







# A Complementary Approach in the Analysis of the Human Gut Microbiome Applying Self-organizing Maps and Random Forest

Valeria Burgos<sup>1</sup>(✉) , Tamara Piñero<sup>1</sup> , María Laura Fernández<sup>2</sup> ,  
and Marcelo Risk<sup>1</sup> 

<sup>1</sup> Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) - CONICET, Hospital Italiano de Buenos Aires, Instituto Universitario del Hospital Italiano, Buenos Aires, Argentina

[valeria.burgos@hospitalitaliano.org.ar](mailto:valeria.burgos@hospitalitaliano.org.ar)

<sup>2</sup> CONICET - Universidad de Buenos Aires, Instituto de Física del Plasma (INFIP), Buenos Aires, Argentina

**Abstract.** The human gastrointestinal tract is colonized by millions of microorganisms that make up the so-called gut microbiota, with a vital role in the well-being, health maintenance as well as the appearance of several diseases in the human host. A data mining analysis approach was applied on a set of gut microbiota data from healthy individuals. We used two machine learning methods to identify biomedically relevant relationships between demographic and biomedical variables of the subjects and patterns of abundance of bacteria. The study was carried out focusing on the two most abundant human gut microbiota groups, *Bacteroidetes* and *Firmicutes*. Both subsets of bacterial abundances together with the metadata variables were subjected to an exploratory analysis, using self-organizing maps that integrate multivariate information through different component planes. Finally, to evaluate the relevance of the variables on the biological diversity of the microbial communities, an ensemble-based method such as random forest was used. Results showed that age and body mass index were among the most important features at explaining bacteria diversity. Interestingly, several bacteria species known to be associated to diet and obesity were identified as relevant features as well. In the topological analysis of self-organizing maps, we identified certain groups of nodes with similarities in subject metadata and gut bacteria. We conclude that our results represent a preliminary approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals personalize patient treatment or support decision making.

**Keywords:** Gut microbiome · Self-organizing maps · Random forest · Precision medicine · Bioinformatics

# 1 Introduction

Data science comprises different scientific fields of knowledge to target the analysis of complex and massive data. In particular, the increased interest on the application of machine learning algorithms to extract hidden associations or patterns in electronic health records, processing of medical images, prediction of a health situation or classification of patients has demonstrated the need for machine learning tools for reliable decision-making in healthcare and handling of biological data. The human gastrointestinal tract harbors millions of microorganisms which includes bacteria, archaea, fungi and viruses, interacting in symbiotic relationships between the host and each microbial community. This is known as the gut microbiota, while the collective genome of all symbiotic and pathogenic microorganisms represents the gut microbiome. The establishment of a large part of the component communities that will remain in the adult life occurs at birth and during the first years of life [1,2] and its composition is shaped not only by the host genetics but also by environmental factors, nutritional status, age and lifestyle. Importantly, the gut microbiome plays an essential role in a number of health-beneficial functions (digestion, synthesis of essential vitamins and amino acids, absorption of calcium, magnesium and iron, fermentation of indigestible components, protection against pathogens, etc.) [3].

The rates of growth and survival of its component populations may fluctuate in response to temporary stressors, such as changes in diet or the consumption of antibiotics [4]. This potential for dynamic restructuring involves two important characteristics of the gut microbiome: plasticity and resilience [5,6]. Ongoing research in human and animal models highlights the importance of a healthy gut microbiome since persistent imbalances in composition and stability, known as dysbiosis, are associated to the onset and progression of chronic diseases that include obesity, irritable bowel syndrome, diabetes, cancer, and neurological diseases such as Parkinson's, among others [7].

The generation of biological knowledge from the large flow of data generated by new technologies in biomedical sciences has accelerated their transformation into data-centered fields. Thus, the study of the human gut microbiome represents a major challenge since it requires an interdisciplinary work between computer science and medicine. The interaction between these two fields will help obtain knowledge about gut bacteria interactions in human health and disease.

In the present work, we analyzed microbiome abundance data and the associated metadata using a machine learning approach: we used the visualization capabilities offered by self-organizing maps to identify patterns of multivariate data stored in multiple layers and additionally, we applied random forest to model the prediction of microbial diversity. For each analysis, we focused on the abundance levels of two major groups of the human gut bacteria, such as *Bacteroidetes* and *Firmicutes*.

## 2 Methods

### 2.1 Dataset

Microarray profiling data of human gut microbiota and anonymized metadata were obtained from the Dryad Digital Repository, as described by [8]. Briefly, the data matrix contained 1172 intestinal samples of western adults. In each sample, bacterial abundances were quantified using the HITChip phylogenetic microarray. This technology allows the assessment of relative abundances of gut bacteria through signal intensities of the targeted 16S rRNA gene, frequently used for the identification of poorly described or non cultured bacteria. Data contained hybridization signals for 130 genus-like phylogenetic groups. Subject metadata included age, sex, nationality, probe-level Shannon diversity, BMI group and subjectID. Geographical origin of the study subjects were: Central Europe (Belgium, Denmark, Germany, the Netherlands), Eastern Europe (Poland), Scandinavia (Finland, Norway, Sweden), Southern Europe (France, Italy, Serbia, Spain), United Kingdom/Ireland (UK, Ireland) and the United States (US). We used VIM and tidyverse R packages [9,10] to check for the presence of missing values (NAs). Records containing NAs were carefully removed without causing bias in the dataset. During the cleaning process, the category ‘Eastern Europe’ was turned out since it was represented by only one complete case. The final dataset to be used was represented by 1056 complete patient records containing 130 bacterial abundance data and subject metadata.

### 2.2 Self-organizing Maps (SOMs)

Self-organizing maps (also known as Kohonen maps) represent an optimal option to organize multidimensional data in a two-dimensional space by using a neural network. SOM uses the vector space as a model to represent data in a two-dimensional lattice: each value through N samples could be referred to as a data point in an N-dimensional space. Thousands of data points would therefore form data clouds in space, with an intrinsic topology due to geometric relationships. From this it follows that the greater the similarity in the data value level, the closer is the geometric space they occupy. To visualize the trained SOM, we used heatmaps for each variable to plot the degree of connectivity between adjacent output neurons through the use of a color intensity panel. In the case of multivariate datasets, the visualization of different heatmaps allows an overall analysis of the relations between the variables since maps are linked to each other by position: in each map, a node in a given location corresponds to the same unit in another map. The SOM map can be implemented in different topologies. Data were divided into two subsets by major groups of gut bacteria (*Bacteroidetes* and *Firmicutes*). We used a regular hexagonal 2D grid consisting of 750 neurons, in  $30 \times 25$  grids. Data was logarithmically-scaled before training. We used the kohonen R package, which provides a standardized framework for SOMs.

## 2.3 Random Forest

Random forest (RF) is an ensemble learning method that can solve regression and classification problems. The algorithm uses a random subset of the training samples for each tree and a random subset of predictors in each step during the training process. These two sources of randomization make the algorithm robust to correlated predictors and more reliable at obtaining average outputs into a model. Data were divided into two subsets by major groups of gut bacteria (*Bacteroidetes* and *Firmicutes*). Bacterial abundance data and metadata were used as RF regressors to generate a diversity prediction model, which was performed using the RF regression algorithm provided by the R interface for h2o [11]. We used 10-fold cross validation for training the regression models and their performance were evaluated using Mean Absolute Error (MAE) as the error metric. After parameter tuning (mainly focused on the number of trees) through cross validation, the best RF regression model was selected.

## 3 Results

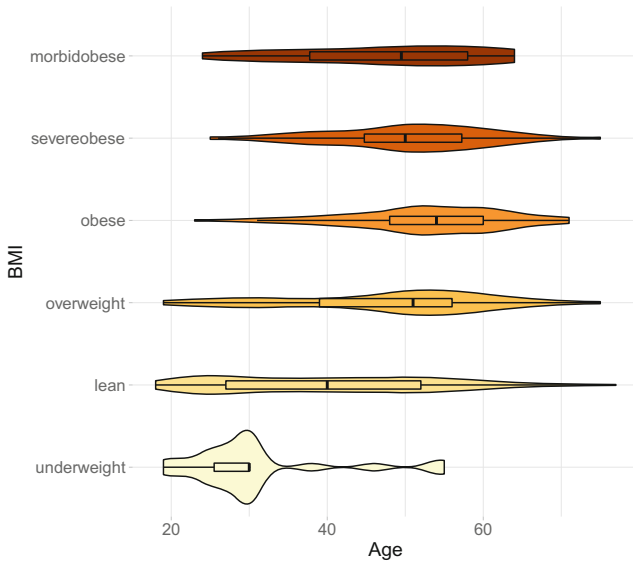
### 3.1 Characteristics of the Study Population

The degree of obesity is a relevant aspect in gut microbiome studies in terms of its influence on the microbiota composition [12, 13]. This parameter, that can be obtained through the body mass index (BMI), indicates the nutritional status of an individual. Descriptive analysis of the study population showed that lean individuals were homogeneously distributed in all age groups, while overweight, obese and severe obese categories were more abundant in 45–60 year-old individuals. A large proportion of the underweight population was represented between 20–30 years old (Fig. 1).

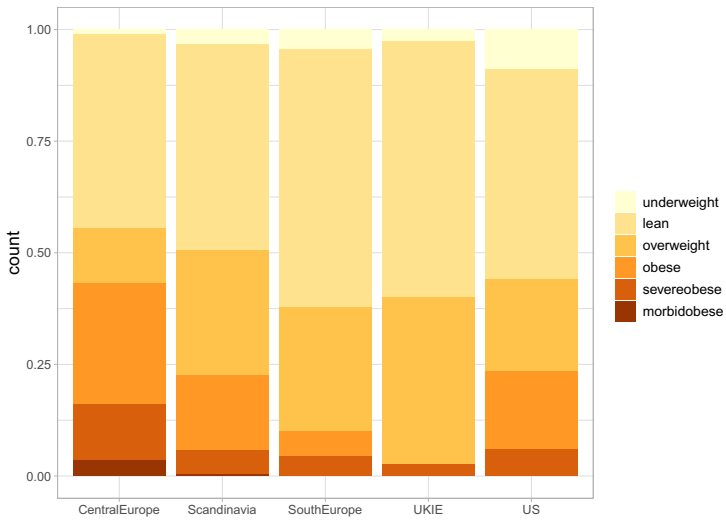
The distribution of the different BMI categories in each geographic region showed that lean individuals represented approximately half of the proportions for all locations. For Scandinavia, Southern Europe, UK/Ireland (UKIE) and the US, the following proportion was represented by overweight subjects. In contrast, in Central Europe, the second proportion after lean individuals was represented by obese individuals. Morbid obese subjects were present only in Central Europe (Fig. 2).

### 3.2 SOM Analysis

Each component plane or map in a SOM represents one type of data: a two-dimensional lattice for each metadata variable (BMI, nationality, age, sex and diversity) as well as for each bacteria (whose relative abundance is represented in expression levels of the 16S rRNA gene). Since each map preserves shape and density, exploration of the geometric relationships between nodes allows a direct identification of similarities and differences between the layers.



**Fig. 1.** Distribution of the BMI categories across age intervals in the study population.



**Fig. 2.** Proportion of each BMI category of the study subjects across different geographic locations.

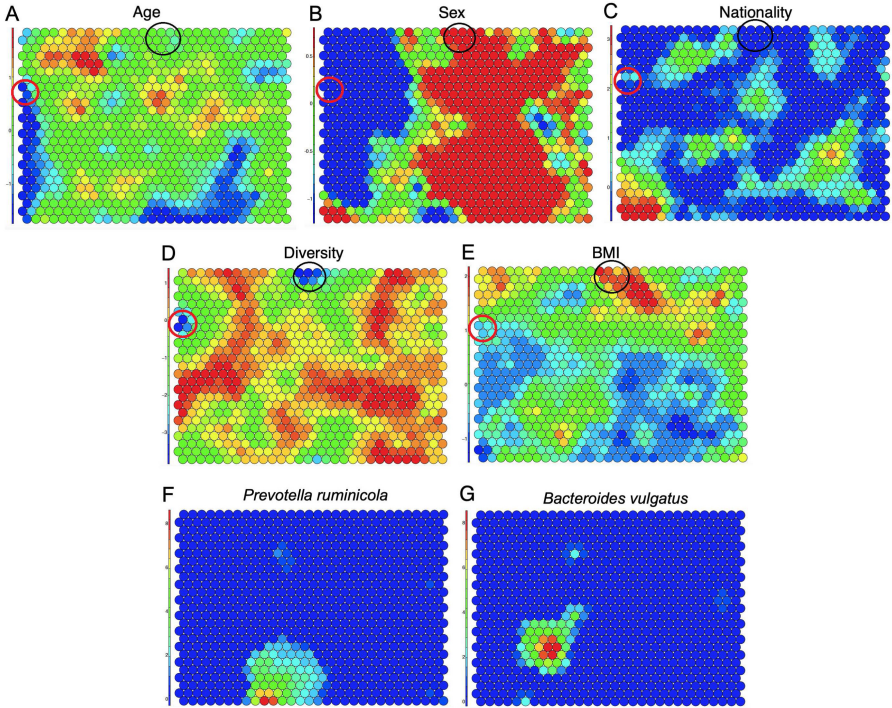
***Bacteroidetes***. After running the SOM algorithm using the *Bacteroidetes* subset, the different regions in each map indicated that the age distributed into two well-defined subregions of nodes for younger ages (around 20 years old), quite separated from a small group of nodes representing older individuals (older than 65 years old). Nodes representing 40–50 year old subjects were scattered throughout the map (Fig. 3A). Men and women were clearly distributed in two large, separated areas in the map (Fig. 3B). Scandinavian individuals mapped in a few groups of isolated nodes while Central European subjects were homogeneously distributed. Interestingly, the map identified an isolated group of nodes corresponding to individuals from the United States (Fig. 3C). Additionally, underweight and lean subjects mapped in two wide groups, while severe and morbid obese individuals mapped mainly in a small and defined group of nodes (Fig. 3E). The distribution of microbial diversity showed that two small and defined groups of nodes mapped low diversity values while higher values distributed into larger and clearly defined subsets of nodes across the map (Fig. 3D).

After SOM training, the abundance levels of each bacteria species of the *Bacteroidetes* phylum was also represented in a map. In the present dataset, several members belonging to this phylum were identified but no overlay between any bacterial map was observed (data not shown). However, since many members of the *Bacteroidetes* community have a relevant role in the host health, we chose to analyze two prominent bacteria whose relative abundances are known to be influenced by the host lifestyle and diet: a high fat and protein intake is associated with elevated microbial presence of *Bacteroides* species, while a high fiber intake is associated with high microbial levels of *Prevotella* species [14, 15]. It appears that the abundances of these two *Bacteroidetes* members showed no overlay between any host metadata map because there are no coincidences in location (Fig. 3F and G).

The superimposition of the multiple maps described above allows to obtain some clear aspects of the data from the perspective of the *Bacteroidetes* subset:

- Low diversity values overlaps with a lower BMI (lean subjects) corresponding to young men from Central Europe and Scandinavia (indicated by a red circle in Fig. 3A–E).
- Interestingly, another subset of low diversity values overlaps with high BMI values (that is, severe to morbid obese) corresponding to middle-aged female individuals from Central Europe (indicated by a black circle in Fig. 3A–E).
- The small proportion of young subjects is equally represented in both men and women, while older individuals correspond exclusively to men.
- There are no women older than 65 years.
- US nationality corresponds to lean women in the range of medium values of microbial diversity.

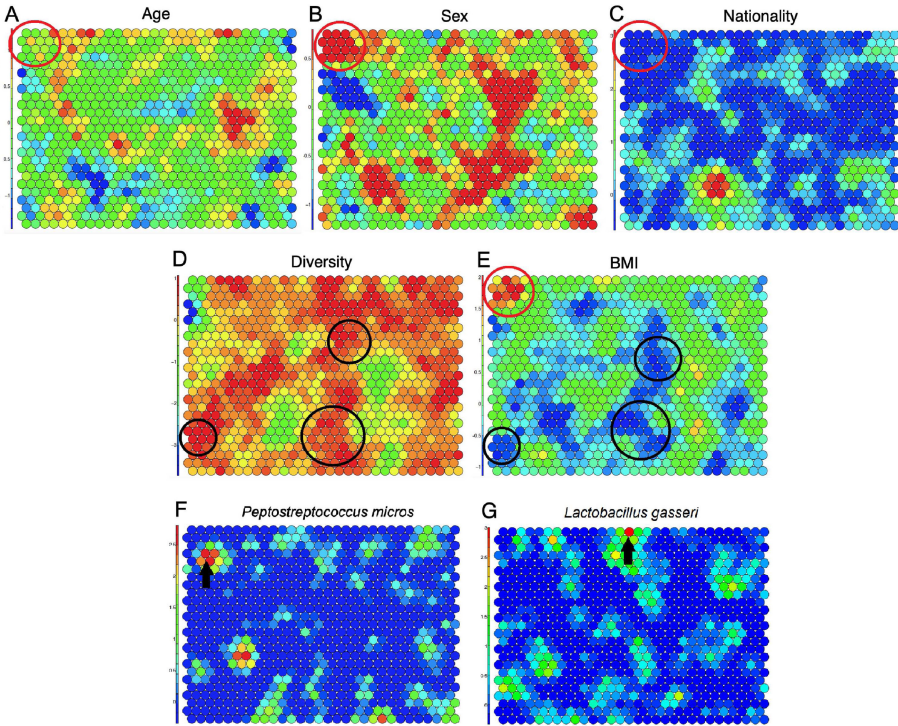
***Firmicutes***. When subject metadata was trained using the *Firmicutes* subset, the different layers showed a more disperse variable distribution in the maps: middle-age individuals were homogeneously represented throughout the lattice while younger (around 20 years old) and older ages (older than 65 years) mapped



**Fig. 3.** SOM results for metadata variables associated with *Bacteroidetes*. The color index of each map is established based on the values for each variable. A) Age, scale adjusted for a range from 18 to 77 years old, blue = younger, red = older adults. B) Sex, blue = male, red = female. C) Nationality, color gradient beginning at Central Europe (level 0, color blue), Scandinavia (level 1, light blue), Southern Europe (level 2, green), UK/Ireland (level 3, orange) and the US (level 4, red). D) Diversity, scale adjusted for a range of 4.7 to 6.3 diversity index values beginning at blue (lowest diversity) to red (highest diversity). E) BMI, color gradient beginning at underweight (color blue), lean (light blue), overweight (green), obese (yellow), severe obese (orange) and morbid obese (red). In F) *P. ruminicola* and G) *B. vulgatus* the color gradient represents the level of abundance, blue = low, red = high. (Color figure online)

in small, discrete groups of nodes (Fig. 4A). Men and women were not as clearly separated in their node distribution as in the *Bacteroidetes* subset (Fig. 4B). Regarding nationality, a node pattern similar to *Bacteroidetes* was observed (Fig. 4C). High microbial diversity values predominated throughout the map while only three nodes mapped for low diversity values (Fig. 4D). Additionally, lean and underweight subjects predominated in most of the nodes and only a very small group of nodes grouped the highest BMI values, corresponding to the severe obese category (Fig. 4E).

The phylum *Firmicutes* is made up of around 250 different genera of bacteria, such as *Lactobacillus*, *Bacillus*, *Clostridium*, *Enterococcus*, and *Ruminococcus*,



**Fig. 4.** SOM results for metadata variables associated with *Firmicutes*. The color index of each map is established based on the values for each variable. A) Age, scale adjusted for a range from 18 to 77 years old, blue = younger, red = older adults. B) Sex, blue = male, red = female. C) Nationality, color gradient beginning at Central Europe (level 0, color blue), Scandinavia (level 1, light blue), Southern Europe (level 2, green), UK/Ireland (level 3, orange) and the US (level 4, red). D) Diversity, scale adjusted for a range of 4.7 to 6.3 diversity index values beginning at blue (lowest diversity) to red (highest diversity). E) BMI, color gradient beginning at underweight (color blue), lean (light blue), overweight (green), obese (yellow), severe obese (orange) and morbid obese (red). In F) *P. micros* and G) *L. gasseri* the color gradient represents the level of abundance, blue = low, red = high. (Color figure online)

among other important members. In the present dataset, 74 members belonging to this phylum were identified and consequently, each generated a heatmap after SOM training (data not shown). However, superimposing multiple maps revealed that only one node corresponding to overweight individuals slightly coincided with higher values of abundance of a single species, *Peptostreptococcus micros* (indicated by an arrow in Fig. 4F). This represents an interesting result since several other members of *Firmicutes* have previously been associated to obesity [16,17]. Additionally, an overlay between high microbial diversity and higher values of abundance for *Lactobacillus gasseri* was observed (indicated by an arrow in Fig. 4G).



The map overlay between diversity and BMI categories allowed to observe that:

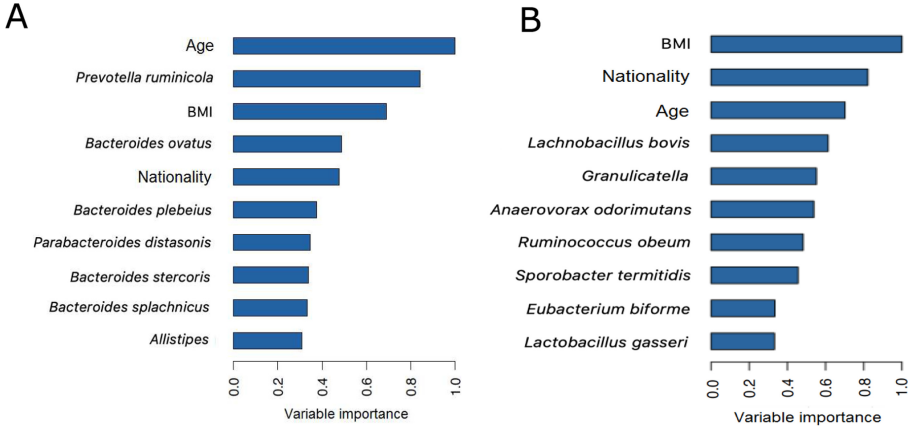
- Severe to morbid obese individuals were middle-age women from Central Europe (indicated by a red circle in Fig. 4A, B, C and E).
- Many of the maximum diversity values superimpose with the lowest BMI categories (indicated by a black circle in Fig. 4D and E).

### 3.3 Random Forest

Diversity is an important variable in microbiome research because it describes the richness (the number of classes) and distribution of the component microorganisms among classes. Understanding diversity in the intestinal microbial community also allows us to understand the impact of factors on bacteria distribution, such as the use of antibiotics, type of diet, degree of obesity, medical interventions and environmental factors, among others (reviewed in [18]). We developed a RF regression model of microbiome diversity. A very useful visual way to interpret RF results of the prediction model is through a ranking list of feature importance, which refers to the relative influence of each feature on the target variable. It considers whether a variable was selected to split and how much the squared error (over all trees) improved as a result.

***Bacteroidetes***. We observed that the age of the individuals was the most important factor in the regression model to predict the diversity of the gut microbiota. Regarding the bacterial abundances, it was observed that two species had the greatest relative importance in the identification of diversity: *Prevotella ruminicola* and *Bacteroides ovatus*. On the other hand, of the remaining metadata variables, the BMI of individuals had a remarkable position in the ranking of importance, followed by the geographic location. Gender was not important in the prediction of diversity. While the list of importance for each variable is informative, a better interpretation is obtained after scaling between 0 and 1 in a descending order of importance (Fig. 5A).

***Firmicutes***. When data was RF-trained considering this subset, the first three positions of the list were occupied by the BMI of the individuals, followed by Nationality and Age. Regarding the abundances of the 74 types of bacteria, it was observed that two members had the greatest relative importance in predicting diversity: *Lachnobacillus bovis* and *Granulicatella* (Fig. 5B).



**Fig. 5.** Scale of the relative importance of the first ten variables on the diversity of *Bacteroidetes* (A) and *Firmicutes* (B) after implementing random forest.

## 4 Discussion

We sought to characterize the relations between subject metadata and specific bacterial members of the gut microbiome in a thousand western adults through the use of two machine learning methods that provide robust analytical visualizations, such as self-organizing maps and random forest. Bacteria of the human intestinal microbiome are taxonomically classified into six large groups or phyla, which in turn are subclassified into classes, orders, families, genera and species. The present work addressed the analysis on two of the most abundant phyla, *Bacteroidetes* and *Firmicutes*, which represent 90% of the gut microbiota [19].

The configuration of the SOM outcome maintains the topological structure of the input multidimensional data, in which similar values are mapped in the same or near node in a two dimensional map. Such topological preservation is of particular significance in the exploratory phase of omics data mining since there is generally no *a priori* knowledge of data structure. We presented the visualization of different superimposed heatmaps that allowed the exploration of relationships between input variables. This way of presenting SOM outcomes is similar to previous studies [20].

Our results showed that only one species of *Firmicutes*, *Peptostreptococcus micros*, was slightly associated to nodes that grouped overweight individuals, mostly middle-aged women. Notably, several previous studies have shown that *Peptostreptococcus micros* (later classified as *Parvimonas micra*) is one of various colorectal cancer (CRC) microbial markers [13,21]. This bacteria has also been reported to have a pathogenic role in periodontal diseases [22]. Considering the behavior of the oral microbiome during periodontal infection and its influence in health complications, such as diabetes, cardiovascular disease, and obesity [23], the significance of the results obtained here provides a basis for future studies on the possible role of gut bacteria as biological markers of a

developing overweight condition. Among bacteria with known beneficial roles, we observed that high abundance of *Lactobacillus gasseri* was related to nodes that grouped high diversity values. This result supports the previously reported role of *Lactobacillus gasseri* in the management of obesity and probiotic properties [24,25]. The topological structures of the metadata variables were slightly different depending whether *Bacteroidetes* or *Firmicutes* subsets were used for SOM training. For both age and sex, mapping distribution of the study subjects was more effective using the *Bacteroidetes* data. In the case of the different BMI categories, subject distribution was more effective using *Firmicutes* data.

Although the SOM results presented here allowed us to gain insight into the different regions of matching information underlying host metadata and the relative abundance of bacteria, we consider that a deeper approach of the use of SOM is needed, in terms of parameter configuration, such as size, dimensionality, shape, learning rate, among others. Considering that the relationships between gut microbiome and host BMI are dynamic and complex, self-organizing maps provide an excellent tool of visualization and dimensionality reduction that could serve as a complementary tool in a biomedical report.

Analysis of the microbiome diversity in the human body is essential to understand the structure, biology and ecology of its component communities. This analysis represents a critical first step in microbiome studies. When supervised learning through a regression random forest algorithm was used to determine which variables were important in the prediction of microbial diversity, we observed that both age and BMI category of the individuals appeared as the most relevant in the regression models generated for *Bacteroidetes* and *Firmicutes* subsets. The contribution of these physiological factors in shaping the gut microbiome has been reported previously: with age, the beneficial functions provided by a healthy gut microbiome begin to decrease in association to an increasing frequency of inflammatory processes and disease, especially in the elderly. Regarding the influence of BMI, various studies that compare the intestinal microbiota between obese and lean individuals indicate that the variation in the degree of diversity is associated with body weight (obese individuals present a low diversity, which means a higher BMI) [26–29].

Random forest regression also indicated that two *Bacteroidetes* species were the most relevant on diversity: *Prevotella ruminicola* is involved in the response of individuals to dietary supplements [30] and *Bacteroides ovatus* is a dominant species in the human intestine, previously identified as a next generation probiotic due to its preventive effects on intestinal inflammation (reviewed in [31]). On the other hand, two members of the *Firmicutes* group were identified as important at predicting diversity in this subset: *Lachnobacillus bovis* and *Granulicatella* whose relative abundances are reported to be influenced by the type of diet and the use of antibiotics in obese individuals [32,33]. Hence, some of the results obtained in both RF regression models are consistent with published microbiome research, indicating the robustness of the RF regression algorithm. A further analysis is needed that involves a complete parameter tuning so as to characterize the most accurate RF setting for a microbiome project.

In the last decade, the impulse provided by innovative developments in technology and the generation of large volumes of microbiome data has caused an increase in the use of machine learning methods in this field, such as microbial ecology, identification of certain bacteria to cancer and forensics, among others [34–36]. We consider that our study represents the start of a contribution to the vast field of microbiome research, although we need further refinements of the methodology used in order to validate the obtained models and improve performances.

The accumulating research of the gut microbiome and its influence on health and disease has accelerated the need for integration of multidisciplinary fields in its analysis. In general, health professionals (medical doctors, nurses, biochemists) are not prepared to work in all the steps along the data analysis process (cleaning, filtering, choice of algorithms, interpretation, etc.). Therefore, data science and machine learning can contribute to the translation of innovative results into valuable knowledge that provide decision support in microbiome-based precision medicine.

## 5 Conclusions

We used two robust computer science-based methods, such as self-organizing maps and random forest, to study the relationships between gut microbiome data and host information. Our results represent a preliminary approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals to individualize patient treatment or support decision making. Additionally, this work contributes to the increasingly growing area of gut microbiome interactions on human health and disease. However, further studies using other machine learning algorithms to validate the results obtained here are required.

## References

1. Dominguez-Bello, M.G., et al.: Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci.* **107**(26), 11971–11975 (2010)
2. Koenig, J.E., et al.: Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci.* **108**(Supplement 1), 4578–4585 (2011)
3. Rowland, I., et al.: Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.* **57**(1), 1–24 (2017). <https://doi.org/10.1007/s00394-017-1445-8>
4. Kiewiet, M., et al.: Flexibility of gut microbiota in ageing individuals during dietary fiber long-chain inulin intake. *Molecular Nutrition & Food Research*, p. 2000390 (2020)
5. Ruggles, K.V., et al.: Changes in the gut microbiota of urban subjects during an immersion in the traditional diet and lifestyle of a rainforest village. *Msphere*, vol. 3(4) (2018)
6. Liu, H., et al.: Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut* **68**(12), 2254–2255 (2019)

7. Floch, M.H., Ringel, Y., Walker, W.A.: The microbiota in gastrointestinal pathophysiology: implications for human health, prebiotics, probiotics, and dysbiosis. Academic Press (2016)
8. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., De Vos, W.M.: Tipping elements in the human intestinal ecosystem. *Nat. Commun.* **5**(1), 1–10 (2014)
9. Kowarik, A., Templ, M.: Imputation with the R package VIM. *J. Stat. Softw.* **74**(7), 1–16 (2016). <https://doi.org/10.18637/jss.v074.i07>
10. Wickham, H., et al.: Welcome to the tidyverse. *J. Open Source Softw.* **4**(43), 1686 (2019). <https://doi.org/10.21105/joss.01686>, <http://dx.doi.org/10.21105/joss.01686>
11. LeDell, E., et al.: h2o: R interface for 'h2o'. R package version 3(0.2) (2018)
12. Haro, C., et al.: Intestinal microbiota is influenced by gender and body mass index. *PLoS ONE* **11**(5), e0154090 (2016)
13. Yu, J., et al.: Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**(1), 70–78 (2017)
14. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., Lionetti, P.: Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci.* **107**(33), 14691–14696 (2010)
15. Wu, G.D., et al.: Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**(6052), 105–108 (2011)
16. Koliada, A., et al.: Association between body mass index and firmicutes/bacteroidetes ratio in an adult ukrainian population. *BMC Microbiol.* **17**(1), 120 (2017)
17. Crovesy, L., Masterson, D., Rosado, E.L.: Profile of the gut microbiota of adults with obesity: a systematic review. *Eur. J. Clin. Nutr.* **74**(9), 1251–1262 (2020)
18. Reese, A.T., Dunn, R.R.: Drivers of microbiome biodiversity: a review of general rules, feces, and ignorance. *MBio* **9**(4), e01294-18 (2018)
19. Arumugam, M., et al.: Enterotypes of the human gut microbiome. *Nature* **473**(7346), 174–180 (2011)
20. Qian, J., et al.: Introducing self-organized maps (som) as a visualization tool for materials research and education. *Results Mater.* **4**, 100020 (2019)
21. Xu, J., et al.: Alteration of the abundance of parvimonas micra in the gut along the adenoma-carcinoma sequence. *Oncol. Lett.* **20**(4), 1 (2020)
22. Nagarajan, M., Prabhu, V.R., Kamalakkannan, R.: Metagenomics: implications in oral health and disease. In: *Metagenomics*, pp. 179–195. Elsevier (2018)
23. Goodson, J., Groppo, D., Halem, S., Carpino, E.: Is obesity an oral bacterial disease? *J. Dent. Res.* **88**(6), 519–523 (2009)
24. Selle, K., Klaenhammer, T.R.: Genomic and phenotypic evidence for probiotic influences of lactobacillus gasseri on human health. *FEMS Microbiol. Rev.* **37**(6), 915–935 (2013)
25. Mahboubi, M.: Lactobacillus gasseri as a functional food and its role in obesity. *Int. J. Med. Rev.* **6**(2), 59–64 (2019)
26. Turnbaugh, P.J., et al.: A core gut microbiome in obese and lean twins. *Nature* **457**(7228), 480–484 (2009)
27. Yatsunencko, T., et al.: Human gut microbiome viewed across age and geography. *Nature* **486**(7402), 222–227 (2012)
28. Dominianni, C., et al.: Sex, body mass index, and dietary fiber intake influence the human gut microbiome. *PLoS ONE* **10**(4), e0124599 (2015)
29. Bosco, N., Noti, M.: The aging gut microbiome and its impact on host immunity. *Genes & Immunity*, pp. 1–15 (2021)

30. Chung, W.S.F., et al.: Relative abundance of the prevotella genus within the human gut microbiota of elderly volunteers determines the inter-individual responses to dietary supplementation with wheat bran arabinoxylan-oligosaccharides. *BMC Microbiol.* **20**(1), 1–14 (2020)
31. Tan, H., et al.: Pilot safety evaluation of a novel strain of bacteroides ovatus. *Front. Genet.* **9**, 539 (2018)
32. Salonen, A., et al.: Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *ISME J.* **8**(11), 2218–2230 (2014)
33. Reijnders, D., et al.: Effects of gut microbiota manipulation by antibiotics on host metabolism in obese humans: a randomized double-blind placebo-controlled trial. *Cell Metab.* **24**(1), 63–74 (2016)
34. Ai, D., Pan, H., Han, R., Li, X., Liu, G., Xia, L.C.: Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. *Genes* **10**(2), 112 (2019)
35. Thompson, J., Johansen, R., Dunbar, J., Munsky, B.: Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. *PLoS ONE* **14**(7), e0215502 (2019)
36. Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., IV., Wiens, J., Schloss, P.D.: A framework for effective application of machine learning to microbiome-based classification problems. *MBio* **11**(3), e00434-20 (2020)