CHALMERS | GÖTEBORGS UNIVERSITET

Statistical analysis and modelling of gene count data in metagenomics

VIKTOR JONSSON

Thesis for the degree of Doctor of Philosophy to be defended in public on
Friday, February 17, 2017 at 10.00 in Pascal,
Department of Mathematical Sciences, Chalmers tvärgata 3, Göteborg.
Faculty opponent is Professor Yudi Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

The thesis is available at
The Department of Mathematical Sciences
Chalmers tekniska högskola och Göteborgs universitet
SE-412 96 Göteborg
Telefon: 031-772 1000

Statistical analysis and modelling of gene count data in metagenomics

Viktor Jonsson

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Microorganisms form complex communities that play an integral part of all ecosystems on Earth. Metagenomics enables the study of microbial communities through sequencing of random DNA fragments from the collective genome of all present organisms. Metagenomic data is discrete, high-dimensional and contains excessive levels of both biological and technical variability, which makes the statistical analysis challenging.

This thesis aims to improve the statistical analysis of metagenomic data in two ways; by characterising the variance structure present in metagenomic data, and by developing and evaluating methods for identification of differentially abundant genes between experimental conditions. In Paper I we evaluate and compare the statistical performance of 14 methods previously used for metagenomic data. In Paper II we implement an overdispersed Poisson model and use it to show that the biological variability varies considerably between genes. The model is used to evaluate a range of assumptions for the variance parameter, and we show that correct modelling of the variance is vital for reducing the number of false positives. In Paper III we extend the model used in Paper II to incorporate zero-inflation. Using the extended model, we show that metagenomic data does indeed contain substantial levels of zero-inflation. We demonstrate that the new model has a high power to detect differentially abundant genes. In Paper IV we suggest improvements to the annotation and quantification of gene content in metagenomic data. Our proposed method, HirBin, uses a data-centric approach to identify effects at a finer resolution, which in turn allows for more accurate biological conclusions.

This thesis highlights the importance of statistical modelling and the use of appropriate assumptions in the analysis of metagenomic data. The presented results may also guides researchers to select and further refine statistical tools for reliable analysis of metagenomic data.

Keywords: metagenomics, statistical modelling, hierarchical statistical models, gene ranking, overdispersion, zero-inflation, false discovery rate, receiver operating characteristic curves.