Internal Report CS Bioinformatics Track 17-02 August 2017

Leiden University

Computer Science

Bioinformatics Track

Integrating the Microbiome and Metabolome Using Datamining and Network Analysis Approaches

B.ASc. A.A.B. Versteeg

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Integrating the Microbiome and Metabolome Using Datamining and Network Analysis Approaches

Master's thesis

Leiden Institute of Advanced Computer Science (LIACS) – Leiden University

Internship provider: BaseClear B.V. Leiden

Author: B.ASc. A.A.B. Versteeg
s1203428@umail.leidenuniv.nl

University mentor: Dr. E.M. Bakker
Internship mentors: Dr. W. Pirovano
Dr. B. van den Bogert

August 14, 2017

Abstract

This document describes a study on the integration of microbial 16S rRNA gene data with targeted metabolomics data, in the context of bacterial vaginosis (BV), to investigate microbial traits and metabolic particularities involved with this disorder. Multiple computational methods were applied, ranging from simple statistical tests to different applications of decision trees and most importantly a newly devised network analysis approach that is based on the traditional PageRank algorithm, which we named *PageRank based metabolic modeling* (PBMM).

The PBMM algorithm is capable of predicting metabolite profiles based on metagenomic data and hence cannot be directly applied on amplicon data. Therefore, metagenomes were predicted using PICRUSt software package. The predicted metagenomes were combined with KEGG data to construct a reaction network on which the PBMM algorithm was applied. Predictions were assessed using observed metabolite profiles.

The findings are that the newly devised method is particularly good at predicting metabolite profiles for patients that were not diagnosed with BV, whereas the method's performance on is reasonable for patients positively diagnosed. This distinct difference is likely to be contributed to the increased community complexity that comes with this disorder.

Contents

1 Introduction

The process of unraveling the mysteries surrounding the symbiosis with our microbiomes has recently grown in importance due to the emerging understanding of the strong influence that this symbiosis has on the human health. For instance, the gut microbiome carries out various functions involved with digestion of food, provision of essential nutrients and vitamins and a line of defense against pathogenic microbiota[1]. Not surprisingly, the intestinal microbiota are occasionally considered as an additional human organ[2], that takes care of a great variety of biochemical transformations, more than any other human system[3].

While this collaborative relationship is crucial for both sides, it possibly comes with downsides as well, the opposite named dysbiosis. Such balance defects as well as typical microbiome compositions were linked to a multitude of diseases, such as inflammatory diseases[4], neurological disorders[5], metabolic diseases[6], obesity and cardiovascular diseases[7–9] and last but not least cancer[10, 11].

A recent study on this topic performed by Raes and Bork[12] has shown the complexity of community functions; a multitude of species is believed to lead to the difference between symbiosis and dysbiosis, meaning a single species is not necessarily the perpetrator of a disease as the rational is all about a global balance. A microbial ecosystem, while differing at levels of taxonomy, carries out the various tasks as a system altogether[12]. Microbiota exchange metabolites (i.e. cross-feeding[13]) by excretion and absorption to perform intracellular biochemical transformations catalyzed with their enzymes.

Sequencing technologies

The advent of high-throughput sequencing technologies, provides novel insights into the complexities that shape host-microbe interactions. The intestine can now be profiled using various DNA sequencing techniques. The binning of the sequence data into operational taxonomic units (OTUs) is ideally done up to strain level, as on species level the the organisms may already exhibit very different traits. For example, most strains of the very common *Escherichia coli* bacteria are non-pathogenic, however, the O157:H7 strain has resulted in multiple epidemic outbreaks during the past decades [14].

Microbial profiling is most commonly performed by amplification and sequencing of the 16S rRNA gene, though this is a cost-effective and practical approach, it merely allows for an optimal taxonomic identification on species level. In order to allow for identification on strain level, a common resort is Whole Metagenome Shotgun (WMS) sequencing, even though this comes with a greater cost in total as well as per nucleotide. In addition, cares must be taken to account for the contamination of nucleic acid by the host. Despite the relatively high costs, sequencing whole metagenomes allows for a more profound analysis of the data.

Multi-omic data integration

To obtain a better understanding of the community functional activities, as well as a mechanistic understanding of the links underlying the sym- and dysbiosis, the quantitative taxonomic data is often combined with other -omic data types, such as transcriptomics[15], proteomics[16] and metabolomics [17, 18].

While (meta)genomics only gives insight into the genetic content of the organisms, this data complemented with proteome or metabolome data can give much deeper insights as not only the genetic content is analyzed, but also the organisms' biochemical functional activity.

The metabolome, representing the collection of all metabolites involved with both the microbiota as well as the host, are both the end-products as well as the precursors of cellular processes. Generally, only the extracellular metabolome is measured, meaning no precise data is available on the intracellular biochemical situation, which can introduce a complexity. Mainly, because various organism's, while genetically different, have different influences on shared parts of the metabolome, i.e. one metabolite may be consumed or produced by several types of organisms, introducing a dimensional problem; the responsible for a decrease or increase of a metabolite can be difficult or even impossible to designate.

With more advanced measurement methods the extracellular metabolome can be filtered and the retained cells be lysed to also measure the intracellular metabolome[19]. While this enhances the data's depth (disregarding the biases induced by different lysis specificities), contributing the increase or decrease of a certain metabolite to one or more organisms (i.e. consumers/producers respectively) can remain challenging as all intracellular metabolites are mixed prior to the measuring.

Depending on the goal of the study it might be favorable to resort to transcriptomics to obtain data on the level of gene expression. This combined with whole metagenome sequencing can optimally yield gene expression levels for each organism, while taking into account strain specific variation.

When comparing the metabolome with the transcriptome and proteome, the latter two are better considered as a more intermediate measurement, meaning the data is closer to organism specific activity, with a trade-off that the data gives less direct insight into the community-wide phenotypic traits. To this end, a resort to (targeted) metabolomics can be a viable direction.

Flux balance analysis

Flux balance analysis (FBA) is generally considered as the current state-of-the-art metabolic modeling technique. The technique has proven to be capable of efficiently computing the metabolic fluxes. This is done using genome-scale reconstructions of metabolic models, which are formalized as a set of variables that represent the fluxes and a set of constraints that are based on the reactants and products in combination with the stoichiometrics of each catalyzed reaction. For some metabolites also transport reactions are involved as constraints with at least the lower or upper limit set free (i.e. consumption, production or both). The mathematical composite can then be evaluated using linear programming, where the objective function to be optimized is typically based on the precursors of biomass[20–22].

Even though FBA as technique itself is relatively simple, it has several requisites; firstly the genome-scale model has to be very accurate and therefore experimentally validated (e.g. uptake and production rates). This requirement makes the technique deemed to fail within the context of complex microbial communities. Merely well defined mock communities are currently considered as feasible, new techniques are to be developed first[23]. Very obstructively is the great number of poorly defined community members. These members are not often classified on species level, while strain precision is necessary. Last but not least, for many organisms genome-scale metabolic models are yet to be constructed. Automated procedures for model construction have been proposed[24], however, the laborious process of manual curation remains essential. A model not accurated, is likely to return invalid results and probably even will not converge to an optimum.

FBA alternative

The analysis of high-dimensional microbial community data comes with many complexities. Regular datamining techniques are not applied trivially and are possibly even not appropriate, due to the *curse of dimensionality* aspect. Of these issues FBA can only solve some, but its requirements are too strict to be useful within the context of complex gut or other communities.

To this end new techniques and methods should be devised and investigated, to allow for thorough analysis of complex microbial communities, while involving the entire microbiome and metabolome in attempt to utilize all information available. The technique's yielded output should give at least the opportunity of making interpretations on a community-level, and optimally on organism level with the goal of elucidating member-member specific interactions.

2 Literature review

The present literature reveals several papers of studies on the topic of unraveling the complexities involved with gut or other microbial communities, using both quantitative taxonomic and metabolomic data. Studies differ in the approach of withdrawing information from these two sources; in a simplified manner by evaluating both datasets independently or more advanced by making an attempt at effectively integrating both data types, with the goal to obtain information not detectable when evaluating the datasets independently, but only when leveraging from the combined potential.

2.1 Metabolic Signatures of Bacterial Vaginosis

Dataset's properties

The paper by Srinivasan et al. [25] describes a study on the microbial and metabolic signatures of bacterial vaginosis (BV). 70 patients were involved during the study of whom each a sample was collected from the cervicovaginal lavage fluid. The samples were sequenced by targeting the V3-V4 region of the 16S rRNA gene with broad-range PCR pyrosequencing (Roche 454 Life Sciences titanium technology). The metabolite concentrations were measured using a LC-MS analysis (i.e. targeted metabolomics).

All the 70 patients were tested for BV using two clinical criteria, the Amsel[26] and Nugent[27] criterion. With the Amsel criterion, 26 patients were diagnosed negatively versus 44 patients positively. Unlike the binary outcome of the Amsel criterion, the Nugent clinical criterion is based on a numeric score, ranging from 0-10, with class ranges as follows: negative $< 4 \leq$ intermediate $\leq 7 <$ positive. Despite this difference between the two criteria, the two methods did gave corresponding results (i.e. were not in conflict), the Nugent's intermediate group was composed of 4 negative and 6 positive patients, based on the Amsel criterion.

Data preprocessing methods

After all data processing steps, 171 OTUs were found using QIIME's[28] OTU picking pipeline. However, the number of distinct taxonomies is only 63, as the vast majority is unclassified at the level of species or even higher. The concentration profiles of 279 metabolites were measured, of which only 96 were made publicly available, possibly due to missing KEGG references.

Metabolite data was normalized by median centering followed by log transforms. Missing values were imputed using the lowest occurrence value of that metabolite. According to the paper the OTU abundances were only log transformed and thus no measurements were taken to account for the relativeness.

Findings

The paper states that the metabolic signatures specific to BV were strong across multiple metabolic pathways and in turn associated with the presence or absence of certain bacteria. In summary, 62% of the 279 measured metabolites (of which only 78 were mapped to the KEGG database) turned out to be significant. Lower concentrations of amino acids were associated with BV, not surprisingly the same applied to the amino acid successors, dipeptides. In contrast, higher levels of amino acid catabolites and polyamines were measured for the BV positive.

More specifically, certain bacteria caught the eye in particular, being more specific to BV than other bacteria. Such as Lactobacillus, which tented to be absent in case of BV. This characteristic of BV was observed earlier during other studies [29, 30]. This absence/presence pattern of Lactobacillus is also reflected to the metabolome, where lower levels of lactate are specific to the control group, not surprisingly as Lactobacillus as bacterial lineage is known to be the key contributor of this metabolite. Contrary to lactate, succinate was measured in higher concentrations for BV positive diagnosed patients, previously also observed during several other studies[31–34].

2.2 Model-Based Integration of Taxonomic and Metabolomic Profiles

Probably the most related study, performed by Noecker et al. [18], demonstrates an analytical framework that can link community composition with metabolic shifts. The framework is used to analyze four independently acquired datasets, each pairing taxonomic/metagenomic data with metabolic profiles. One of their datasets was acquired from the study by Srinivasan et al.[25] (previously described in section 2.1). This same dataset was also analyzed during this thesis project (see section 4.1).

Their study was primarily focused on two datasets, the one relating to bacterial vaginosis, and the other from mice guts, from a study that investigated *Clostridium difficile* susceptibility in case of antibiotics treatment.

The study is based on the presumption that the stress applied on or present in a community is likely to induce a variation in the data, that eases the process of statistical association. For instance, shifts in both the microbiome and metabolome are easier to couple if prevalent than if weak. Coupling the shifts between datasets should give insights into species specific metabolic functions. While such data does increase the associative potential of the entities involved, it is also prone to spurious correlations, e.g.: two very different organisms responding similarly to a condition, will therefore also correlate to each others metabolites. Finally, the biochemical context in case of high stress might be very different from normal situations and can hence can be very unrepresentative for the reality.

The framework's core principle is to predict metabolite abundances from the 16S rRNA gene abundances. Predictions are then assessed by correlating to the observed metabolome. For the well predicted metabolites this should indicate that the species used for prediction contribute to the production or absorption of this metabolite, which is quantified by correlating each species' abundance to the predicted metabolite concentration.

The workflow used to predict the metabolites consists of various steps, integrating multi-omic data from various sources. See figure 1 for a schematic display of the workflow.

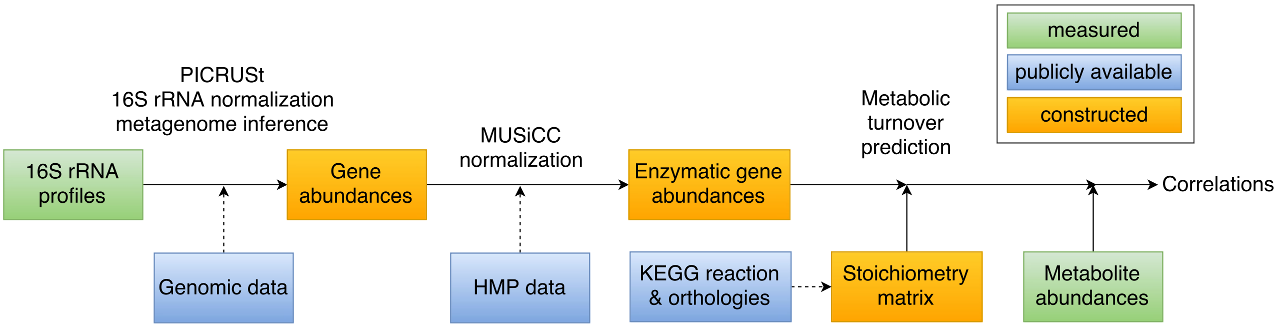


Figure 1: Workflow used by Noecker et al.

Following an explanatory text for each step of the workflow:

1. Using the PICRUSt [35] framework:

   (a) The 16S rRNA gene abundances are normalized based on their copy-number variations, by dividing each abundance by the known/predicted gene copy-number (see section 4.3).

   (b) Next, for each OTU its genetic content is predicted using a reference database, resulting in a table pairing each sample with KEGG orthology (KO) identifiers with corresponding abundances (see section 4.4).

2. Predicted KO abundances are then normalized using the MUSiCC [36] package, which corrects for several biases; sample sequencing depth, average genome size, the species richness and average genome mappability (i.e. genome complexity due to repeats resulting in an assembly bias [37]).

3. At this point the transition from metagenomic to metabolic data is made. The microbial community-wide metabolic potential is predicted using an adapted version of the previously developed method by Larsen et al. [38]. The method's base line is making an attempt of predicting metabolite abundances based on the on the predicted KO abundances, using pathway specific (i.e. catalyzed by certain enzymes) reaction data obtained from the KEGG[39] database.

   Predicting the metabolic turnover is done using a stoichiometry matrix S_{mn}, containing the quantitative relationships between m metabolites and n genes (i.e. enzymatic products). Thus, each cell S_{ij} represents the turnover of metabolite i by gene j. The matrix construction proceeds as follows:

- For each enzyme e coded by gene j:
  - For each irreversible reaction r catalyzed by enzyme e:
    - For each reactant i subtract its coefficient from S_{ij}
    - For each product i add its coefficient to S_{ij}

Only irreversible reactions are taken into account, as the authors consider only these to contribute to the metabolite predictions, i.e. true reaction direction depends on many factors and cannot not be predicted trivially. Also the reversible reactions cannot be incorporated into the stoichiometry matrix easily, as the addition and subtraction of the same stochiometric coefficient makes no difference.

According to the authors, this stoichiometry matrix based approach accounts for metabolic fluxes, as the used reactions are pathway specific in a sense that the enzymes, present in the corresponding pathway, catalyze the reactions and hence require less reactants to allow the reaction to occur. To this end, no bounds on the fluxes are considered, meaning reactions are assumed to be able to occur without limits and also reaction rates are not considered.

Metabolites involved – as either reactant or product – in reactions that are catalyzed by enzymes coded by 30 or more genes (named "currency" metabolites) are excluded from further analysis, following the method previously demonstrated by Greenblum et al. [40] and Taxis et al. [41]. Furthermore, the matrix is normalized compound-wise (i.e. row-wise) over a range of [-1..1], following the study of Larsen et al. [38].

A side note to the approach of constructing this stoichiometry matrix, is that it differs from the traditional stoichiometry matrix in a sense that the genes are normally put against reactions, instead of metabolites in this case.

Community-wide metabolic potential (CMP) is then computed by multiplying the gene per sample abundance matrix with the stoichiometry matrix.

4. After all the data transformations, correlations were made using Mantel tests (see section 4.6), in order to estimate the correlation between pairs of CMP matrices. The authors state that regular correlations are not applicable as the CMP scores represent relative predictions.

5. Well predicted metabolites do not yet give a direct indication of the contribution of involved OTUs to that metabolite. The predicted metabolite concentration is derived from an arbitrary set of OTUs. Backtracking the workflow can be somewhat cumbersome as many variables and also multiple normalization techniques are involved during the process. To this end, the authors proposed an simplified but viable approach. For each OTU involved with a metabolite its abundances over the samples is correlated with the predicted concentration (which is based on all OTUs involved that metabolite). The correlation coefficient is said to give an estimate of that OTUs degree of involvement, with that well predicted metabolite.

6. Spurious findings are ruled out by shuffling the network's edges (i.e. null model) and re-performing the analysis. The paper states no correlations were found worth mentioning.

An obvious caveat of this approach is that each step possibly comes with a bias due to missing data. The paper's authors stated they were not able to detect these events and hence the severity of the bias remained unclear.

So called currency metabolites, that occur in reactions associated with 30 or more genes (that code for the enzymes catalyzing these reactions), were excluded from the prediction of the relative metabolic turnover (RMT), following the studies by [40, 41]. Understandably, of such metabolites it is difficult to predict the concentration profiles accurately, since a relatively high number of OTUs have influence on the concentration. However, exclusion using this criterion therefore filters metabolites difficult to predict and hence can give a deceiving impression on the method's performance.

Metabolites transformed intracellular by microbiota can, after excretion, be absorbed by other microbiota and consequently be further transformed; a phenomena also known as cross-feeding[42]. This aspect is currently not captured in the metabolic turnover. The method used for predicting the relative metabolic turnover, is non-iterative, as it makes predictions using only an one-time conversion. For instance, the excretion of one metabolite by a organism, which in turn is consumed by another organism (i.e. cross-feeding) is not accounted for, while generally believed to be of great influence and thus importance.

3 Definitions and terminology

We define a network as $G = (V, E)$ which consists of a set of nodes V and a set of edges E. Edge weights for node v and u are noted as $W_{(v,u)}$.

The samples collected from patients who were diagnosed for BV as negative, intermediate or positive, may also be referred to as the negative/intermediate/positive samples or group.

4 Methodology

4.1 Dataset

The dataset analyzed during this study pairs 16S rRNA gene profiles together with metabolite concentration profiles. The dataset was obtained from the study performed by Srinivasan et al.[25] (see section 2.1). It consists of 70 samples collected from 70 women, with the purpose to investigate metabolomic and taxonomic traits surrounding bacterial vaginosis (BV). Sample meta-data stating the diagnosis for each patient/sample, based on both the Nugent and Amsel criterion, were provided by the authors upon request. The raw 16S rRNA data was deposited in the NCBI Short Read Archive under accession number SRP056030. In terms of sparsesity, the taxonomic data consists of 89.6% zeros, whereas the metabolic data has 2.9%.

4.2 Technical setup

The devised framework was implemented on a multi-node computing cluster, using the Jupyter IPython[43] environment. This setup is very suitable for projects with a focus on data analysis. Additionally, this setup works very well together with the IPyParallel package (official extension of IPython) that simplifies the implementation of code that must run in parallel over multiple cores or nodes.

The framework's main process runs in a so called Docker[44] container. This setup is similar to the better known virtual machines, as it also supports isolated execution in an enclosed environment, allowing for parallel installation/execution of software that could possibly conflict otherwise. Furthermore, Docker container instances run on the same kernel, which greatly reduces the computational overhead that does come with the traditional virtual machines.

The Docker container enclosing the framework functions as the starting point of all analyses performed, hence also the execution of other pipelines are initiated from here. To this end, the core process/container is also made capable of starting new containers from in itself. Factually, these newly initiated containers do not run from inside the parent container, instead the parent container connects to the Docker daemon (i.e. manager of the entire Docker system) running outside all containers, which in turns starts a new container. In other words, while the processes are started in a nested way, the execution proceeds in parallel.

4.3 Data preprocessing

Taxonomic data

Quantitative taxonomic data should be normalized in attempt to account for the relativeness and uneven total abundance sums per sample. Although, several approaches were experimented with, the most reasonable approach seemed to be the one that is provided in the QIIME package[28], a utility capable of normalizing OTU tables using the metagenomeSeq's CSS[45] (cumulative sum scaling) transformation method.

The CSS method does not only re-scale abundance, in order to correct for uneven sequencing depths, it also attempts to correct the compositionality by rarefying the abundances per sample to a count that is constant for all samples while discarding low abundant OTUs. This is done with the goal to correct for the bias also induced by uneven sequencing depths; a sample with a greater sequencing depth has a greater chance to contain data on rare OTUs than a sample with a low sequencing depth. Many downstream analyses are sensitive to incorrect absences of low abundant OTUs. Rarefaction can give insight in to the possible error caused by this, by re-sampling with incrementing depths and re-performing the calculations (e.g. species diversity/richness) for each iteration. The error for each iteration potentially gives insight into the bias introduced by the uneven sequencing depths [46].

In the approach demonstrated by Srinivasan et al. abundances are log-transformed, in attempt to normalize the data. However, a caveat with this approach is that the data does not necessarily conform to normality. Secondly, the data is very sparse (i.e. contains many zero's, see section 4.1), and thus regular log transforms are not appropriate.

To cope with these aspects another approach was employed, one based on the Box-Cox power transformation, originally developed by Box and Cox[47], which is formalized as:

$$g(y,\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{when } \lambda > 0 \\ \log(y) & \text{when } \lambda = 0 \end{cases} \tag{1}$$

Where λ is the power by which all values will be raised to optimize the distribution towards normality. If $\lambda = 0$ then the logarithm of y will be used to transform, otherwise it would always result in 1. Finally, because the transformation requires positive non-zero data, a shifting constant c can be used to make the values great enough prior to the transformation. For this study c was set to 1.

Metabolic data

Metabolite concentration values were normalized by dividing each value by the mean of the corresponding sample. Zero-values were set to the lowest non-zero value of that metabolite. Finally, the data is log transformed. Regular log transform is appropriate in this case, because the zeros are not present in the data anymore.

4.4 Metagenome prediction

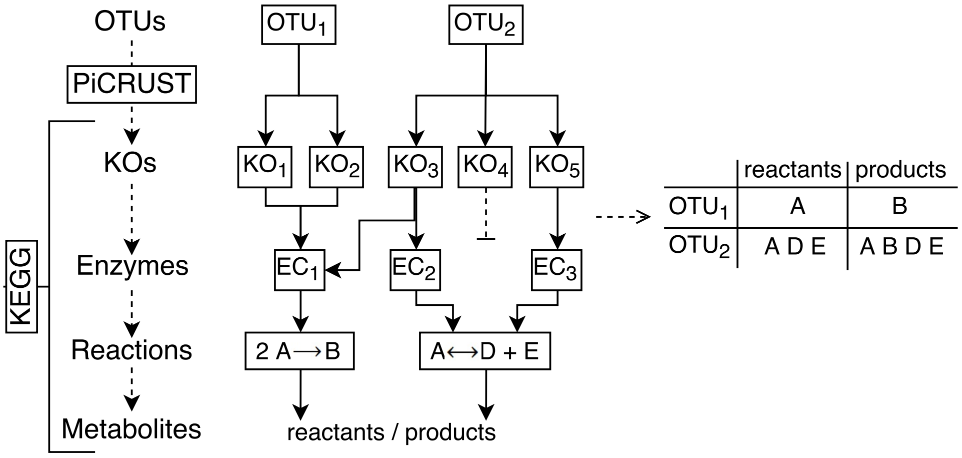
Insight into the functional capabilities of the community's organisms is essential for the process of elucidating the different functions and interactions of the community members. The 16S rRNA profiles merely give information into the composition. To this end, the PICRUSt [48] package was used to predict the metagenomes. This tool uses a vast reference database to link GreenGenes and KEGG orthology (KO) identifiers. Furthermore, the database contains data on the corresponding copy number variations of the marker genes as well as other genes (identified using KOs), this data is used to correct the OTU abundances and KO abundances (i.e. multiple marker genes give rise to over-quantification).

4.5 Involving KEGG data

The PICRUSt's predicted metagenomes are not yet coupled with biochemical data, which is essential when investigating species specific biochemical characteristics. The metagenomes are presented by KEGG orthology (KO) identifiers and per sample corresponding abundances. A KO identifier can represent multiple orthologous genes, which generally have similar chemical properties and therefore complement each other for missing data. The entire set of genes code for a variety of enzymes that potentially catalyze different biochemical reactions. The linkage between these different entities can be made using data obtained from the KEGG application programming interface (API). This source does not contain information on reaction directionality. KEGG does make this data available through their FTP interface, stored in the file called "reaction_mapformula.lst". This tab-delimited file states the reaction identifier, pathway identifier, and the reaction itself (e.g. C00022 => C06033). This representation of the reaction does not contain the compound coefficients, which therefore have to be obtained using the KEGG API and coupled with the corresponding compounds that occur in the reaction. After joining together the information from the different sources, the reactions are hashed to filter duplicate reactions. This hash operation accounts for possible different order of the compounds or a mirrored notation of the reaction formula, and thus only filters reactions that are duplicate from a biochemical perspective.

Due to license issues only the 2010 version of the "reaction_mapformula.lst" file was usable. The KEGG API, on the other hand, does not require a license for academic users.

After all data instances are connected via hash maps, connections between any entity can be made directly or indirectly, e.g. which OTU is believed to produce/consume which metabolites as well as vice versa. Figure 2 shows a schematic overview from the mappings made between the different entities.



	reactants	products
OTU1	A	B
OTU2	A D E	A B D E

Figure 2: schematic over simplified example of how different entities can relate to each other. Starting with the predicted metagenomic genetic content, represented as KOs, and further down to the encoded enzymes and finally the reaction and involved compounds.

4.6 Mantel test based correlation

In the paper by Noecker et al.[18] a demonstration is given of the application of a Mantel test[49] based Pearson correlation method. Mantel tests allow for computing the correlation coefficient between two distance matrices. This adapted approach of Pearson correlation aims at a greater degree of robustness, in particular for relative data.

The computation of the correlation coefficient between two data vectors v and w, starts by constructing for each vector a distance matrix D_1 and D_2, in which both columns and rows represent the samples. The cell values are established by subtraction of the two vector values that correspond to the two samples:

$$D_x = \begin{bmatrix} s_1 - s_1 & s_1 - s_2 & ... & s_1 - s_n \\ s_2 - s_1 & s_2 - s_2 & ... & s_2 - s_n \\ ... & ... & ... & ... \\ s_n - s_1 & s_n - s_2 & ... & s_n - s_n \end{bmatrix} \tag{2}$$

The Mantel test implemented into the package developed by Carr[50] was used during this project to compute the results.

Optionally, vectors that contain less nonzero values than a certain threshold could be discarded (threshold used by Noecker et al. was ≤ 5). This will however only discard vectors that would correlate badly anyways, meaning setting such threshold might give a deceiving impression that primarily strongly (positively or negatively) correlations were made.

The Pearson or Spearman correlation coefficient is then computed using a Mantel correlation test. Also a p-value and z-value is calculated by permutating the matrices and testing for a similar correlation coefficient, which is repeated 5000 times. A threshold for significance was set on <0.05 for the p-value and >1.96 for the absolute z-value. The p-values were corrected for multiple testing using the Bonferroni method, which is done by multiplying the p-values by the number of tests.

4.7 Identifying the signatures of bacterial vaginosis

4.7.1 Singular based signatures

The degree to which two distributions – of either a metabolite or OTU – differ between two different classes (e.g. control and experimental group) can give biological relevant insights into the data. The paper by Srinivasan et al. shows results that were obtained using the Welch's two sample t-test, quantifying for the OTUs and metabolites the specificity to BV (i.e. is a metabolite or OTU significantly absent/present with respect to the diagnosis). Only the results from the t-test results of the metabolites were reported in detail.

While the t-test is a common approach to test the specificity, it assumes that the distributions are normally distributed, because the method bases its calculations on the non-intersecting surfaces between the two distributions, while taking variances (allowed to be unequal if using Welch's t-test) into account and using the mean as center point.

Visual inspection as well as normality tests (using SciPY[51] normaltest[52]) showed otherwise (see section 5.3.2), the distributions are not normal in the majority of the cases, which in particular applies to the OTUs. To this end, results were reproduced using also other metrics that are valid in case of distributions not normal. Firstly, a Mann-Whitney U test was used, secondly Pearson and Spearman correlations were calculated between the OTU/metabolite abundances and the Nugent numeric score. Again the p-values were corrected for multiple testing using the Bonferroni method.

4.7.2 Multiplex signatures using decision trees

Previous section describes approaches to detect signatures of BV in a singular fashion, i.e. how does a single OTU or metabolite relate to the condition. These approaches ignore possible relations between a multitude of OTUs and/or metabolites, that could be of biological significance. To this end, the following section describes approaches to investigate more complex signatures of BV.

Estimating integration profitability

To consolidate the assumption that the microbiome and metabolome truly have a combined potential, for predicting and thus relating to the condition, an own devised method was applied on the data. This method leverages from the power of great numbers, by randomly constructing a great number of decision trees, each with a randomized composition of features (i.e. arbitrary OTUs and metabolites). Each tree was then evaluated by assessing its performance in predicting the condition per patient. Finally, the prediction scores are plotted versus the feature composition ratio, where more or less equal proportions of OTUs and metabolites indicates an information gain, and hence an increased combined potential. The highly iterative approach is supposed to ensure validity, since invalid interpretations due to coincidences are very unlikely.

The iteration proceeds as follows:

1. The joint dataset is composed of more OTUs than metabolites. Therefore, random sampling from both feature types should be corrected to prevent a bias towards OTU features. This is done by sampling M OTUs where M is the total number of metabolites, which in turn is joined with the metabolites to compose the joint set S where $|S| = M \cdot 2$.

2. Then using the Scikit's[53] ExtraTreeClassifier model – *"An extremely randomized tree classifier"* – random trees are built using the set of features S.

3. The constructed tree is trained to predict the condition and tested, using 10-fold cross-validation (i.e. the thresholds are reset and reestablished optimally for each fold). The prediction error is stored together with the ratio of OTUs and metabolites.

After repeating the steps numerous times, the results are plotted.

Using decision trees to detect OTU/metabolite interaction patterns

Previous section describes an approach to use decision trees to estimate the combined potential of the two datasets. Even though, this approach does yield trees that can predict the condition very well, these trees are not to be used for elucidating the complex interaction patterns of multiple OTUs as well as multiple metabolites. The trees are searched for in a high frequency fashion, without any pruning operations and hence the results are highly susceptible to overfitting. To this end, it is necessary to take preventive measures for overfitting in attempt to increase the likelihood that the trees generalize well on future and thus previously unseen data.

To decrease the chance on overfitting the first measure taken is imposing a size restriction on the tree's clades, in addition to limiting the depth of the tree, both operations prevent the tree from becoming fragmented and hence adjusting too much to the variance and noise in the data. Last but not least, the model is each time cross-validated using the leave-one-out fashion. The high number of folds is chosen to keep the amount of training data high.

4.8 Estimating per OTU contribution of metabolites

The contribution of an OTU to the production or consumption of a metabolite was previously investigated by Noecker et al.[18] (see section 2.2). The authors implemented a framework capable of predicting the relative metabolic turnover (RMT), of which the well predicted metabolites are then used to base interpretations on that relate to the contribution of an OTU with respect to that metabolite.

In contrast to the setup used by Noecker et al., this experiment is not based on a stoichiometry matrix. Instead, simple mappings are made which couple OTUs with metabolites as producers/consumers or both. The mappings are then used to predict an abundance for each metabolite for each sample, by summing and subtracting the producing and consuming OTU abundances, respectively.

This simplified approach excludes several components of the approach that were proposed by the authors, such as:

- Correcting for copy number variations of the genes represented by each KO.
- Exclusion of currency metabolites.
- Stoichiometric compounds coefficients.

The approach of leaving out these quantitative adjustments, could be defended with the hypothesis that a bacterium cannot synthesize twice the amount if the copy-number or stoichiometric is 2 instead of 1. Possibly, the production or consumption rate is primarily limited by the availability of the metabolic precursors, i.e. high concentrations of a required reactant of a reaction increases its rate.

4.9 An adapted PageRank algorithm

This section describes how the PageRank algorithm can be applied within the context of metabolic modeling, by adapting it only slightly. The traditional PageRank algorithm was used to compute a score of importance for web pages, to sort the results of Google's first search engine[54]. The algorithm is typically applied on a directed network where nodes represent websites that are connected with edges if one website links to another via at least one hyperlink. More important websites are more often linked than less important websites. Also, the importance of the source website is used to determine the target website's importance; a website is more important if linked by an important website than unimportant. Therefore, this is a circular problem, which is solved using an iterative approach, more thoroughly explain in section 4.9.2.

This topic shows resemblance with the problem of predicting metabolite concentrations from OTU abundances; the production rate of one metabolite is dependent on another, which in turn is dependent on another, and so on.

4.9.1 Reaction network

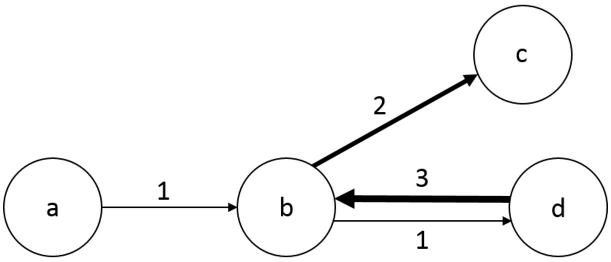
The directed network that PageRank is applied on, is constructed using the samples' taxonomic compositions in combination with the KEGG entity mappings. The nodes of the network represent the chemical compounds, whereas an edge between them represents a reaction involving them both, one as reactant and the other as product, corresponding to the edge's source and target node, respectively. This reaction must be catalyzed by at least one enzyme that is encoded by one of the OTUs. The edge weights are based on the sum of abundances of the OTUs that are responsible for that reaction. Greater edge weights result in a greater throughput of Page score. Therefore, a greater observed OTU abundance will lead to higher predicted metabolite concentrations.

Different setups were experimented with, in which only irreversible reactions were used, but also both reversible and non-reversible. In case of the latter, reversible reactions, meaning two nodes would be connected by two edges in both directions possibly of different weights.

4.9.2 Traditional PageRank algorithm

This section describes the traditional PageRank algorithm as proposed by Page et al. [54]. Each step of the algorithm is explained using over-simplified example data.

#1



Example of a reaction network, where N nodes $v \in G$ and edges $(v, w) \in E$ represent metabolite and reactions/enzymes, respectively.

#2

source	target	weight
a	b	1
b	c	2
b	d	1
d	b	3

Edges connecting nodes v, w with corresponding edge weight vector $\mathbf{w}$.

#3

Target

Source	**a**	**b**	**c**	**d**
a	0	1	0	0
b	0	0	2	1
b	0	0	0	0
d	0	3	0	0

Weighted adjacency matrix $M_{N,N}$

#4

$$\mathbf{s} = 1 \div \sum_{i=1}^{N} M_{i,:} = 1 \div \begin{bmatrix} 1 \\ 3 \\ 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{3} \\ 0 \\ \frac{1}{3} \end{bmatrix}$$

Calculate edge fraction vector $\mathbf{s}$.

#7

$M = M \cdot \mathbf{s} =$

	a	**b**	**c**	**d**
a	0	1	0	0
b	0	0	$\frac{2}{3}$	$\frac{1}{3}$
b	0	0	0	0
d	0	1	0	0

Dot product of M with vector $\mathbf{s}$ makes the outgoing edge weights per node proportional to one.

#8

$$\mathbf{p} = \mathbf{d} = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & ... & N \end{bmatrix}$$
$$\mathbf{x} = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & ... & N \end{bmatrix}$$

Where $\mathbf{p}$ and $\mathbf{d}$ are the personalization and dangling weights vector, respectively, which are set to $\frac{1}{N}$ if not specified otherwise. Finally, vector $\mathbf{x}$ will be used to store the PageRank scores.

#9

$$\mathbf{x}_{dangling} = \begin{cases} 0 & \mathbf{x}_i > 0 \\ 1 & \mathbf{x}_i = 0 \end{cases} \forall i \in \{1, 2, ..., N\}$$

Binary vector $\mathbf{x}_{dangling}$ represents the Booleans of a node being dangling (i.e. no outgoing edges) or not.

#10

$$\mathbf{x}_{previous} = \mathbf{x}$$
$$\mathbf{x} = \alpha \cdot (\mathbf{x} \cdot M + \mathbf{d} \cdot \sum(\mathbf{x} \cdot \mathbf{x}_{dangling})) + (1 - \alpha) \cdot \mathbf{p}$$
$$L_1 = \frac{1}{N} \cdot \sum|\mathbf{x} - \mathbf{x}_{previous}|$$

- Vector $\mathbf{x}_{previous}$ keeps track of the previous vector $\mathbf{x}$ so the change can be quantified for each iteration.
- Variable α represents the damping factor, which is also the proportion of the total PageRank score that is assigned to the personalized nodes according to personalization fractions in vector $\mathbf{p}$.
- Above calculations should be repeated until convergence which is if L_1 is lower than specified threshold T.

The algorithm is constructed such that the sum of PageRank scores is meant to sum up to one exactly, for each iteration. If this check fails, the algorithm is implemented wrongly and likely to yield incorrect results.

4.9.3 Algorithm adaptations

The original PageRank algorithm was adapted in multiple ways to make it more applicable within the context of metabolic modeling, resulting in the own devised method named "PageRank based metabolic modeling". Firstly, the before mentioned step #**4** is replaced with:

$$\mathbf{s} = 1 \div \sum_{i=1}^{N} M_{:,i} = 1 \div \begin{bmatrix} 0 \\ 4 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{4} \\ \frac{1}{2} \\ 1 \end{bmatrix}$$

Summing over the columns instead of the rows, gives the summed weights of outgoing edges instead of incoming edges.

Secondly, step #**5** is replaced with, where M is multiplied by **s** such that all resulting columns sum up to one, and therefore introducing proportionality. In terms of graph edges, this means that for each node the summed weight of the outgoing edges sum up to one:

$$M = M \cdot \mathbf{s} = \begin{array}{|c|c|c|c|c|} \hline & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} \\ \hline \mathbf{a} & 0 & \frac{1}{4} & 0 & 0 \\ \hline \mathbf{b} & 0 & 0 & 1 & 1 \\ \hline \mathbf{b} & 0 & 0 & 0 & 0 \\ \hline \mathbf{d} & 0 & \frac{3}{4} & 0 & 0 \\ \hline \end{array}$$

From a biological perspective, this means that the flux of a metabolite is primarily restricted by the abundance of the OTUs producing that metabolite, rather than the consuming OTU abundances.

Lastly, step #**8** is replaced with:

$$\mathbf{p} = \begin{bmatrix} \frac{\bar{p}}{N} & \frac{\bar{p}}{N} & ... & N \end{bmatrix}$$
$$\mathbf{d} = \begin{bmatrix} \frac{\bar{d}}{N} & \frac{\bar{d}}{N} & ... & N \end{bmatrix}$$
$$\mathbf{x} = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & ... & N \end{bmatrix}$$

Where $\bar{p}$ and $\bar{d}$ are constants that can be adjusted. Both the constants may be set to a value greater than one, meaning that the sum of personalization and dangling values per node may rise above one. This adjustments breaks the base principle of the original algorithm, that ensures all PageRank scores sum up to one. With this setup scores are believed to have more freedom, i.e. they are not restricted to a [0..1] range, which is deemed to be very restrictive when applying the algorithm to multiple samples.

4.10 Generating null-distributions

To validate the findings made using this approach two different methods will be used to generate null-distributions. The first approach aims at randomizing while retaining network topology, using edge shuffling. The second approach shuffles data as well, but only inside each class of the condition, with the goal to retain the bias that comes with the class separation (e.g. two OTUs that are similarly specific to BV are likely to correlate with each others metabolites and hence possibly falsely indicate consumption/production of each others metabolites).

Network shuffling

Network shuffling was performed by interchanging the source of an edge pair (i.e. a reaction between two metabolites), iteratively for the whole network. This approach ensures that the degree distribution is not affected and hence topology of the network is retained. Significant findings will indicate that the method is prone to spurious findings. Difference between the distributions of both the findings as well as the null distribution can be used to quantify a score relative to the null hypothesis.

Class corrected null-distribution

In attempt to construct a null-distribution, that is corrected for class induced biases, we propose the following algorithm that is based on the previously described (see section 4.8) method used for quantifying the contribution per OTU of each metabolite.

The algorithm operates according to the following steps:

1. Repeat for n iterations:
    (a) For each observed metabolite m:
        i. Acquire the set of OTUs o that is believed to be capable of producing metabolite m (inferred from both bi- and unidirectional reactions, see figure 2 of section 4.5).
        ii. Sum the abundances of the OTUs o per sample and use as predicted concentration for metabolite m.
        iii. Permutate the per sample predicted concentrations, only inside each class, i.e. permutate the values of the two classes separately.
        iv. Correlate the predicted concentrations with the observed concentrations for metabolite m and store the correlation coefficient, coupled with the link to metabolite m.

2. Take the mean of the correlation coefficients per metabolite and return the correlation coefficients for each metabolite.

The process is repeated until the calculated mean prediction scores have converged.

5 Results

This section describes the various results obtained using the methods described in the previous section. The results are presented primarily in a comparative manner with the studies performed by Srinivasan et al.[25] and Noecker et al.[18]. Furthermore, it demonstrates novel approaches for obtaining insights into the relations between the microbiome and metabolome with respect to bacterial vaginosis. The first part focuses on characteristics of the dataset such as composition of the microbiomes and metabolomes.

5.1 Exploring the dataset

5.1.1 Composition analysis

Figure 3 gives a first glimpse on the composition of the microbiome and metabolome. The samples were (horizontally) sorted using their mutual Bray-Curtis distance calculated between their taxonomic profiles. Since this sorting is NP-hard, a heuristic was used which starts with the most similar sample pair, which in turn is extended by attaching most similar samples on both sides until all samples are processed, this for both sample classes apart. The OTUs and metabolites were vertically sorted (per sample) based on the mean abundance/concentration over all samples (i.e. most prevalent appear at the middle of the diagram).

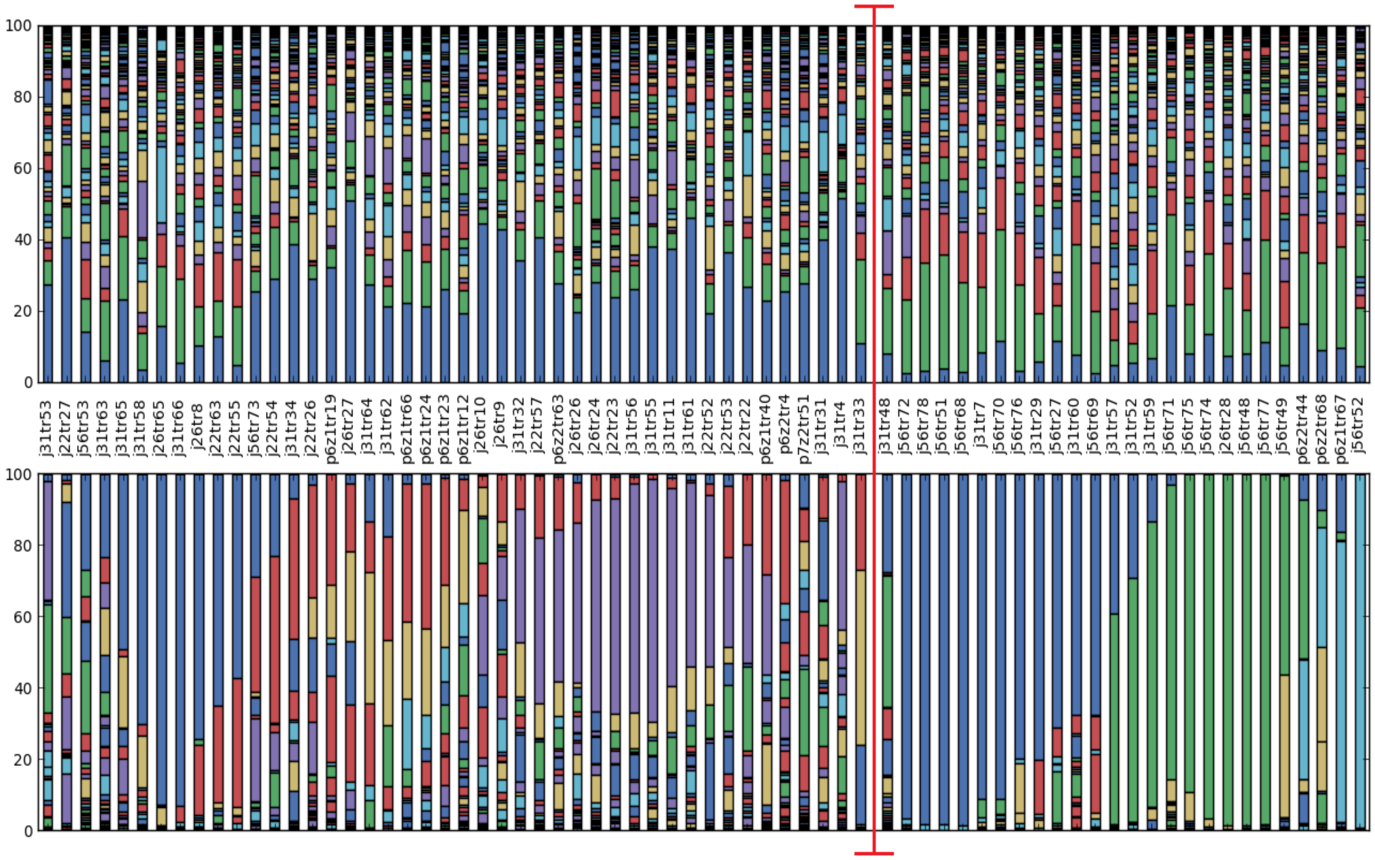


Figure 3: metabolic (top) and taxonomic (bottom) composition for each sample. Samples grouped per class, BV positive and negative, are shown left and right, respectively.

Based on only a visual inspect it already becomes clear that there is a distinct difference between the two classes, from both microbial and metabolic compositional perspective. Finally, the positive class appears to exhibit more divergence, by containing fewer very prevalent species.

5.1.2 Species diversity and richness

Figure 4 shows the species ecosystem's biodiversity and richness for all samples with respect to BV. Results were computed using the *core_ diversity_ script.py* utility of QIIME[28].

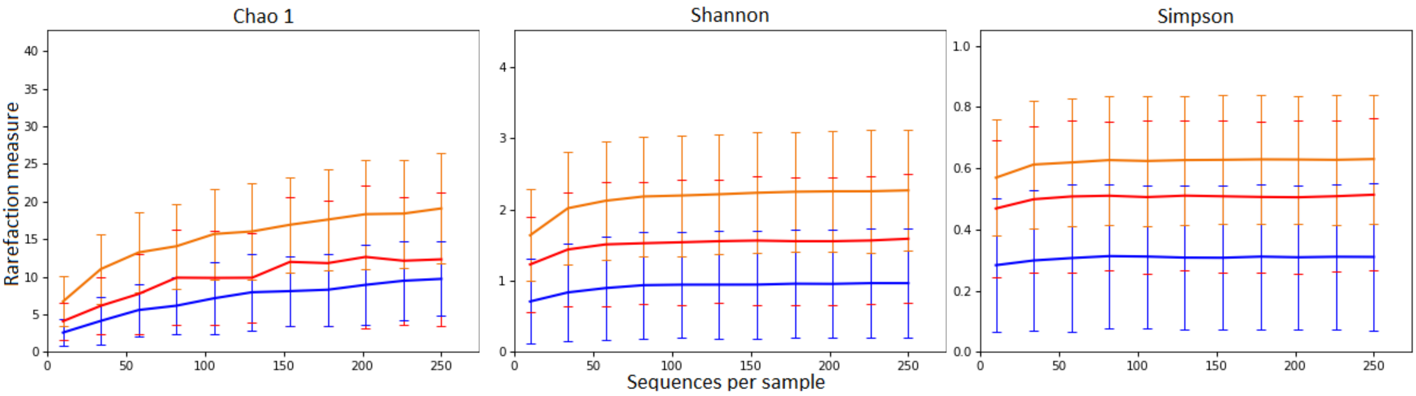


Figure 4: Showing three different species diversity/richness measures. Where the orange, red and blue line depict the positive, intermediate and negative group.

The Chao1 gives insight into the total richness by predicting the true number of OTUs, which is done by extrapolating on the number of rare OTUs, to estimate the number of OTUs missed out on due to undersampling. On the other hand, the Shannon (H') and Simpson's (D) index both serve the same goal of quantifying species diversity. However, the two measures are prone to different biases. The Shannon index gives more weight to rare than common species. It is also stated to favor consistency of species abundances in OTUs, whereas the Simpson's index discriminates the rare species, i.e. more strongly weighting the dominant/abundant. Also, a notable difference is that the latter represents a probability whereas the Shannon index does not[55–57].

The measures are plotted using rarefaction curves, to provide insight into possible deviations that are a consequence of too shallow sequencing. The curves do not show discrepancies between the different sampling points and hence it can be concluded that the measurements are unlikely to be invalid due to too shallow sequencing. Furthermore, a clear separation between the three different classes is visible. The communities of the positive group are most divers, with the intermediate group at the second place. This confirms the observations done previously at figure 3.

5.1.3 Data dimensionality & reaction direction

As previously mentioned, the dimensionality (i.e. number of features) of this dataset is very high in contrast to the number of samples. Also, the number of OTUs that operate with each metabolite relates to this same aspect; currency metabolites are likely to be involved by many organisms, which complicates the analysis. At the same time, reaction directionality is a complicating aspect. Firstly, reversible reactions are possibly very difficult to involve in the analysis, as the direction often depends on the current biochemical situation. Disregarding the reversible reactions, certain organisms may still synthesize as well as consume certain metabolites due to non-reversible reactions for that metabolite in both directions.

Figure 5 displays a summary of the number of OTUs involved with each metabolite, with corresponding role of each OTU, i.e. consuming, producing and both.

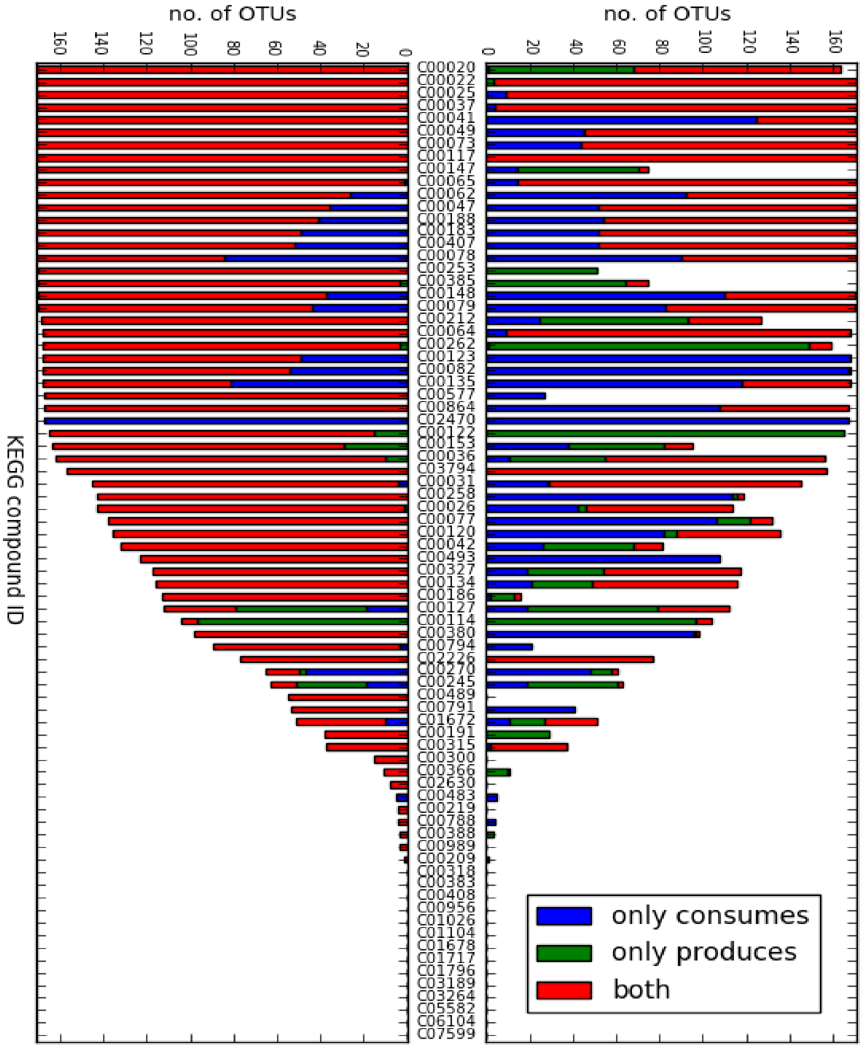


Figure 5: For each metabolite is shown how many OTUs operate with it, either consuming, producing or both. The left side shows this for all reactions, whereas the right side only presents the non-reversible reactions.

The figure shows that almost every metabolite is both consumed as well as produced, by at least one of the observed OTUs, but in almost half of the cases it is interacted with by roughly all OTUs. The plot shows a slight dominance for only consuming OTUs compared to only producing OTUs.

Excluding the reversible reactions decreases the number of OTUs that operate with the metabolite in both ways, significantly. On the other hand, the total number of OTUs that interact in at least one way is still quite high for most metabolites, from this perspective the dimensionality did not decrease much. Finally, the plot shows a clear dominance for OTUs only consuming in comparison to OTUs only producing.

5.2 Reproducing taxonomic classifications

The paired dataset was made available in different formats. The data was already processed to OTU abundances, where manual inspection showed that this representation of the data is very sparse. To ensure reasonable quality of the OTU classification, it was redone using the QIIME metagenomics analysis package [28]. The samples were first pre-processed by trimming of adapters and primers. The reads were then processed using the open reference OTU picking tool contained in QIIME. The similarity threshold that determines a match between a read or a reference OTU, was incremented iteratively and plotted against the performance measure (see figure 6). This performance measure (described in methodology section 4.8) gives an indication of how well the 16S rRNA profiles correspond to the metabolite concentrations, allowing for an assessment of the methods that preprocess the data.

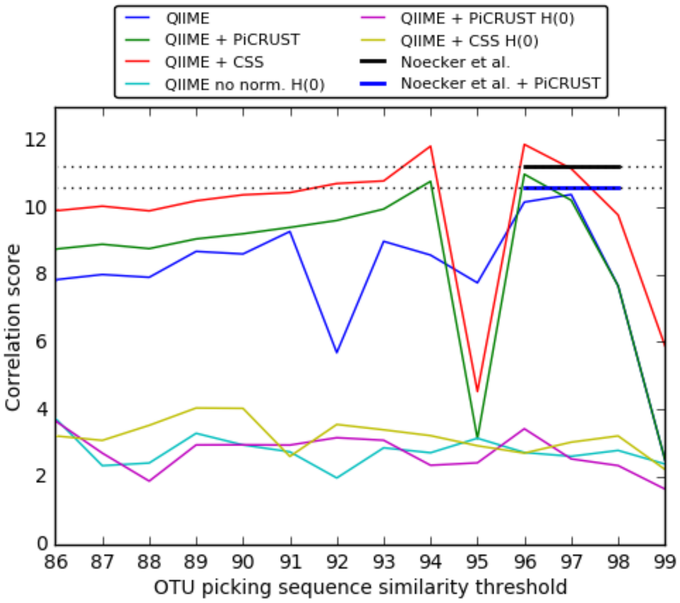


Figure 6: OTU classification quality based on a metric (see section 4.8) that gives a score indicating the correspondence between microbiome and metabolome data sets. Scores for the method used by Noecker et al. are only shown for 97% similarity.

The figure shows the best score for the OTU profiles that were self picked using the QIIME pipeline and normalized using the QIIME's built-in CSS normalization utility. For the same data the PICRUSt normalization tool did perform almost as well. The OTUs picked by Noecker et al.[18] turned out slightly inferior. The sole difference in setup is that Noecker et al. also rarefied the OTU counts to the total abundance of the sample with the lowest sequencing depth, which might degrade the succeeding normalization methods and possibly explain the slight degradation.

Comparing the CSS and PICRUSt normalization methods to the self picked OTUs without normalization, shows that both normalization methods appear to correct for fluctuations with a negative effect on the metabolite predictability. One peculiar outlier is shown at 95% similarity, where the non-normalized score makes a relatively small drop, but is amplified greatly by both normalization methods. No clear explanation is present for this yet. However, here it is important to note that missing KEGG data might be of influence. Also, the sensitive nature of clustering algorithms can lead to very different results if a tipping point is exceeded.

Finally, the null distributions (constructed as described in section 4.10) show that there is indeed a bias involved with this performance score, but not in close range of the experiments.

5.3 Signatures of bacterial vaginosis

This section describes obtained results that relate to biological patterns specific to BV. Both the metabolic and taxonomic characteristics were investigated using multiple methods, ranging from traditional PCA methods, to T-tests, to decision trees. Using these different approaches the signatures of BV are investigated for both metabolites and OTUs in a one-to-one fashion as a many-to-many fashion.

5.3.1 Principal component analysis

See figure 7 and 8 for a first insight into the separation of the BV positively and negatively diagnosed patients using the Amsel and Nugent criteria, respectively. The spatial separation is based on the 16S rRNA profiles together with the metabolite concentrations, joined using PCA of which the two most prevalent directions of variation define the 2D spatial separation.

The two plots clearly show two clusters for both classes, of which especially the negative cluster is relatively dense, whereas the positive cluster is sparser by showing a greater spread. Also is the positive class blend into the negative class and not vice versa, indicating the diseased patients are more variant in biological terms.

The quantitative score of the Nugent criterion shows coherence with the plot as the smaller sized nodes, and hence also the intermediates, tend to the left and blend into the negative class. Apparently, these patients were not separable from the negative class based on only a two component PCA dimension reduction. A better split might be achievable with the use of more advanced methods.

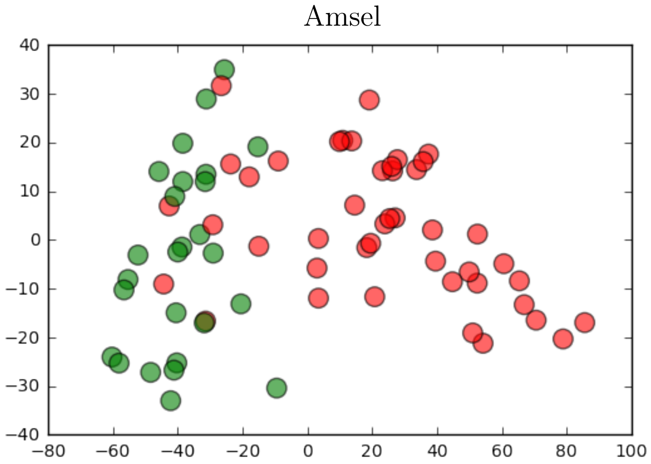


Figure 7: Scatter plot showing samples of BV positively (red) and negatively (green) diagnosed patients based on the Amsel criterion. Samples (i.e. data point) 2D locations were based on the first two components of a PCA applied on both datasets (i.e. 16S rRNA profiles and metabolite concentrations).

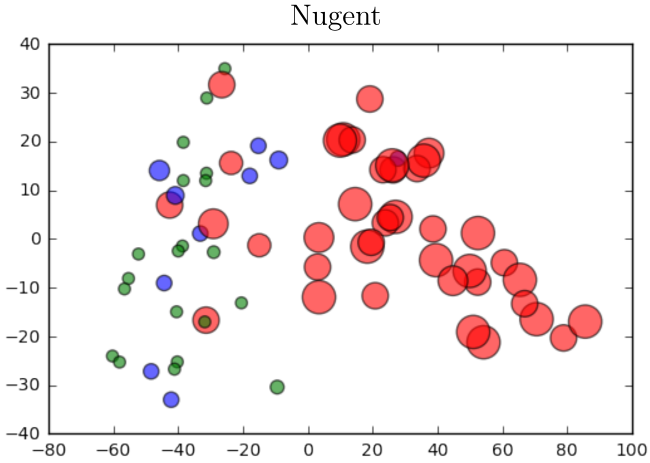


Figure 8: Same setup as figure 7, except the Nugent criterion was used instead om Amsel. Furthermore, the blue dots depict samples with intermediate Nugent scores. Finally, the dot sizes correspond to the numeric Nugent score.

5.3.2 Singular based signatures

Following the research by Srinivasan et al.[25], for each metabolite and OTU has been computed to what degree it is associated with BV. As explained in section 4.7.1, not only the Welch's t-test has been used, since this metric assumes normality and the tested distributions do not necessarily conform to this.

See table 1 for the top 30 metabolites most specific and nonspecific to BV, where the rank was based on the Mann-Whitney U test in ascending order.

See table 2 for the 30 most BV specific OTUs. In addition to this table, there were also OTUs transformed into an phylogenetic tree (Newick format) and visualized using the Dendroscope package[58], see figure 9.

Results show similar findings as done in the study done by Srinivasan et al. Also, no strong discrepancies are observed between the results of the different metrics. To conclude, the fact that these distributions rarely conform to normality, does not lead to significant problems when analyzed using metrics that do assume normality.

KEGG ID	Compound name	Mann–Whitney		Welch's t-test		Pearson		Spearman		Normality p-value
		U	p-val	t-val	p-val	r	p-val	r	p-val	
C01672	Cadaverine	6.0	0.00	17.45	0.00	0.85	0.00	0.78	0.00	0.00 / 0.00 / 0.03
C00134	Putrescine	6.0	0.00	17.18	0.00	0.81	0.00	0.72	0.00	0.00 / 0.02 / 0.05
C00483	Tyramine	8.0	0.00	11.99	0.00	0.76	0.00	0.71	0.00	0.00 / 0.20 / 0.90
C00082	L-Tyrosine	15.0	0.00	-11.94	0.00	-0.82	0.00	-0.79	0.00	0.01 / 0.22 / 0.57
C00794	D-Sorbitol	12.0	0.00	-11.09	0.00	-0.80	0.00	-0.79	0.00	1.00 / 0.60 / 0.01
C00489	Glutarate	17.0	0.00	10.79	0.00	0.79	0.00	0.78	0.00	0.05 / 0.51 / 0.36
C03189	DL-Glycerol 1-phosphate	39.0	0.00	10.45	0.00	0.72	0.00	0.69	0.00	0.00 / 0.11 / 0.01
C00209	Oxalate	41.0	0.00	-10.10	0.00	-0.73	0.00	-0.70	0.00	0.01 / 0.47 / 0.33
C00388	Histamine	43.0	0.00	9.46	0.00	0.71	0.00	0.69	0.00	0.00 / 0.44 / 0.03
C00079	L-Phenylalanine	48.0	0.00	-9.11	0.00	-0.78	0.00	-0.77	0.00	0.00 / 0.00 / 0.08
C00078	L-Tryptophan	53.0	0.00	-8.98	0.00	-0.79	0.00	-0.80	0.00	0.02 / 0.14 / 0.04
C00031	D-Glucose	55.0	0.00	-8.92	0.00	-0.71	0.00	-0.70	0.00	0.00 / 0.36 / 0.26
C00577	D-Glyceraldehyde	33.0	0.00	8.87	0.00	0.73	0.00	0.74	0.00	0.66 / 0.85 / 0.63
C00270	N-Acetylneuraminate	49.0	0.00	8.68	0.00	0.67	0.00	0.65	0.00	0.00 / 0.21 / 0.57
C00042	Succinate	55.0	0.00	8.38	0.00	0.72	0.00	0.74	0.00	0.00 / 0.39 / 0.01
C00315	Spermidine	54.0	0.00	8.24	0.00	0.67	0.00	0.67	0.00	0.24 / 0.60 / 0.44
C01026	N,N-Dimethylglycine	59.0	0.00	7.97	0.00	0.66	0.00	0.66	0.00	0.01 / 0.04 / 0.01
C07599	Alloxanthine	75.0	0.00	7.84	0.00	0.69	0.00	0.67	0.00	0.07 / 0.25 / 0.19
C00065	L-Serine	63.0	0.00	-7.73	0.00	-0.70	0.00	-0.74	0.00	0.13 / 0.92 / 0.09
C03264	D-2-Hydroxyisocaproate	80.0	0.00	6.82	0.00	0.65	0.00	0.62	0.00	0.05 / 0.01 / 0.90
C00047	L-Lysine	76.0	0.00	-6.25	0.00	-0.65	0.00	-0.63	0.00	0.58 / 0.00 / 0.55
C00073	L-Methionine	100.0	0.00	-5.74	0.00	-0.63	0.00	-0.64	0.00	0.04 / 0.00 / 0.01
C00408	L-Pipecolate	88.0	0.00	-5.71	0.00	-0.62	0.00	-0.61	0.00	0.83 / 0.01 / 0.49
C00123	L-Leucine	111.0	0.00	-5.67	0.00	-0.57	0.00	-0.56	0.00	0.07 / 0.02 / 0.52
C00188	L-Threonine	116.0	0.00	-5.54	0.00	-0.58	0.00	-0.58	0.00	0.29 / 0.64 / 0.62
C01717	4-Hydroxy-2-quinolinecarboxylic acid	115.0	0.00	5.31	0.00	0.60	0.00	0.58	0.00	0.01 / 0.05 / 0.39
C00318	L-Carnitine	132.0	0.00	-5.05	0.00	-0.52	0.00	-0.55	0.00	0.71 / 0.54 / 0.27
C00380	Cytosine	128.0	0.00	5.04	0.00	0.56	0.00	0.50	0.00	0.02 / 0.27 / 0.67
C00186	(S)-Lactate	120.0	0.00	-5.03	0.00	-0.54	0.00	-0.52	0.00	0.01 / 0.00 / 0.81
C02470	Xanthurenic acid	123.0	0.00	-4.90	0.00	-0.54	0.00	-0.57	0.00	0.05 / 0.08 / 0.02

Table 1: Table shows the 30 metabolites most specific to BV, sorted on the scores computed using the Mann-Whitney U test. Normality p-values are calculated for the distributions of all, only negative and only positive samples, respectively. See table 21 (appendix 6.2) for the whole table.

Taxonomy	Mann–Whitney		Welch's t-test		Pearson		Spearman		Normality p-value
	U	p-val	t-val	p-val	r	p-val	r	p-val	
f:Bifidobacteriaceae g:Gardnerella	16.0	0.00	7.26	0.00	0.73	0.00	0.49	0.00	0.00 / 0.00 / 0.00
f:Veillonellaceae g:Dialister	18.5	0.00	12.71	0.00	0.84	0.00	0.82	0.00	0.00 / 0.00 / 0.00
f:Lactobacillaceae g:Lactobacillus	24.0	0.00	-9.67	0.00	-0.77	0.00	-0.79	0.00	0.00 / 0.16 / 0.02
f:Prevotellaceae g:Prevotella	31.0	0.00	11.74	0.00	0.80	0.00	0.73	0.00	0.00 / 0.01 / 0.00
o:Coriobacteriales f:Coriobacteriaceae	33.5	0.00	13.99	0.00	0.81	0.00	0.69	0.00	0.00 / 0.00 / 0.00
f:Veillonellaceae g:Megasphaera	47.0	0.00	11.09	0.00	0.77	0.00	0.77	0.00	0.63 / 0.00 / 0.00
f:Leptotrichiaceae g:Sneathia	56.0	0.00	11.70	0.00	0.76	0.00	0.73	0.00	0.76 / 0.00 / 0.00
f:Aerococcaceae g:Aerococcus	71.0	0.00	9.60	0.00	0.61	0.00	0.40	0.00	0.00 / 0.00 / 0.00
f:[Tissierellaceae] g:Peptoniphilus	93.0	0.00	7.99	0.00	0.68	0.00	0.67	0.00	0.99 / 0.00 / 0.01
f:Gemellaceae g:Gemella	100.0	0.00	10.02	0.00	0.65	0.00	0.61	0.00	0.00 / 0.58 / 0.05
f:[Tissierellaceae] g:Parvimonas	105.5	0.00	7.08	0.00	0.64	0.00	0.69	0.00	0.00 / 0.00 / 0.09
f:Clostridiaceae g:Clostridium	110.0	0.00	9.88	0.00	0.73	0.00	0.79	0.00	0.52 / 0.58 / 0.01
f:Actinomycetaceae g:Mobiluncus	180.0	0.00	6.54	0.00	0.58	0.00	0.64	0.00	0.00 / 0.58 / 0.00
f:Porphyromonadaceae g:Porphyromonas	200.0	0.00	6.01	0.00	0.53	0.00	0.57	0.00	0.00 / 0.58 / 0.00
f:Lachnospiraceae g:Shuttleworthia	210.0	0.00	5.77	0.00	0.57	0.00	0.64	0.00	0.00 / 0.58 / 0.00
g:Lactobacillus s:iners	220.0	0.00	-1.16	18.00	-0.22	5.04	-0.49	0.00	0.00 / 0.00 / 0.00
f:Actinomycetaceae g:Actinomyces	240.0	0.00	4.99	0.00	0.39	0.00	0.35	0.00	0.00 / 0.58 / 0.00
g:Lactobacillus s:reuteri	240.0	0.00	-3.50	0.00	-0.52	0.00	-0.48	0.00	0.00 / 0.01 / 0.58
f:[Tissierellaceae] g:Anaerococcus	253.5	0.72	2.70	0.72	0.33	0.72	0.33	0.00	0.00 / 0.02 / 0.00
f:Actinomycetaceae g:Arcanobacterium	260.0	0.00	4.31	0.00	0.35	0.00	0.37	0.00	0.00 / 0.58 / 0.02
c:Bacteroidia o:Bacteroidales	260.0	0.00	4.51	0.00	0.49	0.00	0.58	0.00	0.00 / 0.58 / 0.00
c:Clostridia o:Clostridiales	260.0	0.00	4.46	0.00	0.33	0.72	0.31	0.72	0.00 / 0.58 / 0.01
o:I025 f:Rs-045	290.0	0.72	3.78	0.00	0.35	0.00	0.39	0.00	0.00 / 0.58 / 0.01
g:Lactobacillus s:coleohominis	292.0	0.00	-2.51	1.44	-0.38	0.00	-0.39	0.00	0.00 / 0.05 / 0.00
g:Peptostreptococcus s:anaerobius	294.0	1.44	2.53	0.72	0.25	2.16	0.22	5.04	0.00 / 0.00 / 0.00
o:Lactobacillales f:Streptococcaceae	300.0	0.00	-2.46	1.44	-0.36	0.00	-0.32	0.72	0.00 / 0.03 / 0.58
f:Fusobacteriaceae g:Fusobacterium	300.0	0.72	3.36	0.00	0.32	0.72	0.35	0.00	0.00 / 0.58 / 0.00
f:Mycoplasmataceae g:Mycoplasma	310.0	0.72	3.17	0.00	0.30	0.72	0.31	0.72	0.00 / 0.58 / 0.00
o:Bacteroidales f:S24-7	310.0	0.72	3.30	0.00	0.32	0.72	0.35	0.00	0.00 / 0.58 / 0.00
f:Lachnospiraceae g:Moryella	310.0	0.72	3.31	0.00	0.35	0.00	0.41	0.00	0.00 / 0.58 / 0.00

Table 2: Table shows the 30 OTUs most specific to BV, sorted on the scores computed using the Mann-Whitney U test. Normality p-values are calculated for the distributions of all, only negative and only positive samples, respectively. See table 22 (appendix 6.2) for the whole table.

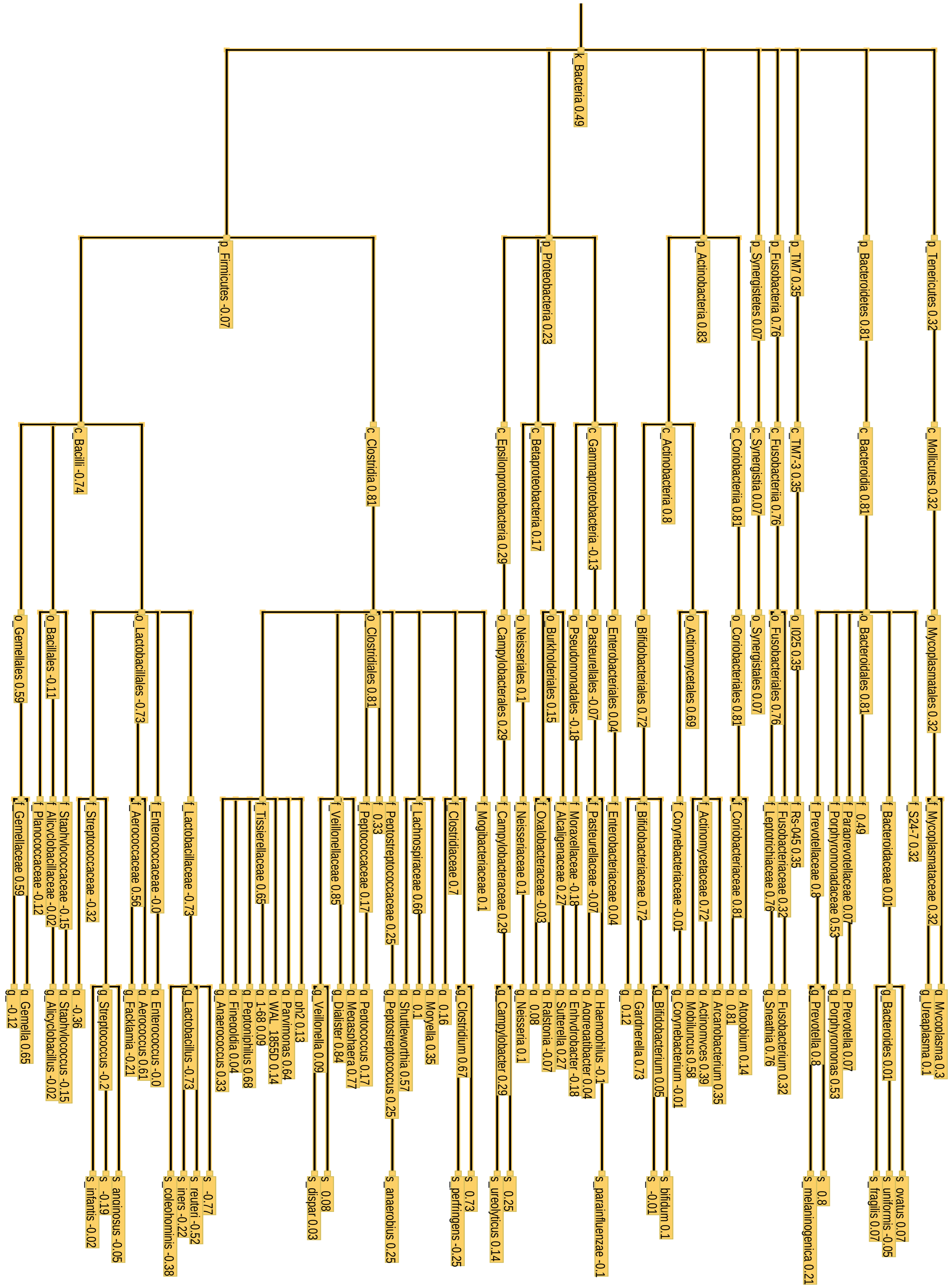


Figure 9: Phylogenetic tree of Pearson correlations also shown in table 2.

5.3.3 Multiplex signatures

Previous section describes the detected signatures of bacterial vaginosis in a singular fashion, i.e. how a single OTU or chemical compound relates to the condition. Also of interest is the way how different entities are specific to the condition with respect to each other. For instance, how one or more OTUs correspond to the quantified symptoms with respect to one or more metabolites. Detected patterns of such kind possibly give a reflection of the activity of the organisms that are involved. The patterns are searched for using decision tree classifiers. This is a flexible approach that keeps the results very interpretable.

Prior to the analysis the data is normalized to account for the relativeness of the quantities. This is done in the same way as in the preceding experiments, but followed up by scaling each abundance/concentration between the range of 0 and 1, meaning for each entity its lowest occurrence value is transformed to zero, whereas the highest is transformed to one. This eases the interpretation of the decision trees, but is of no further influence, since decision trees can work with fractions evenly well.

Estimating integration profitability
Firstly, a experiment (see section 4.7.2) was performed to investigate the information gain made by combining the taxonomic and metabolic data. This was done by generating more than 200k random trees. For each tree the performance was determined using 10-fold cross-validation. The tree prediction scores were then plotted versus the feature composition ratio (i.e. proportion of metabolites and OTUs), shown in figure 10.

The figure shows that well predicted trees generally favor somewhat equal proportions of metabolite and OTU features. It also shows that the not well generalizing trees generally have a higher proportion of metabolites as features, which suggests that OTUs have a greater relevant information density than metabolites. This might be a result of the metabolite specific nature of targeted metabolomics, as it measures only a small subset of the entire collection of present metabolites. Also does a OTU represent a greater number of metabolites by consuming/producing it, meaning it could therefore harbor more information.

The bottom section of the plot, shows that the standard deviation of the feature ratio decreases with an increasing tree score, indicating that a greater tree score comes with consistency in favorability for feature composition, which in turn substantiates the observations stated in the preceding paragraph.

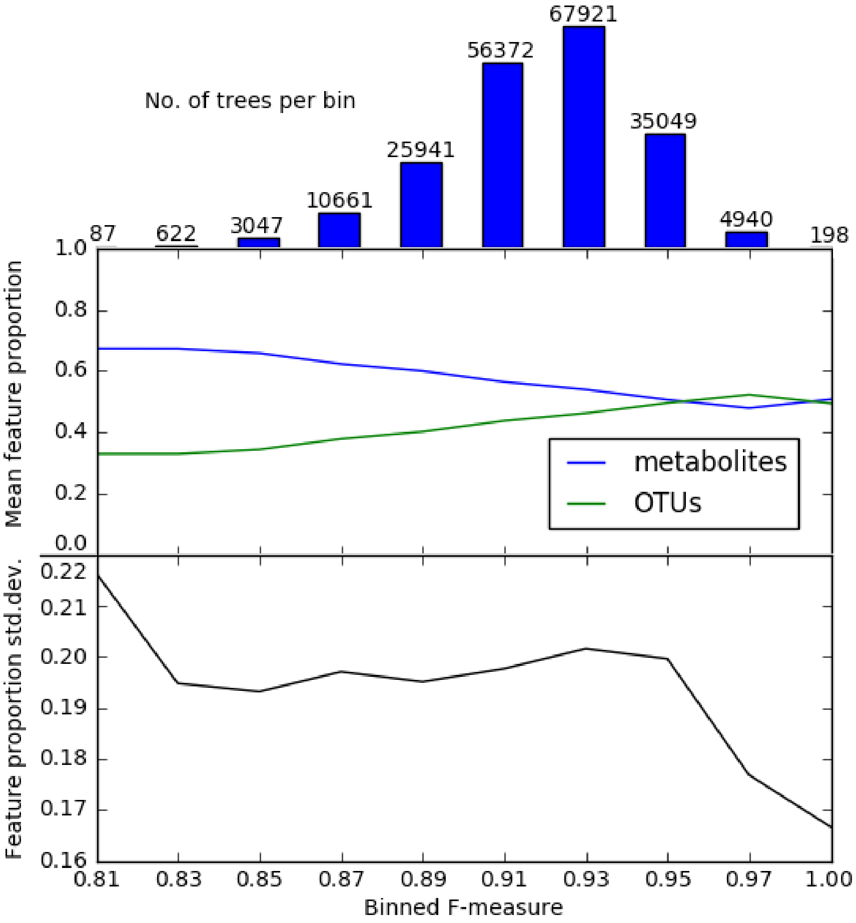


Figure 10: The three subplots, from top to bottom, present: the number of trees per bin, the bin average feature composition proportions and the feature ratio standard deviation.

Detecting multiplex relations using decision trees

This section describes detected biological patterns that involve a multitude of OTUs and/or metabolites, obtained using decision trees as described in section 4.7.2. Due to the non-negligible chance on overfitting, two preventive measures were taken, the trees are only allowed to grow 4 levels deep (i.e. longest path from root node to a node). Furthermore, each leave must represent at least 4 samples. This approach prevents the tree to sprout very specific and fragmented branches. Finally, all tree scores are determined while using leave-one-out cross-validation.

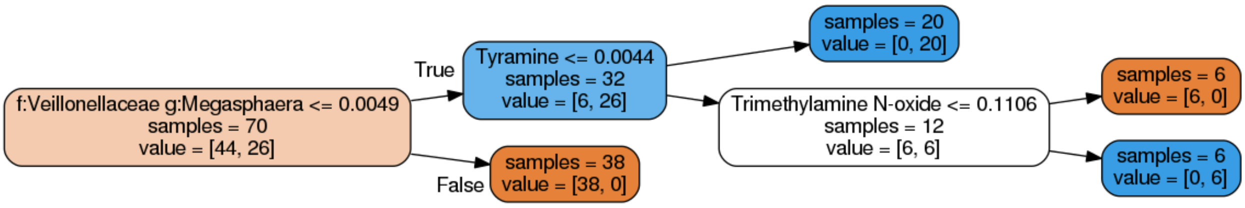


Figure 11: the first tree showing a biological pattern, relating to bacterial vaginosis, involving multiple metabolites as well as OTUs.

Figure 11 shows a tree scoring well in predicting BV. According to the tree the genus *Megasphaera* was found to have the tendency to be present in case of a positive diagnosis. Referring back to table 2 the same interpretations can be made based on the two-sample t-tests. More specifically, if *Megasphaera* genus class is relatively absent, the condition is likely to be negative, but only if trimethylamine N-Oxide is relatively absent.

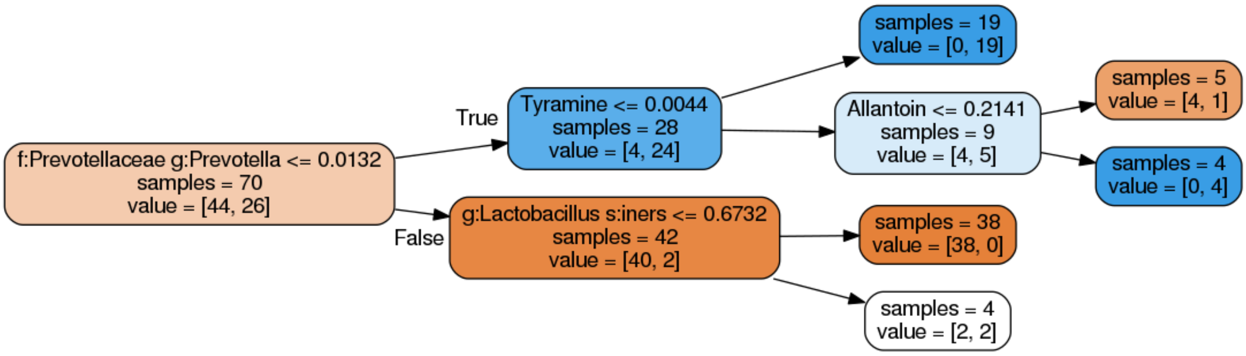


Figure 12: second tree, same setup, except a different pattern.

Figure 12 shows the second biological pattern detected, stating that a presence of the *Prevotella* genus class is associated with a positive diagnosis of BV. On the other hand, when the *Prevotella* class is relatively absent, it strongly associates to a negative diagnosis if tyramine is absent. If the latter is relatively present, the pattern states that an absence of allotoin corresponds to a positive diagnosis.

More trees were found that perform good in predicting bacterial vaginosis, but do not involve both metabolites as well as OTUs and thus do not give insight into potential interaction patterns of OTUs via metabolites. Due to the high dimensional nature of the data, it is likely to also find irrelevant trees.

5.4 Estimating OTU contribution of metabolites

This section presents the results of the OTU/metabolite contribution approach as explained in section 4.8. The method's performance is assessed by comparing it with a null-distribution and also with the method proposed by Noecker et al.

Confirming joint OTU prediction potential

The results of the following experiment were acquired to confirm that combining OTUs increases the coupling force with metabolites, in contrast to one-to-one relations between OTUs and metabolites. The former statement is in line with the presumption that most metabolites are interacted with by a multitude of OTUs, whereas the latter statement represents the opposite. The difference between the two is investigated by making correlations in two different ways:

1. For each metabolite the mean correlation between the metabolite's concentration and the abundances of the OTUs that operate with the metabolite.

2. For each metabolite the correlation between the metabolite's concentration and the sum of abundances of the OTUs that operate with the metabolite.

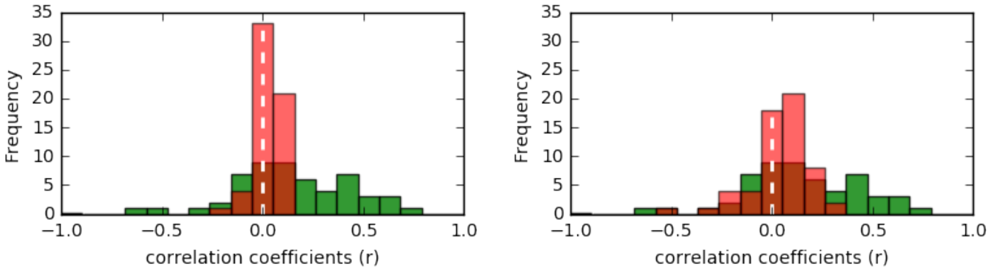


Figure 13: For both plots the prediction correlation coefficients are shown in green, whereas the OTU/metabolite correlation coefficients, averaged per metabolite, are shown in red. Left a normal mean and right a weighted mean based on the mean OTU abundance.

The plot shows a greater spread for the predictions (green) than the individual correlations (red), disregarding the use of mean or weighted mean, indicating that correlating sums of involved OTU abundances with metabolites is stronger than correlating OTUs with metabolite in a one-to-one fashion.

Adapted method for predicting relative metabolic turnover

This section section describes the results obtained with the reproduced algorithm by Noecker et al. as well as the simplified version as explained in section 4.8. The results are each time compared between two methods, of which one can be a null-distribution that is constructed as explained in section 4.10. Predictions are assessed by correlating the observed and predicted metabolite concentrations using Pearson's correlation. The obtained coefficients are visually compared with scatter plots, where

	H(0)	Abundances summed	H(0)	Noecker et al.	Noecker et al.	Abundances summed
$\mu(r < 0)$	-0.07	-0.18	-0.05	-0.18	-0.18	-0.18
$\mu(r > 0)$	0.09	0.20	0.10	0.18	0.18	0.23
$\mu\ r$	0.03	0.10	0.05	0.02	0.02	0.13

Table 3: all scores for figures 14 and 15 respectively.

each of the two axes (i.e. x and y) represents a method. In addition, the performance scores are for each method numerically summarized using three metrics (see table 3), the mean of: only negative correlation coefficients, only positive coefficients and finally all the coefficients.

Figure 14 shows two scatter plots comparing the prediction results for the own devised method and that of Noecker et al., both compared with the class-corrected null-distribution.

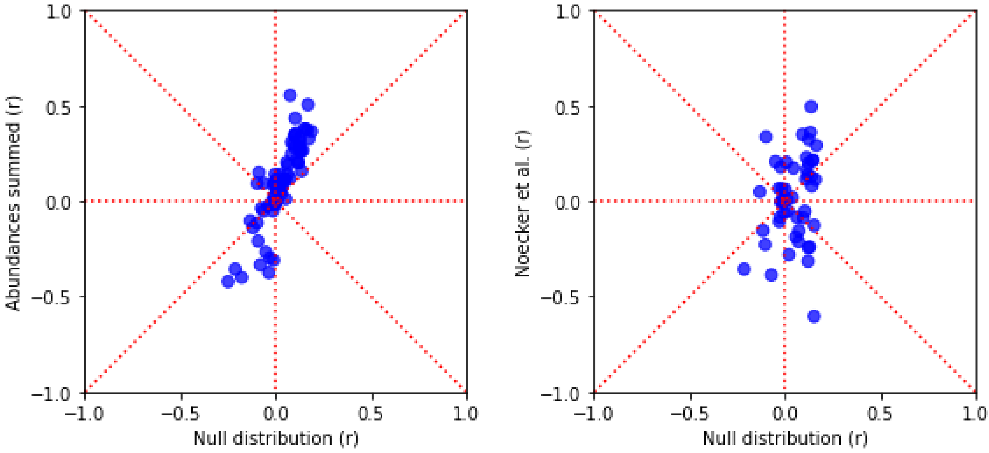


Figure 14: Left plot shows the Pearson correlation scores of the predicted metabolite concentrations using the own devised method versus the null-distribution. Right plot is set up in the same way, except the method assessed is that of Noecker et al.

For both methods the plots clearly show a vertical stretch, indicating that both methods are capable of predicting metabolites better than random, i.e. generating patterns biologically more relevant than the null-distribution, either negatively or positively. The own devised method also appears to have a shift to the higher correlation coefficients, meaning less metabolites are negatively predicted and the positive correlations are stronger. To consolidate this observation, the methods are also compared with each other in figure 15.

The plot shows that a small majority of the metabolites – predicted positively by both methods – were predicted better using the own devised method. It also shows that the own devised method was more capable of positively predicting metabolites which the method by Noecker et al. predicted negatively. Last but not least, the own devised method was capable of predicting 77 metabolites versus 49 metabolites using the method by Noecker et al., being a consequence of technical aspects as explained in 2.2.

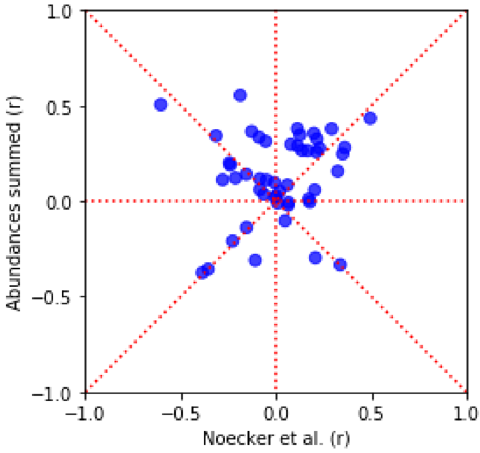


Figure 15: This plot shows the Pearson correlation coefficients of both methods plotted against each other.

5.5 PageRank based metabolic modeling

This section describes the results that were found using the own devised *PageRank based metabolic modeling* (PBMM) algorithm, as described in section 4.9.3. Predictions were again assessed by computing Pearson correlations coefficients between the observed and predicted metabolite concentrations.

The first two plots (figure 16) show for both the PBMM method and the method of Noecker et al., the performance compared against the null-distribution.

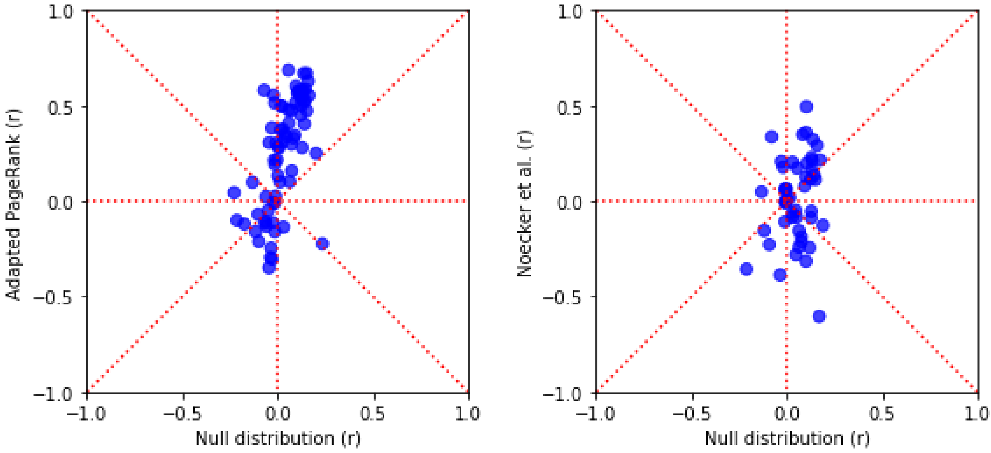


Figure 16: performance for the PBMM method (left) and method by Noecker et al. (right) compared with the null-distribution.

The plots for both methods show a vertical stretch, indicating predictions are better than random. Furthermore, PBMM method has greater correlation coefficients for negatively as well as positively predicted metabolites, whereas the method by Noecker et al. is more balanced in both directions. These visual observations are confirmed by table 4.

	H(0)	PBMM	H(0)	Noecker et al.	Noecker et al.	PBMM	PBMM positives	PBMM negatives	Noecker positives	Noecker negatives
$\mu(r<0)$	-0.06	-0.15	-0.06	-0.18	-0.18	-0.11	-0.13	-0.17	-0.14	-0.14
$\mu(r>0)$	0.09	0.41	0.10	0.18	0.18	0.45	0.27	0.48	0.23	0.20
$\mu\ r$	0.03	0.27	0.05	0.02	0.02	0.34	0.16	0.38	0.05	-0.01

Table 4: all scores for figures 16, 17 and 18, respectively.

Figure 17 shows the same predictions for both methods but compared against each other.

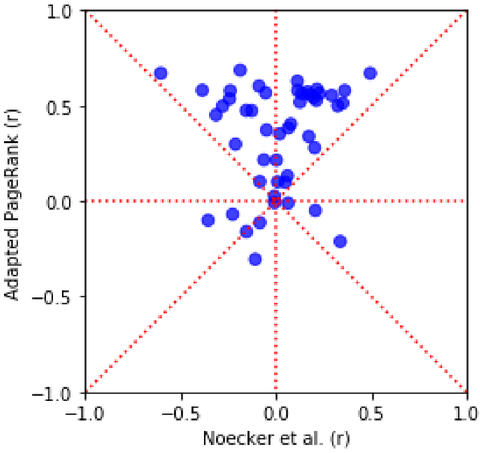


Figure 17: performance of the PBMM algorithm plotted versus the performance of the method proposed by Noecker et al.

The plot shows a significant greater performance for the PageRank based approach. Not only are the positive correlations stronger, but also are the negative correlations fewer.

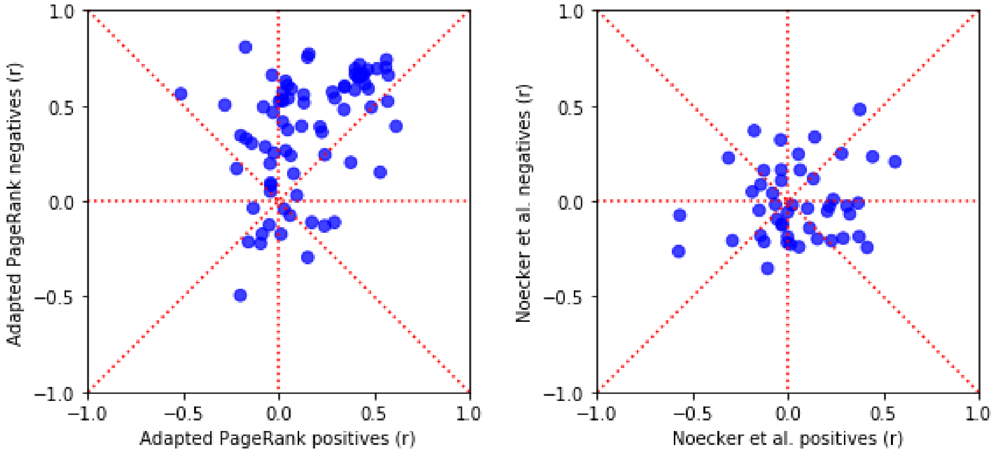


Figure 18: both scatter plots show the correlations coefficients for the positive class (y-axis) plotted against those of the negative class (x-axis). Left: PBMM algorithm. Right: method proposed by Noecker et al.

Figure 18 shows for both methods the difference of performance when applied only on one of the two classes separately, i.e. comparing the correlations coefficients of observed and predicted metabolite concentrations of only positively diagnosed patients versus only negatively diagnosed patients using the Amsel criterion. This comparison aims at determining whether the methods have a preference – in terms of performance – for one of the two classes.

In case of the PBMM algorithm, there is a significant difference. The metabolites are predicted more accurately for the negatively diagnosed patients. This remarkable difference is not shown by the method of Noecker et al. This significant finding might relate to the nature of the algorithm; converging to a balance before returning the predictions. This assumption possibly coincides with observations done in previous sections 5.1.1 and 5.1.2, where it became clear that the positive class is more divergent from both a microbial as well as metabolic perspective. Similarly, the PCA plots in figure 7 (section 5.3.1) display a greater cluster density for the negative class, contrary to the positive class which exhibits a greater variance by blending into the negative class.

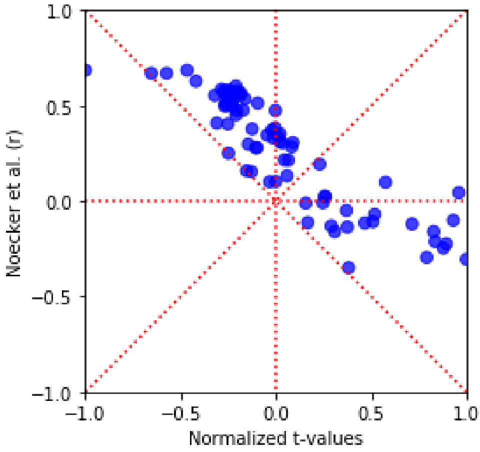


Figure 19: this comparative scatter plot shows how the metabolite concentration predictions relate to BV specificity quantified by two-sample t-tests.

To obtain a more precise insight into the class specific performance of the PBMM method, figure 19 shows a comparative scatter plot between the metabolites predicted over all samples versus metabolite t-values, obtained as explained in section 5.3.2. The plot shows that BV aspecific metabolites tend to be very well predicted up to $r \approx 0.8$, where the prediction performance appears to level off. Similarly, the metabolites specific for BV tend be predicted very badly, since all predictions approximate the zero point.

To further confirm the BV class specific performance difference, figure 20 shows the metabolite correlation coefficients plotted against the PageRank iterations, predicted on both patient classes separately, as well as combined. The three plots confirm the previous findings. In addition, the course of convergence per class is elucidated. The negative class initiates a stronger convergence at iteration 11, whereas the positive class does that at iteration 14.

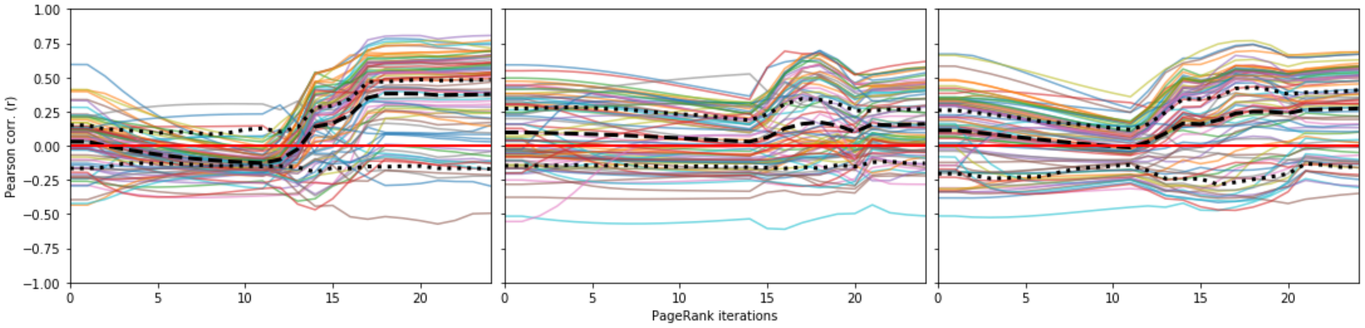


Figure 20: the above plots show the correlation coefficients plotted over the PageRank iterations, for each metabolite separately, tested on the negative, the positive and all samples, in respective order. The black lines represent the means in respective order of the positive, negative and all correlation coefficients.

Not negligible is the observation that the negative class does not exceeds its optimum, whereas the positive class does decrease in performance around iteration 20.

6 Discussion

6.1 Concluding the findings

With this study we showed that integrating the microbiome and metabolome gives opportunity to extract more information, in addition to the opportunity to use observed metabolite profiles to assess and validate predictions made using the microbial data.

Firstly, we showed that observed metabolite profiles can be used to assess taxonomic classifications. This showed a discrepancy at a similarity threshold of 97%, where the OTU picking assessment method indicated an abrupt degradation of taxonomic quality. This is most likely caused by a significant different taxonomic classification. Possibly, also missing KEGG data plays a role, the true cause is however yet to be pinpointed. Nonetheless, the method of assessing showed that the normalization methods proved capable of mitigating the negative effects, in addition to improving the performance for other similarity thresholds. Of the two normalization methods the QIIME CSS method proved superior in comparison to the PICRUSt normalization method, for at least this dataset as well as technical setup.

Next, we confirmed the study by Srinivasan et al. by finding similar metabolic and microbial signatures of bacterial vaginosis (BV). More interestingly, the multiplex patterns detected with decision trees, gave insights into both the data as well as the condition on another level. To confirm the potential of combining the two types of data, a method was devised that scores randomly constructed decision trees and reflects the performances back to the proportions of the feature types (i.e. taxonomic or metabolic features). The results show that an information gain is to be made by combining the two datasets. In turn, decision trees were constructed to detect patterns involving both microbial and metabolic species, possibly elucidating complex interaction patterns between them. Even though, preventive measures were taken to prevent overfitting, this approach remains prone to overfitting as well as accidental findings due to the high dimensions of the data in combination with the limited number of samples.

Estimating OTU contribution of consumption/production of metabolites

The paper by Noecker et al.[18] states how predicting the relative metabolic turnover (RMT) can be used to couple OTUs with metabolites as the responsible consumers or producers. Interpretations are only stated to be justified if metabolites are predicted well. In section 5.4 we showed that their approach to predicting the RMT can be improved by simplifying it. For instance, not involving copy-number variations (of both 16S rRNA gene and enzyme encoding genes) and not involving stoichiometrics (i.e. excluding compound reaction coefficients) showed that metabolites can be predicted more accurately. Adjustments to their algorithm possibly gives opportunity to new interpretations.

PageRank based metabolic modeling algorithm

Finally, we showed that an algorithm invented decades ago can be revived and applied on biological data by adapting it only slightly, with the goal to make it applicable within the context of metabolic modeling. We named this variant the *PageRank based metabolic modeling* (PBMM) algorithm.

The algorithm was originally designed to aim for and convergence to a steady network state. Both the microbiome and metabolome are generally believed to subsist also on balance, whether or not possibly disturbed by a medical condition or other external factors. Our method showed greatest promise (see section 5.5) on the samples obtained from patients diagnosed negatively for bacterial vaginosis (BV), whereas it lost performance on the samples that were obtained from patients diagnosed positively. This prevalent and distinct difference is most likely to be contributed to the increased community complexity (i.e. richness and diversity) that comes with BV (see section 5.1.2). Another possibility, considered less likely, is that the algorithm cannot work well with communities that are out of balance due to a disease, and subsequently results in an algorithmic convergence that is not a good representative of reality. This two possible explanations can also very likely both be true, whether or not to different degrees. Further investigation is required to give confidence on this matter. Future studies should also be performed on a biological replicate to exclude possible miscomprehensions due to peculiarities in this data. Our believe is that stress applied on the communities do allow for strong patterns to be detected, but should be limited to not run into irrelevant situations. For instance, the study performed by Noecker et al. also applied their algorithm on gut communities treated with antibiotics, which in our humble opinion is a dataset that gives rise to spurious findings.

The comparable method, flux balance analysis, requires high quality data as well as manually curated models, to be able to return sensible results. The PBMM algorithm showed to that it can operate on data lacking quality in multiple perspectives and hence considered robust with respect to poor data quality. After all, the 16S rRNA amplicon data does not allow for identification on strain level, in addition to being dependent on vast reference databases such as PICRUSt. Moreover, Roche 454 pyrosequencing data is not anymore considered as high standard due to the shallow depth of the sequencing, which possibly explains the very sparse data. Last but not least, the KEGG data from 2010 is likely to miss out significantly, and better results are likely to be obtained when the latest version is used.

6.2 Proposition for future research

As stated before, the PBMM method should be reproduced using a biological replicate and ideally with data of higher quality (e.g. Whole metagenome sequencing paired with large-scale metabolomics).

Since relatively rare interactions between the community members can have strong influences on successive biochemical events, it is believed that also the not so prevalent patterns are of great interest and thus must also be made detectable. To make such achievements feasible either a greater dataset should be obtained and/or better performing techniques be used for making the measurements. Whole metagenome sequencing would yield better data, yielding insights into the precise genetic content of the different community members.

In addition to the metagenome and metabolome, we believe involving and integrating also the transcriptome is of much potential in this context. The sequences transcripts could be coupled with specific species, whereas the metabolome gives a community-wide overview. The latter does complicate the pinpointing of species specific biochemical activity, but does give good insights into community-wide biochemical activity and is very close to the phenotypic traits. The combination of all three mentioned data would therefore complement each other well. A public dataset pairing these datasets would allow the research community to investigate this topic thoroughly.

Future studies could also aim at elucidating the distinct difference in performance of the PBMM algorithm, when applied on samples of the control and diseased group. Even though, this odd finding is likely related to the community's complexity, it might also relate to some extent to the degree of balance and efficiency of the community, which is disrupted by disorders. Future research must shed light on this matter.

At the stage that the algorithm is thoroughly evaluated and the predictions are of even better quality, then the research should focus at extraction of interaction patterns between the detected organisms, that occur via the chemical compounds. Metabolic fluxes could be determined by measuring the throughput of each edge in the reaction network. Finally, random adjustments (i.e. removal of an edge/reaction) could be made to the network to evaluate the consequences of this.

Last but not least, we see much virtue in applying flux balance analysis methods in the context of metabolic modeling. While there are many hurdles to be overcome we deem this a very viable direction as various factors are accounted for and the method is capable of yielding accurate results. On the other hand, this approach does require accurate and manually curated models or errors will be propagated through the model during computation and lead to invalid results. To this end, we suggest to approach the problem using an algorithm more similar to the PBMM algorithm, which appears more robust to low-quality data.

References

[1] Lora V Hooper, Dan R Littman, and Andrew J Macpherson. "Interactions between the microbiota and the immune system". In: *Science* 336.6086 (2012), pp. 1268–1273.

[2] Fernando Baquero and Cesar Nombela. "The microbiome as a human organ". In: *Clinical Microbiology and Infection* 18.s4 (2012), pp. 2–4.

[3] Junjie Qin et al. "A metagenome-wide association study of gut microbiota in type 2 diabetes". In: *Nature* 490.7418 (2012), pp. 55–60.

[4] Xochitl C Morgan et al. "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment". In: *Genome biology* 13.9 (2012), R79.

[5] Yan Wang and Lloyd H Kasper. "The role of microbiome in central nervous system disorders". In: *Brain, behavior, and immunity* 38 (2014), pp. 1–12.

[6] Petia Kovatcheva-Datchary and Tulika Arora. "Nutrition, the gut microbiome and the metabolic syndrome". In: *Best Practice & Research Clinical Gastroenterology* 27.1 (2013), pp. 59–72.

[7] Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. "An obesity-associated gut microbiome with increased capacity for energy harvest". In: *nature* 444.7122 (2006), pp. 1027–131.

[8] Peter J Turnbaugh et al. "A core gut microbiome in obese and lean twins". In: *nature* 457.7228 (2009), pp. 480–484.

[9] Jose M Ordovas and Vincent Mooser. "Metagenomics: the role of the microbiome in cardiovascular diseases". In: *Current opinion in lipidology* 17.2 (2006), pp. 157–161.

[10] Tiffany L Weir, Daniel K Manter, Amy M Sheflin, Brittany A Barnett, Adam L Heuberger, and Elizabeth P Ryan. "Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults". In: *PloS one* 8.8 (2013), e70803.

[11] Jiyoung Ahn, Rashmi Sinha, Zhiheng Pei, Christine Dominianni, Jing Wu, Jianxin Shi, James J Goedert, Richard B Hayes, and Liying Yang. "Human gut microbiome and risk of colorectal cancer". In: *Journal of the National Cancer Institute* (2013), djt300.

[12] Jeroen Raes and Peer Bork. "Molecular eco-systems biology: towards an understanding of community function". In: *Nature Reviews Microbiology* 6.9 (2008), pp. 693–699.

[13] Erica C Seth and Michiko E Taga. "Nutrient cross-feeding in the microbial world". In: (2014).

[14] Patricia M Griffin and Robert V Tauxe. "The epidemiology of infections caused by Escherichia coli O157: H7, other enterohemorrhagic E. coli, and the associated hemolytic uremic syndrome". In: *Epidemiologic reviews* 13.1 (1991), pp. 60–98.

[15] Eric A Franzosa et al. "Relating the metatranscriptome and metagenome of the human gut". In: *Proceedings of the National Academy of Sciences* 111.22 (2014), E2329–E2338.

[16] Alison R Erickson et al. “Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn’s disease”. In: *PloS one* 7.11 (2012), e49138.

[17] Ian H McHardy et al. “Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships”. In: *Microbiome* 1.1 (2013), p. 1.

[18] Cecilia Noecker, Alexander Eng, Sujatha Srinivasan, Casey M Theriot, Vincent B Young, Janet K Jansson, David N Fredricks, and Elhanan Borenstein. “Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation”. In: *mSystems* 1.1 (2016), e00013–15.

[19] Arne Buchholz, Ralf Takors, and Christian Wandrey. “Quantification of intracellular metabolites in Escherichia coli K12 using liquid chromatographic-electrospray ionization tandem mass spectrometric techniques”. In: *Analytical biochemistry* 295.2 (2001), pp. 129–137.

[20] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. “Advances in flux balance analysis”. In: *Current opinion in biotechnology* 14.5 (2003), pp. 491–496.

[21] Jong Min Lee, Erwin P Gianchandani, and Jason A Papin. “Flux balance analysis in the era of metabolomics”. In: *Briefings in bioinformatics* 7.2 (2006), pp. 140–150.

[22] Karthik Raman and Nagasuma Chandra. “Flux balance analysis of biological systems: applications and challenges”. In: *Briefings in bioinformatics* 10.4 (2009), pp. 435–449.

[23] Stefanie Widder et al. “Challenges in microbial ecology: building predictive understanding of community function and dynamics”. In: *The ISME journal* (2016).

[24] Finja Buchel et al. “Path2Models: large-scale generation of computational models from biochemical pathway maps”. In: *BMC systems biology* 7.1 (2013), p. 116.

[25] Sujatha Srinivasan, Martin T Morgan, Tina L Fiedler, Danijel Djukovic, Noah G Hoffman, Daniel Raftery, Jeanne M Marrazzo, and David N Fredricks. “Metabolic signatures of bacterial vaginosis”. In: *MBio* 6.2 (2015), e00204–15.

[26] Richard Amsel, Patricia A Totten, Carol A Spiegel, Kirk CS Chen, David Eschenbach, and King K Holmes. “Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations”. In: *The American journal of medicine* 74.1 (1983), pp. 14–22.

[27] Robert P Nugent, Marijane A Krohn, and Sharon L Hillier. “Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation.” In: *Journal of clinical microbiology* 29.2 (1991), pp. 297–301.

[28] J Gregory Caporaso et al. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nature methods* 7.5 (2010), pp. 335–336.

[29] David A Eschenbach, Pamela R Davick, Betsy L Williams, Seymour J Klebanoff, Karen Young-Smith, CM Critchlow, and King K Holmes. "Prevalence of hydrogen peroxide-producing Lactobacillus species in normal women and women with bacterial vaginosis." In: *Journal of Clinical Microbiology* 27.2 (1989), pp. 251–256.

[30] Sharon L Hillier, Marijane A Krohn, Lorna K Rabe, Seymour J Klebanoff, and David A Eschenbach. "The normal vaginal flora, H2O2-producing lactobacilli, and bacterial vaginosis in pregnant women". In: *Clinical Infectious Diseases* 16.Supplement 4 (1993), S273–S281.

[31] Carol A Spiegel, Richard Amsel, David Eschenbach, Fritz Schoenknecht, and King K Holmes. "Anaerobic bacteria in nonspecific vaginitis". In: *New England Journal of Medicine* 303.11 (1980), pp. 601–607.

[32] CA Spiegel. "Bacterial vaginosis." In: *Clinical Microbiology Reviews* 4.4 (1991), pp. 485–502.

[33] Jack D Sobel. "Bacterial vaginosis". In: *Annual review of medicine* 51.1 (2000), pp. 349–356.

[34] Steven S Witkin and William J Ledger. "Complexities of the uniquely human vagina". In: *Science Translational Medicine* 4.132 (2012), 132fs11–132fs11.

[35] Morgan GI Langille et al. "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences". In: *Nature biotechnology* 31.9 (2013), pp. 814–821.

[36] Ohad Manor and Elhanan Borenstein. "MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome". In: *Genome biology* 16.1 (2015), p. 1.

[37] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. "Sequencing depth and coverage: key considerations in genomic analyses". In: *Nature Reviews Genetics* 15.2 (2014), pp. 121–132.

[38] Peter E Larsen, Frank R Collart, Dawn Field, Folker Meyer, Kevin P Keegan, Christopher S Henry, John McGrath, John Quinn, and Jack A Gilbert. "Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset". In: *Microbial informatics and experimentation* 1.1 (2011), p. 1.

[39] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[40] Sharon Greenblum, Peter J Turnbaugh, and Elhanan Borenstein. "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease". In: *Proceedings of the National Academy of Sciences* 109.2 (2012), pp. 594–599.

[41] Tasia M Taxis et al. "The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity". In: *Nucleic acids research* (2015), gkv973.

[42] GR Glenn and MB Roberfroid. “Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics”. In: *J. nutr* 125 (1995), pp. 1401–1412.

[43] M Ragan-Kelley, F Perez, B Granger, T Kluyver, P Ivanov, J Frederic, and M Bussonier. “The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication.” In: *AGU Fall Meeting Abstracts*. Vol. 1. 2014, p. 07.

[44] Dirk Merkel. “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux Journal* 2014.239 (2014), p. 2.

[45] Joseph N Paulson, O Colin Stine, Hector Corrada Bravo, and Mihai Pop. “Differential abundance analysis for microbial marker-gene surveys”. In: *Nature methods* 10.12 (2013), pp. 1200–1202.

[46] Robert K Colwell, Anne Chao, Nicholas J Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L Chazdon, and John T Longino. “Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages”. In: *Journal of plant ecology* 5.1 (2012), pp. 3–21.

[47] George EP Box and David R Cox. “An analysis of transformations”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1964), pp. 211–252.

[48] Morgan GI Langille et al. “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences”. In: *Nature biotechnology* 31.9 (2013), pp. 814–821.

[49] Nathan Mantel. “The detection of disease clustering and a generalized regression approach”. In: *Cancer research* 27.2 Part 1 (1967), pp. 209–220.

[50] Jon Carr. *MantelTest*. 2017. URL: http://jwcarr.github.io/MantelTest/.

[51] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. [Online; accessed 2017-03-20]. 2001. URL: http://www.scipy.org/.

[52] Ralph B d’Agostino. “An omnibus test of normality for moderate and large size samples”. In: *Biometrika* (1971), pp. 341–348.

[53] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[54] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, Nov. 1999. URL: http://ilpubs.stanford.edu:8090/422/.

[55] Noah Fierer and Robert B Jackson. “The diversity and biogeography of soil bacterial communities”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.3 (2006), pp. 626–631.

[56] Ludmila I Kuncheva and Christopher J Whitaker. “Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy”. In: *Machine learning* 51.2 (2003), pp. 181–207.

[57] Elizabeth A Grice, Heidi H Kong, Gabriel Renaud, Alice C Young, Gerard G Bouffard, Robert W Blakesley, Tyra G Wolfsberg, Maria L Turner, and Julia A Segre. “A diversity profile of the human skin microbiota”. In: *Genome research* 18.7 (2008), pp. 1043–1050.

[58] Daniel H Huson, Daniel C Richter, Christian Rausch, Tobias Dezulian, Markus Franz, and Regula Rupp. “Dendroscope: An interactive viewer for large phylogenetic trees”. In: *BMC bioinformatics* 8.1 (2007), p. 460.

[59] Sujatha Srinivasan, Martin T Morgan, Tina L Fiedler, Danijel Djukovic, Noah G Hoffman, Daniel Raftery, Jeanne M Marrazzo, and David N Fredricks. “Metabolic signatures of bacterial vaginosis”. In: *MBio* 6.2 (2015), e00204–15.

Appendices

Univariate tests – BV signatures

KEGG ID	Compound name	Mann–Whitney		Welch's t-test		Pearson		Spearman		Normality p-value
		U	p-val	t-val	p-val	r	p-val	r	p-val	
C01672	Cadaverine	6.0	0.00	17.45	0.00	0.85	0.00	0.78	0.00	0.00 / 0.00 / 0.03
C00134	Putrescine	6.0	0.00	17.18	0.00	0.81	0.00	0.72	0.00	0.00 / 0.02 / 0.05
C00483	Tyramine	8.0	0.00	11.99	0.00	0.76	0.00	0.71	0.00	0.00 / 0.20 / 0.90
C00082	L-Tyrosine	15.0	0.00	-11.94	0.00	-0.82	0.00	-0.79	0.00	0.01 / 0.22 / 0.57
C00794	D-Sorbitol	12.0	0.00	-11.09	0.00	-0.80	0.00	-0.79	0.00	1.00 / 0.60 / 0.01
C00489	Glutarate	17.0	0.00	10.79	0.00	0.79	0.00	0.78	0.00	0.05 / 0.51 / 0.36
C03189	DL-Glycerol 1-phosphate	39.0	0.00	10.45	0.00	0.72	0.00	0.69	0.00	0.00 / 0.11 / 0.01
C00209	Oxalate	41.0	0.00	-10.10	0.00	-0.73	0.00	-0.70	0.00	0.01 / 0.47 / 0.33
C00388	Histamine	43.0	0.00	9.46	0.00	0.71	0.00	0.69	0.00	0.00 / 0.44 / 0.03
C00079	L-Phenylalanine	48.0	0.00	-9.11	0.00	-0.78	0.00	-0.77	0.00	0.00 / 0.00 / 0.08
C00078	L-Tryptophan	53.0	0.00	-8.98	0.00	-0.79	0.00	-0.80	0.00	0.02 / 0.14 / 0.04
C00031	D-Glucose	55.0	0.00	-8.92	0.00	-0.71	0.00	-0.70	0.00	0.00 / 0.36 / 0.26
C00577	D-Glyceraldehyde	33.0	0.00	8.87	0.00	0.73	0.00	0.74	0.00	0.66 / 0.85 / 0.63
C00270	N-Acetylneuraminate	49.0	0.00	8.68	0.00	0.67	0.00	0.65	0.00	0.00 / 0.21 / 0.57
C00042	Succinate	55.0	0.00	8.38	0.00	0.72	0.00	0.74	0.00	0.00 / 0.39 / 0.01
C00315	Spermidine	54.0	0.00	8.24	0.00	0.67	0.00	0.67	0.00	0.24 / 0.60 / 0.44
C01026	N,N-Dimethylglycine	59.0	0.00	7.97	0.00	0.66	0.00	0.66	0.00	0.01 / 0.04 / 0.01
C07599	Alloxanthine	75.0	0.00	7.84	0.00	0.69	0.00	0.67	0.00	0.07 / 0.25 / 0.19
C00065	L-Serine	63.0	0.00	-7.73	0.00	-0.70	0.00	-0.74	0.00	0.13 / 0.92 / 0.09
C03264	D-2-Hydroxyisocaproate	80.0	0.00	6.82	0.00	0.65	0.00	0.62	0.00	0.05 / 0.01 / 0.90
C00047	L-Lysine	76.0	0.00	-6.25	0.00	-0.65	0.00	-0.63	0.00	0.58 / 0.00 / 0.55
C00073	L-Methionine	100.0	0.00	-5.74	0.00	-0.63	0.00	-0.64	0.00	0.04 / 0.00 / 0.01
C00408	L-Pipecolate	88.0	0.00	-5.71	0.00	-0.62	0.00	-0.61	0.00	0.83 / 0.01 / 0.49
C00123	L-Leucine	111.0	0.00	-5.67	0.00	-0.57	0.00	-0.56	0.00	0.07 / 0.02 / 0.52
C00188	L-Threonine	116.0	0.00	-5.54	0.00	-0.58	0.00	-0.58	0.00	0.29 / 0.64 / 0.62
C01717	4-Hydroxy-2-quinolinecarboxylic acid	115.0	0.00	5.31	0.00	0.60	0.00	0.58	0.00	0.01 / 0.05 / 0.39
C00318	L-Carnitine	132.0	0.00	-5.05	0.00	-0.52	0.00	-0.55	0.00	0.71 / 0.54 / 0.27
C00380	Cytosine	128.0	0.00	5.04	0.00	0.56	0.00	0.50	0.00	0.02 / 0.27 / 0.67
C00186	(S)-Lactate	120.0	0.00	-5.03	0.00	-0.54	0.00	-0.52	0.00	0.01 / 0.00 / 0.81
C02470	Xanthurenic acid	123.0	0.00	-4.90	0.00	-0.54	0.00	-0.57	0.00	0.05 / 0.08 / 0.02
C00036	Oxaloacetate	119.0	0.00	4.88	0.00	0.56	0.00	0.54	0.00	0.08 / 0.02 / 0.94
C00077	L-Ornithine	145.0	0.00	-4.64	0.00	-0.50	0.00	-0.41	0.00	0.07 / 0.07 / 0.71
C00135	L-Histidine	155.0	0.00	-4.50	0.00	-0.56	0.00	-0.57	0.00	0.05 / 0.19 / 0.03
C00062	L-Arginine	158.0	0.00	-4.44	0.00	-0.56	0.00	-0.57	0.00	0.92 / 0.55 / 0.87
C00049	L-Aspartate	129.0	0.00	-4.33	0.00	-0.55	0.00	-0.59	0.00	0.59 / 0.00 / 0.71
C00383	Malonate	154.0	0.00	4.23	0.00	0.55	0.00	0.54	0.00	0.19 / 0.84 / 0.00
C00037	Glycine	176.0	0.00	-4.18	0.00	-0.50	0.00	-0.59	0.00	0.35 / 0.70 / 0.42
C00191	D-Glucuronate	177.0	0.00	-4.11	0.00	-0.47	0.00	-0.50	0.00	0.04 / 0.56 / 0.40
C00120	Biotin	186.0	0.00	-4.04	0.00	-0.33	0.00	-0.28	0.02	0.30 / 0.72 / 0.51
C00025	L-Glutamate	182.0	0.00	-3.98	0.00	-0.50	0.00	-0.53	0.00	0.01 / 0.31 / 0.42
C00300	Creatine	200.0	0.00	-3.65	0.00	-0.41	0.00	-0.50	0.00	0.71 / 0.11 / 0.34
C00956	L-2-Aminoadipate	195.0	0.00	3.64	0.00	0.39	0.00	0.40	0.00	0.58 / 1.00 / 0.77
C00022	Pyruvate	209.0	0.00	3.33	0.00	0.37	0.00	0.43	0.00	0.43 / 0.29 / 0.29
C00407	L-Isoleucine	214.0	0.00	-3.28	0.00	-0.34	0.00	-0.32	0.01	0.01 / 0.08 / 0.14
C00366	Urate	216.0	0.00	-3.28	0.00	-0.35	0.00	-0.37	0.00	0.80 / 0.18 / 0.96
C06104	Adipate	205.0	0.00	3.19	0.00	0.45	0.00	0.44	0.00	0.03 / 0.29 / 0.00
C00989	4-Hydroxybutanoic acid	214.0	0.00	3.17	0.00	0.44	0.00	0.45	0.00	0.16 / 0.79 / 0.00
C00245	Taurine	216.0	0.00	3.09	0.00	0.31	0.01	0.34	0.00	0.74 / 0.60 / 0.25
C00253	Nicotinate	196.0	0.00	3.04	0.00	0.43	0.00	0.46	0.00	0.20 / 0.65 / 0.01
C00117	D-Ribose 5-phosphate	226.0	0.00	2.97	0.01	0.42	0.00	0.43	0.00	0.22 / 0.55 / 0.00
C00122	Fumarate	236.0	0.01	2.96	0.01	0.41	0.00	0.37	0.00	0.24 / 0.51 / 0.01
C02226	2-Methylmaleate	228.0	0.00	2.90	0.01	0.39	0.00	0.37	0.00	0.23 / 0.60 / 0.01
C00262	Hypoxanthine	237.0	0.01	-2.85	0.01	-0.36	0.00	-0.32	0.01	0.20 / 0.19 / 0.70
C00026	2-Oxoglutarate	235.0	0.00	2.81	0.01	0.41	0.00	0.40	0.00	0.34 / 0.91 / 0.01
C00219	Arachidonate	242.0	0.01	2.77	0.01	0.39	0.00	0.41	0.00	0.12 / 0.45 / 0.00
C05582	Homovanillate	242.0	0.01	2.76	0.01	0.39	0.00	0.38	0.00	0.23 / 0.42 / 0.01
C00114	Choline	241.0	0.01	2.51	0.02	0.27	0.02	0.23	0.05	0.95 / 0.40 / 0.32
C01796	D-Erythrose	248.0	0.01	2.49	0.02	0.32	0.01	0.32	0.01	0.98 / 0.76 / 0.65
C00385	Xanthine	279.0	0.03	2.26	0.03	0.34	0.00	0.32	0.01	0.09 / 0.31 / 0.00
C00791	Creatinine	270.0	0.02	-2.22	0.03	-0.23	0.05	-0.21	0.08	0.62 / 0.77 / 0.72
C02630	2-Hydroxyglutarate	261.0	0.01	-2.04	0.05	-0.31	0.01	-0.38	0.00	0.00 / 0.10 / 0.40
C00041	L-Alanine	295.0	0.05	1.90	0.07	0.24	0.04	0.22	0.07	0.99 / 0.73 / 0.64
C00788	L-Adrenaline	300.0	0.06	1.79	0.08	0.24	0.05	0.24	0.05	0.22 / 0.67 / 0.05
C00327	L-Citrulline	299.0	0.06	1.72	0.09	0.11	0.37	0.07	0.57	1.00 / 0.81 / 0.88
C00147	Adenine	314.0	0.09	1.62	0.11	0.16	0.19	0.15	0.22	0.71 / 0.31 / 0.86
C00212	Adenosine	311.0	0.08	-1.30	0.20	-0.09	0.46	-0.06	0.61	0.54 / 0.67 / 0.75
C00493	Shikimate	325.0	0.12	-1.25	0.22	-0.18	0.15	-0.18	0.15	1.00 / 0.81 / 0.89
C00020	AMP	305.0	0.07	-1.07	0.29	-0.12	0.32	-0.12	0.31	0.79 / 0.41 / 0.26
C00148	L-Proline	330.0	0.14	-1.00	0.32	-0.11	0.35	-0.17	0.15	0.83 / 0.83 / 0.12
C01104	Trimethylamine N-oxide	344.0	0.19	-0.89	0.38	-0.07	0.54	-0.05	0.66	0.60 / 0.92 / 0.99
C00153	Nicotinamide	352.0	0.23	0.81	0.42	0.13	0.27	0.17	0.15	0.27 / 0.39 / 0.10
C00864	Pantothenate	377.0	0.36	0.60	0.55	-0.00	0.98	-0.06	0.62	0.10 / 0.61 / 0.06
C00183	L-Valine	371.0	0.33	-0.55	0.58	-0.03	0.82	0.01	0.91	0.31 / 0.54 / 0.07
C03794	N6-(1,2-Dicarboxyethyl)-AMP	374.0	0.34	-0.37	0.71	-0.08	0.49	-0.09	0.47	0.91 / 0.44 / 0.88
C01678	Cysteamine	378.0	0.37	0.36	0.72	0.03	0.78	0.02	0.90	0.93 / 1.00 / 0.96
C00258	D-Glycerate	385.0	0.41	0.19	0.85	0.01	0.97	0.02	0.89	0.36 / 0.62 / 0.57
C00064	L-Glutamine	390.0	0.44	-0.11	0.91	-0.05	0.71	-0.06	0.63	0.82 / 0.57 / 0.88
C00127	Glutathione disulfide	395.0	0.47	0.06	0.95	0.01	0.95	0.01	0.93	1.00 / 0.68 / 0.68

Figure 21: Table showing several metrics of association between each metabolite and bacterial vaginosis. See section 5.3.2 for a more complete explanation.

Taxonomy	Mann–Whitney		Welch's t-test		Pearson		Spearman		Normality p-value
	U	p-val	t-val	p-val	r	p-val	r	p-val	
f:Bifidobacteriaceae g:Gardnerella	16.0	0.00	7.26	0.00	0.73	0.00	0.49	0.00	0.00 / 0.00 / 0.00
f:Veillonellaceae g:Dialister	18.5	0.00	12.71	0.00	0.84	0.00	0.82	0.00	0.00 / 0.00 / 0.00
f:Lactobacillaceae g:Lactobacillus	24.0	0.00	-9.67	0.00	-0.77	0.00	-0.79	0.00	0.00 / 0.16 / 0.02
f:Prevotellaceae g:Prevotella	31.0	0.00	11.74	0.00	0.80	0.00	0.73	0.00	0.00 / 0.01 / 0.00
o:Coriobacteriales f:Coriobacteriaceae	33.5	0.00	13.99	0.00	0.81	0.00	0.69	0.00	0.00 / 0.00 / 0.00
f:Veillonellaceae g:Megasphaera	47.0	0.00	11.09	0.00	0.77	0.00	0.77	0.00	0.63 / 0.00 / 0.00
f:Leptotrichiaceae g:Sneathia	56.0	0.00	11.70	0.00	0.76	0.00	0.73	0.00	0.76 / 0.00 / 0.00
f:Aerococcaceae g:Aerococcus	71.0	0.00	9.60	0.00	0.61	0.00	0.40	0.00	0.00 / 0.00 / 0.00
f:[Tissierellaceae] g:Peptoniphilus	93.0	0.00	7.99	0.00	0.68	0.00	0.67	0.00	0.99 / 0.00 / 0.01
f:Gemellaceae g:Gemella	100.0	0.00	10.02	0.00	0.65	0.00	0.61	0.00	0.00 / 0.58 / 0.05
f:[Tissierellaceae] g:Parvimonas	105.5	0.00	7.08	0.00	0.64	0.00	0.69	0.00	0.00 / 0.00 / 0.09
f:Clostridiaceae g:Clostridium	110.0	0.00	9.88	0.00	0.73	0.00	0.79	0.00	0.52 / 0.58 / 0.01
f:Actinomycetaceae g:Mobiluncus	180.0	0.00	6.54	0.00	0.58	0.00	0.64	0.00	0.00 / 0.58 / 0.00
f:Porphyromonadaceae g:Porphyromonas	200.0	0.00	6.01	0.00	0.53	0.00	0.57	0.00	0.00 / 0.58 / 0.00
f:Lachnospiraceae g:Shuttleworthia	210.0	0.00	5.77	0.00	0.57	0.00	0.64	0.00	0.00 / 0.58 / 0.00
g:Lactobacillus s:iners	220.0	0.00	-1.16	18.00	-0.22	5.04	-0.49	0.00	0.00 / 0.00 / 0.00
f:Actinomycetaceae g:Actinomyces	240.0	0.00	4.99	0.00	0.39	0.00	0.35	0.00	0.00 / 0.58 / 0.00
g:Lactobacillus s:reuteri	240.0	0.00	-3.50	0.00	-0.52	0.00	-0.48	0.00	0.00 / 0.01 / 0.58
f:[Tissierellaceae] g:Anaerococcus	253.5	0.72	2.70	0.72	0.33	0.72	0.33	0.00	0.00 / 0.02 / 0.00
f:Actinomycetaceae g:Arcanobacterium	260.0	0.00	4.31	0.00	0.35	0.00	0.37	0.00	0.00 / 0.58 / 0.02
c:Bacteroidia o:Bacteroidales	260.0	0.00	4.51	0.00	0.49	0.00	0.58	0.00	0.00 / 0.58 / 0.00
c:Clostridia o:Clostridiales	260.0	0.00	4.46	0.00	0.33	0.72	0.31	0.72	0.00 / 0.58 / 0.01
o:I025 f:Rs-045	290.0	0.72	3.78	0.00	0.35	0.00	0.39	0.00	0.00 / 0.58 / 0.01
g:Lactobacillus s:coleohominis	292.0	0.00	-2.51	1.44	-0.38	0.00	-0.39	0.00	0.00 / 0.05 / 0.00
g:Peptostreptococcus s:anaerobius	294.0	1.44	2.53	0.72	0.25	2.16	0.22	5.04	0.00 / 0.00 / 0.00
o:Lactobacillales f:Streptococcaceae	300.0	0.00	-2.46	1.44	-0.36	0.00	-0.32	0.72	0.00 / 0.03 / 0.58
f:Fusobacteriaceae g:Fusobacterium	300.0	0.72	3.36	0.00	0.32	0.72	0.35	0.00	0.00 / 0.58 / 0.00
f:Mycoplasmataceae g:Mycoplasma	310.0	0.72	3.17	0.00	0.30	0.72	0.31	0.72	0.00 / 0.58 / 0.00
o:Bacteroidales f:S24-7	310.0	0.72	3.30	0.00	0.32	0.72	0.35	0.00	0.00 / 0.58 / 0.00
f:Lachnospiraceae g:Moryella	310.0	0.72	3.31	0.00	0.35	0.00	0.41	0.00	0.00 / 0.58 / 0.00
f:Alcaligenaceae g:Sutterella	330.0	2.16	2.86	0.72	0.27	2.16	0.26	2.16	0.00 / 0.58 / 0.00
f:Veillonellaceae g:Veillonella	350.0	3.60	2.30	2.16	0.08	36.00	0.05	47.52	0.00 / 0.58 / 0.00
f:Streptococcaceae g:Streptococcus	353.0	7.92	-1.33	14.40	-0.19	7.92	-0.21	6.48	0.00 / 0.00 / 0.00
f:Staphylococcaceae g:Staphylococcus	359.0	6.48	-1.23	16.56	-0.15	15.84	-0.16	12.96	0.00 / 0.00 / 0.00
f:Peptococcaceae g:Peptococcus	360.0	5.76	2.06	3.60	0.17	10.80	0.14	17.28	0.00 / 0.58 / 0.00
g:Prevotella s:melaninogenica	360.0	5.76	1.98	4.32	0.21	5.76	0.24	3.60	0.00 / 0.58 / 0.00
f:Campylobacteraceae g:Campylobacter	360.0	5.76	2.07	3.60	0.25	2.16	0.31	0.72	0.00 / 0.58 / 0.00
f:Aerococcaceae g:Facklamia	360.0	1.44	-1.45	11.52	-0.21	5.76	-0.19	8.64	0.00 / 0.00 / 0.58
g:Clostridium s:perfringens	360.0	1.44	-1.45	11.52	-0.25	2.16	-0.23	4.32	0.00 / 0.00 / 0.58
f:[Tissierellaceae] g:WAL_1855D	362.0	10.80	0.89	27.36	0.14	18.00	0.23	4.32	0.00 / 0.00 / 0.00
f:Coriobacteriaceae g:Atopobium	370.0	7.92	1.70	7.20	0.14	18.00	0.14	18.72	0.00 / 0.58 / 0.00
g:Campylobacter s:ureolyticus	370.0	7.92	1.78	5.76	0.14	18.72	0.16	12.96	0.00 / 0.58 / 0.00
o:Clostridiales f:Clostridiaceae	370.0	7.92	1.78	5.76	0.16	14.40	0.14	18.72	0.00 / 0.58 / 0.00
f:Oxalobacteraceae g:Ralstonia	378.0	19.44	-0.58	41.04	-0.07	38.88	-0.06	44.64	0.00 / 0.00 / 0.00
o:Bifidobacteriales f:Bifidobacteriaceae	380.0	11.52	1.43	11.52	0.12	22.32	0.10	29.52	0.00 / 0.58 / 0.00
o:Clostridiales f:Lachnospiraceae	380.0	11.52	1.43	11.52	0.10	30.24	0.05	48.24	0.00 / 0.58 / 0.00
o:Clostridiales f:[Mogibacteriaceae]	380.0	11.52	1.42	11.52	0.10	30.24	0.05	48.24	0.00 / 0.58 / 0.00
f:Moraxellaceae g:Enhydrobacter	380.0	5.76	-1.00	23.76	-0.18	10.08	-0.16	12.96	0.00 / 0.00 / 0.58
g:Bifidobacterium s:bifidum	380.0	11.52	1.43	11.52	0.10	29.52	0.06	45.36	0.00 / 0.58 / 0.00
o:Bacillales f:Planococcaceae	380.0	5.76	-1.00	23.76	-0.12	23.76	-0.10	29.52	0.00 / 0.00 / 0.58
o:Gemellales f:Gemellaceae	380.0	5.76	-1.00	23.76	-0.12	23.76	-0.10	29.52	0.00 / 0.00 / 0.58
o:Burkholderiales f:Oxalobacteraceae	382.0	20.16	0.61	39.60	0.08	38.16	0.08	38.16	0.00 / 0.00 / 0.00
f:Alicyclobacillaceae g:Alicyclobacillus	389.5	22.32	-0.59	40.32	-0.02	64.80	0.04	51.84	0.00 / 0.00 / 0.00
g:Haemophilus s:parainfluenzae	389.5	22.32	-0.46	46.80	-0.10	28.80	-0.13	19.44	0.00 / 0.00 / 0.00
f:Pasteurellaceae g:Aggregatibacter	390.0	18.00	1.00	23.04	0.04	54.00	-0.02	60.48	0.00 / 0.58 / 0.00
c:Synergistia o:Synergistales	390.0	18.00	1.00	23.04	0.07	41.04	0.04	54.72	0.00 / 0.58 / 0.00
g:Veillonella s:dispar	390.0	18.00	1.00	23.04	0.03	58.32	0.02	61.92	0.00 / 0.58 / 0.00
f:Neisseriaceae g:Neisseria	390.0	18.00	1.00	23.04	0.10	29.52	0.10	28.08	0.00 / 0.58 / 0.00
f:Bifidobacteriaceae g:Bifidobacterium	390.0	18.00	1.00	23.04	-0.01	66.96	-0.08	38.16	0.00 / 0.58 / 0.00
g:Bacteroides s:fragilis	390.0	18.00	1.00	23.04	0.07	41.04	0.04	54.72	0.00 / 0.58 / 0.00
o:Enterobacteriales f:Enterobacteriaceae	390.0	18.00	1.00	23.04	0.04	54.00	-0.02	60.48	0.00 / 0.58 / 0.00
f:[Tissierellaceae] g:ph2	390.0	18.00	1.00	23.04	0.13	20.16	0.16	12.24	0.00 / 0.58 / 0.00
g:Bacteroides s:ovatus	390.0	18.00	1.00	23.04	0.07	41.04	0.04	54.72	0.00 / 0.58 / 0.00
f:[Paraprevotellaceae] g:[Prevotella]	390.0	18.00	1.00	23.04	0.07	41.04	0.04	54.72	0.00 / 0.58 / 0.00
f:Mycoplasmataceae g:Ureaplasma	390.0	18.00	1.00	23.04	0.10	29.52	0.10	28.08	0.00 / 0.58 / 0.00
f:[Tissierellaceae] g:1-68	390.5	26.64	0.57	41.04	0.09	32.40	0.10	29.52	0.00 / 0.00 / 0.00
g:Streptococcus s:infantis	390.5	23.04	-0.35	52.56	-0.02	61.20	-0.04	51.84	0.00 / 0.00 / 0.00
f:[Tissierellaceae] g:Finegoldia	396.0	33.84	0.16	62.64	0.04	55.44	0.02	64.08	0.00 / 0.01 / 0.00
g:Streptococcus s:anginosus	399.0	35.28	-0.17	62.64	-0.05	48.96	-0.08	38.16	0.00 / 0.00 / 0.00
f:Enterococcaceae g:Enterococcus	400.0	35.28	0.08	67.68	-0.00	71.28	-0.05	47.52	0.00 / 0.00 / 0.00
f:Corynebacteriaceae g:Corynebacterium	400.0	35.28	-0.02	70.56	-0.01	67.68	-0.01	65.52	0.00 / 0.00 / 0.00
g:Bacteroides s:uniformis	NaN	NaN	NaN	NaN	-0.05	46.80	-0.07	40.32	0.00 / 0.58 / 0.58

Figure 22: Table showing several metrics of association between each OTU and bacterial vaginosis. See section 5.3.2 for a more complete explanation.