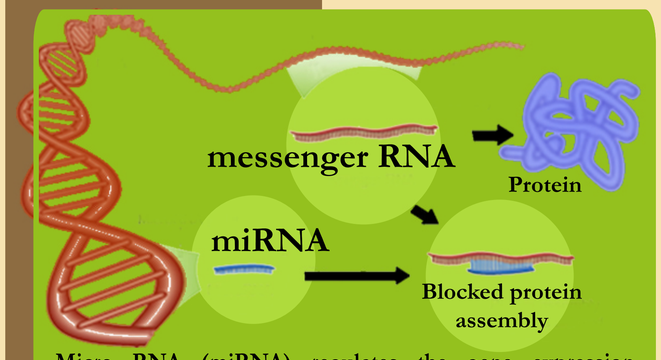
Quality Control and Visualization of miRNA data

Guillermo Carrasco, Sofia Bergström and Yim Wing Chow
guillermo.carrasco@scilifelab.se, sobergst@kth.se, ywchow@kth.se
In collaboration with Marc Friedländer and Phil Ewels

Introduction



Micro RNA (miRNA) regulates the gene expression and it is thus important to study. These small RNAs are around 22 nucleotides long and have the ability to silence targeted messenger RNAs which prevents them from folding properly into proteins. A single miRNA can target hundreds of different messenger RNAs.

113 samples from six different projects have been used in this project. HiSeq 2500 Illumina platforms were used to sequence those projects. Sequencing starts with a library preparation, followed by amplification and finally the actual sequencing part.

When you get miRNA data you need to know how many of the reads actually corresponds to miRNA. Other RNA types can be present or the reads can have a low quality, and are thus not of interest to study. In this project we have applied different types of sensible controls for miRNA data.

Aims

1. Data exploration: How does the miRNA data for these projects look like?
2. QC and visualisation: What are the peculiarities for miRNA data?

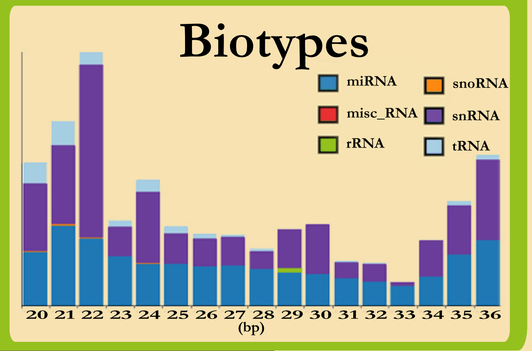
Method

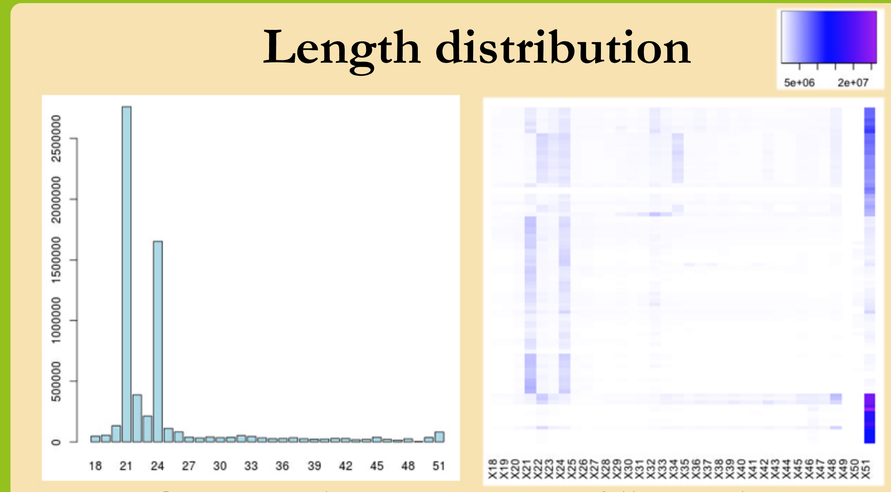
Six projects from SciLifeLab with a total of 113 samples have been used.

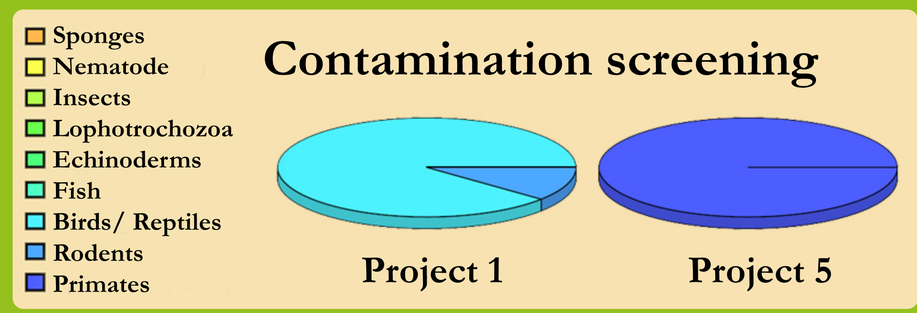
- Remove low quality reads
- Remove adapters
- Merge resulting FASTA files
- FastQC
- Contamination control

Results

The length distribution plots indicates that quite many read length are around 20-25 nucleotides, but there are even more reads that are around 51 nt. Other kinds of small RNA types are mixed within each sample. Project 1 mapped to birds and reptiles compared to most other projects that mapped to primates.







Conclusion

Quality control and data curation are very important steps prior to any analysis. Bad libraries or sequencing, contamination from other species or just wrong classification can lead to wrong conclusions. In this case, read filtering gave us a sense of the library quality, and contamination screening helped us realize that we were not only analyzing human data.

Source for introduction picture:
http://cnx.org/contents/45283c3f-c8cc-4659-94b5-d2faa9c2cf2f@2/Analysis_of_miRNA_and_mRNA_ass