A comparison of statistical tests in microbial ecology

Stine Katrine Rye Østergaard

Master student at Bioinformatics Research Centre, Aarhus University. Student number: 201706401

Supervisors

Thomas Bataillon Associate professor at Bioinformatics Research Centre, Aarhus University

Leendert Vergeynst. Assistant professor at Biological and Chemical Engineering, Aarhus University

Ioannis Kampouris. Postdoc at Biological and Chemical Engineering, Aarhus University

Spring semester 2022

Preface

I would like to thank to my three supervisors Thomas Bataillon, Leendert Vergeynst and Ioannis Kampouris for all the patience and feedback throughout the project.

Abstract

Microorganisms occupy a broad range of ecological habitats and play a large role in shaping the different ecosystems of the planet. Differential abundance analysis can be used to detect which organisms increase or decrease in abundance as a result of a change in the environmental conditions. This statistical procedure gets complicated due to the compositional nature and large variance of microbial abundance data, and this complexity of the data has also prevented the acceptance of a universal model of microbial abundance data in the literature. Different simulation approaches are thus used in the different benchmark studies found in the literature, and this is a problem because some simulation approaches favor methods if they are based on the same statistical assumptions.

In this study, a novel simulation approach, based on general simulation trends in the literature, is used to investigate how the statistical assumptions made during simulation affect the structure of the simulated data. A benchmark study of the five statistical methods DESeq2, ANCOM-BC, ALDEx2, two-sample t-test, and Wilcoxon rank-sum test is performed to investigate their performance on the simulated data, especially when a low number of biological replicates are available. The performance of the five methods was also evaluated on a real data set obtained from a microbial ecology study in the Disko Bay, Greenland.

The simulation approach created data with less variation than other simulation approaches in the literature and is not appropriate for benchmarking the different methods, but inspection of the results still revealed how the assumptions made during simulation affect the data generated. It was still possible to rediscover a general trend from the literature, that the sensitivity of the methods depends on the number of replicates available, and this constraint differs between methods. The key result obtained from the analysis of real data is that the number of positive hits gets reduces tremendously when the number of replicates is reduced from three to two.

Table of contents

1 Abbreviations and important definitions

ASV	Amplicon sequencing variant. Groups of reads from a DNA sequencing that are clustered based on 100% similarity.
FDR	False discovery rate. Defined as: $\frac{False\ positives}{False\ positives + true\ positives}$
FPR	False-positive rate. Defined as: $\frac{False\ positives}{False\ positives + true\ negatives}$
Library size	The total count of reads in each sample
OTU	Operational taxonomical unit. Groups of reads from a DNA sequencing that are clustered based on 97% similarity.
Sensitivity	$\frac{True\ positives}{True\ positives+false\ negatives}$
Specificity	$\frac{True\ negatives}{True\ negatives+false\ positives}$
Taxa	Plural of taxon
Taxon	Taxonomical unit. Can be any class such as species, genus, or kingdom.