ABSTRACT

High-throughput sequencing and small non-coding RNAs

David Langenberger

In this thesis the processing mechanisms of short non-coding RNAs (ncRNAs) is investigated by using data generated by the current method of high-throughput sequencing (HTS). The recently adapted short RNA-seq protocol allows the sequencing of RNA fragments of microRNA-like length (~18-28nt). Thus, after mapping the data back to a reference genome, it is possible to not only measure, but also visualize the expression of all ncRNAs that are processed to fragments of this specific length.

Short RNA-seq data was used to show that a highly abundant class of small RNAs, called microRNA-offset-RNAs (moRNAs), which was formerly detected in a basal chordate, is also produced from human microRNA precursors. To simplify the search, the blockbuster tool that automatically recognizes blocks of reads to detect specific expression patterns was developed. By using blockbuster, blocks from moRNAs were detected directly next to the miR or miR* blocks and could thus easily be registered in an automated way.

When further investigating the short RNA-seq data it was realized that not only microRNAs give rise to short ~22nt long RNA pieces, but also almost all other classes of ncRNAs, like tRNAs, snoRNAs, snRNAs, rRNAs, Y-RNAs, or vault RNAs. The formed read patterns that arise after mapping these RNAs back to a reference genome seem to reflect the processing of each class and are thus specific for the RNA transcripts of which they are derived from. The potential of this patterns in classification and identification of non-coding RNAs was explored. Using a random forest classifier which was trained on a set of characteristic features of the individual ncRNA classes, it was possible to distinguish three types of ncRNAs, namely microRNAs, tRNAs, and snoRNAs.

To make the classification available to the research community, the free web service ‘DARIO’ that allows to study short read data from small RNA-seq experiments was developed.

The classification has shown that read patterns are specific for different classes of ncRNAs. To make use of this feature, the tool deepBlockAlign was developed. deepBlockAlign introduces a two-step approach to align read patterns with the aim of quickly identifying RNAs that share similar processing footprints.
In order to find possible exceptions to the well-known microRNA maturation by Dicer and to identify additional substrates for Dicer processing the small RNA sequencing data of a Dicer knockdown experiment in MCF-7 cells was re-evaluated. There were several Dicer-independent microRNAs, among them the important tumor supressor mir-663a.

It is known that many aspects of the RNA maturation leave traces in RNA sequencing data in the form of mismatches from the reference genome. It is possible to recover many well- known modified sites in tRNAs, providing evidence that modified nucleotides are a pervasive phenomenon in these data sets.