Abstract

In most cases, modern RNA sequencing methods are used to assess the level of gene expression and quantitatively compare expression in different biological samples. However, transcriptomic data can also be considered in a qualitative context, when we find differences, similarities and draw conclusions based on them.

The first goal of this work was to develop a method for qualitative analysis of transcriptomic sequence data in multifactorial experiments. The second goal was to add a method for interpreting biological effects to this statistical analysis. Finally, the third goal was to check whether qualitative comparison can also improve or introduce a new context to quantitative analysis.

To obtain comparisons of data sets - in the context of jointly and disjointly occurring nucleotide sequences - the Mash distance was used. It is the basis for the subsequent analysis of qualitative variability of transcriptomes and identification of statistically significant effects due to their diversity. Furthermore, the use of the GOSlim GOA ontology was proposed to obtain a more detailed biological interpretation of the results. Examples of application of the methods on real experimental data are presented and the results obtained in this way are discussed.

Real transcriptomic data from a three-factor experiment (factors: genotype, day, temperature), conducted on barley (Mikołajczak i in., 2022) were obtained by Illumina sequencing. These data were subjected to quality control and assembled into scaffolds. For each pair of sets, the Mash distance was calculated, resulting in a distance matrix that was subjected to PCoA analysis and then univariate and multivariate analysis of variance. Calculations were performed for raw data, for scaffolds, and jointly for the raw and assembled data sets, in order to compare the results. The results of the analyses showed the greatest influence of the temperature factor and the day factor on the qualitative differences in the transcriptome sets. The assembly step noticeably differentiated the sets from each other, while the results for scaffolds and raw data separately were similar.

The assembled data sets were divided into subsets based on the GOSlim terms assigned to the individual sequences. For each term for which the subsets of data were non-empty, the analysis described above was performed. As for the entire sets, the effects of the temperature factor and the day factor were significant.

The data were subjected to quantitative differential analysis using two models: a linear analysis of variance model and an analysis of covariance model, where the explanatory variables were the principal coordinates. The numbers of genes showing differential expression levels obtained using the information contained in the Mash distances (i.e., in the analysis of covariance model) were lower than those obtained using the linear model.

The most important general conclusions resulting from the presented work are: Mash distance is a suitable tool for introducing a qualitative context to the analysis of transcriptomic data, the developed method cost is relatively low in terms of time and resources, and the use of the GOSlim ontology allowed enriching the analysis of transcriptomic data with a functional element. Using the results from PCoA performed on the Mash distance matrix allowed improving the results of differential quantitative analysis.

Keywords: distance, Mash, transcriptome, GO, GOSlim, factorial experiments, genotype by environment interaction