Bacterial Communities in Women with Bacterial Vaginosis: High Resolution Phylogenetic Analyses Reveal Relationships of Microbiota to Clinical Criteria

Seminar Report

Pierre Barbera

pierre.barbera@student.kit.edu

Abstract. This work represents a report on the study *Bacterial Communities in Women with Bacterial Vaginosis: High Resolution Phylogenetic Analyses Reveal Relationships of Microbiota to Clinical Criteria* by Srinivasan et al.. In it we examine the wet lab and bioinformatics methods used to generate and analyze the metagenomic data. We give a special focus to taxonomic classification, a core concept in metagenomic studies as a whole. We also provide a short summary of the authors' results and observations. We conclude the report by pointing out similar studies and their core differences.

1 Introduction

Metagenomics, the study of the collective genetic material of entire bacterial communities, is an emergent field driven by recent innovations in DNA sequencing. One of its central goals is to allow scientific inquiry into the role of bacteria in their environmental context.

Using traditional laboratory methods this has been difficult, as for the majority of bacterial species, culturing is not yet possible. Modern metagenomic techniques circumvent the need to culture by utilizing next generation sequencing (NGS) to process the genetic information of an entire environmental sample. Usually, it is combined with techniques to filter for specific genes within a sample.

A common application of NGS-driven study is profiling the bacterial composition of an environmental sample. In clinical studies, the purpose of this is typically to identify correlations between bacterial composition of a human microbial subcommunity and a persons health. Ultimately, the hope is to find causal links between presence of specific species of bacteria and recorded diseases.

In this report, we will investigate the methods of one such clinical study in detail. In particular we will elaborate on the computational methods used. The study in question investigates the vaginal bacterial composition of women diagnosed with *bacterial vaginosis*, a common infection of the vagina [15]. Its methods highlight the basic framework of metagenomic studies.

2 Preparatory and Wet-lab work

The first step in the execution of any study is gathering the data in a controlled, well documented fashion. In metagenomics this primarily involves physically taking the sample, whether it is from rain forest top soil or the human vagina, and documenting any possibly valuable meta data.

When employing computational phylogenetic methods, bringing physical samples over into the digital realm is a necessity. This usually involves concentrating the sample to the DNA region of interest, sequencing that region and preparing the data for processing.

This chapter will highlight the just mentioned processes, as the authors applied them in the study.

2.1 Sample Collection and Diagnosis

The basis for the study were vaginal samples from 242 female patients from a clinic for sexually transmitted diseases (STD) in a major US city, collected over the course of about four years. The authors also recorded disease status of the patients, using established clinical tests for bacterial vaginosis (BV), primarily *Amsel's criteria* [1] with further confirmation using the *Nugent score* [11].

For a positive diagnosis using **Amsel's criteria**, at least three of the following have to hold true:

1. homogeneous, white to yellow vaginal discharge
2. the presence of *clue cells*, cells that are apparently multi-shaded, indicating an abundant presence of bacteria on their surface
3. vaginal fluid observing a pH-level greater than 4.5
4. positive *whiff test*, which is the presence of a fishy odor when potassium hydroxide solution is added to vaginal discharge

Diagnosis using the **Nugent score** is based on *Gram staining*, a laboratory method that uses color staining of bacterial samples to differentiate them into three groups: Gram positive, Gram negative and the intermediate of the two, Gram variable.

When building the Nugent score of a sample, three Gram stains are performed. Each determines a sub-score, and added together they form a total ranging from 0 to 10. A higher score corresponds to a higher indication of the presence of BV. The sub-scores increase with the absence of *Lactobacillus spp.* and presence of *Gardnerella* and *Bacteroides spp.*.

2.2 16S rRNA Gene Amplification

To obtain the bacterial composition using phylogenetic trees, there are two basic approaches. The first applies random *shotgun* sequencing to the environmental sample, resulting in reads from all over the bacterial genomes. A database is then assembled of a set of marker genes, from a number of reference species. Reads can then be compared to these marker sequences, allowing classification [10].

The method used in this study is more directed. First, a gene is identified that is present in all species that have to be classified. Ideally this gene should be highly conserved, allowing for very general targeting, while also containing regions with higher diversity to be able to differentiate between species.

The **16S rRNA**-gene is a popular choice that satisfies the criteria. The 16S rRNA forms part of the small subunit of the bacterial ribosome, and as such is an integral part of bacterial life. Not only do all bacteria possess this gene, it is also highly conserved, as a high rate of mutation would quickly lead to the death of its host. It also contains nine *hypervariable* regions with a higher relative chance of mutation. This makes it an ideal basis for phylogenetic analysis.

To be able to efficiently perform analysis, it is preferable to only sequence the 16S rRNA genes of a sample. To do this, the authors use the **polymerase chain reaction** (PCR). During PCR, the number of copies of specified DNA subregions are amplified by several orders of magnitude. It uses a pair of *primers*, short DNA sequences that delimit the target subregion. They serve as the starting point for DNA synthesis, similar to primers in natural DNA replication within cells.

PCR functions by creating successive steps of thermal conditions for the sample. First, the sample is heated to allow the DNA in to denature. This allows the primers to attach to their specific locations at the start and the end of the targeted region. Crucially, they are built such that they match to the 5' end of the strand of their location, as Figure 1 illustrates. Conditions are then altered to allow the polymerase enzyme to work. It attaches at the primers' 3' end and begins copying the target region of DNA, extending the primer in the process.

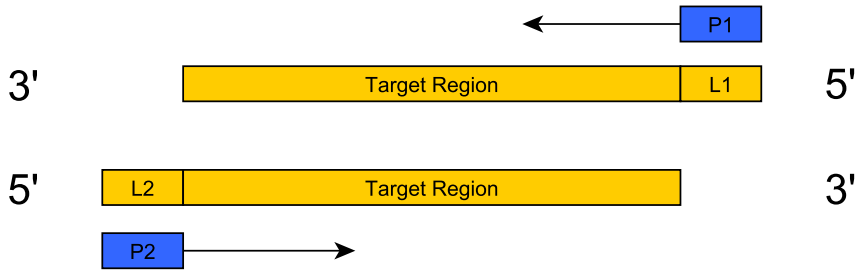


Fig. 1. Primer annealing phase of the polymerase chain reaction. Shown are the two, denatured, strands of the target DNA. The two primers, `P1` and `P2`, anneal to their counterpart nucleotide sequences at the 5' ends of the target region. Arrows show the direction of polymerization following immediately after this phase.

After this elongation phase, the strands are once again separated, and the process repeats. In each iteration, the number of template strands to which primers can anneal doubles, allowing the number of copies of the target region to grow exponentially. This means, that the genetic material in the target region

quickly outnumbers any other genetic material in the sample, improving the conditions for sequencing.

In the case this study, the target encompassed the V3 and V4 hypervariable regions of the 16S gene. The study's PCR step specifically targets them, as any differences between DNA sequences stemming from different species are much more likely to occur within them. Additionally, as the computational work of genomic studies usually scales with the length of the reads used as input, this subregion focus can lighten the load.

2.3 Sequencing and Preprocessing

Following the amplification of the samples' genomic content, the authors performed DNA sequencing using the 454 Life Sciences FLX process [17]. In preparation, the previous PCR step has added *adapter sequences* to either end of the 16S sequences (Figure 2). They are an integral part of 454 pyrosequencing, as they allow for *emulsion PCR*.

Fig. 2. Resulting PCR amplicon, as required by 454 Life Sciences pyrosequencing. The amplified sub-region of the 16S gene is extended by a DNA barcode for read identification, as well as adapter sequences on both ends. Pyrosequencing uses them to 1) allow the sequence to anneal to a bead and 2) enable primers to anneal, allowing copying by DNA polymerase.

During emulsion PCR, DNA fragments are suspended in a solution with microscopic plastic beads. The surface of these beads is covered with copies of a short polynucleotide, to which one end of the DNA fragments, called the adapter, corresponds to. This allows the fragments to attach to the beads. Moreover, it is assured that most beads will only capture one fragment. Synthetic oil is added, allowing for the formation of suspended oil droplets containing single beads. Standard PCR protocols are then followed to amplify the fragments, such that they cover the surface of their beads.

The resulting beads are then spread onto a special platter, made up of wells that fit single beads. Then, conditions are set for DNA replication. In successive iterations, the machine adds different *deoxynucleoside triphosphates* (dNTPs), which are modified NTPs, that start a light emitting chain reaction when added to the end of a DNA strand. A photo detector records this light for every bead, indicating whether or not the nucleotide was the correct next choice. By repeating this addition and detection, FLX sequencing forms a full DNA string readout. This full sequence, now digital data, is also called a **read**.

Additional to adding sequencing adapters, PCR has added a *DNA barcode*. This barcode allows the positive identification of correct reads obtained through sequencing, discarding any accidental byproducts. To ease further processing, the barcode as well as the two adapter and primer sequences are then removed from the reads.

3 Taxonomic Classification

Now that sequence data is available digitally, the bioinformatics portion of the study can begin. Primarily, this involves identifying what species are present in a sample. The authors build a tree from species, that they know to be important to the environmental niche under study. They then place the reads on the tree, allowing classification in relation to this set of known species.

3.1 Building the Reference Tree

In order to identify the obtained reads taxonomically, first their relation to each other, and previously known species, has to be determined. In the study, the authors do this by placing the reads on a phylogenetic tree. They obtain this *reference tree* by first assembling a list of bacteria that are known, through previous study, to inhabit the human vagina.

For each species on the list, they downloaded the associated 16S sequence from the website of the Ribosomal Database Project (RDP) [2]. The data obtained from the RDP also included NCBI taxonomic annotation, which they used to label the sequences. However, as the authors note, such publicly available sequences are often falsely labeled. As a consequence, they performed *mislabel detection* on them.

To achieve this, pairwise distances between all sequences labeled with the same taxon are computed. A *primary reference sequence* `S` is then selected, having the minimum median distance to all other sequences. Figure 3 illustrates this principle. Sequences that are further from `S` than some cutoff δ are discarded as mislabeled. According to the authors, their threshold was found empirically and can be regarded such that all remaining sequences are at least 98.5% identical.

To form the final set of reference sequences, the authors first selected all `S`-sequences identified during mislabel detection. Additionally they selected any sequence belonging to a reference strain. Reference strains are living cultures, that serve as a reference for a bacterial species, especially for the assignment of taxonomic labels [4].

To complete the collection, the authors selected a varying number of sequences per taxon. This was on average five sequences, with the overall goal of maximizing the sum of pairwise distances between them, resulting in maximal diversity within the sets. The number of sequences per taxon also varied, depending on whether a taxon was of lower relative biological importance (less sequences) or where higher taxonomic resolution was required (more sequences).

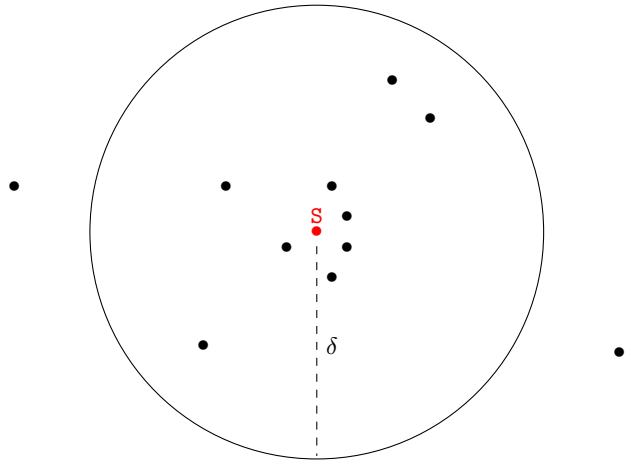


Fig. 3. Visualization of the principle of mislabel detection. First, pairwise distances are computed between all sequences labeled as belonging to the same taxon. A *primary reference sequence* `S`, having the smallest median distance to all other sequences, is then identified. Finally, sequences with a distance to `S` greater than some selected cutoff δ are discarded as mislabeled.

They then fed the resulting set of 16S sequences into a standard pipeline for generating phylogenetic trees: first they built a multiple sequence alignment (MSA) of all sequences. Using this, they built the tree using `RAxML` [16], a maximum likelihood phylogenetic tree inference suite.

3.2 Read Placement using `pplacer`

In the study, the authors used `pplacer` [9] to find optimal placements of all reads on the reference tree. The software is able to perform this operation using either of two different criteria: Bayesian posterior probability or maximum-likelihood.

When using the Bayesian method,it calculates the posterior probability of a placement dependent on the reference tree. This is the probability that the read placed at the given edge is the correct choice. In contrast, the maximum-likelihood method produces the set of confidence, or likelihood weight ratios, across all edges of the reference tree.

Placing a read does not change the reference tree. Doing so simply gives the read a placement location. This in turn means, that multiple reads can share the same location. It is noteworthy, that placing the reads on the reference tree *without extending* it is exactly what makes this technique computationally efficient. This is because the cost of finding a maximum likelihood placement on a phylogenetic tree grows exponentially with the number of taxa on the tree. As environmental samples usually contain thousands of unique sequences, this step would otherwise incur a higher computational workload. Making the placement operations independent of each other also makes it more easy to parallelize.

The result of taxonomic placement of all reads belonging to a metagenomic sample is their coordinate mapping on the reference tree. More specifically, as the paper uses the Bayesian mode, each sequence now has a set of possible locations on the tree, each with it's associated posterior probability. To shorthand, we will subsequently call one such tree with its associated placements of one sample a **sample-tree**.

3.3 Taxonomic Assignment

After obtaining the possible placement locations for each read in the sample, there still remains the task of assigning taxonomic labels to the sequences. As the paper focuses on analysis of community composition, correctness of this step is crucial. At its most basic, taxonomic assignment, in this case, is derived from the taxonomic assignment of the edges of the reference tree.

Because the authors built the reference tree from known, previously taxonomically classified sequences, its leafs already have assignments. As for the inner edges of the tree, the protocol is as follows: for each side of an edge, investigate what the most specific taxonomic rank is, that all leafs on that side share. The label assigned to the edge is then whichever is the most specific rank of either side. Figure 4 illustrates this.

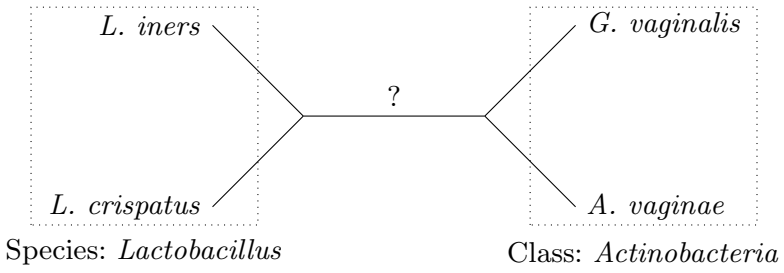


Fig. 4. Taxonomic assignment of inner branches of a reference tree of known species. Leaves are the known species and are labeled accordingly. Boxes are labeled as the most specific taxonomic rank that is shared between all leaves contained. The label of the central edge, indicated by a question mark, is assigned as the most specific rank shared by all leaves on either side of the edge. In this instance, this would be *Lactobacillus spp.*.

For the taxonomic assignment of the sample sequences, we need to recall that sequence placements are represented as posterior probabilities of all placement locations. Using this information, combined with the taxonomically labeled reference tree, `pplacer` calculates the posterior probability of each assignment of a label to a sequence. It does this by taking the sum of all placement probabilities of a sequence that share the same taxonomic assignment.

The authors enhance this method by assigning compound names for genus and species level labels of ambiguous sequences. On these ranks they select up to three labels, that have a posterior probability of greater than 0.05. The result of this are labels of the form *Streptococcus mitis/oralis*.

4 Correlation Analysis

With taxonomic classification concluded, data analysis can proceed. Methods highlighted in this section focus on visualization of bacterial composition of the samples, and more importantly finding correlations between bacterial prevalence and whether a patient is BV positive.

4.1 Phylogenetic Kantorovic-Rubinstein Distance

In order to cluster any defined entities, there has to be a distance metric between them. For the previously described sample-trees, Evans and Matsen describe the phylogenetic Kantorovic-Rubinstein (K-R) distance [3]. It builds on the commonly used *UniFrac* distance [6], a metric aiming to define a biologically meaningful distance between two metagenomic samples.

Conceptually, the K-R distance metric is similar to the *earth-mover distance* (EMD), in that it defines distances between probability distributions. It does this using a mass-transport analogy: differences of probability density at a specific interval, are seen as mass that has to be transported through one distribution, so as to match the second distribution. Doing so requires work: mass times distance. EMD defines the distance between two distributions as the minimum amount of work required to transform one into another.

To apply this distance to the aforementioned sample-trees, the authors define a probability density on them. As Figure 5 shows, it is almost trivial: they simply label every edge of the reference tree with the fraction of sequences that were mapped to that edge. They do this individually for every sample.

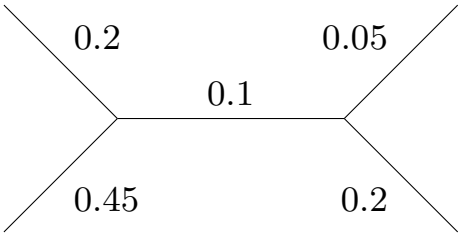


Fig. 5. Example of a probability distribution mapped to a trivial phylogenetic tree. When representing a metagenomic sample, the edge labels specify the fraction of sequences that have mapped to that edge in the reference tree.

When computing the K-R distance between two such trees, they can be thought of as two networks of roads, with their probability "mass" as sand on the roads. Further, as both trees are based on the reference tree, they are topologically identical. The mass is then moved along the roads in one tree, such that both trees are identical in distribution as well. Again, the distance is defined as the minimum amount of work required to do so.

As Evans and Matsen point out, this procedure can be computed efficiently via integral over the tree in linear time, making it suitable for the data requirements of metagenomic studies.

4.2 Squash Clustering

With the sample-trees represented by probability distributions on reference trees, and having defined a distance metric between them (Section 4.1), the authors perform *hierarchical clustering*. The result of this procedure is the organization of sample-trees into a hierarchy, drawn again as a tree, called a *cluster tree*.

Hierarchical clustering can be exemplified by the basic steps of the Neighbor Joining algorithm [14]: first, build a *Pairwise Distance Matrix* (PWD), a matrix filled with the distances between all possible pairs of entities. From this matrix select the smallest entry. The pair of entities that produced the entry is selected as the first pair to be joined, or clustered. A new entity is made representing the merging of the two entities, forming their parent node when drawing the result as a cluster tree. The branch lengths between the children and their parent are then assigned using a defined scheme. In Neighbor Joining, this is simply half the distance between the original two entities. Finally the PWD is remade, replacing the merged entities with their newly formed parent entity.

Squash Clustering [8], a clustering scheme for sample-trees, works similarly. Here too a PWD is constructed, however it is done using all sample-trees, labeled using the probability distribution of their reads (Figure 5). Again, the smallest entry is selected, this time according to the K-R distance. Their corresponding sample-trees are then merged by *squashing* the two trees together, as illustrated in the upper part of Figure 6: visually, the two trees are laid over each other, so that identical branches are next to each other. As a reminder, this is possible because the sample-trees have identical topology. They are the reference tree, enhanced with additional sample-specific probability distributions as edge labels. A weighted average is then applied to edge labels of overlaid branches. If the two trees have m and n number of discrete masses in their distributions, and their distributions are denoted as μ and τ respectively, then the weighted average is given as

$$\frac{m}{m+n}\mu + \frac{n}{m+n}\tau \tag{1}$$

The resulting averaged tree represents the merged entity, and forms the parent node of the two original trees in the clustering tree.

As in Neighbor Joining, branch lengths have to be assigned between parent and child. In Squash Clustering, this is simply the K-R distance (Section 4.1)

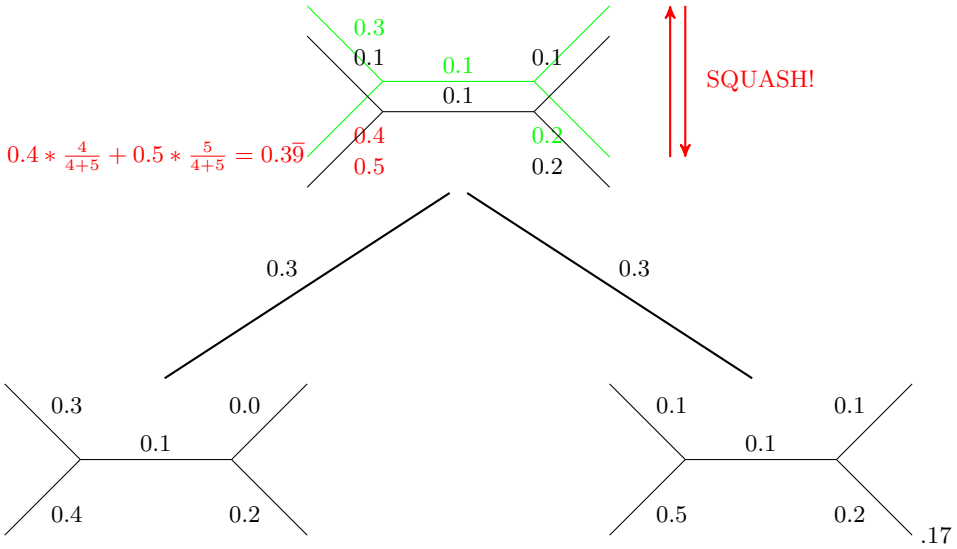


Fig. 6. Squash Clustering. The upper part shows the mechanism of squashing, wherein two topologically identical trees are combined by a weighted average of their branches. For this average, weight is the number of mass distributions of the current tree divided by the sum of mass distributions of both trees (see Equation 1). The full graphic shows the result of one full merging step, with the two trees that are merged connected to their parent node, the resulting of squashing. Branch lengths in this cluster tree are assigned as the Kantorovic-Rubinstein distance between the nodes.

between them. Figure 6 shows the result of one clustering iteration. Finally, the PWD is remade, just as mentioned for general hierarchical clustering, and the process starts over until the root of the cluster tree is formed.

4.3 Edge-PCA

In statistics, when faced with data of high dimensionality, a common way of reducing volume is Principal Component Analysis (PCA). In it, correlation between all pairs of variables is evaluated, to find the pair with the highest variance, or difference. If the data allows it, PCA can help to identify the variables primarily responsible for differentiating clusters.

In practice, for clinical metagenomics studies like the one covered here, the variables are the species present in a sample, and the clusters may correspond to disease status. Finding the variables, or species, responsible for a disease is usually the main subject of investigation, and as such the authors use PCA as a primary analytical technique.

First however, PCA has to be made applicable to the data at hand: phylogenetic trees. Matsen and Evans tackle this by introducing *Edge-PCA* [8]. In principle, it tries to find those edges in the tree that represent the best split, such that a maximum number of reads on either side of it belong to either of two groups.

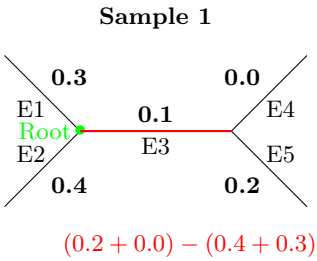
To understand this more easily we consider the following example, as it is also relevant to the study itself. The tree we operate on is the reference tree as described in Section 3.1. Through phylogenetic placement, outlined in Section 3.2, we have found a most likely location for every read on the tree.

Edge-PCA will begin by building a matrix, each row generated from a different sample. The columns represent the edges of the common reference tree. If the reference tree is unrooted, a root node can be chosen arbitrarily, however the choice has to be consistent throughout the process. Entries in the matrix are computed thusly: for the given edge, remove the edge from the tree. Then, subtract the total fractions of edge placements on the sub-tree containing the root from the total fractions of edge placements on the complementary sub-tree. This process is illustrated in Figure 7.

The resulting matrix is then fed into the standard PCA algorithm: a *covariance matrix* is derived and its eigenvectors extracted. These resulting eigenvectors represent the principal components. In their implementation, Matsen and Evans go one step further and map them back onto the tree for easier visualization, the general idea of which is visualized in Figure 8. This allows easy identification of possible differences in species prevalence between groups, such as healthy and diseased subjects.

5 Results and Discussion

Srinivasan et al. used Squash Clustering (Section 4.2) to organize their samples into a cluster tree. By visualizing this tree next to the patients disease status



Datamatrix

Sample	E1	E2	E3	E4	E5
1	-0.7	-0.6	-0.5	...	
2					
⋮					

Fig. 7. Building the data matrix from the sample-trees for Edge-PCA. Each row is built from a sample-tree and indexed accordingly. Each column represents an edge of the reference tree. Entries are computed by removing the current edge, then taking the sum of number of reads from all edges on the non-root side and subtracting the number of reads from the edges of the root side. The root can be chosen arbitrarily from the set of nodes, but has to be consistent throughout the process [8].

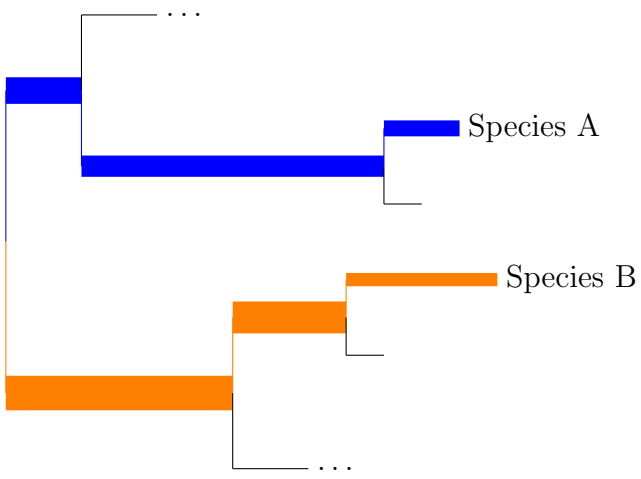


Fig. 8. Example of edge principal component, mapped to a phylogenetic tree. Color distinguishes sign, orange being positive and blue negative. Thickness represents how extreme the value is. Assume the graphic displays the first principle component, mapped onto the reference tree. A possible conclusion in this example could be, that prevalence of Species A vs. prevalence of Species B separates the data into two important clusters.

and bacterial composition of the sample (Appendix A), the authors were able to draw a first conclusion: the vaginal bacterial environment of women diagnosed with BV is highly diverse. In contrast, women diagnosed as healthy had a very homogeneous bacterial composition, dominated by *Lactobacillus spp.*, specifically either *L. iners* or *L. crispatus*.

Correlation analysis using Edge Principal Components Analysis (Section 4.3) supported this observation, showing a strong positive correlation between absence of BV and presence of *L. crispatus*.

The authors also investigated co-occurence of bacteria using Pearson correlation coefficients [12]. Again this showed that lactobacilli were strongly anti-correlated with bacteria that are believed to be associated with BV, splitting the bacteria into two distinct clusters. Between BV-associated bacteria, the authors found sub clusters, suggesting cooperation within, as well as indicating possible competition between these groups.

Finally, the authors examined correlations between bacterial taxa and the four Amsel criteria (Section 2.1), the mainstay diagnostic test for BV. They found, that only two bacteria were positively associated with all four criteria, despite the high bacterial diversity in BV positive samples. They also observed, that several other bacteria were highly associated with one to three criteria, as shown in the Venn diagram in Appendix B.

6 Related Studies and Recent Development

Since its publication, several comparable studies continued the investigation of the causes of BV. One study in particular, by Macklaim et al., took a different approach: instead of targeting the 16S gene for taxonomic profiling, they sequenced the *collective transcriptome* [7]. This allowed them to find functional differences between bacterial communities in women with and without BV. For example, they found evidence for a heightened anti-viral defense occurring in *L. iners* when residing in the BV microbiome. In addition, the transcriptomic approach allowed them to establish a bacterial taxonomic diversity essay, like the one produced in the study under focus in this work. They also found, that a very heterogeneous microbiome, not dominated by *L crispatus*, characterizes BV.

Another notable study analyzed daily change in vaginal bacterial composition of women with BV over a ten week period [13]. Methods were essentially the same as in this study, except the authors targeted the V1-V3 hypervariable regions of the 16S gene instead. They observed that, while current treatment does reduce the population of anaerobic bacteria and increase the population of *Lactobacillus spp.*, conditions quickly reverted back. Also they observed, that bacterial composition and propensity of change seems to be highly individualized.

A third study used *quantitative PCR* (qPCR) instead of sequencing to establish bacterial compositions [5]. This technique uses PCR primers targeting whole branches of a phylogeny. The process is monitored in real-time and bacte-

rial composition is inferred. Again, the authors took samples from patients over a multiple week long period. Their most notable observation was a effect they call *conversion*, a period of abnormality in the bacterial composition just prior to occurrence of BV symptoms.

Bibliography

[1] Amsel, R., Totten, P.A., Spiegel, C.A., Chen, K.C.S., Eschenbach, D., Holmes, K.K.: Nonspecific vaginitis. The American Journal of Medicine (1983), http://dx.doi.org/10.1016/0002-9343(83)91112-9

[2] Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A., McGarrell, D.M., Marsh, T., Garrity, G.M., et al.: The ribosomal database project: improved alignments and new tools for rrna analysis. Nucleic acids research 37(suppl 1), D141–D145 (2009)

[3] Evans, S.N., Matsen, F.A.: The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples (May 2010), http://arxiv.org/abs/1005.1699

[4] Kyrpides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Gker, M., Parker, C.T., Amann, R., Beck, B.J., Chain, P.S.G., Chun, J., Colwell, R.R., Danchin, A., Dawyndt, P., Dedeurwaerdere, T., DeLong, E.F., Detter, J.C., De Vos, P., Donohue, T.J., Dong, X.Z., Ehrlich, D.S., Fraser, C., Gibbs, R., Gilbert, J., Gilna, P., Glckner, F.O., Jansson, J.K., Keasling, J.D., Knight, R., Labeda, D., Lapidus, A., Lee, J.S., Li, W.J., MA, J., Markowitz, V., Moore, E.R.B., Morrison, M., Meyer, F., Nelson, K.E., Ohkuma, M., Ouzounis, C.A., Pace, N., Parkhill, J., Qin, N., Rossello-Mora, R., Sikorski, J., Smith, D., Sogin, M., Stevens, R., Stingl, U., Suzuki, K.i., Taylor, D., Tiedje, J.M., Tindall, B., Wagner, M., Weinstock, G., Weissenbach, J., White, O., Wang, J., Zhang, L., Zhou, Y.G., Field, D., Whitman, W.B., Garrity, G.M., Klenk, H.P.: Genomic encyclopedia of bacteria and archaea: Sequencing a myriad of type strains. PLoS Biol 12(8), e1001920 (08 2014), http://dx.doi.org/10.1371%2Fjournal.pbio.1001920

[5] Lambert, J.A., John, S., Sobel, J.D., Akins, R.A.: Longitudinal analysis of vaginal microbiome dynamics in women with recurrent bacterial vaginosis: Recognition of the conversion process. PLoS ONE 8(12), e82599 (12 2013), http://dx.doi.org/10.1371%2Fjournal.pone.0082599

[6] Lozupone, C., Knight, R.: Unifrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology 71(12), 8228–8235 (2005)

[7] Macklaim, J.M., Fernandes, A.D., Di Bella, J.M., Hammond, J.A., Reid, G., Gloor, G.B., et al.: Comparative meta-rna-seq of the vaginal microbiota and differential expression by lactobacillus iners in health and dysbiosis. Microbiome 1(1), 12 (2013)

[8] Matsen, F.A., Evans, S.N.: Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. PLoS ONE 8(3), e56859 (03 2013), http://dx.doi.org/10.1371%2Fjournal.pone.0056859

[9] Matsen, F.A., Kodner, R.B., Armbrust, E.V.: pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences

onto a fixed reference tree. BMC bioinformatics 11(1), 538 (Jan 2010), http://www.biomedcentral.com/1471-2105/11/538

[10] von Mering, C., Hugenholtz, P., Raes, J., Tringe, S., Doerks, T., Jensen, L., Ward, N., Bork, P.: Quantitative phylogenetic assessment of microbial communities in diverse environments. Science (New York, N.Y.) 315(5815), 1126–1130 (2007), http://dx.doi.org/10.1126/science.1133420

[11] Nugent, R.P., Krohn, M.A., Hillier, S.L.: Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. Journal of clinical microbiology 29(2), 297–301 (1991)

[12] Pearson, K.: Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London pp. 240–242 (1895)

[13] Ravel, J., Brotman, R.M., Gajer, P., Ma, B., Nandy, M., Fadrosh, D.W., Sakamoto, J., Koenig, S.S., Fu, L., Zhou, X., et al.: Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. studies 19, 20 (2013)

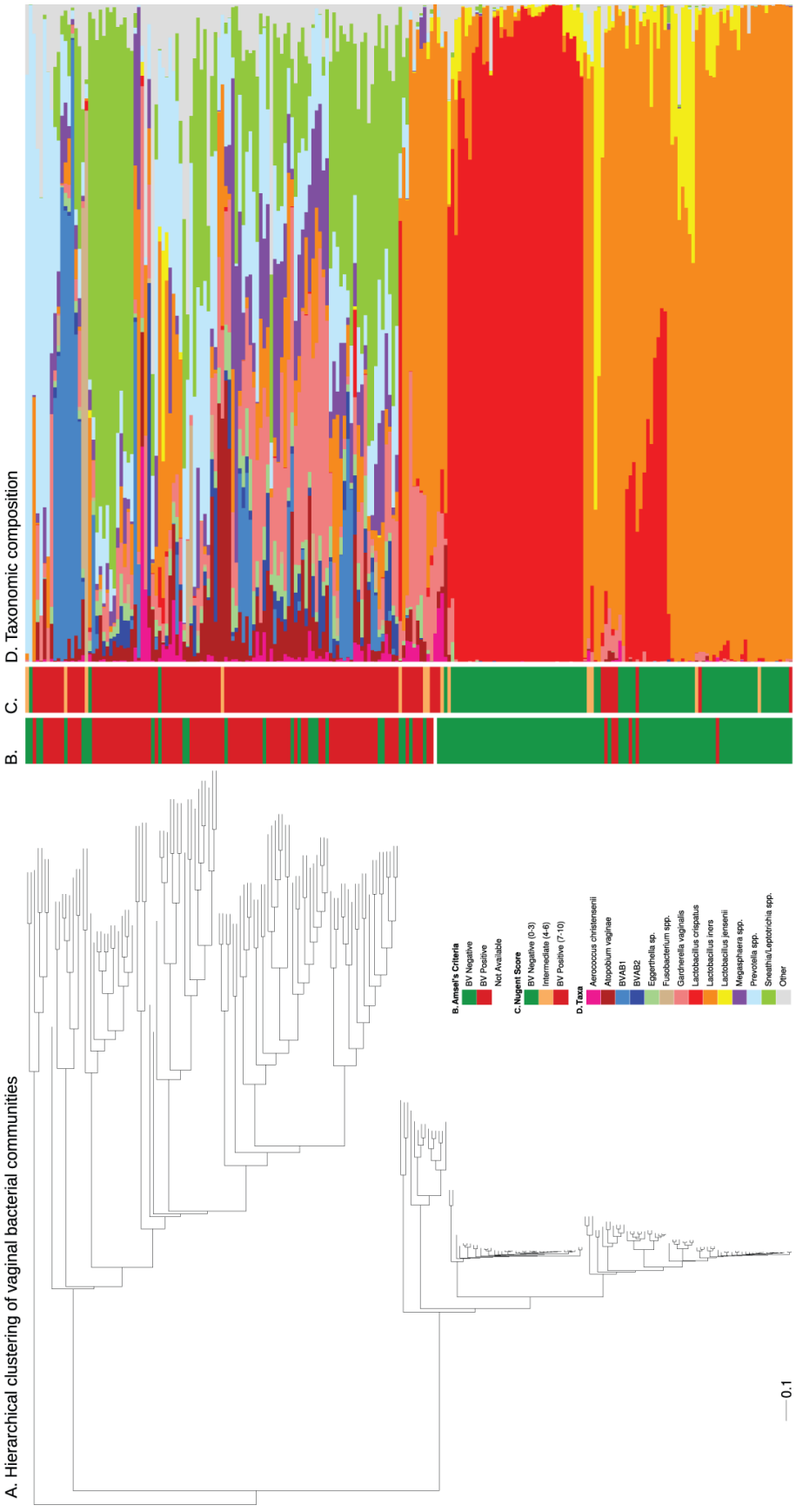
[14] Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4(4), 406–425 (1987)

[15] Srinivasan, S., Hoffman, N.G., Morgan, M.T., Matsen, F.A., Fiedler, T.L., Hall, R.W., Ross, F.J., McCoy, C.O., Bumgarner, R., Marrazzo, J.M., Fredricks, D.N.: Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS ONE 7(6), e37818 (06 2012), http://dx.doi.org/10.1371%2Fjournal.pone.0037818

[16] Stamatakis, A.: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21), 2688–2690 (2006), http://dx.doi.org/10.1093/bioinformatics/btl446

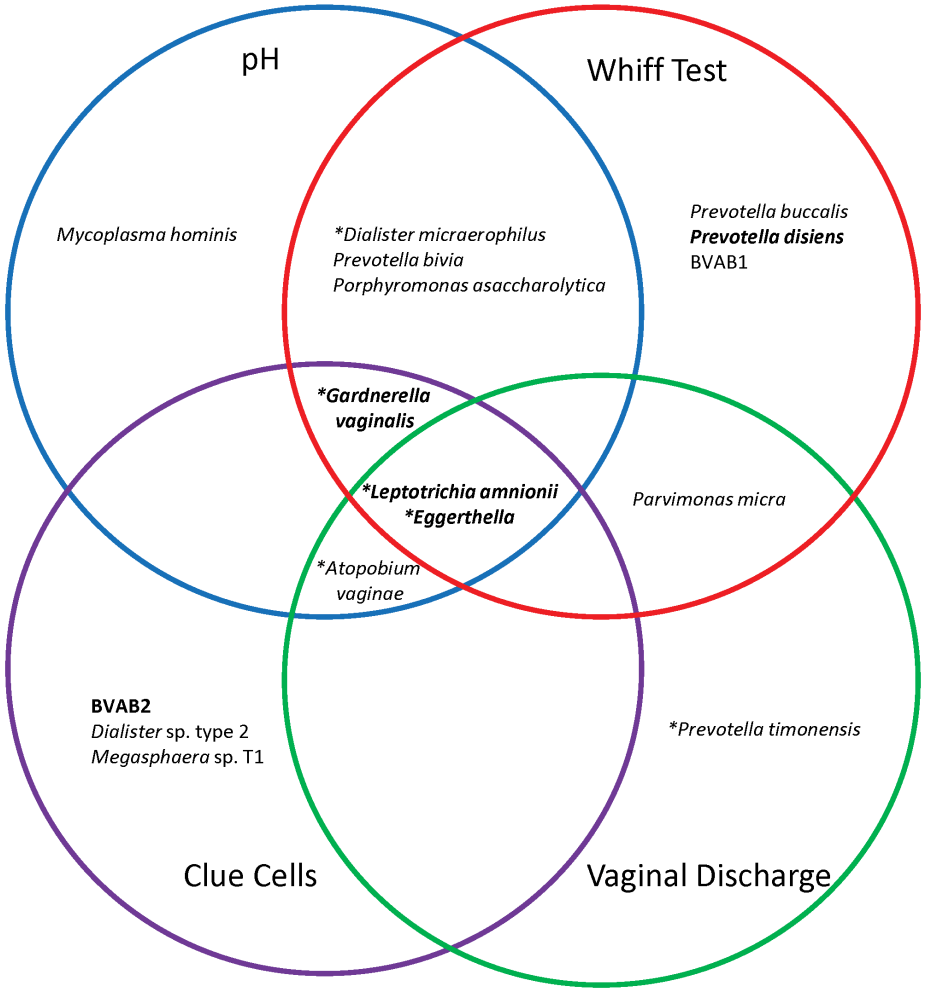
[17] Voelkerding, K.V., Dames, S.A., Durtschi, J.D.: Next-generation sequencing: from basic research to diagnostics. Clinical chemistry 55(4), 641–658 (2009)

Appendix A Clustering Tree



Taken from Srinivasan et al. [15].

Appendix B Associations of taxa with Amsel's criteria



Taken from Srinivasan et al. [15].