Dottorati Innovativi con caratterizzazione industriale - XXXIV ciclo

POR Campania FSE 2014-2020

Oral plaque microbiome resilience under external dietary stimuli

Renato Giliberti

This research project was conducted during my doctoral studies in the “Food science” PhD program at the Department of Agricultural Sciences, University of Naples Federico II, headed by prof. Amalia Barone and before by prof. Danilo Ercolini. I performed my activity in the metagenomics group at the same department under the supervision of prof. Edoardo Pasolli. The data collection process was conducted during the 20 months spent in the PreBiomics S.r.l. company (Trento, Italy) under the supervision of Doc. Mattia Bolzan (CTO). I also spent 3 months abroad at EMBL Heidelberg (Germany) working under the supervision of Prof. Georg Zeller.

In the last twenty years, the study of human microbiome composition and its relationship with host health has involved the scientific community with growing interest. Current knowledge allows us to predict host phenotype based on the microbiome composition. However, most of the studies have been focused on gut microbiome characterization, while other biological niches have been less characterized.

Of particular interest is the oral microbiome, widely recognized as resilient and as responsible for major oral diseases.

In the last twenty years, there has been a massive spreading in dental implant use as a substitution to lost or heavily diseased teeth. The increasing use of dental implants has gone hand in hand with the spreading of implants-related diseases like mucositis and peri-implantitis. It has been demonstrated that implant diseases have, at least in part, a bacterial origin. Last studies have demonstrated how we can predict disease state starting from oral microbiome composition. Oral microbiome characterization has been made mainly using 16S sequencing techniques, by limiting the analysis to genus-level taxonomic resolution and to small cohorts. External factors that may have an influence on microbiome composition and expression have been considered only marginally.

Dietary habits influence on oral microbiome composition has been defined marginally and with obsolete techniques. The objective of this research project is to fulfil the lack in terms of dietary habits influence on oral microbiome using the latest sequencing technologies and using statistical, computational and machine learning approaches.

The study was conducted on a cohort composed of 451 subjects with dental implants, acquired thanks to the involvement of 48 dental clinics on the whole italian territory. The cohort included healthy (156), peri-implant (152) and mucositis (143) subjects. A large set of clinical variables along with the information about dietary habits were collected. The dietary information was collected through the administration of a food frequency questionnaire (draft by EPIC). A subgingival plaque sample was collected and sequenced with shotgun techniques for 121 (healthy) subjects. The study was conducted with the following objectives: 1) evaluation of the relationship between dietary habits and health status and between dietary habits and clinical variables

(chapter 3); 2) oral microbiome characterization and evaluation of its relationship with dietary habits (chapter 4); 3) accomplishment of objectives 1) and 2) with the use of statistical and machine learning methods optimized in chapter 2.

In chapter 2, it was conducted an analysis voted to: 1) evaluating the predictive power of presence/absence profiles in contrast to more canonical relative abundance profiles; 2) evaluating and comparing different classification algorithms to identify the best one in terms of classification accuracy. The analysis was conducted on 4128 shotgun samples from 25 publicly available dataset, and extended to 4026 16S samples from 30 publicly available datasets. Relative abundances profiles generated through MetaPhlAn3 were converted in presence/absence profiles by considering different thresholds and different taxonomic resolution levels. Five different classification algorithms (i.e., Random Forest; Lasso; ENet; SVM; LSVM) were compared and their accuracies were evaluated in terms of AUC.

Results showed that classification made on presence/absence profiles has comparable results with the ones obtained on relative abundances. The Random forest algorithm resulted as the less sensible to data transformation and as the classifier able to maximize classification performances. Results were confirmed on the considered 16S datasets.

In chapter 3 an analysis voted to the identification of dietary habits influence on health status was conducted. The analysis was done during the Beta Program at PreBiomics where I spent 18 months of my research activity.

Multiple analyses voted to identify the relationship between dietary habits and disease state were made: ordination analysis; correlations coefficient computation; linear discriminant analysis; logistic regression; Random forest classification. Results highlighted weak correlations between dietary habits and implant disease state. The use of dietary habits data did not improve predictive power performances.

In absence of strong correlations between dietary habits and peri-implant health status, possible associations between dietary habits and oral microbiome composition were studied. This was limited to healthy subjects. Multiple statistical significant correlations were found between dietary habits and microbial species, and more specifically between carbohydrates sources and some potentially pathogenic species.

The extension of the analysis to metabolic potential allowed us to identify the correlations between carbohydrates and fat sources consumption and different metabolic pathways involved in different pathogenic processes

The analysis suggested that dietary habits have a limited effect on the peri-implant diseases manifestation. At the same time, dietary habits seemed to influence the oral microbial community in a way that can trigger a selective push towards potentially pathogenic species involved in different oral pathogenic processes.

The absence of strong correlations between dietary habits and disease state may depend on: 1) site specificity of the metagenomic samples involved in the analysis (i.e., sub-gingival plaque); 2) relative low quality of the dietary habits data (the collection of questionnaires was strongly impacted by COVID-19); 3) nature of dietary data collected via FFQs known to be based on estimations with possible strong differences from real consumptions.

Further investigations may be performed by extending the study to the saliva microbiome and by acquiring additional information about the oral environment (such as pH and nitrate content). A better definition of the diet effect on microbiome could be achieved also by acquiring microbiome information of diseased subjects and by sequencing longitudinal metagenomic data.

Contents

1 Introduction

1.1 An introduction to the human microbiome

It has become common knowledge that we are symbiotic creatures, accommodating a large number of microbial hosts. Since the discoveries of Leeuwenhoek and de Bary on bacteria in human mouths and algal-fungal partnerships that constitute lichens, the research on symbiosis has progressed remarkably. We discovered that microbial symbionts have been shaping the make-up of the eukaryotic influencing growth, development, energy metabolism, nutrition, digestion and defence of eukaryotes [1,2]. The set of microorganisms colonising a given biological niche is called microbiome. The microbiome, composed by species living in a symbiosis status, establishes with the host an endosymbiosis relationship. The endosymbiosis has deeply impacted the evolution of life and continues to shape the ecology of countless species. The evolution has been traditionally seen as a largely bifurcating pattern, reflecting mutations and other changes in existing genetic information and the occasional speciation and divergence of lineages. While lineage bifurcation has clearly been important in evolution, the endosymbiosis has also made profound contributions to evolutionary novelty with the merging of different lineages during the endosymbiosis relationship [1,3].

In the last 20 years, the study of the human microbiome has assumed a central role in the scientific community [4–9]. The microbiome, represented by the set of microbial species colonising different biological niches of the human body (e.g., skin, gut, mouth, and vagina), has been largely studied in terms of its composition and its influence on human health and disease. The ever increasing knowledge of the microbiome allowed scientists to identify several microbiome traits (e.g., taxonomic-level relative abundances of some species) associated with several diseases [10]. This has been unlocked also thanks to the possibility of estimating host phenotypes from microbiome data using computational approaches involving machine learning techniques [10,11], large-scale analyses [12,13], and meta-analyses aiming at integrating multiple cohorts and data types.

1.2 The oral microbiome and its role in disease

In the literature, most of the attention has been given to the characterization of the gut microbiome (36,000 entries in Scopus as of August 2022), although other biological niches of the human body have deserved attention. Of particular interest is the oral microbiome (6,00 entries in Scopus as of August 2022) represented by the microbial population living in our mouths. Oral microbiome is recognized to be remarkably resilient and quite diverse in terms of bacterial, archaeal, viral, and fungal component [14–16]. The resilience of the oral microbiome mostly derives from the slow growth rate of saliva and crevicular fluid, the two principal substrates available in our mouths.

Study and characterization of the oral microbiome went hand in hand with the linkage of its composition with some oral diseases. The use of culture-independent methods allied with next generation DNA sequencing methods is providing a far deeper analysis than hitherto possible. In contrast to the commensal microbiota found in other body sites that typically live in harmony with the host, the normal microbiota of the mouth is responsible for the two commonest diseases – dental caries and periodontal diseases [17], in addition to multiple other diseases. As example, the dissolution of tooth structures by acids produced as a result of the fermentation of dietary carbohydrates by oral bacteria [18–20]; endodontic infections or infections and death of the teeth pulp [21–23]; periodontitis or the loss of the attachment between gingiva and teeth that can bring to the loss of the teeth due to an heavy colonisation of the gap by anaerobic bacteria [24–26]. Moreover, the oral microbiome is known as a reservoir of infection for other body sites gaining access to the bloodstream through carious lesions. Oral microbiome is indeed correlated to infectious endocarditis [27], and brain and liver abscesses [28,29]. Oral bacteria have been detected in the lungs in cystic fibrosis patients [29]. In addition, it has been suggested that the oral microbiome may be linked to diseases that affect other body sites either in a causative way or as a reflection of systemic changes in the body [30]. In the latter case, oral microbiome analysis could be used as a diagnostic tool to detect other diseases.

1.2.1 Oral microbiome and peri-implantitis

In the last two decades, there has been an increasing use of dental implants to replace missing teeth [31]. The growing diffusion of dental implants led to a spread of implant diseases such as peri-implantitis, which affects both soft and hard tissues surrounding the implant, and mucositis, which precedes peri-implantitis and involves, instead, only soft tissues [32–36]. The clinical characterization of these two implant pathologies is represented by a continuum of inflammation, tissue destruction, and microbial pressure. Such symptoms are also a consequence of host-specific immune-mediated responses, genetics, as well as lifestyle and environmental factors [31,37]. The trigger of peri-implant disease is still unknown and under investigation. It is still unclear if it is the microbial challenge or the hyperinflammatory state itself [38]. The association between disease state and oral microbiome composition was and still is largely investigated [31,39–57]. Some authors emphasise that the shifting of microbiome composition during the pathogenesis of peri-implant diseases own a mutual relationship with the hyperinflammatory state typical of the disease in a kind of reciprocal causation and tries to shape the typical peri-implant microbiome[39,47,51,58,59]. Other authors identify the shifting in microbial composition as a cause to the development of peri-implant diseases and underlines the possibility to prevent peri-implantitis by looking at oral microbial composition [31,49,50,52,53].

1.3 Metagenomic analyses of the oral microbiome

In the literature, most of the works that investigated the relationship between peri-implant diseases and oral microbiome relied on analyses based on 16S rRNAs sequencing. Despite the scientific goals that have been reached through such an approach in this and other contexts, it basically limits the analysis to taxonomic characterization at genus level, preventing finer analyses such as species-level, strain-level, and functional-potential. Moreover, they rely on a quite small number of samples [41–46,48,59–62]. Additionally, external factors that may have an influence on the oral microbiome composition have been taken into account only marginally. The human microbiome, including the oral one, is an evolving community that responds to the push of external factors. Extrinsic host factors refer to external stimuli that have the

potential to affect the individual microbiome and can exert an influence on the host characteristics. These factors that are not inherent biological characteristics of the host can be modulated directly (e.g., lifestyle and behaviour) or not intentionally (e.g., environmental factors) [63] by the host. Such factors can contribute to an imbalance of the oral environment and potentially lead to a diseased state [64]. Only scattered studies have been previously conducted to evaluate effects of lifestyle on the oral microbiome composition [65–67].

1.4 The effect of diet on the oral microbiome

Despite major dietary changes that happened in history, the oral microbiome stayed relatively stable, going through a compositional shift and a decrease in diversity [68]. Moreover, adaptation of bacteria to these new environmental conditions [69,70] and continuous changes in the environment and lifestyle of the host [71] are believed to have greatly contributed to the present configuration of the oral microbiome in humans. To date, due to the resilient nature of the oral microbiome [14], there are only a few studies investigating this relationship. The disentanglement of the effect of the consumption of different foods on the oral microbiome composition is still ongoing. Some evidence indicates that a frequent consumption of fermentable carbohydrates is one of the causes of dental caries, driving the plaque ecology towards a state of dysbiosis [65,72,73]. The fermentation of the carbohydrates by the oral microbiota led to the formation of organic acids that can lower the oral pH if the buffer effect of the saliva is overwhelmed, creating a selective push for the acid tolerant bacteria involved in cariogenic process [74]. The differences in oral microbiome between different diet regimes (omnivore and vegan diet) were shaped. This kind of variances were attributed to the ingestion of specific macro and micronutrients, this evidence is not conclusive, and it is not sure these types of diets are able to modulate the oral microbiota [75]. Other researchers tried to investigate the shifts in the oral microbiome of elite male endurance race walkers from Europe, Asia, the Americas and Australia, in response to one of three dietary patterns often used by athletes during a period of intensified training: a High Carbohydrate diet, a Periodised Carbohydrate diet or a ketogenic Low Carbohydrate High Fat diet [76]. Furthermore, different dietary components have been studied in search for modulatory effects on the oral

microbiome. Inhibitory properties of short-, medium- and long-chain fatty acids consumption on the oral microbiome have been described, suggesting a modulating effect on the oral ecology [77]. So far, the available evidence to assess the real impact of dietary preferences and different nutrients on the oral microbiome is still insufficient and more studies are needed. The lack of coverage in terms of the number of studies investigating the relation between oral microbiome and diet goes hand in hand with the obsolete technique used to shape the oral microbiome.

The totality of the works aiming at investigating the relationship between oral microbiome and diet relies on 16S data, which is limited to genus-level taxonomic resolution. With whole-genome sequencing (WGS) techniques such as shotgun sequencing it is possible to reach species- and strain- level resolution by enabling a more detailed characterization of the oral microbiome and its more in-depth link with dietary patterns.

We will fill this gap in the present study by acquiring a new cohort associated with dietary information and oral microbiome data and analysing it through state-of-the-art techniques involving computational tools, machine learning (ML) approaches, and downstream statistical analyses.

1.5 Summary of my doctoral research activity

My doctoral research activity has been focused on multiple research questions. The overall experimental design with generated data and performed analyses is summarised in Fig 1.1. In brief, we collected an Italian cohort of 451 subjects involving people with dental implants from 48 dental clinics (thanks to the BetaProgram involving the PreBiomics company). Such a population spanned healthy people (N = 156) along with patients affected by peri-implantitis (N = 152) and mucositis (N = 143). We collected a rich set of clinical variables along with dietary habits through the administration of the EPIC FFQ validated questionnaire. We also collected the plaque sample for 121 out of 156 healthy subjects and sequenced their microbiome through shotgun sequencing. Main research questions involved the evaluation of dietary patterns in our cohort along with their relationship with clinical variables and health

status (Chapter 3). We also characterised the oral microbiome and evaluated their potential links with diet in the healthy population (Chapter 4). Such goals were accomplished through the involvement of statistical approaches and machine learning methods, which were partially optimised in Chapter 2. Finally, conclusions and future research lines are drawn in Chapter 5.

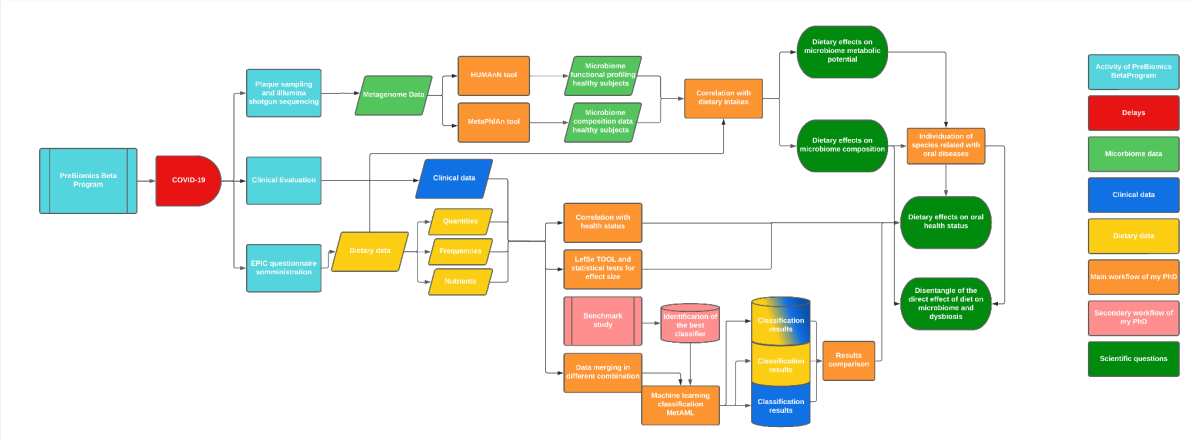


Fig 1.1. Flowchart of my PhD research project. Graphical representation of the data involved in my PhD research project and the analyses made to answer my scientific questions.

2 Benchmarking of multiple machine learning classification methods and feature extraction strategies

This chapter reports the analyses and results described in the article “Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa” that I authored as first author and that was published in *PLOS Computational Biology* on April 21st, 2022 [13].

2.1 Introduction and scientific rationale

Evidence has linked the human microbiome, the large set of microorganisms that reside in our body, with health and disease conditions [10]. Several diseases have been associated with microbiome traits and estimation of host phenotypes from microbiome composition has received remarkable attention in the community. In this regard, growing attention has been given to predicting host phenotypes using machine-learning based approaches, and in which adoption of classification methodologies for case-control studies has represented the most investigated scenario [11]. Classification represents a practical approach to implicitly integrate multiple characteristics (i.e., features, such as the case of combination of hundreds of microbial relative abundances) and get evaluation metrics of relatively easy interpretation. This is the case of the area under the receiver operating characteristic curve (AUC), the most used metric in the microbiome field for binary classification problems [11], which ranges in value from 0 to 1 with better accuracy when moving towards one.

Focusing on case-control studies, machine learning methods have been involved in two main types of analyses. The first has relied on applying established methodologies to newly generated data, which has allowed researchers to provide evidence of the predictability of host phenotypes from microbiome data for several different diseases including inflammatory bowel disease [78], obesity [79], type-2 diabetes [80], colorectal cancer [81], and paved the way to the potential use of the microbiome as a

diagnostic tool [82,83]. The increasing number of large population studies [84,85] has also enabled the implementation of several (large-scale) meta-analyses aiming at validating findings across independent cohorts. Besides analyses based on 16S rRNA data [86–88], similar efforts have been extended more recently to shotgun data [12,89–91], while extension to other-omics data has been more challenging[12]. The second group of analyses has been focused on the proposal of new methodologies in two main directions: extraction of better feature representations or optimization at classifier level [92]. While classification can be applied on the original set of features, improvements can be obtained by reducing the dimensionality of the feature space (for example by selecting or extracting specific operational taxonomic units (OTUs) or microbial taxa). Examples include feature subset selection [93], recursive feature elimination [89], and hierarchical feature engineering [94]. Different (supervised) methods have been adopted for classification purposes. Some widely used strategies are represented by logistic regression [95], support vector machines (SVMs) [78], k-nearest neighbours [96], and random forests (RFs) [88]. Comparisons among different classifiers have also been performed, with ensemble methods such as RFs and extreme gradient boosting decision trees that have exhibited in general the best performances [97]. Recently, different solutions based on deep learning approaches have been also proposed [98,99], including methods to transform high-dimensional data into robust low-dimensional representations [100], although challenges still arise due to the limited amount of labelled information that is typically available in case-control microbiome studies [101].

Despite the different methodologies adopted along the classification pipeline, classification models have been typically built by considering OTU or relative abundance profiles as input features. However, such types of data are intrinsically sparse, therefore this potentially enables to make inferences from the presence/absence of microbial taxa rather than their relative abundance values. This also poses the question whether it is the presence of particular taxa rather their abundance values to be relevant for discrimination purposes. Surprisingly, this aspect has not been investigated yet.

In this chapter, we aim at filling this gap by presenting a meta-analysis on publicly available datasets from both shotgun and 16S rRNA data. Such analysis was carried

out for two main reasons: 1) To test the hypothesis that classification on microbiome data could be done on presence/absence profiles without loss or even with an improvement in terms of classification performances; 2) To evaluation and compare different classification algorithms to find the best choice in terms of classification performances. Such findings will be beneficial to perform the ML-based analysis conducted in Chapter 3 and Chapter 4.

2.2 Materials and methods

2.2.1 Data Availability

The data and source code used to produce the results and analyses presented in this manuscript are available on a GitHub repository at https://github.com/RGilib/giliberti-meta-analysis-2022.

2.2.2 The considered publicly available metagenomic and 16S rRNA datasets

In this chapter, we conducted a meta-analysis on publicly available human metagenomic datasets for host phenotype classification. More specifically, we considered 4,128 samples coming from 25 shotgun metagenomic studies/datasets as summarised in Table 2.1 and Fig 2.1A. Twenty-one studies were devoted to the characterization of the gut microbiome in association with different diseases (i.e., case-control studies). Two additional datasets were case-control studies (peri-implantitis, mucositis, and schizophrenia) from oral metagenomes. We also considered a dataset aiming at characterizing changes in the human microbiome due to consumption of cephalosporins, while the last dataset was devoted to the discrimination between body sites (i.e., stool vs oral) in the Human Microbiome Project (HMP) dataset. Metagenomic samples were processed to generate species-level taxonomic profiles through MetaPhlAn3 [102]. Species abundances are expressed as real numbers in the range [0,1] with values that sum to 1 for each sample. Generation of relative abundances at other taxonomic levels (i.e., genus, family, and order) was also extracted from the MetaPhlAn3 output. Metadata information in terms of disease

status or body site for the HMP dataset are available in the curatedMetagenomicData package [103].

We additionally analysed 4,026 16S rRNA samples coming from 30 publicly available case-control studies (S1 Table and Fig 2A). We considered the same set of gut samples considered in [88] with metadata information in terms of disease status as follows: autism spectrum disorder (ASD), *Clostridioides difficile* infection (CDI), CRC, enteric diarrheal disease (EDD), human immunodeficiency virus (HIV), IBD, liver cirrhosis (CIRR), minimal hepatic encephalopathy (MHE), non-alcoholic steatohepatitis (NASH), obesity (OB), Parkinson disease, psoriatic arthritis (PSA), rheumatoid arthritis (RA), and T1D. 16S rRNA samples were pre-processed following the same procedure adopted in [88]. More specifically, we discarded samples with fewer than 100 reads and removed OTUs with less than 10 reads and/or present in less than 1% of the samples. After calculating the relative abundance of each OTU, OTUs were collapsed to genus level by summing their relative abundance values and by discarding any OTUs which were un-annotated at the genus level.

Table 2.1. Summary of the 25 classification tasks derived from metagenomic datasets for case-control prediction. ACDV: Atherosclerotic cardiovascular disease, AD: Alzheimer's disease, BD: Behcet's disease, CRC: Colorectal cancer, IBD: irritable bowel disease, T1D: Type 1 diabetes, T2D: Type 2 diabetes. We additionally considered the HMP_2012 dataset [84] for body site discrimination between gut (N = 414) and oral (N = 147) samples.

Dataset name	body site	# controls	Cases	# cases	Reference
JieZ_2017	Gut	171	ACVD	214	[104]
ChngKR_2016	Skin	40	AD	38	[105]
YeZ_2018	Gut	45	BD	20	[106]
RaymondF_2016	Gut	36	cephalosporins	36	[107]
QinN_2014	Gut	114	cirrhosis	123	[108]
FengQ_2015	Gut	61	CRC	46	[109]
GuptaA_2019	Gut	30	CRC	28	[110]
HanniganGD_2017	Gut	28	CRC	27	[111]
ThomasAM_2018a	Gut	24	CRC	29	[112]
ThomasAM_2018b	Gut	28	CRC	32	[112]
VogtmannE_2016	Gut	52	CRC	52	[113]
WirbelJ_2018	Gut	65	CRC	60	[114]
YachidaS_2019	Gut	251	CRC	258	[115]
YuJ_2015	Gut	53	CRC	75	[116]
ZellerG_2014	Gut	61	CRC	53	[81]
LiJ_2017	Gut	41	hypertension	99	[117]
IjazUZ_2017	Gut	38	IBD	56	[118]
NielsenHB_2014	Gut	248	IBD	148	[119]
GhensiP_2019_m	Oral	49	mucositis	20	[31]
GhensiP_2019	Oral	49	peri-implantitis	23	[31]
Castro-NallarE_2015	Oral	16	schizophrenia	16	[120]
Heitz-BuschartA_2016	Gut	26	T1D	27	[121]
KosticAD_2015	Gut	89	T1D	31	[122]
KarlssonFH_2013	Gut	43	T2D	53	[123]
QinJ_2012	Gut	174	T2D	170	[124]

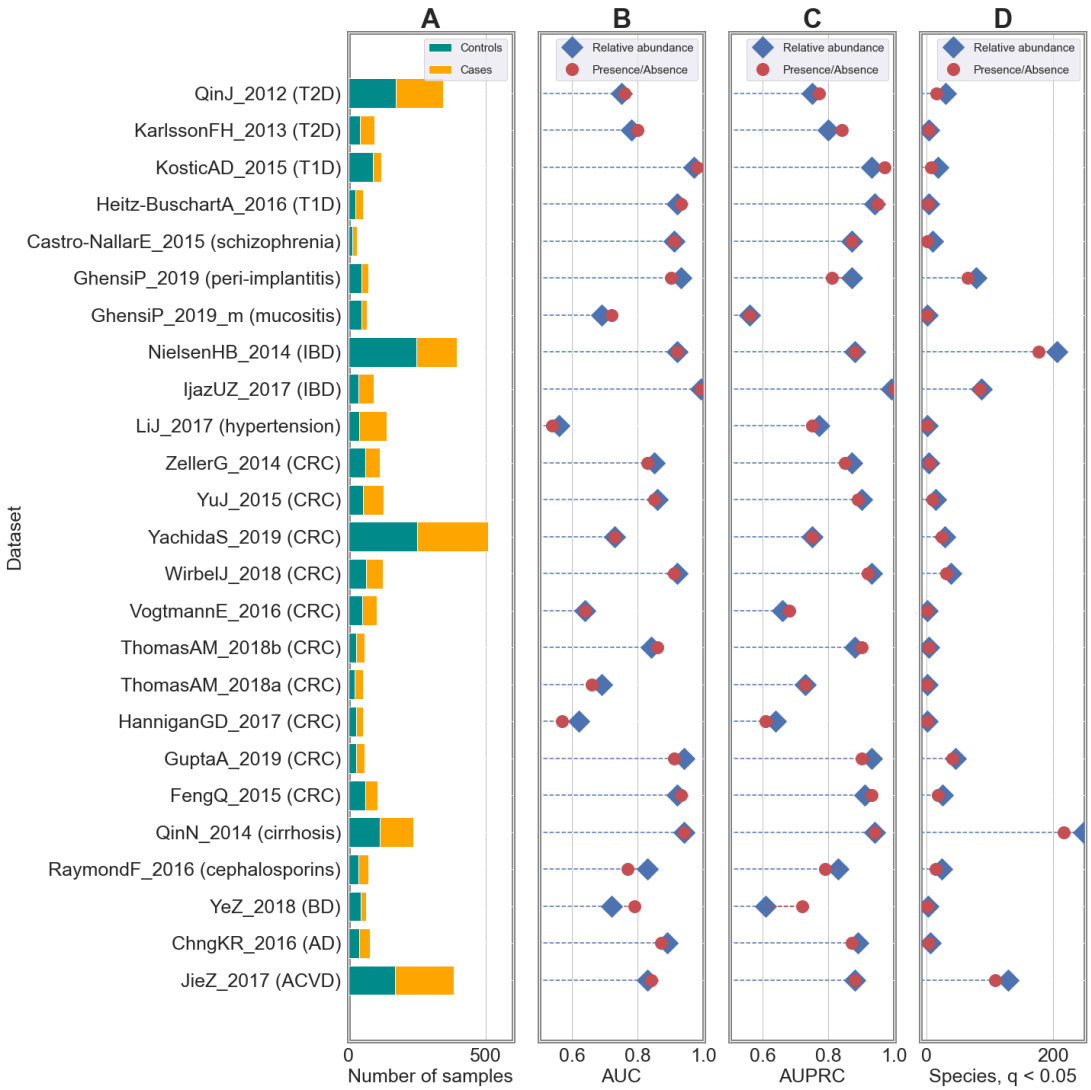


Fig 2.1. Classification accuracies are robust to degradation from species-level relative abundance to presence/absence profiles in shotgun datasets. Results obtained on 25 case-control studies for host phenotype classification from human microbiomes. (A) Number of case and control samples across the different studies. (B) AUC and (C) AUPRC scores using RF as back-end classifiers on species-level taxonomic profiles. Comparison between relative abundance (in blue) and presence/absence (in red) profiles highlighted negligible differences and no statistical differences in none of the studies (see S2.1 Fig for AUC scores and S2.2 Table for p-values). Metrics of comparison in terms of AUC, AUPRC, precision, recall, and F1 are summarised in S2.2 Fig and S2.2 Table is represented a comparison between AUC and AUPCR scores. (D) Number of statistically significant taxa from relative abundance (in blue) and presence/absence (in red) profiles.

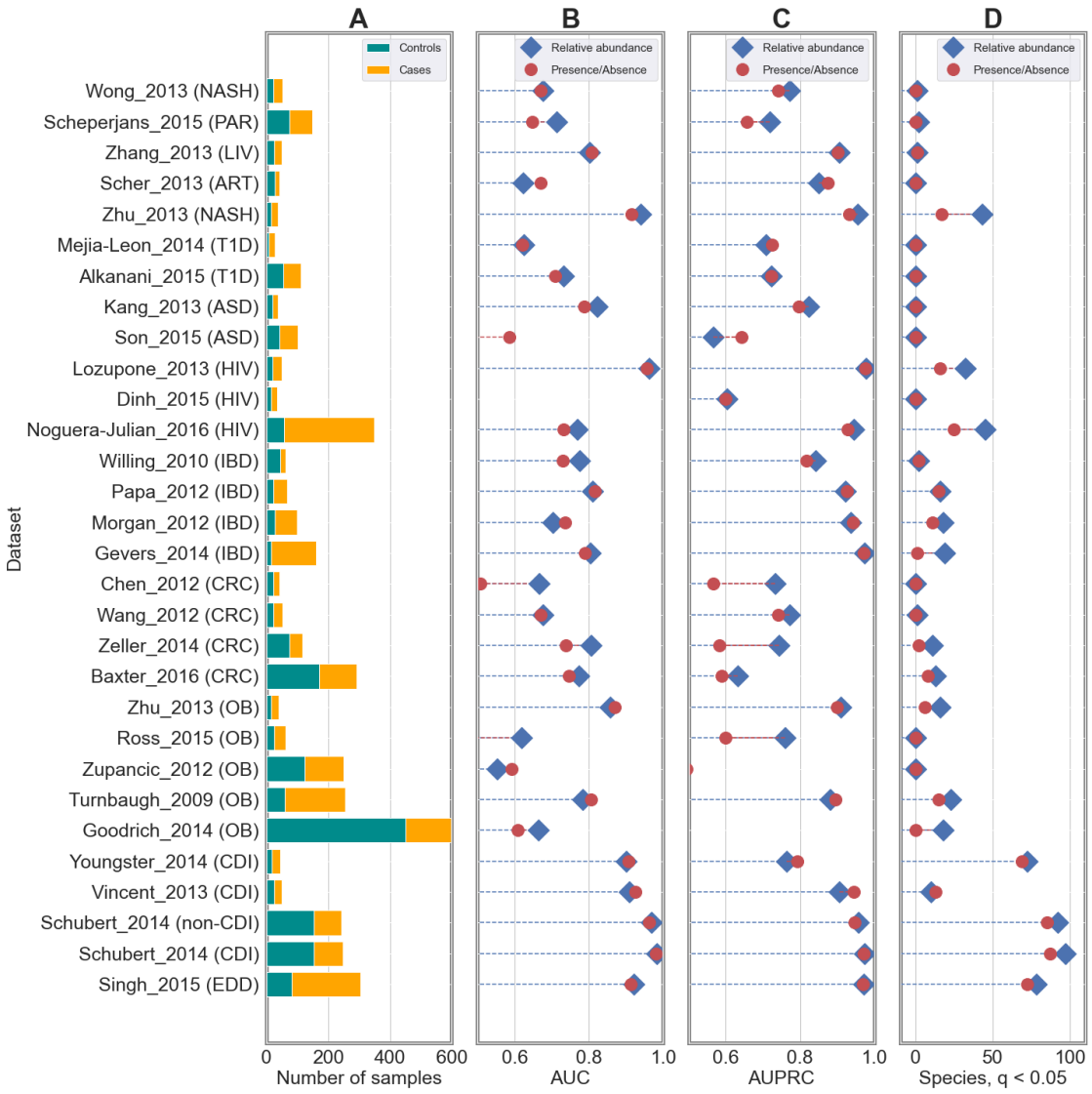


Fig 2.2. Classification accuracies are robust to degradation from genus-level relative abundance to presence/absence profiles in 16S rRNA datasets. Results obtained on 30 case-control studies for host phenotype classification from human microbiomes. (A) Number of case and control samples across the different studies. (B) AUC and (C) AUPRC scores using RF as back-end classifiers on species-level taxonomic profiles. Comparison between relative abundance (in blue) and presence/absence (in red) profiles highlighted negligible differences and no statistical differences in none of the studies (see S2.5 Table for p-values) as found also in shotgun datasets (see Fig 2.1). Metrics of comparison in terms of AUC, AUPRC, precision, recall, and F1 are summarised in S2.5 Table. (C) Number of statistically significant taxa from relative abundance (in blue) and presence/absence (in red) profiles.

2.2.3 The adopted machine learning methods

The classification tasks on both shotgun and 16S rRNA data were carried out by considering the already developed and validated MetAML (Metagenomic prediction Analysis based on Machine Learning) tool [89]. Main analyses were conducted by using Random Forests (RFs) as back-end classifiers, and validations were extended to other three classifier types: support vector machines with linear (denoted with LSVM in this chapter) and RBF (denoted with SVM in this chapter) kernel, Lasso, and Elastic Net (ENet).

Free parameters of the classifiers were set as follows. For RF, i) the number of trees was set to 500, ii) the number of features to consider when looking for the best split was equal to the root of the number of original features, and iii) the Gini impurity criterion was used to measure the quality of a split. For Lasso and ENet, the regularisation parameters were obtained using a 5-fold stratified cross-validation approach. For Lasso the alpha parameter was found in the set $\{10^{-4}, \ldots, 10^{-0.5}\}$ with 50 uniform steps. For ENet, besides the alpha parameter, also the L1_ratio parameter was chosen in the set [0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0].

2.2.4 Validation and evaluation strategies

We conducted two main types of analysis: i) cross-validation and ii) cross-study analysis. In cross-validation, samples were randomly divided into k (with k = 10 in our case) folds by considering a stratified cross-validation approach to preserve the percentage of samples of each class. Results were repeated and averaged on 20 independent runs. Different models were trained on the same cross-validation splits. We also considered a cross-study analysis in order to evaluate robustness of the prediction when transferring models from a source to a target domain. In this setting, the classification model was trained on the source dataset and accuracy was evaluated on a different independent dataset.

Classification accuracies were evaluated in terms of five main metrics: area under the curve (AUC), area under the precision-recall curve (AUPRC), precision, recall, and F1. We calculated mean difference and standard error for each 10-fold CV and averaged across the 20 repetitions. We calculated the 95% confidence interval on the difference

in AUC performance between two classifiers as done in [89]using the t-distribution with df = 9:

$$95\%CI : \frac{1}{20}\frac{1}{10}\sum_{i=1}^{20}\sum_{i=1}^{10}(AUC_{1ij} - AUC_{2ij}) \pm 2.26 \times \frac{\sigma_j}{\sqrt{10}} \quad (1)$$

where AUC1ij and AUC2ij are the AUC of two classifiers in fold i of repetition j, and σj is the standard deviation of the AUC1ij−AUC2ij across i = 1…10 folds in repetition j. We computed the p-values from the t-statistics from mean difference and standard error smoothed over the 20 repetitions:

$$t = \frac{\frac{1}{2010}\sum_{i=1}^{20}\sum_{i=1}^{10}(AUC_{1ij} - AUC_{2ij})}{\frac{1}{20}\sum_{i=1}^{20}\frac{\sigma_j}{\sqrt{10}}} \quad (2)$$

We used a two-tailed t-test with df = 9.

2.2.5 Experimental setting for shotgun datasets

Most of the analyses on shotgun datasets were conducted by considering a cross-validation approach. Twenty-four classification tasks were devoted to the discrimination of healthy from diseased subjects (i.e., case-control studies), while the HMP dataset was used to perform body site discrimination between gut and oral samples. We also considered the ten independent datasets associated with CRC and evaluated prediction capabilities in a cross-study setting.

Baseline results were obtained by considering the original relative abundance profiles at species-level resolution provided by MetaPhlAn3 [102] as features and using RF as back-end classifier. This is the setting that was successfully deployed and validated in multiple meta-analyses such as the ones presented in [31,89,103,112]. At this point, multiple comparisons were performed: i) starting from the original species-level relative abundance profiles (one profile for each sample), we generated presence/absence profiles by simply thresholding the relative abundance values at 0%. This generated a set of boolean profiles where 1 indicated the presence of the species regardless of its relative abundance in the considered sample, while 0 was associated with its absence. The same approach based on RF was applied on this set of newly generated profiles and compared with the results obtained on the original relative abundances. Results are summarised in Figs 2.1B, 2.1C, S2.1 and S2.2; ii)

the same procedure described in i) was applied again by thresholding the relative abundance profiles at different values to assess sensitivity of classification to low abundant species. We considered these values as threshold levels: 0.0001%, 0.001%, 0.01%, and 0.1%. Results using RF as classifier are summarised in Figs 2.3A and S2.3A; iii) we extended the comparison done at species-level between original relative abundance and boolean (with threshold = 0%) profiles to three other taxonomic levels (i.e., genus, family, and order) to evaluate sensitivity of classification when moving from species to coarser taxonomic resolutions. Results with RF classification are summarised in Figs 2.4 and S2.3B; iv) we finally assessed robustness of our findings to the choice of the classification method. We compared RF results with the ones obtained by other four classifier algorithms (i.e., SVM with linear kernel, SVM with RBF kernel, Lasso, ENet) for both relative abundance and presence/absence profiles (Figs 2.5 and S2.3C). While we report in main figures only comparisons in terms of AUC, comparisons for the other three metrics (i.e., precision, recall, and F1) are reported in S2.2 Table.

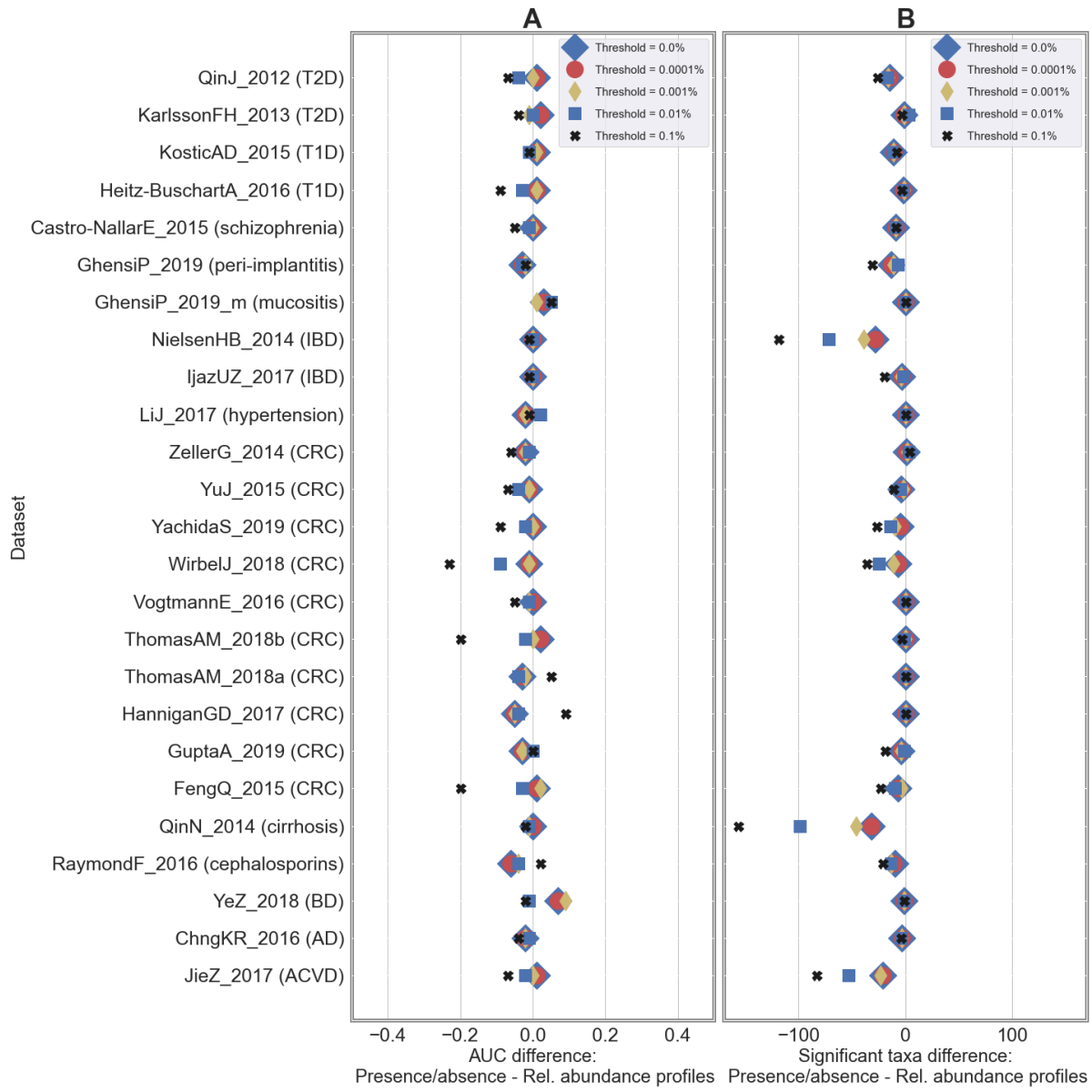


Fig 2.3. Classification accuracies are not impacted when relative abundances are thresholded up to 0.001%.

Results on the 25 case-control shotgun studies by comparing the baseline (i.e., species-level relative abundance profiles) with the presence/absence profiles generated by thresholding at different relative abundance values (ranging from 0% to 0.1%). (A) Difference in AUC between the presence/absence and the relative abundance RF classification result. A positive value indicates that presence/absence outperforms relative abundance data. AUC scores at different thresholds are summarised in S2.2 Table. (B) Difference in number of statistically significant taxa (numbers summarised in S2.7 Table).

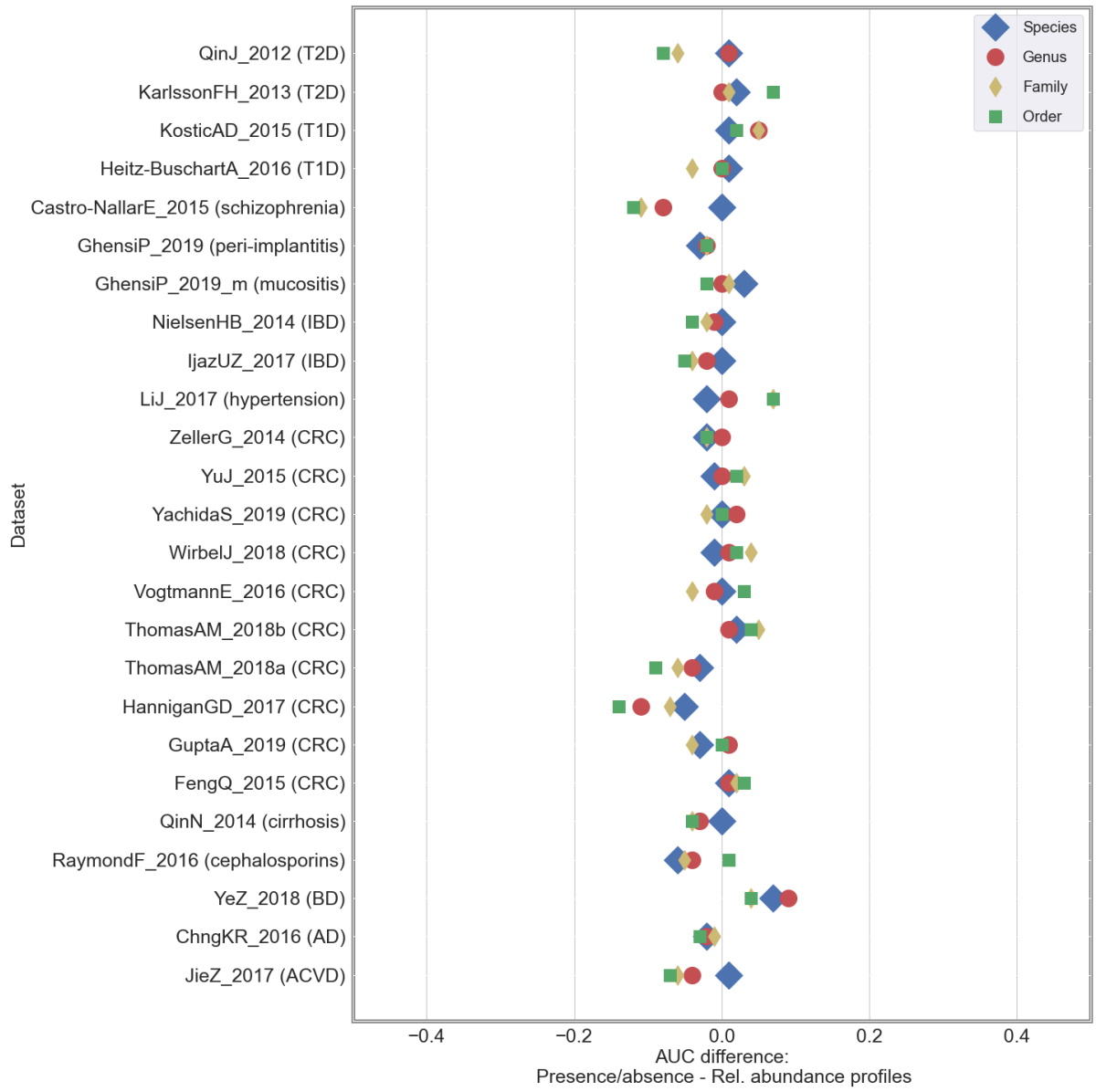


Fig 2.4. Classification results are more impacted to relative abundance degradation at coarser taxonomic resolution.

Results on the 25 case-control shotgun studies by comparing the baseline (i.e., relative abundance profiles) with the presence/absence profile generated by thresholding at 0.0% and varying taxonomic resolution from species to order level. Difference in AUC between the presence/absence and the relative abundance RF classification result. A positive value indicates that presence/absence outperforms relative abundance data.

2.2.6 Experimental settings for 16S rRNA datasets

For 16S rRNA datasets we carried out only cross-validation analyses. From the genus-level profiles generated as described in the section “The considered publicly available metagenomic and 16S rRNA datasets”, we generated the boolean profiles (with threshold = 0%) as similarly done for shotgun data. We compared the two types of profiles using a RF classifier (results in Figs 2.2B, 2.2C and S2.4), and were then extended also to the other classifier types (results in S2.3 Table).

2.2.7 Statistical tests

On the same set of scenarios in which we compared classification accuracies, we conducted statistical tests to evaluate to which extent degradation from relative abundance to boolean profiles can impact the identification of differentially abundant species. We used Mann-Whitney U test to identify the set of significant taxa when relative abundance profiles were involved, while we adopted Fisher exact test to deal with presence/absence data. Although it is out of the scope of the present study to perform a comprehensive evaluation of available statistical tests, further investigation taking into account alternatives including methodologies that can deal with compositional issues [125,126] is warranted. Finally, false detection rate (FDR) was applied for multiple testing correction, and corrected p-values < 0.05 identified significant taxa.

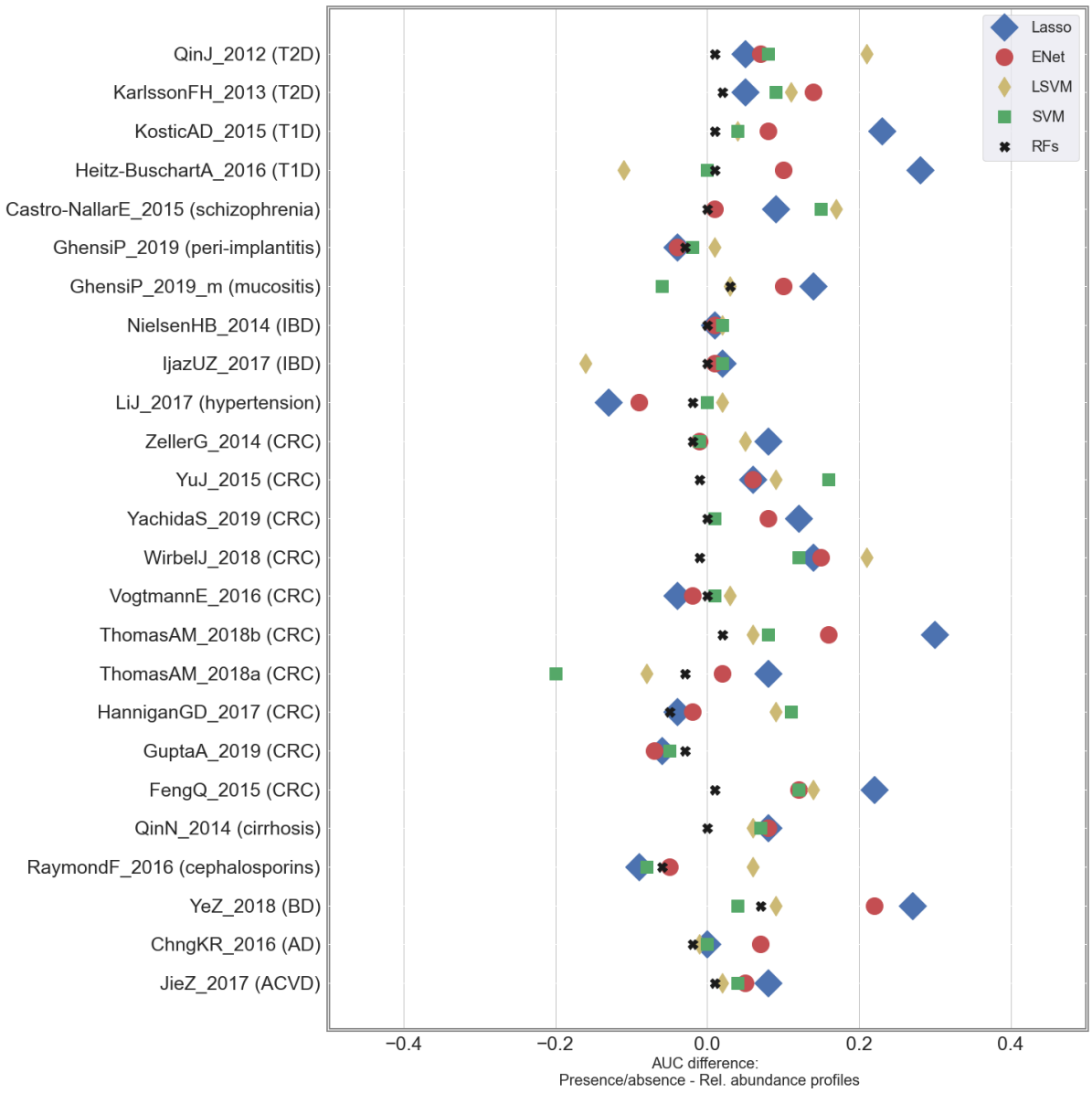


Fig 2.5. Findings in terms of stability of the classification accuracy are robust to the classifier choice.

Differences in terms of AUC between presence/absence and relative abundance profiles for the 25 case-control shotgun datasets at varying classification algorithms. ENet: Elastic Net; LSVM: SVM with linear kernel; SVM: SVM with RBF kernel; RFs: Random Forests.

2.2.8 Rarefaction analysis

We further performed rarefaction analysis by: i) considering the three datasets having the highest number of significant species from relative abundance profiles (i.e., JieZ_2017, NielsenHB_2014, and QinN_2014); ii) rarefying raw reads (using https://github.com/lh3/seqtk) and considering 1M reads for each metagenome; iii) applying the same pipeline to generate taxonomic profiles through MetaPhlAn3; iv) applying the same pipeline to build classification models and identifying statistically significant species.

2.3 Results and discussion

In this chapter, we conducted a meta-analysis aiming at evaluating to which extent degradation from relative abundance to presence/absence of microbial taxa can impact host phenotype classification from human metagenomes. The analysis was conducted on 4,128 public available metagenomes coming from 25 datasets (Table 1 and Fig 2.1A). Metagenomes were uniformly processed to generate species-level taxonomic profiles with MetaPhlAn3 [102] (see Materials and Methods) with metadata information available in the curatedMetagenomicData package [102]. From relative abundance profiles, expressed as real numbers in the range [0, 1], we generated presence/absence profiles by simply thresholding the relative abundance values at 0%. This generated a set of boolean profiles where one indicated the presence of the species regardless of its relative abundance in the considered sample, while zero was associated with its absence.

2.3.1 Baseline classification results replicate original findings

As baseline, we considered the classification approach that we originally proposed in [89] and that was then used for different tasks such as detection of microbial signatures linked to colorectal cancer (CRC) from human metagenomes [112], characterization of the oral microbiome in dental implant diseases [31], and identification of changes associated with dietary interventional studies [127]. More specifically, we considered a RF classifier applied on the species-level relative abundance profiles, and evaluated classification accuracies in terms of multiple metrics (i.e., area under the ROC curve (AUC), area under the precision-recall curve (AUPRC), precision, recall, and F1) using

a cross-validation (CV) approach (see Materials and Methods). We obtained variable accuracies ranging from 0.56 (in terms of AUC) for hypertension in the LiJ_2017 dataset [119] to 0.99 for IBD in the IjazUZ_2017 dataset [118], with an average AUC across the 25 case-control studies equal to 0.83 (S2.4 Table). Such values were in line with what reported in the original publications, although a fair comparison is difficult to be performed due to differences in terms of adopted algorithms and input features. On the 17 publications that reported classification results on the same samples here considered, we obtained an average AUC of 0.80 in comparison to the average of 0.83 reported in the original publications (S2.4 Table).

2.3.2 Degradation from species-level relative abundance to presence/absence profiles does not worsen classification accuracies

We applied the same classification approach on the same set of samples to the presence/absence profiles (Materials and Methods). In this way, we evaluated to which extent moving from relative abundance to presence/absence information could impact classification accuracies. Surprisingly, we observed negligible differences between the two experimental settings (Figs 2.1B, 2.1C and S2.1 and S2.2 Table). In both cases (i.e., using presence/absence or relative abundance profiles), we obtained an average AUC of 0.83 (AUPRC = 0.83) across the 25 case-control studies, with AUC and AUPRC values strongly correlated (S2.2 Fig; Spearman correlation = 0.918). Some variations were observed at dataset-level (relative abundance outperformed presence/absence at a maximum of 0.06 in terms of AUC in the RaymondF_2016 dataset [106], while the opposite case was verified in YeZ_2018 [106] for an AUC difference of 0.07), however these were likely due to random perturbations and in none of the cases they were associated with statistically significant differences ($p > 0.05$, S2.2 Table). This was also confirmed in terms of the other metrics of comparison (i.e., precision, recall, and F1), with no significant differences between the two profile types (S2.2 Table). In a similar setting, we performed body site discrimination (oral vs stool samples) in the HMP dataset [84], with a value of AUC equal to 1.00 for both profile types. Therefore, such findings suggested that it was more the presence of same taxa rather than their actual relative abundance to be relevant for discrimination purposes.

We extended this analysis to 16S rRNA samples. More specifically, we considered the same set of 30 case-control studies for a total of 4,026 samples that were originally collected and analysed in [88] (Fig 2.2A and S2.1 Table). We applied the same pre-processing procedure adopted in [88] (Materials and Methods), and performed the prediction tasks by adopting the classification pipeline already considered for shotgun data. We obtained results similar to the ones presented in [88] on the genus-level relative abundance profiles (average AUC across the 30 datasets equal to 0.76 and 0.74 in our analysis and in [88], respectively) (S5 Table), although some differences could occur due to the different code implementations. By degrading relative abundance to presence/absence profile, we obtained few differences in the classification results between the two profile types (Figs 2.2B, 2.2C and S2.4 and S2.5 Table). Average AUC across the 30 studies was quite close (0.76 for relative abundance and 0.75 for presence/absence profiles), with differences that were statistically significant in only 3 out of 30 cases (S2.5 Table). Such differences, albeit impacting a limited number of datasets, may be due to the coarser taxonomic resolution and the higher noise component associated with 16S data.

2.3.3 Statistically significant taxa are consistent between relative abundance and presence/absence profiles

We extended the analysis from classification to identification of differentially abundant/present taxa (i.e., possible biomarkers) through statistical testing (Materials and Methods). By comparing the sets of statistically significant species in the different case-control studies ($q < 0.05$; using Mann-Whitney U test for relative abundance and Fisher exact test for presence/absence profiles, both corrected through false detection rate (FDR), S2.6 Table) we found similar numbers (Fig 2.1D and S2.7 Table), with values more driven by disease and dataset types than average number of reads (S2.5 Fig). On average, we found 39 and 32 significant species from relative abundance and presence/absence profiles, respectively. We may hypothesise that diseases that rely on rare biomarkers are less affected by degradation to presence/absence profiles than the ones that are characterised by stronger community shifts in abundant and prevalent taxa. Although this is not sufficiently supported by our data, further investigation in this direction is warranted.

On a per dataset basis, p-values associated with statistically significant species correlated well between relative abundance and presence/absence profiles (S2.6 Fig). This was reflected also by the high percentage of taxa (78%) that were detected as significant in both cases, which was further confirmed by performing hierarchical clustering on the set of statistically significant taxa coming from relative abundance and presence/absence profiles (S2.7 Fig). Conversely, we identified discrepancies between case-enriched and control-enriched taxa in only 1.74% of the statistically significant features, which were coming from just 5 of the 24 analysed datasets (S2.8 Fig). Moreover, we didn't identify any taxa for which the two tests disagreed across datasets (S2.8 Fig).

Focusing on the gut microbiome datasets, we also identified the species that were mostly associated with disease or health (S2.7 Fig). The species most enriched in cases was *Clostridium bolteae* (significant in 78% of the diseases), followed by *Streptococcus anginosus* group (55%), *Ruthenibacterium lactatiformans* (55%), *Hungatella hathewayi* (55%), and *Eisenbergiella tayi* (55%) with all of them already reported in the literature as possible biomarkers for different disease conditions [81,88,112,114,128]. Similarly, species most enriched in controls were *Anaerostipes hadrus* (significant in 66% of the diseases), *Roseburia faecis* (55%), *Roseburia intestinalis* (55%), *Prevotella copri* (44%), and *Eubacterium hallii* (44%) [81,84,112,129].

Consistence between relative abundance and presence/absence outcomes was finally obtained on the 16S data, with 20 and 15 genera that were found to be significant on average from relative abundance and presence/absence profiles, respectively (Fig 2.2D and S2.8 Table).

2.3.4 Relative abundance values lower than 0.001% do not impact classification outcomes

We evaluated how different values in thresholding relative abundance profiles could impact classification results. We thresholded the abundances at different values (i.e., moving from a threshold equal to 0%—which corresponded to the presence/absence scenario discussed in the previous section—to 0.0001%, 0.001%, 0.01%, and 0.1%,

Materials and Methods), meaning that values below the chosen threshold were forced to zero. We did not observe changes in the classification accuracy when the threshold was set to 0.0001% and 0.001% (Figs 2.3A and S2.3A and S2.2 Table). In both cases, we got an average AUC = 0.83 across the 25 case-control studies as obtained on the relative abundance profiles and using a threshold equal to zero, and no statistically significant differences were found. This was reflected by the number of statistically significant species (Fig 2.3B and S2.7 Table) that decreased very marginally from 32 (average value by considering 0% or 0.0001% as threshold) to 31 (threshold = 0.001%). Although very low abundant species may be actual biomarkers, they did not contribute to improving classification accuracies which was likely due to the impossibility to estimate their presence and relative abundance properly as being below or close to the limit of detection, which we quantified in this setting to be around 0.001% (with an average number of reads across our considered metagenomes equal to 47.5M). Major differences were obtained when thresholding at higher values (i.e., 0.01% and 0.1%). In these cases, average AUC decreased to 0.81 (threshold = 0.01%) and 0.78 (threshold = 0.1%), with significant differences in 3 and 6 cases, respectively.

Results on rarefied reads (Materials and Methods) showed, as expected, a slight decrease in terms of classification accuracies and number of detected biomarkers with respect to the original data set, although patterns in function of the thresholding value when going from relative abundance to presence/absence data were confirmed (S2.9 Table).

2.3.5 Coarser taxonomic levels are less robust to profile degradation

We further tested to which extent classification accuracy was affected by the taxonomic resolution level considered to feed the classifier. By considering original relative abundance profiles, average AUC moved from 0.83 (species-level resolution) to 0.80 (with 3 statistically significant cases), 0.78 (6), and 0.76 (11) for genus, family, and order levels, respectively (S2.10 Table). Such differences, albeit not too strong, suggested species as “optimal” level to optimise classification accuracies, with further improvements that may be obtained—although not tested here due to methodological limitations—with sub-species- or strain-level resolutions.

Similarly, we compared classification accuracies between relative abundance and presence/absence profiles at different taxonomic levels. While no differences were obtained at species-level (as already discussed in Fig 2.1), we observed that coarser resolutions brought increasing AUC differences (Figs 2.4 and S2.3B and S2.11 Table). An average AUC difference of 0.022, 0.041, and 0.061 was obtained for genus, family, and order, respectively (with 0, 1, and 2 statistically significant cases, respectively). Similar patterns were observed in terms of number of statistically significant features (S2.7 Table).

2.3.6 Findings are robust to cross-study analysis and to the classifier choice

We applied the same approach on a cross-study setting. We considered the ten independent metagenomics studies associated with CRC for a total of 1313 samples (Table 1) and applied a leave-one-dataset-out (LODO) approach in which the model was built on all datasets but the single dataset used for testing (Materials and Methods). As previously reported [112,114], we observed an overall moderate decrease of the accuracy when moving from CV (average AUC equal to 0.80) to LODO (average AUC equal to 0.76; S2.9 Fig and S2.12 Table). More importantly, we confirmed previous findings in terms of stability of the accuracy when moving from relative abundance to presence/absence profiles at species-level resolution (Fig 2.6A). The average AUC remained stable at 0.76 for the presence/absence profiles at threshold equal to 0%, 0.0001%, and 0.001%, while it decreased to 0.74 and 0.73 when thresholding at 0.01% and 0.1%, respectively. We also confirmed that the better taxonomic resolution was associated with smaller classification performance differences between relative abundance and presence/absence data (Fig 2.6B and 2.6C).

We finally tested if the choice of the classification method could impact the findings described in the previous sections in terms of degradation from relative abundance to presence/absence profiles. First, we confirmed [89] the superiority of RF with respect to other four classification methods (i.e., Lasso [130], Elastic Net [131], and support vector machines (SVMs) with linear and RBF kernels [132]) on both relative abundance (S2.10A Fig and S2.13 Table) and presence/absence profiles (S2.10B Fig

and S2.13 Table), and this was also verified on 16S rRNA data (S2.3 Table). On average, thresholding of relative abundance values did not negatively impact classification accuracies, instead it generally improved results in a quite unexpected way (Figs 2.5 and S2.3C). Higher differences were observed for Lasso, with an average AUC equal to 0.79 and 0.72 for presence/absence and relative abundance data, respectively, and the same pattern was obtained for the other classifier methods (with an average difference in terms of AUC equal to 0.05, 0.03, and 0.02 for ENet, LSVM, and SVM, respectively). We observed a greater variability of the classification accuracies with respect to what was observed for RF classification. In fact, we obtained statistically significant differences in Lasso, ENet, LSVM, and SVM studies for 10, 6, 5, and 6, respectively, however always in majority in favour of the presence/absence data. We therefore conclude that, despite a few differences occurred in a limited number of cases, maximization of classification accuracies was generally made possible through presence/absence profiles.

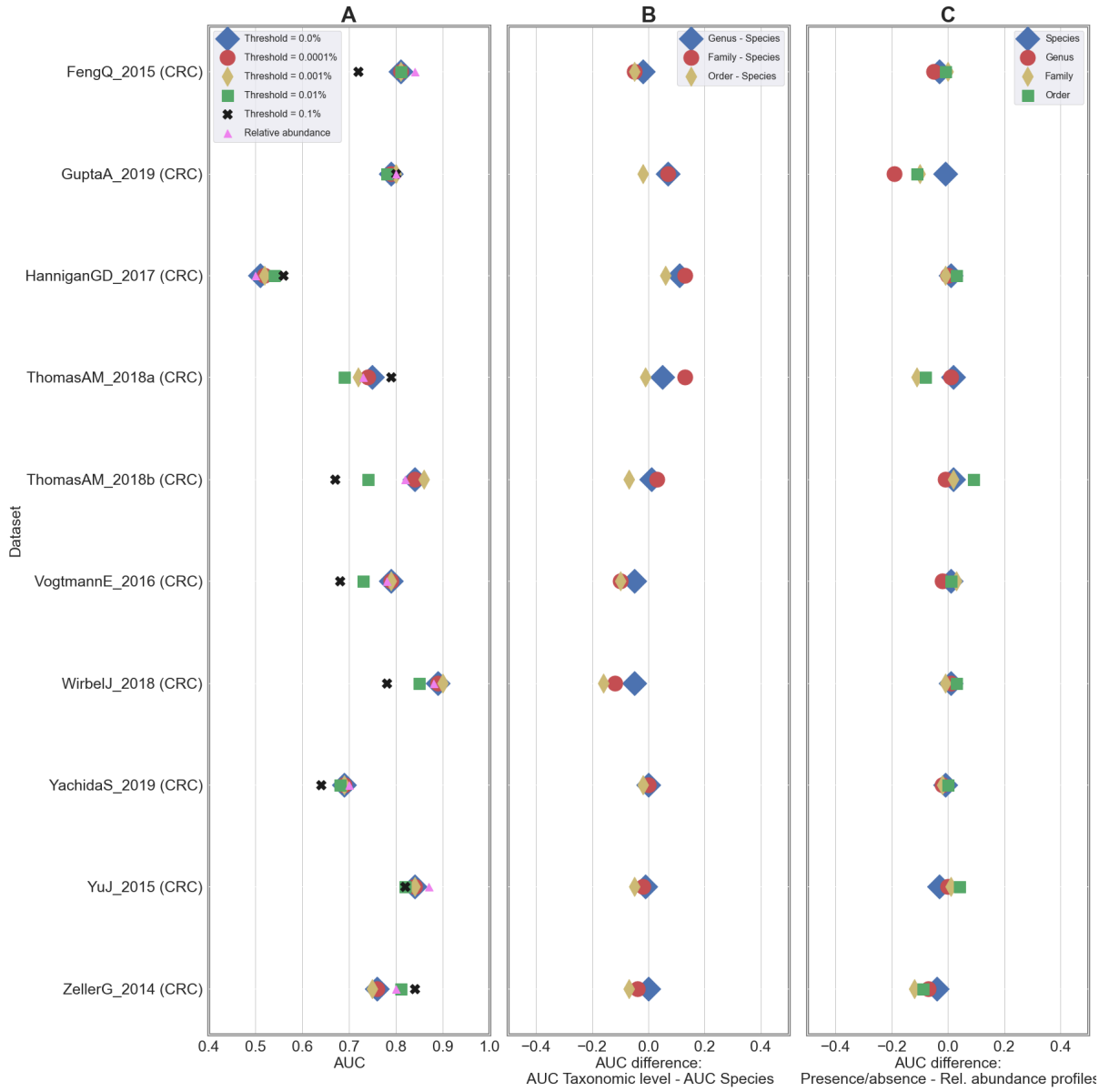


Fig 2.6. Degradation of relative abundance profiles does not impact LODO classification.

Results in terms of leave-one-dataset-out (LODO) validation on 10 CRC shotgun datasets. (A) AUC scores using RF as back-end classifiers on species-level relative abundance (in pink) and presence/absence profiles generated at different threshold values. (B) Difference in AUC between species and other taxonomic-level resolutions. A negative value indicates that species-level outperforms the comparison level. (C) Difference in AUC between presence/absence and relative abundance classification results at varying taxonomic levels.

2.4 Conclusions

In the present study, we conducted a meta-analysis on 25 publicly available datasets spanning more than 4,000 shotgun metagenomes and associated with different case-control studies. By applying species-level taxonomic profiling and machine-learning based classification approaches based on state-of-the-art methodologies we demonstrated that the presence of microbial taxa is sufficient to maximise classification accuracies. This was accomplished by degrading original relative abundance data to presence/absence profiles by considering different threshold values. We estimated a value of 0.001% in terms of relative abundance as limit of detection, meaning that although very low abundant species may be actual biomarkers they were not useful to improve classification accuracy. Results were robust to the choice of the classifier. This was obtained by considering different traditional classification algorithms that are designed for continuous data and potentially "suboptimal" when applied on binary data. This actually reinforces our findings, meaning that accuracies may be even better when models on presence/absence profiles are trained using classifiers more designed for binary data. Moreover, although doing an extensive evaluation of existing classifiers is out of the scope of the present study, maximisation of classification accuracies may be reached by adopting other classification approaches including the ones specifically proposed for microbiome data analysis [133,134]. Findings were finally extended from cross validation to cross study analysis and confirmed on 16S rRNA data associated with a compendium of more than 4,000 samples coming from 30 public studies.

The growing literature aiming at identifying microbial biomarkers for different diseases opened the possibility to build non-invasive diagnostic tools from microbiome data. To this purpose, much superior accuracy can be achieved by considering multi-feature rather than single biomarkers diagnostic models, and in which machine learning-based classification approaches have a fundamental role in building such models. Moreover, maximal accuracy can usually be achieved by using a limited number of features (in the order of ten or twenty). Such findings recently presented in the literature in addition to outcomes of our study, which suggest that the detection of microbial taxa is sufficient to maximise classification accuracies, are important steps toward the development of fast and inexpensive tests applied on stool samples for diagnostic purposes.

Regarding the specific tasks involved in this doctoral research activity, we found RF classifier as the best one for maximising classification accuracies and quite robust to multiple data transformation techniques. This classification algorithm will be adopted to build the classification models presented in Chapters 3 and 4.

3 Collection and evaluation of dietary intakes and clinical data in their relationship with the peri-implant health status

3.1 Introduction and scientific rationale

Dietary recommendations have changed substantially. In the past, high-carbohydrate diets were recommended as heart-friendly. In 2002, sugary items were still recommended as heart-healthy snacks because they were free of saturated fats. Recommendation of this in the direction of sugar consumption quickly became obsolete when in 2009 the WHO recommended restricting sugar intake. Moreover, the most recent recommendations advise a restriction of not only added sugars, but also refined grains. Such dietary recommendations provide a starting point to an optimum diet for preventing dental caries and improving oral health. Studies investigating the relationship between oral diseases and diet are limited. To date the relationship between oral health and diseases focused on the identification of the dietary effects on caries formation [135–137] underlining how fermentable sugar consumption could led to the formation of caries and on the protective role of protein consumption and reduced vegetable fats assumption on the development of periodontal diseases [138,139]. No studies investigated the relationship between dietary habits and peri-implant diseases.

In this chapter we have tried to advance the field in this topic by shaping the relationship between dietary habits and peri-implant oral health. Such analysis was carried out for two main reasons: 1) To evaluate the effects of dietary habits on peri-implant diseases; 2) To evaluate the potential capabilities to predict the disease state from dietary habits.

3.2 Materials and methods

3.2.1 Introduction to the PreBiomics Beta Program

A pilot study has provided evidence of the link between microbiome composition and peri-implantitis diseases [31]. Main results i) identified a site-specific definite microbial firm characterising the subgingival plaque of implants affected by peri-implantitis; ii) identified *Fusobacterium nucleatum* as a keystone coloniser in the intermediate condition of mucositis; iii) assessed the accuracy of ML-based classification models for the identification of peri-implant diseases; iv) detected an uncharacterized sub-species of *F. nucleatum* as significantly associated with peri-implant diseases.
Such research has been expanded by PreBiomics thanks to the BetaProgram with the aim of developing a commercial kit for peri-implant diseases based on the microbiome composition for diagnostic and prognostic purposes.

In this BetaProgram study (still ongoing) the aim is to collect 2,000 implant subgingival plaque samples from more than 40 Italian dental clinics along with a rich set of clinical variables. Specific to my doctoral activity, there is also the acquisition of dietary habits through FFQ questionnaire and their characterization towards health statuses and microbiome composition.

3.2.2 Subject recruitment and collection of clinical variables

For the purpose of this thesis, we sampled 451 subjects from 28 Italian dental clinics. Inclusion criteria of the patients involved in the study were:

- Being at least 18 years of age;
- Not being pregnant;
- Having at least one dental implant;
- Having not used local or systemic antibiotics in the two weeks before the sampling;
- Not being affected by acquired immune deficiency syndrome.

For each patient, it was collected by clinicians a rich set of clinical metadata about patient clinical history and patient condition at the time of sampling:

- Demographic data: sex, age, weight, and height;
- Medical history: smoking habit, diabetes, autoimmune diseases or other systemic diseases, alcohol consumption, and medications taken;
- Dental history: current and past periodontal status, number of remaining teeth, number of implants, previous peri-implantitis, frequency of home oral care, hours since last toothbrush, and chlorhexidine usage. Clinical parameters included implant, site of sampling, diagnosis of implant age (time from installation), implant system used and nature of reconstruction (single implant, fixed or removable), type of implant retention (screw, cement, conometric), radiographic peri-implant bone loss, width of the keratinized mucosa, as well as peri-implantitis probing depth (PPD), plaque index (PI), bleeding on probing (BOP), and suppuration (SUP). The latter four parameters were measured in each patient at the buccal, mesial, lingual, and distal sites of the experimental implant. PI, BOP, and SUP were recorded on a binary scale (presence/absence) for each surface and PPD was measured to the nearest millimetre on the scale. In case of mucositis and peri-implantitis, any eventual subsequent therapy was noted.

All patients were anonymized in the clinic by assigning a unique subject ID to each subject. Downstream analyses were performed using the anonymized metadata.

Each sampled subject was included in three different groups, depending on her/his disease phenotype and oral health status: healthy, mucositis, peri-implantitis. To avoid biases across clinics, meetings were organised to instruct dentists on a common protocol for the examination, collection, and measurement procedures. Follow-up meetings were organised after 6 months to ensure consistency of the sampling and the inclusion criteria of the three disease phenotypes.

During my twenty months stay at PreBiomics I contributed to the acquisition of the metadata information by checking and reporting errors in their collection. This task was accomplished by constant monitoring of the data collected by dentists, by generating weekly reports, and by producing and sending warning emails. Finally, meetings and phone calls were organised to collect opinions and critical issues identified by the dentists.

3.2.3 Dietary data collection and characteristics of the FFQ approach

Analytical studies about diet composition and its link with health status needs information about daily dietary intakes on an individual basis [140]. In literature, different methods of daily dietary intakes have been assessed. The nature of the data itself and the collection techniques, that generally relies on estimates made by the interviewed, are the cue of differences between the collected data and the true daily dietary intake ranging from 4% to 400% [141]. Even if the amount and the kind of food consumed varies between subjects, people rarely perceive what they eat and how much they do [141]. Among the available dietary assessment methods, the food frequency questionnaire (FFQ) has been widely used in large epidemiological studies. A FFQ is an advanced form that asks respondents how often and how much they eat over a specific period. A food frequency questionnaire (FFQ) is considered semi-quantitative if the instrument addresses both the frequency and the amount of each food item consumed [142]. The advantages of a food frequency method are: i) standardisation of the questionnaire; ii) simple automation of the method; iii) cheapness of the method; iv) lack of influence on eating behaviours. FFQs should be developed specifically for each study group and research purposes because diet is influenced by multiple factors including ethnicity, culture, individual's preference, economic status, etc. [143].

3.2.4 Acquisition and processing of the EPIC questionnaires

Along with clinical metadata, we acquired daily dietary intakes from subjects involved in the study. We relied on a validated FFQ questionnaire designed by EPIC (The European Prospective Investigation into Cancer and Nutrition), one of the largest cohort studies in the world, with more than half a million (521 000) participants recruited across 10 European countries and followed for almost 15 years [144]. The EPIC FFQ was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors, and the incidence of cancer and other chronic diseases. EPIC provided an Italian version (written in Italian and representative of the more common Italian foods) of the validated FFQ with the possibility to fill the questionnaire through a web platform.

The EPIC FFQ is composed of 164 questions. Along with an introductory survey, questions are organised in different sections related to 16 main food categories: dry sauce pasta; soups; meat; fish; raw vegetables; potatoes and cooked vegetables; eggs; sandwiches; cured meat and appetiser; cheeses; fruit; bread and wine; coffee, milk and sweets; spices; soy products and whole cereals; cooking methods). Questions range in three types: i) the usual dimension of the portion for a given food; ii) the frequency of consumption of a given food; and iii) the preferences of different foods belonging to the same food category (e.g., for the “Pasta” category preferences in the consumption of “pasta with tomato sauce”, “white pasta”, “pasta soup”, etc. were asked).
A fac-simile of the questionnaire is available in S3.1 Fig.

The original plan to acquire questionnaires directly in the clinics through tablets was reshaped due to COVID-19 pandemic restrictions. We provided a paper version of the questionnaire to each subject, which was therefore filled at home, gave back to the clinic, and put into the web platform by us. We administered a total of 1,157 questionnaires (involving 40 dental clinics) and got back 451 questionnaires from 28 clinics.

The web platform processed raw data provided by users to generate three refined data products: i) food daily intakes; ii) weekly frequency in food consumption; iii) daily intakes of micro and macro nutrients.

3.2.5 Generation of additional dietary indexes

In addition to the three data types furnished by the EPIC platforms, we computed a new set of indexes to better understand the level of fitness of the subjects to the quality of their eating habits. More specifically, we computed two well-recognized indexes: 1) the MIDI index [145]. This score is based on intakes of typical Mediterranean foods ranging from 0 to 11 (where 11 reflects a “perfect” Mediterranean diet). The score is increased when the considered food is in a given range as follows: i) pasta (at least 73 grams/day), ii) vegetables (at least 162 grams/day), iii) potatoes (less than 17 grams/day), iv) fruits (at least 392 grams/day), v) legumes (at least 23 grams/day), vi) fish (at least 38 grams/day), vii) red meat (less than 70 grams/day), viii) olive oil (at

least 30 grams/day), ix) butter (less than 1.7 grams/week), x) sugared/carbonated beverages (0 grams/day), xi) alcohol (less than 12 grams/day); 2) the HDI index [146]on the dietary recommendations for the prevention of chronic diseases provided by the WHO study group [147]. It is in the range between 0 and 9 (where 9 reflects a “healthy” diet) and composed of these nine components: i) saturated fatty acids (less than 10 % of energy intake), ii) polyunsaturated fatty acids (between 3% and 7% of energy intake), iii) proteins (between 10% and 15% of energy intake), iv) simple sugars (less than 10% of energy intake), v) complex sugars (between 50% ad 75% of energy intake), vi) dietary fibre (between 27 and 40 grams/day), vii) fruit and vegetables (at least 400 grams/day), viii) legumes nuts and dried fruit (at least 30 grams/day), ix) cholesterol (less than 300 milligrams/day).

We finally calculated the percentage of daily energy intakes coming from the different nutrients categories: carbohydrates, lipids, and proteins. This decomposition was made to check adherence of diet habits to the Nutrient Reference Intake Levels suggested by the Italian society of human nutrition (*Società italiana di nutrizione umana* S.I.N.U.) in the LARN tables proposed in 2014.

3.2.6 Strategies to deal with missing data in clinical variables

Despite efforts in getting clinical variables in an exhaustive way, some missing values were present in the data collected by dentists. This was especially true for those variables for which more than one measurement was required (e.g., for the peri-implantitis probing depth dentists were asked to take different measurements in four different sites: buccal, mesial, lingual, and distal site). In multivariate analyses, the naive strategy of dropping observations with missing values would lead to a sensible reduction in the sample size, so we investigated alternative solutions to deal with this issue: i) substitution of the missing value with the mean value of the considered feature (defined as substitution in the rest of the text); ii) iterative imputation of the missing value by estimating its value as a function of other features (defined as imputation). This was obtained through the sklearn.impute.IterativeImputer python package. Iterative imputation may generate negative values for some clinical features that do not allow negative values for definition. As an example, the peri-implantitis probing depth represents the depth in millimetres of the peri-implant pocket and cannot

assume negative values being a measure of length). Features after imputation were considered only when building classification models.

3.2.7 Ordination techniques for multivariate analysis

Multivariate analysis was conducted by considering two popular ordination techniques such as principal component analysis (PCA) and principal coordinates analysis (PCoA). PCA was computed through the sklearn.decomposition.PCA python package, while PCoA using the skbio.stats.ordination.pcoa python package. For PCoA we considered and compared four distance metrics: i) Euclidean distance; ii) Manhattan distance; iii) Jaccard distance; and iv) Bray-Curtis distance.

3.2.8 Generation of the dietary habits report

We generated and provided a dietary habits report to each person involved in the study to give him/her information about the healthiness of his/her dietary habits. From the processed dietary data, we automatically generated the report in PDF format using the python module pyfpdf2, a library for simple and fast PDF document generation. The report was structured in four main sections:

- an introductory section that gives information about the data processor and the origin of the data. We also provided here a warning message to suggest not to use the information given in the report as an indicator to change dietary habits since the report is based on information obtained through a self-assessment of the eating habits. We also reported the percentage of completeness of the questionnaire, calculated as the ratio between the number of answered questions and the number of questions on the questionnaire;
- the MIDI index [145]. In this section, the MIDI index is introduced in terms of meaning, way of calculation and foods categories involved in the calculation. The score is provided along with the foods categories for which the person showed an out-of-range consumption;
- the HDI index [146]. This section was structured as the previous one related to the MIDI index;
- the SINU LARN table. In this section, we reported the contribution in percentage terms to the daily energy intakes for each macronutrient: lipids, sugars, and

proteins. We also compared the percentage composition with the values suggested as healthy intakes by the SINU.

A fac-simile of the dietary report is available in S3.2 Fig.

3.2.9 Finding correlations between clinical/dietary features and health status

We identified associations between clinical/dietary features and health status by computing Spearman correlation coefficients among them. Correlations were computed by representing the study condition value as 0 (for healthy), 1 (for mucositis), and 2 (for peri-implantitis). This was possible due to the nature of the health status variable that can be considered as a categorical ordered variable. Correlations were also computed for the different binary settings (i.e., healthy vs peri-implantitis, healthy vs mucositis, mucositis vs peri-implantitis). For such binary cases we also computed point biserial correlation coefficients. Moreover, we identified the features differently abundant among the three health statuses by using Mann-Whitney U (for continuous variables) and Fisher exact (for categorical variables) tests. All p-values were adjusted for multiple hypothesis testing with false discovery rate (FDR) using the statsmodels.stats.multitest.fdrcorrection python module. FDR correction was applied separately for the three dietary data types (i.e., food daily intakes, weekly frequency of food consumption, and daily nutrient intakes of nutrients) and the clinical variables. Q-values < 0.05 identified statistically significant variables. The obtained results were compared with the clinical characterisation of the peri-implant diseases available in literature [35,36,148–150].

3.2.10 Dealing with multicollinearity in dietary data

Quantities and frequencies of consumption generated by the processing of the FFQ questionnaires are multicollinear by definition. The multicollinearity is due to the fact that some variables are representative of subcategories. For example, the “pasta” macro category is also reported as “dry sauce pasta”, “pasta soup”, “white pasta”, etc. This can bring distortions in the downstream statistical analyses. We dealt with this issue by adopting the variance inflation factor (VIF) approach through the python module statsmodels.stats.outliers_influence.variance_inflation_factor. We skimmed

features having VIF < 10 [151–153]. Multicollinearity corrected data were considered to further identify differentially abundant dietary data based on logistic regression and linear discriminant analysis as described in the next paragraphs.

3.2.11 LDA on clinical and dietary data

We identified the variables characterised by the largest effect sizes in function of the health status through a linear discriminant analysis (LDA) implemented in the LefSe tool [147]. LDA is a generalisation of Fisher's linear discriminant to find a linear combination of variables that separates two or more classes. LDA was applied on clinical and dietary data separately, with the latter one pre-processed in advance to remove multicollinearity issues as described in the previous section.

3.2.12 Identification of possible confounding factors in dietary data

We identified possible confounding factors in dietary data by looking at beta diversity differences in function of sex, patient age and BMI score. This was accomplished by applying a permutational multivariate analysis of variance (PERMANOVA) test through the python module skbio.stats.distance.permanova. We considered Euclidean distance and 999 permutations.

For sex, which resulted in the only variable with statistically significant differences, we also identified the discriminative features by considering a Mann-Whitney U test with FDR correction Q-values <= 0.05 identified statistically significant variables.

3.2.13 Logistic regression on clinical and dietary data

We applied logistic regression on both clinical and dietary data to identify the effect of the variation of variables on the probability to belong to each study group (i.e., healthy, mucositis, and peri-implantitis). We considered the python module sklearn.linear_model.LogisticRegression. We considered the different binary settings of health status (i.e., healthy vs peri-implantitis, healthy vs mucositis, and mucositis vs peri-implantitis) as dependent variables. Health conditions were coded as an ordered categorical variable by substituting “0” and “1” to the condition in order of condition severity. For dietary data, the logistic regression was applied on the multicollinearity-corrected data as described previously.

Evaluation of the portion of variation in the dependent variable predictable from the independent variables was made by computing the Efron's pseudo $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \underline{y})^2}$, where *n* is the number of observations, *y* is the dependent variable, $\underline{y}$ is the mean of the *y* values, and $\hat{y}$ is the value predicted by the model. Statistical significance was evaluated by computing the χ^2 test for each involved variable and by adjusting using the FDR correction. Q-values < 0.05 identified statistically significant variables. Evaluation of the effects of the significant variables was made by looking at the odds ratio: $OR = \frac{odds(x+1)}{odds(x)} = e^{\beta_1}$, where β_0 and β_{1x} are the linear regression coefficients for a given variable and the odds are the outputs generated by the model.

3.2.14 Generation of prediction models through ML-based classification

We considered a ML-based classification approach to build classification models for different tasks and multiple data types. The classification tasks were performed using the already developed and validated Metagenomic prediction Analysis based on Machine Learning (MetAML) tool [89]. More specifically, we considered Random Forests (RFs) as classification algorithm, which was the best classifier identified in Chapter 2 of this thesis.

All analyses were performed through a cross-validation approach. In cross-validation, samples were randomly divided into k (with k = 10 in our case) folds by considering a stratified cross-validation approach to preserve the percentage of samples of each class. Free parameters of the classifier were set as follows. i) the number of trees was set to 500, ii) the number of features to consider when looking for the best split was equal to the root of the number of original features, and iii) the Gini impurity criterion was used to measure the quality of a split. Results were repeated and averaged on 20 independent runs.

Evaluation of the classification performances was based on these main metrics: i) area under the curve (AUC), ii) area under the precision-recall curve (AUPRC), iii) precision, iv) recall and v) F1.

The classification process on the two different clinical datasets (clinic substitution data and clinic imputation data) was conducted considering the three health conditions at once and the different paired combinations of health statuses (healthy vs peri-implantitis, healthy vs mucositis and mucositis vs peri-implantitis).
The classification approach was considered to predict the health status from the three dietary data separately (i.e., quantities of consumption, frequency of consumption, and daily intakes of micro and macro nutrients). As a baseline, we also built models to predict the health status from the clinical variables.

3.3 Results and discussion

3.3.1 Overview of the sampled cohort

The study involved 451 subjects enrolled in the BetaProgram and all having at least one dental implant. This cohort was recruited from 28 dental clinics distributed among 10 Italian regions: Trentino Alto Adige (N = 9 clinics); Veneto (N = 5); Lombardy (N = 4); Friuli Venezia Giulia (N = 2); Marche (N = 2); Campania (N = 1); Emilia Romagna (N = 1); Sicily (N = 1); Piedmont (N = 1); Apulia (N = 1); Lazio (N = 1). Subjects involved in the study are evenly distributed between male (46%) and females (54%). Age ranged from 23 to 85 with a mean value of 60.

3.3.2 Multiple clinical variables are associated with the health status

Subjects were grouped in three main categories according to the health of their dental implants: i) healthy (subjects with at least one healthy implant and no diseased implants; 34.5%), ii) mucositis (patients with at least one implant with mucositis and no implants with peri-implantitis; 31.5%), and iii) peri-implantitis (patients with at least one implant with peri-implantitis; 33.7%). Inclusion of subjects in one of the groups was based on characteristics that are normally adopted by dentists to diagnose mucositis and peri-implantitis according to the criteria delineated by the Consensus Report on Peri-implant Diseases [33]: marginal bone level, clinical signs of inflammation, presence of suppuration.

We evaluated to which extent clinical variables were associated with the health status. Among the 37 variables, we assessed statistically significant differences ($p < 0.05$, ANOVA test) in 26 cases (Table 3.1). Such variables can be grouped in three main groups: i) variables representative of personal data, which resulted non-significant apart from the patient age (Fig. 3.1); ii) variables representative of the main clinical signs of the disease state (i.e., bleeding on probing; plaque index; suppuration; bone loss; peri-implantitis probing depth), which resulted to be all significant (Fig 3.2); iii) variables not strictly correlated with the disease state but potentially related to the general health conditions. Here we identified as significant the smoking state ($p = 0.010$) and the average number of cigarettes per day ($p = 0.024$), which are largely recognized as risk factors for peri-implant diseases [154–158] and overall involved in oral health [159–163]. Interestingly, no significance was instead achieved in terms of years after quitting smoking. We also detected significant other variables less recognized in the literature such as the number of dental implants, the year of implantation, the number of teeth left, and having precedent episodes of peri-implantitis (Fig 3.3). Finally, occasional use of alcohol was not related to the health status.

We extended the analysis to the identification of differentially abundant features in the binary combinations of health statuses through statistical testing using Mann-Whitney U test for continuous features and Fisher exact test for categorical features. We found that the clinical data showed a number of significant features going from 29 to 18 for the healthy vs peri-implantitis and the mucositis vs peri-implantitis case respectively. The results obtained by the tests show the same pattern encountered in computing the correlations in terms of statistical significance. A summary of the p-value obtained by the tests is available in Table 3.3.

Table 3.1 Summary of the statistical tests aiming at discriminating the health status from the clinical variables. P-values are obtained through ANOVA test and FDR correction. Q-value <= 0.05 denotes a statistically significant variable for which we indicate the class of enrichment. Clinical data are grouped as personal, peri-implantitis related, and health status related

Personal data		
Variable	**P-value**	**Higher in**
patient age when visited	0.010	peri-implantitis
patient sex	0.855	
patient height	0.680	
patient weight	0.261	
sampled element	0.158	
Peri-implantitis related data		
Variable	**P-value**	**Higher in**
ppd b	0.000	peri-implantitis
ppd m	0.000	peri-implantitis
ppd d	0.000	peri-implantitis
ppd l	0.000	peri-implantitis
bone loss m	0.000	peri-implantitis
bone loss d	0.000	peri-implantitis
plaque index b	0.000	peri-implantitis
plaque index m	0.000	peri-implantitis
plaque index d	0.000	peri-implantitis
plaque index l	0.000	peri-implantitis
bleeding on probing b	0.000	mucositis
bleeding on probing m	0.000	mucositis
bleeding on probing d	0.000	mucositis
bleeding on probing l	0.000	mucositis
suppuration b	0.000	peri-implantitis
suppuration m	0.000	peri-implantitis
suppuration d	0.000	peri-implantitis
suppuration l	0.000	peri-implantitis
Health status related data		
Feature	**P-value**	**Higher in**
already had peri-implantitis	0.000	peri-implantitis
avg cigarettes per day	0.024	peri-implantitis
avg weekly sport activity	0.238	
electronic toothbrush	0.680	
mucous thickness b	0.991	
mucous thickness l	0.934	
number of dental implants	0.000	peri-implantitis
number of teeth left	0.000	healthy
occasional use of alcohol	0.795	
smoking state	0.010	peri-implantitis
width km b	0.003	healthy
width km l	0.934	
year of implantation	0.000	healthy
years quit smoking	0.795	

Table 3.2 P-values of Mann-Whitney U and Fisher-exact tests on clinical features. Summary of the Mann-Whitney U and Fisher-exact tests p-values after FDR correction relative to the clinical features. The features were tested in the three possible paired combinations of health status (HVSP, MVSP, HVSM). The p-values were filtered for p-value <= 0.05, an empty table cell means the non-significant difference in the feature distribution between the two tested groups. All: three-class scenario with the three health status categories; hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

Features	hvsp	hvsm	mvsp
already had periimplantitis	5E-09	4E-03	3E-03
avg cigarettes per day	2E-03		
bleeding on probing b	7E-27	3E-18	
bleeding on probing d	3E-29	2E-25	
bleeding on probing l	5E-27	4E-12	1E-04
bleeding on probing m	3E-26	8E-16	2E-02
bone loss d	2E-39	1E-06	7E-25
bone loss m	3E-42	1E-05	2E-29
mucous thickness b	9E-03	5E-03	2E-02
mucous thickness l	2E-04	1E-03	5E-04
number of dental implants	5E-10	6E-05	2E-02
number of teeth left	5E-15	2E-04	3E-04
patient age when visited	3E-03		
plaque index b	1E-11	4E-06	
plaque index d	1E-07	2E-06	
plaque index l	4E-10	4E-06	
plaque index m	2E-08	5E-06	
ppd b	6E-35	3E-12	3E-15
ppd d	1E-34	3E-12	9E-17
ppd l	1E-32	4E-10	2E-13
ppd m	4E-33	3E-12	3E-15
sampled element	4E-02		
smoking state	2E-03		
suppuration b	4E-18	2E-02	3E-12
suppuration d	8E-11		6E-07
suppuration l	2E-06		2E-04
suppuration m	6E-14		4E-10
width km b	1E-03	5E-02	
width km l	1E-05	1E-06	3E-03

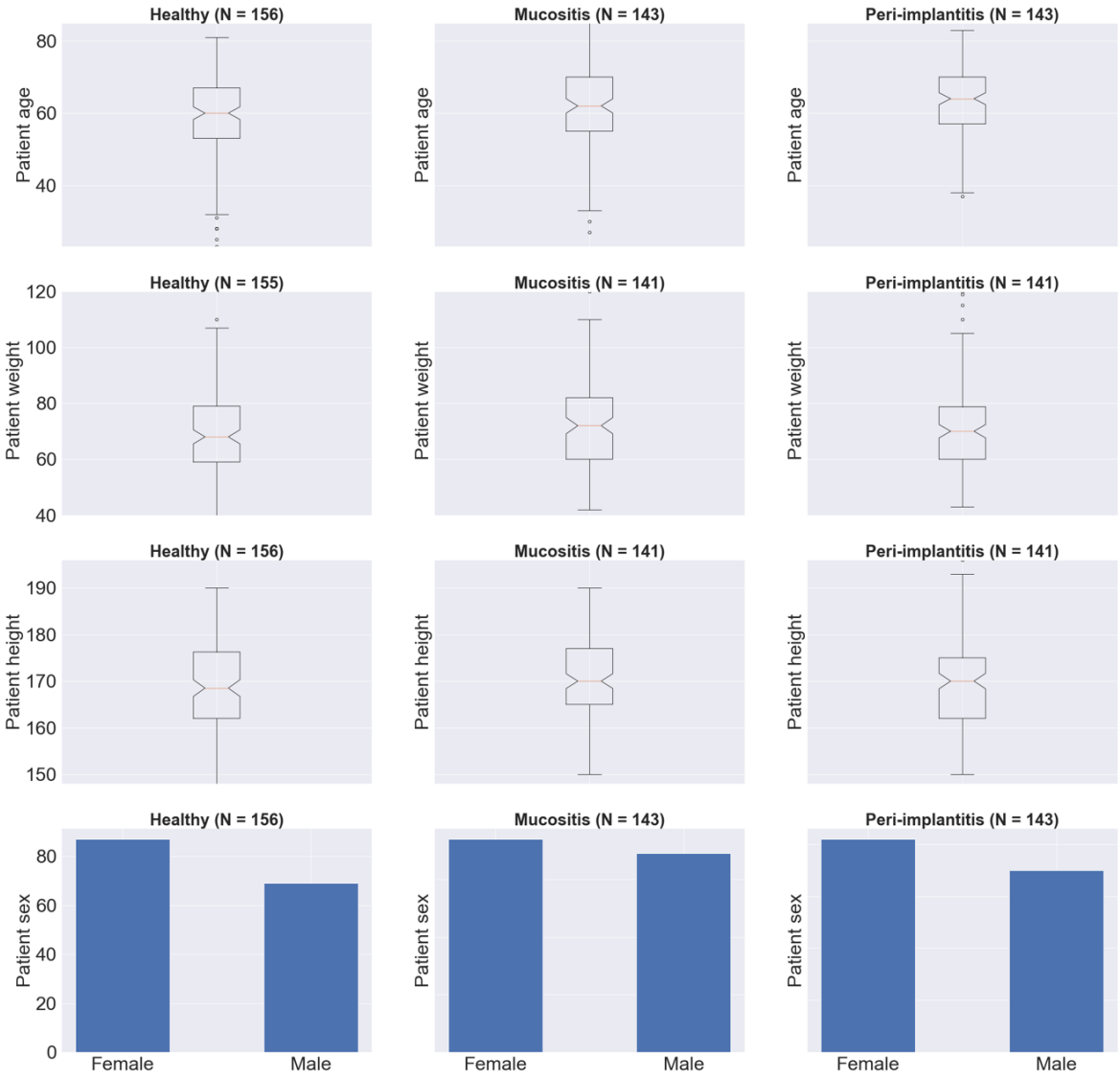


Fig 3.1 Differences in the clinical variables in terms of personal data among the three health study groups. Only statistically significant variables according to ANOVA test are reported (see Table 3.1 for the complete list of p-values).

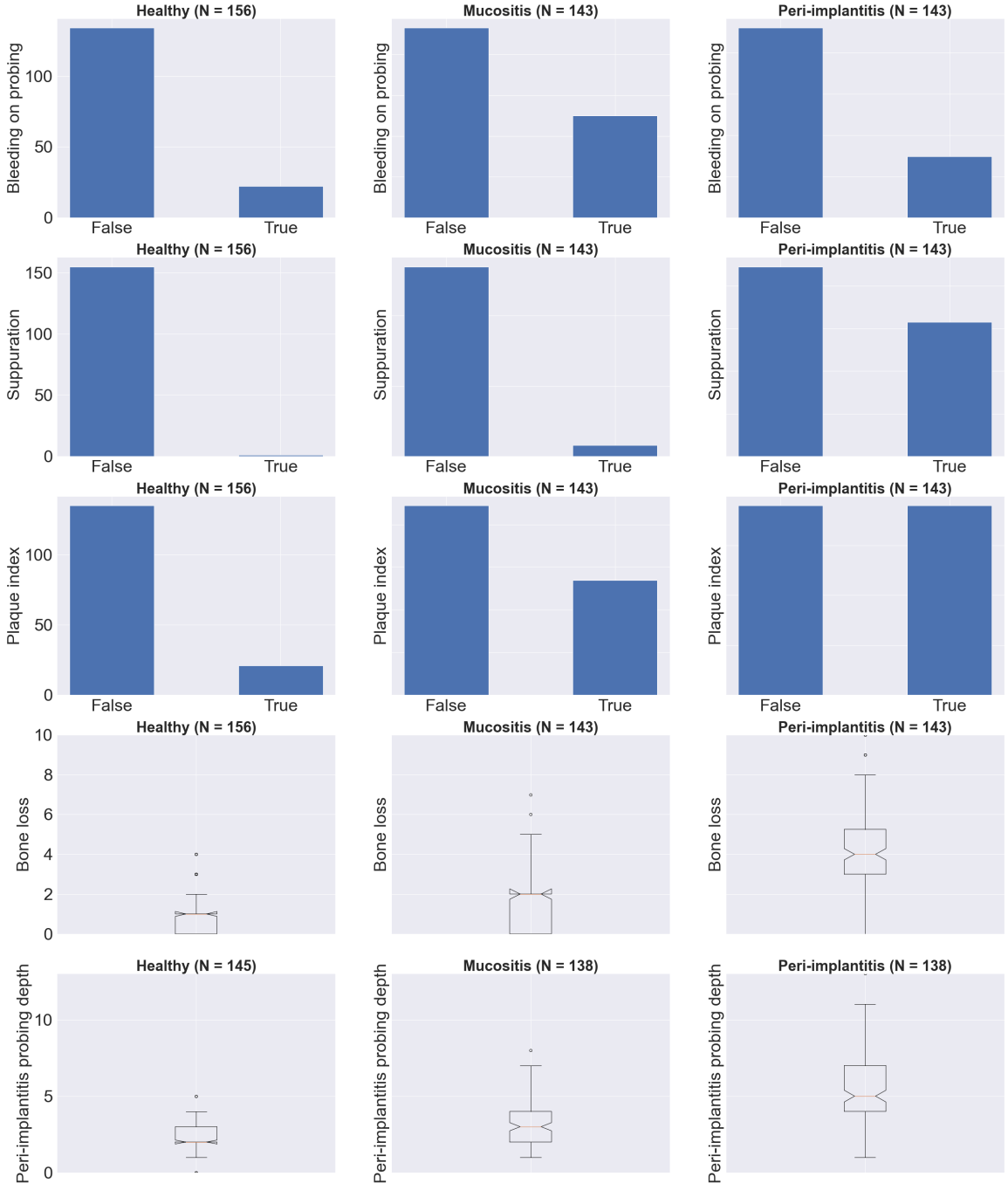


Fig 3.2 Differences in the clinical variables in terms of peri-implantitis related data among the three health study groups. Only statistically significant variables according to ANOVA test are reported (see Table 3.1 for the complete list of p-values).

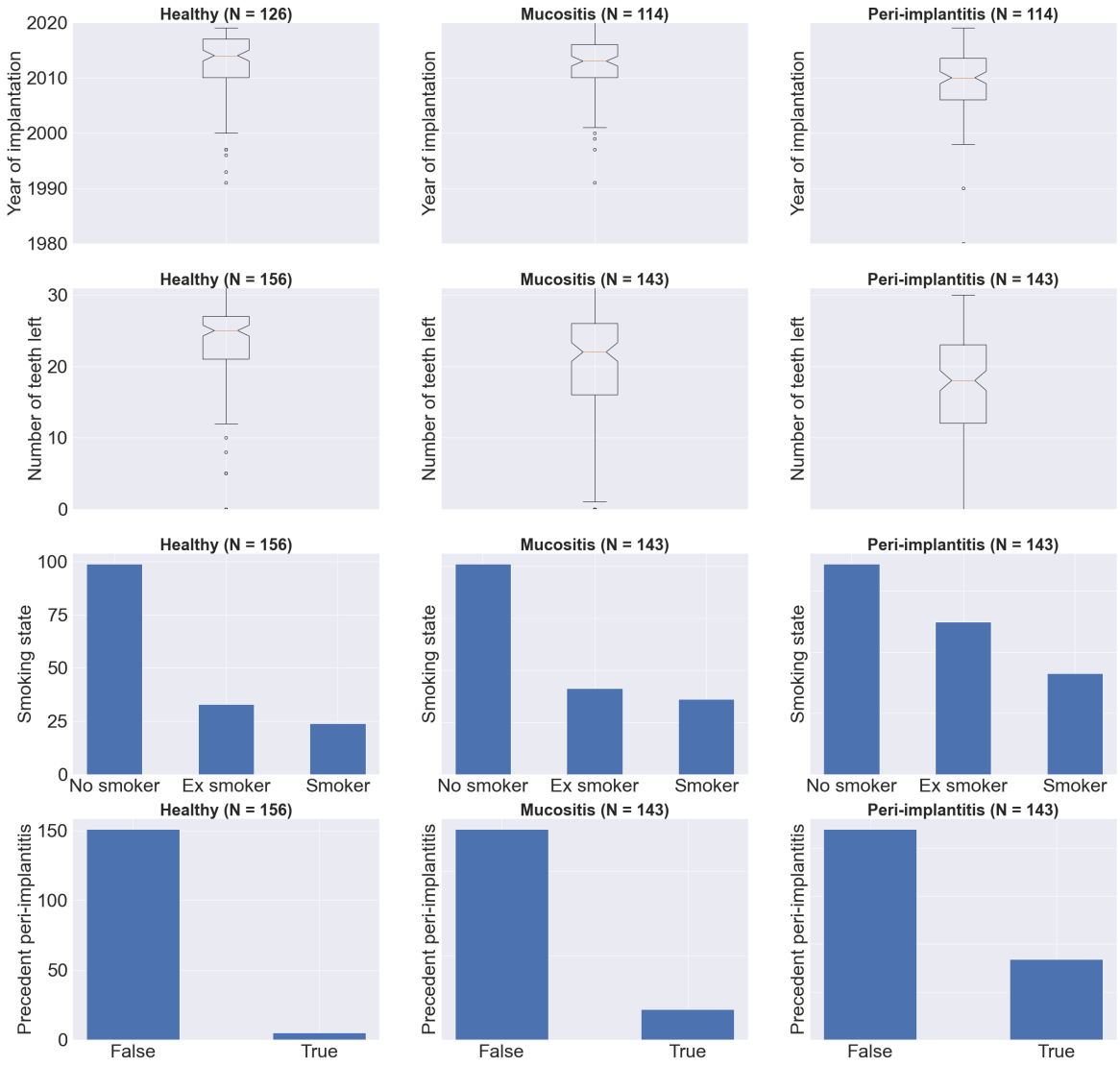


Fig 3.3 Differences in the clinical variables in terms of general health status related data among the three health study groups. Only statistically significant variables according to ANOVA test are reported (see Table 3.1 for the complete list of p-values).

3.3.3 Associations between clinical features and health status are confirmed by correlation analyses

We extended the analysis discussed in the previous section by computing Spearman and point biserial correlations between clinical variables and the health status.. Results confirmed what was discussed in the previous section, with several clinical variables showing associations with the health status. (Table 3.3). We found 27 out of the 37 clinical variables correlated significantly with the health status. More specifically, 25 and 2 were enriched in disease and health, respectively. We found such variables enriched in diseased subjects, which were already found in the literature as clinically relevant for the peri-implantitis disease status [150,164,165]: already had peri-implantitis (representative of the peri-implant clinical story of the patients); bleeding on probing for all the four sites of measurement (representative of the gingival bleeding during the sampling procedure); bone loss for the two measurement sites (representative of the severity of the gingival bone loss); plaque index for all the four measurement sites (representative of the presence of subgingival plaque); suppuration for all the four measurement sites (representative of the presence of suppuration).

Additionally, we found other variables higher in disease subjects which were not strictly representative of peri-implant but more related to the general oral health: number of dental implants (representative of a proxy of the general health status); patient age when visited (representative of the age of the patients in the moment of the clinical evaluation; age is a recognized risk factor for peri-implantitis [150,164–166]); average cigarettes per day (representative of the average number of cigarettes smoked in a day; also smoke is a risk factor for peri-implantitis [150,154,167]); more general smoking state (representative of the appartenance to the smoker group).
On the other hand, we found few variables enriched in the healthy subjects:

- number of teeth left (as a proxy of the health status);
- width km (representative of the width of the keratinized mucosa in the buccal site; recognized as an implant health indicator [168]);
- year of implantation (representative of the age of the implant).

We further confirmed Spearman correlation results by considering logistic regression to compute the effect of variation of the clinical features on the probability to belong to one of the three health statuses. Correctness of the model was evaluated in terms of Efron's R^2, while variable Statistical significance was computed by χ^2 tests. Results obtained by logistic regression (Table 3.4) were inline with findings obtained by Spearman correlation. More specifically we found an agreement in terms of statistically significant variables in 65% of the cases. The proportion of the variance for the health status that was explained in the regression model by the clinical variables was high in all settings. More specifically, the models trained on the clinical features showed a large portion of variance explained by the dependent variable for the three binary combinations healthy vs peri-implantitis (R^2= 0.87); mucositis vs peri-implantitis (R^2=0.57); and healthy vs mucositis R^2=0.49.

We finally identified the the effect size of clinical features in discriminating the three health categories by considering LDA implemented into the LefSe tool [169]. Results confirmed previous findings based on statistical testing and correlation coefficients when the three health conditions are considered simultaneously (Fig 3.4) as well for the binary combinations (i.e., healthy vs peri-implantitis Fig 3.5; healthy vs mucositis Fig 3.6; and mucositis vs peri-implantitis Fig 3.7).
Results achieved by the different approaches were concordant in finding strong associations between multiple clinical variables and the health status, with some significant variables already recognized in the literature as well as other more related to the general oral health.

Table 3.3 Point biserial and Spearman correlation coefficients between clinical features and the health status. The health status was codified as a numerical ordered categorical feature with values assigned in order of condition severity. Empty cells represent correlation coefficients associated with non-significant correlations (p-value after FRD correction > 0.05). All: three-class scenario with the three health status categories; hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

Correlation							
coefficient	**Point biserial**			**Spearman**			
case	**hvsp**	**hvsm**	**mvsp**	**all**	**hvsp**	**hvsm**	**mvsp**
feature							
already had periimplantitis	0.340	0.176	0.187	0.287	0.340	0.176	0.187
avg cigarettes per day	0.179			0.150	0.182		
bleeding on probing b	0.619	0.523	0.116	0.511	0.619	0.523	0.116
bleeding on probing d	0.648	0.623		0.533	0.648	0.623	
bleeding on probing l	0.622	0.414	0.241	0.512	0.622	0.414	0.241
bleeding on probing m	0.612	0.483	0.154	0.514	0.612	0.483	0.154
bone loss d	0.723	0.306	0.588	0.668	0.762	0.297	0.617
bone loss m	0.744	0.283	0.637	0.691	0.792	0.264	0.674
number of dental implants	0.331	0.200	0.145	0.307	0.361	0.241	0.154
number of teeth left	-0.412	-0.232	-0.168	-0.374	-0.454	-0.223	-0.226
patient age when visited	0.181			0.141	0.171		
plaque index b	0.393	0.280	0.123	0.321	0.393	0.280	0.123
plaque index d	0.305	0.288		0.253	0.305	0.288	
plaque index l	0.364	0.279		0.298	0.364	0.279	
plaque index m	0.325	0.274		0.268	0.325	0.274	
ppd b	0.664	0.433	0.460	0.632	0.711	0.415	0.467
ppd d	0.668	0.395	0.495	0.629	0.699	0.401	0.495
ppd l	0.620	0.362	0.427	0.589	0.676	0.364	0.432
ppd m	0.652	0.378	0.468	0.612	0.682	0.396	0.472
sampled element	0.119			0.099	0.119		
smoking state	0.173		0.123	0.148	0.178		0.122
suppuration b	0.503	0.145	0.419	0.458	0.503	0.145	0.419
suppuration d	0.305	0.117	0.305	0.346	0.378	0.117	0.305
suppuration l	0.231		0.231	0.258	0.274		0.231
suppuration m	0.377	0.122	0.377	0.406	0.435	0.122	0.377
width km b				-0.158	-0.188	-0.123	
year of implantation	-0.221		-0.221	-0.276	-0.326		-0.237

Table 3.4 Odds ratio obtained by considering the logit model on the clinical variables and the health status of the subject Empty cells are associated with odds ratio non-statistically significant (p-value >= 0.05 after FDR correction). The health status is codified as a numerical ordered categorical variable with values assigned in order of condition severity. hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

Features	odds peri-implantitis HVSP	odds peri-implantitis MVSP	odds mucositis HVSM
avg cigarettes per day	0.989		1.004
bleeding on probing b	1.069		1.137
bleeding on probing d	1.079		1.181
bleeding on probing l	1.075	1.023	1.109
bleeding on probing m	1.075	1.015	1.136
bone loss d	1.347	1.208	1.166
bone loss m	1.426	1.315	1.145
mucous thickness b	0.992	0.998	0.998
mucous thickness l	1.011	0.991	0.995
number of dental implants	1.004		
occasional use of alcohol	1.025	1.018	1.017
patient height	0.998		
plaque index b	1.040		1.052
plaque index d	1.033		1.068
plaque index l	1.038		1.063
plaque index m	1.039		1.065
ppd b	1.342	1.112	1.275
ppd d	1.310	1.125	1.249
ppd l	1.281	1.101	1.231
ppd m	1.307	1.106	1.260
smoking state	0.997		
suppuration b	1.038	1.040	1.012
suppuration d	1.020	1.019	
suppuration l	1.013	1.012	
suppuration m	1.032	1.037	
width km b	0.937	0.966	0.942
width km l	0.988	1.010	0.988
years quit smoking	0.965		

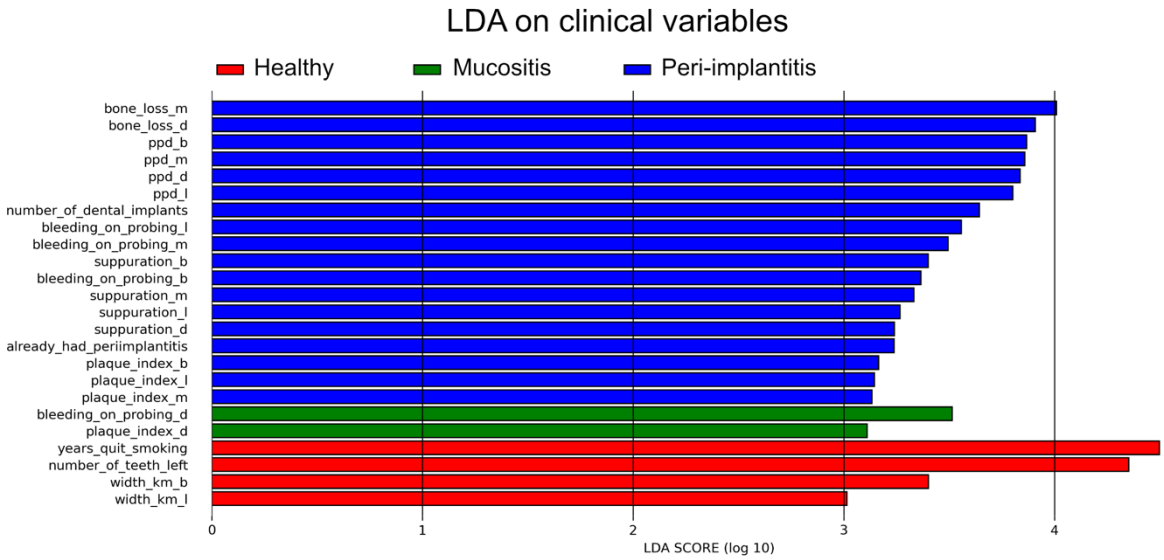


Fig 3.4 Variables associated with the largest effect size by applying LDA on the clinical features by considering simultaneously the three health statuses. Only variables having LDA score >= 2 are reported. The three colours are associated with the category of enrichment.

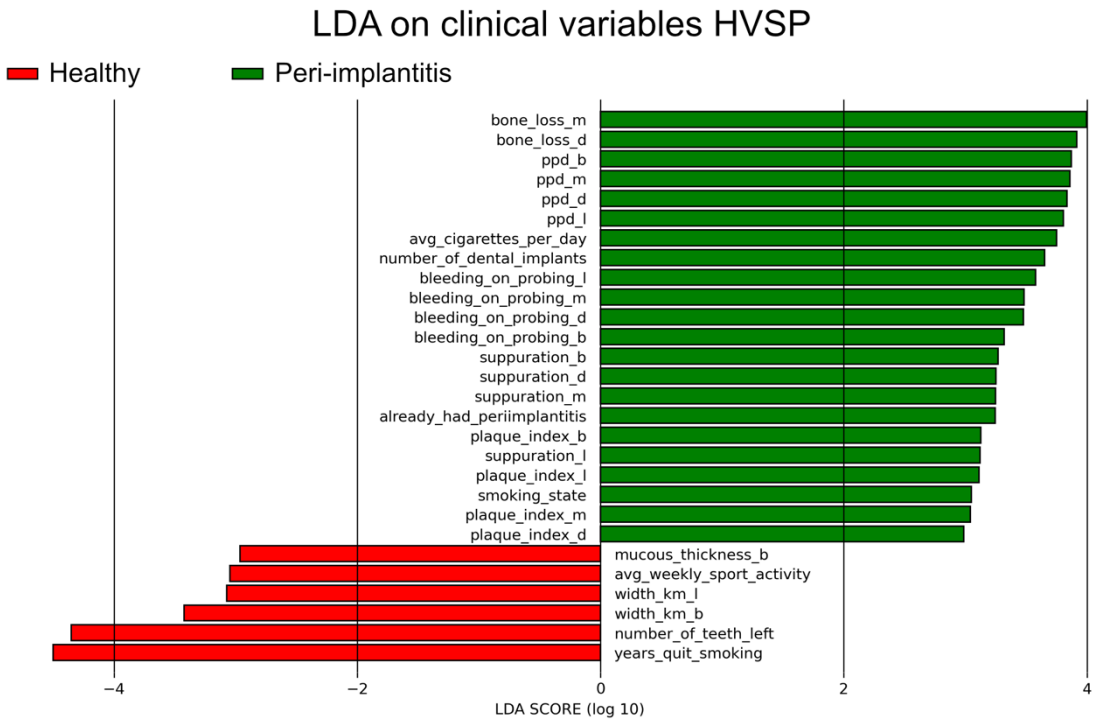


Fig 3.5 Largest effect sizes obtained by LDA on the clinical features by comparing healthy with peri-implantitis subjects. Effect sizes were filtered for LDA score >= 2. The two colours are associated with the category of enrichment.

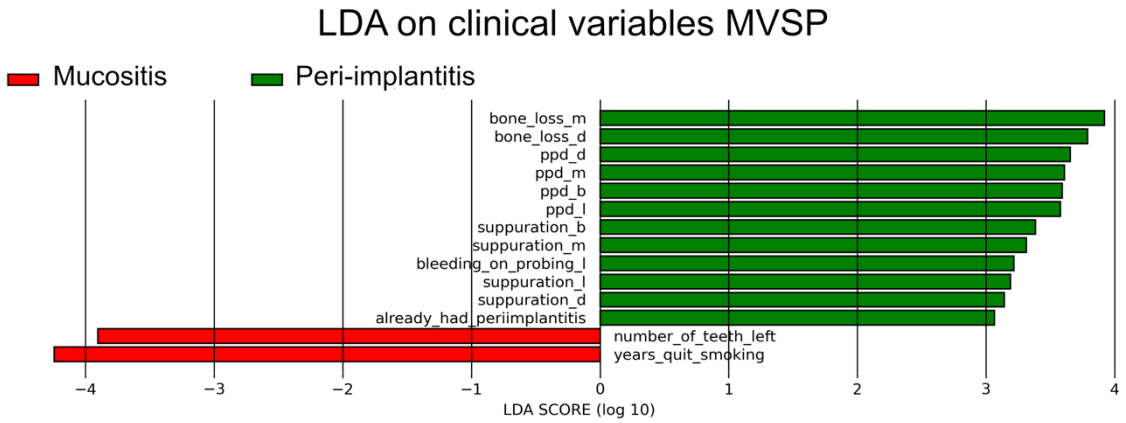


Fig 3.6 Largest effect sizes obtained by LDA on the clinical features by comparing healthy vs peri-implantitis subjects. Effect sizes were filtered for LDA score >= 2. The two colours are associated with the category of enrichment.

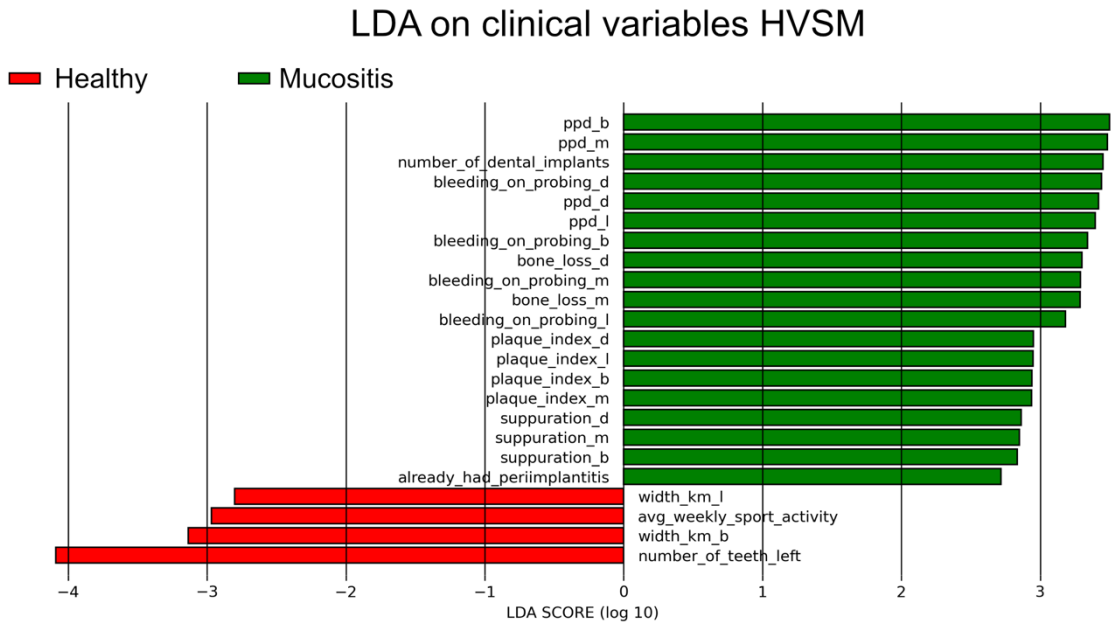


Fig 3.7 Largest effect sizes obtained by LDA on the clinical features by comparing combination of health statuses healthy with mucositis subjects. Effect sizes were filtered for LDA score >= 2. The two colours are associated with the category of enrichment

3.3.4 Evaluation of the quality of the FFQ questionnaires

We extended the analysis to the dietary data which were acquired from the exact set of subjects (N = 451) thanks to administration of FFQ. Due to the self-administration of the FFQ, some of the questionnaires were returned incomplete with some unfilled questions or sections. 133 out of 451 questionnaires (29.5%) were complete, while mean value of completeness was equal to 92%; 16.5% of the questionnaires resulted in a completeness value < 80% (Fig 3.8).

We evaluated to which extent the completeness level of the returned questionnaires was influenced by two main demographic factors such as age and sex, along with the information about the dental clinic. No differences were related to the sex of the patients ($p = 0.503$, Mann-WhitneyU test; Fig 3.9), while significant differences were assessed in terms of age ($p = 0.012$, ANOVA test; Fig 3.9). More specifically, mean completeness ranged from 99% for young people (< 30 years old) to 90% for elderly people (> 80 years old) which was in agreement with expected behaviour and existing literature [170,171]. Significant differences were also obtained in terms of dental clinics ($p = 0.000$, ANOVA test; Fig. 3.9) which was likely related to the differences in providing support in filling the questionnaires by the different dental clinics.

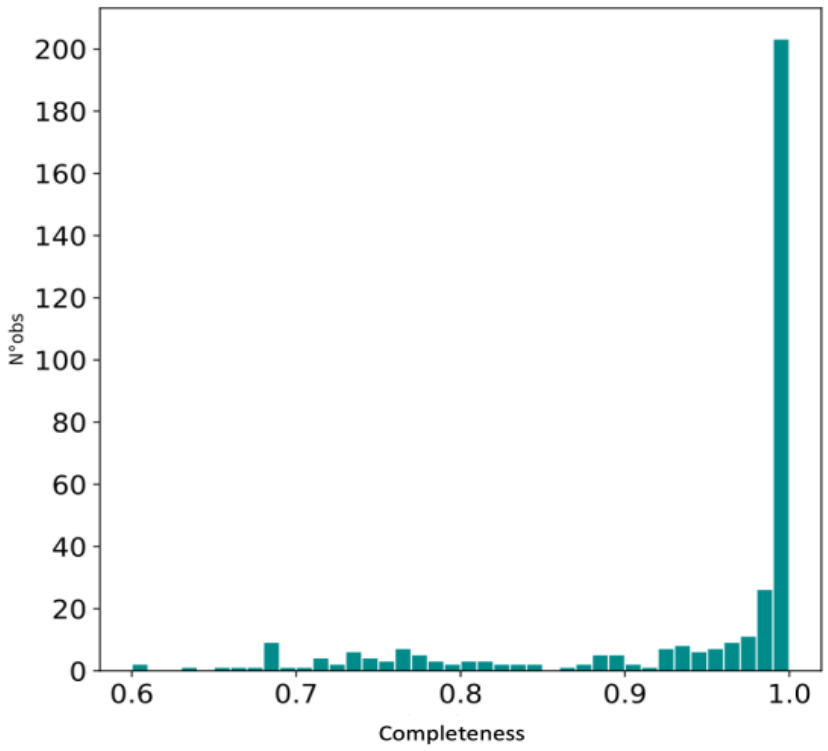


Fig 3.8 Histogram with the level of completeness for the FFQ questionnaires that was filled by the 451 subjects included in this analysis.

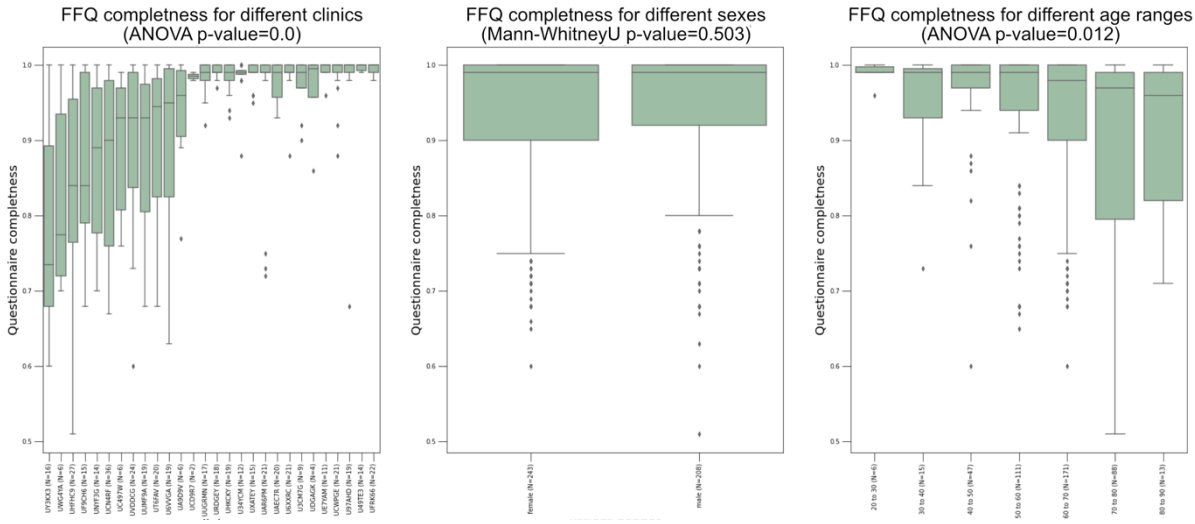


Fig 3.9 FFQ questionnaire completeness across different demographic characteristics. Statistically significant levels of completeness were assessed across different clinics (panel A; p = 0.0 ANOVA test) and age ranges (panel C; p = 0.012 ANOVA test), while no differences were obtained in terms of sex (panel B; Mann-Whitney U test).

3.3.5 Ordination analysis highlights patterns related to dietary indexes

We applied ordination analysis to evaluate potential differences in the composition of the different dietary datasets across multiple demographic data (i.e., age, sex, and BMI), health status, and dietary indexes (i.e., MIDI and HDI). We considered two main techniques, PCA and PCoA, and for the latter ones we compared four distance metrics: i) Euclidean, ii) Manhattan, iii) Jaccard; and iv) Bray-Curtis. For the different tested scenarios, we obtained comparable results across the four distance metrics (see an example in Fig. 3.10), so we report only the results using Euclidean distance in the following analyses for conciseness. All the analyses were repeated on the three main dietary data types, i.e., quantities, frequencies, and nutrients.

We evaluated strong patterns in terms of the two main dietary indexes involved in the analysis, i.e., MIDI and HDI. This was something expected since the indexes are directly computed from the dietary data. Subjects characterised by different MIDI and HDI scores showed differences in dietary habits in terms of quantities (Fig 3.11 A,B; Fig 3.12 A,B), frequencies (Fig 3.13 A,B; Fig 3.14 A,B), and daily intakes of micro and macronutrients (Fig 3.15 A,B; Fig 3.16 A,B).

A similar analysis was conducted in terms of age. In this case no differences were verified in terms of quantities (Fig 3.11 C; Fig 3.12 C), frequencies (Fig 3.13 C; Fig 3.14 C) or daily intakes of micro and macro nutrients (Fig 3.15 C; Fig 3.16 C) across different ages. This may be related also to the fact that the majority of the individuals are elderly (age median value equal to 62), while major shifts in dietary habits in the direction of uniformity happen between 40 and 60 years old [172,173]. In the same way, also differentiations in terms of BMI were not evident across the three data types: Fig 3.11 D and Fig 3.12 D for quantities; Fig 3.13 D and Fig 3.14 D for frequencies; Fig 3.15 D and Fig 3.16 D for daily intakes of micro and macro nutrients). This may be also due to the simplicity in calculating the BMI score which does not take into account other important factors linked to the health status of the individuals such as metabolism, level of physical activity, genetics, drug administration, etc. that can induce shifting in subject weight without being correlated to dietary habits [174–176].

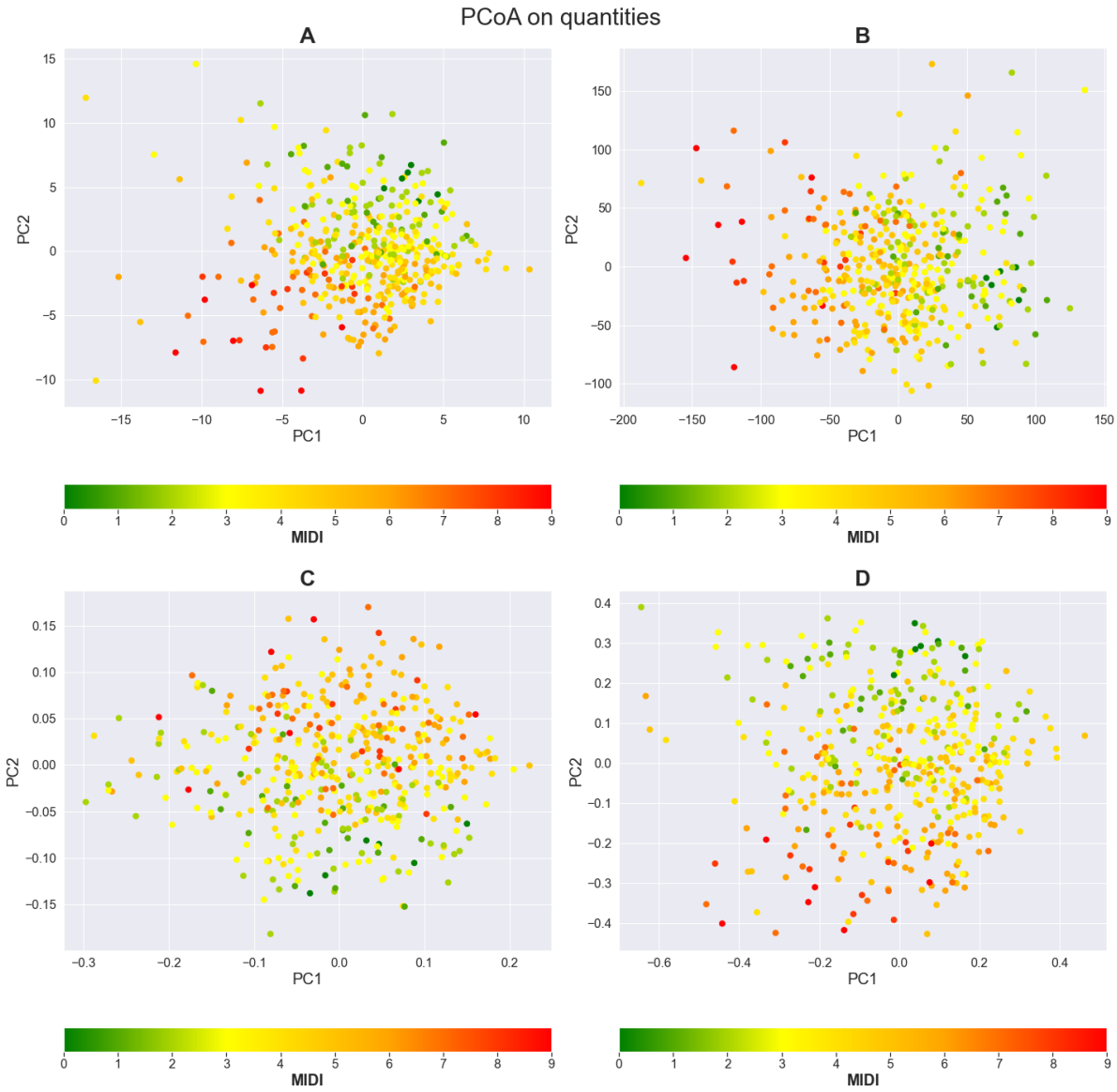


Fig 3.10 PCoA applied on dietary data in terms of quantities at varying distance matrices and coloured by MIDI. Four distances are considered: (A) Euclidean; (B) Manhattan; (C) Jaccard; and (D) Bray-Curtis.

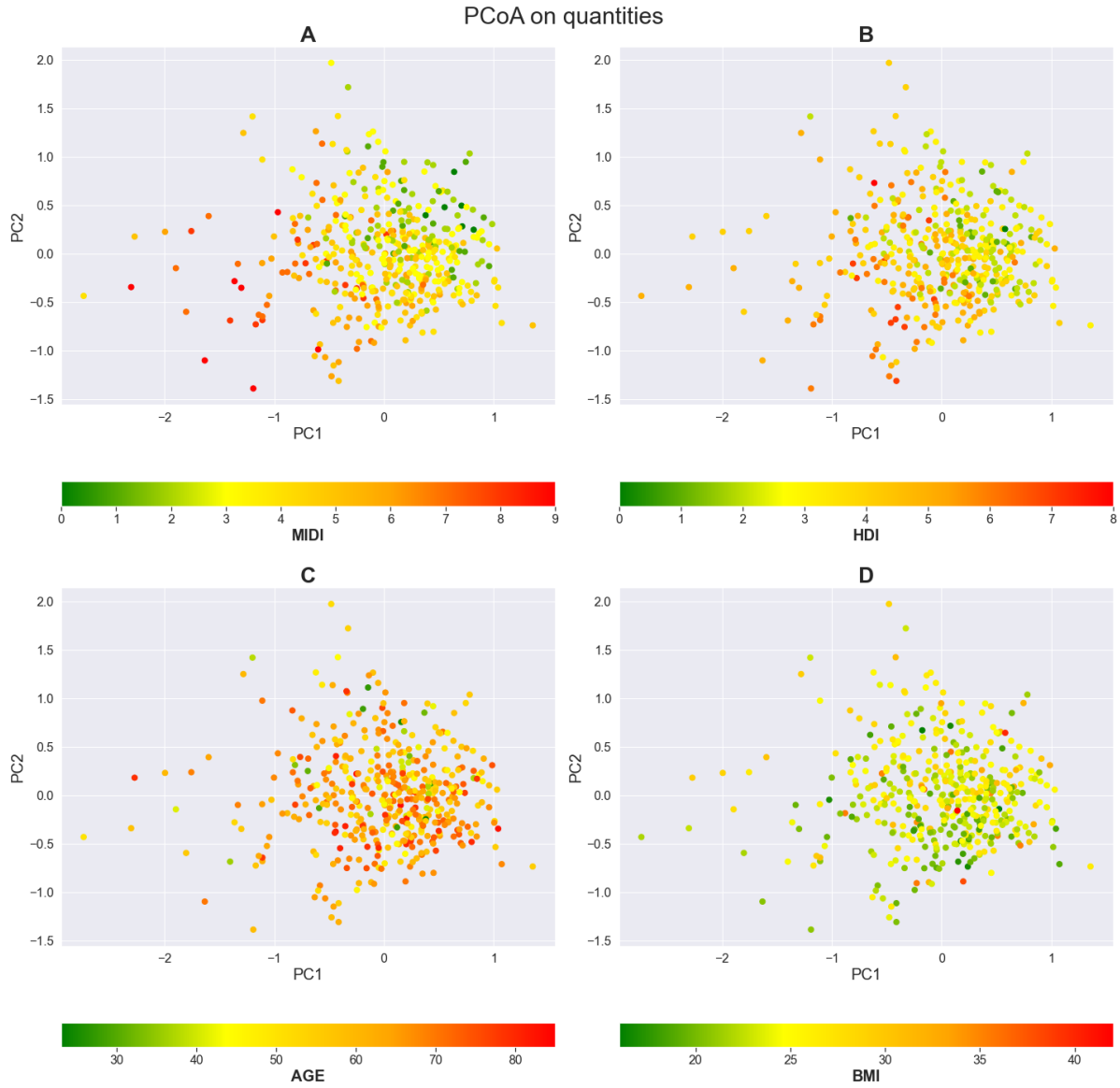


Fig 3.11 PCoA applied on dietary data in terms of consumption quantities coloured by different subject characteristics. Four subject characteristics are considered: A) MIDI score; B) HDI score; C) subject age; D) BMI score. Results are obtained using Euclidean distance.

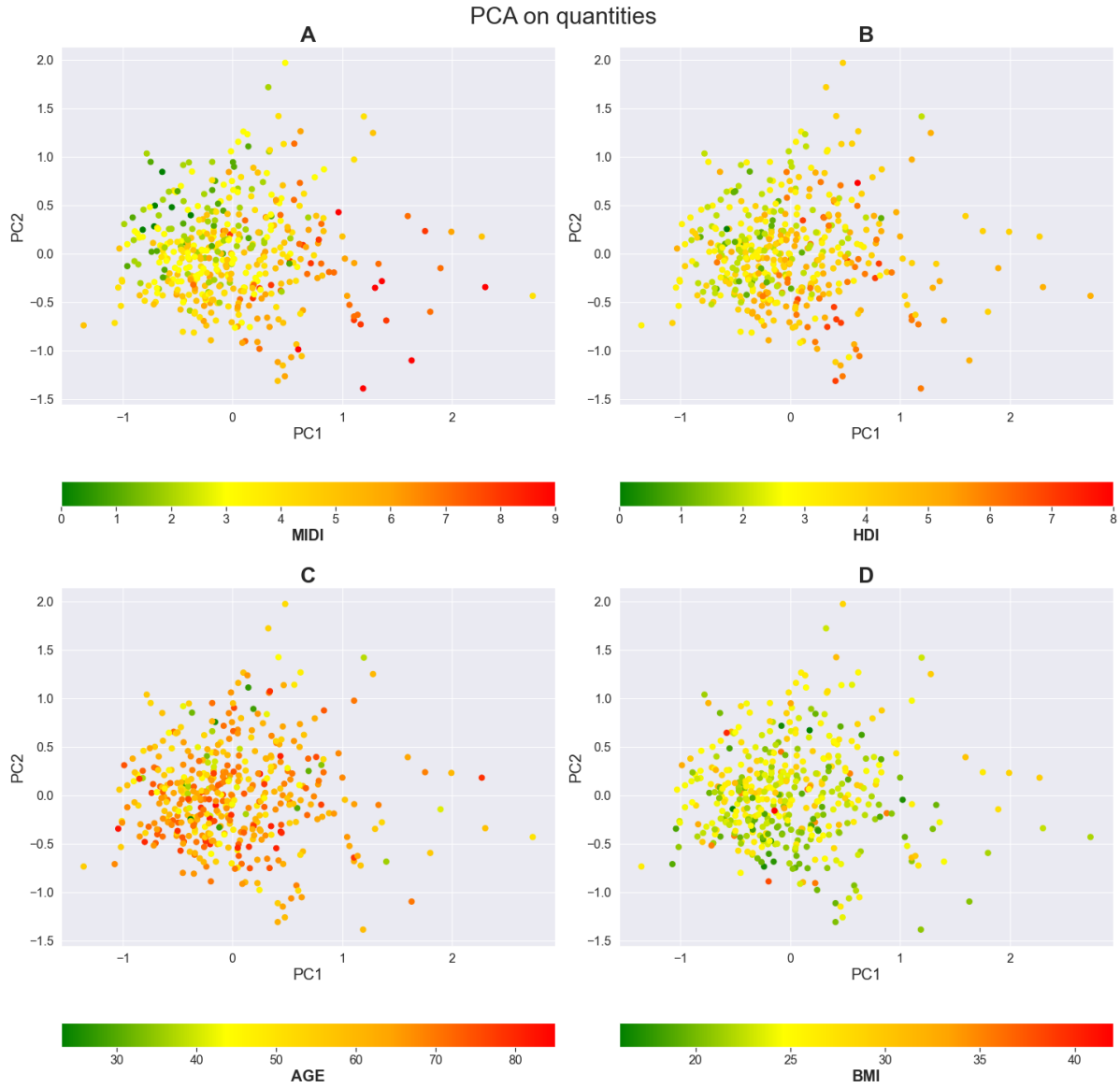


Fig 3.12 PCoA applied on dietary data in terms of consumption frequencies coloured by different subject characteristics. Four subject characteristics are considered: A) MIDI score; B) HDI score; C) subject age; D) BMI score. Results are obtained using Euclidean distance.

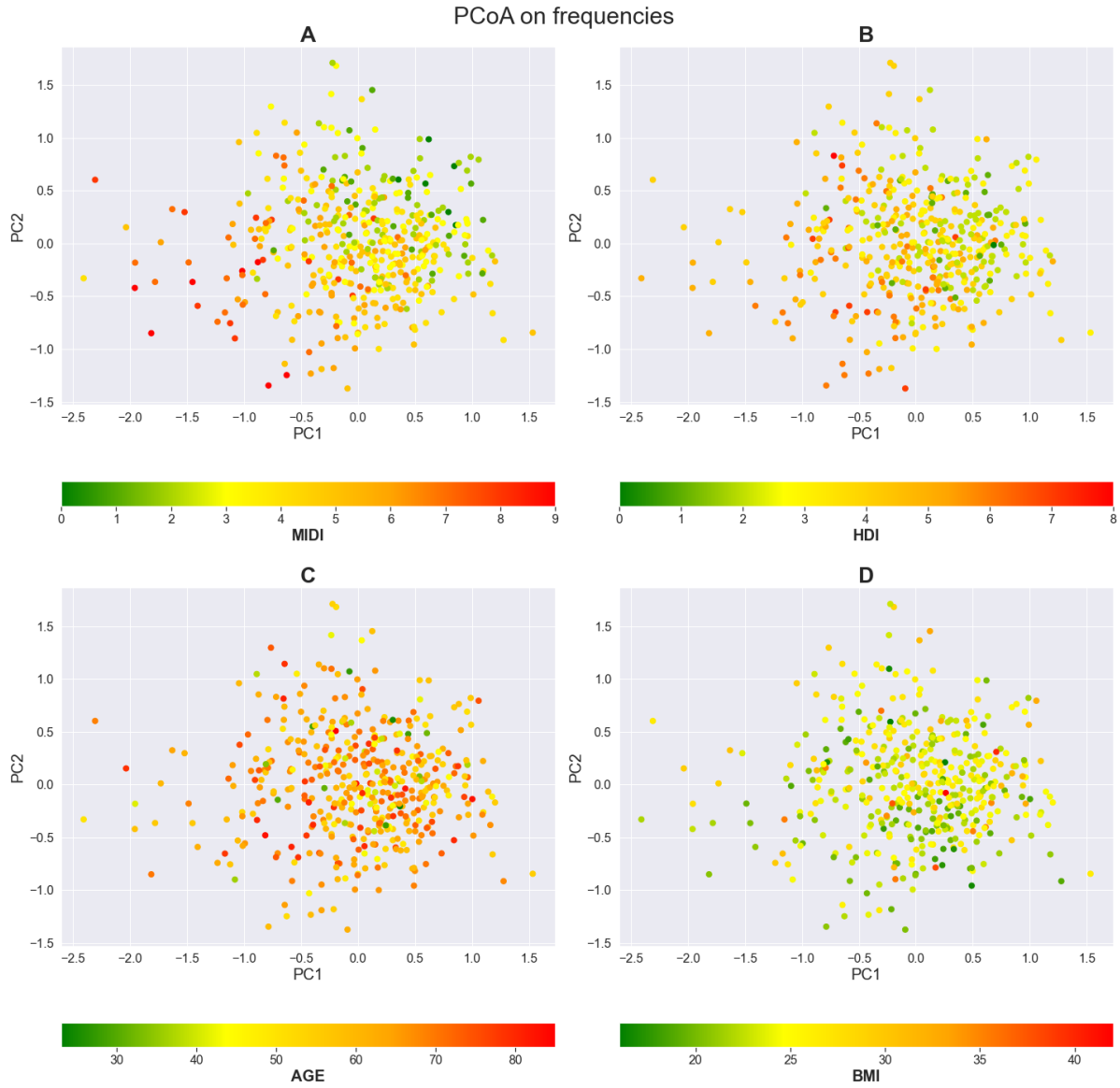


Fig 3.13 PCoA applied on dietary data in terms of daily intakes of micro and macro nutrients coloured by different subject characteristics. Four subject characteristics are considered: A) MIDI score; B) HDI score; C) subject age; D) BMI score. Results are obtained using Euclidean distance.

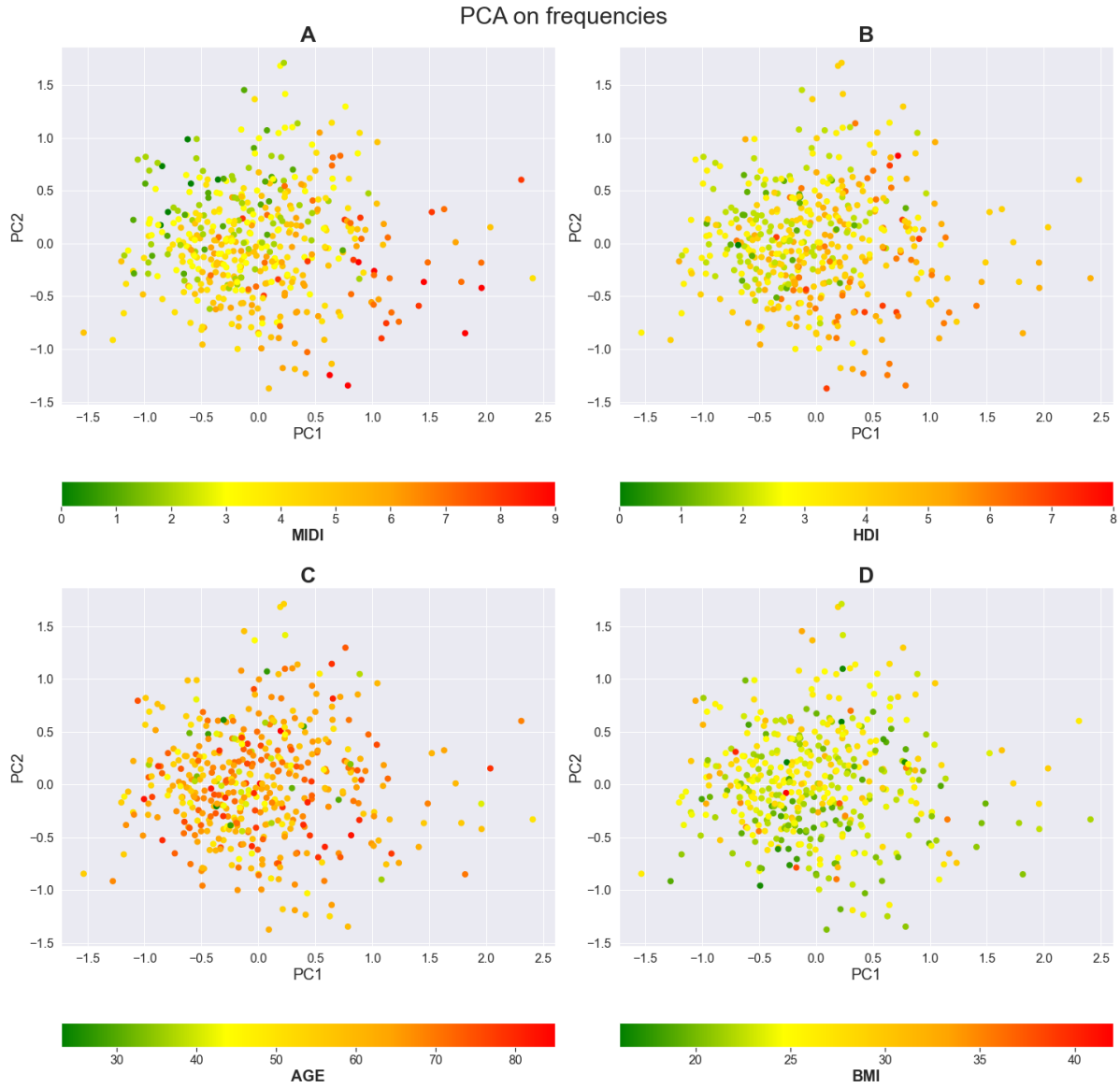


Fig 3.14 PCA applied on dietary data in terms of consumption quantities coloured by different subject characteristics. Four subject characteristics are considered: A) MIDI score; B) HDI score; C) subject age; D) BMI score.

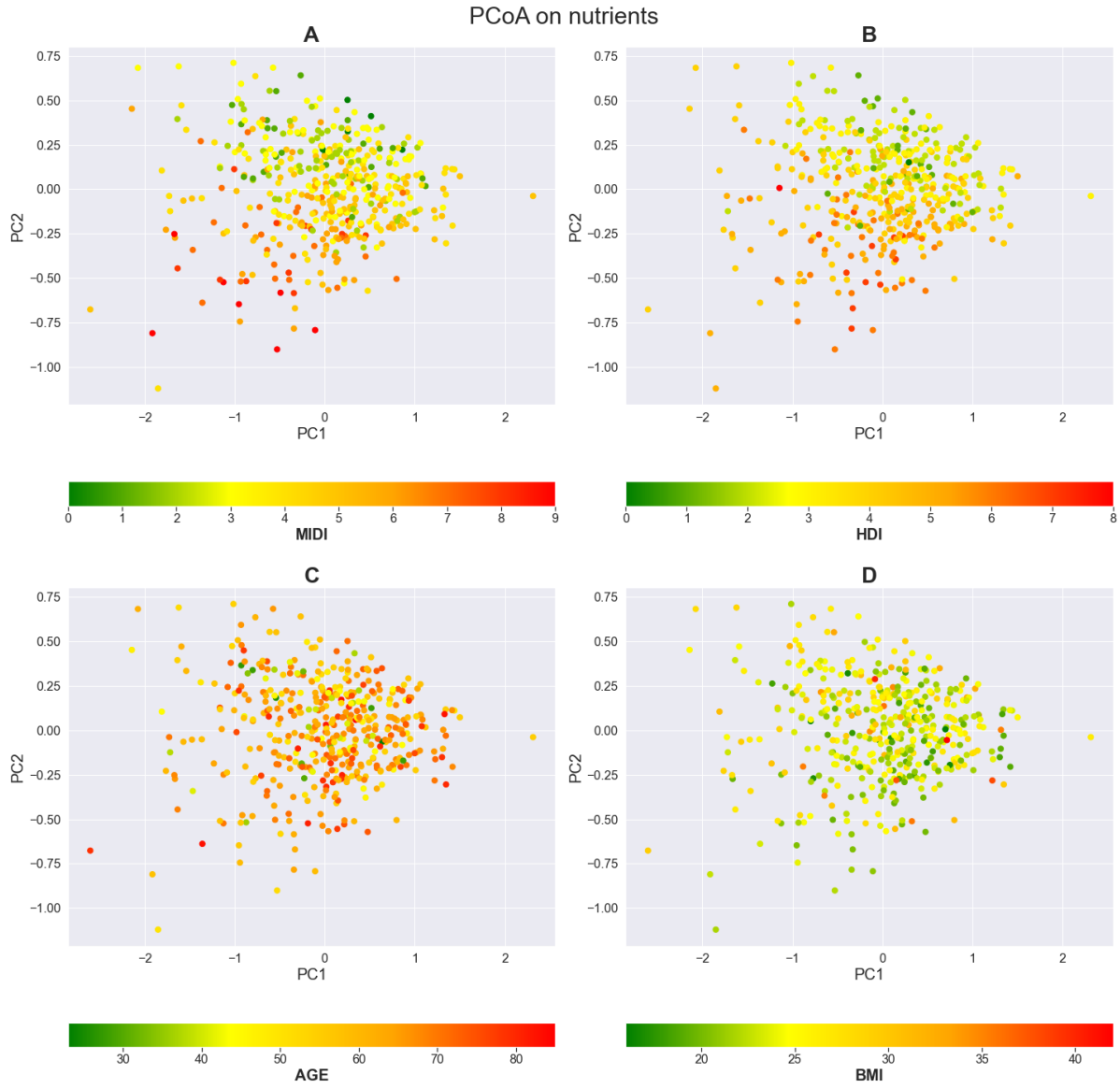


Fig 3.15 PCA applied on dietary data in terms of consumption frequencies coloured by different subject characteristics. Four subject characteristics are considered: A) MIDI score; B) HDI score; C) subject age; D) BMI score.

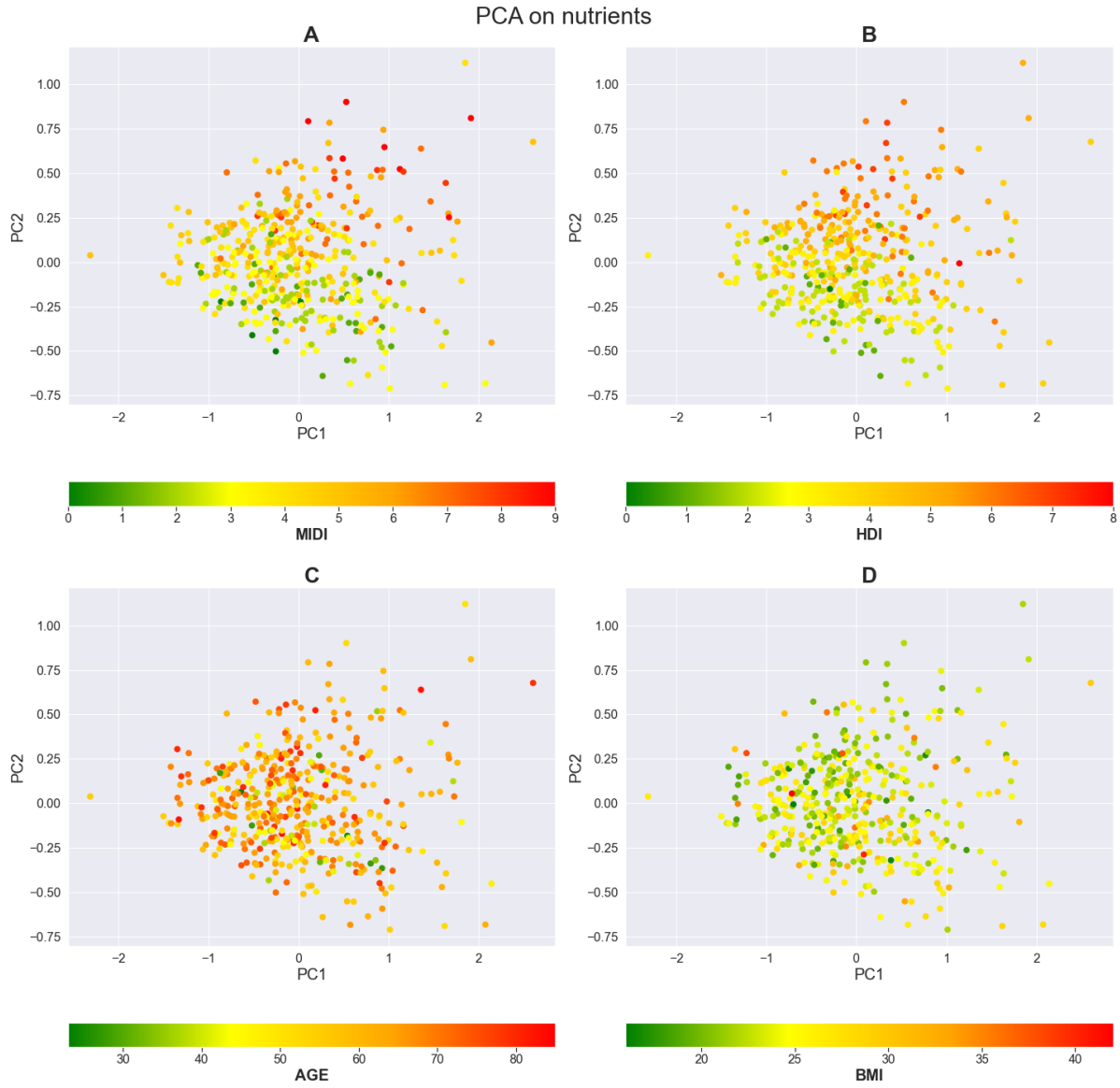


Fig 3.16 PCA applied on dietary data in terms of daily intakes of micro and macro nutrients coloured by different subject characteristics. Four subject characteristics are considered: A) MIDI score; B) HDI score; C) subject age; D) BMI score.

3.3.6 Sex of the subjects as a possible confounding factor in dietary habits

While we did not find patterns in function of subject age and BMI as described in the previous section, strong differences were instead found based on the sex of the subjects enrolled in the study. Such differences were consistent using both PCA and PCoA techniques and across the three dietary data types: quantities (Fig 3.17 A,D; $p < 0.001$ PERMANOVA Table 3.5), frequencies (Fig 3.17 B,E; $p < 0.001$ PERMANOVA), and daily intakes of micro and macro nutrients (Fig 3.17 C,F; $p < 0.001$ PERMANOVA). This is consistent with known eating styles between the two sexes, with women that generally have eating styles characterised by healthier choices [177].

We went further by identifying differentially distributed variables between the two sexes through Mann-Whitney U test. We found 28 and 27 food categories with different distribution in terms of quantities (Table 3.5) and frequencies (Table 3.6), respectively, between the two sexes. We also detected 14 daily intakes of micro and macro nutrients differently distributed between the two groups (Table 3.7). More specifically, food categories enriched in males were mainly sources of simple fermentescible sugars and fats, consistent with what reported in the literature [178,179]. Instead, categories enriched in females comprised mainly vegetables, in both cooked and raw forms, as already reported in the literature [180].

The differences in dietary habits between males and females may act as a confounding factor and could contribute to misidentifying the relationship among variables involving dietary data. However, we note the two sexes are evenly distributed among the three study groups (Table 3.8), which should reduce biases in other analyses [181,182].

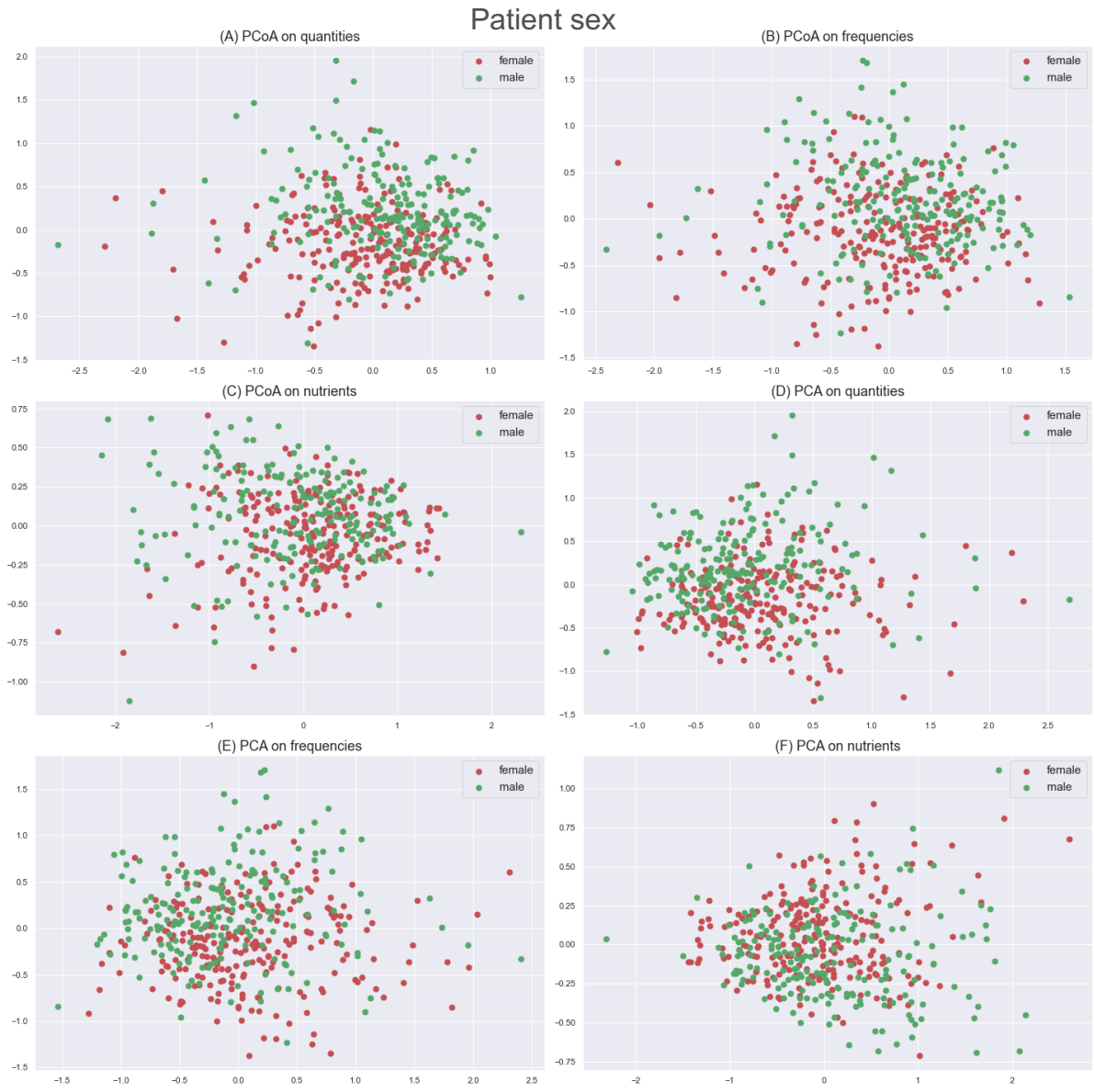


Fig 3.17 PCoA and PCA on dietary data coloured by sexes of the subject. A) PCoA on quantities; B) PCoA on frequencies; C) PCoA on daily intakes of micro and macro nutrients; D) PCA on quantities; E) PCA on frequencies; F) PCA on daily intakes of micro and macro nutrients.

Table 3.5 List of the variables in terms of quantities of food consumption unevenly distributed between the two sexes. P-values were computed using the Mann-Whitney U test and FDR correction.

Variable	P-value	Higher in
OTHER VEGETABLES	0.001	females
TOMATOES-COOKED	0.003	females
ROOT VEGETABLES	0.013	females
CABBAGES	0.009	males
STALK VEGETABLES, SPROUTS	0.000	females
MIXED SALAD, MIXED VEGETABLES	0.001	females
PASTA, OTHER GRAIN	0.000	males
GRAINS, WHOLEMEAL	0.015	females
BREAD	0.011	males
BREAKFAST CEREALS	0.004	females
BEEF	0.000	males
PORK	0.001	males
MUTTON/LAMB	0.012	males
HORSE	0.014	males
GOAT	0.012	males
PROCESSED MEAT	0.000	males
OFFALS	0.001	males
OTHER ANIMAL FAT	0.000	males
CONFECTIONERY NON CHOCOLATE	0.001	males
CARBONATED/SOFT/ISOTONIC DRINKS, DILUTED SYRUPS	0.001	females
TEA	0.001	females
HERBAL TEA	0.001	females
WINE	0.000	males
BEER, CIDER	0.000	males
SPIRITS, BRANDY	0.000	males
MAYONNAISES AND SIMILARS	0.021	males
SOUPS	0.000	females
PIZZA	0.040	females

Table 3.6 List of the frequencies of food consumption unevenly distributed between the two sexes. P-values were computed using the Mann-Whitney U test and FDR correction.

Variable	P-value	Higher in
OTHER VEGETABLES	0.001	females
ROOT VEGETABLES	0.001	females
CABBAGES	0.001	females
STALK VEGETABLES, SPROUTS	0.000	males
MIXED SALAD, MIXED VEGETABLES	0.002	females
NUTS AND SEEDS (+ NUT SPREAD)	0.019	females
PASTA, OTHER GRAIN	0.031	males
GRAINS, WHOLEMEAL	0.016	females
BREAD	0.012	males
BREAKFAST CEREALS	0.004	females
BEEF	0.001	males
PORK	0.000	males
MUTTON/LAMB	0.006	males
HORSE	0.019	males
GOAT	0.006	males
PROCESSED MEAT	0.000	males
OFFALS	0.006	males
OTHER ANIMAL FAT	0.000	males
CONFECTIONERY NON CHOCOLATE	0.001	males
CARBONATED/SOFT/ISOTONIC DRINKS, DILUTED SYRUPS	0.001	females
TEA	0.001	females
HERBAL TEA	0.001	females
WINE	0.000	males
BEER, CIDER	0.000	males
SPIRITS, BRANDY	0.000	males
MAYONNAISES AND SIMILARS	0.013	males
SOUPS	0.000	females

Table 3.7 List of the daily intakes of micro and macro nutrients unevenly distributed between the two sexes. P-values were computed using the Mann-Whitney U test and FDR correction.

Variable	P-value	Higher in
Total proteins	0.001	males
Animal proteins	0.001	males
Animal fats	0.003	males
Other polyunsaturated fats	0.004	males
Cholesterol	0.011	males
Simple carbohydrates	0.002	males
Starch	0.000	females
Alcohol	0.000	males
Sodium	0.000	males
Phosphorus	0.033	females
Zinc	0.001	females
Niacin	0.001	females
Retinol	0.033	females
beta-Carotene	0.026	females

Table 3.8 Percentage of male and female subjects across the three health status conditions

	healthy	mucositis	peri-implantitis
male	44.2%	48.3%	46.1%
females	55.8%	51.7%	53.9%

3.3.7 ML-based classification highlights good accuracy in predicting sex from dietary data

We used a ML-based approach to build classification models aiming at predicting sex from the dietary data. We used a cross-validation approach based on RF classification. The three dietary data types were compared, and results were in agreement with what found from ordination analysis (Fig. 3.18). The highest accuracy (AUC = 0.83) was achieved using consumption quantities as input features. Frequency of consumption gave an AUC = 0.81, while features representative of micro and macro nutrient intakes reduced the AUC to 0.73.

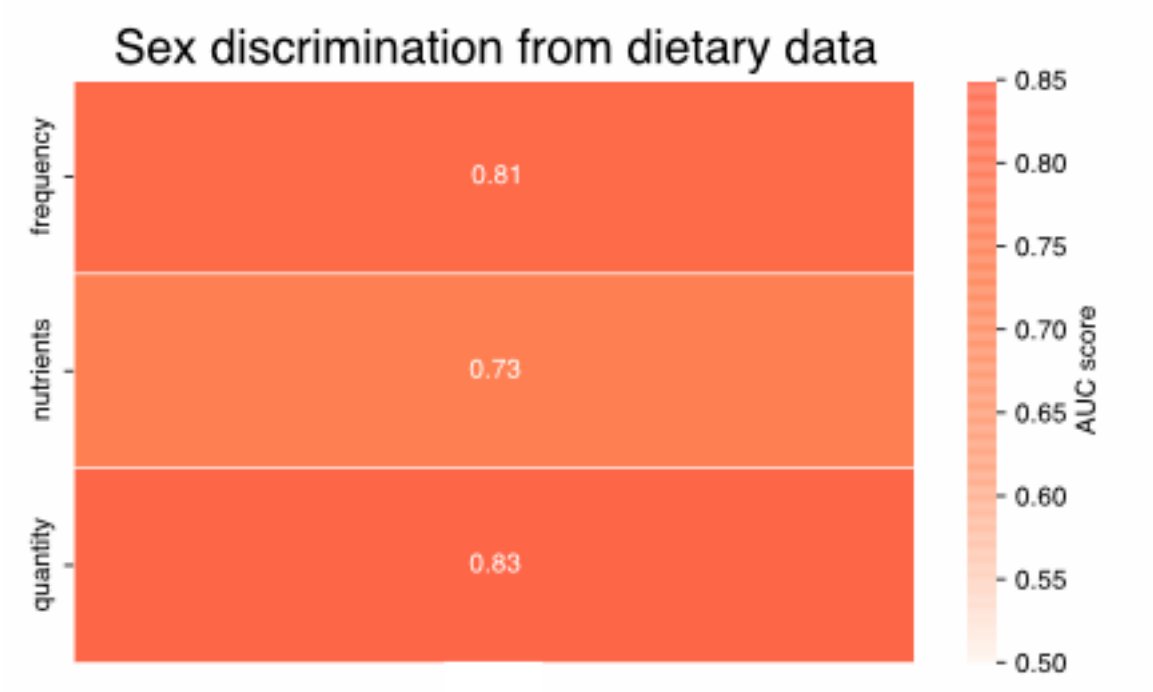


Fig 3.18 Classification results in terms of AUC obtained by predicting the sex from the dietary data. AUC were estimated by applying RF classifier in a cross-validation approach. The three dietary data types were compared.

3.3.8 Dietary habits are weakly associated with the health status

We finally extended the approaches discussed in the previous sections by evaluating differences among health statuses from dietary data. Ordination analysis didn't show differences among subjects characterised by different health statuses in terms of quantities (Fig 3.19 A, D), frequencies (Fig 3.19 B,E), or daily intakes of micro and macro nutrients (Fig 3.19 D,F).

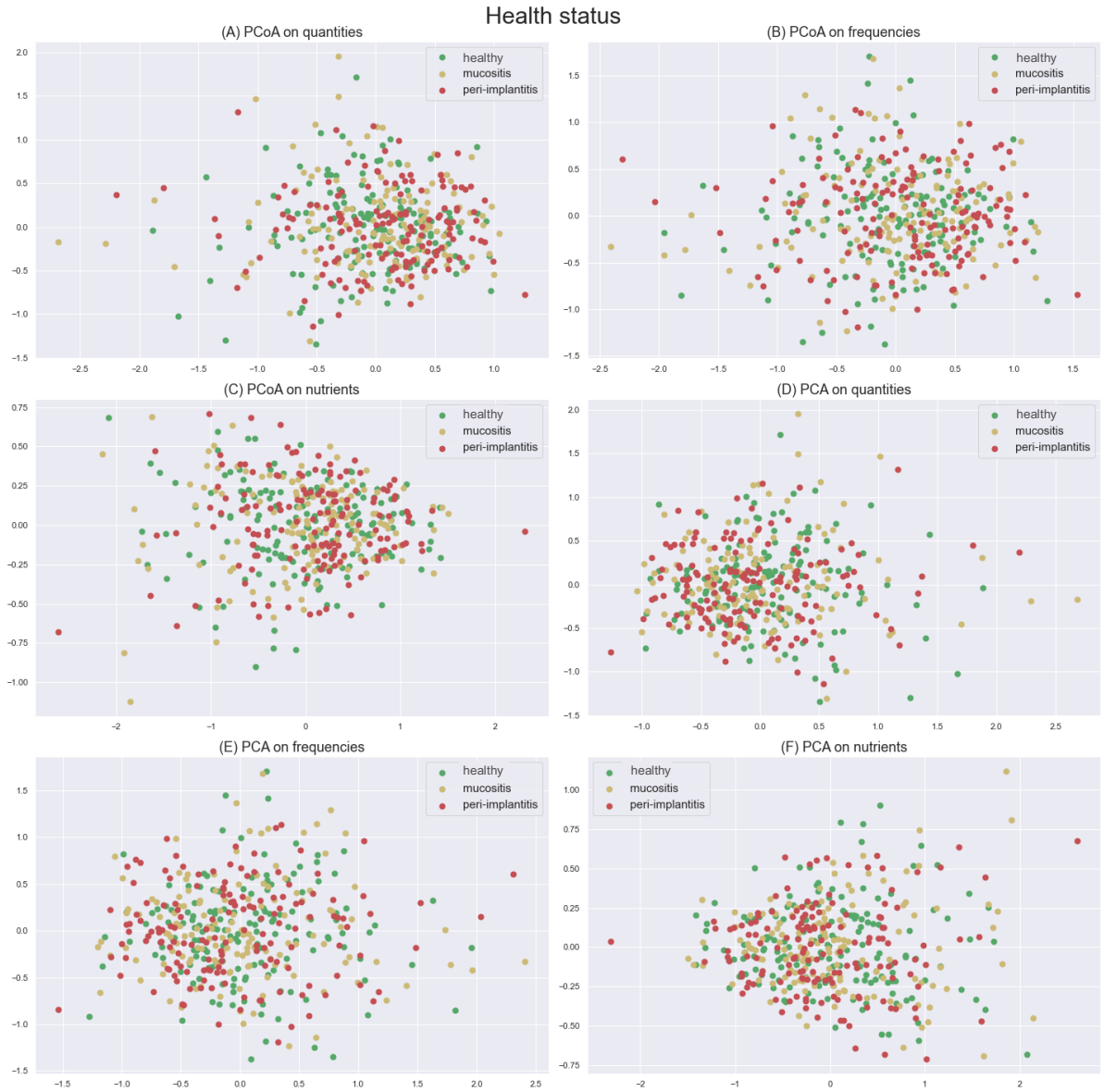


Fig 3.19 PCoA and PCA on dietary data coloured by health status. A) PCoA on quantities; B) PCoA on frequencies; C) PCoA on daily intakes of micro and macro nutrients; D) PCA on quantities; E) PCA frequencies; F) PCA on daily intakes of micro and macro nutrients.

We computed correlation coefficients between dietary data and the health status (we considered a q-value ≤ 0.1 to identify statistically significant variables). No significant correlations were found between health status and daily intakes, while some quantity and frequency features resulted to be significantly correlated. Such correlations were found when considering the three health conditions simultaneously (Fig 3.20), as well as when considering the binary case in comparing healthy and peri-implantitis subjects (Fig 3.21). Correlations showed the magnitude of variation in consumption of legumes, white yoghurt and white meat; variation in quantity and frequency of consumption of seafood and extra virgin olive oil (in cooked not fried foods) were associated with the healthy condition. From such results, we can say that peri-implantitis is not strongly driven by the dietary habits of the patients.

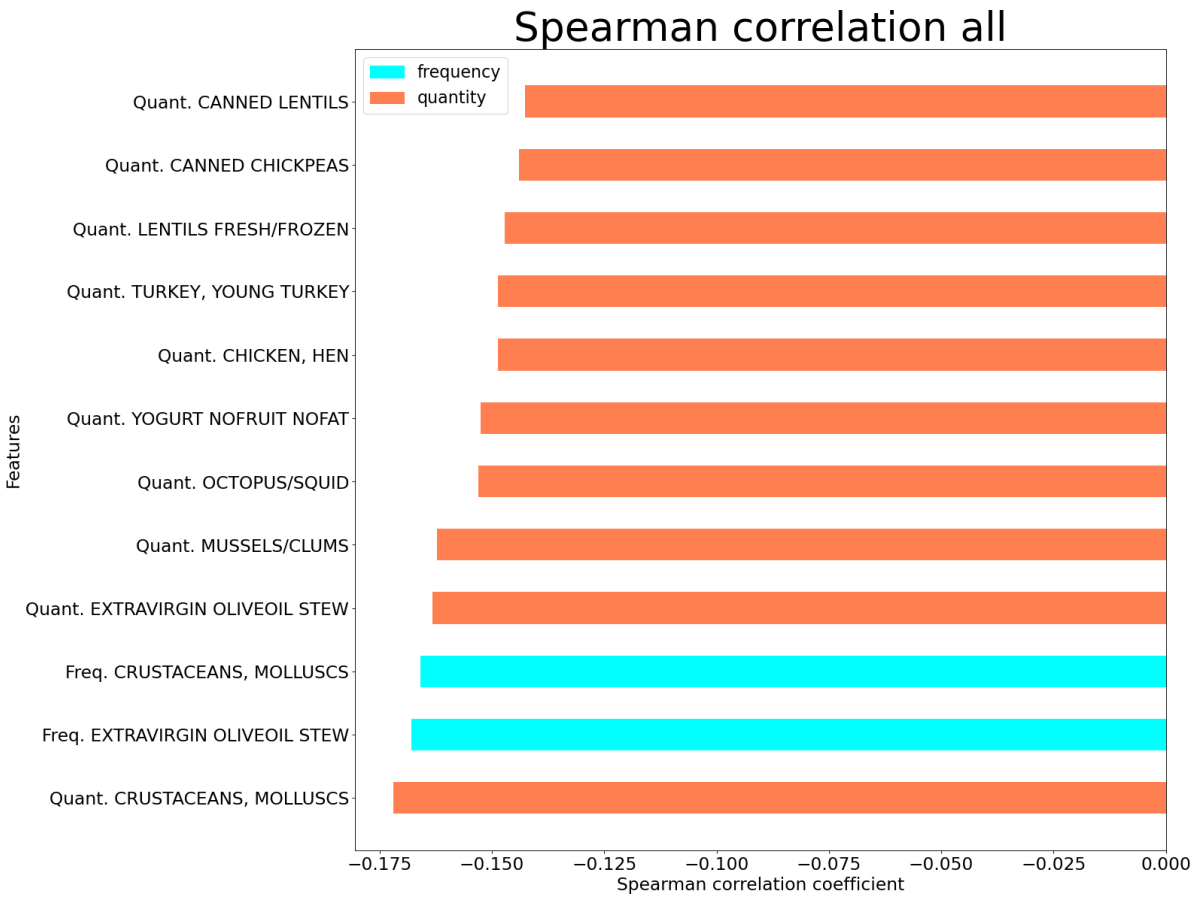


Fig 3.20 Spearman correlation coefficients between dietary features and patient's health status in the dataset with the three health conditions. Spearman correlation coefficients obtained by computing the spearman correlation between the clinical features and the health status of the patient. The correlations were computed considering the three health conditions codified as a numeric ordered categorical variable. The represented coefficients were filtered for p-value of the correlation <= 0.1. The different colours represent different kinds of features.

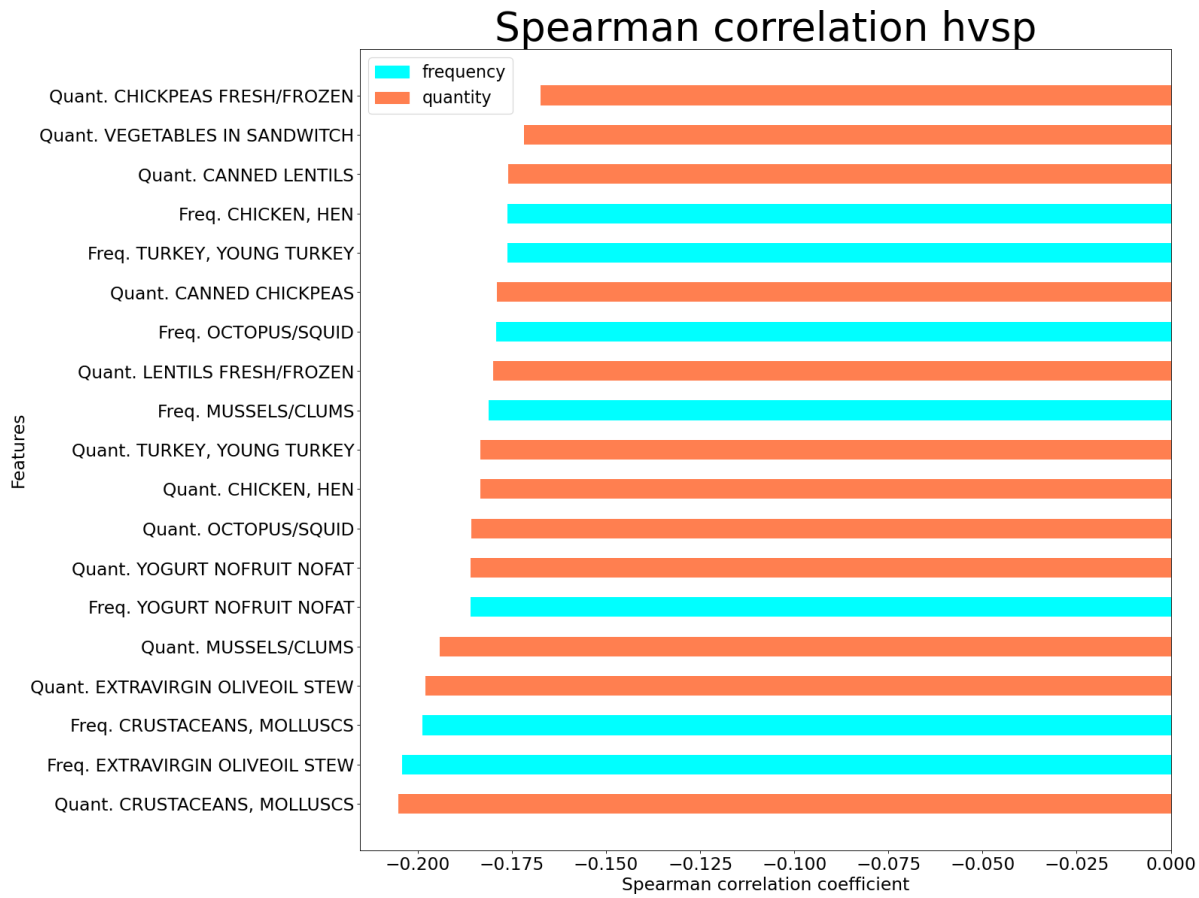


Fig 3.21 Spearman correlation coefficients between dietary features and patient's health status in the healthy vs peri-implantitis dataset. Spearman correlation coefficients obtained by computing the spearman correlation between the clinical features and the health status of the patient. The correlations were computed considering only the healthy and the peri-implantitis health conditions codified as a numeric ordered categorical variable. The represented coefficients were filtered for p-value of the correlation <= 0.1. The different colours represent different kinds of features.

3.3.9 Correction of dietary data for multicollinearity led to a sensible reduction in the number of variables

We corrected multicollinearity in frequencies and quantities of consumption data using the VIF approach by skimming features with VIF > 10. This correction led to a sensible reduction in the number of features by removing features representative of food subcategories. The number of features was reduced from 378 to 67 for both data types.

We used the same approach discussed in the previous sections to identify discriminatory variables from data corrected from multicollinearity issues. Also in this scenario, no micro or macro nutrient variables resulted statistically significant; on the other hand some significance was obtained for quantities (Fig 3.22) and frequencies (Fig 3.23) of food consumption. More specifically, frequencies and quantities of consumption of white meat, fish, and leafy vegetables characterised the dietary habits of the healthy subjects, confirming the evidence found in literature about the importance of protein consumption in maintaining a good oral health [138,183]. Dietary habits of mucositis subjects compared to t peri-implantitis subjects were characterised by the consume of sources of available carbohydrates (e.g., confectionery, pizza and bread) known as a risk factor for the development of microbial conveyed oral diseases [184] and as a main cause of acidification of oral pH with consequent selective push to acidophilic species and oral dysbiosis [136,185,186].

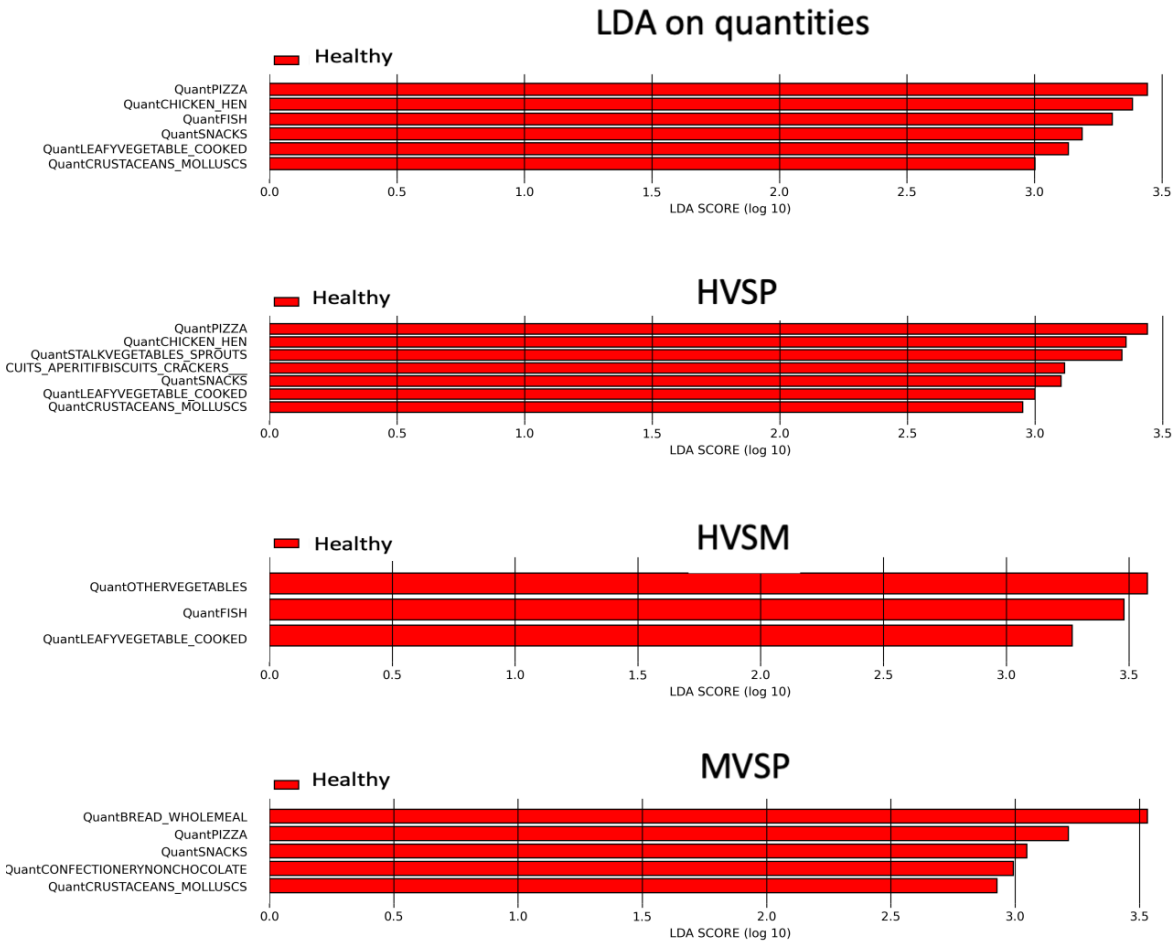


Fig 3.22 Largest effect size features obtained by LDA on the quantities of food consumption by comparing subjects in function of the health status.

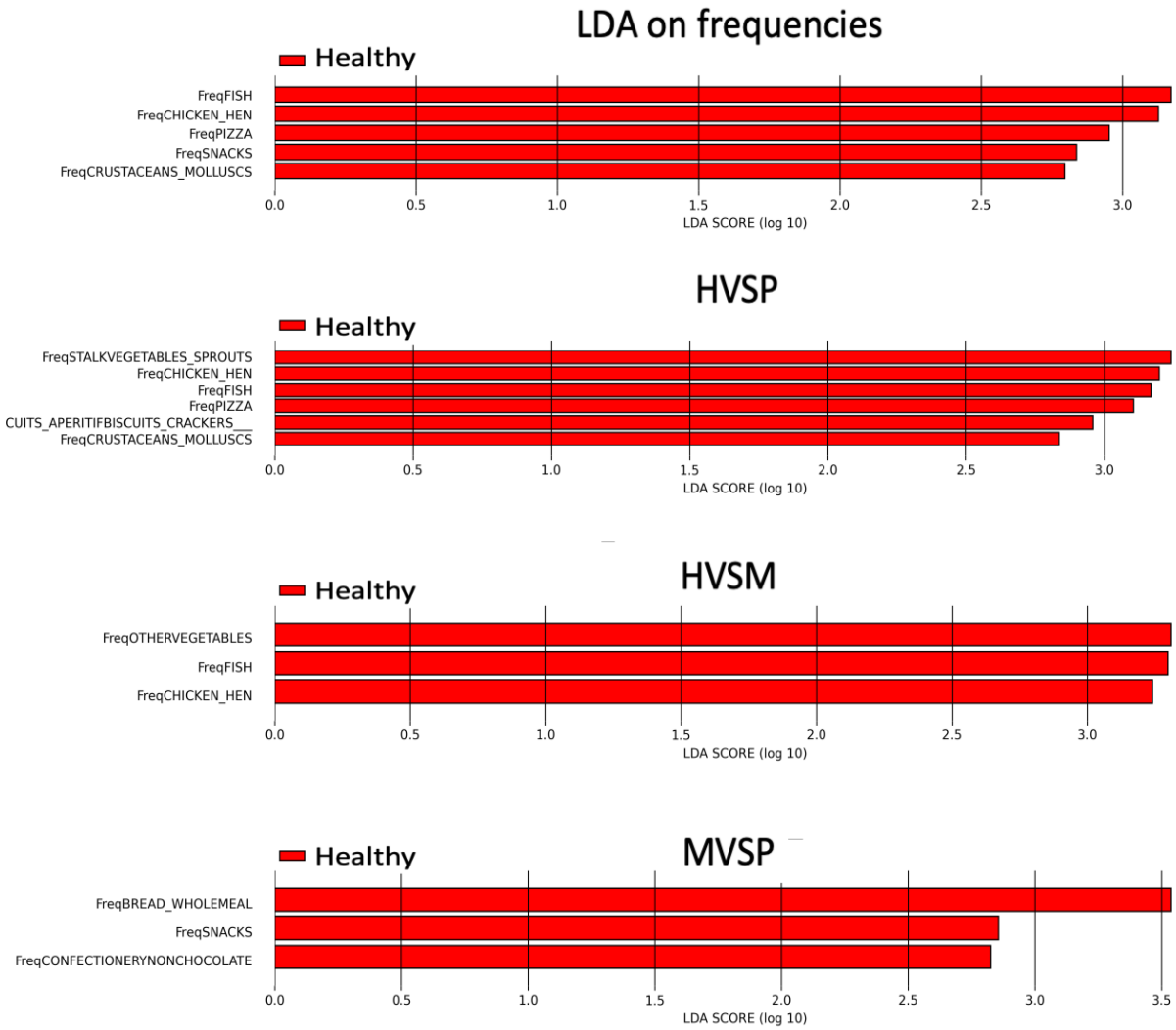


Fig 3.23 Largest effect size features obtained by LDA on the frequencies of food consumption by comparing subjects in function of the health status.

The same approach used in the training of the logistic regression model on the clinical features was considered to evaluate effects of dietary features variation on the probability to belong to the three health statuses. Also in this case, the model was built on the datasets after multicollinearity correction. The logit models trained on the different dietary data types did not produce any statistically significant results (Table 3.9), which further confirmed the weak correlation between dietary habits and peri-implant disease states.

Table 3.9 R2 relative to the logit model trained on the different binary combinations of health statuses from dietary data. hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

implementation	R2
quantities hvsp	0.23
quantities hvsm	0.21
quantities mvsp	0.21
frequencies hvsp	0.19
frequencies hvsm	0.18
frequencies mvsp	0.17
nutrients hvsp	0.06
nutrients hvsm	0.05
nutrients mvsp	0.05

3.3.10 Clinical features guarantee high accuracies for health status classification

We considered the same classification approach based on RF classifier used for sex prediction to predict the health status from the set of clinical features. We compared the two different strategies to deal with missing values. The results in terms AUC are reported in Fig 3.24 and a summary of the other metrics is available in Table 3.10. We obtained variable accuracies ranging from 0.91 for classification made on the imputed clinical dataset (i.e., when missing value were substituted by iterative imputation) between mucositis and peri-implantitis subjects (MVSP) to 0.99 when iterative imputation and mean substitution was applied to discriminate healthy and peri-implantitis subjects (HVSP). Such values were in line with what obtained by the correlations analysis previously reported. We used these results as a baseline to compare classification models trained from dietary data

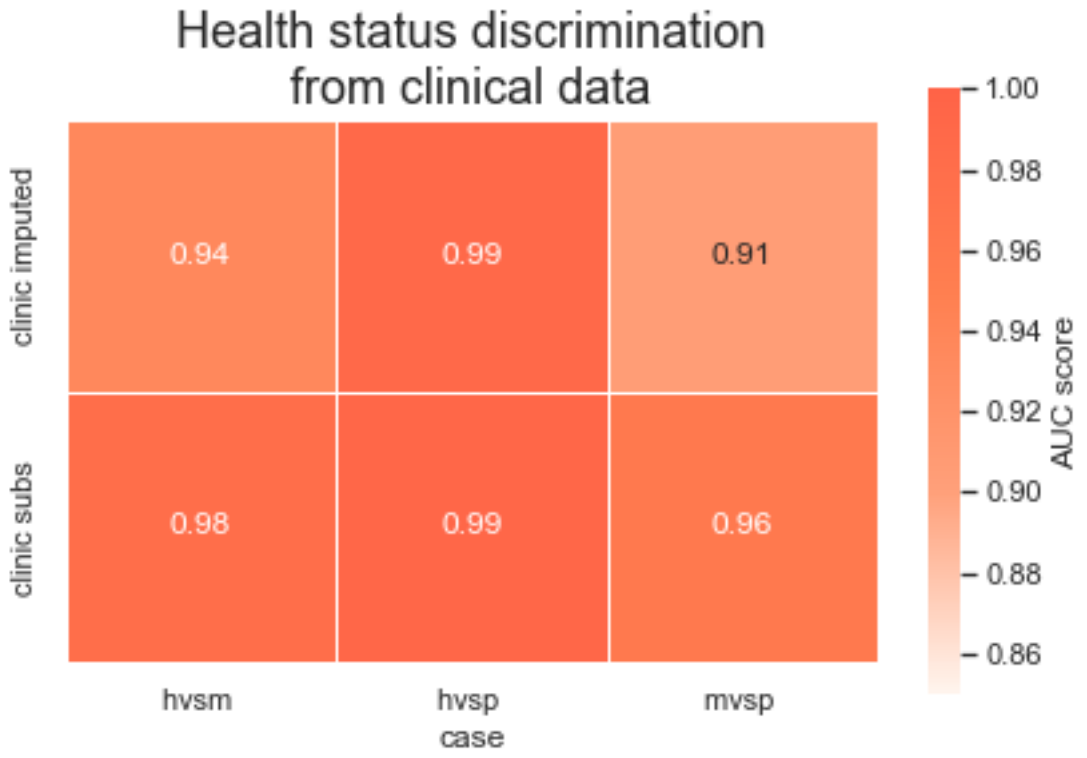


Fig 3.24 Classification results in terms of AUC scores obtained by predicting the health status from the clinical data. Models were built by RF classifier on the clinical features by considering the two different missing data management approaches. hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

Table 3.10 Summary of the classification results obtained by predicting the health status from the clinical data. Results obtained by RF classifier on the different clinical datasets obtained by the two different missing data management approaches, and reported in terms of AUC, AUPCR, F1, precision, and recall. All: three-class scenario with the three health status categories; hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

AUC				
case	**all**	**hvsm**	**hvsp**	**mvsp**
clinic imputed		0.94	0.99	0.91
clinic subs		0.98	0.99	0.96
AUPRC				
clinic imputed		0.93	0.99	0.91
clinic subs		0.98	0.99	0.97
F1				
clinic imputed	0.79	0.86	0.95	0.82
clinic subs	0.88	0.94	0.96	0.90
Precision				
clinic imputed	0.80	0.86	0.95	0.83
clinic subs	0.89	0.94	0.96	0.90
Recall				
clinic imputed	0.79	0.86	0.95	0.82
clinic subs	0.88	0.94	0.96	0.90

3.3.11 Classification on dietary data confirms weak capabilities for health state prediction

Finally, the same approach adopted to predict the health status from the clinical features was applied to the dietary datasets. In agreement with statistical testing and correlation analysis, we obtained low accuracy values (Fig. 3.25 and Table 3.11). For the majority of the cases, we obtained AUC values close to 0.5, which is associated with the randomic classification case. Only classifications made on frequency and quantity for discriminating between healthy and peri-implantitis subjects exhibited a moderate level of prediction with an AUC score of 0.61 and 0.60, respectively. Classification on the same set of data and by splitting for patient sex did not produce any sensible improvement in the classification performances.

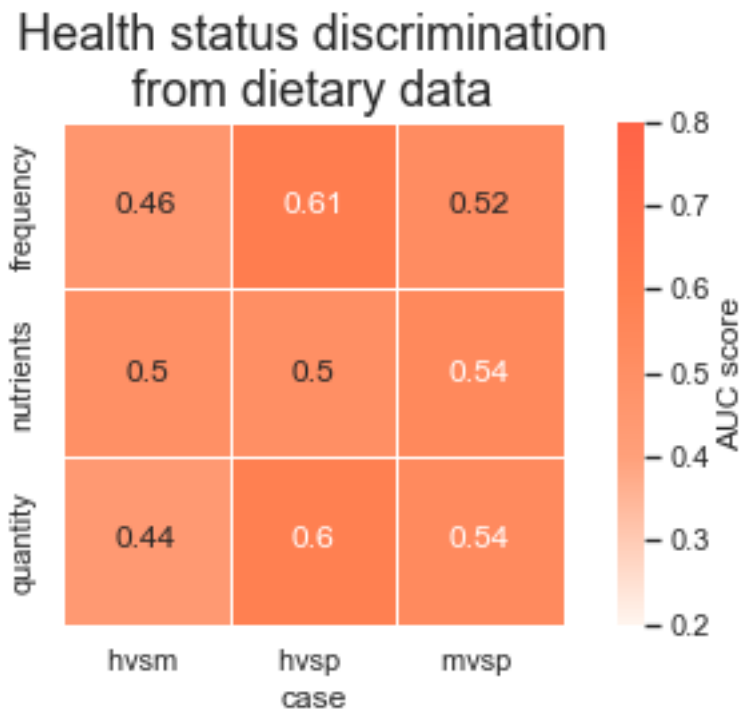


Fig 3.25 Classification results in terms of AUC scores obtained by predicting the health status from the dietary data. Models were built by RF classifier on the three dietary features datasets. hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

Table 3.11 Summary of the classification results obtained by predicting the health status from the dietary data. Results obtained by RF classifier on the different dietary datasets, and reported in terms of AUC, AUPCR, F1, precision and recall. All: three-class scenario with the three health status categories; hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

AUC				
	all	**hvsm**	**hvsp**	**mvsp**
frequency		0.46	0.61	0.52
micronutrients		0.50	0.50	0.54
quantity		0.44	0.60	0.54
AUPRC				
frequency		0.49	0.61	0.55
micronutrients		0.50	0.53	0.57
quantity		0.48	0.60	0.57
F1				
frequency	0.35	0.48	0.58	0.51
micronutrients	0.33	0.49	0.50	0.54
quantity	0.34	0.47	0.57	0.52
Precision				
frequency	0.36	0.49	0.59	0.52
micronutrients	0.34	0.49	0.50	0.55
quantity	0.34	0.47	0.58	0.53
Recall				
frequency	0.37	0.49	0.59	0.52
micronutrients	0.34	0.49	0.50	0.55
quantity	0.35	0.48	0.57	0.53

3.3.12 Combination of dietary data with clinical features does not improve classification accuracies

We finally evaluated if classification accuracies could be improved by combining dietary data with clinical features. The adding of the dietary features to the classification process did not improve the baseline results obtained on clinical data only (Fig. 3.26). Differences in classification accuracies occurred for the case in discriminating healthy and mucositis subjects (with an AUC worsening from 0.98 to 0.97), and mucositis and peri-implantitis subjects (with a small AUC improvement from 0.96 to 0.97).

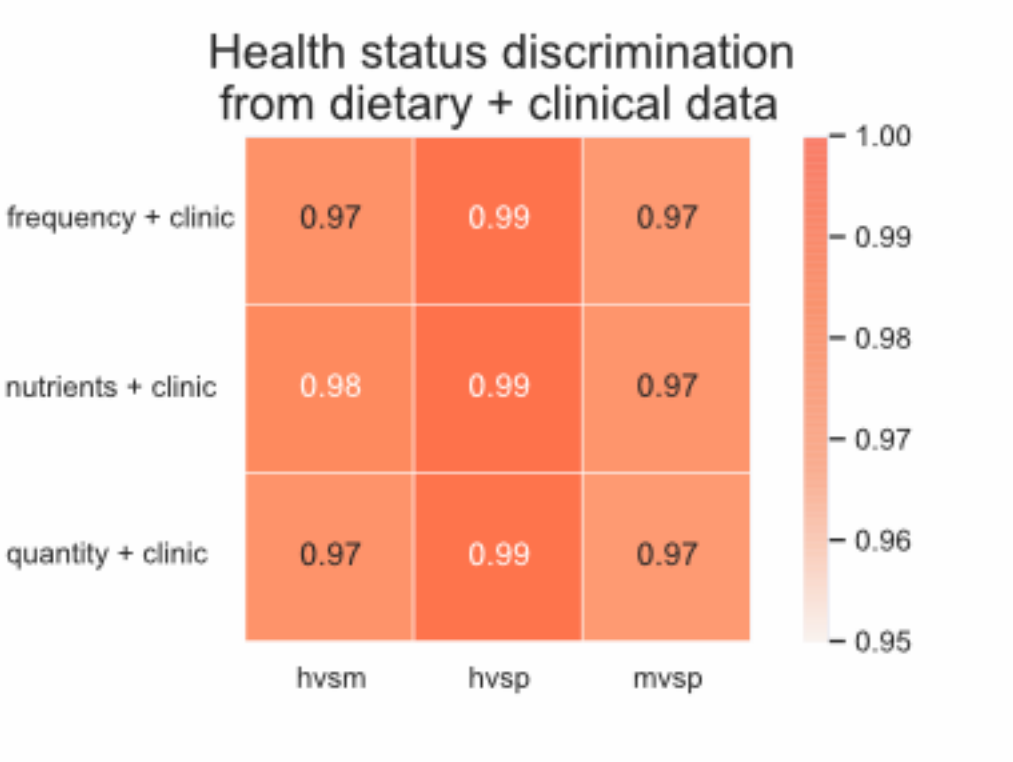


Fig 3.26 Classification results in terms of AUC scores obtained by predicting the health status from combining clinical and dietary data. Models were built by RF classifier on the three different combinations of dietary and clinical features. Hvsp: healthy vs peri-implantitis; hvsm: healthy vs mucositis; mvsp: mucositis vs peri-implantitis.

3.3.13 Conclusions

From the extensive set of analyses aiming at linking dietary data with health status information we found that dietary habits seem to not have a direct effect on the manifestation of peri-implant diseases. Such results were obtained by considering statistical testing and correlation analysis as well as by building prediction models based on cross-validation analysis.

4 Linking diet and oral microbiome

4.1 Introduction and scientific rationale

While the role of diet in the shaping of the gut microbiome was largely investigated, showing a strong influence of the dietary habits on the gut microbiome composition [187–202], the role of diet on the shaping of the oral microbiome is still under investigation. As already debated in the first chapter of this thesis, efforts are given by the scientific community to try to disentangle effects of the consumption of different foods on the oral microbiome composition. In the literature, findings identified relationships between the consumption of foods and nutrients and the pathogenetic process of certain oral diseases [203–205]. This was mainly attributed to the creation of a prosperous environment for the spreading and the proliferation of microbial species involved in the pathogenetic process. However, a proper characterisation of the species more linked to food consumption has been scarcely performed. It is still unclear if the push of the dietary habits is the primary driver to the proliferation of species directly related to the oral diseases or if it is one of the key stimuli to the instauration of a metabolic and microbial dysbiosis as a condition favourable to promote the spreading of pathogens species, as already hypothesised in periodontitis and osteoporosis [206,207], oral candidiasis [208], and pancreatic cancer and liver cirrhosis [209]. In this chapter, we gave a first characterisation of the influence of dietary intakes on the microbial composition and the metabolic potential of the oral microbiome. We also tried to disentangle the effect of the dietary intakes between the promotion of pathogenic species and the creation of a favourable scenario for the instauration and the proliferation of disease related species through a metagenomic analysis of functional metabolic pathways associated with the oral microbiome.

4.2 Materials and methods

4.2.1 Metagenomic sample collection

For the purposes of this thesis, we collected 121 metagenomic samples from healthy subjects (65 females and 56 males) coming from 28 Italian dental clinics. The sampling protocol followed in this study was based on the one validated by the Human Microbiome Project (HMP) consortium [210]. A single implant was sampled from each selected patient even if the patient has more than one implant with the same tested condition; in such cases one implant was chosen randomly for the sampling. To access submucosal and subgingival plaque samples, saliva was excluded from the selected sites using cotton rolls and an air syringe, supramucosal and supragingival plaque was removed with sterile cotton pellets. The technician in charge of the sampling had to collect the plaque from the deepest probing site with individual sterile titanium Gracey curettes. The use of Gracey curettes was preferred to the use of sterile paper points according to HMP [210] and to avoid potential contamination [211]. After the collection, samples were immediately placed in separate Eppendorf 1.5-mL microcentrifuge tubes (Eppendorf, Hamburg, Germany) containing sterile SCF-1 buffer solution (50 mM Tris-HCl, pH 7.5; 1 mM EDTA, pH 8.0; 0.5% Tween-20) [212] and preserved from the clinics till later retire by PreBiomics for later analysis, clinics were acknowledged that samples must be preserved away from heat sources and preferentially in fridge to preserve the DNA from thermal decay. Total genomic DNA was isolated using the Qiagen Power Soil Pro Kit (Qiagen, Hilden, Germany): an additional enzymatic disruption step for complete lysis of Gram-positive and Gram-negative species was performed, following the manufacturer's protocol. Isolated DNA was stored at −20 °C. Laboratory. Metagenome samples were quantified and the libraries were prepared using the Illumina DNA Prep Kit (Illumina Inc., San Diego, CA, USA) using the manufacturer's protocol. Libraries were sequenced on the NovaSeq-6000 platform (2 x 150bp reads). Shotgun metagenomics generated an initial set of 942 samples. The raw metagenomes generated by the Illumina sequencing were processed with Trim Galore (v. 0.6.6) with the following parameters: "--nextera --stringency 5 --length 75 --nextseq 20 --max_n 2 --trim-n --dont_gzip --no_report_file --suppress_warn". Human and bacteriophage phiX174 DNA (Illumina spike-in) was then removed using BowTie2 [213] (v. 2.3.4.3) by mapping the reads against the corresponding reference genomes. We used MetaPhlAn [214] (v. 4) for the taxonomic

characterization of the sampled microbial community and by setting “--stat_q 0.2”. We used HUMAnN 2.0 [215] for the metabolic potential of the microbiome. HUMAnN 2.0 is a pipeline for efficiently and accurately profiling the presence/absence and abundance of microbial pathways in a community from metagenomic or metatranscriptomic sequencing data (typically millions of short DNA/RNA reads). This process, referred to as functional profiling, aims to describe the metabolic potential of a microbial community and its members.

4.2.2 Evaluate sex as a possible confounding factor

Differences in eating habits between the two sexes, already identified in chapter 3, may act as a possible confounding factor in the subsequent analyses. To avoid possible distorting effects, we searched for differences between microbiome composition and microbiome metabolic potential between the two sexes using two different approaches: i) PCoA ordination technique for Multivariate analysis using the python package skbio.stats.ordination.pcoa on the Euclidean distance matrix computed among samples; ii) PERMANOVA test using the python module skbio.stats.distance.permanova on the Euclidean distances matrix and by considering 999 permutations.

4.2.3 Computing correlations between dietary data and oral microbiome

We identified associations between frequencies of food consumption and the microbiome composition/metabolic potential of healthy subjects by computing Pearson correlation coefficients between them. We adjusted the p-values with FDR for multiple hypothesis testing using the python module statsmodels.stats.multitest.fdrcorrection. Q-values < 0.05 identified statistically significant variables. We considered only food frequencies since, as already demonstrated in gut microbiome analyses, is more the frequency and occurrence of external factors to affect microbial composition more than the magnitude of such external factors [216,217]. For food frequency data we considered multicollinearity-corrected data using VIF < 10. In terms of microbiome data, correlations were computed by considering taxonomic profiles generated by MetaPhlAn as well function profiles generate by HUMAnN as described in the previous paragraph.

4.3 Results

4.3.1 Oral microbiome composition is not linked with multiple host characteristics

We considered a subset of the 451 subjects involved in the analysis presented in Chapter 3, for which we acquired the oral plaque microbiome through shotgun metagenomics. More specifically, we considered 121 subjects, all belonging to the "healthy" category, to have their characterization in terms of microbiome composition (generated by the MetaPhlAn tool) and functional potential (generated by the HuMANN tool).

In the first set of analyses, we evaluated to which extent microbiome composition could be linked to host characteristics. We performed clustering analysis and generated the heatmap representing the 50 microbial species having higher mean abundance across samples (Fig 4.1). We overimposed metadata information in terms of different host characteristics (i.e., subject sex, BMI, HDI score, and MIDI score) and no particular patterns were found. No particular patterns were verified clustering was shown between patients identified by different characteristics.

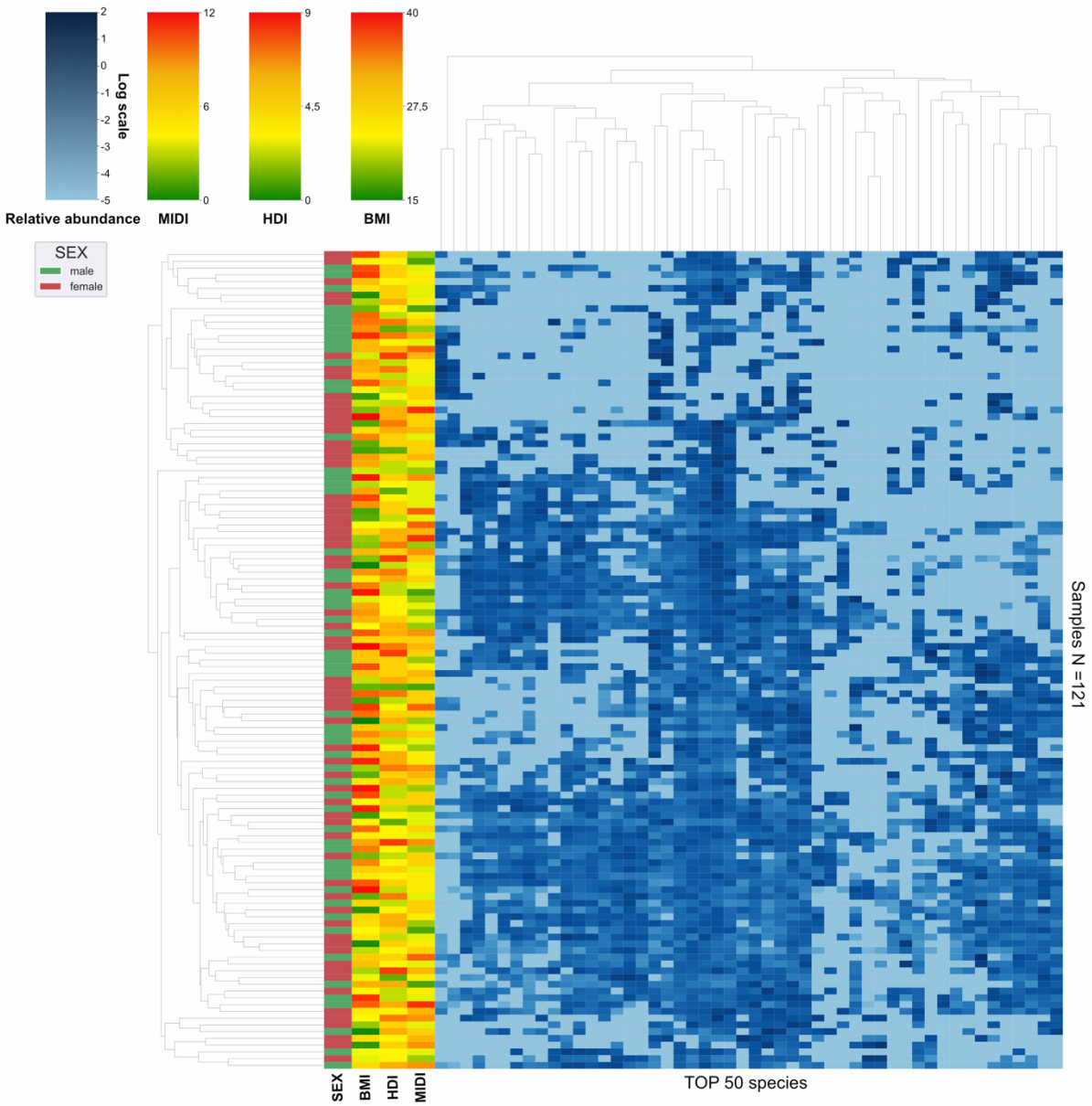


Fig 4.1 Taxonomic profiles generated from oral plaque microbiomes associated with 121 healthy subjects. The heatmap shows the relative abundances generated by MetaPhlAn for the 50 most abundant species across the samples. The left-most colorbar identifies the sex of the subjects; the three right-most colorbars indicate different dietary related features (i.e., BMI, HDI score, MIDI score).

4.3.2 Plaque microbiome does not differ for composition and metabolic potential in function of the sex

Given our findings that associated dietary patterns with the sex of the subjects involved in the study, we evaluated to which extent the composition of the plaque microbiome may be driven by the sex. Ordination analysis based on PCoA on taxonomic composition (Fig 4.2 A) and metabolic potential (Fig 4.2 B) did not show any differentiation among male and female subjects. These results were confirmed by PERMANOVA tests (p-value > 0.05 for both taxonomic and functional profiles). The absence of differences in microbiome encountered between the two sexes gives us a hint on the uncorrelated nature of dietary habits (that resulted different between the two sexes) and sub gingival plaque oral microbiome (that does not show any difference between the two sexes).

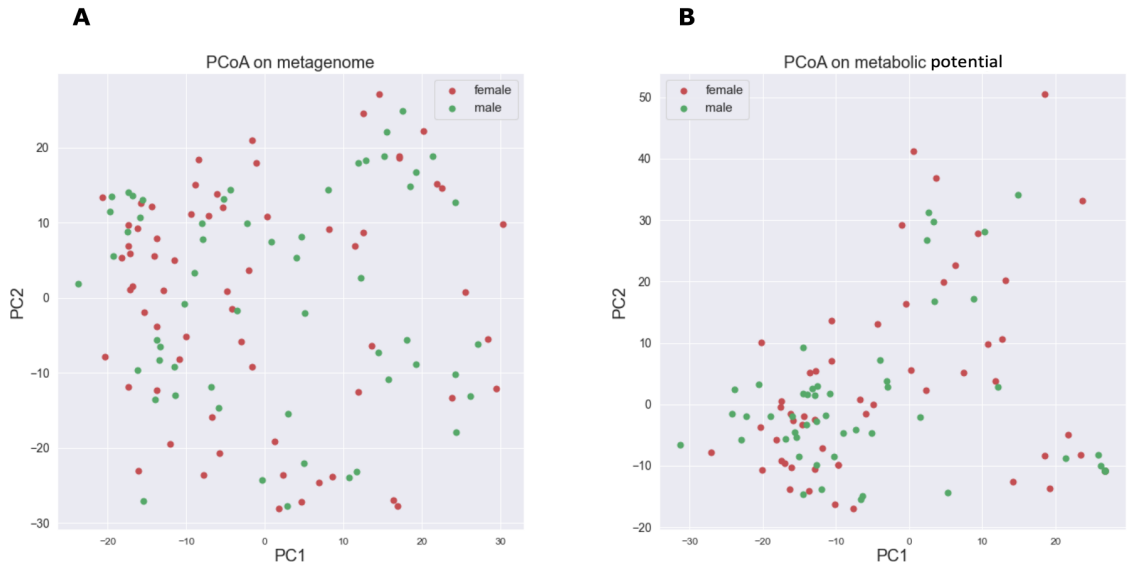


Fig 4.2 PCoA on (A) taxonomic and (B) metabolic profiles generated by shotgun metagenomes and coloured by sex of the subjects Results are obtained by considering Euclidean distance.

4.3.3 Sugary foods are main drivers of acidophilic pathogenic microbial species

We evaluated effects of the variation in frequencies of food consumption on the oral plaque microbiome composition by computing the Pearson correlation. We identified 48 microbial species correlated in a significant way with a set of frequencies of food consumption (Table 4.1). Among them, we identified multiple species that were already reported in the Potentially pathogenic microbial species include i) *Bulleidia extructa* a gram-positive, anaerobic and non-spore-forming bacterium, already identified as an etiological factor of the periodontal diseases, dental caries and dental abscess [218,219] correlated with the frequencies of consumption of wine; ii) *Campylobacter spp.* a gram-negative microaerophilic genus of bacteria, identified in literature as a possible cause of periodontitis[220], correlated with the frequency of consumption of citrus fruit; iii) *Capnocytophaga spp.* A gram-negative, CO2 dependant microbial genus; it is a typical species of oral microbiome with opportunistic pathogenic role in periodontal diseases [221] correlated with the frequencies of consumption of beef and sugary beverages; iv) *Fusobacterium nucleatum* a Gram negative, anaerobic oral bacterium identified as involved in the pathogenic process of the periodontitis and the peri-implantitis [31,222] resulted correlated with the frequency of consumption of chocolate and candies and with the frequency of consumption of wholemeal cereals; v) *Leptotrichia spp.* recognized in literature as a genera associated with a large set of oral diseases such as peri-implantitis, gingivitis, halitosis, and oral cancer [223,224] correlated with the frequencies of consumption of chocolate and candies, breakfast cereals, salty biscuits and cracker; vi) *Oribacterium spp*. A gram-positive strictly anaerobic microbial genera identified in literature as involved in caries pathogenic process [225] resulted correlated with the frequencies of consumption of sweets; vii) *Prevotella spp.* A gram-negative, anaerobic microbial genera identified in literature as etiological agent of peri-implant diseases [31] and as an indicator of general poor oral health [226] resulted correlated with the frequencies of consumption of mutton meat, pizza, sweets, fruit and spirits; viii) *Selenomonas spp. A* gram-negative anaerobic microbial genera recognized in literature as associated with *Fusobacterium spp.* in the pathogenic process of periodontal diseases[227] resulted correlated with the frequency of consumption butter, milk, sweets and vegetables oils not coming from olives; ix) *Treponema lecithinolyticum* a gram negative, facultative anaerobe microbial

species identified in literature as etiological agent of peri-implant diseases [31] and endodontic infections [228] resulted correlated with the frequency of consumption of sugary beverages. All the statistically significant associations were characterised by a positive correlation coefficient meaning that an increase in frequencies of food consumption are associated with an increase of the relative abundance of the involved species. The majority of frequencies of food consumption showing statistically significant associations with potentially pathogenic microbial species are relative to sources of simple fermentescible carbohydrates. This is in agreement with findings that showed how frequent consumption of fermentable carbohydrates could be one of the causes of the development of dental caries, driving the plaque ecology towards a state of dysbiosis [65,72,73]. The fermentation of the carbohydrates by the oral microbiota led to the formation of organic acids that can lower the oral pH if the buffer effect of the saliva is overwhelmed, creating a selective push for the acid tolerant bacteria involved in the cariogenic process [74]. However, a rise in the saliva pH may not necessarily be safe for oral health harbouring several periodontitis-assorted species as described in [225]. This suggest how probably there is not a favourable general condition to maintain an healthy oral microbiome and that a varied diet rich in vegetables could maintain a balance in the pH of saliva also thanks to the provision of dietary nitrate able to contrast an acid shifting of the saliva pH [229,230], exerting control power on the develop of potentially pathogenic species[231]. The lack of information about saliva pH did not allow us to go further in this direction. Moreover, the source of our microbiological sample (i.e., the subgingival plaque) represented a different biological niche than the saliva and to date few studies focused on the sub gingival plaque microbiome have tried to link their composition with the dietary patterns [232,233].

Table 4.1 Correlation coefficients between frequencies of food consumption and oral plaque microbiome composition. Coefficients are computed through Pearson correlation. P-values are FDR corrected. Only statistically significant correlation (Q-value <= 0.05) are reported.

Frequencies	Species	Correlation value	p-value
MIXED SALAD, MIXED VEGETABLES	Actinobaculum_sp_oral_taxon_183_t__SGB15892	0.37	0.01
VEAL	Bifidobacterium_dentium_t__SGB17234	0.45	0.00
WINE	Bulleidia_extructa_t__SGB6820	0.32	0.05
CITRUS FRUIT	Campylobacter_gracilis_t__SGB19300	0.32	0.04
SOUPS	Campylobacter_rectus_t__SGB19315	0.34	0.03
BEEF	Capnocytophaga_sp_oral_taxon_338_t__SGB2478	0.43	0.00
CARBONATED/SOFT/ISOTONIC DRINKS, DILUTED SYRUPS	Capnocytophaga_sp_oral_taxon_338_t__SGB2478	0.44	0.00
SOYA PRODUCTS	Cardiobacterium_valvarum_t__SGB9416	0.34	0.03
MILK	Colibacter_massiliensis_t__SGB5869	0.34	0.03
RABBIT (DOMESTIC)	Cutibacterium_acnes_t__SGB16955	0.46	0.00
GRAINS, WHOLEMEAL	Fusobacterium_nucleatum_t__SGB6007	0.34	0.02
CHOCOLATE, CANDY BARS, PASTE, CONFETTI/FLAKES	Fusobacterium_nucleatum_t__SGB6011	0.33	0.04
HORSE	GGB12790_SGB19844_t__SGB19844	0.48	0.00
SOUPS	GGB12790_SGB19845_t__SGB19845	0.33	0.04
FRUIT AND VEGETABLE JUICES	Isoptericola_variabilis_t__SGB17153_group	0.40	0.00
MIXED FRUITS	Kytococcus_sedentarius_t__SGB17151	0.36	0.01
PIZZA	Kytococcus_sedentarius_t__SGB17151	0.45	0.00
BREAD, WHOLEMEAL	Lachnoanaerobaculum_sp_ICM7_t__SGB4494_group	0.37	0.01
SOUPS	Lachnoanaerobaculum_sp_ICM7_t__SGB4494_group	0.35	0.02
SOYA PRODUCTS	Lachnoanaerobaculum_sp_ICM7_t__SGB4494_group	0.34	0.02
CONFECTIONERY NON CHOCOLATE	Lancefieldella_parvula_t__SGB964	0.39	0.00
CONFECTIONERY NON CHOCOLATE	Lautropia_dentalis_t__SGB13164	0.42	0.00
CHOCOLATE, CANDY BARS, PASTE, CONFETTI/FLAKES	Leptotrichia_hongkongensis_t__SGB6059	0.35	0.02
SALTY BISCUITS, APERITIF BISCUITS, CRACKERS,..	Leptotrichia_sp_oral_taxon_212_t__SGB6070_group	0.33	0.04
CONFECTIONERY NON CHOCOLATE	Leptotrichia_sp_oral_taxon_498_t__SGB6053	0.33	0.03
BREAKFAST CEREALS	Leptotrichia_wadei_t__SGB6055	0.35	0.02
GRAINS, WHOLEMEAL	Mycoplasma_salivarium_t__SGB5934	0.47	0.00
TOMATOES-RAW	Neisseria_subflava_t__SGB9450_group	0.38	0.01
CONFECTIONERY NON CHOCOLATE	Oribacterium_sp_oral_taxon_078_t__SGB7282	0.40	0.00
RABBIT (DOMESTIC)	Paraburkholderia_bryophila_t__SGB32753	0.33	0.04
RABBIT (DOMESTIC)	Paraburkholderia_fungorum_t__SGB13048	0.46	0.00
CITRUS FRUIT	Peptidiphaga_gingivicola_t__SGB15894	0.53	0.00
ICE CREAM	Peptidiphaga_gingivicola_t__SGB15894	0.38	0.01
LEGUMES	Porphyromonas_pasteri_t__SGB2043	0.37	0.01
VEGETABLE OILS (NO OLIVE)	Prevotella_baroniae_t__SGB1533	0.43	0.00
MUTTON/LAMB	Prevotella_intermedia_t__SGB1560	0.38	0.01

Frequencies	Species	Correlation value	p-value
PIZZA	Prevotella_melaninogenica_t__SGB1552	0.35	0.02
MIXED FRUITS	Prevotella_oulorum_t__SGB1520	0.33	0.04
CONFECTIONERY NON CHOCOLATE	Prevotella_oulorum_t__SGB1520	0.35	0.02
CITRUS FRUIT	Prevotella_pleuritidis_t__SGB1526	0.37	0.01
SPIRITS, BRANDY	Prevotella_sp_oral_taxon_376_t__SGB21554	0.37	0.01
BUTTER	Selenomonas_artemidis_t__SGB5878	0.40	0.00
CONFECTIONERY NON CHOCOLATE	Selenomonas_flueggei_t__SGB5882	0.40	0.00
VEGETABLE OILS (NO OLIVE)	Selenomonas_sp_oral_taxon_126_t__SGB5888	0.53	0.00
MILK	Selenomonas_sp_oral_taxon_920_t__SGB25061	0.32	0.05
PORK	Shuttleworthia_satelles_t__SGB7281	0.32	0.05
BREAKFAST CEREALS	Streptococcus_anginosus_t__SGB8028_group	0.40	0.00
OFFALS	Streptococcus_anginosus_t__SGB8028_group	0.32	0.05
RICE	Streptococcus_constellatus_t__SGB8026	0.41	0.00
MUSHROOMS	Streptococcus_gordonii_t__SGB8053	0.41	0.00
PROCESSED CHEESE	Streptococcus_gordonii_t__SGB8053	0.53	0.00
OFFALS	Streptococcus_gordonii_t__SGB8053	0.36	0.01
SNACKS	Streptococcus_gordonii_t__SGB8053	0.37	0.01
OTHER VEGETABLES	Streptococcus_infantis_t__SGB8095	0.36	0.02
CABBAGES	Streptococcus_infantis_t__SGB8095	0.50	0.00
FISH	Streptococcus_infantis_t__SGB8095	0.33	0.04
PROCESSED CHEESE	Streptococcus_mitis_t__SGB8163	0.42	0.00
HORSE	Streptococcus_mitis_t__SGB8163	0.35	0.02
OFFALS	Streptococcus_mitis_t__SGB8163	0.34	0.02
RABBIT (DOMESTIC)	Streptococcus_salivarius_t__SGB8007_group	0.52	0.00
FRUIT AND VEGETABLE JUICES	Streptococcus_sanguinis_t__SGB8047	0.38	0.01
RABBIT (DOMESTIC)	Streptococcus_sp_A12_t__SGB8059_group	0.62	0.00
SPIRITS, BRANDY	Tannerella_sp_oral_taxon_808_t__SGB2047	0.42	0.00
TOMATOES-COOKED	Tannerella_sp_oral_taxon_HOT_286_t__SGB2048	0.33	0.04
CARBONATED/SOFT/ISOTONIC DRINKS, DILUTED SYRUPS	Treponema_lecithinolyticum_t__SGB3587	0.59	0.00
SPIRITS, BRANDY	Treponema_sp_OMZ_804_t__SGB3607	0.43	0.00

4.3.4 Animal fats and sugary foods are associated with potential pathogenic metabolic pathways

We extended the analysis on linking oral plaque microbiome and diet by finding associations with the potential microbial metabolic pathways. We found 22 microbial pathways that correlated in a statistically significant way to a set of frequencies of food consumption (Table 4.2). Pathways of interest among them are represented by: i) L-ornithine biosynthesis I and L-arginine degradation XIII correlated with butter consumption already identified in literature as associated with *Seolmonas spp*. and involved in the production of malodorous gases characterising the halitosis, gingivitis and periodontitis diseases [234]; ii) L-arginine biosynthesis I, L-ornithine biosynthesis I and L-arginine degradation XIII correlated with sugary foods consumption and already identified in literature as one of the causes of defections in formation of *Streptococcus spp.* biofilm, useful to protect the teeth enamel and contrast the formation of caries and the appearance of periodontal diseases [235]; iii) (S)-propane-1,2-diol degradation correlated with cereal consumption already identified as involved in the carbohydrates fermentation of oral pathogenic species such as *Fusobacterium nucleatum* [236,237]; iv) heme b biosynthesis correlated with the consumption of lamb meat and already identified as involved in the iron metabolism of anaerobic species entangled in the pathogenic process of periodontal diseases [238–240]; v) other metabolic pathways correlated with frequencies of food consumption are characterised by an acidification of the substrate due to the release of organic acids as final products. Moreover, the selective push of the acidification of the saliva in favour of acidophilic bacteria is one of the main reasons for reduction of biodiversity in the oral microbiome, reducing the set of metabolic pathways expressed [14,241].

Table 4.2 Correlation coefficients between frequencies of food consumption and oral plaque microbiome metabolic potential. Coefficients are computed through Pearson correlation. P-values are FDR corrected. Only statistically significant correlation (Q-value <= 0.05) are reported.

Frequency	Pathway	Correlation	P value
MUTTON/LAMB	1CMET2-PWY: folate transformations III (E. coli)	0.35	0.01
CONFECTIONERY NON CHOCOLATE	ARGSYN-PWY: L-arginine biosynthesis I (via L-ornithine)	0.44	0.00
CONFECTIONERY NON CHOCOLATE	ARGSYNBSUB-PWY: L-arginine biosynthesis II (acetyl cycle)	0.37	0.01
BUTTER	GLUTORN-PWY: L-ornithine biosynthesis I	0.32	0.04
CONFECTIONERY NON CHOCOLATE	GLUTORN-PWY: L-ornithine biosynthesis I	0.47	0.00
PROCESSED CHEESE	GLYCOLYSIS: glycolysis I (from glucose 6-phosphate)	0.35	0.01
SNACKS	GLYCOLYSIS: glycolysis I (from glucose 6-phosphate)	0.50	0.00
BREAD	HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)	0.37	0.01
MUTTON/LAMB	HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)	0.71	0.00
CHICKEN, HEN	HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)	0.33	0.03
PROCESSED MEAT	HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)	0.31	0.05
CRUSTACEANS, MOLLUSCS	HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)	0.34	0.02
OTHER ANIMAL FAT	HEME-BIOSYNTHESIS-II: heme b biosynthesis I (aerobic)	0.51	0.00
BREAD	PHOSLIPSYN-PWY: superpathway of phospholipid biosynthesis I (bacteria)	0.32	0.03
WINE	PHOSLIPSYN-PWY: superpathway of phospholipid biosynthesis I (bacteria)	0.37	0.01
MIXED SALAD, MIXED VEGETABLES	PWY-2941: L-lysine biosynthesis II	0.49	0.00
RICE	PWY-2941: L-lysine biosynthesis II	0.33	0.03
PROCESSED CHEESE	PWY-3001: superpathway of L-isoleucine biosynthesis I	0.44	0.00
PIZZA	PWY-3001: superpathway of L-isoleucine biosynthesis I	0.32	0.04
OTHER VEGETABLES	PWY-5484: glycolysis II (from fructose 6-phosphate)	0.32	0.05
SALTY BISCUITS, APERITIF BISCUITS, CRACKERS,..	PWY-5484: glycolysis II (from fructose 6-phosphate)	0.40	0.00
FRUIT AND VEGETABLE JUICES	PWY-5484: glycolysis II (from fructose 6-phosphate)	0.36	0.01
BOUILLON	PWY-5667: CDP-diacylglycerol biosynthesis I	0.39	0.00
VEGETABLE OILS (NO OLIVE)	PWY-5981: CDP-diacylglycerol biosynthesis III	0.46	0.00
CONFECTIONERY NON CHOCOLATE	PWY-6630: superpathway of L-tyrosine biosynthesis	0.40	0.00
GRAINS, WHOLEMEAL	PWY-7013: (S)-propane-1,2-diol degradation	0.44	0.00
BUTTER	PWY-7198: pyrimidine deoxyribonucleotides de novo biosynthesis IV	0.32	0.04
SOYA PRODUCTS	PWY-7282: 4-amino-2-methyl-5-diphosphomethylpyrimidine biosynthesis II	0.35	0.01
VEAL	PWY-7357: thiamine phosphate formation from pyrithiamine and oxythiamine (yeast)	0.31	0.05
BUTTER	PWY-8187: L-arginine degradation XIII (reductive Stickland reaction)	0.35	0.01
CONFECTIONERY NON CHOCOLATE	PWY-8187: L-arginine degradation XIII (reductive Stickland reaction)	0.36	0.01
VEAL	PWY-I9: L-cysteine biosynthesis VI (from L-methionine)	0.33	0.03
SALTY BISCUITS, APERITIF BISCUITS, CRACKERS,..	PWY0-1296: purine ribonucleosides degradation	0.35	0.01
PIZZA	PWY0-1296: purine ribonucleosides degradation	0.32	0.04
BOUILLON	PWY0-1319: CDP-diacylglycerol biosynthesis II	0.37	0.01

Frequency	Pathway	Correlation	P value
CONFECTIONERY NON CHOCOLATE	PWY0-162: superpathway of pyrimidine ribonucleotides de novo biosynthesis	0.37	0.01
CITRUS FRUIT	THISYNARA-PWY: superpathway of thiamine diphosphate biosynthesis III (eukaryotes)	0.33	0.03

4.3.5 Discussion

This chapter provides one of the firsts diet-oral microbiome studies voted to the characterization of the sub gingival plaque microbiome composition and potential in association with daily dietary intakes, multiple measures of diet quality and different host characteristics. These associations were studied on a newly acquired cohort of peri-implant patients on the whole Italian territory. Despite the evidence found in literature about the different eating styles characterising the two different sexes[177–179] and confirmed in the analysis carried on in chapter 3, microbiome did not differ between the two sexes in terms of composition and potential. The microbiome was particularly unrelated to dietary habits confirming the previous findings stating the oral microbiome resilience over different external stimuli [14,15,242].

Some already known pathogenic microbial species correlated with the frequencies of consumption of main sources of dietary carbohydrates. A high level of consumption of carbohydrates was already defined as one of the only external stimuli able to influence microbiome composition towards a dysbiosis state [14,184,241,243], enhancing acidification of the oral environment. An unbalanced eating style enriched in sugars and animal fat was correlated with the expression of some metabolic pathways related to some microbial species already identified as involved in the pathogenic process of some of the main oral diseases. Dietary habits were identified as not correlated with microbiome composition. However unbalanced dietary habits could be the main trigger to the disruption of the health-maintaining mechanisms that limit the effect of disease drivers including the complex set of metabolic and functional interrelationships that develop within dental biofilms and between biofilms and the host.

5 Conclusions

In the present study, we conducted an analysis on a cohort of 451 Italian subjects for which there was available information about their dietary habits and their oral peri-implant clinical conditions. We also collected the oral plaque microbiome for 121 of them. By applying different inferential statistics techniques, species-level taxonomic and functional profiling and machine-learning based classification approaches based on state-of-the-art methodologies we demonstrated that dietary habits are scarcely correlated with the presence of the peri-implant diseases and with the composition and metabolic potential of the oral microbiome. No further recommendation about a dietary regime useful to preserve good oral health can be done beyond those already present in literature about a “healthy” eating style [242,243].

Due to COVID-19 spreading our experimental design changed ongoing during the sample collection phase. The collection of dietary data made by the administration of a validated FFQ questionnaire provided by EPIC, was planned as a supervised procedure. Patients should have answered the questionnaire questions with the supervision of the dental clinics involved in the study using a web platform implemented with a data capture system not allowing them to skip to the next question without an answer. The spreading of covid changed the experimental design, passing from a supervised web filling of the questionnaires to an unsupervised filling of a paper version of the questionnaire. The questionnaires were given to the patient at the moment of the recruitment by the clinic with the direction to fill it at home and deliver it back to the clinic at the first possible occasion. The lack of contextuality between the moment of the sampling and the filling of the questionnaire led to a high loss in observation due to the non-delivery of a large number of questionnaires. Moreover, the delivered questionnaires were characterised by large portions of unanswered questions. The questionnaires were inserted on the web platform after the delivery of the questionnaire without the availability of the patient. The logic of the web platform in which the questionnaires were inserted led to a large number of zeros across the three dietary datasets. This generated a large amount of noise on the dietary datasets and the loss of information. For these reasons together with the nature of the dietary data itself and the collection techniques, that generally relies on estimates made by the interviewed generating differences between the collected data and the observed

daily dietary intake ranging from 4% to 400% [141], we were not able neither to identify the relationship between dietary habits and oral microbiome already identified in literature.

Diet is a major lifestyle related risk factor. Dietary habits have been found correlated with incidence of cancer [244] and the dietary composition information has been useful to predict cardiovascular diseases risk [245]. It is largely demonstrated that fermentescible sugar consumption exert a selective push towards cariogenic species and acidification of the oral environment with consequent disruption of the dental enamel [136,185,186,246] and that dietary nitrite intakes along with right amount of protein assumption play a key role in protection versus periodontal diseases and carious lesions [247–252]. The absence of strong correlations between dietary habits and composition and metabolic potential of the microbiome could be also given by the low quality of the dietary data collected in our study. Moreover the specificity of the collection site for oral microbiome (subgingival plaque) characterised by high level of stability under external stimuli over time [253] this together with the resilience characterising oral microbiome [14,15] could have led to the under identification of statistical significant correlations.

The growing attention of the scientific community towards oral microbiome as an indicator of the oral and systemic health status together with the always growing awareness of the complex interactions between microbiome, host and environment are playing a key role in moving the interest of the scientific community towards the characterisation of the microbiomes hosted in different biological niches of our body. The diet as a major lifestyle related risk factor is being taken into account in the characterisation of the interactions. The absence of information regarding the oral environment (e.g., oral pH, saliva iron concentration, and saliva nitrate content) together with low quality of the dietary data and the specificity of the sample site are a major limit to the results reached in this study. Moreover, recent approaches to the study of microbiome are considering nonlinear correlation as the best choice to investigate intermicrobiome correlations among taxas [254]. This kind of approach is promising for the identification of perturbations or changes in the interactions among microbiota within and between ecosystem(s) and could be helpful in the fulfilling of the

objectives of this work identifying new relationship between food consumption and microbiome composition. Further investigations in this direction are needed.

6 Supporting informations

S2.1 Table. Summary of the 30 classification tasks derived from 16S rRNA datasets for casecontrol prediction. ASD: Autism spectrum disorder, CD: Crohn disease, CDI: Clostridium difficile infection, CIRR: Cirrhosis, MHE: Minimal hepatic encephalopathy, CRC: Colorectal cancer, EDD: enteric diarrheal disease, HIV: human immunodeficiency virus, NASH: non-alcoholic steatohepatitis, OB: obesity, PAR: Parkinson's disease, PSA: psoriatic arthritis, RA: Rheumatoid arthritis, T1D: type-1 diabetes, UC: ulcerative colitis. Non-CDI controls are patients with diarrhoea who tested negative for C. difficile infection.

S2.2 Table. Results obtained from the classification process done on the shotgun datasets. Comparison in terms of AUC, AUPRC, F1, precision, recall between relative abundance and presence/absence profiles at different threshold levels. Results are obtained using RF classification at the species-level taxonomic resolution.

S2.3 Table. Comparison in terms of AUC between relative abundance and presence/absence profiles with different classification algorithms (RF: Random Forest; Lasso; ENet: Elastic Net; LSVM: SVM with linear kernel; SVM: SVM with RBF kernel).

S2.4 Table. Comparison in terms of AUC between our results (using RF classification on the relative abundance profiles) and the ones reported in the original publications. In most of the cases, different classifier algorithms and/or input features were used in the original analysis. Original papers that did not conduct a classification analysis are not included in this table.

S2.5 Table. Results obtained from the classification process done on the 16s datasets. Comparison in terms of AUC, AUPRC, F1, precision, recall between relative abundance and presence/absence profiles at different threshold levels. Results are obtained using RF classification at the species-level taxonomic resolution.

S2.6 **Table. P-values (after FDR correction) obtained by testing differences in abundance of each species between controls and cases.**

S2.7 **Table. Number of statistically significant taxa (q< = 0.05) between cases and controls for each shotgun dataset and at varying input features (relative abundance vs presence/ absence profiles) and taxonomic level.**

S2.8 **Table. Number of statistically significant taxa (q< = 0.05) between cases and controls for each 16s dataset and at varying input features (relative abundance vs presence/absence profiles).**

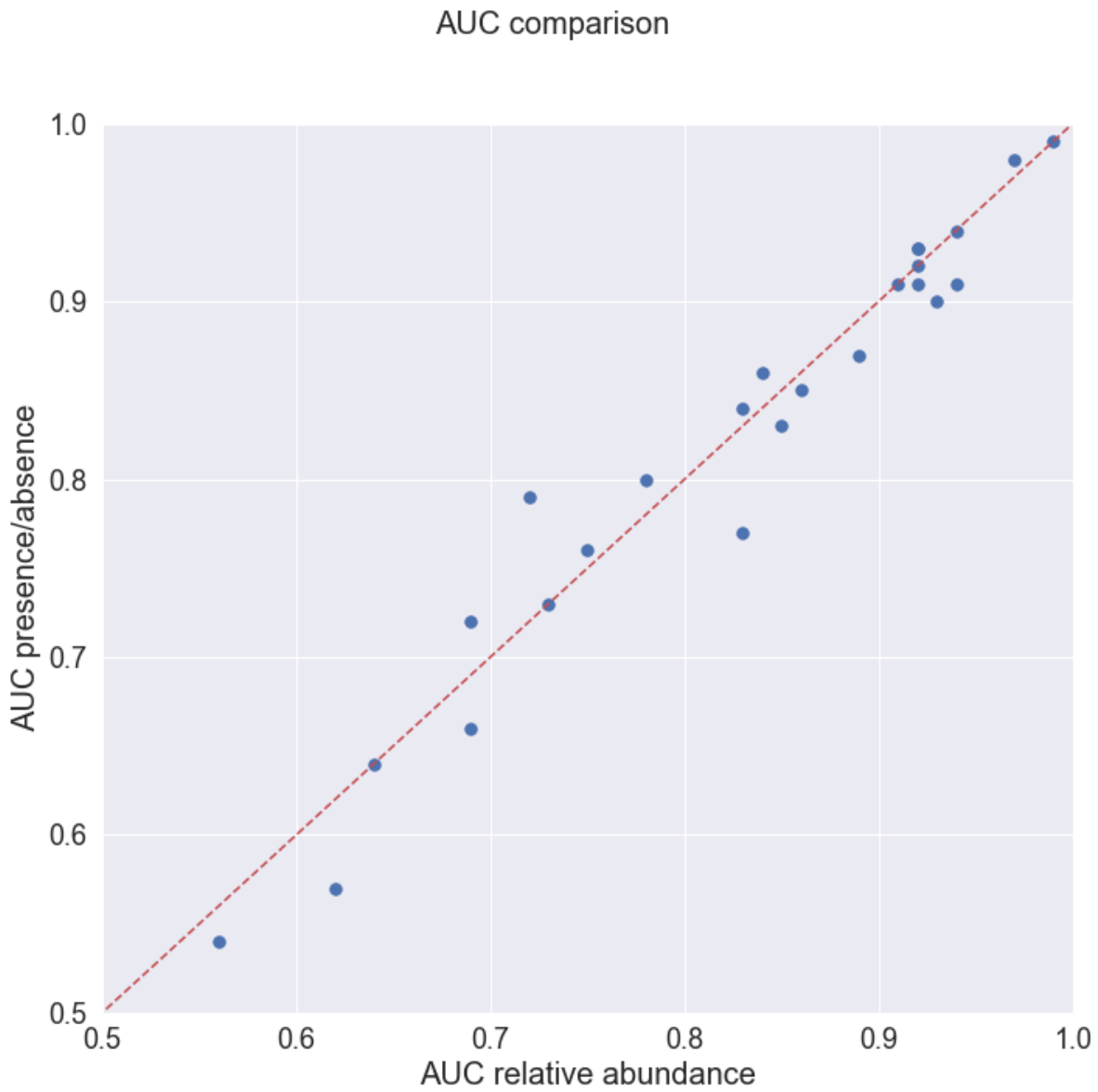
S2.9 **Table. Results obtained on three selected shotgun datasets after rarefying metagenomes at 1M reads.** Comparison in terms of AUC, F1, precision, recall, in addition to number of statistically significant taxa (q< = 0.05), between the results obtained classifying on the abundances matrix and the classification made on the presence/absence boolean matrix at different taxonomic levels (only at species level).

S2.10 **Table. Results obtained from the classification process done on the shotgun datasets.** Comparison in terms of AUC between the results obtained classifying at different taxonomic resolution levels. The results are obtained using the RF classifier on the relative abundances matrixes

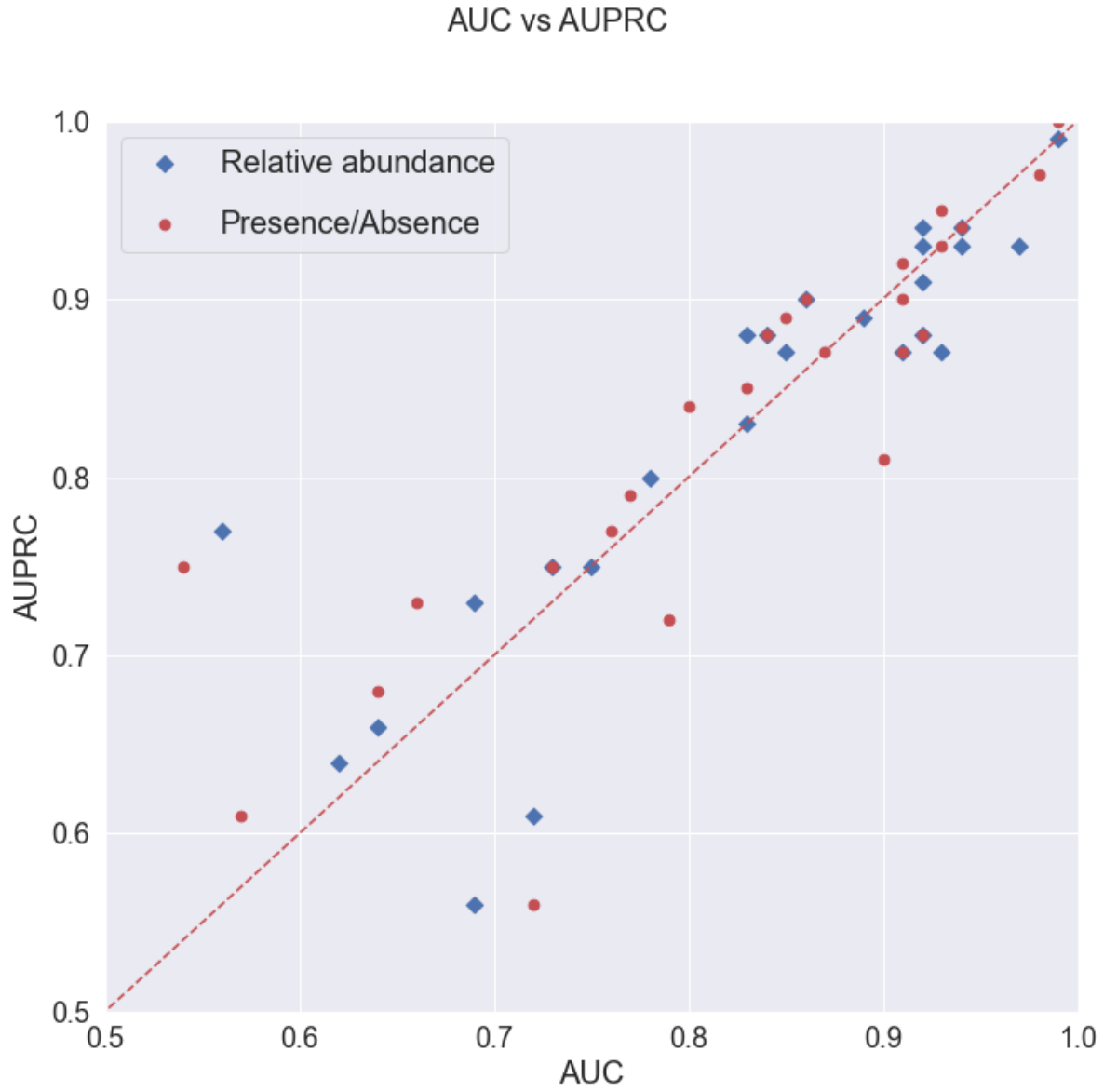
S2.11 **Table. Results obtained from the classification process done on the shotgun dataset.** Comparison in terms of AUC, F1, precision, recall between the results obtained classifying on the abundances matrix and the classification made on the presence/absence boolean matrix at different taxonomic levels (species, genus, etc).

S2.12 **Table. Results obtained by the LODO classification for datasets associated with CRC.** Comparison in terms of AUC obtained classifying thresholding the dataset at different levels and at different taxonomic levels

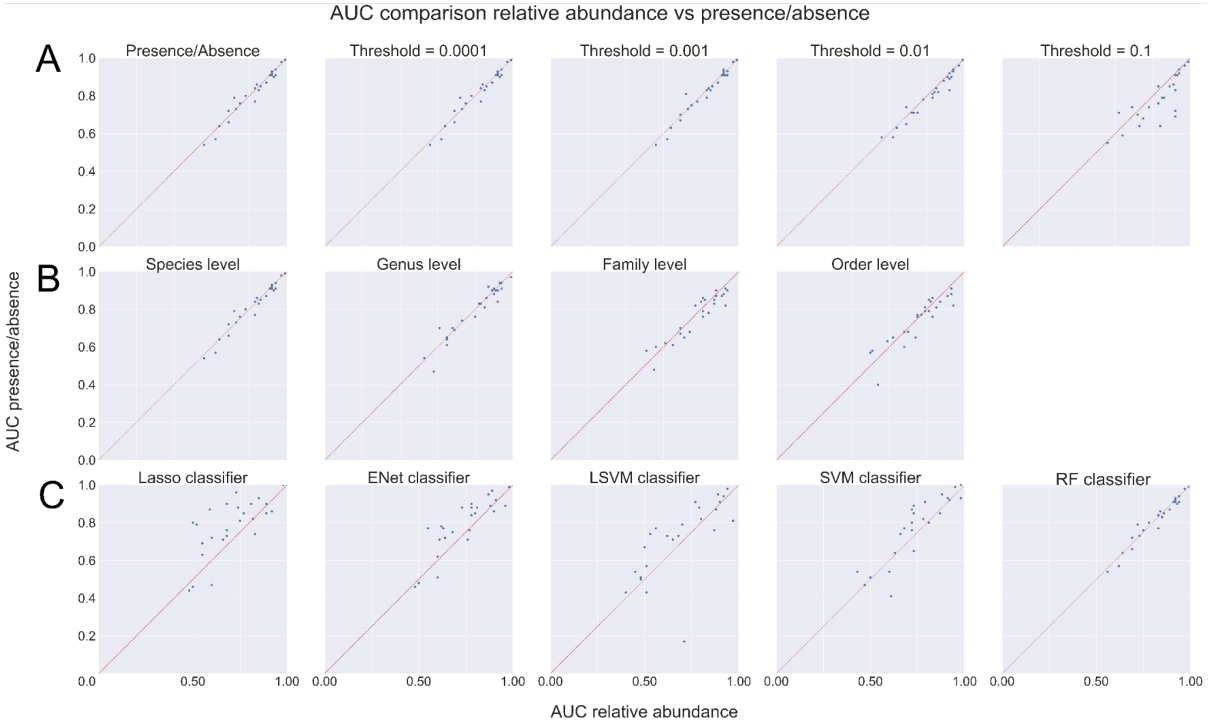
S2.13 Table. Comparison in terms of AUC, F1, precision, recall between the results obtained from different classifiers on the relative abundances matrix and on the presence absence boolean matrix (only at species level).



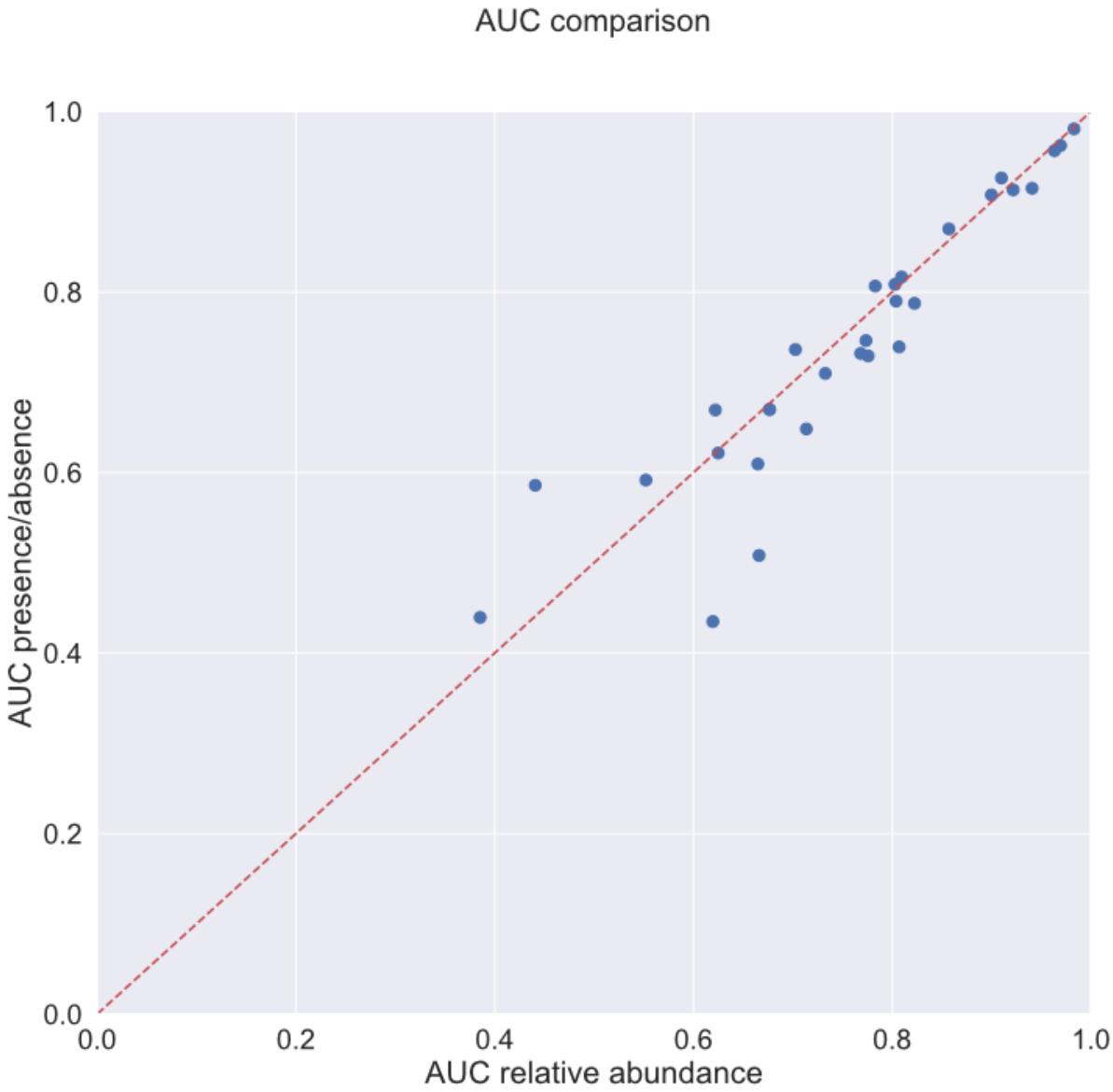
S2.1 Fig. Classification accuracies are robust to degradation from species-level relative abundance to presence/absence profiles in shotgun datasets. Comparison in terms of AUC between presence/absence and relative abundance profiles for the 25 case-control shotgun datasets.



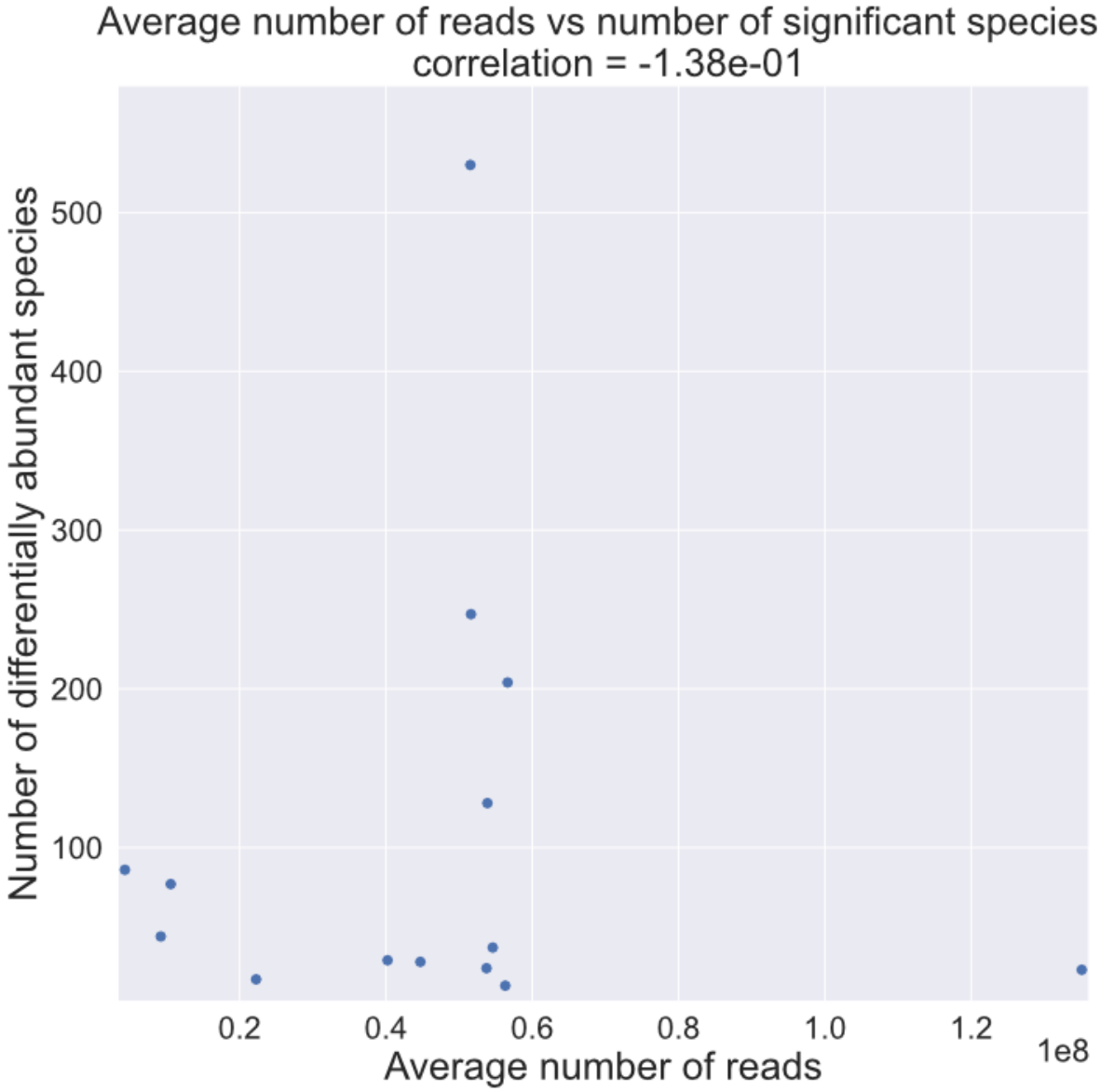
S2.2 Fig. AUC correlates well with AUPRC. Comparison in terms of classification accuracies between AUC (area under the curve) and AUPRC (area under the precision-recall curve) for the 25 case-control shotgun datasets and by considering relative abundance (in blue; Spearman correlation = 0.889) and presence/absence (in red; Spearman correlation = 0.918) profiles.



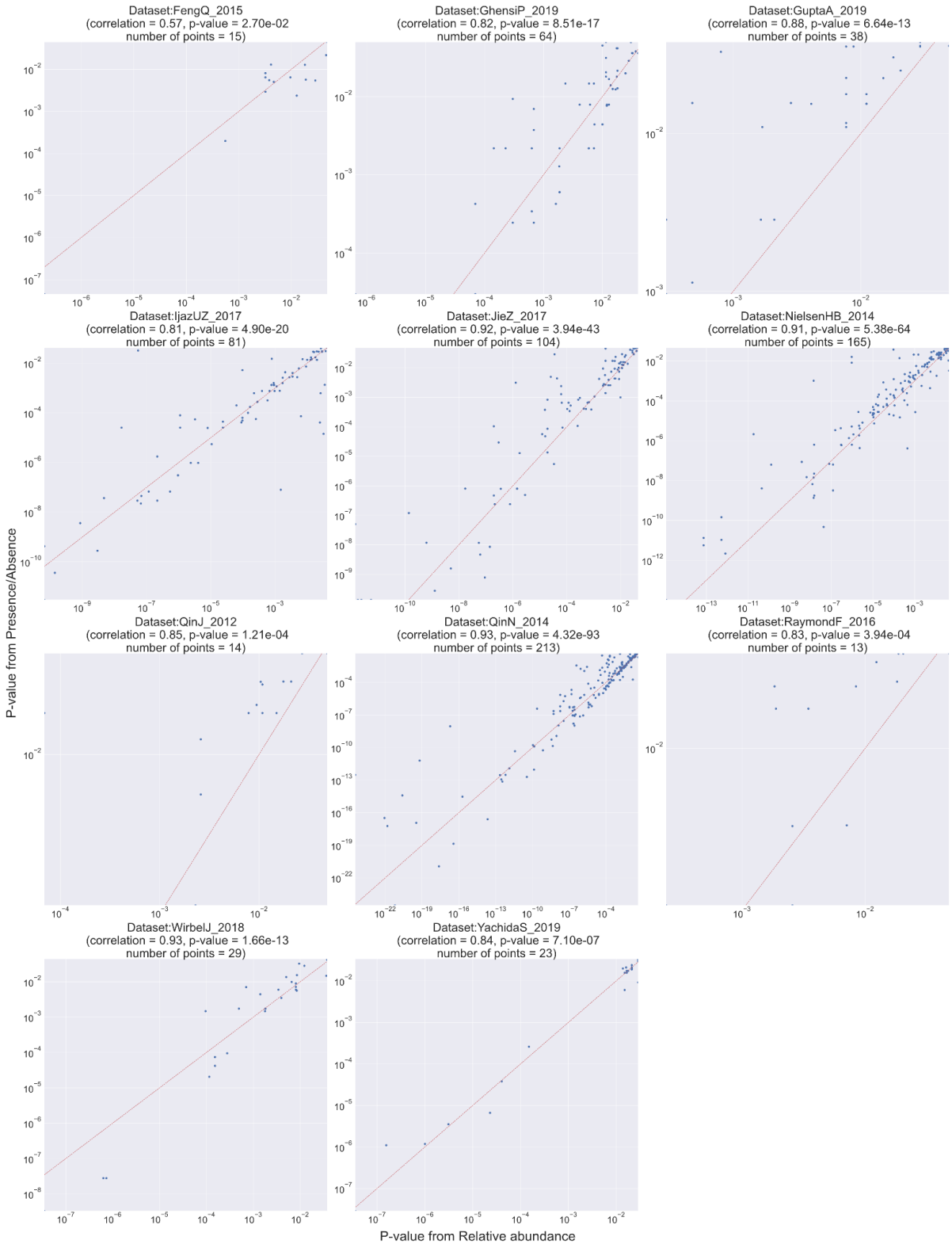
S2.3 Fig. Classification accuracies are robust to degradation from species-level relative abundance to presence/absence profiles in shotgun datasets. Comparison in terms of AUC between presence/absence and relative abundance profiles for the 25 case-control shotgun datasets by (A) thresholding at different relative abundance values (ranging from 0% to 0.1%), (B) changing taxonomic resolution (from species to order level), and (C) changing classification algorithm.



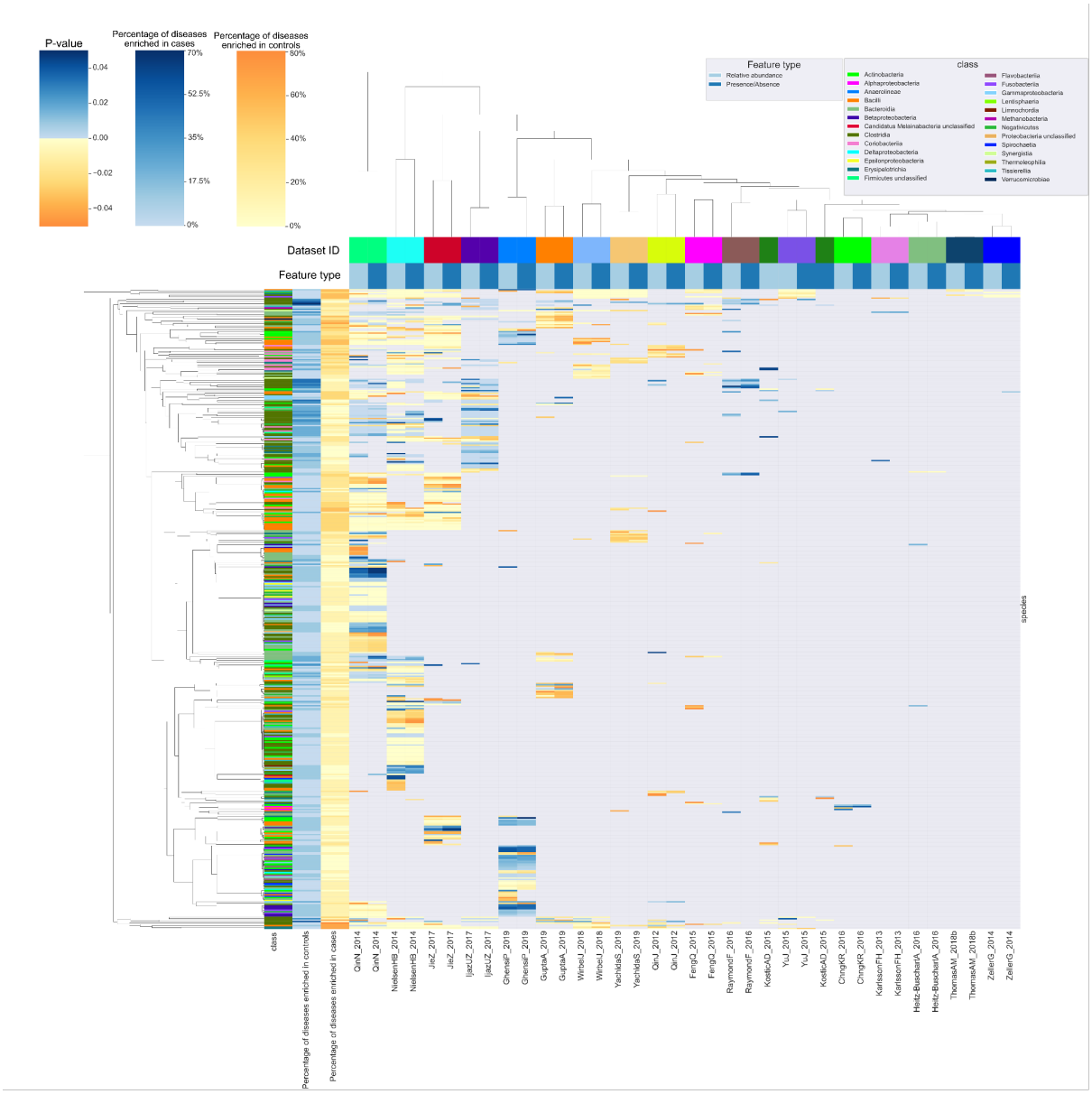
S2.4 Fig. Classification accuracies are robust to degradation from species-level relative abundance to presence/absence profiles in 16S rRNA datasets. Comparison in terms of AUC between presence/absence and relative abundance profiles for the 30 case-control 16 rRNA datasets.



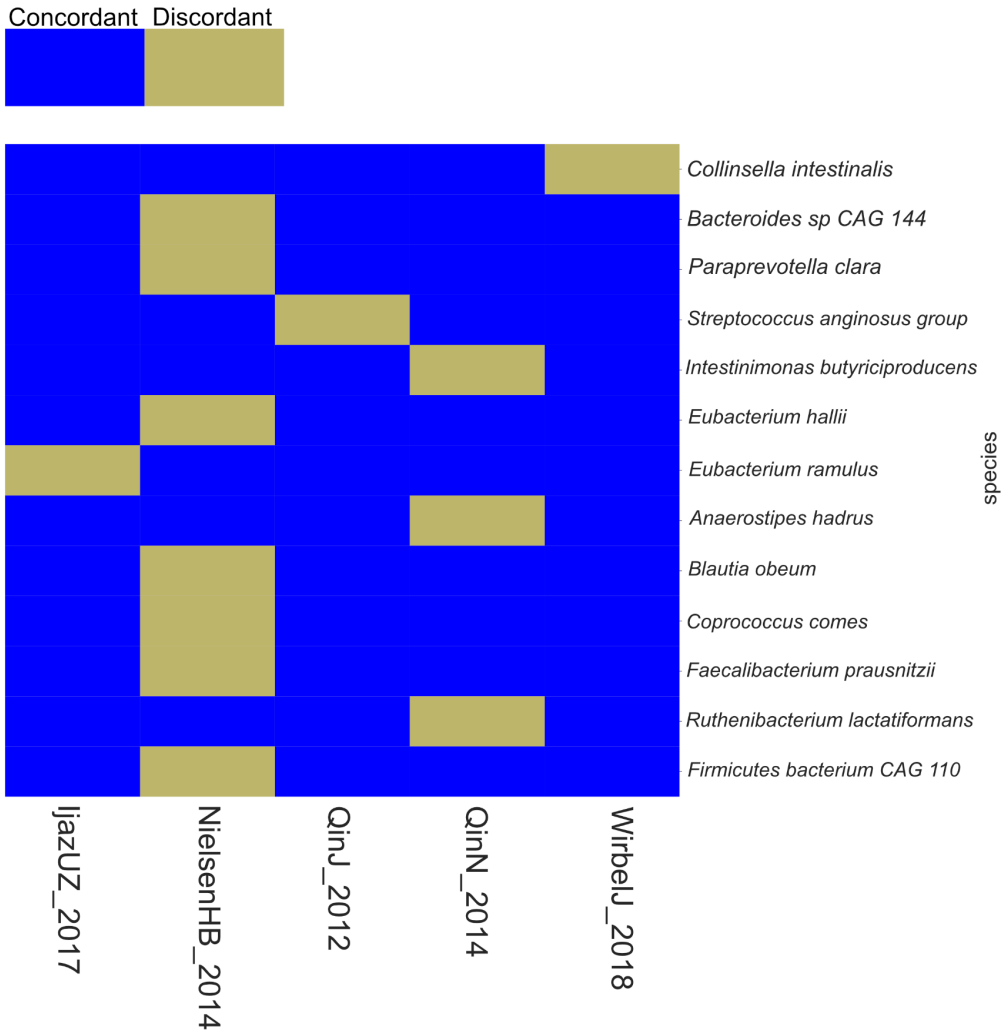
S2.5 Fig. Number of differentially abundant species has weak correlation with the average number of reads. Each dot represents one of the 26 case-control shotgun studies. The number of statistically significant species is computed on relative abundance profiles.



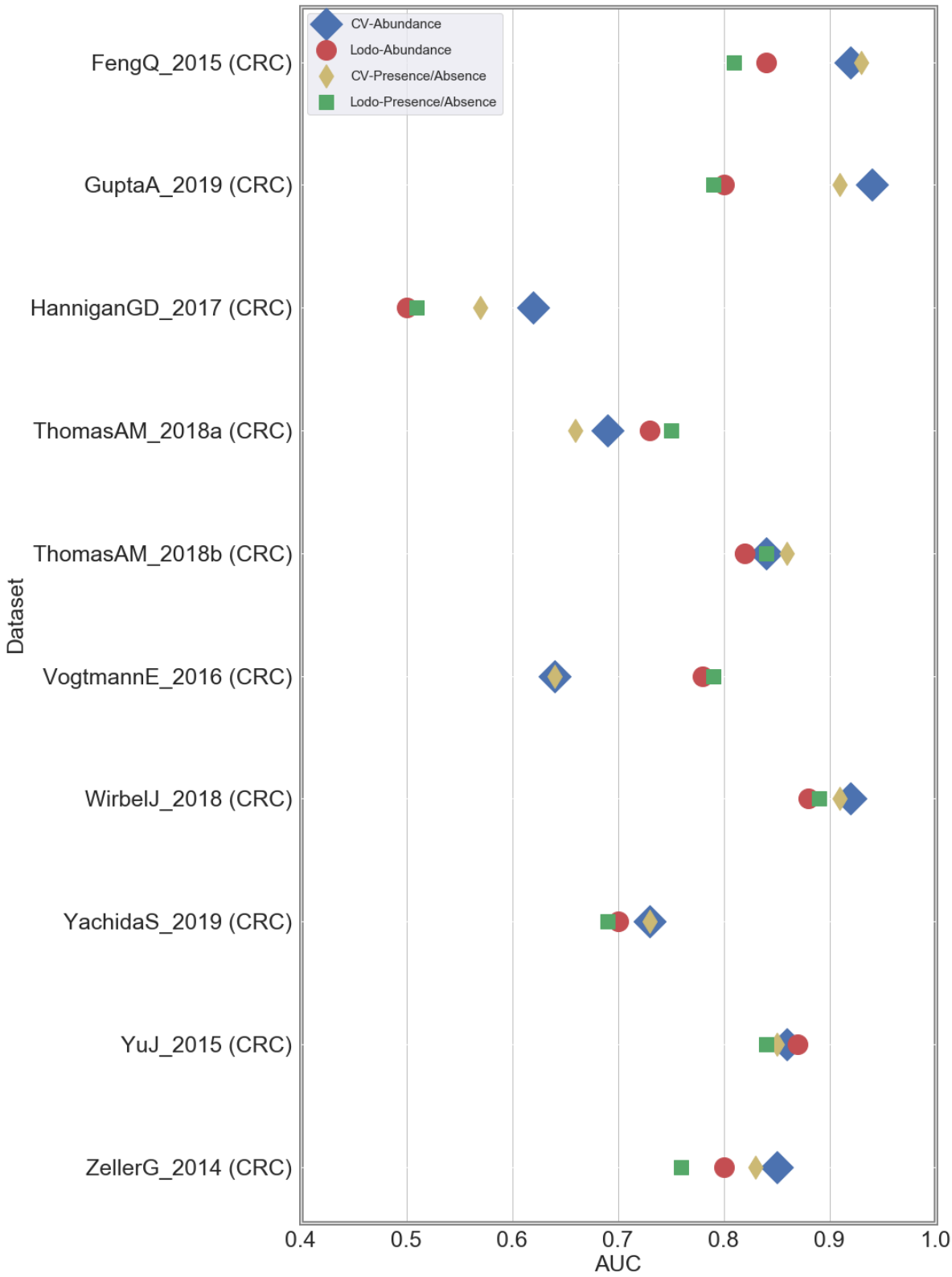
S2.6 Fig. P-values associated with statistically significant species correlate well between relative abundance and presence/absence profiles. Each dot represents a different taxa (i.e., species) and we report only species significant in at least one of the two data types. Only datasets with at least ten data points are shown.



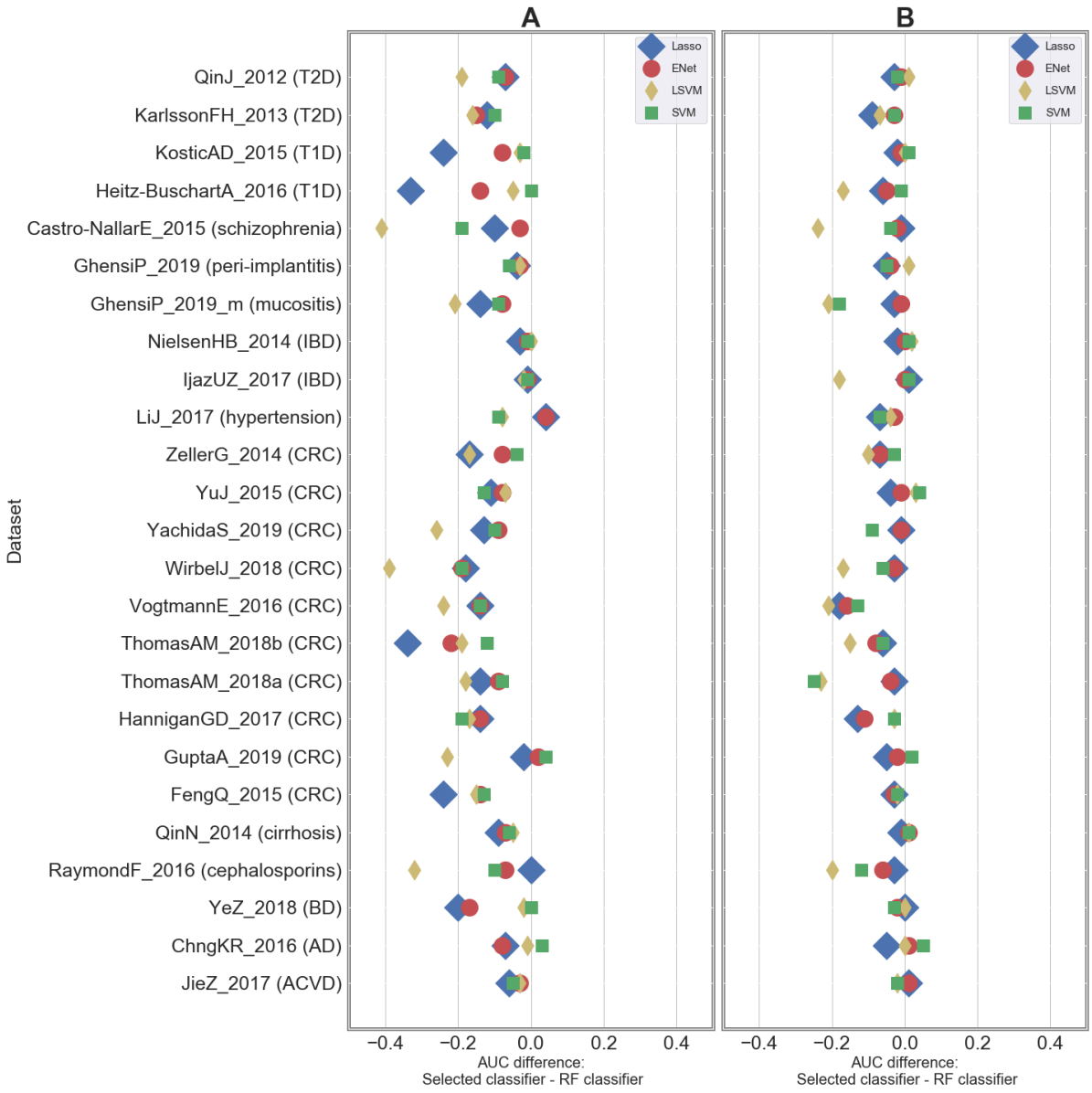
S2.7 Fig. Statistically significant taxa are consistent between relative abundance and presence/absence data on a per dataset basis. Heatmap generated on the p-values (after FDR correction; $p > 0.05$ in grey) obtained by applying statistical tests on the case-control metagenomic datasets. Only the 18 datasets with at least one discriminative taxa are reported. Left-most colorbar identifies the taxonomic class of each taxa. The two right-most colorbars indicate the percentage of diseases for which the species resulted to be enriched in controls (in green) and in cases (in red). This percentage is computed on a per disease basis, when multiple datasets are available for the same disease, the taxa is considered significant when detected as significant in at least one dataset.



S2.8 Fig. Statistically significant taxa from relative abundance and presence/absence profiles did not disagree across datasets. We identified discrepancies between case-enriched and control-enriched taxa derived from relative abundance and presence/absence data in only 1.74% of the statistically significant features, which were coming from just 5 datasets. No taxa disagreed across datasets



S2.9 Fig. Degradation of relative abundance profiles has a limited impact on both CV and LODO classification. AUC scores using RF as back-end classifiers on species-level relative abundance and corresponding presence/absence profiles in CV and LODO settings.



S2.10 Fig. RFs generally outperform other classifiers. Results on the 25 case-control shotgun studies by considering different classification algorithms. Difference in AUC between RFs and other classification methods on (A) the relative abundance and (B) the presence/absence profiles. A positive value indicates that the comparison method outperforms RFs

S3.1 Fig. FFQ questionnaire used to collect dietary data. Fac-simile of the paper version of the questionnaire (in Italian) that we administered to the subjects involved in the study to collect dietary habit information. This is a version provided by EPIC and that we adjusted graphically.

S3.2. Fig. Dietary habits report. Fac-simile of the dietary habits report provided to the subjects involved in the study that filled in the FFQ questionnaire.

7 Bibliography

1. Wernegreen JJ. Endosymbiosis. Curr Biol. 2012;22: R555–61.

2. McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, et al. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci U S A. 2013;110: 3229–3236.

3. Russell JA, Dubilier N, Rudgers JA. Nature's microbiome: introduction. Mol Ecol. 2014;23: 1225–1237.

4. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24: 392–400.

5. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449: 804–810.

6. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev. 2012;70 Suppl 1: S38–44.

7. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016;8: 51.

8. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. The application of ecological theory toward an understanding of the human microbiome. Science. 2012;336: 1255–1262.

9. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13: 260–270.

10. Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med. 2016;375: 2369–2379.

11. Zhou Y-H, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet. 2019;10: 579.

12. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol. 2021;22: 93.

13. Giliberti R, Cavaliere S, Mauriello IE, Ercolini D, Pasolli E. Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa. PLoS Comput Biol. 2022;18: e1010066.

14. Wade WG. Resilience of the oral microbiome. Periodontol 2000. 2021;86: 113–122.

15. Rosier BT, Marsh PD, Mira A. Resilience of the Oral Microbiota in Health: Mechanisms That Prevent Dysbiosis. J Dent Res. 2018;97: 371–380.

16. Laheij AMGA, Rozema FR, Brennan MT, von Bültzingslöwen I, van Leeuwen SJM, Potting C, et al. Long-Term Analysis of Resilience of the Oral Microbiome in Allogeneic Stem Cell Transplant Recipients. Microorganisms. 2022;10. doi:10.3390/microorganisms10040734

17. Wade WG. The oral microbiome in health and disease. Pharmacol Res. 2013;69: 137–143.

18. Downes J, Mantzourani M, Beighton D, Hooper S, Wilson MJ, Nicholson A, et al. Scardovia wiggsiae sp. nov., isolated from the human oral cavity and clinical material, and emended descriptions of the genus Scardovia and Scardovia inopinata. Int J Syst Evol Microbiol. 2011;61: 25–29.

19. Kaur R, Gilbert SC, Sheehy EC, Beighton D. Salivary levels of Bifidobacteria in caries-free and caries-active children. Int J Paediatr Dent. 2013;23: 32–38.

20. Munson MA, Banerjee A, Watson TF, Wade WG. Molecular analysis of the microflora associated with dental caries. J Clin Microbiol. 2004;42: 3023–3029.

21. Munson MA, Pitt-Ford T, Chong B, Weightman A, Wade WG. Molecular and cultural analysis of the microflora associated with endodontic infections. J Dent Res. 2002;81: 761–766.

22. Razavi A, Gmür R, Imfeld T, Zehnder M. Recovery of Enterococcus faecalis from cheese in the oral cavity of healthy subjects. Oral Microbiol Immunol. 2007;22: 248–251.

23. Kampfer J, Göhring TN, Attin T, Zehnder M. Leakage of food-borne Enterococcus faecalis through temporary fillings in a simulated oral environment. Int Endod J. 2007;40: 471–477.

24. Armitage GC. Development of a classification system for periodontal diseases and conditions. Ann Periodontol. 1999;4: 1–6.

25. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL Jr. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998;25: 134–144.

26. Kumar PS, Griffen AL, Barton JA, Paster BJ, Moeschberger ML, Leys EJ. New bacterial species associated with chronic periodontitis. J Dent Res. 2003;82: 338–344.

27. Mylonakis E, Calderwood SB. Infective endocarditis in adults. N Engl J Med. 2001;345: 1318–1330.

28. Marques da Silva R, Caugant DA, Josefsen R, Tronstad L, Olsen I. Characterization of Streptococcus constellatus strains recovered from a brain abscess and periodontal pockets in an immunocompromised patient. J Periodontol. 2004;75: 1720–1723.

29. Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Kehagia V, et al. Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with cystic fibrosis. J Clin Microbiol. 2006;44: 2601–2604.

30. Ahn J, Chen CY, Hayes RB. Oral microbiome and oral and gastrointestinal cancer risk. Cancer Causes Control. 2012;23: 399–404.

31. Ghensi P, Manghi P, Zolfo M, Armanini F, Pasolli E, Bolzan M, et al. Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. NPJ Biofilms Microbiomes. 2020;6: 47.

32. Heitz-Mayfield LJA. Peri-implant diseases: diagnosis and risk indicators. J Clin Periodontol. 2008;35: 292–304.

33. Lindhe J, Meyle J, Group D of European Workshop on Periodontology. Peri-implant diseases: Consensus Report of the Sixth European Workshop on Periodontology. J Clin Periodontol. 2008;35: 282–285.

34. Zitzmann NU, Berglundh T. Definition and prevalence of peri-implant diseases. J Clin Periodontol. 2008;35: 286–291.

35. Ramanauskaite A, Becker K, Schwarz F. Clinical characteristics of peri-implant mucositis and peri-implantitis. Clin Oral Implants Res. 2018;29: 551–556.

36. Renvert S, Persson GR, Pirih FQ, Camargo PM. Peri-implant health, peri-implant mucositis, and peri-implantitis: Case definitions and diagnostic considerations. J Periodontol. 2018;89 Suppl 1: S304–S312.

37. Bartold PM, Van Dyke TE. An appraisal of the role of specific bacteria in the initial pathogenesis of periodontitis. J Clin Periodontol. 2019;46: 6–11.

38. Van Dyke TE, Bartold PM, Reynolds EC. The Nexus Between Periodontal Inflammation and Dysbiosis. Front Immunol. 2020;11: 511.

39. Faveri M, Figueiredo LC, Shibli JA, Pérez-Chaparro PJ, Feres M. Microbiological diversity of peri-implantitis biofilms. Adv Exp Med Biol. 2015;830: 85–96.

40. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 2009;19: 1141–1152.

41. Heuer W, Kettenring A, Stumpp SN, Eberhard J, Gellermann E, Winkel A, et al. Metagenomic analysis of the peri-implant and periodontal microflora in patients with clinical signs of gingivitis or mucositis. Clin Oral Investig. 2012;16: 843–850.

42. Kumar PS, Mason MR, Brooker MR, O'Brien K. Pyrosequencing reveals unique microbial signatures associated with healthy and failing dental implants. J Clin Periodontol. 2012;39: 425–433.

43. Dabdoub SM, Tsigarida AA, Kumar PS. Patient-specific analysis of periodontal and peri-implant microbiomes. J Dent Res. 2013;92: 168S–75S.

44. Tamura N, Ochi M, Miyakawa H, Nakazawa F. Analysis of bacterial flora associated with peri-implantitis using obligate anaerobic culture technique and 16S rDNA gene sequence. Int J Oral Maxillofac Implants. 2013;28: 1521–1529.

45. Schaumann S, Staufenbiel I, Scherer R, Schilhabel M, Winkel A, Stumpp SN, et al. Pyrosequencing of supra- and subgingival biofilms from inflamed peri-implant and periodontal sites. BMC Oral Health. 2014;14: 157.

46. Albertini M, López-Cerero L, O'Sullivan MG, Chereguini CF, Ballesta S, Ríos V, et al. Assessment of periodontal and opportunistic flora in patients with peri-implantitis. Clin Oral Implants Res. 2015;26: 937–941.

47. Li Z-J, Wang S-G, Li Y-H, Tu D-X, Liu S-Y, Nie H-B, et al. [Study on Microbial Diversity of Peri-implantitis Subgingival by High-throughput Sequencing]. Sichuan Da Xue Xue Bao Yi Xue Ban. 2015;46: 568–572.

48. Belibasakis GN, Mir-Mari J, Sahrmann P, Sanz-Martin I, Schmidlin PR, Jung RE. Clinical association of Spirochaetes and Synergistetes with peri-implantitis. Clin Oral Implants Res. 2016;27: 656–661.

49. Apatzidou D, Lappin DF, Hamilton G, Papadopoulos CA, Konstantinidis A, Riggio MP. Microbiome of peri-implantitis affected and healthy dental sites in patients with a history of chronic periodontitis. Arch Oral Biol. 2017;83: 145–152.

50. Sanz-Martin I, Doolittle-Hall J, Teles RP, Patel M, Belibasakis GN, Hämmerle CHF, et al. Exploring the microbiome of healthy and diseased peri-implant sites using Illumina sequencing. J Clin Periodontol. 2017;44: 1274–1284.

51. Schincaglia GP, Hong BY, Rosania A, Barasz J, Thompson A, Sobue T, et al. Clinical, Immune, and Microbiome Traits of Gingivitis and Peri-implant Mucositis. J Dent Res. 2017;96: 47–55.

52. Sousa V, Nibali L, Spratt D, Dopico J, Mardas N, Petrie A, et al. Peri-implant and periodontal microbiome diversity in aggressive periodontitis patients: a pilot study. Clin Oral Implants Res. 2017;28: 558–570.

53. Al-Ahmad A, Muzafferiy F, Anderson AC, Wölber JP, Ratka-Krüger P, Fretwurst T, et al. Shift of microbial composition of peri-implantitis-associated oral biofilm as revealed by 16S rRNA gene cloning. J Med Microbiol. 2018;67: 332–340.

54. Kröger A, Hülsmann C, Fickl S, Spinell T, Hüttig F, Kaufmann F, et al. The severity of human peri-implantitis lesions correlates with the level of submucosal microbial dysbiosis. J Clin Periodontol. 2018;45: 1498–1509.

55. Pimentel SP, Fontes M, Ribeiro FV, Corrêa MG, Nishii D, Cirano FR, et al. Smoking habit modulates peri-implant microbiome: A case-control study. J Periodontal Res. 2018;53: 983–991.

56. Koyanagi T, Sakamoto M, Takeuchi Y, Maruyama N, Ohkuma M, Izumi Y. Comprehensive microbiological findings in peri-implantitis and periodontitis. J Clin Periodontol. 2013;40: 218–226.

57. Yu X-L, Chan Y, Zhuang L, Lai H-C, Lang NP, Keung Leung W, et al. Intra-oral single-site comparisons of periodontal and peri-implant microbiota in health and disease. Clin Oral Implants Res. 2019;30: 760–776.

58. Flamm SL. Hepatitis C virus infection in special populations. Gastroenterol Hepatol . 2013;9: 823–825.

59. Zheng H, Xu L, Wang Z, Li L, Zhang J, Zhang Q, et al. Subgingival microbiome in patients with healthy and ailing dental implants. Sci Rep. 2015;5: 10948.

60. Rahim MI, Winkel A, Ingendoh-Tsakmakidis A, Lienenklaus S, Falk CS, Eisenburger M, et al. Bacterial-Specific Induction of Inflammatory Cytokines Significantly Decreases upon Dual Species Infections of Implant Materials with Periodontal Pathogens in a Mouse Model. Biomedicines. 2022;10. doi:10.3390/biomedicines10020286

61. Tsigarida AA, Dabdoub SM, Nagaraja HN, Kumar PS. The Influence of Smoking on the Peri-Implant Microbiome. J Dent Res. 2015;94: 1202–1217.

62. Yu X-L, Chan Y, Zhuang L-F, Lai H-C, Lang NP, Lacap-Bugler DC, et al. Distributions of Synergistetes in clinically-healthy and diseased periodontal and peri-implant niches. Microb Pathog. 2016;94: 90–103.

63. Cornejo Ulloa P, van der Veen MH, Krom BP. Review: modulation of the oral microbiome by the host to promote ecological balance. Odontology. 2019;107: 437–448.

64. Kilian M, Chapple ILC, Hannig M, Marsh PD, Meuric V, Pedersen AML, et al. The oral microbiome - an update for oral healthcare professionals. Br Dent J. 2016;221: 657–666.

65. Chapple ILC, Bouchard P, Cagetti MG, Campus G, Carra M-C, Cocco F, et al. Interaction of lifestyle, behaviour or systemic diseases with dental caries and periodontal diseases: consensus report of group 2 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. J Clin Periodontol. 2017;44 Suppl 18: S39–S51.

66. Barbour SE, Nakashima K, Zhang JB, Tangada S, Hahn CL, Schenkein HA, et al. Tobacco and smoking: environmental factors that modify the host response (immune system) and have an impact on periodontal health. Crit Rev Oral Biol Med. 1997;8: 437–460.

67. Heasman PA, Hughes FJ. Drugs, medications and periodontal disease. Br Dent J. 2014;217: 411–419.

68. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45: 450–5, 455e1.

69. Cross KL, Chirania P, Xiong W, Beall CJ, Elkins JG, Giannone RJ, et al. Insights into the Evolution of Host Association through the Isolation and Characterization of a Novel Human Periodontal Pathobiont, Desulfobulbus oralis. MBio. 2018;9. doi:10.1128/mBio.02061-17

70. Cornejo OE, Lefébure T, Bitar PDP, Lang P, Richards VP, Eilertson K, et al. Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans. Mol Biol Evol. 2013;30: 881–893.

71. Gillings MR, Paulsen IT, Tetu SG. Ecology and Evolution of the Human Microbiota: Fire, Farming and Antibiotics. Genes . 2015;6: 841–857.

72. Marsh P, Martin MV, Lewis MAO, Williams D, Rogers H, Wilson M. Marsh and Martin’s Oral Microbiology. Elsevier; 2016.

73. Pitts NB, Zero DT, Marsh PD, Ekstrand K, Weintraub JA, Ramos-Gomez F, et al. Dental caries. Nat Rev Dis Primers. 2017;3: 17030.

74. Takahashi N, Nyvad B. The role of bacteria in the caries process: ecological perspectives. J Dent Res. 2011;90: 294–303.

75. Hansen TH, Kern T, Bak EG, Kashani A, Allin KH, Nielsen T, et al. Impact of a vegan diet on the human salivary microbiota. Sci Rep. 2018;8: 5847.

76. Murtaza N, Burke LM, Vlahovich N, Charlesson B, O’Neill HM, Ross ML, et al. Analysis of the Effects of Dietary Pattern on the Oral Microbiome of Elite Endurance Athletes. Nutrients. 2019;11. doi:10.3390/nu11030614

77. Huang CB, Alimova Y, Myers TM, Ebersole JL. Short- and medium-chain fatty acids exhibit antimicrobial activity for oral microorganisms. Arch Oral Biol. 2011;56: 650–654.

78. Cui H, Zhang X. Alignment-free supervised classification of metagenomes by recursive SVM. BMC Genomics. 2013;14: 641.

79. Sze MA, Schloss PD. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. MBio. 2016;7. doi:10.1128/mBio.01018-16

80. Vatanen T, Franzosa EA, Schwager R, Tripathi S, Arthur TD, Vehik K, et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. Nature. 2018;562: 589–594.

81. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10: 766.

82. Eloe-Fadrosh EA, Rasko DA. The human microbiome: from symbiosis to pathogenesis. Annu Rev Med. 2013;64: 145–163.

83. McCoubrey LE, Elbadawi M, Orlu M, Gaisford S, Basit AW. Harnessing machine learning for development of microbiome therapeutics. Gut Microbes. 2021;13: 1–20.

84. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486: 207–214.

85. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464: 59–65.

86. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23: 1704–1714.

87. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome. 2013;1: 11.

88. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8: 1784.

89. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. 2016;12: e1004977.

90. Armour CR, Nayfach S, Pollard KS, Sharpton TJ. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. mSystems. 2019;4. doi:10.1128/mSystems.00332-18

91. Vangay P, Hillmann BM, Knights D. Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. Gigascience. 2019;8. doi:10.1093/gigascience/giz042

92. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of Machine Learning in Human

Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Front Microbiol. 2021;12: 634511.

93. Ditzler G, Morrison JC, Lan Y, Rosen GL. Fizzy: feature subset selection for metagenomics. BMC Bioinformatics. 2015;16: 358.

94. Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. BMC Bioinformatics. 2018;19: 227.

95. Wu H, Cai L, Li D, Wang X, Zhao S, Zou F, et al. Metagenomics Biomarkers Selected for Prediction of Three Different Diseases in Chinese Population. Biomed Res Int. 2018;2018: 2936257.

96. Bang S, Yoo D, Kim S-J, Jhang S, Cho S, Kim H. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. Sci Rep. 2019;9: 10189.

97. Wang X-W, Liu Y-Y. Comparative study of classifiers for human microbiome data. Med Microecol. 2020;4. doi:10.1016/j.medmic.2020.100013

98. LaPierre N, Ju CJ-T, Zhou G, Wang W. MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods. 2019;166: 74–82.

99. Díez López C, Vidaki A, Ralf A, Montiel González D, Radjabzadeh D, Kraaij R, et al. Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. Forensic Sci Int Genet. 2019;41: 72–82.

100. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep. 2020;10: 6026.

101. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15. doi:10.1098/rsif.2017.0387

102. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife. 2021;10. doi:10.7554/eLife.65088

103. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14: 1023–1024.

104. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun. 2017;8: 845.

105. Chng KR, Tay ASL, Li C, Ng AHQ, Wang J, Suri BK, et al. Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. Nat Microbiol. 2016;1: 16106.

106. Ye Z, Zhang N, Wu C, Zhang X, Wang Q, Huang X, et al. A metagenomic study of the gut microbiome in Behcet's disease. Microbiome. 2018;6: 135.

107. Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. 2016;10: 707–720.

108. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513: 59–64.

109. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun. 2015;6: 6528.

110. Gupta A, Dhakan DB, Maji A, Saxena R, P K VP, Mahajan S, et al. Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. mSystems. 2019;4. doi:10.1128/mSystems.00438-19

111. Hannigan GD, Duhaime MB, Ruffin MT 4th, Koumpouras CC, Schloss PD. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. MBio. 2018;9. doi:10.1128/mBio.02248-18

112. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25: 667–678.

113. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, et al. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One. 2016;11: e0155362.

114. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25: 679–689.

115. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25: 968–976.

116. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2017;66: 70–78.

117. Li J, Zhao F, Wang Y, Chen J, Tao J, Tian G, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017;5: 14.

118. Ijaz UZ, Quince C, Hanske L, Loman N, Calus ST, Bertz M, et al. The distinct features of microbial “dysbiosis” of Crohn’s disease do not occur to the same extent in their unaffected, genetically-linked kindred. PLoS One. 2017;12: e0172605.

119. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32: 822–828.

120. Castro-Nallar E, Bendall ML, Pérez-Losada M, Sabuncyan S, Severance EG, Dickerson FB, et al. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. PeerJ. 2015;3: e1140.

121. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2: 16180.

122. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015;17: 260–273.

123. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498: 99–103.

124. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490: 55–60.

125. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Nat Commun. 2019;10: 2719.

126. Ling W, Zhao N, Plantinga AM, Launer LJ, Fodor AA, Meyer KA, et al. Powerful and robust non-parametric association testing for microbiome data via a zero-inflated quantile approach (ZINQ). Microbiome. 2021;9: 181.

127. Meslier V, Laiola M, Roager HM, De Filippis F, Roume H, Quinquis B, et al. Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. Gut. 2020;69: 1258–1268.

128. Pandit L, Cox LM, Malli C, D'Cunha A, Rooney T, Lokhande H, et al. Clostridium bolteae is elevated in neuromyelitis optica spectrum disorder in India and shares sequence similarity with AQP4. Neurol Neuroimmunol Neuroinflamm. 2021;8. doi:10.1212/NXI.0000000000000907

129. Tamanai-Shacoori Z, Smida I, Bousarghin L, Loreal O, Meuric V, Fong SB, et al. Roseburia spp.: a marker of health? Future Microbiol. 2017;12: 157–170.

130. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. 1996;58: 267–288.

131. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67: 301–320.

132. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297.

133. Reiman D, Metwally AA, Sun J, Dai Y. PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data. IEEE J Biomed Health Inform. 2020;24: 2993–3001.

134. Rahman MA, Rangwala H. IDMIL: an alignment-free Interpretable Deep Multiple Instance Learning (MIL) for predicting disease from whole-metagenomic data. Bioinformatics. 2020;36: i39–i47.

135. Bang G, Kristoffersen T. Dental caries and diet in an Alaskan Eskimo population. Scand J Dent Res. 1972;80: 440–444.

136. Touger-Decker R, van Loveren C. Sugars and dental caries. Am J Clin Nutr. 2003;78: 881S–892S.

137. Sheiham A, James WPT. A new understanding of the relationship between sugars, dental caries and fluoride use: implications for limits on sugars consumption. Public Health Nutr. 2014;17: 2176–2184.

138. Pindborg JJ, Bhat M, Roed-Petersen B. Oral changes in South Indian children with severe protein deficiency. J Periodontol. 1967;38: 218–221.

139. Hujoel PP, Lingström P. Nutrition, dental caries and periodontal disease: a narrative review. J Clin Periodontol. 2017;44 Suppl 18: S79–S84.

140. Bingham SA. Limitations of the various methods for collecting dietary intake data. Ann Nutr Metab. 1991;35: 117–127.

141. Mackerras D, Margetts BM. Nutritional Epidemiology. Handbook of Epidemiology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. pp. 999–1042.

142. Willett W. Nutritional Epidemiology. OUP USA; 2013.

143. Teufel NI. Development of culturally competent food-frequency questionnaires. Am J Clin Nutr. 1997;65: 1173S–1178S.

144. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5: 1113–1124.

145. Agnoli C, Krogh V, Grioni S, Sieri S, Palli D, Masala G, et al. A priori-defined dietary patterns are associated with reduced risk of stroke in a large Italian cohort. J Nutr. 2011;141: 1552–1558.

146. Huijbregts PP, Feskens EJ, Räsänen L, Fidanza F, Alberti-Fidanza A, Nissinen A, et al. Dietary patterns and cognitive function in elderly men in Finland, Italy and The Netherlands. Eur J Clin Nutr. 1998;52: 826–831.

147. Huijbregts P, Feskens E, Räsänen L, Fidanza F, Nissinen A, Menotti A, et al. Dietary pattern and 20 year mortality in elderly men in Finland, Italy, and The Netherlands: longitudinal cohort study. BMJ. 1997;315: 13–17.

148. Berglundh T, Jepsen S, Stadlinger B, Terheyden H. Peri-implantitis and its prevention. Clin Oral Implants Res. 2019;30: 150–155.

149. Khammissa RAG, Feller L, Meyerov R, Lemmer J. Peri-implant mucositis and periimplantitis : clinical and histopathological characteristics and treatment : review. In: South African Dental Journal [Internet]. [cited 23 Aug 2022]. Available: https://journals.co.za/doi/10.10520/EJC144737

150. Schwarz F, Derks J, Monje A, Wang H-L. Peri-implantitis. J Clin Periodontol. 2018;45 Suppl 20: S246–S266.

151. O'brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual Quant. 2007;41: 673–690.

152. Tay R. Correlation, Variance Inflation and Multicollinearity in Regression Model. Journal of the Eastern Asia Society for Transportation Studies. 2017;12: 2006–2015.

153. Marcoulides KM, Raykov T. Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods. Educ Psychol Meas. 2019;79: 874–882.

154. Sgolastra F, Petrucci A, Severino M, Gatto R, Monaco A. Smoking and the risk of peri-implantitis. A systematic review and meta-analysis. Clin Oral Implants Res. 2015;26: e62–e67.

155. Ferreira SD, Silva GLM, Cortelli JR, Costa JE, Costa FO. Prevalence and risk variables for peri-implant disease in Brazilian subjects. J Clin Periodontol. 2006;33: 929–935.

156. Roos-Jansåker A-M, Lindahl C, Renvert H, Renvert S. Nine- to fourteen-year follow-up of implant treatment. Part II: presence of peri-implant lesions. J Clin Periodontol. 2006;33: 290–295.

157. Koldsland OC, Scheie AA, Aass AM. Prevalence of peri-implantitis related to severity of the disease with different degrees of bone loss. J Periodontol. 2010;81: 231–238.

158. Rodriguez-Argueta OF, Figueiredo R, Valmaseda-Castellon E, Gay-Escoda C. Postoperative complications in smoking patients treated with implants: a retrospective study. J Oral Maxillofac Surg. 2011;69: 2152–2157.

159. Al-Shammari KF, Moussa MA, Al-Ansari JM, Al-Duwairy YS, Honkala EJ. Dental patient awareness of smoking effects on oral health: Comparison of smokers and non-smokers. J Dent. 2006;34: 173–178.

160. Bloom B, Adams PF, Cohen RA, Simile C. Smoking and oral health in dentate adults aged 18-64. NCHS Data Brief. 2012; 1–8.

161. Telivuo M, Kallio P, Berg MA, Korhonen HJ, Murtomaa H. Smoking and oral health: a population survey in Finland. J Public Health Dent. 1995;55: 133–138.

162. Setia S, Pannu P, Gambhir RS, Galhotra V, Ahluwalia P, Sofat A. Correlation of oral hygiene practices, smoking and oral health conditions with self perceived halitosis amongst undergraduate dental students. J Nat Sci Biol Med. 2014;5: 67–72.

163. Ide R, Mizoue T, Ueno K, Fujino Y, Yoshimura T. [Relationship between cigarette smoking and oral health status]. Sangyo Eiseigaku Zasshi. 2002;44: 6–11.

164. Mombelli A, Müller N, Cionca N. The epidemiology of peri-implantitis. Clin Oral Implants Res. 2012;23 Suppl 6: 67–76.

165. Algraffee H, Borumandi F, Cascarini L. Peri-implantitis. Br J Oral Maxillofac Surg. 2012;50: 689–694.

166. Dreyer H, Grischke J, Tiede C, Eberhard J, Schweitzer A, Toikkanen SE, et al. Epidemiology and risk factors of peri-implantitis: A systematic review. J Periodontal Res. 2018;53: 657–681.

167. Haas R, Haimböck W, Mailath G, Watzek G. The relationship of smoking on peri-implant tissue: a retrospective study. J Prosthet Dent. 1996;76: 592–596.

168. Adibrad M, Shahabuei M, Sahabi M. Significance of the width of keratinized mucosa on the health status of the supporting tissue around implants supporting overdentures. J Oral Implantol. 2009;35: 232–237.

169. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12: R60.

170. Yan T, Tourangeau R. Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. Appl Cogn Psychol. 2008;22: 51–68.

171. Gigliotti L, Dietsch A. Does Age Matter? The Influence of Age on Response Rates in a Mixed-Mode Survey. Hum Dimensions Wildl. 2014;19: 280–287.

172. Whitelock E, Ensaff H. On Your Own: Older Adults' Food Choice and Dietary Habits. Nutrients. 2018;10. doi:10.3390/nu10040413

173. Sergi G, Bano G, Pizzato S, Veronese N, Manzato E. Taste loss in the elderly: Possible implications for dietary habits. Crit Rev Food Sci Nutr. 2017;57: 3684–3689.

174. Nuttall FQ. Body Mass Index: Obesity, BMI, and Health: A Critical Review. Nutr Today. 2015;50: 117–128.

175. Grier T, Canham-Chervak M, Sharp M, Jones BH. Does body mass index misclassify physically active young men. Prev Med Rep. 2015;2: 483–487.

176. Abramowitz MK, Hall CB, Amodu A, Sharma D, Androga L, Hawkins M. Muscle mass, BMI, and mortality among adults in the United States: A population-based cohort study. PLoS One. 2018;13: e0194697.

177. Bärebring L, Palmqvist M, Winkvist A, Augustin H. Gender differences in perceived food healthiness and food avoidance in a Swedish population-based survey: a cross sectional study. Nutr J. 2020;19: 140.

178. Rolls BJ, Fedoroff IC, Guthrie JF. Gender differences in eating behavior and body weight regulation. Health Psychol. 1991;10: 133–142.

179. Manippa V, Padulo C, van der Laan LN, Brancucci A. Gender Differences in Food Choice: Effects of Superior Temporal Sulcus Stimulation. Front Hum Neurosci. 2017;11: 597.

180. Chung S-J, Hoerr S, Levine R, Coleman G. Processes underlying young women's decisions to eat fruits and vegetables. J Hum Nutr Diet. 2006;19: 287–298.

181. de Graaf MA, Jager KJ, Zoccali C, Dekker FW. Matching, an appealing method to avoid confounding? Nephron Clin Pract. 2011;118: c315–8.

182. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench. 2012;5: 79–83.

183. Seck T, Moreau JL. [Dental lesions. After-effects of kwashiorkor]. Inf Dent. 1982;64: 1259–1268.

184. Navia JM. Carbohydrates and dental health. Am J Clin Nutr. 1994;59: 719S–727S.

185. Meurman JH, Rytömaa I, Kari K, Laakso T, Murtomaa H. Salivary pH and glucose after consuming various beverages, including sugar-containing drinks. Caries Res. 1987;21: 353–359.

186. Burt BA, Pai S. Sugar consumption and caries risk: a systematic review. J Dent Educ. 2001;65: 1017–1023.

187. Shen T-CD. Diet and Gut Microbiota in Health and Disease. 2017. pp. 117–126.

188. Ramos S, Martín MÁ. Impact of diet on gut microbiota. Current Opinion in Food Science. 2021;37: 83–90.

189. Zhang N, Ju Z, Zuo T. Time for food: The impact of diet on gut microbiota and human health. Nutrition. 2018;51-52: 80–85.

190. Nakayama J, Yamamoto A, Palermo-Conde LA, Higashi K, Sonomoto K, Tan J, et al. Impact of Westernized Diet on Gut Microbiota in Children on Leyte Island. Front Microbiol. 2017;8: 197.

191. Flint HJ, Duncan SH, Scott KP, Louis P. Links between diet, gut microbiota composition and gut metabolism. Proc Nutr Soc. 2015;74: 13–22.

192. Murphy EA, Velazquez KT, Herbert KM. Influence of high-fat diet on gut microbiota: a driving force for chronic disease risk. Curr Opin Clin Nutr Metab Care. 2015;18: 515–520.

193. Moreira APB, Texeira TFS, Ferreira AB, Peluzio M do CG, Alfenas R de CG. Influence of a high-fat diet on gut microbiota, intestinal permeability and metabolic endotoxaemia. Br J Nutr. 2012;108: 801–809.

194. Kim K-A, Gu W, Lee I-A, Joh E-H, Kim D-H. High fat diet-induced gut microbiota exacerbates inflammation and obesity in mice via the TLR4 signaling pathway. PLoS One. 2012;7: e47713.

195. Wu GD, Compher C, Chen EZ, Smith SA, Shah RD, Bittinger K, et al. Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. Gut. 2016;65: 63–72.

196. Sanz Y. Effects of a gluten-free diet on gut microbiota and immune function in healthy adult humans. Gut Microbes. 2010;1: 135–137.

197. Klement RJ, Pazienza V. Impact of Different Types of Diet on Gut Microbiota Profiles and Cancer Prevention and Treatment. Medicina . 2019;55. doi:10.3390/medicina55040084

198. Scott KP, Gratz SW, Sheridan PO, Flint HJ, Duncan SH. The influence of diet on the gut microbiota. Pharmacol Res. 2013;69: 52–60.

199. De Almeida CV, de Camargo MR, Russo E, Amedei A. Role of diet and gut microbiota on colorectal cancer immunomodulation. World J Gastroenterol. 2019;25: 151–162.

200. Guirro M, Costa A, Gual-Grau A, Herrero P, Torrell H, Canela N, et al. Effects from diet-induced gut microbiota dysbiosis and obesity can be ameliorated by fecal microbiota transplantation: A multiomics approach. PLoS One. 2019;14: e0218143.

201. De Angelis M, Garruti G, Minervini F, Bonfrate L, Portincasa P, Gobbetti M. The Food-gut Human Axis: The Effects of Diet on Gut Microbiota and Metabolome. Curr Med Chem. 2019;26: 3567–3583.

202. Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med. 2021;27: 321–332.

203. Moynihan PJ. The role of diet and nutrition in the etiology and prevention of oral diseases. Bull World Health Organ. 2005;83: 694–699.

204. Rugg-Gunn AJ. Nutrition, diet and oral health. J R Coll Surg Edinb. 2001;46: 320–328.

205. Scardina GA, Messina P. Good oral health and diet. J Biomed Biotechnol. 2012;2012: 720692.

206. Contaldo M, Itro A, Lajolo C, Gioco G, Inchingolo F, Serpico R. Overview on Osteoporosis, Periodontitis and Oral Dysbiosis: The Emerging Role of Oral Microbiota. NATO Adv Sci Inst Ser E Appl Sci. 2020;10: 6000.

207. Jiao Y, Hasegawa M, Inohara N. The Role of Oral Pathobionts in Dysbiosis during Periodontitis Development. J Dent Res. 2014;93: 539–546.

208. Villar CC, Dongari-Bagtzoglou A. Fungal diseases: Oral dysbiosis in susceptible hosts. Periodontol 2000. 2021;87: 166–180.

209. Mohammed H, Varoni EM, Cochis A, Cordaro M, Gallenzi P, Patini R, et al. Oral Dysbiosis in Pancreatic Cancer and Liver Cirrhosis: A Review of the Literature. Biomedicines. 2018;6. doi:10.3390/biomedicines6040115

210. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biol. 2015;16: 67.

211. van der Horst J, Buijs MJ, Laine ML, Wismeijer D, Loos BG, Crielaard W, et al. Sterile paper points as a bacterial DNA-contamination source in microbiome profiles of clinical samples. J Dent. 2013;41: 1297–1301.

212. Tett A, Pasolli E, Farina S, Truong DT, Asnicar F, Zolfo M, et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. NPJ Biofilms Microbiomes. 2017;3: 14.

213. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25.

214. Blanco-Miguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. bioRxiv. 2022. p. 2022.08.22.504593. doi:10.1101/2022.08.22.504593

215. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15: 962–968.

216. Hasan N, Yang H. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ. 2019;7: e7502.

217. Dalton A, Mermier C, Zuhl M. Exercise influence on the microbiome-gut-brain axis. Gut Microbes. 2019;10: 555–568.

218. Posalski I, Morgan MA, Riley ME, Goldstein EJC. Bulleidia extructa Prosthetic Hip Infection After a Dental Procedure: Potential Need for Prophylaxis. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2021. pp. e849–e850.

219. Morgan MA, Goldstein EJ. Bulleidia extructa: An underappreciated anaerobic pathogen. Anaerobe. 2021;69: 102339.

220. Macuch PJ, Tanner AC. Campylobacter species in health, gingivitis, and periodontitis. J Dent Res. 2000;79: 785–792.

221. Holdeman LV, Moore WE, Cato EP, Burmeister JA, Palcanis KG, Ranney RR. Distribution of capnocytophaga in periodontal microfloras. J Periodontal Res. 1985;20: 475–483.

222. Han YW. Fusobacterium nucleatum: a commensal-turned pathogen. Curr Opin Microbiol. 2015;23: 141–147.

223. Eribe ERK, Olsen I. Leptotrichia species in human infections. Anaerobe. 2008;14: 131–137.

224. Eribe ERK, Olsen I. Leptotrichia species in human infections II. J Oral Microbiol. 2017;9: 1368848.

225. Zhou J, Jiang N, Wang Z, Li L, Zhang J, Ma R, et al. Influences of pH and Iron Concentration on the Salivary Microbiome in Individual Humans with and without Caries. Appl Environ Microbiol. 2017;83. doi:10.1128/AEM.02412-16

226. Yamashita Y, Takeshita T. The oral microbiome and human health. J Oral Sci. 2017;59: 201–206.

227. Kolenbrander PE, Andersen RN, Moore LV. Coaggregation of Fusobacterium nucleatum, Selenomonas flueggei, Selenomonas infelix, Selenomonas noxia, and Selenomonas sputigena with strains from 11 genera of oral bacteria. Infect Immun. 1989;57: 3194–3203.

228. Siqueira JF Jr, Rôças IN. PCR-based identification of Treponema maltophilum, T amylovorum, T medium, and T lecithinolyticum in primary root canal infections. Arch Oral Biol. 2003;48: 495–502.

229. Rosier BT, Palazón C, García-Esteban S, Artacho A, Galiana A, Mira A. A Single Dose of Nitrate Increases Resilience Against Acidification Derived From Sugar Fermentation by the Oral Microbiome. Front Cell Infect Microbiol. 2021;11: 692883.

230. Li H, Thompson I, Carter P, Whiteley A, Bailey M, Leifert C, et al. Salivary nitrate--an ecological factor in reducing oral acidity. Oral Microbiol Immunol. 2007;22: 67–71.

231. Jockel-Schneider Y, Schlagenhauf U, Stölzel P, Goßner S, Carle R, Ehmke B, et al. Nitrate-rich diet alters the composition of the oral microbiota in periodontal recall patients. J Periodontol. 2021;92: 1536–1545.

232. Khocht A, Orlich M, Paster B, Bellinger D, Lenoir L, Irani C, et al. Cross-sectional comparisons of subgingival microbiome and gingival fluid inflammatory cytokines in periodontally healthy vegetarians versus non-vegetarians. J Periodontal Res. 2021;56: 1079–1090.

233. Woelber JP, Gärtner M, Breuninger L, Anderson A, König D, Hellwig E, et al. The influence of an anti-inflammatory diet on gingivitis. A randomized controlled trial. J Clin Periodontol. 2019;46: 481–490.

234. Hampelska K, Jaworska MM, Babalska ZŁ, Karpiński TM. The Role of Oral Microbiota in Intra-Oral Halitosis. J Clin Med Res. 2020;9. doi:10.3390/jcm9082484

235. Zhu B, Macleod LC, Kitten T, Xu P. Streptococcus sanguinis biofilm formation & interaction with oral pathogens. Future Microbiol. 2018;13: 915–932.

236. Buckel W, Martins BM, Messerschmidt A, Golding BT. Radical-mediated dehydration reactions in anaerobic bacteria. Biol Chem. 2005;386: 951–959.

237. Buckel W. Energy Conservation in Fermentations of Anaerobic Bacteria. Front Microbiol. 2021;12: 703525.

238. Kwack KH, Jang E-Y, Yang SB, Lee J-H, Moon J-H. Genomic and phenotypic comparison of Prevotella intermedia strains possessing different virulence in vivo. Virulence. 2022;13: 1133–1145.

239. Slakeski N, Dashper SG, Cook P, Poon C, Moore C, Reynolds EC. A Porphyromonas gingivalis genetic locus encoding a heme transport system. Oral Microbiol Immunol. 2000;15: 388–392.

240. Dashper SG, Cross KJ, Slakeski N, Lissel P, Aulakh P, Moore C, et al. Hemoglobin hydrolysis and heme acquisition by Porphyromonas gingivalis. Oral Microbiol Immunol. 2004;19: 50–56.

241. Millen AE, Dahhan R, Freudenheim JL, Hovey KM, Li L, McSkimming DI, et al. Dietary carbohydrate intake is associated with the subgingival plaque oral microbiome abundance and diversity in a cohort of postmenopausal women. Sci Rep. 2022;12: 2643.

242. Bacci G, Mengoni A, Emiliani G, Chiellini C, Cipriani EG, Bianconi G, et al. Defining the resilience of the human salivary microbiota by a 520-day longitudinal study in a confined environment: the Mars500 mission. Microbiome. 2021;9: 152.

243. Hujoel P. Dietary carbohydrates and dental-systemic diseases. J Dent Res. 2009;88: 490–502.

244. Food and Agriculture Organization of the United Nations, World Health Organization. Sustainable healthy diets: Guiding principles. Food & Agriculture Org.; 2019.

245. Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst. 1981;66: 1191–1308.

246. Baik I, Cho NH, Kim SH, Shin C. Dietary information improves cardiovascular disease risk prediction models. Eur J Clin Nutr. 2013;67: 25–30.

247. Sunitha S, Prashanth GM, Shanmukhappa, Chandu GN, Subba Reddy VV. An analysis of concentration of sucrose, endogenous pH, and alteration in the plaque pH on consumption of commonly used liquid pediatric medicines. J Indian Soc Pedod Prev Dent. 2009;27: 44–48.

248. Bedale W, Sindelar JJ, Milkowski AL. Dietary nitrate and nitrite: Benefits, risks, and evolving perceptions. Meat Sci. 2016;120: 85–92.

249. Hohensinn B, Haselgrübler R, Müller U, Stadlbauer V, Lanzerstorfer P, Lirk G, et al. Sustaining elevated levels of nitrite in the oral cavity through consumption of nitrate-rich beetroot juice in young healthy adults reduces salivary pH. Nitric Oxide. 2016;60: 10–15.

250. Archer DL. Evidence that ingested nitrate and nitrite are beneficial to health. J Food Prot. 2002;65: 872–875.

251. Sánchez GA, Miozza VA, Delgado A, Busch L. Total salivary nitrates and nitrites in oral health and periodontal disease. Nitric Oxide. 2014;36: 31–35.

252. Lingström P, Moynihan P. Nutrition, saliva, and oral health. Nutrition. 2003;19: 567–569.

253. Gondivkar SM, Gadbail AR, Gondivkar RS, Sarode SC, Sarode GS, Patil S, et al. Nutrition and oral health. Dis Mon. 2019;65: 147–154.

254. Tamashiro R, Strange L, Schnackenberg K, Santos J, Gadalla H, Zhao L, et al. Stability of healthy subgingival microbiome across space and time. Sci Rep. 2021;11: 23987.

255. Lin H, Eggesbø M, Peddada SD. Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. Nat Commun. 2022;13: 4946.

8 List of related pubblications

Giliberti R, Cavaliere S, Mauriello IE, Ercolini D, Pasolli E. Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa. PLoS Comput Biol. 2022;18: e1010066.

Overall evaluation of the PhD thesis

PhD course in ....Food Science.....

Cycle .....

REVIEWER (*first and last name*) .....GEORG ZELLER.....

AFFILIATION: .....EMBL HEIDELBERG, GERMANY.....

PhD candidate (*first and last name***)** ....RENATO GILIBERTI.....

Overall Evaluation of PhD thesis.....The thesis addresses novel questions focusing on the oral microbiome. It uses state-of-the-art methods with an excellent combination of methodological and application-centric sub-projects. The thesis is well-structured, approaches are clearly motivated and described and findings are well embedded in the literature. Results show in-depth investigations presented in detail. Discussion contains well supported conclusions and is balanced also acknowledging limitations. Overall the thesis clearly shows the scientific maturity of the candidate. I therefore recommend Renato Giliberti's admission to the defence.

Therefore:

☒ The PhD candidate is admitted to the public defence of the thesis in the ordinary final exam session;

☐ The PhD candidate is admitted to the public defence of the thesis **with postponement to the extraordinary session of the final exam***;

Date 29. Oct. 2022

Signature

G. Zeller

* **the extraordinary session of the final exam** wil be set by October 31st 2022 (or with an extension, according to the **D.L.n. 34 of 19/05/2020, art. 236, comma 5)**

Overall evaluation of the PhD thesis
PhD course in Food Science
Cycle 34

REVIEWER (*first and last name*) Prof. Levi Waldron

AFFILIATION: Department of Epidemiology and Biostatistics, Graduate School of Public Health and Health Policy, City University of New York

PhD candidate (*first and last name***)** Renato Giliberti

Overall Evaluation of PhD thesis *This is a thesis of outstanding quality and high relevance both methodologically and in advancing the understanding of causes of peri-implantitis. I learned several important things from the methodological investigation of machine learning that will impact my own work; in particular: that sensitive thresholding of taxonomic relative abundance at species or genus level has little or no impact on prediction model accuracy, that such thresholding in fact improves the performance of regression-based methods like Lasso, Elastic Net, and SVM, that thresholding begins to degrade performance at higher taxonomic levels or less-sensitive thresholds, but that such degradation is lessened in Leave-one-dataset-out (LODO) performance compared to cross-validation (CV), that the non regression-based Random Forest continues to be the most universal high-performance method for prediction from microbiome data, and that also differential abundance analysis is hardly affected by thresholding. These insights are of high novelty and help to understand what are the issues most relevant to prediction modeling in microbiome data and will impact the development of future statistical methodologies. The per-implantitis study involved collection of a large dataset of questionnaires, oral health, and oral microbiome data from dental clinics across Italy, an extremely ambitious PhD project. Analysis of this study follows good statistical principles and is a thorough investigation of the dataset, identifying clinical and microbiome correlates of peri-implantitis, including the insight that diet is not strongly driven by diet, but is extremely accurately predicted by clinical variables. Some suggestions for this chapter are 1) it is missing a Discussion section, 2) (recommendation only) use color blind-friendly palettes in the PCoA plots, as 10% of men cannot read a red-green color scale, and 3) (recommendation only) use the word "sex" instead of "gender" if referring to a biological variable (see e.g. http://orwh.od.nih.gov/sex-gender) –this is only recommended because this distinction is not likely of great*

importance to this study. The final chapter investigates correlations between the dietary questionnaire and submucosal and supragingeval plaque microbiomes and microbial functional potential. The analysis is basic but reasonable. Shortcomings in the dietary dataset are discussed, as the pandemic necessitated at-home questionnaires instead of supervised. I note that dimension reduction of the dietary dataset, e.g. using PCoA, might have helped address limitations in its accuracy by allowing focus on broad dietary patterns that may be more accurate/robust than individual responses. This can be considered for publication but is not necessary for the thesis.

Therefore:

☒ The PhD candidate is admitted to the public defence of the thesis in the ordinary final exam session;

☐ The PhD candidate is admitted to the public defence of the thesis **with postponement to the extraordinary session of the final exam**.

Date October, 31st 2022

Signature

2