Ruprecht-Karls-Universität Heidelberg

Faculty of Engineering Sciences

Master Program Molecular Biotechnology

The human gut in smoking research - Current findings and potential for future research

Lukas Fesenmeier

Internship protocol

Bioinformatics

15.04.2024 - 15.06.2024

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

This review consolidates current research on the effects of smoking on the human gut microbiome, emphasizing the complex interactions between smoking and microbial communities. By summarizing results from several related studies, it assesses the impact of smoking on the human holobiont, the integrated organism formed by the human host and its gut microbiota.

Additionally, this work provides an overview of state-of-the-art methodologies in smoking related microbiome research, covering high-throughput sequencing approaches, bioinformatics workflows, typical statistical challenges associated with microbiome data, and various tools for biomarker discovery.

Furthermore, the review highlights the potential of incorporating advanced machine learning tools, such as random forest classifiers, and intermediate data integration methods to address multi-omics questions. It provides promising findings from disease research and tries to emphasize the potential of these applications in life-style related microbiome research.

This work aims to inspire researchers in the field of smoking-related studies to adopt these innovative approaches, ultimately enhancing our understanding of how smoking affects the gut microbiome and its interplay with other systems, such as the immune system, metabolome, and brain-gut axis.

Contents

2 Introduction

In recent years, the human microbiome has gained increasing attention. Our understanding has evolved from viewing microbial species living in and on our bodies as merely parasitic or insignificant co-inhabitants, to recognizing humans as holobionts (*Postler and Ghosh 2017*. A concept which implies an ecological unit comprising a multicellular eukaryotic host closely associated with its microbiome, forming a single, integrated and complex system, going further than simply suggesting only a symbiotic relationship. Although the concept of the holobiont remains controversial, especially from an evolutionary standpoint, the significant associations between host species and their microbiomes are undeniable and a current subject of ongoing research.

Recent research (*Sender, Fuchs, and Milo 2016*) indicates that 0.2 kg of of a 70 kg is comprised of microbiome communities within the human species, accounting for about 0.4% of the total body mass. They estimate the total number of microorganisms to 3.0 $\cdot 10^{13}$, which is more than actual body cells. The microorganisms considered in this study have been found to be aiding digestion in the gut, providing protection on the skin and even having a role in internal and reproductive organs. Moreover there is growing evidence that they interact closely with the host and within their communities, providing a highly sophisticated multi-faceted interaction that goes far beyond their originally intended function (*Huttenhower et al. 2012*).

With the advent of high-throughput sequencing technologies, such as 16S-rRNA profiling and whole-genome shotgun sequencing (WGS), our ability to characterize and describe microbial communities has revolutionized (*Escobar-Zepeda, Vera-Ponce de León, and Sanchez-Flores 2015*). These tools enable researchers to quantify species abundances based on DNA presence and characterize the functionality of metagenomes. Advances in machine learning and statistical modeling further enhance our understanding, helping to identify biomarkers and elucidate complex relationships within the microbiome (*Papoutsoglou et al. 2023*). Additionally, integrating datasets from metabolomics or related data can complement microbiome research, creating a more comprehensive picture. Combining different omics technologies has proven to be a valid approach, yielding sophisticated insights, especially with metabolomics data in disease research (*Shaffer et al. 2017*).

The human gut microbiome has attracted research interest due to its critical role in energy production, enzyme synthesis, and its heavy microbial load. Gut microbes digest different dietary compounds to produce essential amino acids and finally short chain fatty acids (SCFA) like acetate. SCFAs represent the major carbon flux from the diet through the gut microbiota to the host and evidence is emerging for a regulatory role of SCFA in local, intermediary and peripheral metabolism (*Morrison and Preston 2016*). Further, gut microbiota was shown to promote host immunity by affecting antigen-presenting cells and help maintaining the mucosal barrier. More specific microbes were found to modulate and communicate with intestinal epithelial cells and dentritic cells to maintain homeostasis (*Swiatczak and Rescigno 2012*). Dysbiosis, the change of gut microbial composition and functional ability has been shown to effect host health. Moreover it is the relationship between the gut microbiata and the host immunesystem, which is assumed to have an effect on autoimmune diseases, inflammatory bowel disease (IBD), infectious diseases, diabetis or obesity (*Yoo et al. 2020*). Besides that, research results indicate a strong link of colorectal cancer (*Rebersek 2021*) and cardio vascular diseases with microbiome alterations (*Fromentin et al. 2022*). Researchers even suspect that emotions and stress are modulated by the gastrointestinal tract through the gut-brain axis, potentially explained to the extensive network of neurons in the intestines. For instance, *Ke et al. 2023* reported associations of positive and negative emotion with specific metabolic pathways and species abundances. Therefore, connecting gut microbiomes to specific phenotype or diseases is a promising way to uncover new biomarkers and provide insights into disease origins.

While most studies focus on microbiome alterations related to specific diseases, one ongoing question is how the gut microbial compositions is influenced by various lifestyle factors. For instance, it was reported that different diets strongly impact the microbiome of humans and other vertebrates. Closely related, monkeys on a low fiber/high fat diet have a strongly different microbiome compared to a group with a high fiber/low fat diet (*Y. Yan et al. 2022*). Similar observations were reported for geographic location, ethnic differences and amount of physical exercise (*Parizadeh and Arrieta 2023* and *Walker et al. 2021*).

One of these lifestyle related factors is the consumption of tobacco in form of smoking cigarettes. Despite the fact that smoking is associated with numerous health problems such as cancer, cardiovascular disease and respiratory illnesses like lung cancer, it is estimated that each year eight million people die from the tobacco use every year (*WHO 2024*). Lighting a cigarette produces 98 toxic gases including carbon monoxide, hydrogen sulfide, ammonia and nitrogen oxides (*Talhout et al. 2011*). When entering the circulatory system, they induce inflammatory changes and heavily effect oxygen transport (*R. Sun et al. 2020*). From the circulatory system they not only affect the respiratory system, but also systematically damaging all other organs. Additionally, nicotine, the substance responsible for smoking addiction, has been shown to affect the regulation of neurotransmitters, therefore playing an important role in mood, cognition and learning patterns (*Singer et al. 2004*). In conclusion, smokers expose their bodies to extremely severe conditions compared to non-smokers, which results in potential selection pressure in the intestinal tract, reshaping the human microbiome (e.g. *Shima et al. 2019*).

This work specifically reviews studies examining the effect of smoking tobacco on the human gut microbiome. Investigating the effects of smoking on the gut microbiome is important for several reasons. Understanding how smoking alters the gut microbiome can reveal additional ways through which smoking affects overall health. Additionally, it can help identifying potential biomarkers for smoking-related diseases, offering new avenues for early detection and improve understanding of the holobiont human in relationship with gut microbes. Further this review aims to provide a fundamental overview about the statistical techniques, which were previously applied to microbiome data to answer questions regarding smoking, facilitating new research and an overview which what has been done so far. Finally, it suggests techniques which were successfully applied to disease related research but might also yield insightful results when applied to smoking research.

3 Technical Challenges and Aspects

3.1 Study Design

Defining a person a smoker isn't as trivial as some might think. The amount of cigarettes consumed varies widely. There are people smoking only on weekends for example. Others quit smoking at some point of there life, spending a considerable amount of time as smoker and as non-smoker. To overcome this issue some studies define smoking as a continuous variable. Often they multiply the number of cigarettes or packs daily multiplied by years of active smoking (*Kato et al. 2010* and *Nolan-Kenney et al. 2020*). This approach is similar to calculating to the Brinkman index, often associated with lung cancer research (*Arumsari et al. 2023*). Other studies define a harsh cutoff, e.g. $\geq$ 15 cigarettes a day is a smoker for at least 25 years of smoking and create a binary problem (*Zhu et al. 2024*). To achieve better results it might be advisable to choose populations of heavy smoker and strict non-smokers. This might solve the issue but could introduce other biases like a general more healthy lifestyle and diet of non-smokers. *Stewart et al. 2018* also included a Fagerstrom test to test for nicotin dependence. Another interesting approach to quantify smoking is measuring the concentration of nicotine metabolites, for example cotinine in urine, and use it as a biomarker for tobacco smoke exposure (*Pérez-Castro et al. 2024*).

Often studies try to eliminate confounding effects during sample selection. It as been shown that antibiotics, gender and systematic diseases, e.g. IBD affect the microbiome. For that reason studies were often particularly designed to exclude these predisposed patients (*Lee et al. 2018*, *S. Yan et al. 2021*). Moreover, females are also often excluded due to lower representation in the population especially in asian studies and more complex regulatory hormonal systems (*Li et al. 2023*, *Lee et al. 2018*).

3.2 Sequencing Technologies

When considering sequencing technologies, several methods are employed, offering distinct advantages that should also be considered during study design. 16S rRNA sequencing is a robust and affordable technique, sequencing the one of the hypervariable regions of the prokaryotic ribosome with previously synthesized DNA-primers (*Escobar-Zepeda, Vera-Ponce de León, and Sanchez-Flores 2015*). Usually the V3 or V4 region is used for that purpose. This is a simple and effective technique, but could lead to PCR biases. Then DNA fragments are clustered into operational taxonomic units (OTUs) at a given similarity threshold. Fragments corresponding to the same OTU are considered from the same species or genus and count tables can be constructed by mapping the OTUs to a reference database. By sequence similarity phylogenetic trees can be constructed. In most of studies reviewed here the Quime2 pipeline is used for these steps (*Hall and Beiko 2018*). Despite the obvious advantages the OTU approach has low phylogenetic power at species level and poor discriminatory power for some specific genera, because it clusters highly similar sequences together (*Janda and Abbott 2007*).

Alternatively further differentiation can be achieved with amplicon sequence variants (ASV). Algorithms seek to incorporate sequencing error profiles in order to correct sequences to exact actual sequences. Instead of grouping similar sequences, reads are thereby resolved to the single nucleotide level, allowing precise detection of differences. However, they have a limited capability to deal with small or missing overlapping regions and cannot deal with undefined bases, which is no problem for the OTU approach (*Jeske and Gallert 2022*).

Shotgun metagenomic sequencing or WGS on the other hand overcomes the previously mentioned limitations of 16S rRNA sequencing. WGS relies on random DNA fragments from the environmental sample, which are binned to form overlapping bins and then associated with their taxa of origin (*Quince et al. 2017*). From there several software approaches are available to identify corresponding taxa from the reference database. However, mostly the MetaPhlAn2 software from the Huttenhower lab is used for species level resolution by mapping species specific markergenes to the bins (*Truong et al. 2015*). The major advantage of WGS is that it is not restricted to ribosomal RNA and allows conclusions about functional metabolomic profiles. Some studies reviewed here take advantage and also implement pathway abundances in their studies, integrating more data for their hypothesises. However, it is noteworthy that most of the smoking studies analysed here used 16S rRNA sequencing, while only a few WGS studies were performed (*S. Yan et al. 2021*). This preference might be due to the more complex computational workflow and generally higher costs associated with WGS. However, the situation is improving, as the costs for sequencing are experiencing a downward trend. As researchers agree upon software pipelines used for WGS, trends may shift more towards this technology.

Another option to analyze microbiome data is provided with YIF Scan, which uses predefined primers to quantify DNA with RT-qPCR (*Shima et al. 2019*). However, this approach is limited to the primers used in the experiment and does not allow the detection of all microbial communities in a general setting.

3.3 Downstream Analysis

After performing sequencing and running a pipeline the result are usually count-tables. A table in which all the sequencing reads associated with a specific feature (in our case microbial species) are counted for each sample. For instance, a data matrix consists of n taxa in the rows and m samples in the columns, resulting in a datamatrix with dimensions $n \times m$. For most studies, the platform of choice for data analyis was R. R not only provides tools for sophisticated multivariate analysis but also libraries specifically dedicated to microbiome research, facilitating research and achieving consensus. The extensive amount of R libraries for microbiome research is overwhelming, a good starting point might be the review, 'The best practice for microbiome analysis'(*Wen et al. 2023*), pointing out common practice for several research questions. Following these guidelines one can start with microbiome data analysis in R. Besides R, researchers also used SPSS in different versions to apply statistical tests to answer their research questions. Additionally, some reliable and widely applied tools are also implemented on the Galaxy-server, available for use for a non-coding audience (*The Galaxy Community 2024*).

3.4 General problems with microbiome data

Working with microbiome data introduces several major problems: the high inter sample heterogeneity, the sparsity of data matrices, the dimensions of the data matrices and the compositional nature of the data (*Papoutsoglou et al. 2023*). Microbiomes in human gut and their associated metagenome tend to be as individual as fingerprints, showcasing unique compositions different from patient to patient (*Schloissnig et al. 2013*). Species have been found associated with ethnicity, others with diets and diseases. Since factors influencing human life and choices are highly individual and compositions reflect everyday decisions of every individual, the identification of general trends is hampered.

This results in sparse data matrices, species abundant in one individual may not be found in another, even though life circumstances might appear similar. Species might be sub-dominant for a given sample and not appear at the depth of coverage for a given sample since sequencing depth also varies between samples (*Papoutsoglou et al. 2023*. Due to the excess of zeros within the data, statistical distributions are far from gaussian and the application of statistical tools is hampered.

Another problem is the 'curse of dimensionality', data sets usually have a high number of features (taxa) relative to the number of samples, posing significant challenges for data analysis. This can lead many statistical methods to overfit or to produce artifactual results (*Armstrong et al. 2022*).

Further, data is of compositional nature. As mentioned earlier count tables represent the absolute number of counts associated within a sample. The total counts of a sample (sequencing depth) varies between samples. Therefore absolute count numbers are either divided by the total counts to get inter sample comparable relative abundances or rarefied to a specific amount of reads per sample (*Armstrong et al. 2022*). However, by doing so all relative abundances add up to one, introducing a compositional nature of the data. Compositionality introduces issues, especially for the statistics. For instance, by assuming that one taxa is more represented in one sample or treatment condition, another taxa has to be reduced to still add up to one. Consequently, alterations in the abundance of one sequence necessitate compensatory changes in the abundance of other sequences (*Papoutsoglou et al. 2023*).

Several preprocesing techniques have been proposed to overcome and address these challenges. Studies have tried to preselect groups based on gender, ethnicity or adjusted the data with linear models and perform data analysis on the residuals. However, when looking into recent results of non adjusted data and adjusted data, results are sometimes similar and sometimes differ a lot.

A common approach so far is filtering the data during preprocessing for a fixed criteria, for example discarding all taxa which are not found in at least 10% of the samples or found to have a lower total abundance of 1e-5% in total (*Pérez-Castro et al. 2024*, *Prakash et al. 2021*). This not only helps discarding taxa found in only few individuals which are probably not informative regarding a general condition but also reduces dimensions of data. For example, studies regarding the smoking study have only a total of 30 samples but hundreds of associated taxa. Filtering out variables helps forming matrices which are better applicable to statistical tools, mitigating the curse of dimensionality.

To address the non-Gaussian distributions and compositional nature of the data, several transformation methods have been proposed. The most popular method is the Centered Log Ratio (CLR) transformation, which takes the log-ratio of counts and their geometric means. This is done within each sample based on relative abundances (*Aitchison 1982*).

4 Smoking induced differences in gut microbiome

4.1 In sample diversity (α diversity)

Biological diversity is often associated with healthiness. A healthy ecosystem is characterised by biodiversity in ecological balance. Alpha diversity metrics or in-sample diversity metrics combine richness, the amount of different species with evenness, the proportions of the individual groups, providing an effective way to quantify biological diversity in a system. A commonly used diversity metrics is the Shannon index. It measures the probability that two individuals randomly selected from a sample will belong to the same species. Usually statistical significance was assessed with an unparametric test like the Mann-Whitney or the Wilcoxon rank sum test. Most studies did not find a significant difference in Shannon index between smoker and non-smoker (*Li et al. 2023*, *Prakash et al. 2021*, *Zhu et al. 2024*, *Harakeh et al. 2020*, *R. Lin et al. 2020*, *Biedermann et al. 2013*). However *Zhu et al. 2024* and *Ishaq et al. 2017* found that diversity in smoker is reduced, but not significantly. On the other hand *Stewart et al. 2018* and *Zhang et al. 2019* claim to have found statistical significant alterations in alpha diversity, whereby the significance of *Zhang et al. 2019* barely passed the threshold. No contradictory results were found with other indexes like the Chao1, the Simpson or the Pilou diversity-indexes. To summarise, there is generally little to no evidence for a difference in in-sample diversity indices. Smokers do not have statistically less biodiversity in their gut and the hypothesis that smoking might reduce species richness remains speculative. However, there is a weak trend, which could indicate a slightly lower diversity.

4.2 Community diversity (β diversity)

The term beta diversity is used when comparing different habitats. The aim is to either reduce dimensionality for a visual representation or applying multivariate statistical tests. Traditionally Principal Component Analysis (PCA) can be used for that. PCA is used to reduce the dimensionality of data while preserving as much variance as possible. For instance *Pérez-Castro et al. 2024* were able to show that children microbiomes are significantly altered when their mother used to smoke during pregnancy. In microbiome data analysis, however, the more generalised approach of principle Coordinate Analysis (PCoA) has become established. PCoA is more flexible since it can handle any distance matrix. In contrast, PCA is limited to Euclidean distances because of its reliance on the covariance matrix. As mentioned in earlier (section 3.4), microbiome data is of compositional nature. To overcome this particular challenge the brays-curtis distance matrix or the jaccard distance metrix is usually used instead. Both methods take care of the compositional nature and are followed up by a non metric multidimenional scaling (nMDS) or a PCoA to create a visual low dimensional representation of the data (*Lee et al. 2018*, *Pérez-Castro et al. 2024*, *Harakeh et al. 2020*, *R. Lin et al. 2020*,*Nolan-Kenney et al. 2020*). In the low-dimensional representation, hardly any differences between smokers and non-smokers can be recognised at first glance; in almost all studies, all samples overlap and no obvious separation can be seen. The best separation between smokers and non-smokers is achieved by *S. Yan et al. 2021*, who are able to show the difference in community diversities with help of confidence ellipses in their PCoA-Plot. They were also able to demonstrate that the difference between the first two principle components is statistically significant.

In addition to compositional distance matrices, weighted unifrac and unweighted unifrac distance matrices are also widely applied. Unifrac is a distance metric which considers the phylogentic lineage of the taxas. Both methods rely on a phylogenetic tree which needs to be constructed previously. Technically it identifies the branch lengths unique to a community and compares the sum of them to the total branch length of the tree. The weighted version of unifrac additionally includes the relative abundances of the species to calculate distances. Unifrac distance matrices were calculated by *Nolan-Kenney et al. 2020*, *Pérez-Castro et al. 2024*, *Biedermann et al. 2013* and *Zhu et al. 2024*. However, visualisation results after nMDS or PCoA were not much different compared to brays-curtis distance matrices. It is notable that *Nolan-Kenney et al. 2020* also combined results from unifrac weighted, unweighted and brays-curtis into a omnibus-test.

Although smokers and non-smokers are shown in a low-dimensional representation, distance matrices were also used to quantify the differences between the communities using multivariate statistical tests. *Stewart et al. 2018* and *Lee et al. 2018* were able to show that microbiome of smokers is significantly altered when comparing to never-smokers using a permutational multivariate analysis of variance (PERMANOVA). Similar results were achieved with a MANOVA from (*Pérez-Castro et al. 2024*. However results could not be confirmed by *Zhu et al. 2024*, who also used PERMANOVA and an additional analysis of similarities, which is an unparametric analogue of PERMANOVA based on a ranked dissimilarity values. With this in mind, we can suggest that microbiomes differ between smokers and non-smokers regarding community diversity. However, it should not be ignored that not all studies can confirm this assumption.

4.3 Taxonomic Differences

Microbes are classified and structured based on their ribosomal RNA in a taxonomic ranks. Starting from the kingdom of bacteria, microbial species are classified as phylum, class, order, family, genus to species. Depending on the research question, read counts of a specific level are added together and compared between the study conditions. The most basic approach is to simply compare different study groups after normalisation with a statistical test. This can be done by using a conventional t-test under the assumption that abundances are normally distributed or a non-parametric Wilcoxon test. Another option is to use software designed for differential gene expression analysis like edgeR (*Robinson, McCarthy, and Smyth 2010*) or DESeq2 (*Love, Huber, and Anders 2014*). In recent years however more sophisticated tools have been developed specifically for microbiome data. Three tools frequently used and also applied to microbiome data to answer the question how smoking alters the microbiome are Linear discriminant analysis Effect Size (LEfSe), Analysis of Composition of Microbiomes (ANCOM) and

Microbiome Multivariable Associations with Linear Models (MaAsLin). LEfSe determines features most likely to explain the differences between classes by integrating statistical tests with effect sizes from Linear Discriminant Analysis (*Segata et al. 2011*). Additionally LEfSe visualizes differential abundant features across all taxonomic orders using a cladogram and the option to evaluate LDA effect scores with a barchart (Figure 1). ANCOM on the other hand is specifically designed to handle compositional data matrices by pairwise log-ration testing (*Mandal et al. 2015*). For multivariable associations MaAsLin was developed taking confounds and complex experiment conditions into account (*Mallick et al. 2021*). Essentially all three of these tools are designed to identify biological markers at all taxonomic levels.

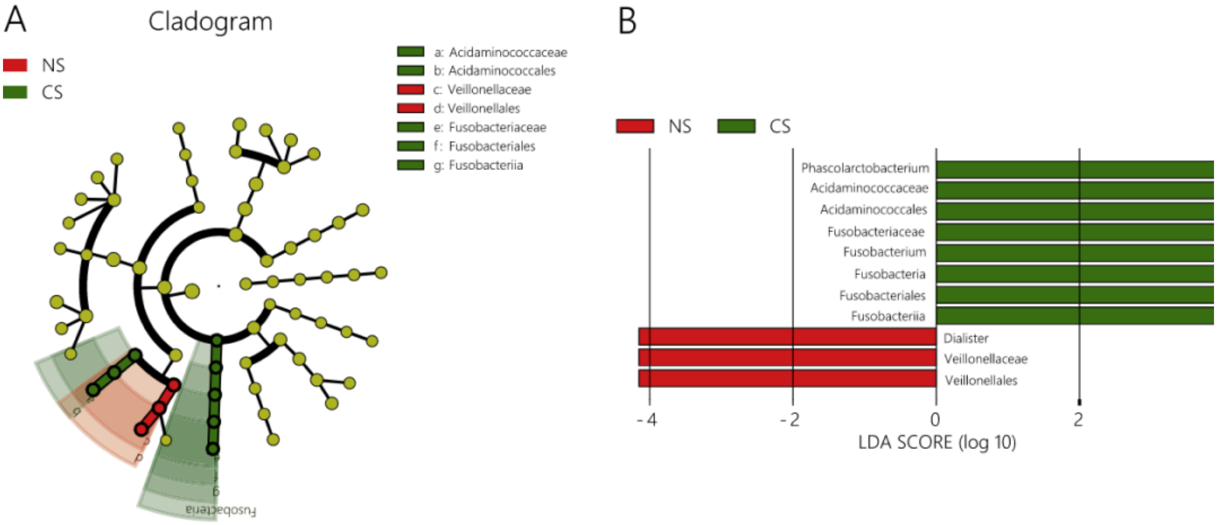

Figure 1: Exemplary LEfSe Output results; A: Cladogram of differential abundant features. The further inwards, the higher the taxonomic order. Enriched features are highlighted in green, depleted features are highlighted in red. B: LDA Effect scores for differential abundant features are visualized with a barchart (NS: non-smoker, CS: current smoker)(*Zhu et al. 2024*)

S. Yan et al. 2021, one of the few studies who obtained WGS data, found a total of 94 differentially abundant species with LEfSe. Other studies using 16S rRNA Sequencing found less features. Notably the results from *S. Yan et al. 2021* show a strong enrichment of *Rumminococcus gnavus* and *Bacteroides vulgatus* in the smoking population. Across most studies the genus of *Bacteroides* was found to be significantly enriched in smokers including *Li et al. 2023*, *R. Lin et al. 2020*, *Lee et al. 2018* and non significantly increased in *Ishaq et al. 2017*. Further *Kato et al. 2009* and *H. Lin and Peddada 2020* both show a positive linear relationship between packyears of smoking and the *Bacteroides* abundance, indicating a strong connection.

The higher abundance of *Bacteroides* also suggests a shift in balance between the two phyla *Firmicutes* and *Bacteroidetes*. This ratio in particular is interesting because it was previously suggested as a potential biomarker for obesity and other diseases (*Ahmad et al. 2023*).*Lee et al. 2018*, *Stewart et al. 2018* and *H. Lin and Peddada 2020* report that this ratio is significantly affected by smoking, indicating general health impact. It is unclear however if smokers generally have a more unhealthy lifestyle or if this ration is direct affected by the smoking. Interestingly *Lee et al. 2018* report that the organisms in the intestines of smoker and non-smoker are similar back at family level even though they are distinct at phylum level.

Similarly the finding of *Ruminococcus gnavus* by *S. Yan et al. 2021* was confirmed by *R. Lin et al. 2020*. They hypothesize that the increased abundance of *Ruminococcus gnavus* has an inflammatory effect because this prokaryote produces specific antigens and stimulates immune cells to produce corresponding antibodies, boosted by metabolits of tobacco consumption *S. Yan et al. 2021*. Other deferentially abundant genera, identified by *S. Yan et al. 2021* include *Prevotella*, which was also found by *Prakash et al. 2021* and *Stewart et al. 2018*, *Lachnospiracea*, which was reported by *Prakash et al. 2021* and *Lee et al. 2018* and *Clostridiales*, found by *Zhu et al. 2024* and *Li et al. 2023*.

On the other hand the genus *Dialister* was found to be enriched in non-smokers (*Harakeh et al. 2020*, *Zhu et al. 2024*). Also *Tenericutes* were found depleted in smokers and therefore to be associated with non-smoking by *Prakash et al. 2021* and *Harakeh et al. 2020*.

Even if the studies use different sequencing technologies or apply different statistical tools or preprocessing transformations, similarities in results are found across studies, indicating biological truth. Nevertheless, it must be kept in mind that of about 10 studies on the topic, often only 2-3 have similar findings. This may be due to more profound underlying confounds like diet, race or individual genetics or technical biases and challenges.

5 Cessation and former smoking

5.1 Immediate result of Cessation

Quitting smoking is often seen as a big milestone for a person. Not only does the subject overcome their physical and psychological addiction, but their physical health already improves within a very short time. It has been shown that health-promoting effects occur after a short time. Just one day after quitting smoking, high blood pressure and the risk of a heart attack begin to fall. After one year, the risk of a heart attack or coronary heart disease has already halved compared to a long-term smoker. After 15 years, the risk approaches that of a person who has never smoked (Timeline after quitting smoking *2018*).

It has been shown that this fundamental change also has an effect on the gut microbiome. *Biedermann et al. 2013* report that pylogenetic diversity metrics has increased after 4 weeks cessation. This trend was confirmed by their samples taken after 8 weeks, also showing an increasing trend. With PCA on weighted unifrac distances they were further able to show a clear separation of the intervention group compared to the continuous smoking group. The highest distance values regarding inter sample diversity were determined for subjects undergoing smoking cessation between the 4 weeks sample and the sample before smoke cessation, indicating a microboial shift in the gut (*Biedermann et al. 2013*). They found an increase of *Firmicutes* and *Actinobacteria* and an simultaneously decrease of *Proteobacteria* and *Bacteroidetes*, providing additional evidence that *Bacteroides*, the main contribution to the *Bacteroidetes* phylum are highly associated with pack-years of smoking (*Kato et al. 2009*).

5.2 Long term effects

Even though an increased alpha diversity after smoking cessation was reported, there cannot be made conclusions about longterm effects because of the limited observation period (*Biedermann et al. 2013*). Especially when considering section 4.1, it remains questionable if smoking or smoke cessation has actual impact on in-sample diversity. However differences between samples were confirmed. *Prakash et al. 2021* report a significant difference between never smoker and current smokers, but no differences between former smoker and never smoker. They hypothesise that the gut microbiome shifts back towards the non-smoking state after quitting smoking. They provide evidence for this hypothesis by organising former smokers into groups based on years of abstinence from smoking. By assessing the normalised differences of several genera between the individual groups, they were able to show that the normalised difference to never-smokers decreases over the years (Figure 2)(*Prakash et al. 2021*).

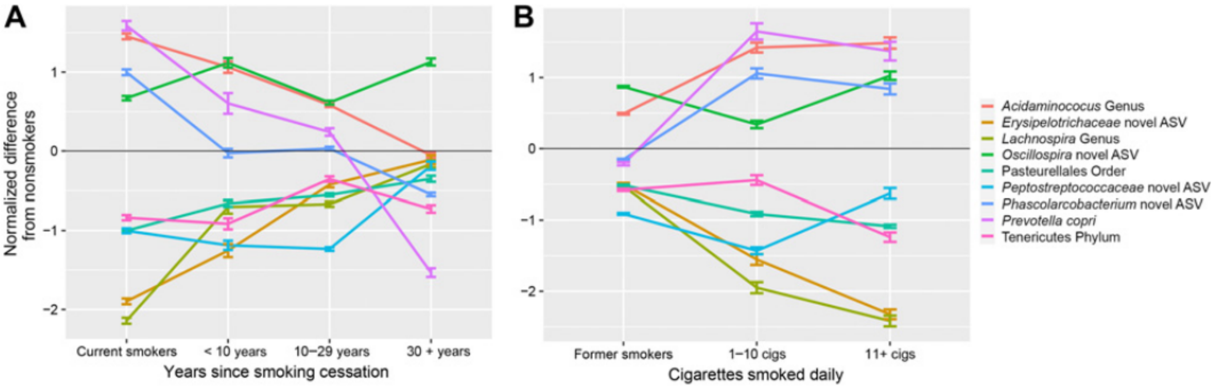

Figure 2: Lineplots demonstrating the effects of former smoking over the years. A: Development of chosen species over the years after smoking cessation, binned into age groups. A trend can be observed, The more time passes, the smaller the distance becomes. B: Smoking more than 10 cigarettes daily introduces more differences compared to former smokers. (source: *Prakash et al. 2021*)

6 Machine Learning in microbiome research

6.1 Classification

In recent years machine learning has revolutionized the field of microbiome research, enabling scientists to uncover complex patterns and make predictions that were previously unattainable. Machine learning has always been linked to sequence classification: tools like QIIME (*Hall and Beiko 2018*) and MetaPhlAn (*Truong et al. 2015*) use supervised learning algorithms to classify microbial sequences into taxonomic categories. These models are trained on reference databases and can identify the presence and abundance of microbial taxa in a sample. Additionally machine learning models can identify patterns associated with diseases such as IBD, diabetes, and cancer. These models can predict disease risk based on microbiome composition, offering potential for early diagnosis and personalized medicine. In recent years several supervised methods have been proposed and used in microbiome research including Logistic regression, Linear Discriminant Analyis, k-nearest neighbors, support vector machines and naive bayes classifiers (*Marcos-Zambrano et al. 2021*). Also deep learning approaches has been made, but data matrices usually do not include enough samples for good performance. Hence, the performance of these models is often limited by the fact that data matrices typically do not include a sufficient number of samples.

Unfortunately, due to the curse of dimensionality and the unknown patterns in the data, one cannot provide specific guidance on choosing a predictive modeling algorithm based on the number of features. Moreover, the so-called No Free Lunch Theorem in machine learning states that there is no single “best” method that can universally excel in solving all types of problems. The selection of an appropriate algorithm needs to consider the specific characteristics and constraints of the task at hand. Nevertheless, a combination of feature selection and a suitable performance estimation protocol can enhance a classifier’s performance in a high-dimensional setting (*Papoutsoglou et al. 2023*).

6.2 Random Forests

However in practice ensemble learning methods have been shown to have a reliable and good performance. Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. One of the most popular methods is the random forest. What has greatly contributed to the popularity of forests is the fact that they can be applied to a wide range of prediction problems and have few parameters to tune. The method is not only easy to use, but also widely recognised for its accuracy and ability to handle small sample sizes and high-dimensional feature spaces (*Biau and Scornet 2016*). Therefore, they are perfectly suited for microbiome data, which has been demonstrated using a benchmark dataset where they show reliably high sensitivity and specificity (*Papoutsoglou et al. 2023*) and outperform other models (*Wang, F. Sun, and Luan 2024*).

At the very bottom end random forest models rely on multiple decision trees. A decision tree is a tree-like model of decisions and their possible consequences. Each internal node represents a ”test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (in classification) or a continuous value (in regression).

Random forests utilize bootstrap aggregation (bagging), where multiple datasets are generated by random sampling with replacement from the original dataset. Each of these datasets is used to train a different decision tree. This helps in reducing variance and avoiding overfitting. In addition to bagging, random forests introduce another layer of randomness by selecting a random subset of features for each tree. This further decorrelates the trees, which enhances the overall predictive performance of the forest.

During the training phase, random samples are drawn with replacement from the training set (bootstrap samples). For each bootstrap sample, a decision tree is built. During the construction, a random subset of features is chosen at each split to determine the best split. Each tree is grown to the largest extent possible without pruning. In the prediction phase, each tree in the forest makes a prediction when a new input is presented. The predictions from all the trees are aggregated. For classification, this is done by majority voting, and for regression, by averaging the outputs.

Random forests offer several advantages. By combining multiple decision trees, random forests usually achieve higher accuracy and robustness compared to individual trees. The process of random sampling and feature selection helps in reducing overfitting. Random forests can handle a large number of input features without the risk of dimensionality. They also provide a way to estimate the importance of different features in the prediction process. However, there are also disadvantages. The model is more complex compared to a single decision tree, making it less interpretable. Building multiple trees and maintaining them can be resource-intensive in terms of both time and memory. Since each new input has to be passed through many trees, prediction can be slower compared to simpler models.

Random forest models have been used to diagnose colorectal cancer (CRC) based on fecal shotgun sequencing data (*Ai et al. 2019*). Other studies use the interpretability of their classification model to identify biomarkers (*Gupta et al. 2019*). Their have been applied in several microbiome scenarios, not limited to the gut, for instance it has also been used to analyze the effects of cohabitants on human skin (*Ross, Doxey, and Neufeld 2017*). An extensive review about applications of random forests is avaiable (*Marcos-Zambrano et al. 2021*).

6.3 Normalization

Before training a model it is advisable to normalize data. In differential analysis, the main objective of normalization among different datasets is to remove or mitigate spurious associations between microbes and diseases. On the other hand, the main objective of normalization for phenotype prediction is to increase prediction accuracy, robustness, reliability and generalizability of the trained model to the unseen testing data (*Wang, F. Sun, and Luan 2024*). Therefore the aim of normalization is to minimize technical biases, but to retain biological variance. Several methods have been proposed to do so. Generally normalization methods can be classified into Scaling and Transformation methods. The basic idea of Scaling is to divide the count table by a scaling factor to remove biases. The most basic one is to divide by the total number of reads (relative abundance). Another popular option is the trimmed mean of M-values (TMM), implemented in edgeR (*Robinson, McCarthy, and Smyth 2010*). Transformation methods apply a logarithmic or count based transformation to account for the skewed nature of microbiome data. Centered log ratio is done on top of total sum scaling and accounts for the compositional nature of the data.

6.4 SIAMCAT - a versitile tool to implemement machine learning

SIAMCAT, which stands for Statistical Inference of Associations between Microbial Communities And host phenoTypes, is a comprehensive tool designed to facilitate the statistical analysis and modeling of microbiome data (*Wirbel et al. 2021*). It is particularly effective in associating microbial community compositions with host phenotypes, such as diseases or other biological conditions. One of the primary features of SIAMCAT is its ability to perform feature selection, allowing users to identify the most relevant microbial taxa that are informative for distinguishing between different host phenotypes.

The structure of the SIAMCAT class is designed to integrate seamlessly with the popular R package phyloseq (*McMurdie and S. Holmes 2013*), which is commonly used for microbiome analysis. SIAMCAT builds upon the phyloseq object, leveraging its data structures to manage and manipulate microbiome data efficiently. The SIAMCAT class encapsulates various components of the analysis workflow, including data preprocessing, which was discussed in section 6.3), model training, and evaluation, within a single, cohesive framework. This design allows users to transition smoothly from data import and initial exploration in phyloseq to more advanced statistical modeling and machine learning tasks in SIAMCAT. By maintaining compatibility with phyloseq, SIAMCAT ensures that users can take advantage of the extensive ecosystem of tools and functions available in phyloseq, while also extending their capabilities to include sophisticated association analyses and predictive modeling. The overall structure of the SIAMCAT class is illustrated in figure 3.

Further it supports several various machine learning algorithms, including regularized logistic regression, support vector machines, and random forests (section 6.2), enabling robust modeling of the relationships between microbiome composition and host phenotypes. SIAMCAT includes rigorous cross-validation procedures to prevent overfitting and ensure that the models are generalizable. Performance evaluation metrics, such as receiver operating characteristic (ROC) curves and precision-recall curves, are also provided to assess model accuracy.

To handle common issues in microbiome data, such as varying sequencing depths and compositionality, SIAMCAT offers options for data normalization and transformation. The tool also includes visualization capabilities, enabling users to create feature importance plots, heatmaps, and ordination plots that help in interpreting the associations between microbial features and host phenotypes.

Additionally, SIAMCAT emphasizes reproducibility by allowing users to generate detailed analysis reports and documentation that can be easily shared and reviewed. Its flexibility and comprehensive set of features make SIAMCAT a powerful tool for researchers working with microbiome data to uncover meaningful biological insights and associations.

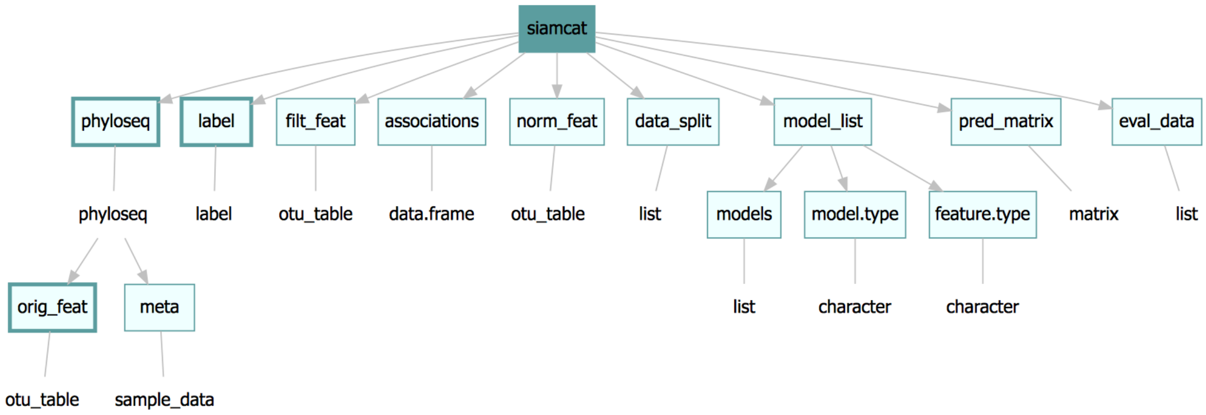

Figure 3: The SIAMCAT object is built on top of the phyloseq object. It includes all phyloseq functionality and adds functionality useful for machine learning. It is possible to train multiple models on different parameters within the same siamcat object (source: *Wirbel et al. 2021*).

6.5 Machine learning in smoking research

Even though unsupervised methods such as PCoA and methods for biomarker identification such as LEfSe are established as a standard in smoking research, there are hardly any studies that have attempted predictive modelling. This hypothesis was confirmed here with a quick Google Scholar search using the keywords *microbiome*, *machine learning*, *classification* and *smoking*. Only one study was found to predict smoking habits from microbiome data in saliva microbiome (*Díez López et al. 2022*). They evaluate seven machine learning methods (logistic regression, k-nearest neighbour, support vector machines with linear kernel, support vector machines with radial kernel, decision trees, random forests and extreme gradient boosting). Additionally they apply data-augmentation techniques to improve classifier performance. Data augmentation techniques do not necessarly improve the AUC which can be seen in figure 4. In that study random forests and support vector machines with linear kernel show the best performances. However the best performing models have AUCs around 0.7.

7 Integration with other datatypes

7.1 Multi-omics in disease research

The growing recognition of the gut microbiome's importance and complexity has led to a surge in multi-omic microbiome studies. These studies apply several molecular assays to the same set of samples, aiming to capture multiple layers of information regarding the microbiome's involvement in disease *Muller, Shiryan, and Borenstein 2024*. A common objective of these multi-omic microbiome analyses is to identify disease-associated markers—specific features from various omics (e.g., certain species, pathways, or metabolites) whose measured abundances are strongly associated with the disease in question *Qin et al. 2012*. For example, *Yachida et al. 2019* combine metagenomic and metabolomic analyses to reveal biomarkers at distinct stage-specific phenotypes in colorectal cancer. Unfortunately, while such multi-omic data offer an exciting opportunity to study the microbiome and its role in human

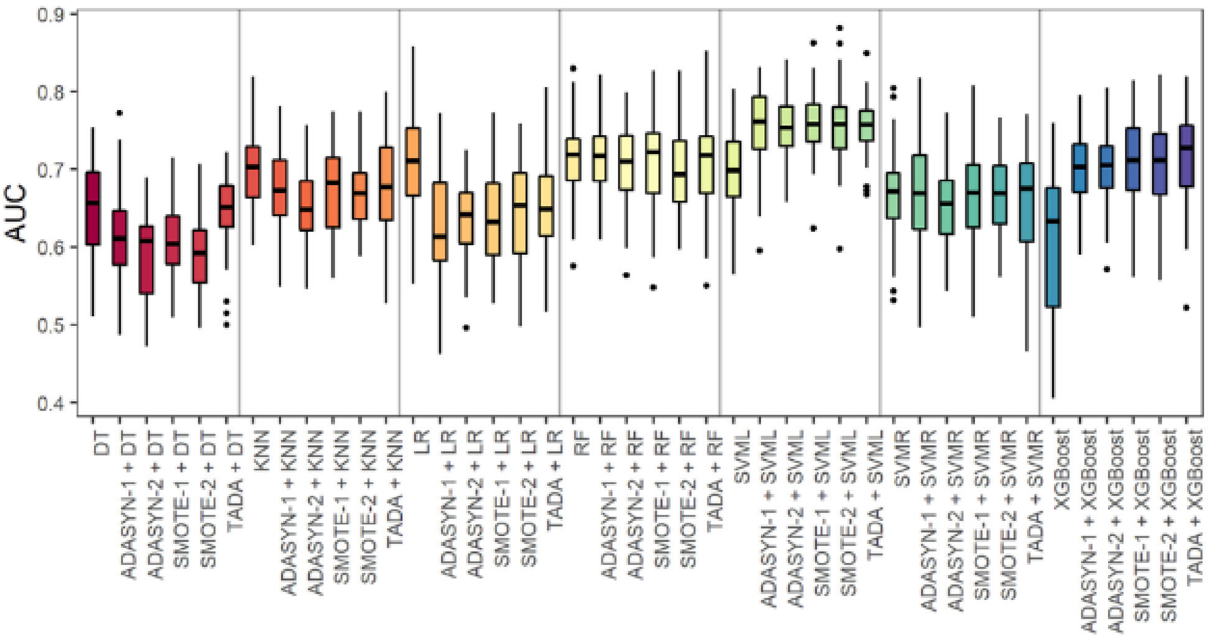

Figure 4: Performances of machine learning classifiers to predict smoking habits from saliva microbiome data. Performance metric here is the area under curve metric (AUC). Most models achieve performances of around 0.7, however they are not much improved by data augmentation techniques.(source: *Díez López et al. 2022*)

health, rigorous integrative analysis of such data remains highly challenging, as does using such data to gain a systems-level understanding of the microbiome *Chong and Xia 2017*.

Several preliminary attempts directed at addressing this challenge and considering both cross-omic dependencies and associations with a disease state have been introduced. For example, some microbiome studies took a two-step approach by first identifying features that are associated with the disease, and then clustering these features based on pairwise correlations between features *Pedersen et al. 2018*. This approach, however, may fail to identify features that are not sufficiently informative by themselves but can be incorporated into larger modules that, as a whole, are strongly predictive of the disease *Muller, Shiryan, and Borenstein 2024*.

To address these challenges several methods have been proposed, including canonical correlation analysis (CCA), Co-Inertia Analyis, partial least squares and more (*Sankaran and S. P. Holmes 2019*). A relatively recent approach based on CCA is MintTea (Multi-omic INTegration Tool for microbiomE Analysis), which relies on CCA. CCA in general, an intermediate integration method that receives two feature tables, and outputs a linear transformation per table so that the resulting latent variables are maximally correlated (*Witten and Tibshirani 2009*). Additionally to make analysis more robust the CCA is repeated several times on subsets of the data and putative modules are extracted. Next features from the putative modules are put into context with a co-occurrence network. Features found in the same putative module are further extracted to disease associated modules (Figure 5). MintTea has achieved remarkable performances and was able to reliable extract biological valuable information from a total of nine studies (*Muller, Shiryan, and Borenstein 2024*).

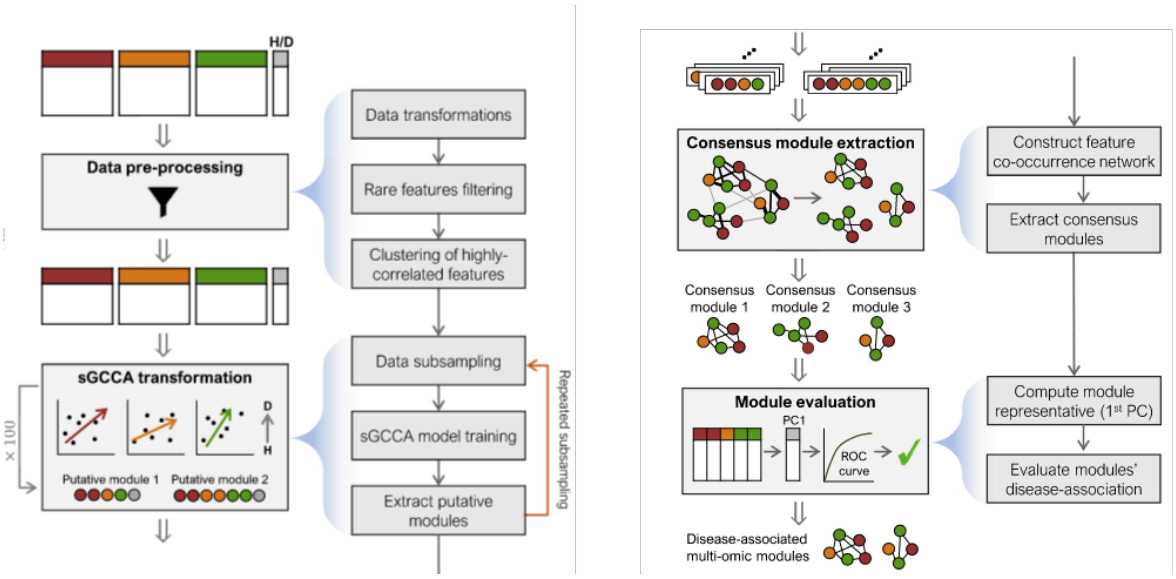

Figure 5: The general MintTea workflow. After dataspecific preprocessing 100 sGCCA transformations on randomly selected 80% of the data are run. Features found to have a relationship to the label of interest are extracted into putative moduls. Next a consensus network is constructed from features which are found in similar putative moduls. Dimensions of Consensus modules are further reduced using PCA and the predictive power of the first principle component is evaluated (source: *Muller, Shiryan, and Borenstein 2024*).

7.2 Integrating smoker's microbiome with other data

So far microbiomes in association to smoking have been connected to Magnetic resonance imaging (*Curtis et al. 2019*). The authors postulate that Bacteria contributing to the supply of essential aminoacids and neurotransmitters have an impact on Resting State functional connectivity. They report a significant cluster for the connectivity between the left anterior insula and the left operculum associated with *Bacteroides*. For *Prevotella* they found a significant cluster for the connectivity between the left anterior insula and the right occipital cortex. Both features have been reported to be deferentially abundant in smokers (see section 4.3), affect these brain areas.

Li et al. 2023 analyzed the relationship between host transcriptome data with the microbiome. They performed bidirectional mediation analysis and found that the p-values is are more lower indicating that genes act as mediators and shape the microbiome. They also found that in transcriptomic data immune-related pathways are mainly affected by smoking, for example, immune cell proliferation and differentiation, immune-related signaling pathways, and cytokine production, emphasizing an effect of tobacco smoking on the immune system.

Another study combines biochemistral measurements from the blood with microbiome data. More specific they look at the cell counts of different blood cells, metabolites and tumor markers in blood. They report significant differences in these clinical characteristics in high-density lipoproteins (HDL) and carcinoembryonic antigen (CEA). They report a negative correlation between *Ruminococcus gnavus* and HDL.

When using WGS as sequencing technology, instead of having only sequences according to the hypervariable regions of the ribosome, read contigs cover the entire metagenome allowing functional profiling. For instance *S. Yan et al. 2021* report that several pathways, whose functions included nucleoside synthesis, biosynthesis, degradation of carbohydrates, degradation of amino acids and nucleotides, and generation of precursor metabolites and energy, were individually enriched in smokers. Interestingly, nine of the distinct pathways that were enriched in smokers were responsible for amino acid biosynthesis, including L-serine, glycine, L-isoleucine, L-valine, L-methionine, L-lysine, L-threonine and aspartate. They hypothesize is that accumulation of amino acids, the basic structures of proteins, may be due to the increased demand for amino acids caused by changes in the bacterial abundances due to smoking, thereby increasing the amino acid abundances. However when using 16SrRNA Sequencing data there is also the option to infer pathway abundances from the phylogenetic taxa information using a reference database *Douglas et al. 2020*. By doing so *Prakash et al. 2021* they found relative enrichment of oxidative and degradative metabolism superpathway members, such as lactose and galactose degradation or formaldehyde oxidation, and depletion of biosynthesis superpathway members, such as peptidoglycan biosynthesis (Fig. 4A). Pathways outside of these superpathway classes, like glycolysis and reductive acetyl coenzyme A pathway, were also increased in current smokers relative to former/never smokers. Additionally they correlated the pathways with there deferentially abundant species and concluded which taxa is responsible for the observation.

So far data in smoking research integration of human gut microbiome with other data mostly relies on pairwise correlation analysis. However more sophisticated methods were addressed in section 7.1 potentially uncovering new insights across multiple omics and organs throughout the human body.

8 Conclusions and final remarks for further research

In this report differences between smokers and non-smokers across numerous studies have been reviewed and findings were summarized. Although no substantial differences where found for alpha diversity indexes (most often the shannon index was used), researchers disagree regarding the population wide differences. Multivariate statistical tools like PERMANOVA are often able to elucidate class differences but not consistent across all studies. Even when statistical differences are found, populations are often not separable in lower dimensional representations like PCoA or nMDS, indicating only small differences between the smoking and non-smoking populations. This fact is also confirmed by small R^2 values in PERMANOVA or small R values in ANOSIM.

When analyzing the differential abundant features across all taxonomic orders similar features are found throughout multiple studies, indicating biological truth. Especially the species *Bacteroides* and *Ruminococcus gnavus* seem to be consistenly differentially abundant. In addition several genera (*Dialister, Tenericutes, Prevotella, Clostridiales*) have been shown to be different across multiple studies, statistical tests and sequencing platforms.

In disease research for colorectal cancer and IBD, advanced machine learning classifiers and data integration tools are routinely applied to microbiome data, yielding insightful results about interactions with other systems or organs. In contrast, smoking research has largely overlooked this potential so far, with few studies contextualizing their findings alongside other datasets.

There is significant potential in applying sophisticated tools, such as Random Forest classifiers or tools for intermediate data integration (like MintTea), to uncover interactions between the microbiome, the host's metabolic system, and the immune system in the context of smoking. However, further studies are required to fully realize the potential.

9 References

Ahmad, Manal Ali et al. (Aug. 2023). “Association of the gut microbiota with clinical variables in obese and lean Emirati subjects”. English. In: *Frontiers in Microbiology* 14. Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2023.1182460. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1182460/full (visited on 07/15/2024).

Ai, Dongmei et al. (Feb. 2019). “Using Decision Tree Aggregation with Random Forest Model to Identify Gut Microbes Associated with Colorectal Cancer”. eng. In: *Genes* 10.2, p. 112. ISSN: 2073-4425. DOI: 10.3390/genes10020112.

Aitchison, J. (1982). “The Statistical Analysis of Compositional Data”. en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1982.tb01195.x, pp. 139–160. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1982.tb01195.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1982.tb01195.x (visited on 07/17/2024).

Armstrong, George et al. (Feb. 2022). “Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data”. English. In: *Frontiers in Bioinformatics* 2. Publisher: Frontiers. ISSN: 2673-7647. DOI: 10.3389/fbinf.2022.821861. URL: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.821861/full (visited on 07/17/2024).

Arumsari, Dessy et al. (Apr. 2023). “The description of smoking degree based on brinkman index in patients with lung cancer”. en. In: 7.3. Number: 3 Publisher: Universitas Airlangga. ISSN: 2541-092X. URL: https://e-journal.unair.ac.id/JBE/article/view/12184 (visited on 07/17/2024).

Biau, Gérard and Erwan Scornet (June 2016). “A random forest guided tour”. en. In: *TEST* 25.2, pp. 197–227. ISSN: 1863-8260. DOI: 10.1007/s11749-016-0481-7. URL: https://doi.org/10.1007/s11749-016-0481-7 (visited on 07/21/2024).

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.

Chong, Jasmine and Jianguo Xia (Nov. 2017). “Computational Approaches for Integrative Analysis of the Metabolome and Microbiome”. In: *Metabolites* 7.4, p. 62. ISSN: 2218-1989. DOI: 10.3390/metabo7040062. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5746742/ (visited on 07/25/2024).

Curtis, K et al. (Aug. 2019). “Insular resting state functional connectivity is associated with gut microbiota diversity”. en. In: *The European journal of neuroscience* 50.3. Publisher: Eur J Neurosci. ISSN: 1460-9568. DOI: 10.1111/ejn.14305. URL: https://pubmed.ncbi.nlm.nih.gov/30554441/ (visited on 07/07/2024).

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).

Douglas, Gavin M. et al. (June 2020). “PICRUSt2 for prediction of metagenome functions”. en. In: *Nature Biotechnology* 38.6. Publisher: Nature Publishing Group, pp. 685–688. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0548-6. URL: https://www.nature.com/articles/s41587-020-0548-6 (visited on 07/26/2024).

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).

Gupta, Ankit et al. (Nov. 2019). "Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India". In: *mSystems* 4.6. Publisher: American Society for Microbiology, 10.1128/msystems.00438–19. DOI: 10.1128/msystems.00438-19. URL: https://journals.asm.org/doi/10.1128/msystems.00438-19 (visited on 07/21/2024).

Hall, Michael and Robert G. Beiko (2018). "16S rRNA Gene Analysis with QIIME2". en. In: *Microbiome Analysis: Methods and Protocols*. Ed. by Robert G. Beiko, Will Hsiao, and John Parkinson. New York, NY: Springer, pp. 113–129. ISBN: 978-1-4939-8728-3. DOI: 10.1007/978-1-4939-8728-3_8. URL: https://doi.org/10.1007/978-1-4939-8728-3_8 (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).

Huttenhower, Curtis et al. (June 2012). "Structure, function and diversity of the healthy human microbiome". en. In: *Nature* 486.7402. Publisher: Nature Publishing Group, pp. 207–214. ISSN: 1476-4687. DOI: 10.1038/nature11234. URL: https://www.nature.com/articles/nature11234 (visited on 07/16/2024).

Ishaq, Hafiz Muhammad et al. (2017). "Molecular Characterization Of Fecal Microbiota Of Healthy Chinese Tobacco Smoker Subjects In Shaanxi Province, Xi'an China". eng. In: *Journal of Ayub Medical College, Abbottabad: JAMC* 29.1, pp. 3–7. ISSN: 1025-9589.

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).

Jeske, Jan Torsten and Claudia Gallert (Apr. 2022). "Microbiome Analysis via OTU and ASV-Based Pipelines—A Comparative Interpretation of Ecological Data in WWTP Systems". en. In: *Bioengineering* 9.4. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 146. ISSN: 2306-5354. DOI: 10.3390/bioengineering9040146. URL: https://www.mdpi.com/2306-5354/9/4/146 (visited on 07/17/2024).

Kato, Ikuko et al. (Dec. 2009). "Smoking and other personal characteristics as potential predictors for fecal bacteria populations in humans". en. In: *Medical Science Monitor* 16.1. Publisher: International Scientific Information, Inc., CR1–CR7. ISSN: 1234-1010, 1643-3750. URL: https:///abstract/index/idArt/878298 (visited on 07/07/2024).

– (Jan. 2010). "Smoking and other personal characteristics as potential predictors for fecal bacteria populations in humans". eng. In: *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 16.1, CR1–7. ISSN: 1643-3750.

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).

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).

Lin, Renbin et al. (Nov. 2020). “The effects of cigarettes and alcohol on intestinal microbiota in healthy men”. en. In: *Journal of Microbiology* 58.11, pp. 926–937. ISSN: 1976-3794. DOI: 10.1007/s12275-020-0006-7. URL: https://doi.org/10.1007/s12275-020-0006-7 (visited on 07/07/2024).

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. eng. In: *Genome Biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

Mallick, Himel et al. (Nov. 2021). “Multivariable association discovery in population-scale meta-omics studies”. en. In: *PLOS Computational Biology* 17.11. Publisher: Public Library of Science, e1009442. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009442. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009442 (visited on 07/15/2024).

Mandal, Siddhartha et al. (2015). “Analysis of composition of microbiomes: a novel method for studying microbial composition”. eng. In: *Microbial Ecology in Health and Disease* 26, p. 27663. ISSN: 0891-060X. DOI: 10.3402/mehd.v26.27663.

Marcos-Zambrano, Laura Judith et al. (Feb. 2021). “Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment”. English. In: *Frontiers in Microbiology* 12. Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2021.634511. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.634511/full (visited on 03/28/2024).

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).

Morrison, Douglas J. and Tom Preston (May 2016). “Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism”. In: *Gut Microbes* 7.3. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19490976.2015.1134082, pp. 189–200. ISSN: 1949-0976. DOI: 10.1080/19490976.2015.1134082. URL: https://doi.org/10.1080/19490976.2015.1134082 (visited on 07/17/2024).

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).

Nolan-Kenney, Rachel et al. (July 2020). “The Association Between Smoking and Gut Microbiome in Bangladesh”. eng. In: *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco* 22.8, pp. 1339–1346. ISSN: 1469-994X. DOI: 10.1093/ntr/ntz220.

Papoutsoglou, Georgios et al. (Sept. 2023). “Machine learning approaches in microbiome research: challenges and best practices”. English. In: *Frontiers in Microbiology* 14. Publisher: Frontiers. ISSN:

1664-302X. DOI: 10.3389/fmicb.2023.1261889. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1261889/full (visited on 03/26/2024).

Parizadeh, Mona and Marie-Claire Arrieta (Oct. 2023). "The global human gut microbiome: genes, lifestyles, and diet". English. In: *Trends in Molecular Medicine* 29.10. Publisher: Elsevier, pp. 789–801. ISSN: 1471-4914, 1471-499X. DOI: 10.1016/j.molmed.2023.07.002. URL: https://www.cell.com/trends/molecular-medicine/abstract/S1471-4914(23)00152-1 (visited on 07/17/2024).

Pedersen, Helle Krogh et al. (Dec. 2018). "A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links". eng. In: *Nature Protocols* 13.12, pp. 2781–2800. ISSN: 1750-2799. DOI: 10.1038/s41596-018-0064-z.

Pérez-Castro, Sonia et al. (Jan. 2024). "Influence of perinatal and childhood exposure to tobacco and mercury in children's gut microbiota". English. In: *Frontiers in Microbiology* 14. Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2023.1258988. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1258988/full (visited on 05/16/2024).

Postler, Thomas Siegmund and Sankar Ghosh (July 2017). "Understanding the Holobiont: How Microbial Metabolites Affect Human Health and Shape the Immune System". English. In: *Cell Metabolism* 26.1. Publisher: Elsevier, pp. 110–130. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2017.05.008. URL: https://www.cell.com/cell-metabolism/abstract/S1550-4131(17)30296-6 (visited on 07/16/2024).

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.

Qin, Junjie et al. (Oct. 2012). "A metagenome-wide association study of gut microbiota in type 2 diabetes". en. In: *Nature* 490.7418. Publisher: Nature Publishing Group, pp. 55–60. ISSN: 1476-4687. DOI: 10.1038/nature11450. URL: https://www.nature.com/articles/nature11450 (visited on 07/25/2024).

Quince, Christopher et al. (Sept. 2017). "Shotgun metagenomics, from sampling to analysis". en. In: *Nature Biotechnology* 35.9. Publisher: Nature Publishing Group, pp. 833–844. ISSN: 1546-1696. DOI: 10.1038/nbt.3935. URL: https://www.nature.com/articles/nbt.3935 (visited on 03/26/2024).

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).

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth (Jan. 2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/ (visited on 07/16/2024).

Ross, Ashley A., Andrew C. Doxey, and Josh D. Neufeld (July 2017). "The Skin Microbiome of Cohabiting Couples". In: *mSystems* 2.4. Publisher: American Society for Microbiology, 10.1128/msystems.00043–17. DOI: 10.1128/msystems.00043-17. URL: https://journals.asm.org/doi/10.1128/msystems.00043-17 (visited on 07/21/2024).

Sankaran, Kris and Susan P. Holmes (Aug. 2019). "Multitable Methods for Microbiome Data Integration". English. In: *Frontiers in Genetics* 10. Publisher: Frontiers. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00627. URL: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00627/full (visited on 04/18/2024).

Schloissnig, Siegfried et al. (Jan. 2013). "Genomic variation landscape of the human gut microbiome". en. In: *Nature* 493.7430. Publisher: Nature Publishing Group, pp. 45–50. ISSN: 1476-4687. DOI: 10.1038/nature11711. URL: https://www.nature.com/articles/nature11711 (visited on 07/17/2024).

Segata, Nicola et al. (June 2011). “Metagenomic biomarker discovery and explanation”. eng. In: *Genome Biology* 12.6, R60. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-6-r60.

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).

Shaffer, Michael et al. (Nov. 2017). “Microbiome and metabolome data integration provides insight into health and disease”. In: *Translational Research*. Metabolomics, Epigenomics, RNASEQ, and Novel Tissue Imaging Techniques: Translational Aspects 189, pp. 51–64. ISSN: 1931-5244. DOI: 10.1016/j.trsl.2017.07.001. URL: https://www.sciencedirect.com/science/article/pii/S1931524417302323 (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).

Singer, S. et al. (Sept. 2004). “Nicotine-induced changes in neurotransmitter levels in brain areas associated with cognitive function”. eng. In: *Neurochemical Research* 29.9, pp. 1779–1792. ISSN: 0364-3190. DOI: 10.1023/b:nere.0000035814.45494.15.

Stewart, Christopher J. et al. (2018). “Effects of tobacco smoke and electronic cigarette vapor exposure on the oral and gut microbiota in humans: a pilot study”. en. In: *PeerJ* 6. Publisher: PeerJ, Inc. DOI: 10.7717/peerj.4693. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5933315/ (visited on 07/07/2024).

Sun, Rongli et al. (Feb. 2020). “Benzene exposure induces gut microbiota dysbiosis and metabolic disorder in mice”. eng. In: *The Science of the Total Environment* 705, p. 135879. ISSN: 1879-1026. DOI: 10.1016/j.scitotenv.2019.135879.

Swiatczak, Bartlomiej and Maria Rescigno (Feb. 2012). “How the interplay between antigen presenting cells and microbiota tunes host immune responses in the gut”. In: *Seminars in Immunology*. Intestinal Microbiota: shaping the immune system ? 24.1, pp. 43–49. ISSN: 1044-5323. DOI: 10.1016/j.smim.2011.11.004. URL: https://www.sciencedirect.com/science/article/pii/S1044532311001497 (visited on 07/17/2024).

Talhout, Reinskje et al. (Feb. 2011). “Hazardous Compounds in Tobacco Smoke”. en. In: *International Journal of Environmental Research and Public Health* 8.2. Number: 2 Publisher: Molecular Diversity Preservation International, pp. 613–628. ISSN: 1660-4601. DOI: 10.3390/ijerph8020613. URL: https://www.mdpi.com/1660-4601/8/2/613 (visited on 07/17/2024).

The Galaxy Community (July 2024). “The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update”. In: *Nucleic Acids Research* 52.W1, W83–W94. ISSN: 0305-1048. DOI: 10.1093/nar/gkae410. URL: https://doi.org/10.1093/nar/gkae410 (visited on 07/17/2024).

Timeline after quitting smoking (Nov. 2018). en. URL: https://www.medicalnewstoday.com/articles/317956 (visited on 07/18/2024).

Truong, Duy Tin et al. (Oct. 2015). “MetaPhlAn2 for enhanced metagenomic taxonomic profiling”. en. In: *Nature Methods* 12.10. Publisher: Nature Publishing Group, pp. 902–903. ISSN: 1548-7105. DOI: 10.1038/nmeth.3589. URL: https://www.nature.com/articles/nmeth.3589 (visited on 07/17/2024).

Walker, Rebecca L. et al. (Dec. 2021). “Population study of the gut microbiome: associations with diet, lifestyle, and cardiometabolic disease”. en. In: *Genome Medicine* 13.1, p. 188. ISSN: 1756-994X. DOI: 10.1186/s13073-021-01007-5. URL: https://doi.org/10.1186/s13073-021-01007-5 (visited on 07/17/2024).

Wang, Beibei, Fengzhu Sun, and Yihui Luan (Mar. 2024). “Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity”. en. In: *Scientific Reports* 14.1. Publisher: Nature Publishing Group, p. 7024. ISSN: 2045-2322. DOI: 10.1038/s41598-024-57670-2. URL: https://www.nature.com/articles/s41598-024-57670-2 (visited on 04/16/2024).

Wen, Tao et al. (Oct. 2023). “The best practice for microbiome analysis using R”. In: *Protein & Cell* 14.10, pp. 713–725. ISSN: 1674-800X. DOI: 10.1093/procel/pwad024. URL: https://doi.org/10.1093/procel/pwad024 (visited on 07/17/2024).

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.

Witten, Daniela M. and Robert J. Tibshirani (June 2009). “Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data”. en. In: *Statistical Applications in Genetics and Molecular Biology* 8.1. Publisher: De Gruyter. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1470. URL: https://www.degruyter.com/document/doi/10.2202/1544-6115.1470/html (visited on 07/25/2024).

Yachida, Shinichi et al. (June 2019). “Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer”. eng. In: *Nature Medicine* 25.6, pp. 968–976. ISSN: 1546-170X. DOI: 10.1038/s41591-019-0458-7.

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).

Yan, Yueyang et al. (Jan. 2022). “Metagenomic and network analysis revealed wide distribution of antibiotic resistance genes in monkey gut microbiota”. In: *Microbiological Research* 254, p. 126895. ISSN: 0944-5013. DOI: 10.1016/j.micres.2021.126895. URL: https://www.sciencedirect.com/science/article/pii/S0944501321002019 (visited on 04/04/2024).

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).

Zhang, W et al. (Feb. 2019). “Gut microbiota community characteristics and disease-related microorganism pattern in a population of healthy Chinese people”. en. In: *Scientific reports* 9.1. Publisher: Sci Rep. ISSN: 2045-2322. DOI: 10.1038/s41598-018-36318-y. URL: https://pubmed.ncbi.nlm.nih.gov/30733472/ (visited on 07/07/2024).

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).