Identification of heat stress microbial biomarkers using 16S rRNA

Bartosz Czech
bartosz.czech@upwr.edu.pl

1 Summary

1.1 OTU identification

QIIME2 was used for denoising single-end data and paired-end reads. Unfortunately, results show, that reverse reads have poor quality. Therefore, I decided to include only forward reads from paired-end data. Both forward and single-end data were denoised separately. Finally, after the denoising step, reads were merged. Afterwards, denoised reads were clustered based on 99% of similarity. Clusters create Operational Taxonomic Units (OTUs). Set of OTUs was used for calculating phylogenetic tree, core phylogenetic metrics, alpha diversity, beta diversity and for Principal Coordinate Analysis (PCoA). OTUs were then classified to assign taxonomy. We used GreenGenes v13.8 database with sequence groups with 99% of sequence similarity. Since the reference set has a sequence for the whole 16S rRNA gene, in silico PCR was performed to extract only V3-V4 regions. Naive Bayes algorithm was used for the training model. Finally, OTU table, taxonomy table, and unrooted tree were extracted to tsv format.

1.2 Differential abundance

The aim of this step was to find an association between the quantitative trait (EBV, deregressed EBV) and abundance of bacteria. The previous step allowed to identify 46825 unique OTUs for 138 samples.
Firstly, OTU table was cleaned, by removing OTUs with zero variance and small total number of reads per OTU (possible error).
Then, I visualized the OTU table using UMAP technique that allows to do dimensions reduction. UMAP showed, that there is a batch effect – sampling year.
Then, in order to check the distribution of OTU table, I compare their mean and variance. The comparison showed, that distribution of OTUs follows negative binomial distribution (asymptotically higher variances than means).
Because of the distribution, I used a negative binomial distribution regression model, where the dependent variable was normalized counts for given OTU in a

given sample. Independent variables represented deregressed EBV and sampling year. I computed three separate models – for each DEBV.
Models showed 2791, 2676, and 2518 differentially abundant OTUs across DEBV (for rectal temperature, respiratory score, and drooling score, respectively).
Unfortunately, this relationship was superficial. I wisualized the most significant (with lower FDR) OTU and for all the cases, the significance was cause by the outlier value in a single sample.
For this reason, I decided to accumulate OTU tables into the Genus table in order to reduce a rank of matrix and also to have some variability among samples.
Unfortunately, identification of differentially abundant Genus had no effect. On significant level, $\alpha = 0.05$ there were no significant Genus. The minimal FDR was 0.45. Moreover, it was reassuring to see some meaningful relationship between Genus and DEBVs. Based on the literature I found out that DeSeq2 and edgeR might cause false negative values because of the different assumption in 16S rRNA experiments and in RNA-seq. The other problem is that we use a nontypical independent variable – continuous variable. Most experiments focus on case-control studies. Therefore other models are considered.

2 Road map

- linear mixed model with OTU/Genus as a random effect
- robust regression model with Huber statistics
- identification of biomarkers and preparing a machine learning model for prediction of DEBVs
- redoing the whole analysis for other variable regions of 16S rRNA
- integration analysis of different 16S rRNA subregions