Applying compositional data analysis methods to characterize the metatranscriptome and metabolome of the vaginal microbiota

Greg Gloor, Jean Macklaim, Amy McMillan, Mark Sumarah, Gregor Reid

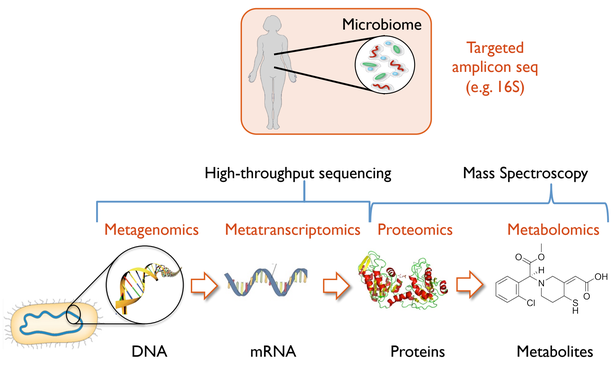
The University of Western Ontario

ggloor@uwo.ca • ggloor.github.io • github.com/ggloor/CoDaSeq

Abstract

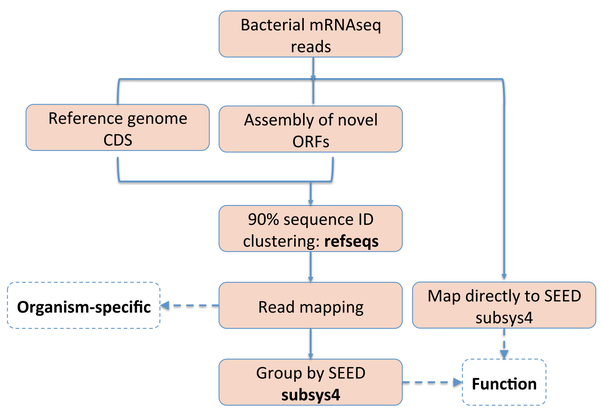
A major challenge of any microbiome investigation is to determine the role of the microbes in the environment, and their effects on the host or system. High-throughput sequencing (HTS) and small molecule analyses provide an overview of the function of the entire microbiome, which can be thought of as a "meta-organism". We used the vaginal microbiome as a model system to demonstrate how a CoDa approach can illuminate the relationship between the bacterially expressed mRNAs (the metatranscriptome), and the products of metabolism (the metabolome). Using a CoDa framework, we identified novel functional profiles of the vaginal microbiome associated with healthy and dysbiotic conditions not identifiable by other methods. We show that the transcriptional components of specific genera (*Megasphaera* and *Prevotella*) associate with the vaginal microbiome subgroups, while others (*Gardnerella*, *Lactobacillus iners*) are only minimally discriminatory. The power to separate subgroups transcriptionally was increased by aggregating reads into functional groups rather than analyze individual organisms. Despite significant taxonomic variability within each subgroup, the approach enabled us to identify core transcriptomic and metabolic products separating health and dysbiosis. Strong correlations were found between the small molecules and the transcripts detected in key metabolic pathways of the condition. This study underscores the power of using a CoDa approach to understand functional differences between microbiome states.

Levels of meta-organism analysis



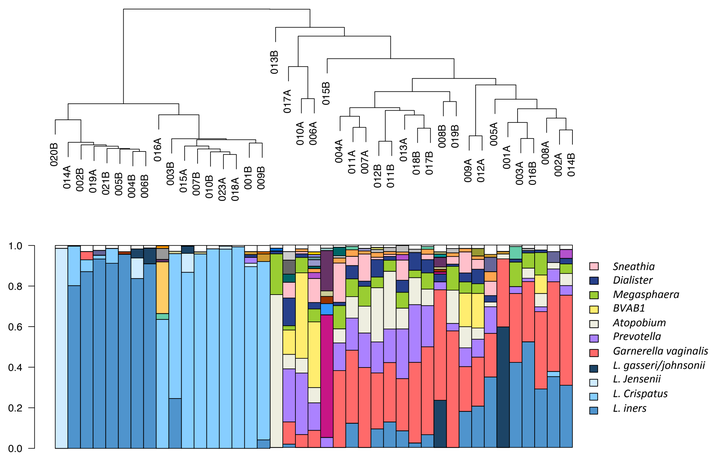
There are multiple levels at which a microbial/host system can be interrogated. Currently, 16S rRNA gene sequencing and metagemone sequencing are the dominant methods. However, using one or more of the functional approaches (eg, metatranscriptome and/or metabolome) give better snapshots of the expressed potential and the expressed result. Integrating multiple data sources also serves as a check and balance that ensures robust conclusions are drawn.

Strategies for mapping reads to functions



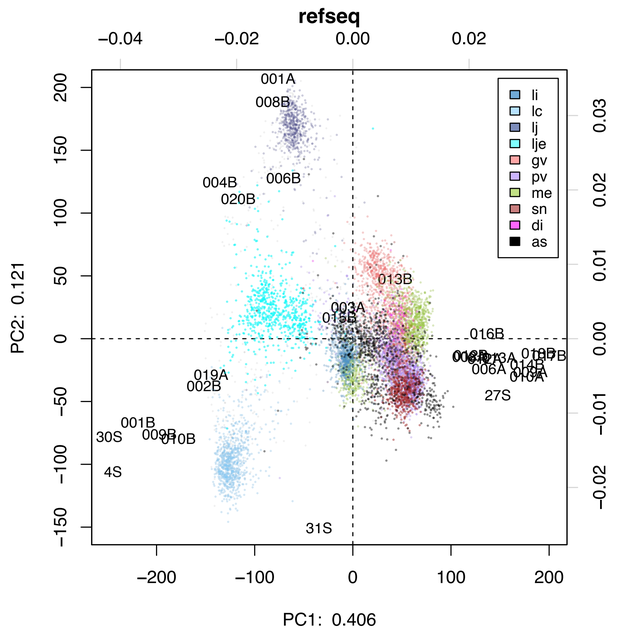
Annotation by function assumes that the ecosystem is the unit of function, and not the individual organism. Total single end Illumina reads were filtered to remove human contamination, then mapped to a set of non-redundant reference ORFs, which were in turn mapped to the SEED subsys4 functional annotation system. Alternatively, reads could be mapped directly to the SEED subsys4 database.

The sample set by 16S rRNA gene sequence



The samples were characterized by 16S rRNA gene sequencing using the V6 variable region. Each bar is a sample, and each color is a taxonomic member of the community. Typically, healthy conditions are dominated by one species of *Lactobacillus* (blue). A dysbiotic shift occurs in bacterial vaginosis (BV) where the community is dominated by a mixed population of anaerobes.

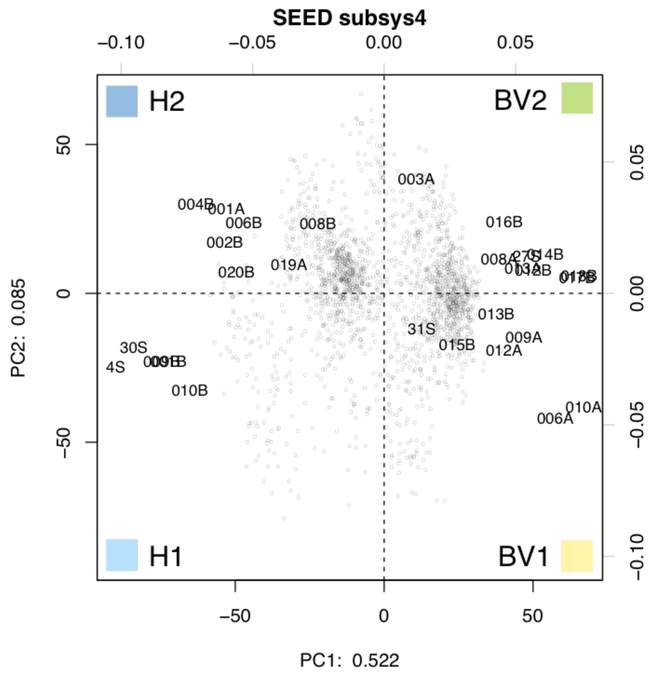
The sample set by mRNA mapping to organism-specific reference sequences



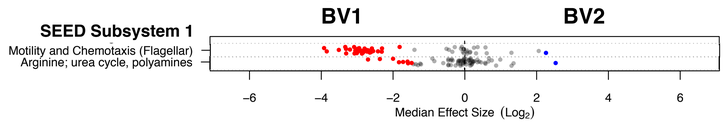
These data were explored by mapping the reads to reference sequences derived from individual organisms followed by the generation of a compositional biplot. Here we are examining the variance in the reference sequences and the samples, not abundance. The first principle component splits the health and BV types, the second principle component splits the different health types. ORFs in *L. iners* are near the center of the plot indicating that this organism is not contributing to the difference between the groups. Note that ORFs from some organisms (e.g. *Megasphaera*) are separated, likely because of different strains with different functions in the ecosystem. Note that the BV samples on the right are not differentiated. Also note, that reads mapping to *Atopobium* were not observed.

1) Aitchison(1986) The statisitical analysis of compositional data (Blackburn Press) and Pawlowsky-Glahn (2015) modeling and analysis of compositional data (Wiley)
2) Fernandes (2013) ANOVA-like differential expression analysis for mixed population RNA-seq data (PLoS ONE)
3) Fernandes (2014) Unifying the analysis of high throughput sequencing datasets (Microbiome)
4) Gloor (2016) It's all relative: analyzing microbiome data as compositions (Annals of Epidemiology)
5) Fernandez (2014) Bayesian-multiplicative treatment of count zeros in compositional data sets (Statistical Modeling)
6) Macklaim (2016) Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis (Microbiome)

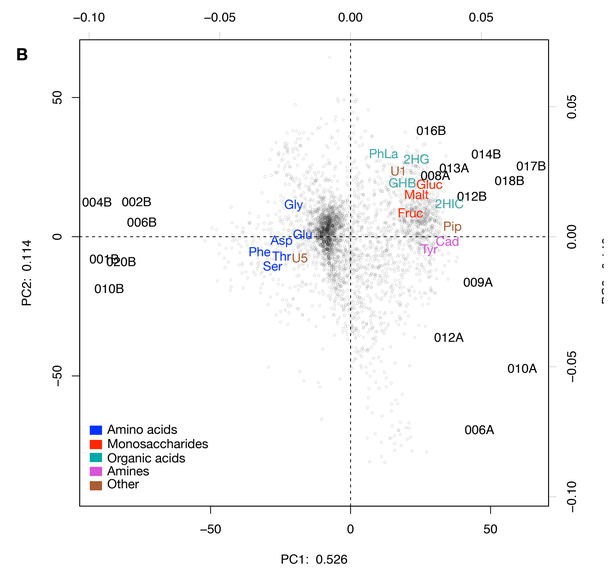
The sample set by mRNA mapping to functions



References sequences were mapped to the SEED subsys4 database and reads mapping to those reference sequences were aggregated at that level. The figure shows a CoDa biplot of that data. The major split in the data is still between the healthy and BV samples. Two key conclusions are that there is redundancy in function despite differences in taxonomic composition, and the the conserved biochemical functions are the major unit of interest – not individual genes.



Note that the BV samples now split into two groups. Comparison of the relative abundance of functions shows that this is driven largely by differential abundance of motility and polyamine-associated functions which can be observed at the SEED subsystem 1 level.



The results of an untargeted metabolomic analysis overlaid on a subset of the data mapped as a CoDa biplot of SEED subsys4 functions shows good agreement of the metabolomic and metatranscriptomic data. The major differences between health and BV samples are amino acid and carbohydrate metabolism by both, and the difference between the two BV types in polyamine production is observed.