Spring 5-1-2022

Alterations of the Gut Mycobiome in Patients with MS - a Bioinformatic Approach

Saumya Shah
saumya.shah@uconn.edu

Alterations of the multiple sclerosis mycobiome:
a bioinformatic approach

Saumya Shah

UConn Department of Computer Science and Engineering, Honors

1 May 2021

Abstract

(123 W)

The mycobiome is the fungal component of the gut microbiome and is implicated in several autoimmune diseases. However, its role in multiple sclerosis (MS) has not been studied. We performed descriptive and formal statistical tests using the R language to characterize the gut mycobiome in people with MS (pwMS) and healthy controls. We found that the microbiome composition of multiple sclerosis patients is different from healthy people. The mycobiome had significantly higher alpha diversity and inter-subject variation in pwMS than controls. Additionally, *Saccharomyces* and *Aspergillus* were over-represented in pwMS. Different mycobiome profiles, defined as mycotypes, were associated with different bacterial abundances. Computer-based analysis of vast sequencing data will continue improving our understanding of the complicated microbiome community and their interactions with the host.

Keywords: microbiome, mycobiome, bioinformatics, computational biology

Introduction

Bioinformatics is an exciting new field that combines statistics, computer science, biology, and mathematics and is expanding as a field because of increased computing power and ability to collect and extract samples. Bioinformatics has focused on a number of research areas including gene expression, metabolism, and microbiome analysis. Human skin, mouth, and gut all contain different types of microbiome. The microbiome can change based on a number of factors, such as antibiotics, disease, diet, hygiene products, or climate.

Despite our growing understanding, the complex interactions between organisms and environment are not widely understood, specifically for the understudied diverse fungi population in the gastrointestinal tract called the gut *mycobiome*. It accounts for ~0.1% of gut microbiota and is ubiquitous in all human populations. The effect of the mycobiome has not yet been studied in patients with Multiple Sclerosis (pwMS). Multiple sclerosis (MS) is an autoimmune, chronic disease affecting the nervous system with unclear causes but it appears gender, genetics, and distance from the equator play a role. However, the gut-brain axis theory and bacteria training the immune system suggests that the gut flora affect systems beyond the intestines [1].

From my work with Dr Yanjiao Zhou on “Alterations of the gut mycobiome in patients with MS” at her laboratory in UConn Health, I will outline the methods and code used to characterize the MS mycobiome. This thesis will focus on the technical aspects of the analysis and critique of the methods used. There will be numerical and categorical data I will describe such as fungi abundance and treatment groups. I will cover how data modeling is interpreted in the framework of ecology and statistics, and what conclusions we draw. I will cover topics such as dimensional reduction, data normalization, connections to linear algebra, and statistics. I will also attempt to give biological and ecological justification for interpretations.

Methods and Results

Data quality and processing

22 healthy controls and 25 people with MS were recruited for the original study. We collected categorical information (MS diagnosis and race) and classified individuals on the binary of pwMS or control.

Stool samples are cultured and sequenced with ITS sequencing. The number of DNA sequences belonging to a certain fungi genus is called "read depth." The median read depth of each sample was 11863 and there were 3 high outliers out of 47 samples. Samples with read depth below 1000 were removed because of low data quality. The range of reads for each sample was 1146 to 2517870. This high variation in read depths is not unusual in microbiome sequencing. To study the microbiome, ecologists often rarefy data to normalize and correct for uneven sample depth through random subsampling. Some researchers argue that rarefying is unnecessary and produces high false positives because ITS sequencing naturally accounts for diversity and abundance in the sample [3][4]. However, other researchers point out that compared to non-normalized data, rarefying data is appropriate for beta diversity analyses and does not affect alpha diversity analysis as long as extremely small sample sizes were excluded [7]. Our study is very interested in the beta diversity of the pwMS and controls, so we determined it was appropriate to rarefy each sample to 10000 reads and convert reads to relative abundances.

To explain concepts in this essay, we will illustrate techniques with a subsample of the data. The original dataset had 22 controls and 25 pwMS and 56 total genera as well as each person's demographics, diet, immune factors, resampling at six-months, and bacteria. For simplicity's sake, I selected four patients, two MS and two controls, and five of the most salient genera to create a 4x5 matrix (Table 1). The numbers represent percent abundance of each fungi genus from each person so that each person's total abundance is 100%. When the values are proportional to each other, the data is called *compositional.* Additionally, the term "sample" can be used interchangeably with "person" in our context. To protect patient privacy, each sample has a codename PID##. Finally, in our example data, the order of samples or the order of bacteria could be sorted differently without losing the meaning of the values.

Because of the nature of sampling and compositional data, drawing conclusions on true abundance is difficult. For example, species abundances often varies widely between samples, as well as among species, where highly abundant species risk burying rarer species. Additionally, when comparing abundances over time, it's unknown if a particular species grew in number *in vivo* or if another species lost abundance and inflated the other's abundance. Compositional data standardizes an equal number of reads across samples, but some of the techniques used to analyze this type of data need further refinement [5].

In real microbiome data sets, most of the data is sparse, meaning that there are many instances of zero counts for several species. It's important to note that zeros in the abundance matrix do not necessarily mean that there were zero of that species in the sample; it appears as zero if the low number of that species is overshadowed by the larger abundance of another species. Also, even if one participant's sample has zero of that fungi, the fungi could still exist in that person and was simply not sequenced deeply enough in that single sample. The value zero should not be treated as an absolute.

The human fungi microbiome is understudied so there are gaps in the genetic databases, leading to a high proportion of unclassified genera. It's unclear whether these unclassified DNA sequences belong to discovered genera, unsequenced genera, produce labels with low confidence, or do not belong to fungi at all.

Table 1	Unclassified	Saccharomyces	Xylaria	Aspergillus	Other
PID1_MS	17	82	0	0	1
PID2_MS	1	94	.01	4	.99
PID3_Ctrl	96	2	0	0	2
PID4_Ctrl	75	15	10	0	0

Table 1 – Subsample of the original abundance matrix. 2 pwMS out of 25 and 2 controls out of 22 have their fungi summarized above. We will use this 4x5 matrix to walkthrough more complex algorithms.

```
set.seed(123)
summary(rowSums(counts))
genus = rrarefy(counts, 10000)
genus = as.data.frame(prop.table(as.matrix(genus), 1)*100)
```

Descriptive analysis

We first examined the gut mycobiome composition in pwMS and healthy control individuals. At the phylum level, 92.5% of phyla were identified. *Ascomycota* and *Basidiomycota* were the two predominant phyla in both pwMS and controls, together accounting for over 80% of total mycobiome population (Figure 1a). At the genus level, we identified 59 genera, of which 25 (excluding the unclassified fungi shared between the two groups) made up 48% of the total relative abundances (Fig 1b). Most sequences can be identified at the phylum level whereas at the more differentiated genera level, the label is unknown.

Saccharomyces, *Xylaria*, *Pencillium*, *Agaricus*, and *Aspergillus* were the top 5 most abundant classified fungi in the gut in both groups. On average, *Saccharomyces* composed 23% and 42% of the gut mycobiome in control and pwMS, respectively. pwMS samples had greater proportions of *Aspergillus* ($p=0.008$, $p_{adj}=0.02$, Wilcoxon) and *Saccharomyces* ($p=0.005$, $p_{adj}=0.02$, Wilcoxon) (Fig 1c). The variance of *Saccharomyces* was significantly higher in pwMS than those in controls ($p=0.0004$, levene test). *Saccharomyces* and *Aspergillus* will be genera of interest in further analysis.

To identify specific genera differences, we performed a Wilcox rank-sum test that tests if the distribution is the same between pwMS and controls. The term “non-parametric” refers to methods that manipulate the data without assuming an underlying probability model. Instead, they use the data to determine the model, making them more flexible but less powerful than parametric methods. Usually the input data is processed as ordinal rather than nominal. Ordinal data is non-metric meaning that one rank does not have a specified magnitude compared to another rank. Using the ordinal methods reduce the dominance of highly-abundant genera on the rest of the dataset.

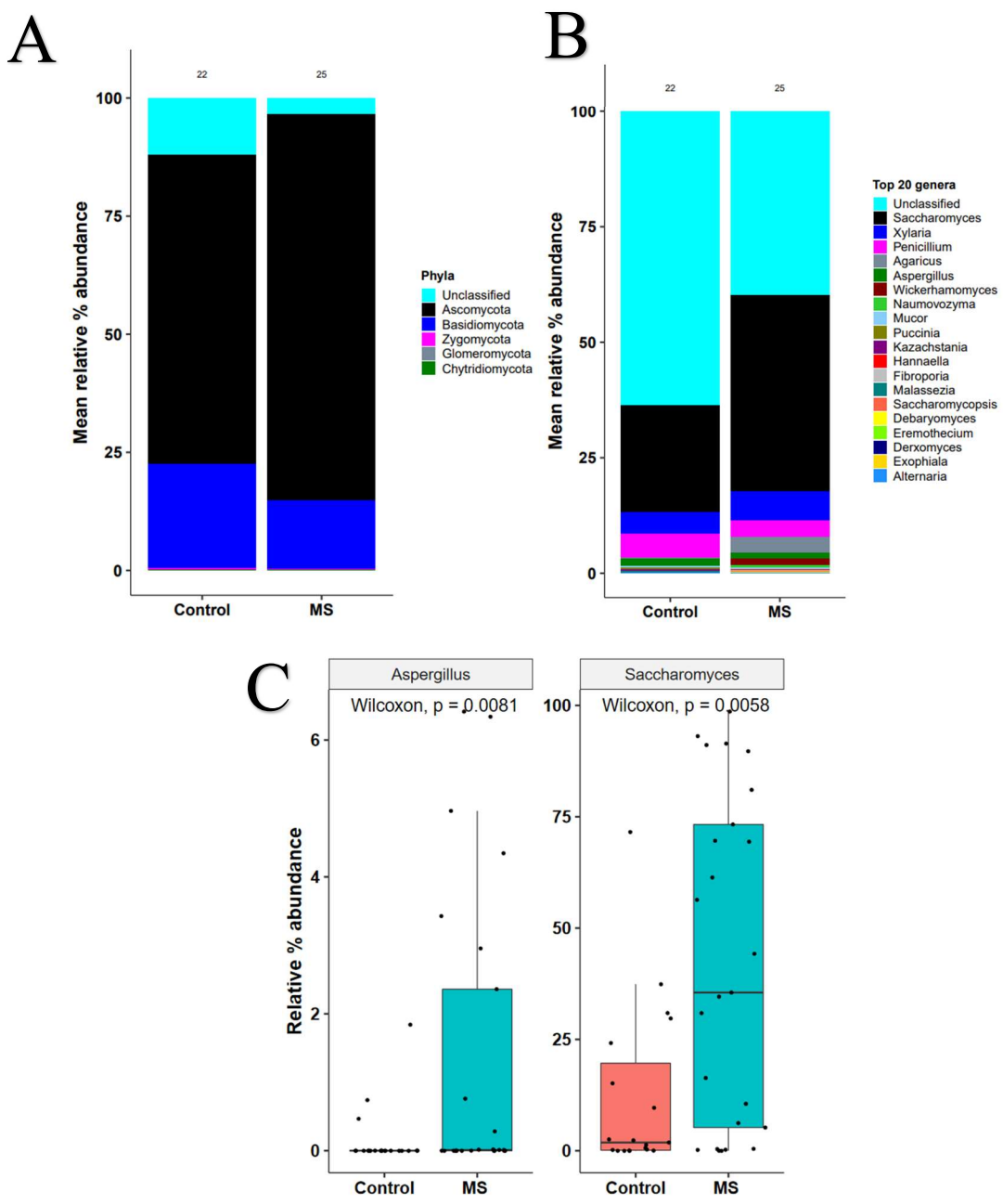

Figure 1: Mycobiome compositions in pwMS and control groups, based on ITS sequencing analysis. (a) Mean relative abundances of fungi phyla in each sample. "Unclassified" represent unknown fungal phyla. (b) Mean relative abundances of the 20 most abundant fungi genera in each sample. "Unclassified" represents unknown fungal genus. (c) Relative abundance of *Aspergillus* (p=0.008, p_{adj}=0.02, Wilcoxon) and *Saccharomyces* (p=0.005, p_{adj}=0.02, Wilcoxon) in the two groups.

```
ggplot(data = top20genus, aes(x = Group, y = value)) + geom_bar(stat = 'identity', aes(fill = variable))+scale_fill_manual(c('Top 20 genera'), values = COLORS)+ ylab('Mean relative % abundance')+ annotate("text", x=1, y=105, label='22')+ annotate("text", x=2, y=105, label='25')+theme_pubr(base_size=22, legend="right")+labs_pubr(base_size=22)+xlab(")
```

```
for (x in colnames(data)){
  wilcox.test(data[,x]~data[,"Group"])["p.value"] }
```

```
#boxplot
ggplot(melt(data, id.vars = ”Group”), aes(x=Group, y=value))+geom_boxplot(outlier.color = 'transparent') +
facet_wrap(~variable)+geom_jitter() + annotate("text", x=1, y=9, label='22')+stat_compare_means() +
annotate("text", x=2, y=9, label='25')
```

Diversity

Alpha diversity is the variation of species within a sample and was evaluated by Shannon diversity and species richness. Stools from pwMS had significantly greater mycobiome richness (p=0.041, Wilcoxon) and Shannon diversity (p=0.043, Wilcoxon) (Fig. 2) than stools from controls, suggesting overgrowth of different types of fungi in the gut mycobiome in pwMS. There were 18 genera exclusive to pwMS samples. For example, *Phlebia* and *Rhizopus* were each detected at very low abundances but in pwMS samples only.

Beta diversity is the compositional differences in subjects in different groups, namely pwMS and controls. We test the beta diversity using permutational multivariate analysis of variance analysis (PERMANOVA). The nonparametric method PERMANOVA performs permutations on the dissimilarity matrix rather than assuming a normal distribution. Specifically it uses the Bray-Curtis dissimilarity matrix which will be explained further in the next section. It calculates the sum of squares of the dissimilarity matrix to compute the F-ratio, a measure of variation between and within groups and compare it with the appropriate F-distrubtion depending on the sample size. The larger the ratio, the greater the variation between groups. The overall mycobiome community differed between pwMS and controls (p=0.04, PERMANOVA).

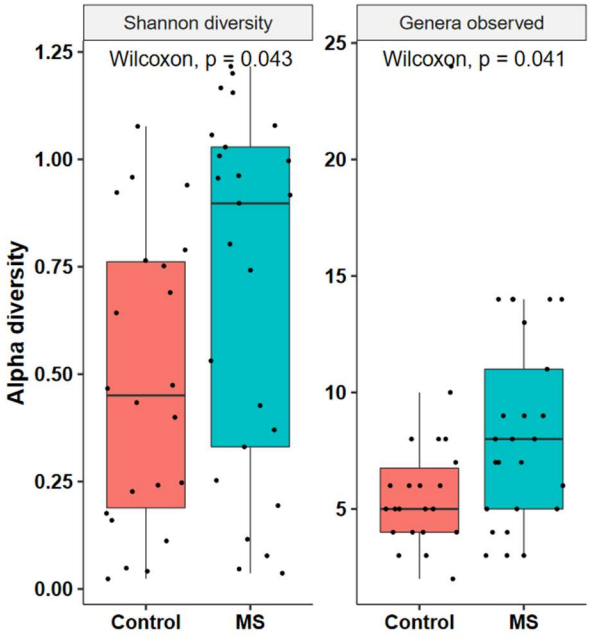

Figure 2 - Alpha diversity variation between the two groups, expressed as Shannon diversity and observed richness

```
#Alpha: Shannon
shannon = as.data.frame(diversity(counts, index='shannon'))
#Alpha: Observed richness
```

```
diver = cbind(Group, shannon, specnumber(counts))
ggplot(melt(diver, id.vars='Group'), aes(x=Group, y=value, fill=Group))+geom_boxplot(outlier.colour =
'transparent')+facet_wrap(~variable, scales='free_y')+geom_jitter(col='black')+ylab('Alpha
diversity')+theme_pubr(base_size=22,legend='none')+labs_pubr(base_size=22)+xlab('')+stat_compare_means()

#Beta
a = adonis(vegdist(dats[,-1], method="bray") ~ Group)
tidy(a[[1]])$p.value[1]
```

Similarity

Similarity is a measure of how each sample compares to each other sample. The Bray-Curtis index for two samples *i* and *j* is calculated with the formula:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where C_{ij} is the sum of lesser counts for bacteria common to each site.

There exist a number of indexes such as Cosine similarity and Minkowski distance for different analysis, but Bray-Curtis is the standard in microbiome and ecological analysis. Bray-Curtis (BC) dissimilarity is also preferred over Euclidean distance $E_{ij} = \sqrt{s_i^2 + s_j^2}$ in biological contexts because our data is compositional, and the Bray-Curtis index is less affected by the number of null values. As with all dissimilarity matriecs, the BC matrix is symmetric. The BC values range from 0 to 1 where 1 means the samples have no bacteria in common and 0 means that they have identical bacteria.

Using the subsampled data matrix from Table 1, the BC index for PID_02_MS and PID_03_Ctrl is the greatest at 0.96 (Table 2). We can verify this qualitatively with the abundance matrix (Table 1) where we see that PID_02_MS and PID3_Ctrl have opposite amounts of *Saccharomyces* and Unclassifed making them very dissimilar. However, the two pwMS samples have very low dissimilarity (0.16) meaning that these samples were quite alike in composition, suggesting that the microbiome composition for pwMS and controls samples are distinct from each other. In the next section, we can continue to visualize the difference between samples in ordination plots.

Table 2	PID_01_MS	PID_02_MS	PID_03_Ctrl	PID_04_Ctrl
PID1_MS				
PID2_MS	.16			
PID3_Ctrl	.80	.96		
PID4_Ctrl	.68	.84	.23	

Table 2 – The Bray-Curtis dissimilarity matrix for the samples from Table 1. Higher values indicate that the sample are more dissimilar.

```
vegdist(dats, method = 'bray')
```

Ordination

Dimensional reduction is key for working with the original dataset that is 47 x 56 dimensions and very sparse. Dimensionality reduction reduces time and space complexity and allows for 2D visualization with NMDS and PCA visualization.

Non-metric Multidimensional scaling (NMDS)

While we were able to verify the dissimilarities using the tables above, more complex data sets will have many more samples and bacteria making it difficult to compare manually. Multidimensional scaling creates 2D plots that condense and show the similarity of several multidimensional samples. Because there is a transformation from multidimensional space to 2D space, there is a numerical value called "stress" that shows how much the coordinates had to change to become 2D. The goal of an MDS algorithm is to minimize stress as much as possible through each iteration. The formula of stress is:

$$\text{Stress} = \sqrt{\frac{\sum \left(f(x) - d\right)^2}{\sum d^2}}$$

where x is a vector and f(x) is the transformation. The isoMDS engine has a $O(n^2)$ runtime which will only become more computationally expensive with larger datasets.

Since Bray-Curtis dissimilarity is the best way to compare ecological samples, we input it into the NMDS algorithm where it calculates stress based on rank orders. All of the samples have been transformed into a 2D coordinates, and the relative distance between the points illustrates how similar points are to each other – albeit with some distortion during NMDS calculation. The axes do not have a physical meaning but are a scale to show the distance between samples. For example, PID2_MS will be displayed far from PID3_Ctrl and close to PID1_MS because the pwMS samples both happened to have similar fungi compositions (Fig 3a). The stress of the 4x5 subsample is very low, nearly 0, meaning that it the NMDS represents the transformed points excellently. This plot confirms what we know about the similarity of the subsampled patients.

We can then create an NMDS for the 47 total samples and look for any separation between the MS and controls (Fig 3b). It seems at first that the pwMS and Control points are neither clearly separated nor mixed. However looking more closely at it, we see areas of mixing on the right side, but also a pwMS cluster toward the top left and a controls cluster on the bottom. We confirmed statistically that they separated with the PERMANOVA test ($p=0.04$). The stress for this plot is 21 which is passable considering that the high dimensionality of the dataset.

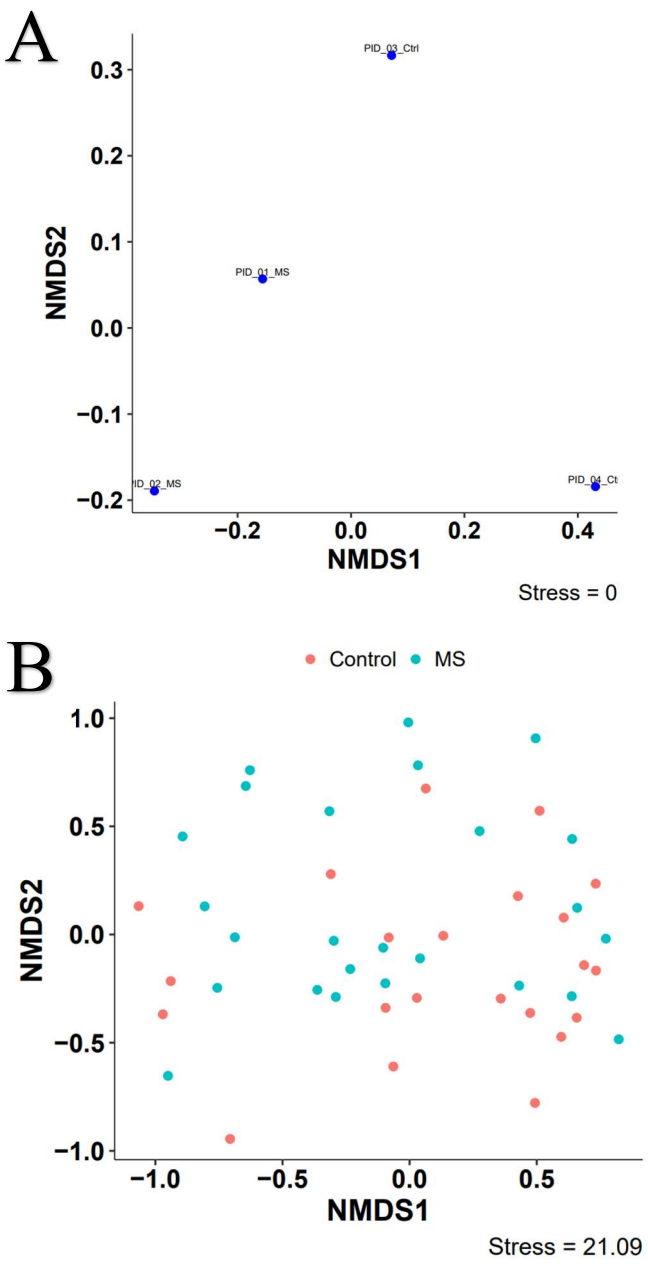

Figure 3 – NMDS plots of Bray-Curtis dissimilarity of controls and pwMS mycobiome. Stress values shown. (a) n=4 subsample. (b) n=22 controls and n=25 pwMS.

```
ord = metaMDS(data, 'bray', engine='isoMDS')
plot(ord, display = 'sites', type='t')
ggplot(as.data.frame(ord[["points"]]), aes(x = V1, y = V2))+ geom_point(size=5,
col='blue')+xlab('NMDS1')+ylab('NMDS2') +theme_pubr() + labs_pubr(caption=paste0('Stress = ',
format(ord[['stress']], digits=4))) +geom_text(aes(label=rownames(data),vjust=-0.5))
```

Principal component analysis (PCA)

Like multidimensional scaling, PCA is another visualization technique that reduces the dimensions and extracts the features to contrast the distance between samples. PCA projects a high-dimensional dataset to 2 dimensions by decomposing the covariance or correlation matrix into eigenvectors. PCA rotates and projects samples on new axes to project the *maximum variance* and to minimize information loss. The eigenvalues determine the magnitude that is the variance explained by each axis.

To normalize the data, we isometric log-ratio transform and scale the data. Then we find the eigenvalues for matrix A by solving $(A-\lambda I)x = 0$ for λ. To get the eigenvector x, simply substitute λ into the previous equation. The second axis is orthogonal to the first axis, and while it is possible to have more than two axes, 2D plots are generally easier to visualize and understand. The total variance is defined as $\sum_{j=1}^{k} s_{jj}$ which equals the sum of eigenvalues for the covariance matrix S.

For the subsample of data, the principle component one explained 72.22% of variance, and component two explained 25.36% which together accounts for 97.58% of the variance. Because a high percentage of the variance is explained through the PCA, a 2D plot is a satisfactory depiction of the data (Fig 4a). The arrangement of points is similar to NMDS where PID2 and PID4 are very different from each other and are thus placed furthest apart. The PCA differs from the NMDS as it places the pwMS samples PID1 and PID2 quite far from each other.

As expected, the n=47 dataset's depictions will differ between NMDS and PCA. The NMDS showed the clustering of pwMS versus controls more clearly and relies on Bray-Curtis dissimilarities rather than Eucliean. The PCA only depicted (21.44 + 15.82=) 37.26% of the variance using the first two principal components as axes. While that is low, it is understandable given that the original dataset that is 47 x 56 dimensions and there are only two dimensions to show the distance between samples. The PCA plot is not as comprehensive as the NMDS plot to show the stretching of the multidimensional, nonlinear dataset.

In both the NMDS and PCA plots, we notice points that are very close together, meaning that the samples have very similar fungal compositions. In the next section we can formally assign points to clusters.

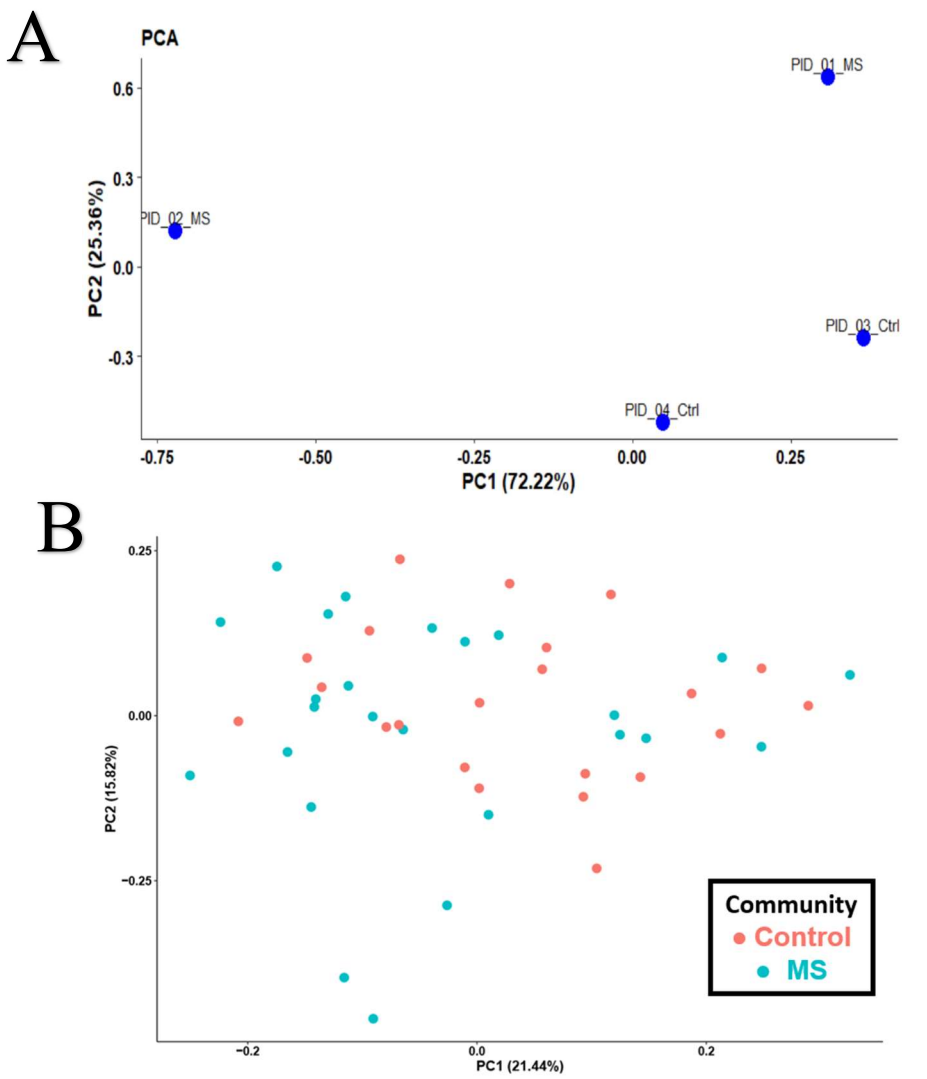

Figure 4 – PCA plot of log-transformed Euclidean distance of controls and pwMS mycobiome. (a) n=4 subsample. (b) n=22 controls and n=25 pwMS.

```
library(ggfortify)
log.AC = ilr(funclust[,-1])
log.AC = as.data.frame(log.AC)
AC.pca = prcomp(log.AC)
autoplot(AC.pca, data= funclust, colour = 'Mclust')
```

Clustering

As we saw in the previous ordination section, we can search for clusters of samples with similar mycobiome profiles. The gut microbiome can be classified into three types (termed enterotypes) based on their relative abundances [8], which are associated with different health and disease conditions. We want to test whether there are fungal clusters ("mycotypes") using clustering techniques. Clustering is useful to find the similarity of samples within the same classification and the dissimilarities to other clusters.

The advantages of the k-means clustering algorithm is that it compresses high-dimensional data to show a small number of clusters of data that is similar to each other. The global objective of partitional algorithms is to minimize distances within clusters and maximize inter-cluster distance. k mean data points are chosen in each iteration to decrease the distance from it and the surrounding data points to make a cluster. We calcualate the sum squared error (SSE) of each sample to the cluster mean and reassign to the closest cluster. The complexity of the algorithm depends on the number of samples, the number k of clusters, the number of iterations, and number of genera. We can increase the iterations for a more precise point where the cluster mean converges. We can interpret the cluster mean as the prototypical member of the mycotype.

k Means is a special case of the Expectation-Maximization (EM) algorithm. The EM algorithm soft assigns points to a cluster by calculating the probability of a point belonging to each cluster using k Gaussian distributions. Against a Gaussian mixture model that combines ('mixes') multiple Gaussian distributions, the k Means algorithm produced tighter clusters with greater average distance between them. Statistics of the k-means clusters are that the average diameter of each cluster is 0.517 and 0.684 and the average distance within clusters was 0.270. The separation between the two clusters was 0.313, and the average distance between clusters was 0.835. The k-Means entropy, which is a measure of cluster's information gain, was slightly higher than that for the Gaussian mixture (0.67 vs 0.64). The Calinski and Harabasz index for k Means (96.95) is much stronger than that of the Gaussian mixture model (2.22). The Dunn index is a similar method to internally validate clustering, and the k-means Dunn index was much higher than that of the Gaussian (0.458 versus 0.004). All of these scores means that k-Means is a stronger choice for for the cluster analysis algorithm than the Gaussian mixture.

k Means operates under an assumption of spherical, similar density clusters, so if the clusters were to have different properties such as size, density, or shape then k-means will struggle to output the right clusters. In that case, we can increase k to find parts of clusters and post-process the data to put the cluster together. One disadvantage is that the user may not initially know the right number of k clusters, but the user can calculate a silhouette score for a range of k and choose the maximum score to indicate good clustering. Using all of the data, we determine that the average silhouette width was 0.665 and that optimized for k=2 clusters (Fig 5a). Manual inspection and testing different k clusters is important to ensure that the optimal parameters are being passed.

Let's examine the k-Means algorithm using the subsample data to better understand how samples are assigned to a particular cluster. In this method of k-Means, the clusters are initialized by a random selection of k rows. The second iteration is sufficient to assign clusters because each sample is already assigned to the cluster its closest to (Table 3). When this algorithm is performed on the larger matrix, more samples and genera make the process more computation and memory intensive.

The computed mycotypes had a reciprocal ratio of pwMS and controls, which supports the hypothesis that the controls and pwMS have different mycobiome compositions. Mycotype 1 consists of 4 controls

and 10 pwMS, whereas Mycotype 2 has 14 controls and 8 pwMS. Mycotype 1 is dominated by *Saccharomyces* and *Pencillium* and is slightly more diverse, whereas Mycotype 2 is dominated by unclassified genera, and a higher average proportion of *Xylaria*, *Agaricus,* and *Aspergillus* (Fig 5b). The clusters created by the k-means are roughly equal density and spherical, with some area of overlap (Fig 5c).

Table 3: Iteration 1	Unclassified	Saccharomyces	Xylaria	Aspergillus	Other	distance from 1	distance from 2	Assignment
PID1	17	82	0	0	1	89.185	20.396	2
PID2	1	94	0.01	4	0.99	108.78	0	2
PID3	96	2	0	0	2	26.721	132.310	1
PID4	75	15	10	0	0	0	108.783	1

Table 3: Updated center	Unclassified	Saccharomyces	Xylaria	Aspergillus	Other
Cluster 1	85.5	8.5	5	0	1
Cluster 2	9	88	0.005	2	0.995

Table 3: Iteration 2	distance from 1	distance from 2	Assignment
PID1	100.60	10.198	2
PID2	120.380	10.198	2
PID3	13.360	122.352	1
PID4	13.360	98.9439	1

Table 3 – Kmeans iterations for the subsample data from Table 1.

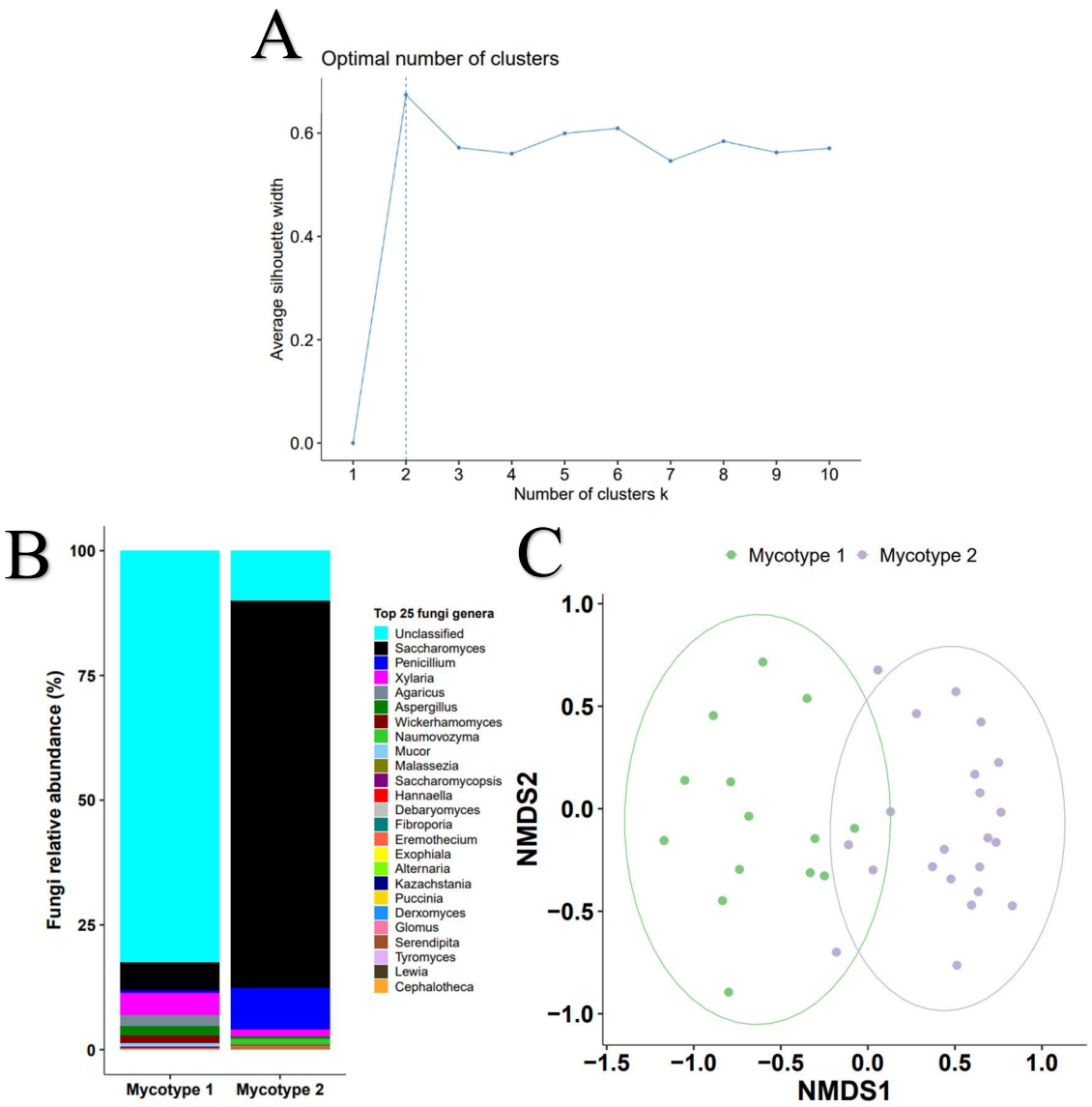

Figure 5 - Interaction between fungal and bacterial microbiome. (a) Silhouette coefficient. (b) Top 20 most abundant mycobiome genera were used to build the model. Barplots were used to show the mean relative abundances of mycobiome in each mycotype. Mean relative abundances of the major fungi in two mycotypes identified by Kmeans partitioning ($n_1 = 22$, $n_2 = 14$). (c) NMDS of n=36 samples colored by Kmeans assignment to cluster 1 or 2.

```
set.seed(123)
fviz_nbclust(funclust, kmeans, method='silhouette')+theme_pubr(base_size = 22)
funclust$Clust = kmeans(funclust,2, iter.max=50)$cluster
cluster.stats(vegdist(funclust), clustering = kmeans(funclust,2)$cluster)

ggplot(bacclust, aes(x=as.factor(-Clust), y=Parabacteroides, fill = as.factor(Clust)))+ geom_boxplot(outlier.color = 'transparent') +
geom_jitter(col='black')+ ggtitle('Wilcox test of bacteria\nbetween mycotypes') + stat_compare_means(size=7) +
scale_fill_brewer(palette='Accent') +
theme_pubr(base_size=22,legend='none')+labs_pubr(base_size=22)+ylab('Parabacteroides')+xlab('')+scale_x_discrete(labels=c("Mycotype 1",
"Mycotype 2"))

ord = metaMDS(dats, 'bray', engine='isoMDS')
mm= as.data.frame(ord[["points"]])[rownames(funclust),]
ggplot(droplevels.data.frame(mm), aes(x = V1, y = V2, col = as.factor(funclust$Clust)))+
geom_point()+xlab('NMDS1')+ylab('NMDS2')+theme_pubr(base_size=22)+labs_pubr(base_size=30)+labs(color='')+scale_color_brewer(palette
='Accent', labels=c('Mycotype 1','Mycotype 2'))+stat_ellipse(type='norm', level=.9)
```

Interactions between the gut mycobiome and bacterial microbiome

With the mycotypes established, we next want to determine how a fungi composition is related to various bacteria, as the human gut is an ecological system with several actors affecting each other. In this clustering analysis, we excluded n=11 participants who took antibiotics because antibiotics would confound the bacterial-fungi correlations and relationships. We compare the mean bacterial abundances of the participants in Mycotype 1 versus those Mycotype 2 (Fig 6a). With a Wilcox test of means, we found that *Parabacteroides* (p=0.045) is significantly greater in Mycotype 1 than in Mycotype 2. In other words, samples dominated by *Saccharomyces* and lower overall fungal diversity, characteristic of Mycotype 2, tend to have less *Parabacteroides* (Fig 6b).

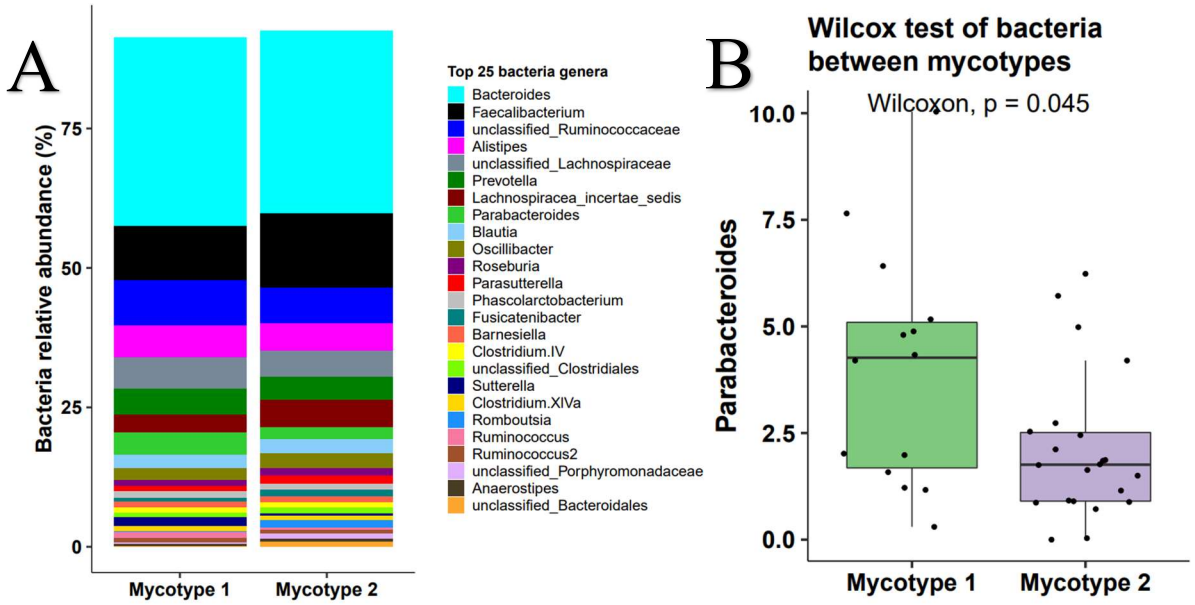

Figure 6 – (a) Mean relative abundances of the gut bacteria in participants belonging to the two mycotypes ($n_1 = 14$, $n_2 = 22$). (b) Significantly over-represented Parabacteroides (p=0.045, Wilcoxon) in Mycotype 1.

```
wilcox.test(diversity(funclust[,-1], index='shannon')~funclust$Clust)
```

```
for (x in colnames(bacclust)){
  wilcox.test(bacclust[,x]~bacclust[,"Clust"])["p.value"] }
```

Correlations

To investigate relationships between fungi and bacteria in the gut, we performed Pearson correlations with all participants at baseline, without discriminating between pwMS and controls. Correlation analysis were performed for two continuous variables correlation. Correlations with correlation coefficient $r < -0.3$ or $r > 0.3$ were reported. The p-values were less than 0.05 in all reported correlations. All the correlations were manually inspected by scatterplotting raw data and correlations driven by one or two data points were not reported.

We found a high correlation between *Mucor* (fungus) and *Fusicatenibacter* (Fig 7, r=0.81, p<0.001). *Fusicatenibacter* is a bacterial genus belonging to *Clostridium cluster XIV* and *Lachnospiracea incertae sedis*. The abundance of the bacterial genus *Prevotella* highly correlated with the fungi *Hannaella* (r=0.62, p<0.001) and *Derxomymces* (r=0.39, p=0.019), both of which are low in abundances. *Alistipes*, one bacterial genus that is associated with MS [40], was positively correlated with *Penicillium* (r = 0.68, p<0.001). When MS and control samples were separated, *Saccharomyces* which had notably greater abundance in MS samples, was negatively correlated with *Lachnospiracea insertae sedis* (r=-0.48, p=0.04) in pwMS. In controls alone, *Saccharomyces* and *Oscillibacter* were positively correlated (r = 0.51, p=0.029). While these correlations do not imply causation, they can be useful for future research to see if there is any biological or metabolic interaction between microorganisms.

There are techniques like Sparse Correlations for Compositional Data (SparCC, SpiecEasi package) that address issues with microbiome datasets, namely sparsity and compositionality [18]. Its robustness comes from log-ratio transformations of the data and $O(n^2)$ iterations of bootstrapping (random resampling with replacement). The input matrix is one abundance table, and the output is a fungi-fungi intracorrelation table. However the correlations for the dataset were weak, suggesting that fungi did not highly associate with each other (Data not shown, $r < |0.55|$).

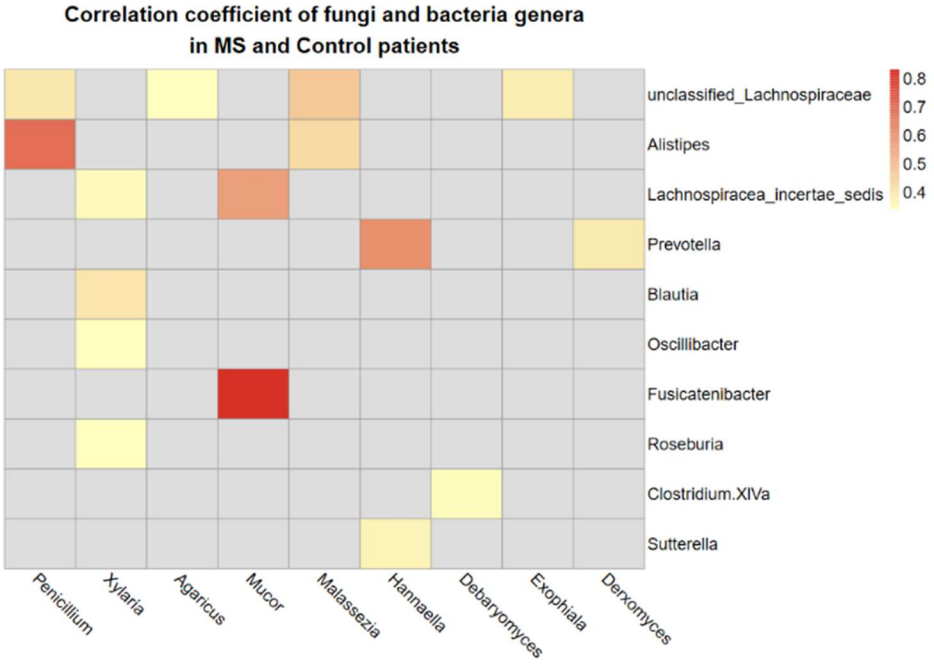

Figure 7 - Interaction between fungal and bacterial microbiome. Heatmap shows fungi-bacteria Pearson correlations. Nonsignificant correlations are omitted.

```
pheatmap(cormat_bac, main = 'Correlation coefficient of fungi and bacteria genera\nin MS and Control patients',
cluster_cols = F, cluster_rows = F, angle_col = 315,fontsize=18, color =
colorRampPalette(c('#ffffbf','#d73027'))(100))
```

```
corr.test(bac[,rownames(cormat_bac)], fungi[,colnames(cormat_bac)],method='pearson',
use='pairwise.complete.obs', adjust='none')
```

Conclusion

This study is the first to define the gut mycobiome of pwMS. Both pwMS and controls were dominated by *Saccharomyces*, *Xylaria*, and *Penicillium*. Nash et. al identified *Saccharomyces*, *Malassezia*, and *Candida* as the dominant genera in healthy samples [9], while Hoffman et. al identified *Saccharomyces*, *Candida*, and *Cladosporium* as the most abundant in healthy control samples [10]. On average, *Candida* and *Malassezia* only account for 0.26% and 0.86% of all our samples, respectively, and *Cladosporium* was not detected at all.

We compared mycobiome diversity and compositions in pwMS and healthy control samples. Overall, the results indicate that pwMS samples had higher mycobiome diversity and greater inter-subject variation compared to the mycobiome of healthy controls. Microbial diversity is an indicator of gut health and it is generally determined by the bacterial diversity. Higher diversity is associated with healthier gut ecosystem. Our data suggests that the mycobiome diversity is an important indicator of the gut microbial diversity in MS patients, and may be more sensitive than the bacterial diversity in response to MS associated gut microbiome changes.

The gut mycobiome accounts for less than one percent of the entire microbiome, therefore it may be considered rare relative to the bacterial microbiome [14]. However, the rare biosphere can have disproportionate effects on health and diseases [15]. We found two commensal gut fungal genera *Saccharomyces* and *Aspergillus* that were more abundant in pwMS. *Saccharomyces* is called the baker and brewer's yeast for its role in food fermentation, as well as commonly being used as a gut probiotics [11]. However, the biological role of *Saccharomyces* in health and diseases is inconsistent across studies, with either protective or detrimental effect on the gut inflammation [12], [13] in animal model of IBD. Our results support a pathogenic correlation of *Saccharomyces* with MS. *Aspergillus* is a genus consisting of several mold species. It is a member of respiratory and gut mycobiome. *Aspergillus* produce aflatoxins and can cause opportunistic infections in humans [16]. Except direct infection in CNS, it is possible that the gut *Asperigllus* activates gut immune response that indirectly affect systemic or CNS inflammation [17].

Future machine learning applications to this dataset is k-Nearest Neighbors or Random Forest classification of patient samples as a potential diagnostic tool for multiple sclerosis. With enough training records of control and MS mycobiomes, we could create a model and test its validity with various mycobiome profiles to see how accurately it predicts an MS or healthy label.

References

All plots were made with “ggplot2” and formatted with “ggpubr” except for heatmaps (“pheatmap”) and PCA plots, which were made with “compositions” to log-transform the data and “ggfortify” to plot it. “Vegan” (v. 2.5-6) was used for count rarefying, Shannon diversity, specnumber, vegdist, adonis, and Bray-Curtis dissimilarity. “fpc” was used for cluster analysis and “psych” was used for matrix-wide correlation analysis. All p-values from multiple comparisons were adjusted with false discovery rate method. All analysis was done in RStudio version 1.2.1 and R 4.0.3.

[1] Carabotti M, Scirocco A, Maselli MA, Severi C. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. Ann Gastroenterol. 2015;28(2):203-209.

[2] Fitzpatrick, Z., Frazer, G., Ferro, A. *et al.* Gut-educated IgA plasma cells defend the meningeal venous sinuses. *Nature* **587,** 472–476 (2020). https://doi.org/10.1038/s41586-020-2886-4

[3] Gloor, G B, Macklaim, J M, Pawlowsky-Glahn V, Egozcue J J (2017) Microbiome Datasets Are Compositional: And This is Not Optional. Frontiers in Microbiology (8). https://doi.org/10.3389/fmicb.2017.02224

[4] McMurdie PJ, Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10(4): e1003531. https://doi.org/10.1371/journal.pcbi.1003531

[5] Tsilimigras, M. C. B. Fodor, A A (2016). Compositional Data Analysis of the microbiome: fundamentals, tools, and challenges. Annals of Epidemiology (26). https://doi.org/10.1016/j.annepidem.2016.03.002.

[6] Yamamoto Y, Osanai S, Fujiuchi S, Akiba Y, Honda H, Nakano H, Ohsaki Y, Kikuchi K. Saccharomyces-induced hypersensitivity pneumonitis in a dairy farmer: a case report. Nihon Kokyuki Gakkai Zasshi. 2002 Jun;40(6):484-8. Japanese. PMID: 12325333.

[7] Ellen S. Cameron, Philip J. Schmidt, Benjamin J.-M. Tremblay, Monica B. Emelko, Kirsten M. Müller. To rarefy or not to rarefy: Enhancing microbial community analysis through next-generation sequencing. bioRxiv 2020.09.09.290049; doi: https://doi.org/10.1101/2020.09.09.290049

[8] M. Arumugam *et al.*, “Enterotypes of the human gut microbiome,” *Nature*, vol. 473, no. 7346, pp. 174–180, May 2011, doi: 10.1038/nature09944.

[9] A. K. Nash *et al.*, “The gut mycobiome of the Human Microbiome Project healthy cohort,” *Microbiome*, vol. 5, no. 1, p. 153, Nov. 2017, doi: 10.1186/s40168-017-0373-4.

[10] M. C. Noverr and G. B. Huffnagle, “Regulation of Candida albicans morphogenesis by fatty acid metabolites,” *Infection and Immunity*, vol. 72, no. 11, pp. 6206–6210, Nov. 2004, doi: 10.1128/IAI.72.11.6206-6210.2004.

[11] T. Kelesidis and C. Pothoulakis, “Efficacy and safety of the probiotic Saccharomyces boulardii for the prevention and therapy of gastrointestinal disorders,” *Therapeutic Advances in Gastroenterology*, vol. 5, no. 2, pp. 111–125, Mar. 2012, doi: 10.1177/1756283X11428502.

[12] T. T. Jiang *et al.*, “Commensal Fungi Recapitulate the Protective Benefits of Intestinal Bacteria,” *Cell Host & Microbe*, vol. 22, no. 6, pp. 809-816.e4, Dec. 2017, doi: 10.1016/j.chom.2017.10.013.

[13] T. R. Chiaro *et al.*, “A member of the gut mycobiota modulates host purine metabolism exacerbating colitis in mice,” *Sci Transl Med*, vol. 9, no. 380, 08 2017, doi: 10.1126/scitranslmed.aaf9044.

[14] G. B. Huffnagle and M. C. Noverr, “The emerging world of the fungal microbiome,” *Trends Microbiol.*, vol. 21, no. 7, pp. 334–341, Jul. 2013, doi: 10.1016/j.tim.2013.04.002.

[15] A. Jousset *et al.*, “Where less may be more: how the rare biosphere pulls ecosystems strings,” *The ISME journal*, vol. 11, no. 4, pp. 853–862, 2017, doi: 10.1038/ismej.2016.174.

[16] R. Pérez-Torrado and A. Querol, “Opportunistic Strains of Saccharomyces cerevisiae: A Potential Risk Sold in Food Products,” *Frontiers in Microbiology*, vol. 6, p. 1522, 2015, doi: 10.3389/fmicb.2015.01522.

[17] M. L. Wheeler *et al.*, “Immunological Consequences of Intestinal Fungal Dysbiosis,” *Cell Host Microbe*, vol. 19, no. 6, pp. 865–873, Jun. 2016, doi: 10.1016/j.chom.2016.05.003

[18] Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687. doi: 10.1371/journal.pcbi.1002687. Epub 2012 Sep 20. PMID: 23028285; PMCID: PMC3447976.