EVALUATION OF STATISTICAL METHODS FOR COMPARATIVE METAGENOMICS

Viktor Jonsson[1], **Olle Nerman**[1] **and Erik Kristiansson**[1]
[1] Mathematical Sciences, University of Gothenburg and Chalmers
University ofTechnology, Sweden.

Metagenomics is the study of microbial communities in their natural state. In comparative metagenomics samples are compared to identify differentially abundant genes and other features between conditions. The data considered is high-dimensional, discrete and highly overdispersed. Several statistical methods, ranging from Fisher's exact test and t-tests to generalized linear models and bootstrapping, have been proposed. However, no comprehensive evaluation has been performed to see which best suits metagenomic data. We present preliminary results from a comparison of several commonly used methods for statistical analysis of metagenomic data. We evaluate the performance of the methods with respect to number of samples, gene abundances, and effect sizes. In order to conserve the variance structure of real metagenomic data the test-data is created by resampling large metagenomic datasets. This study will provide guidance on the choice of statistical methods for future metagenomic studies.

Keywords: Metagenomics, Biostatistics, Bioinformatics.