RNAdeNoise

R function

Ver 1.0

Date: 15.09.2022

User manual

Dr. Igor Deyneko

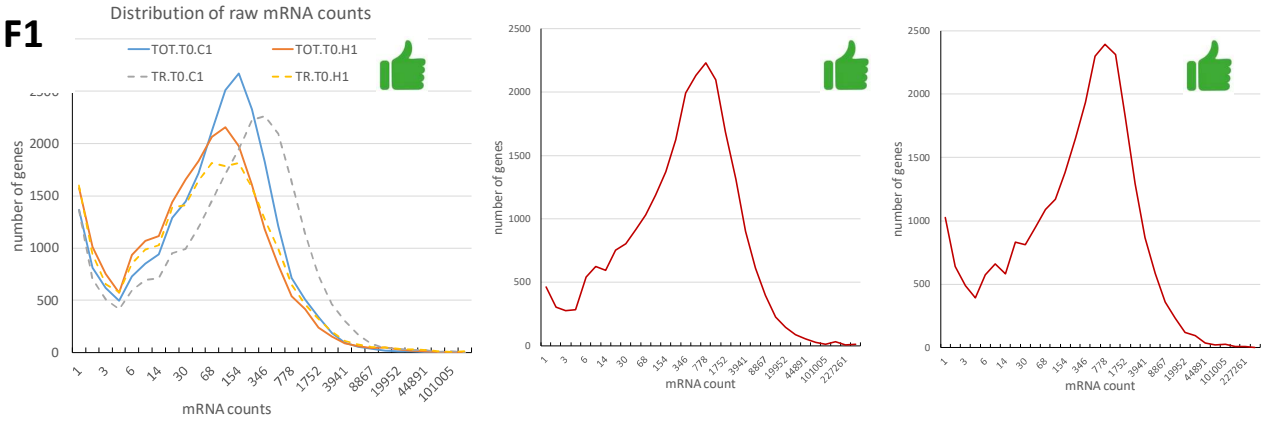
Introduction

We have developed a method for cleaning RNA-seq data, which improves the detection of differentially expressed genes and specifically genes with low to moderate transcription. Using a data modeling approach, parameters of randomly distributed mRNA counts are identified and reads, most probably originating from technical noise, are removed. We demonstrate that the removal of this random component leads to the significant increase in the number of detected differentially expressed genes, more significant p-values and no bias towards low-count genes.

What data can I clean with RNAdeNoise

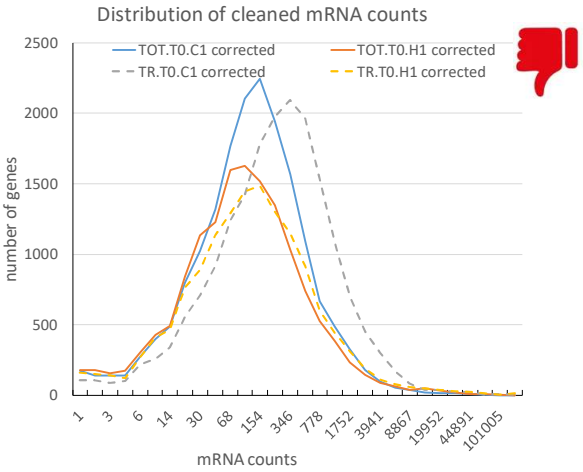
The method was developed to clean RNA-seq data, but can be applied to any data. The prerequisite is a two-peak shape, which is interpreted as consisting of real and random parts. Below are several examples or how the distribution should and should not look like.

F1



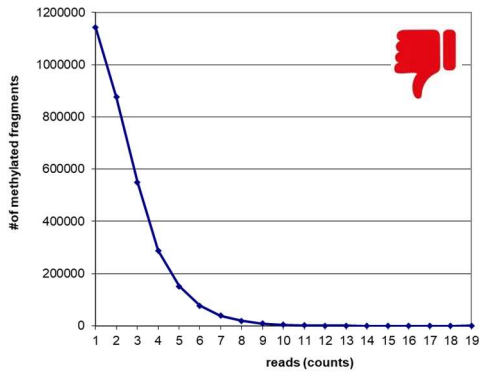
Two-peak distributions, exponential part can be small or even higher than bell-shaped peak, it is correct.

F2



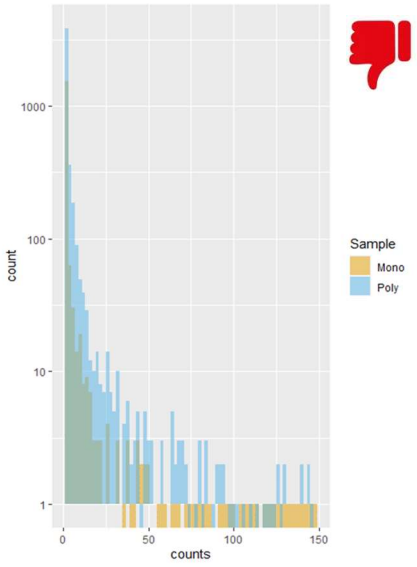
- No exponential part! This data has already been cleaned.

F3



- No second peak, i.e. real signal is so weak that it is masked withing randomly distributed reads.

F4



- Same problem - no second peak, although some mRNAs have high counts, and over 70% of all RNAs have counts above 10 – a very conservative threshold used to cut noisy reads. See explanations in the following text.

In practice it is not always possible to reach sufficient sequencing depth, which can be seen as a reduced or missing second peak. For example, because of the in vivo collection of specific immune cells of mouse thymus, only a very few number of cells can be isolated for sequencing (Fig F3). Thus, the required sequencing depth cannot be achieved and the distribution has only one peak. Another reason could be the sequencing technology. Distribution of mRNA counts of polysomal and monosomal tomato RNA fractions, sequenced on the MinION device (nanoporetech.com) similarly shows only an exponent-shaped distribution (Fig. F4), although many genes have counts far above the commonly used thresholds of 3 to 10. In both above cases it would be wrong to interpret the data to be purely random, but so that it is not possible to separate noise and real reads using statistics.

RNAdeNoise should also not be used if the exponential part is missing, for example, if the data has already been cleaned (Fig F2). Iterative use may result in incorrect exponential model fitting and data corruption.

to be continued…