Ruprecht-Karls-Universität Heidelberg

Faculty of Engineering Sciences

Master Program Molecular Biotechnology

Unraveling the effects of smoking on human gut microbiome and associated immune system interactions

Lukas Fesenmeier

Internship protocol

Bioinformatics

16.06.2024 - 15.08.2024

performed at

Center for applied Mathematics (NOVA Math)
Department of Mathematics
NOVA School of Science and Technology (NOVA FCT)

supervision:

Dr. Marta Belchior Lopes

I herewith affirm that

- I wrote this lab protocol independently under supervision and that I did not use other sources and supporting materials than those indicated
- The adoption of quotation from the literature/internet as well as thoughts from other authors are indicated in the protocol
- I have not submitted this lab protocol for any other examination

I am aware of the fact that a false declaration will have legal consequences.

__________________ __________________

place and date signature

1 Abstract

Dysbiosis, or the alteration of microbiome composition, is frequently linked to various diseases. However, recent research highlights that lifestyle factors, such as smoking, also significantly impact the microbiome. Smoking, in particular, affects the entire body and plays a crucial role in cardiovascular diseases and cancer.

In this study, data from the Human Functional Project was utilized to assess microbiome changes associated with smoking. Additionally, a random forest model was employed to predict smoking habits based on microbiome data. Although the model's performance was suboptimal and the differences between smokers and non-smokers were subtle, the findings were consistent with previous research.

Furthermore, microbiome data was integrated with metabolic and immune data to explore the potential effects of altered microbiomes on the immune system. Three biological modules associated with smoking were identified, encompassing features across multiple omics. These findings illustrate the connection between smoking-related differentially abundant species, cholesterol levels, and various cytokines.

This study serves as a preliminary effort to develop reliable models for predicting smoking status and to identify key interactions between the gut microbiome and the immune system. Further it provides evidence of the association of Cholesterol with Smoking and the gut microbiome.

Contents

2 Introduction

The human microbiome, a vast and intricate ecosystem of microorganisms residing primarily in the gut, plays a crucial role in maintaining overall health and well-being. Comprising bacteria, viruses, fungi, and other microbes, the microbiome is involved in a multitude of bodily functions, including digestion, immune system modulation, and protection against pathogenic invaders (*Sender, Fuchs, and Milo 2016*). Recent advancements in genomic and metagenomic technologies, introducing 16SrRNA Sequencing and whole genome sequencing (WGS) have enabled researchers to delve deeper into the composition and functions of the microbiome, unveiling its significant impact on human health (*Escobar-Zepeda, Vera-Ponce de León, and Sanchez-Flores 2015*)).

Emerging evidence suggests that an imbalanced microbiome, often referred to as dysbiosis, is associated with a range of diseases (*Singh, Proctor, and Willing 2016*). For instance, specific microbial profiles have been identified in patients with IBD and obesity, highlighting the potential of microbiome-targeted therapies (*Yoo et al. 2020*). Other studies indicate a strong link indicate of colorectal cancer (*Rebersek 2021*) and cardio vascular diseases with microbiome alterations (*Fromentin et al. 2022*). Furthermore, the gut-brain axis—a bidirectional communication pathway between the gut microbiome and the brain—underscores the microbiome's influence on neurological health (*Ke et al. 2023*).

Although most studies have focused on disease, environmental and lifestyle factors have also been shown to be important modulators of the microbiome *Shima et al. 2019*. Smoking, a habit that kills around 8 million people every year (*WHO 2024*), is considered to be one of these modulators (e.g. *Prakash et al. 2021*). By exploring how smoking-induced microbiome alterations contribute to disease pathogenesis and how they impact the immune system while fighting diseases, researchers could identify novel therapeutic targets and preventive measures. This knowledge might not only enhances our understanding of the microbiome's role in health and disease but also opens avenues in understanding the tight connection between immunesystem and human microbiome, which was indicated by previous studies (*Yan et al. 2021*, *Lin and Peddada 2020*).

Therefore, microbiome data is contextualized here with potential changes in the immune system using cytokine data, measured by enzyme-linked immunosorbent assays (ELISA) and data from flow injection time-of-flight mass spectrometry (FITOFMS), measured from blood serum. This comprehensive approach not only evaluates changes in the microbiome induced by smoking, but also connects these changes using MintTea, a intermediate integration tool based on canonical correlation analysis (CCA) (*Muller, Shiryan, and Borenstein 2024*).

Additionally most studies in smoking research employ 16SrRNA, which might be more affordable and robust, but could suffer from PCR biases and generally has a poorer resolution, only to genus level (*Janda and Abbott 2007*). To address this research gap and expand current knowledge, this study uses already existing whole-genome sequencing (WGS) data to analyze the relationship between smoking and changes in the human microbiome.

3 Methods

3.1 Dataset selection and preprocessing

Data was selected by querying the database curatedMetagenomicData (*Pasolli et al. 2017*) for samples with associated smoking metadata. Several datasets were identified. Further filtering was applied. Only patients without a disease condition and without current antibiotics usage were considered, further limiting the potential studies. *Schirmer et al. 2016* was chosen as the core study for analysis because of the considerably high sample size (n=483) and the availability of immunological data and mass spectrometry data of the blood. All steps were performed using R, version 4.4 (*R Core Team 2024*). Immunological data and metabolomic data were manually downloaded from the official Human Functional Genomics Project webpage (*Y. Li et al. 2016*). The raw counts of shotgun metagenomic sequencing data were downloaded with the *returnSamples()* function from the R software curatedMetagenomicData (*Pasolli et al. 2017*). The returned data object was converted to a phyloseq object with the software mia (*Ernst et al. 2024*). The phyloseq framework was used to hold the taxonomic data *McMurdie and Holmes 2013*. Samples were filtered to only include samples with associated smoking metadata. Only taxa found in at least 10% of all the samples were retained for analysis to discard non-informative features. The raw count table was converted to relative abundances by dividing with the sequencing depth. In a final filtering step taxa with less than an average relative abundance of 0.0001% were discarded.

3.2 Statistical methods

Alpha diversity indexes were assessed with the vegan package (*Oksanen et al. 2024*). More specific richness, shannon, simson and inverse simpson diversity indexes were calculated for the class labels smoker, non smoker and former smoker. Additionally the Ratio of *Firmicutes* and *Bacteroidetes* was calculated. Class differences were quantified with the wilcoxon ranksum test from base R (*R Core Team 2024*).

Distance matrices to analyze beta diversity were calculated and visualized using buildin function of phyloseq (*McMurdie and Holmes 2013*) and vegan (*Oksanen et al. 2024*). Concretely, the euclidean, the brays-curtis, the jaccard, the weighted and unweighted unifraq distance were calculated. Distances were visualized using Principle Cordinate Analysis (PCoA) and non metric multidimensional scaling (nMDS). Permutational multivariate analysis of variance (PERMANOVA) and analysis of similarities (ANOSIM) were also done with vegan (*Oksanen et al. 2024*). Results were visualized using ggplot.

Association testing was performed with the builtin method of SIAMCAT (*Wirbel et al. 2021*). Under the hood SIAMCAT, calculates fold changes, applies wilcoxon ranksum test between the populations and corrects the false discovery rate with Benjamini Hochberg. Species are considered as differential abundant if the adjusted p-value is lower than threshold (significance level $\alpha \leq 0.05$).

Predictive Modeling was performed with SIAMCAT (*Wirbel et al. 2021*). To account for compositional nature data was previously centered log transformed. Next, data was split for a 5 fold cross validation with stratified samples to address the class imbalances. A random forest classifier was trained and model performance was assessed with the area under curve (AUC) of the receiver operator characteristic (ROC) and the Prediction Recall curve. Further informative features were identified using the median relative Gini coefficient to estimate effect size.

Metabolic data was z-score transformed and each immunological test was divided by the standard deviation for normalization purposes. Spearman Correlations were calculated between microbial abundances and metabolic and immunological features. Results were visualized using heatmaps. Data matrices were integrated using MintTea (*Muller, Shiryan, and Borenstein 2024*) in default settings. Identified MintTea Modules with sufficient predictive power were retained.

3.3 Code availability

All codes and functions used in this work are provided on github (`https://github.com/fesel2/lifestyle_biom`).

4 Results

4.1 General Composition

After filtering 125 taxa and 399 samples are left for analyis, including 60 smokers and 339 never smokers. The majority of reads belongs to the *Firmicutes* phylum. *Bacteroidetes* and *Actinobacteria* are the next most abundant phyla. Additionally some samples belonging to the phyla *Verrucomicrobia* and *Proteobacteria* were found (Figure 1). The pairwise comparison of the *Firmicutes/Bacteroidetes* ratio of the two populations never smoker and current smoker did not yield a significant difference (wilcoxon rank sum test, p-value = 0.72).

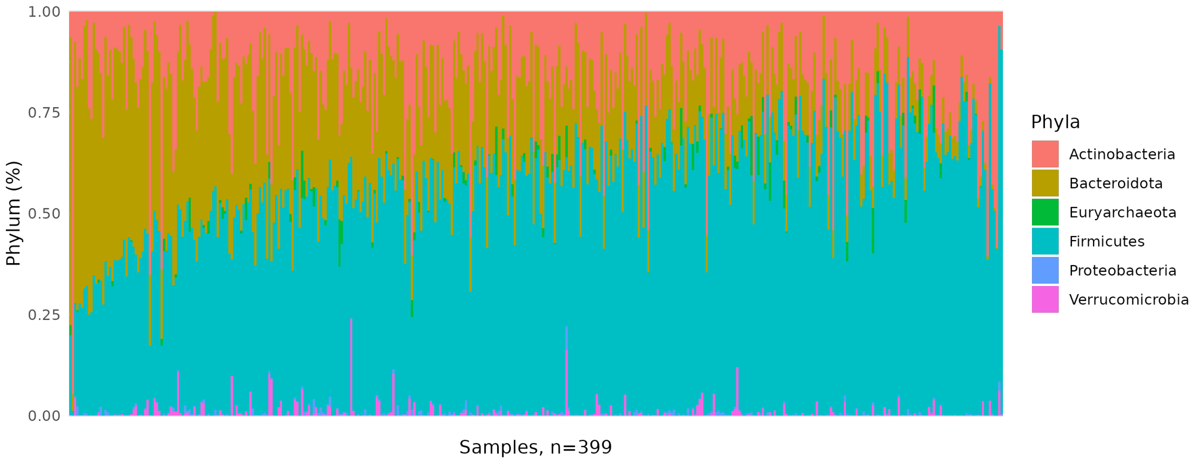

Figure 1: Stacked barplot showing the abundances of the major represented phyla after filtering. Samples are ordered by the *Firmicutes*/*Bacteroidetes* ratio from left to right. High heterogeneity is observed. *Firmicutes* are the most abundant phylum and the *Firmicutes* continuum can be confirmed.

4.2 Alpha diversity

In-sample diversity is assessed with alpha diversity indexes. Smokers generally appear ato have a lesser number of total features, their richness is significantly lower compared to never smoker (wilcoxon rank sum test, p-value = 0.0023). However, Simpson, shannon and inverse simpson do not show a significant difference between populations. Nevertheless, smokers appear to have a slightly lower shannon diversity index. This trend cannot be confirmed with the wilcoxon test (p-value = 0.25)(Figure 2).

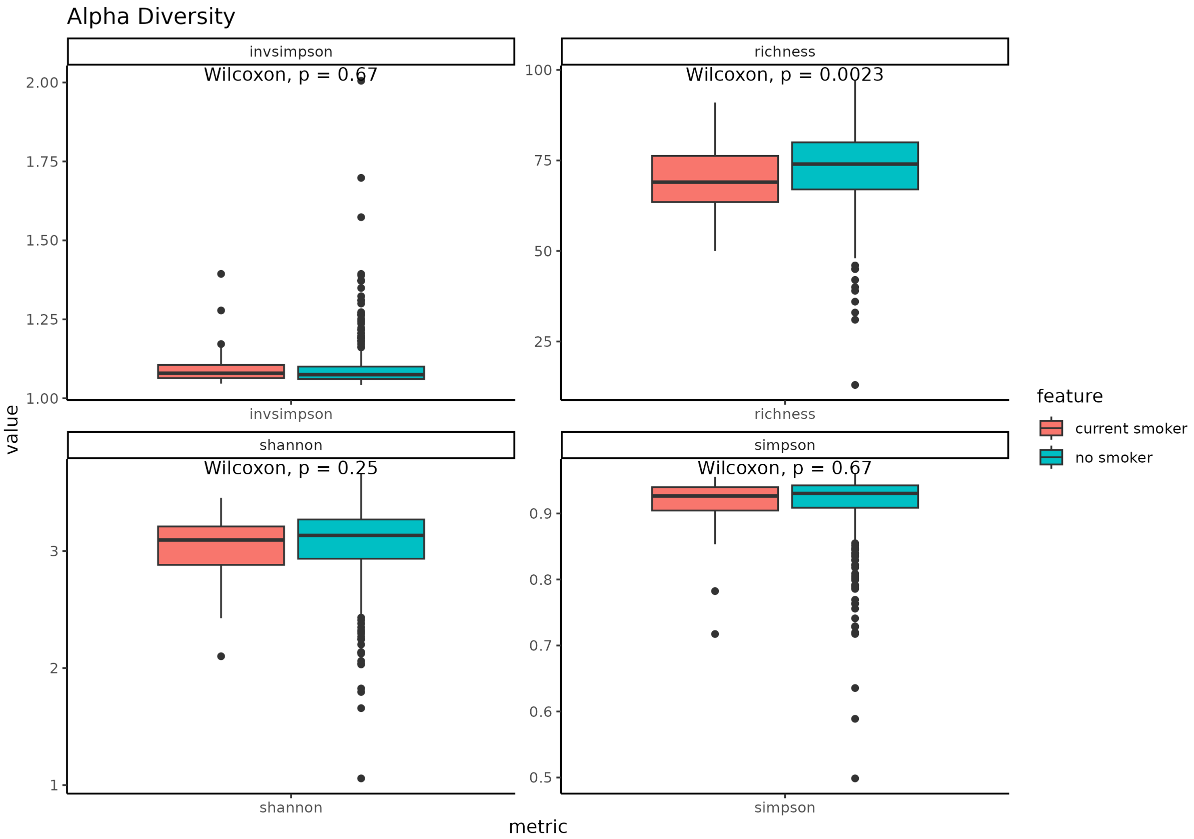

Figure 2: Alpha diversity metrics, Shannon, Simpson, inverse simpson and richness are shown for the two populations current smoker and never smoker with boxplots. Statistical significance is calculated with a pairwise wilcoxon rank sum test.

4.3 Beta diversity

Smoker population is not statistically significant from never smoking population in most of multivariate statistics. Interestingly, results differ when using different distance matrices. No statistically difference (q and p-value) was found for the euclidean, the bray, the jaccard and the weighted unifrac distance matrix. However, when applying PERMANOVA to the unweighted unifrac distance matrix significant community differences between current smokers and never-smokers are discovered, the R^2 remains low though (Table 1). When applying dimension reduction techniques (PCoA and nMDS) to the distance matrices no clear separation between populations can be observed.

Table 1: Results of multivariate statistics (PERMANOVA and ANOSIM) comparing the two populations current smoker vs never smoker

Distance matrix	R_{ANOSIM}	$q_{\text{ ANOSIM}}$	$R^2_{\text{PERMANOVA}}$	$F_{\text{PERMANOVA}}$	$p_{\text{PERMANOVA}}$
euclidean	-0.0176	0.6733	0.0026	1.0518	0.3307
bray	-0.0289	0.8008	0.0035	1.4026	0.0876
jaccard	-0.0289	0.7570	0.0029	1.1424	0.1394
uunifrac	0.0634	0.0717	0.0059	2.3377	0.0080
wunifrac	-0.0265	0.7450	0.0043	1.7269	0.0956

4.4 Differential abundant features

Six species are found to be differential abundant in smokers: *Ruminococcus gnavus*, *Clostridium spiroforme*, *Firmicutes bacterium CAG 94*, *Clostridium inocuum*, *Collinsella intestinalis* and *Collinsella aerofaciens*. One species is specifically enriched in never-smokers and depleted in smokers: *Roseburia hominis*. Foldchanges generally do not exceed a log twofold change of 1 (Figure 3).

Random forest models are trained with all of the 125 taxa abundances. The four most informative features for the model are: *Firmicutes bacterium CAG 94*, *Flavonifractor plautii*, *Clostridium spiroforme*, *Ruminoccocus gnavus*. They all have a median relative Gini coefficient ≥ 0.02. These and another 14 species with a Gini coefficient of at least ≥ 0.01 are shown in Figure 4. Three of these four species were previously identified as differential abundant with statistical tests (Figure 3). However, no clear pattern is observed in the heatmap, there are never smokers with a similar pattern as current smokers structured on the right side.

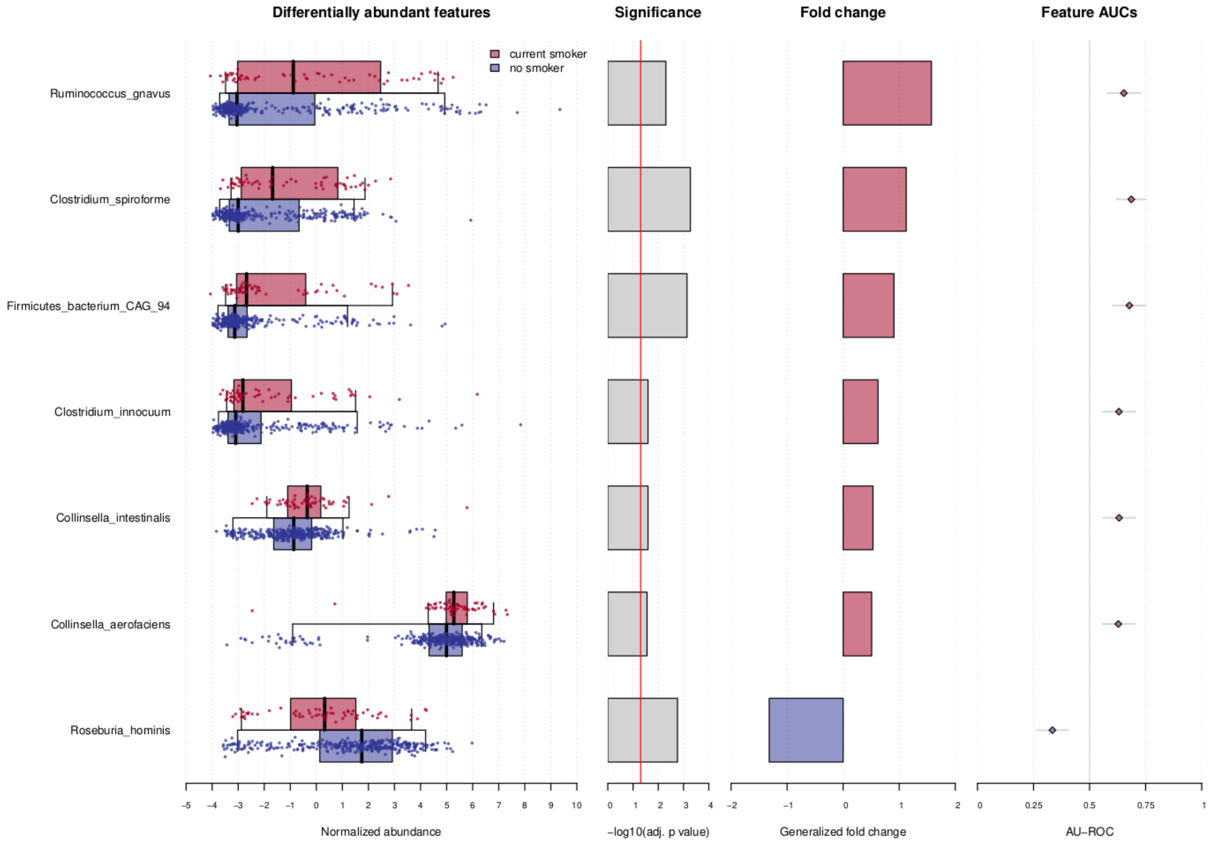

Figure 3: Results from association testing with SIAMCAT. Six species are found to be differential abundant in smokers after FDR correction. One species is found to be depleted in smokers. From left to right: Boxplots of CLR transformed abundance, p-values of wilcoxon test after BH, Foldchange and predictive power.

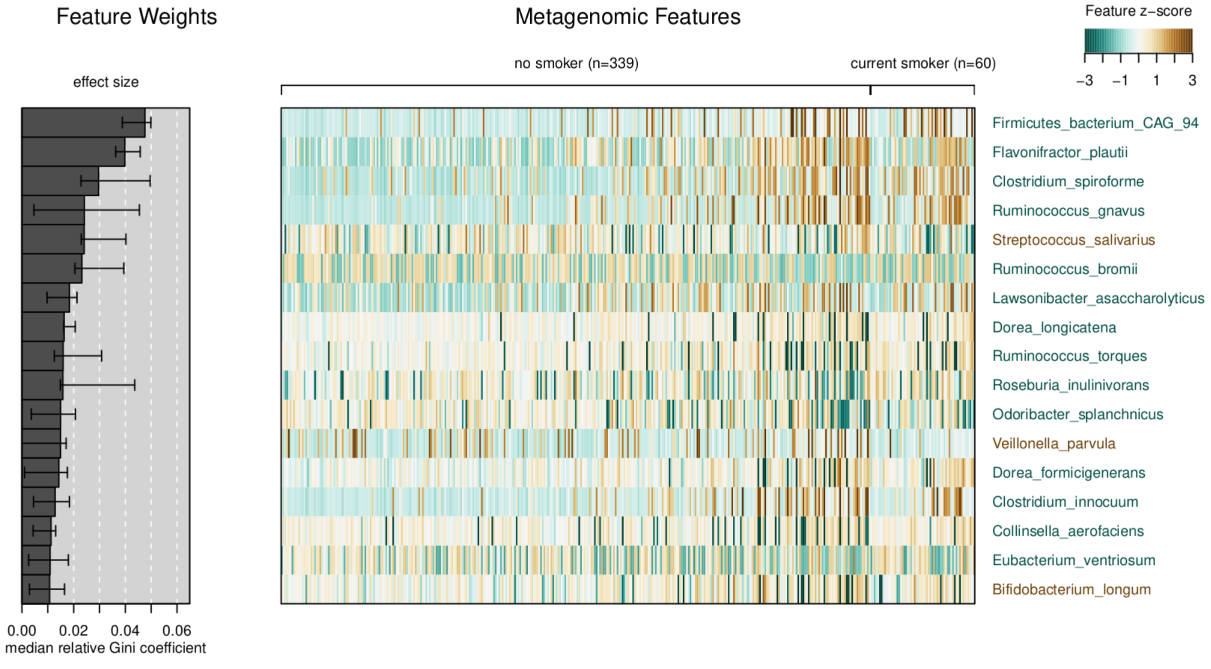

Figure 4: Feature importance of the random forest classifier. Here the 17 features with the heighest feature weight (relative Gini coefficient) are shown (≥ 0.01). Feature expression in the two populations is illustratead with a heatmap pf the z-scored relative abundances.

4.5 Classifier performance

The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) score shows a poor 0.584, suggesting that the model has small ability to distinguish between the positive and negative classes. The ROC curve showed only a slight distinction from the baseline, suggesting only small predictive power (Figure 5).

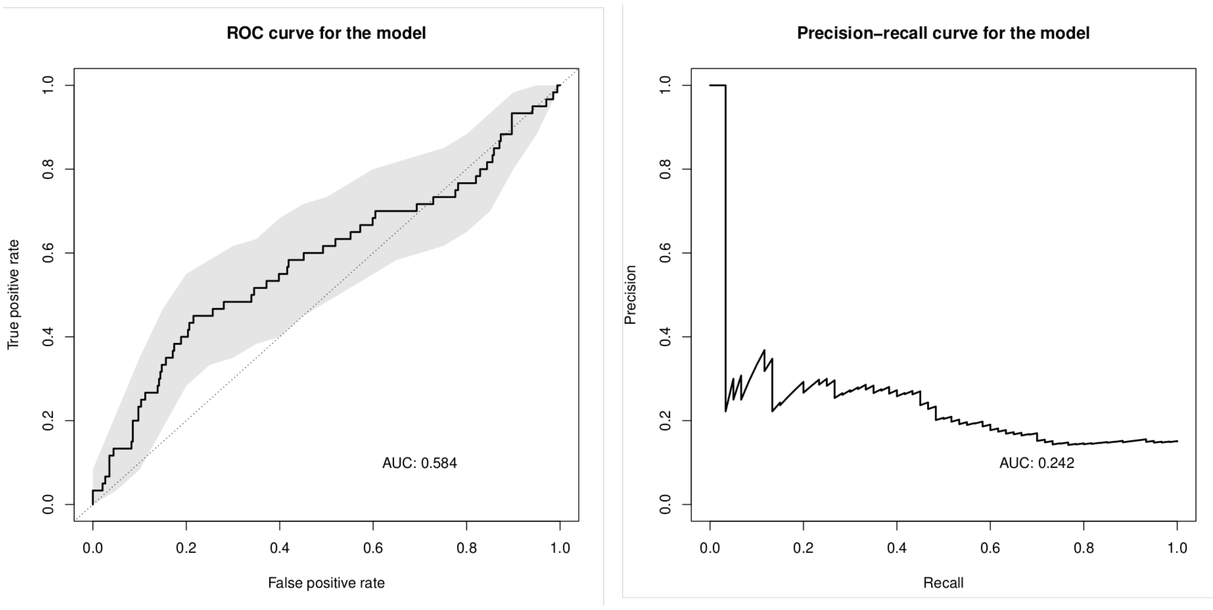

Figure 5: Receiver Operator Characteristic (ROC) and Precision Recall (PR) of the Random Forest model, trained on microbiome data. The ROC plot is the summary of all the cross validations, the grey area covering all.

To ensure the robustness of the model, k-fold cross-validation with k=5 was performed. The model consistently shows slightly higher performance than the baseline across all folds. Since the majority of samples corresponds to the class of never smokers (n=339) and smokers are only n=60, model performance is further assessed with a precision-recall curve. Considering our class balance, the baseline for the precision would be at 60/339 = 0.17, which is the accuracy/prediction when guessing all the samples as current smokers. The precision in Figure (Figure 5) is slightly above the baseline indicating only a slightly better performance than guessing the same class all over. However, in conclusion the classifier fails to predict smoking habits from microbiome data.

4.6 Relation to immune system

Similar Cytokines measurements cluster together in hierarchical clustering. Pearson correlation shows the relationship of microbiome data, more specificly the differential abundant features to Cytokine data. Highest positive correlation is found for several instances of *TNFalpha* and *Ruminococcus gravus* (approximately 0.25 across different supernatent categories), when measured in whole blood, denoted with a *a* in the figure 6. Another positive association is found for *Roseburia hominis* and several Cytokine measurements (*TNFalpha, IL1b, IL6, IL22, IL17*) measured in the supernatant of peripheral blood mononuclear cells (PBMC). On the other hand, negative correlation with *Roseburia hominis* is observed for the Cytokines from the supernatent of serum derived macrophages. Further *Clostridium spiroforme* and *Clostridium innocuum* are both positive correlatated with *IL6* measurements in PBMCs.

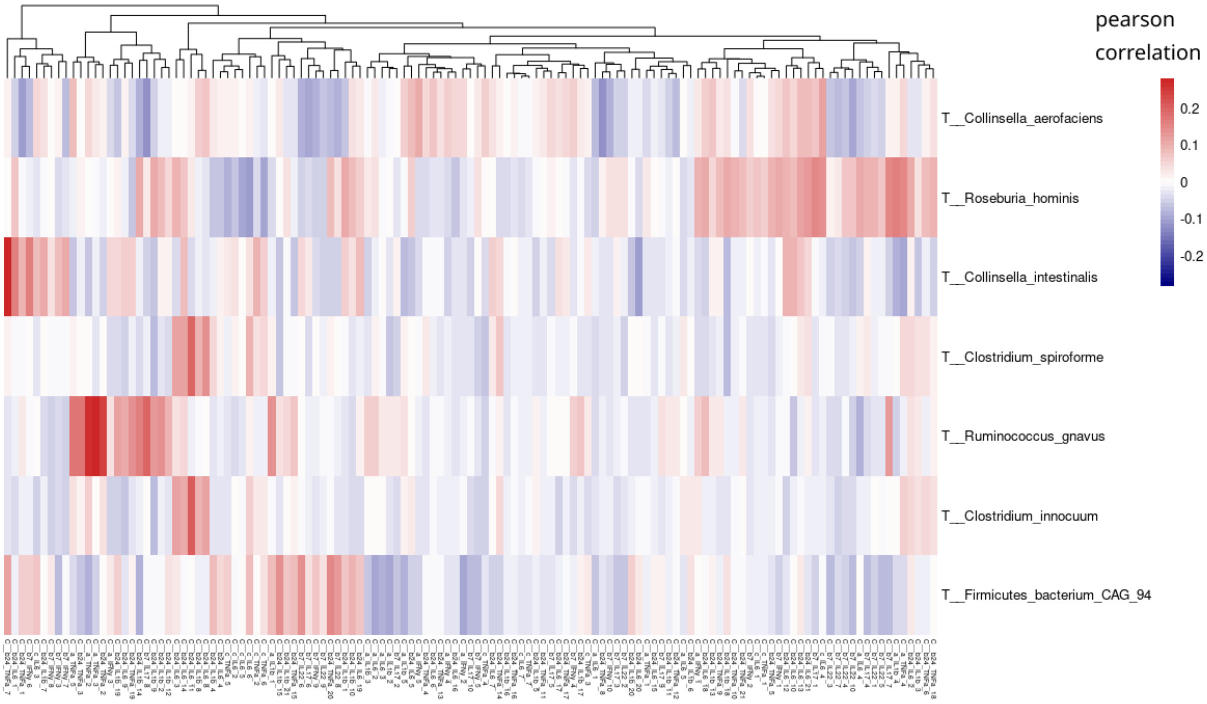

Figure 6: Correlation heatmap of the smoking associated species with the measured Cytokine concentrations. Deprogram based on hierachical clustering is used to group the immunological features. Pearson correlation was used to estimate relationships. Letter notation of immunological features: a: whole blood after 48h, b24: PBMC supernatants after 24h, b7: PBMC supernatants after 7d, c: supernatants of serum derived macrophages after 24h.

Integrating microbiome data with metabolic data and immunological data via MintTea yielded three smoking associated biological modules. Module 1 includes *Alistipes finegoldii* and the differential abundant species *Ruminococcus gnavus*. They are associated with the amino acids Leucin, Valin, Isoleucin and the free cholesterol to total lipids ratio (XL_HDL_FC_percent) in serum. Additionally *Il1b* from PBMS and *TNFalpha* from whole blood is found in connection with these features. These first principle component summarizing all of this features with a PCA has an AUC of 0.66, which is more than the random forest trained on the microbiome data alone.

Module 2 includes the species *Lawsonibacter asaccharolyticus*, *Oscillibacter sp 57 20* and *Ruminococcus torques*. Within this module Phospholipids to total lipids (XL_HDL_PL_percent), Cholesterol esters to total lipids ratio (XL_HDL_CE_percent) and Total cholesterol to total lipids ratio (XL_HDL_C_percent), were found to have a connection. TNFalpha from PBMC, IFNy from PBMC and IL1b from whole blood are found to have a connection with this module. The AUROC of this module is 0.64.

The last module, module 3, consists of immunological features only: IL1b, TNFalpha and IL6 from PBMC supernatent and has an AUROC of 0.64. The three modules are visualized as networks in figure 7.

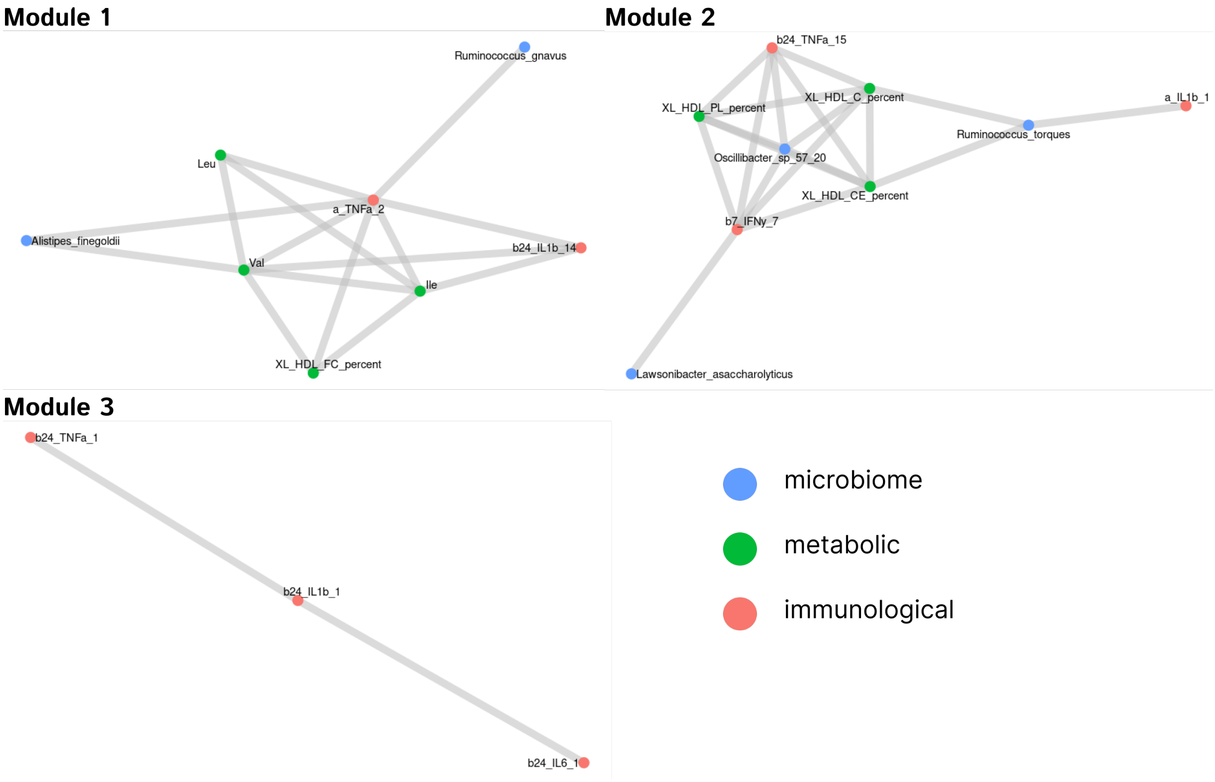

Figure 7: Biological modules associated with smoking identifed by MintTea. Three modules of different sizes and composition are discovered, consisting of microbiome, metabolic and immunological data. Nodes are connected with an edge if they cooccur in the same biological module in at least 80% of the of the CCA runs.

5 Discussion

5.1 Diversity

This study compares differences between smoker and never smokers in the gut microbiome from the Human functional genomic project (*Schirmer et al. 2016*). First, the ecological diversity in gut was analyzed for the two different groups. The shannon and the simpson diversity, both are metrics which are popular in microbiome research, do not appear to be significantly altered (Figure 2). Similar results were reported in previous studies, who did not find any decrease or increase in these metrics (*Biedermann et al. 2013*,*J. Li et al. 2023*, *Prakash et al. 2021*, *Lee et al. 2018*). However, when looking at the richness only, the total number of species in a sample, a lower number is found in smokers. Additionally, a small tendency is observed indicating that smoking might lower other diversity metrics as well. Potentially some of the toxic and carcinogen substances might be toxic to specific bacterial species, leading to the reduction or elimination of rarer microbial species. These potentially eliminated species probably do not alter the diversity of the microbiome to much, indicated by the shannon and simpson diversity metrics.

In regards of beta diversity (community analysis) results vary depending on the dissimilarity matrix used. With unweighted unifrac the smoking group is significantly different from the non-smoking group, confirmed with a PERMANOVA. However R^2 is quite low. Other dissimilarity matrices do not show any difference at all (1). Therefore, the data suggest that differences between the groups are low in multivariate analysis. Similar results were obtained by *J. Li et al. 2023*, *Zhu et al. 2024*, *Harakeh et al. 2020*, who also did not find the groups significantly altered. After putting the data from *Schirmer et al. 2016* in context with the literature, it can be concluded that smoking has none or little effect on microbiome diversity.

5.2 Classifying performance

The hypothesis that the differences between smoker and non-smoke are only minor is also backed up by the performance of the random forest model trained with SIAMCAT, which exhibited poor results on both the receiver operating characteristic (ROC) curve and the precision-recall curve. These metrics indicate that predicting smoking habits from microbiome data was not successful (Figure 5). Several factors likely contributed to these disappointing results. Firstly, the impact of smoking might be to subtle to significantly alter the gut microbiome, making it difficult for the model to detect a clear, distinct microbial signature associated with smoking.

Secondly, the high inter-sample heterogeneity of microbiome data poses a significant challenge. Individual microbiome profiles vary greatly between people, making it difficult for the model to identify consistent patterns associated with smoking. This variability can obscure any potential signals related to smoking habits. Additionally, the high sparsity of the data further complicates the problem. Many microbial species are not present in every sample, leading to incomplete and sparse datasets. This sparsity makes it challenging to build reliable machine learning classifiers, as the model struggles to discern meaningful patterns from the limited and uneven data.

Moreover, the sample size used for training the model might have been insufficient, especially the amount of smokers was way lower than the non-smokers. To overcome this class imbalance SIAMCAT automatically stratified the sample distributions, reducing the number of samples used for training. Thus the model may not have captured the full range of variability and nuances within the microbiome data. Future studies might overcome this problem by selectively choosing patients of interest.

A comparable study used several machine learning models to predict smoking habits from saliva microbiomes (*Díez López et al. 2022*). They achieved considerably higher performances with their models (about 0.7 AUC). This might indicate that smoking induced changes in saliva are considerably higher than those in gut.

5.3 Differential abundant features

Even though changes in the gut microbiome caused by smoking are of small magnitude, seven species were found to be deferentially abundant (Figure 3). *Ruminococus gnavus* is the species associated with the highest logfoldchange in the smoking population. This is in agreement with a previous study, who also identified *Ruminococus gnavus* as one of the key differences in the smoking population (*Yan et al. 2021*). They also used WGS as sequencing technique with the aim to identify differential abundant features. However *Yan et al. 2021* found a total of 94 species deferentially abundant with LEfSe, which is significantly more than the ones identifed here with SIAMCAT. Nevertheless, they might have employed a more lenient threshold for significance or used different statistical methods that are more sensitive but less specific. Techniques such as relaxed p-value cutoffs, or the application of machine learning algorithms capable of detecting subtle patterns, could also have increased the number of identified features. Also, they had a lower sample size and less class imbalance which might have affected the analysis. Interestingly, they also found *Roseburia hominis* to be enriched in non smokers, which was also confirmed by the results here, giving further indication about the biological ground truth.

In comparison with studies employing 16SrRNA Sequenging several other agreements were found. For instance, *Zhu et al. 2024* found the order of *Clostridia* enriched in smokers which agrees with the findings here of *Clostridium inocuum* and *Clostridium spiroforme*. Additionally, the *Collinsella* genera was found to be associated with smoking, confirming the findings of *Collinsella intestinalis* and *Collinsella aerofaciens* (*J. Li et al. 2023*).

Thus, the consistency with previous studies validates the results here and confirms the potential biological significance identified in this study. Despite finding fewer overall features, it is likely that only the most significant alterations were detected due to the large sample size, whereas other studies were conducted on a much smaller scale.

5.4 Relation to immune system

A weak correlation (pearson = 0.25) between *Ruminococcus gravus* and *TNFalpha* is found (Figure 6). Previous research indicates that *Ruminococcus gnavus* contains specific genes that encode superantigens to induce and bind IgA antibody in vivo, stimulating immune cells to produce corresponding antibodies, thus increasing inflammation (*Yan et al. 2021*). Interestingly, *Roseburia hominis*, found in non-smokers was reported to be responsible for butyrate and SCFA production, breaking down polysaccharides and reducing inflammation (*Chu et al. 2019*). Putting that into context, smoking might alter the human gut towards a state modulating the immune system towards higher inflammation. However it also could be vince versa, and the immune system is responsible for alterations in gut microbiome. To elucidate further work is required, potentially applying directed networks to evaluate the causal relationship of these observations.

Interestingly, MintTea associates *Ruminococcus gnavus* and *Allistipes finegoldi* with Leucin, Isoleucin and Valin with smoking (Figure 7). These amino acids all have structural similarity and the might show up due to general inflammation induced by smoking, which could alter amino acid metabolism. Inflammatory cytokines might influence the uptake and utilization of amino acids by tissues, potentially increasing their concentration in the serum. Another reasoning might be that the oxidative stress caused by smoking can affect cellular metabolism. Interestingly, the level of the high density lipoprotein cholesterol is also found within the same module. Previous studies found that smoking is connected with cholesterol levels (*Chadwick et al. 2015*). A connection between cholesterol levels and smoking could therefore be reasonable result and validate the MintTea workflow here. An interesting finding here is the connection to several microbial species, including the differential abundant *Ruminococcus gnavus*. Similar results, a negative correlation between *Ruminococcus gnavus* and HDL are also reported by *Yan et al. 2021*. The second module includes another relative of the *Ruminococcus* family and is also tightly connected to metabolites in relation with cholesterol, which further calls for the question of the connection of this metabolite. All three biological modules have a higher predictive power than the random forest trained on the microbiome data only to differentiate smokers from non-smokers. This indicates that changes to the metabolome and to the cytokines are larger and therefore allow better classification.

6 Conclusion and Outlook

This study compared the microbiome between smokers and non-smokers using existing data. However only subtle differences were found, which was not enough to extract a clear pattern and train a reliable model. Most likely, the biological differences are simply not sufficient and the patterns are far too limited to ensure a reliable classification. It is also possible that the data set was not adequate or that technical bias predominated, since this dataset was not created to answer this research question. To tackle these problems further studies, potentially specifically dedicated to the subject of smoking are required to estimate if prediction is eventually possible. However, the little to minor changes found in this study are consistent with previous research, which indicates that there are few but consistent changes in microbiome associated with smoking.

Additionally, microbiome data was complemented with immunological and metabolic information to assess its impact on the immune system, particularly through cytokines. Integrating these omics datasets successfully revealed three biological modules linking specific microbial species with metabolites and cytokines. Notably, the potential connection between serum cholesterol levels, significant microbiome alterations, and their effects on cytokines is of particular interest. Further research could explore these findings in depth to understand the underlying mechanisms driving these observations.

7 References

Biedermann, Luc et al. (2013). “Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans”. eng. In: *PloS One* 8.3, e59260. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0059260.

Chadwick, Alexandra C. et al. (July 2015). “Acrolein Impairs the Cholesterol Transport Functions of High Density Lipoproteins”. en. In: *PLOS ONE* 10.4. Publisher: Public Library of Science, e0123138. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0123138. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123138 (visited on 07/24/2024).

Chu, Jae Ryang et al. (Oct. 2019). “Prebiotic UG1601 mitigates constipation-related events in association with gut microbiota: A randomized placebo-controlled intervention study”. en. In: *World Journal of Gastroenterology* 25.40. Publisher: Baishideng Publishing Group Inc., pp. 6129–6144. DOI: 10.3748/wjg.v25.i40.6129. URL: https://www.wjgnet.com/1007-9327/full/v25/i40/6129.htm (visited on 07/24/2024).

Díez López, Celia et al. (July 2022). “Prediction of Smoking Habits From Class-Imbalanced Saliva Microbiome Data Using Data Augmentation and Machine Learning”. English. In: *Frontiers in Microbiology* 13. Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2022.886201. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.886201/full (visited on 07/21/2024).

Ernst, Felix G. M. et al. (2024). *mia: Microbiome analysis*. URL: https://github.com/microbiome/mia.

Escobar-Zepeda, Alejandra, Arturo Vera-Ponce de León, and Alejandro Sanchez-Flores (Dec. 2015). “The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics”. English. In: *Frontiers in Genetics* 6. Publisher: Frontiers. ISSN: 1664-8021. DOI: 10.3389/fgene.

2015.00348. URL: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2015.00348/full (visited on 07/16/2024).

Fromentin, Sebastien et al. (Feb. 2022). “Microbiome and metabolome features of the cardiometabolic disease spectrum”. en. In: *Nature Medicine* 28.2. Publisher: Nature Publishing Group, pp. 303–314. ISSN: 1546-170X. DOI: 10.1038/s41591-022-01688-4. URL: https://www.nature.com/articles/s41591-022-01688-4 (visited on 07/17/2024).

Harakeh, S et al. (Apr. 2020). “Impact of smoking cessation, coffee and bread consumption on the intestinal microbial composition among Saudis: A cross-sectional study”. en. In: *PloS one* 15.4. Publisher: PLoS One. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0230895. URL: https://pubmed.ncbi.nlm.nih.gov/32348307/ (visited on 07/07/2024).

Janda, J. Michael and Sharon L. Abbott (Sept. 2007). “16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls”. In: *Journal of Clinical Microbiology* 45.9. Publisher: American Society for Microbiology, pp. 2761–2764. DOI: 10.1128/jcm.01228-07. URL: https://journals.asm.org/doi/full/10.1128/jcm.01228-07 (visited on 07/17/2024).

Ke, Shanlin et al. (Nov. 2023). “Gut feelings: associations of emotions and emotion regulation with the gut microbiome in women”. en. In: *Psychological Medicine* 53.15, pp. 7151–7160. ISSN: 0033-2917, 1469-8978. DOI: 10.1017/S0033291723000612. URL: https://www.cambridge.org/core/journals/psychological-medicine/article/gut-feelings-associations-of-emotions-and-emotion-regulation-with-the-gut-microbiome-in-women/F1AA1EBBD2C4680CEC7310B6FCD95734 (visited on 07/17/2024).

Lee, Su Hwan et al. (Sept. 2018). “Association between Cigarette Smoking Status and Composition of Gut Microbiota: Population-Based Cross-Sectional Study”. eng. In: *Journal of Clinical Medicine* 7.9, p. 282. ISSN: 2077-0383. DOI: 10.3390/jcm7090282.

Li, Jingjing et al. (Oct. 2023). “Heme Metabolism Mediates the Effects of Smoking on Gut Microbiome”. In: *Nicotine & Tobacco Research*, ntad209. ISSN: 1469-994X. DOI: 10.1093/ntr/ntad209. URL: https://doi.org/10.1093/ntr/ntad209 (visited on 05/16/2024).

Li, Yang et al. (Nov. 2016). “A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans”. English. In: *Cell* 167.4. Publisher: Elsevier, 1099–1110.e14. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.10.017. URL: https://www.cell.com/cell/abstract/S0092-8674(16)31400-3 (visited on 05/03/2024).

Lin, Huang and Shyamal Das Peddada (July 2020). “Analysis of compositions of microbiomes with bias correction”. In: *Nature Communications* 11, p. 3514. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17041-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7360769/ (visited on 05/16/2024).

McMurdie, Paul J. and Susan Holmes (Apr. 2013). “phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data”. en. In: *PLOS ONE* 8.4. Publisher: Public Library of Science, e61217. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0061217. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217 (visited on 05/16/2024).

Muller, Efrat, Itamar Shiryan, and Elhanan Borenstein (Mar. 2024). “Multi-omic integration of microbiome data for identifying disease-associated modules”. en. In: *Nature Communications* 15.1. Publisher: Nature Publishing Group, p. 2621. ISSN: 2041-1723. DOI: 10.1038/s41467-024-46888-3. URL: https://www.nature.com/articles/s41467-024-46888-3 (visited on 04/16/2024).

Oksanen, Jari et al. (2024). *vegan: Community Ecology Package*. URL: https://github.com/vegandevs/vegan.

Pasolli, Edoardo et al. (Nov. 2017). “Accessible, curated metagenomic data through ExperimentHub”. en. In: *Nature Methods* 14.11. Publisher: Nature Publishing Group, pp. 1023–1024. ISSN: 1548-7105.

DOI: 10.1038/nmeth.4468. URL: https://www.nature.com/articles/nmeth.4468 (visited on 04/12/2024).

Prakash, Ajay et al. (July 2021). "Tobacco Smoking and the Fecal Microbiome in a Large, Multi-ethnic Cohort". eng. In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 30.7, pp. 1328–1335. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-20-1417.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Rebersek, Martina (Dec. 2021). "Gut microbiome and its role in colorectal cancer". en. In: *BMC Cancer* 21.1, p. 1325. ISSN: 1471-2407. DOI: 10.1186/s12885-021-09054-2. URL: https://doi.org/10.1186/s12885-021-09054-2 (visited on 07/17/2024).

Schirmer, Melanie et al. (Nov. 2016). "Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity". In: *Cell* 167.4, 1125–1136.e8. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.10.020. URL: https://www.sciencedirect.com/science/article/pii/S0092867416314039 (visited on 04/26/2024).

Sender, Ron, Shai Fuchs, and Ron Milo (Aug. 2016). "Revised Estimates for the Number of Human and Bacteria Cells in the Body". In: *PLoS Biology* 14.8, e1002533. ISSN: 1544-9173. DOI: 10.1371/journal.pbio.1002533. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4991899/ (visited on 07/16/2024).

Shima, T. et al. (Dec. 2019). "Association of life habits and fermented milk intake with stool frequency, defecatory symptoms and intestinal microbiota in healthy Japanese adults". de. In: Publisher: Brill. DOI: 10.3920/BM2019.0057. URL: https://brill.com/view/journals/bm/10/8/article-p841_2.xml (visited on 07/07/2024).

Singh, V. P., S. D. Proctor, and B. P. Willing (July 2016). "Koch's postulates, microbial dysbiosis and inflammatory bowel disease". In: *Clinical Microbiology and Infection* 22.7, pp. 594–599. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2016.04.018. URL: https://www.sciencedirect.com/science/article/pii/S1198743X1630115X (visited on 07/24/2024).

WHO (2024). *Tobacco Use*. en. URL: https://www.paho.org/en/enlace/tobacco-use (visited on 07/17/2024).

Wirbel, Jakob et al. (2021). "Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox". In: *Genome Biology*. URL: https://doi.org/10.1186/s13059-021-02306-1.

Yan, Su et al. (July 2021). "Effects of Smoking on Inflammatory Markers in a Healthy Population as Analyzed via the Gut Microbiota". English. In: *Frontiers in Cellular and Infection Microbiology* 11. Publisher: Frontiers. ISSN: 2235-2988. DOI: 10.3389/fcimb.2021.633242. URL: https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2021.633242/full (visited on 07/07/2024).

Yoo, Ji Youn et al. (Oct. 2020). "Gut Microbiota and Immune System Interactions". en. In: *Microorganisms* 8.10. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute, p. 1587. ISSN: 2076-2607. DOI: 10.3390/microorganisms8101587. URL: https://www.mdpi.com/2076-2607/8/10/1587 (visited on 07/17/2024).

Zhu, Zhouhai et al. (June 2024). "Altered interaction network in the gut microbiota of current cigarette smokers". In: *Engineering Microbiology* 4.2, p. 100138. ISSN: 2667-3703. DOI: 10.1016/j.engmic.2024.100138. URL: https://www.sciencedirect.com/science/article/pii/S2667370324000018 (visited on 07/07/2024).