# Gut microbiome co-abundance networks varies with age, sex, **smoking status and body mass index**<br>Christophe Boetto<sup>1</sup>, Violeta Basten Romero<sup>1</sup>, Léo Henches<sup>1</sup>, Arthur Frouin<sup>1</sup>, Antoine Auvergne<sup>1</sup>,

Christophe Boetto<sup>\*</sup>, Violeta Basten Romero<sup>\*</sup>, Léo Henches<sup>\*</sup>, Arthur Frouin<sup>\*</sup>, Antoine Auvergne<sup>\*</sup><br>Etienne Patin<sup>2</sup>, Marius Bredon<sup>3,4</sup>, Milieu Intérieur Consortium, Sean P. Kennedy<sup>1</sup>, Darragh Duffy<sup>5</sup>, Lluis<br>Quintana-,<br>,<br>,<br>,<br>, , Marius Bredon<sup>9,4</sup>, Milieu Intérieur Consortium, Sean P. Kennedy<sup>+</sup><br>ci<sup>2</sup>, Harry Sokol<sup>3,4,6,7</sup>, Hugues Aschard<sup>1,8</sup><br>Université de Paris, Department of Computational Biology, F-75015 Paris, France<br>Université de Paris, De , Darragh Duffy<sup>3</sup><br>;<br>Inflammation Labo<br>.

Institut Pasteur, Université de Paris, Départment of Computational Biology, F-75015 Paris, France, Institut Pasteur, Université de Paris, Department of Genomes and Genetics, Paris, France, I

Quintana-Murci<sup>2</sup>, Harry Sokol<sup>3,4,6,7</sup>, Hugues Aschard<sup>1,8</sup><br><sup>1</sup>Institut Pasteur, Université de Paris, Department of Computation<sup>2</sup><br>Institut Pasteur, Université de Paris, Department of Genomes and<br><sup>3</sup>Sorbonne Université, ,  $\overline{\phantom{a}}$  ,  $\overline{\phantom{a$ Quintana-Murci<sup>2</sup>, Harry Sokol<sup>3,4,6</sup>,7, Hugues Aschard<sup>4,6</sup><br><sup>1</sup>Institut Pasteur, Université de Paris, Department of Computation<sup>2</sup><br>Institut Pasteur, Université de Paris, Department of Genomes ana<sup>3</sup><br>Sorbonne Université, I nstitut Pasteur, Oniversité de Paris, Department of Genomes and Genetics, Paris, France, Inserne Université, Inser<br>Sorbonne Université, Inserm, Centre de recherche Saint-Antoine, CRSA, Microbiota, Gut and Inflammation Labo

Hôpital Saint-Antoine (UMR 3938), 79012, Paris, France,<br><sup>4</sup> Paris Center for Microbiome Medicine, PaCeMM, FHU, F

r ans center for Microbiome Medicine, PaceMM, FHU, Paris, France<br><sup>5</sup>Institut Pasteur, Université de Paris, Translational Immunology Unit,

nstitut Pasteur, Université de Paris, Translational Immunology Unit, Paris, France, Castroenterology Department, AP-HP, Saint Antoine Hospital, 75012, Paris, France, <sup>7</sup>

INRAE Micalis & AgroParisTech, UMR1319, Micalis & AgroParisTech, Jouy en Josas, France,<br><sup>8</sup>Uggest TU Chan School of Bublie Hoalth, Department of Enidemialagu, Besten, United Stat

8 Harvard TH Chan School of Public Health, Department of Epidemiology, Boston, United States of America

### Abstract

);<br>| 1<br>| c || nerent functional communities. Variability in these communities, typically investigated through<br>a co-abundance, might provide critical insights on the biological links between the gut microbiome<br>I human phenotypes. However taxa co-abundance, might provide critical insights on the biological links between the gut microbiome<br>and human phenotypes. However, existing methods to investigate variations in taxa co-abundance<br>suffer multiple limitatio The transmission provide critical insights on the strength and human phenotypes. However, existing methods to investigate variations in taxa co-abundance suffer multiple limitations. Here, we address the simple but challen suffer multiple limitations. Here, we address the simple but challenging question of identifying<br>factors associated with variability in gut microbiome taxa co-abundance using a novel covariance-<br>based method that resolve t such multiple limitations associated with variability in gut microbiome taxa co-abundance using a novel covariance-<br>based method that resolve these limitations. We screened 80 host factors in 938 healthy<br>participants, and participants, and identified associations between taxa co-abundance variability and age, sex, smoking<br>status, and body mass index (BMI) not captured by abundance-based and diversity-based methods.<br>Increased age and smoking participants, and body mass index (BMI) not captured by abundance-based and diversity-based methods.<br>Increased age and smoking were associated with an overall decrease in co-abundance, and<br>conversely BMI with an increase. status, and they made mass in provided with an overall decrease in co-abundance, and conversely BMI with an increase. Finally, we demonstrate that the proposed approach offers a powerful framework for describing taxa netwo Increased age and smoking were associated with an overall associated with an overall development conversely BMI with an increase. Finally, we demonstrate that the proposed approach offers a powerful framework for describin conversely BMI with an increase that  $m_1$ , we demonstrate that the proposed approach offer and proverful framework for describing taxa networks at the individual level and predicting host features. powerful framework for describing taxa networks at the individual level and predicting host features.

**Introduction**<br>Major initiatives such as the Human Microbiome Project<sup>1</sup>, and epidemiological studies revealed Major initiatives such as the Human Microbiome Project<sup>+</sup>, and epidemiological studies revealed<br>
bortant features associated with changes in the composition of the gut microbiome, including<br>  $e^{24}$ , sex<sup>5-7</sup>, BMI<sup>7-9</sup>, s age<sup>2-4</sup>, sex<sup>5-7</sup>, BMI<sup>7-9</sup>, smoking habits<sup>10-12</sup>, long term diet<sup>8,3,3-17</sup>, and host genetics<sup>17-22</sup> in both healthy<br>individuals and disease cases<sup>23-25</sup>. Yet, these studies also highlighted important challenges in<br>dec individuals and disease cases<sup>23-25</sup>. Yet, these studies also highlighted important challenges in deciphering the complex host-microbiome relationship<sup>26,27</sup>. In particular, despite its high dimensional nature, most exist deciphering the complex host-microbiome relationship<sup>20,27</sup>. In particular, despite its high dimensional<br>nature, most existing human microbiome studies still focus on univariate approaches, testing<br>associations between a v nature, most external matrice of interest and each single taxon. To address this limitation, there<br>has been an increasing interest in multivariate approaches. This includes the application of the alpha<br>and beta diversity has been an increasing interest in multivariate approaches. This includes the application of the alpha<br>and beta diversity indexes, which measure intra- and inter-sample distances between microbiome<br>samples<sup>19,22,28</sup>, and and beta diversity indexes, which measure intra- and inter-sample distances between microbiome<br>samples<sup>19,22,28</sup>, and various multivariate mean-based approaches to conduct a joint test of the<br>abundance of multiple taxa wi

samples<sup>19,22,28</sup>, and various multivariate mean-based approaches to conduct a joint test of the<br>abundance of multiple taxa with a predictor<sup>29-31</sup>. Conversely, the study of features associated with<br>variability in the coabundance of multiple taxa with a predictor<sup>29-32</sup>. Conversely, the study of features associated with<br>variability in the co-abundance of taxa remains puzzling<sup>32-34</sup>.<br>The study of gut microbiome taxa co-abundance, which c variability in the co-abundance of taxa remains puzzling<sup>22-34</sup>.<br>The study of gut microbiome taxa co-abundance, which<br>the gut microbiome, is arising as a promising research top<br>ecosystem whose constituents form sub-commun the gut microbiome, is arising as a promising research topic<sup>35-40</sup>. The gut microbiome is a complex ecosystem whose constituents form sub-communities through interactions between individual taxa. Those sub-communities, s the gut microbiome, is arising as a promising research topic<sup>23-3</sup>. The gut microbiome is a complex<br>ecosystem whose constituents form sub-communities through interactions between individual taxa.<br>Those sub-communities, som Those sub-communities, sometimes referred to as guild<sup>37</sup>, or cliques, display co-abundances because<br>they exploit the same class of resources or work together as a coherent functional group<sup>32,38</sup>.<br>Previous studies already they exploit the same class of resources or work together as a coherent functional group<sup>32,38</sup>.<br>Previous studies already illustrated the potential differences in co-abundance networks across inflammatory bowel disease st they exploit the same class of resources or work together as a coherent functional group<sup>-2,50</sup>.<br>Previous studies already illustrated the potential differences in co-abundance networks across<br>inflammatory bowel disease sta inflammatory bowel disease status and body mass index<sup>35</sup>, and geographically diverse populations<sup>39</sup>.<br>However, there is currently no gold standard method to screen for factors associated with changes in<br>the co-abondance the co-abondance of bacteria across individuals in a population. Existing methods are typically based<br>on the inference and pairwise comparison of networks across conditions, using a threshold-based<br>approach to define co-ab the inference and pairwise comparison of networks across conditions, using a threshold-based<br>approach to define co-abundances when the phenomenon is likely continuous. Furthermore,<br>previous works showed that the inferred n approach to define co-abundances when the phenomenon is likely continuous. Furthermore, previous works showed that the inferred networks can vary substantially across approaches<sup>41</sup>. More problematic, by construction thos previous works showed that the inferred networks can vary substantially across approaches<sup>41</sup>. More<br>problematic, by construction those methods are restricted to categorical predictors, they do not<br>allow for variable adjus allow for variable adjustment, and they do not provide a formal global test of association. We<br>recently developed MANOCCA<sup>42</sup>, a formal statistical framework to test the effect of both categorical<br>and continuous predictor

recently developed MANOCCA<sup>42</sup>, a formal statistical framework to test the effect of both categorical<br>and continuous predictors on the covariance matrix of a multivariate outcome  $-a$  metric directly<br>proportional to the corecently developed MANOCCA<sup>22</sup>, a formal statistical framework to test the effect of both categorical<br>and continuous predictors on the covariance matrix of a multivariate outcome –a metric directly<br>proportional to the co-a and comportional to the co-abundance- that addresses these methodological gap.<br>
Here, we applied MANOCCA to study host factors associated with variability in the gut<br>
microbiome co-abundance network of 938 healthy particip proportional to the co-abundance network of 938 healthy participants from the M<br>microbiome co-abundance network of 938 healthy participants from the M<br>and compared results with multiple standard univariate and multivariate Here, we applicant matter of 938 healthy participants from the Milieu Interieur cohort<sup>43</sup>,<br>I compared results with multiple standard univariate and multivariate approaches. We then<br>inducted an in-depth examination of the microbiome co-abundance network of 938 healthy participants from the Milieu Interieur cohort<sup>-5</sup>, and compared results with multiple standard univariate and multivariate approaches. We then conducted an in-depth examinatio and conducted an in-depth examination of the effect of each associated factor on the taxa interaction<br>network, highlighting the key taxa impacted and how these factors shape the microbiome<br>composition. Finally, we used our network, highlighting the key taxa impacted and how these factors shape the microbiome<br>composition. Finally, we used our framework to assess the performances of predictive models of the<br>associated factor based on taxa co-a composition. Finally, we used our framework to assess the performances of predictive models of the<br>associated factor based on taxa co-abundances.<br>**Results**<br>Identifying host factors associated with taxa co-abundance<br>The com

## Results

Examples associated factor based on taxa co-abundances.<br> **Results**<br> **Results**<br> **Composition of the gut microbiome of a single individual can vary with the many host factors**<br> **Composition of the gut microbiome of a single Results<br>
Results**<br> **Results**<br> **Composition of the gut microbiome of a**<br>
driving the molecular environment of the g<br>
characteristics might promote specific mechanis  $\frac{1}{2}$ <br> $\frac{1}{2}$ Identifying host factors associated with taxa co-abundance<br>The composition of the gut microbiome of a single individual can vary with the many host factors The molecular environment of the gut. Exposure to a given environment and host inacteristics might promote specific mechanisms, inducing a collaboration across a limited number species, while another environment would prom characteristics might promote specific mechanisms, inducing a collaboration across a limited number<br>of species, while another environment would promote only a subset of the same species. The overlap<br>of such mechanisms can characteristics might promote represent mechanisms, including a subset of the same species. The overlap<br>of such mechanisms can result in various co-abundances independently of a mean effect of the<br>exposures in question (Fi of such mechanisms can result in various co-abundances independently of a mean effect of the<br>exposures in question (Fig. 1). We screened 80 host variables, including demographics, socio-<br>exposures in question (Fig. 1). We exposures in question (Fig. 1). We screened 80 host variables, including demographics, socio-<br>
exposures in question (Fig. 1). We screened 80 host variables, including demographics, socioexposures in question (Fig. 1). We screened 80 host variables, including demographics, socio-

processional, and dietary habits measurements (Table 31), for association with overall taxa co-<br>abundance network in healthy individuals using MANOCCA. We conducted analyses at the species,<br>genus, and family levels, focus

genus, and family levels, focusing on the most common taxa<sup>44</sup>, including respectively 675, 718, and<br>151 taxa after quality control filtering (Fig. 2 and Table S2). Except when used as predictors, all<br>analyses were adjust 151 taxa after quality control filtering (Fig. 2 and Table S2). Except when used as predictors, an analyses were adjusted for age, sex and body mass index (BMI).<br>
We identified associations with co-abundance variability a We identified associations with co-abundance variability at<br>after correction for multiple testing  $(n = 8,000 \text{ tests}, P \text{ threshold}$ <br> $10^{-55}$ ,  $P_{genus} = 3.5 \times 10^{-56}$ ,  $P_{family} = 9.2 \times 10^{-37}$ ), sex  $(P_{species} = 2.2 \times 10^{-18})$ , smoking  $(P_{species} = 2.8 \$ after correction for multiple testing ( $n = 8,000$  tests, P threshold = 6.25 x 10<sup>-6</sup>) with age ( $P_{species} = 2.0$  x 10<sup>-55</sup>,  $P_{genus} = 3.5$  x 10<sup>-56</sup>,  $P_{family} = 9.2$  x 10<sup>-37</sup>), sex ( $P_{species} = 2.2$  x 10<sup>-17</sup>,  $P_{genus} = 6.3$  x 10<sup>-22</sup>, 10<sup>-5</sup>,  $P_{genus} = 3.5 \times 10^{-5}$ ,  $P_{family} = 9.2 \times 10^{-5}$ ), sex  $(P_{species} = 2.2 \times 10^{-5}$ ,  $P_{genus} = 6.3 \times 10^{-2}$ ,  $P_{family} = 3.1 \times 10^{-18}$ ), smoking  $(P_{species} = 2.8 \times 10^{-14}$ ,  $P_{genus} = 1.6 \times 10^{-20}$ ,  $P_{family} = 5.6 \times 10^{-13}$  (Fig. 3). Analyses 18) level further identified an association with body mass index ( $P_{genus} = 5.9 \times 10^{-6}$ ). Note that MANOCCA uses a dimension-reduction step to address the large number of parameters in the taxa covariance matrix. Varying t level further identified an association with body mass index ( $P_{genus}$  = 5.9 x 10°). Note that MANOCCA uses a dimension-reduction step to address the large number of parameters in the taxa covariance matrix. Varying the par uses a dimension-reduction-reduc matrix. The parametrization of this step provided in Figures S1-S2 and Supplementary Notes. We also conducted sensitivity analyses, assessing the variability in the results when applying MANOCCA to random subsets of availa

analysis are provided in Figures 31-52 and supplementary Notes. We also conducted sensitivity<br>analyses, assessing the variability in the results when applying MANOCCA to random subsets of<br>available taxa. Overall, the larg available taxa. Overall, the larger the set of taxa included, the stronger the association signal (Fig. 53), highlighting a global effect of the four factors on the microbiome co-abundance network.<br>We applied two alternat **avantable taxa.** Overall, the larger the set of taxa included, the stronger the association signal (Fig. 53), highlighting a global effect of the four factors on the microbiome co-abundance network.<br>We applied two altern S3), inginighting a global effect of the four factors on the intertoblome co-abundance network.<br>We applied two alternative multivariate approaches for comparison purposes: a st<br>MANOVA, testing for association between each MOVA, testing for association between each host factor and the joint abundance of taxa, and<br>ha diversity using the Shannon and Simpson indexes (Fig. 3, Table S3). Some factors were<br>nificant after correction for multiple t alpha diversity using the Shannon and Simpson indexes (**Fig. 3, Table S3**). Some factors were<br>significant after correction for multiple testing (*P* threshold of 6.3 x 10<sup>-4</sup> accounting for the 80 tests<br>per approach) but alpha diversity using the Shannon and Simpson indexes (Fig. 3) Table 33). Some factors were<br>significant after correction for multiple testing (P threshold of 6.3 x 10<sup>-4</sup> accounting for the 80 tests<br>per approach) but at a significant after correction for multiple testing (*P* threshold of 6.3 x 10<sup>-</sup> accounting for the 80 tests<br>per approach) but at a much lower significance level. The MANOVA identified an association with age<br>at the specie per approach) but a much lower significance in the species and family level ( $P_{species} = 5.2 \times 10^{-6}$ ,  $P_{family} = 1.0 \times 10^{-29}$ ) and with sex at the family level ( $P_{Simpson} = 3.4 \times 10^{-4}$ ,  $P_{Shannon} = 3.9 \times 10^{-5}$ ) and the Shannon index ( $P = 5.2 \times 10^{-6}$ ). Both Simpson and Shannon indexes identified a signal with age at the family level ( $P_{Simpson} = 3.4 \times 10^{-4}$ ,  $P_{Shannon} = 3.9 \times 10^{-5}$ ) and the Shannon index also identified age at the genus level ( $P_{Shannon} = 1.6$  $(P_{Sinnpon} = 3.4 \times 10^{-7}, P_{Shannon} = 3.9 \times 10^{-7})$  and the Shannon index also identified age at the genus level  $(P_{Shannon} = 1.6 \times 10^{-5})$ . Altogether, the additional signals observed with smoking and BMI, and the stronger association of (*P<sub>shannon</sub>* = 1.6 x 10<sup>-</sup>). Altogether, the additional signals observed with smoking and BMI, and the stronger association of age and sex with the covariance-based approach as compared to mean-based and diversity index a stranger information content of the co-<br>abundances of taxa over these existing metrics to describe the relationship between the gut<br>microbiome of healthy individuals and these host variables.<br>**Contribution of taxa on the c** 

signal can be easily extracted. We derived these weights for the age, sex, smoking and BMI signals, Franchine of taxa on the co-abundance association signal<br>
By construction of taxa on the co-abundance association signal<br>
By construction, the MANOCCA statistic is a weighted sum of the contribution from each pair of<br>
taxa **Contribution of taxa on the co-abundance association**<br>By construction, the MANOCCA statistic is a weighted s<br>taxa (see **Methods**), so that the contribution weight of eac<br>signal can be easily extracted. We derived these we (<br>t si<br>f f **Contribution of taxa on the co-abundance association signal**<br>By construction, the MANOCCA statistic is a weighted sum of the contribution from each pair of By construction, the Construction weight of each taxon on the co-abundance association<br>al can be easily extracted. We derived these weights for the age, sex, smoking and BMI signals,<br>using on the genus level, which display taxa (see Methods), so that the contribution weight of each taxon on the co-abundance association<br>signal can be easily extracted. We derived these weights for the age, sex, smoking and BMI signals,<br>focusing on the genus le signal can be easily and signal can be easier association. As shown in Figure S4a, most<br>focusing on the genus level, which displayed the strongest association. As shown in Figure S4a, most<br>faxa display a non-zero contribut focusing on the genus lever, which displayed the strongest association. As shown in Figure 3-a, most taxa display a non-zero contribution to the association, highlighting again the global effect of these factors on the mi factors on the microbiome composition. Nevertheless, these contributions show substantial<br>heterogeneity, with a limited number of taxa pairs displaying outstanding weights as compared to<br>the expected under the null (Fig. heterogeneity, with a limited number of taxa pairs displaying outstanding weights as compared to<br>the expected under the null (Fig. **S4a-b**). To assess potential links between effects on co-abundance<br>and effects on relativ the expected under the null (Fig. S4a-b). To assess potential links between effects on co-abundance<br>the expected under the null (Fig. S4a-b). To assess potential links between effects on co-abundance<br>and effects on relati the expected under the null (Fig. 34a-b). To assess potential links between entects on co-abundance<br>and effects on relative abundance, we compared the MANOCCA weights against the univariate mean<br>effect *P*-value associati effect *P*-value associations derived using a linear regression (see **Methods** and **Table S4**). As shown in<br>Figure S4c, we observed a positive and significant correlation between the two terms for all four<br>variables (age, Figure S4c, we observed a positive and significant correlation between the two terms for all four variables (age,  $P = 1.6 \times 10^{-12}$ ; sex,  $P = 1.3 \times 10^{-40}$ ; smoking,  $P = 2.1 \times 10^{-5}$ ; BMI,  $P = 4.6 \times 10^{-15}$ ), suggesting a variables (age,  $P = 1.6 \times 10^{-12}$ ; sex,  $P = 1.3 \times 10^{-40}$ ; smoking,  $P = 2.1 \times 10^{-5}$ ; BMI,  $P = 4.6 \times 10^{-15}$ , suggesting a dual impact of these factors on the abundance and the co-abundance of many of these genera, in agr variables (age,  $P = 1.6 \times 10^{-4}$ ; sex,  $P = 1.3 \times 10^{-4}$ ; smoking,  $P = 2.1 \times 10^{-4}$ ; BMI,  $P = 4.6 \times 10^{-4}$ ), suggesting a dual impact of these factors on the abundance and the co-abundance of many of these genera, in agree genera, in agreement with the existing literature<sup>2-12</sup>. The correspondence is especially marked for sex. Several top genera contributing to the co-abundance also pass a stringent Bonferroni correction threshold ( $P < 8.7 \$ genera, in agreement with the existing literature<sup>212</sup>. The correspondence is especially marked for<br>sex. Several top genera contributing to the co-abundance also pass a stringent Bonferroni correction<br>threshold ( $P < 8.7 \times$ threshold ( $P < 8.7 \times 10^{-7}$ ) for univariate association. This includes *Bacteroides* ( $P = 1.9 \times 10^{-7}$ ), threshold  $(P < 8.7 \times 10^7)$  for univariate association. This includes *Bacteroides*  $(P = 1.9 \times 10^7)$ ,

Coprococcus B (P = 7.1 x 10<sup>-</sup>), Anaerotruncus (P = 8.5 x 10<sup>-</sup>), Agathobacter (P = 3.0 x 10<sup>-</sup>), Alistipes (P<br>
= 2.2 x 10<sup>-8</sup>), and Intestinimonas (P = 1.9 x 10<sup>-9</sup>).<br>
We next investigated the characteristics of the top = 2.2 x 10°), and *Intestinimonas* ( $P = 1.9$  x 10°).<br>We next investigated the characteristics of contribution to the variability in co-abundance 7, 9 and 6 overlapping sets of families, covered age, sex, smoking and BMI, We next investigated the characteristics of the characteristics of the top states of 9,<br>
1) and 6 overlapping sets of families, covered 50% or more of those top contributing genera with<br>
1), sex, smoking and BMI, respectiv 2.9 and 6 overlapping sets of families, covered 50% or more of those top contributing genera with<br>age, sex, smoking and BMI, respectively. Those key families include the ones with the highest relative<br>abundance, *Lachnospi* age, sex, smoking and BMI, respectively. Those key families include the ones with the highest relative<br>abundance, Lachnospiraceae (23.6%), Bacteroidaceae (22.4%), Ruminococcaceae (7.8%),<br>Acutalibacteraceae (6.4%), Oscillos abundance, Lachnospiraceae (23.6%), Bacteroidaceae (22.4%), Ruminococcaceae (7.8%),<br>Acutalibacteraceae (6.4%), Oscillospiraceae (5.6%), but also some rare ones: Eggerthellaceae (0.5%),<br>Peptostreptococcaceae (0.6%), Muribac abundance, Lachnospiraceae (23.6%), Bacteroidaceae (22.4%), Naminococcaceae (7.6%),<br>Acutalibacteraceae (6.4%), Oscillospiraceae (5.6%), but also some rare ones: Eggerthellaceae (0.5%),<br>Peptostreptococcaceae (0.6%), Muriba *Reptostreptococcaceae* (0.4%), Oscillospiraceae (5.6%), but also some rare ones. *Eggertheliaceae* (0.5%),<br>*Peptostreptococcaceae* (0.6%), *Muribaculaceae* (0.5%), and four unspecified Co-Abundance Groups<br>(CAG-74, CAG-508 reprostreptococcaceae (0.6%), Muribaculaceae (0.5%), and four dispectived co-Abundance Groups (CAG-74, CAG-508, CAG-272, CAG-138) (Fig. 4a). While the representativity of families involved in co-<br>abundance variability was (CAG-74, CAG-508, CAG-272, CAG-138) (Fig. 4a). While the representativity of families involved in co-<br>abundance variability was somewhat proportional to their relative abundance, we noted several<br>major differences. Some fa major differences. Some families, such as *Bacteroidaceae* are largely underrepresented in the co-<br>abundance signal. Conversely, co-abundance involving the *Oscillospiraceae* family are strongly<br>impacted by all four factor major differences. Some rammes, such as *butteroidatede* are largely didertepresented in the co-<br>abundance signal. Conversely, co-abundance involving the *Oscillospiraceae* family are strongly<br>impacted by all four factors,

abundance signal. Conversely, co-abundance involving the Oscillospiraceae family are strongly<br>impacted by all four factors, and by BMI in particular. Other families also display factor-specific<br>enrichment, including *Pepto* impresentially, including *Peptostreptococcaceae* and *Muribaculaceae* with smoking, two families<br>already reported to be associated with smoking status<sup>45,46</sup>.<br>Finally, we examined the composition of the top contributing p enrichment, including Peptostreptococcaceae and Muridoccalaceae with smoking, two families<br>already reported to be associated with smoking status<sup>45,46</sup>.<br>Finally, we examined the composition of the top contributing pairs of already reported to be associated with smoking status<sup>45,46</sup>.<br>Finally, we examined the composition of the top co<br>whether they involve changes in interaction within the sainteraction between genera from different families ( Finally (intra-family co-abundance) or<br>Finally interaction between genera from different families (cross-family co-abundance) (**Fig. 4b**). Within-<br>hily co-abundance represented a small fraction of all interactions, with th wheteraction between genera from different families (cross-family co-abundance) (Fig. 4b). Within-<br>family co-abundance represented a small fraction of all interactions, with the vast majority of<br>interactions taking place b Interaction between genera from different families (cross-family co-abundance) (Fig. 4b). Within-<br>family co-abundance represented a small fraction of all interactions, with the vast majority of<br>interactions taking place be Families (Fig. 4c-f). Besides a few exceptions<br>(e.g. variability in the co-abundance between taxa from *Oscillospiraceae* and *Acutalibacteraceae*<br>families for BMI), we did not observe any marked pattern.<br>**Network of impac** 

interactions taking place between genera of different families (Fig. 4c-f). Besides a few exceptions<br>(e.g. variability in the co-abundance between taxa from *Oscillospiraceae* and *Acutalibacteraceae*<br>families for BMI), we (e.g. variability in the co-abundance between taxa from Oscillospiraceae and Acutalibacteraceae<br>families for BMI), we did not observe any marked pattern.<br>**Network of impacted taxa**<br>We formed a network of co-abundance varia **Network of impacted taxa**<br>We formed a network of co-abundance variation from<br>to the MANOCCA association signal with age, sex, smoki<br>**Material**). Altogether, these 4,000 pairs involved a total of<br>there was a substantial ov ノ<br>ししょう **Network of impacted taxa**<br>We formed a network of co-abundance variation from the top 1,000 pairs of genera contributing the MANOCCA association signal with age, sex, smoking, and BMI (Fig. 5a, and supplementary<br>terial). Altogether, these 4,000 pairs involved a total of 476 unique genera. As shown in Figure 5b,<br>re was a substantial overlap i to the MANOCCA association signal with age, sex, sinoking, and BMI (Fig. 3a, and supplementary<br> **Material**). Altogether, these 4,000 pairs involved a total of 476 unique genera. As shown in Figure 5b,<br>
there was a substant Material). Altogether, these 4,000 pairs involved a total of 476 unique genera. As shown in Figure 5b, there was a substantial overlap in pairs of co-abundant taxa impacted by sex and BMI (N= 658 pairs, approximately 66% o approximately 66% of sex and BMI associated pairs), and age and smoking habits (N=306 pairs,<br>approximately 31% of age and smoking-associated pairs). Conversely, the overlap across the nine<br>other pairs of factors was null o approximately 31% of age and smoking-associated pairs). Conversely, the overlap across the nine<br>other pairs of factors was null or negligible. At the taxa level, a core of 200 genera were shared<br>across all predictors (Fig. The pairs of factors was null or negligible. At the taxa level, a core of 200 genera were shared<br>across all predictors (Fig. 5c). Other genera were evenly spread across factors, except for age and<br>smoking which involved 49 Find a core of the pairs of factors and production, with a share of the taxa level of the taxa in smoking which involved 49 (13%) and 54 (15%) genera specific to those two factors, respectively.<br>Together, this suggests tha across all predictors (Fig. 5c). Other genera were evenly spread across factors, except for age and<br>smoking which involved 49 (13%) and 54 (15%) genera specific to those two factors, respectively.<br>Together, this suggests t smoking which is suggests that the four factors partly control the interacting partners of this core<br>genera. As shown in Figure 5d, increased age and smoking are mostly associated with a decrease in<br>co-abundances, with 86% The Figure 1.1 Section in Figure 5d, increased age and smoking are mostly associated with a decrease in co-abundances, with 86% and 75% of top pairs displaying negative associations with these two factors, respectively. BM genera. As shown in Figure 5d, increased age and showing are mostly associated with a decrease in<br>co-abundances, with 86% and 75% of top pairs displaying negative associations with these two<br>factors, respectively. BMI exhi

factors, respectively. BMI exhibited an opposing trend, with 72% of top pairs showing an increase in<br>co-abundance with increasing BMI. The sex predictor displayed a more balanced distribution, with a<br>60% decrease and a 40% factors, respectively. Each increasing BMI. The sex predictor displayed a more balanced distribution, with a 60% decrease and a 40% increase of co-abundance in males as compared to females.<br>
Multiple patterns emerged when 60% decrease and a 40% increase of co-abundance in males as compared to females.<br>
Multiple patterns emerged when exploring the contributing genera. Those shared between<br>
smoking and age are especially enriched in the *Osci* Multiple patterns emerged when exploring the contributing genera. Those s<br>smoking and age are especially enriched in the *Oscillospirales* order (e.g. *Massilliocl*<br>180, CAG-1427, *Marseille-P4683, and MGYG-HGUT-03297*), a between and age are especially enriched in the Oscillospirales order (e.g. Massillioclostridium, CAG-<br>
2, CAG-1427, Marseille-P4683, and MGYG-HGUT-03297), and consistently exhibited reduced co-<br>
undances with the core taxa smoking and age are especially enfielded in the oscillospirales order (e.g. Massilinocloschatam, exception, 180, CAG-1427, Marseille-P4683, and MGYG-HGUT-03297), and consistently exhibited reduced co-<br>abundances with the c 180, CAG-1427, Marseine-P4683, and MGTO-HGOT-03237), and consistently exhibited reduced co-<br>abundances with the core taxa. Among genera unique to smoking, *Bacteroides A* genus was by far<br>the most impacted, showing a reduc abundances with the core taxa. Among genera unique to smoking, *bucceroides* A genus was by far<br>the most impacted, showing a reduction of co-abundances with many of the core taxa. Interestingly,<br>the relative abundance of t the relative abundance of this common genus (detected in 99% of participants, **Table S2**) was not associated with smoking status in our data (*P*-value from a linear regression equals 0.26, **Table S4**). This suggests that the relative abundance of this common genus (detected in 99% of participants, Table 32) was not<br>associated with smoking status in our data (*P*-value from a linear regression equals 0.26, **Table S4**).<br>This suggests that sm associated with smoking status in our data (P-value from a linear regression equals 0.20, Table 34).<br>This suggests that smoking might only break some of its interactions with other genera without  $T_{\rm eff}$  suggests that smoking might only break some of its interactions with other genera with

missed by standard abundance-based approaches. A subset of genera contributing to association<br>with BMI, sex and smoking was enriched from the *Lachnospiraceae* family (*Ruminococcus A, Dorea,*<br>Coprococcus B, GCA-900066135, missed by The Mill, sex and smoking was enriched from the *Lachnospiraceae* family (*Ruminococcus A, Dorea,*<br>Coprococcus B, GCA-900066135, Agathobacter), displaying both increased and decreased co-<br>abundances across predic Coprococcus B, GCA-900066135, Agathobacter), displaying both increased and decreased co-<br>abundances across predictors. Both *Lachnospiraceae* relative abundance and co-abundance with<br>other taxa have already been found to b

exprococcus B, GCA 500006135, Agathobacter), displaying both increased and decreased co-<br>abundances across predictors. Both *Lachnospiraceae* relative abundance and co-abundance with<br>other taxa have already been found to b abundances across predictors. Both Lachnosphaceae relative abundance and co-abundance with<br>other taxa have already been found to be associated with human diseases and obesity in<br>particular<sup>35,47,48</sup>.<br>Finally, to assess the particular<sup>35,47,48</sup>.<br>
Finally, to assess the relevance of the covariance-based co-abundances network impacted, we<br>
considered two alternative approaches. A naïve permutation-based approach, inspired from the<br>
existent<sup>41</sup> Finally, the relevance approaches. A naïve permutation-based approach, inspired from the stent<sup>41</sup>, that produces an empirical comparison of pairwise covariance between all taxa (see **thods**), and the commonly used SparCC existent<sup>41</sup>, that produces an empirical comparison of pairwise covariance between all taxa (see **Methods**), and the commonly used SparCC<sup>49</sup> approach (see **Supplementary Notes** for a detailed description of the two approa Methods), and the commonly used SparCC<sup>--</sup> approach (see **Supplementary Notes** for a detailed description of the two approaches). Note that both the SparCC and the permutation-based approaches, like all existing method, ar approaches, like all existing method, are limited to binary predictors and use a threshold to define<br>co-abundance in each group studied. Both methods are meant to detect significant differences in<br>pairwise taxa correlation co-abundance in each group studied. Both methods are meant to detect significant differences in<br>pairwise taxa correlations across values of a categorical predictor, and should in theory detect effect<br>on co-abundance varia co-abundance in each group studied. Both methods are meant to detect significant differences in pairwise taxa correlations across values of a categorical predictor, and should in theory detect effect on co-abundance varia particular correlations are interestingly similar to those detected by MANOCCA. However, as showed in the simulation from **Figure S5**, permutation-based shows poor specificity as compared to MANOCCA, and SparCC shows the simulation from **Figure S5**, permutation-based shows poor specificity as compared to MANOCCA, and<br>SparCC shows the poorest performances in this simulation with almost no power. We applied both<br>methods to the two binary pr Simulation from Figure 35, permutation-based shows poor specificity as compared to MANOCCA, and<br>SparCC shows the poorest performances in this simulation with almost no power. We applied both<br>methods to the two binary pred methods to the two binary predictors, sex and smoking, at the genus level and crossed the results<br>with MANOCCA's top contributing products. The overlap between MANOCA and the two alternative<br>methods was very modest, but h with MANOCCA's top contributing products. The overlap between MANOCA and the two alternative<br>methods was very modest, but highly significant with minimum *P*-values of 1 x 10<sup>-145</sup> and 2 x 10<sup>-153</sup> for<br>sex and smoking res methods was very modest, but highly significant with minimum *P*-values of 1 x 10<sup>-145</sup> and 2 x 10<sup>-153</sup> for sex and smoking respectively for the permutation approach (Fig. S6 a-b), and 1 x 10<sup>-70</sup> and 1 x 10<sup>-82</sup> for sex

sex and smoking respectively for the permutation approach (Fig. **S6 a-b**), and 1 x 10<sup>-0</sup> and 1 x 10<sup>-2</sup><br>for sex and smoking respectively for the SparCC approach (Fig. **S6c-d**), thus, confirming that those<br>three alternati For sex and smoking respectively for the sparce approach (Fig. Sec-d), thus, commining that those<br>three alternative methods do detect some similar network components.<br>**Prediction of individual features based on taxon corre Prediction of individual features based on taxon correlation**<br>Our framework is built out of a linear model where the covariance<br>level. This is a major advantage over existing correlation approaches<sup>41</sup><br>complementary analy ? ( e r c c c **Prediction of individual features based on taxon correlation**<br>Our framework is built out of a linear model where the covariance is defined at the individual el. This is a major advantage over existing correlation approaches<sup>41</sup>, that allows for a range of<br>nplementary analyses. One particularly important extension is the possibility of training a<br>dictive model of an outcome ba level. This is a major advantage over existing correlation approaches<sup>-2</sup>, that allows for a range of complementary analyses. One particularly important extension is the possibility of training a predictive model of an ou predictive model of an outcome based on taxa covariance, so that the outcome in question can be<br>predicted for any new individual based on its microbiome (and conversely). Here, we assessed the<br>accuracy of MANOCCA to predi predicted for any new individual based on its microbiome (and conversely). Here, we assessed the accuracy of MANOCCA to predict the four most associated features (age, sex, smoking and BMI), using taxa from the species, g predict accuracy of MANOCCA to predict the four most associated features (age, sex, smoking and BMI), using taxa from the species, genus and family level and a 30-fold cross-validation. Accuracy was derived using squaredaccuracy of Manuson and Family level and a 30-fold cross-validation. Accuracy was<br>derived using squared-correlation ( $r^2$ ) for continuous outcomes (age and BMI), and using the area<br>under the receiver operating curve (AUC derived using squared-correlation ( $r^2$ ) for continuous outcomes (age and BMI), and using the area<br>under the receiver operating curve (AUC) for binary outcomes (smoking and sex). We compared the<br>covariance-based predicti

under the receiver operating curve (AUC) for binary outcomes (smoking and sex). We compared the covariance-based prediction model against a standard linear model based on the relative abundance of each single taxa.<br>As sho covariance-based prediction model against a standard linear model based on the relative abundance<br>of each single taxa.<br>As showed in **Figure 6**, the MANOCCA strongly outperforms the standard mean-based prediction<br>model, be covariance based prediction model, being significantly more accurate in all scenarios we considered. Gain in prediction model, being significantly more accurate in all scenarios we considered. Gain in prediction was espec As showed in Fig<br>As showed in Fig<br>model, being signifiespecially large for<br>MANOCCA equals 0<br>comparison, the me<br>significantly higher<br>species, genus and As showed in Figure 6, the MANOCCA strongly outperforms the standard mean-based prediction<br>del, being significantly more accurate in all scenarios we considered. Gain in prediction was<br>ecially large for age with up to a t especially large for age with up to a three-fold increase in power. The median of  $r_{age}^2$  from the<br>MANOCCA equals 0.27, 0.25 and 0.18 for models based on species, genus and family, respectively. In<br>comparison, the mean-b especially large for age with up to a three-fold increase in power. The median of  $r_{age}^2$  from the MANOCCA equals 0.27, 0.25 and 0.18 for models based on species, genus and family, respectively. In comparison, the mean-b uye comparison, the mean-based model  $r_{age}^2$  equal 0.10, 0.07 and 0.05, respectively. Prediction was also significantly higher for sex, with AUCs of 0.66, 0.64, and 0.64 for the mean-based model at the species, genus and fam species, genus and family level, respectively, and AUCs of 0.67, 0.69 and 0.70 for the covariance-<br>based model. This confirms the higher information content of co-abundance as compared to<br>abundance, and demonstrates the va uye  $\tilde{\tau}_{ge}$  equal 0.10, 0.07 and 0.05, respectively. Prediction was also<br>is of 0.66, 0.64, and 0.64 for the mean-based model at the<br>ectively, and AUCs of 0.67, 0.69 and 0.70 for the covariance-<br>ther information content of species, genus and family level, respectively, and AUCs of 0.67, 0.69 and 0.70 for the covariance-<br>based model. This confirms the higher information content of co-abundance as compared to<br>abundance, and demonstrates the va abundance, and demonstrates the validity of using covariance-based co-abundances for prediction<br>purposes.<br>The higher information content of co-abundances for prediction<br>purposes.  $\tt purpose.$ purposes.

### **Discussion**

dence that species form functionally coherent groups that work together to exploit the same<br>ources from the local environment<sup>38</sup>. Studying those groups, rather than each single taxa, might<br>p better understand the role of resources from the local environment<sup>38</sup>. Studying those groups, rather than each single taxa, might<br>help better understand the role of the microbiome in human health outcomes. With this same<br>argument, it has already been argument, it has already been proposed to study those groups through variability in the network of co-abundances<sup>35,39</sup>. Although simple in principle the implementation of this objective can be challenging in practice. He ary and alternation of this objective can be challenging in practice. Here, we applied MANOCCA<sup>42</sup>, a recently developed method, that allows to conduct a formal statistical test of association between taxa covariance and

challenging in practice. Here, we applied MANOCCA<sup>42</sup>, a recently developed method, that allows to conduct a formal statistical test of association between taxa covariance and any predictor, whether continuous, categorical continuous, categorical or binary, to investigate host features associated with the gut microbiome<br>co-abundance in 938 healthy individuals from the Milieu Interieur cohort.<br>We identified highly significant associations bet co-abundance in 938 healthy individuals from the Milieu Interieur cohort.<br>We identified highly significant associations between taxa co-abundance variability and age, sex,<br>smoking status and body mass index (BMI). Except f We identified highly significant associations between taxa co-abundar<br>smoking status and body mass index (BMI). Except for BMI, these associ<br>three taxonomic levels studied: species, genus, and family. In compariso<br>and dive between taxonomic levels studied: species, genus, and family. In comparison were detected at all<br>ee taxonomic levels studied: species, genus, and family. In comparison mean-based multivariate<br>diversity-based analyses only three taxonomic levels studied: species, genus, and family. In comparison mean-based multivariate<br>and diversity-based analyses only identified associations with age, and one association with sex at<br>the family level, but at and diversity-based analyses only identified associations with age, and one association with sex at<br>the family level, but at a much lower significance level. For the four associated features, there was a<br>significant correl the family level, but at a much lower significance level. For the four associated features, there was a<br>significant correlation between contribution to co-abundance and univariate effect on relative<br>abundance, suggesting t the families and the families in this study we applied it to are sox smoking and BML For all fortures to huild productive models. In this study we applied it to are sox smoking and BML For all fortures at the huild product significant correlation between the features impact both the abundance and co-abundance of taxa.<br>The network of top contributing genera shows that variability in interactions were concentrated in a<br>limited number of famili The network of top contributing genera shows that variability in interactions were concentrated in a<br>limited number of families and are essentially taking place between genera of different families,<br>rather than between gen Imited number of families and are essentially taking place between genera of different families,<br>
rather than between genera of the same family. The overlap of top contributing taxa over the four<br>
features was substantial, rather than between genera of the same family. The overlap of top contributing taxa over the four<br>features was substantial, especially between age and smoking, and between sex and BMI, suggesting<br>potentially shared mechani reatures was substantial, especially between age and smoking, and between sex and BMI, suggesting<br>potentially shared mechanisms. Finally, we demonstrate that the MANOCCA framework can be used<br>to build predictive models. In Features was substantially shared mechanisms. Finally, we demonstrate that the MANOCCA framework can be used<br>to build predictive models. In this study we applied it to age, sex, smoking and BMI. For all features,<br>the predi

potentially shared meddels. In this study we applied it to age, sex, smoking and BMI. For all features, the predictive power based on co-abundances was significantly and systematically higher than for a standard mean-based the predictive power based on co-abundances was significantly and systematically higher than for a<br>standard mean-based multivariate model, with up to a three-fold increase in r-squared for age.<br>Our study also has limitatio trandard mean-based multivariate model, with up to a three-fold increase in r-squared for age.<br>
Our study also has limitations. First, the approach is not applicable to microbiome data of small<br>
sample sizes. Despite the d Standard mean-based multivariate model, which up to a hindeptoted incredibte to microbiome data of<br>sample sizes. Despite the data reduction steps through principal component analysis, the num<br>PCs analysed should remain sub The sizes. Despite the data reduction steps through principal component analysis, the number of<br>analysed should remain substantially larger than the sample size, thus limiting the application to<br>asets of 100 participants o Between the data reduction in the sample size, thus limiting the application to datasets of 100 participants or more. Hopefully, this will become less of an issue thanks to the increasingly large cohorts available. Second datasets of 100 participants or more. Hopefully, this will become less of an issue thanks to the increasingly large cohorts available. Second, because of that data reduction step, each application requires the selection o increasingly large cohorts available. Second, because of that data reduction step, each application<br>requires the selection of a number of principal components to be kept. While the optimal parameters<br>are likely to change increasingly large contrained to the multiple components to be kept. While the optimal parameters<br>requires the selection of a number of principal components to be kept. While the optimal parameters<br>are likely to change acr are likely to change across data, this analysis suggests that the gain of a systematic screening over a<br>range of PCs can overcome the cost of additional multiple testing corrections. Third, the proposed<br>approach does not m are many continuity on a start with the system many of and the gain of a system in a system corrections. Third, the proposed approach does not model the compositional aspect of the data per se<sup>50</sup>. However, when the dimens range of the data is large enough, as for the analysis of species, genus or family, this issue<br>becomes negligible (Supplementary Notes and Fig. S7). More importantly, the proposed approach<br>assess variability in the covaria approach does not model the compositional aspect of the data per se<sup>30</sup>. However, when the dimension of the data is large enough, as for the analysis of species, genus or family, this issue becomes negligible (**Supplementa** becomes negligible (**Supplementary Notes and Fig. S7**). More importantly, the proposed approach<br>assess variability in the covariance, and under reasonable assumptions, this variability is independent<br>of the absolute correl becomes negligible (**Supplementary Notes and Fig. 37**). Wore importantly, the proposed approach<br>assess variability in the covariance, and under reasonable assumptions, this variability is independent<br>of the absolute correl of the absolute correlation, so that any remaining bias due to the compositional aspect acts as an offset without impacting our test. Fourth, we demonstrated that covariance can be used for prediction purposes, however, th offset without impacting our test. Fourth, we demonstrated that covariance can be used for prediction purposes, however, the implementation of such predictive model will have to be explored further. As for prediction model of the meral interest. The implementation of such predictive model will have to be explored<br>further. As for prediction model based on relative abundance, some species might not be quantified<br>in the targeted samples for pre Further. As for prediction model based on relative abundance, some species might not be quantified<br>in the targeted samples for prediction. This issue will be exacerbated when working with thousands<br>of covariance terms. One in the targeted samples for prediction. This issue will be exacerbated when working with thousands<br>of covariance terms. One possible solution is to develop sparse predictive models focusing on pairs of<br>taxa that are fairly in the targeted samples of covariance terms. One possible solution is to develop sparse predictive models focusing on pairs of taxa that are fairly common, instead of using the entire covariance matrix. Furthermore, we use taxa that are fairly common, instead of using the entire covariance matrix. Furthermore, we used<br>simple linear predictive models for both abundance and co-abundance. Future work might<br>investigate the use of more complex me taxa that are fairly commonly instead of a bundance and co-abundance. Future work might investigate the use of more complex methods<sup>51</sup> to combine the proposed covariance into prediction models. sinvestigate the use of more complex methods<sup>51</sup> to combine the proposed covariance into prediction models. investigate the use of more complex methods51 to combine the proposed covariance into prediction

Through the characterization of the links between variability in gut microbiome taxa coabundance and both binary and continuous host features when all existing methods are restricted<br>co-abundance and both binary and continuous host features when all existing methods are restricted<br>to *ad hoc* comparisons of Existent. First, the proposed approach and both binary and continuous host features when all existing methods are restricted<br>to *ad hoc* comparisons of inferred networks across a limited number of conditions. Second, our<br> to *ad hoc* comparisons of inferred networks across a limited number of conditions. Second, our framework allows for covariate adjustment, so that the respective effects of correlated factors can be deciphered from one an tramework allows for covariate adjustment, so that the respective effects of correlated factors can<br>be deciphered from one another. Third, our covariance-based approach provides a mean to derive a<br>co-abundance metric at th be deciphered from one another. Third, our covariance-based approach provides a mean to derive a co-abundance metric at the individual level, allowing for a range of secondary analyses, including the development of co-abu co-abundance metric at the individual level, allowing for a range of secondary analyses, including the<br>development of co-abundance-based predictive models. Altogether, the proposed approach open<br>paths for various co-abund co-abundance metric at the individual level, allowing for a range of secondary analyses, including the<br>development of co-abundance-based predictive models. Altogether, the proposed approach open co-abundance-based predictive models. Altogether, the proposed approach open<br>paths for various co-abundance analyses. It is highly complementary to recent efforts to develop<br>experimental design to study co-abundance (e.g. paths for various co-abundance-analyses. It is highly complementary to recent efforts to develop<br>experimental design to study co-abundance (e.g.  $^{52}$ ). It can be used to produce new working<br>hypothesis, and assess statis paths for various co-abundance (e.g. <sup>52</sup>). It can be used to produce new working<br>hypothesis, and assess statistical evidence for effect on co-abundance from both observational and<br>experimental data. experimental design to study co-abundance (e.g. <sup>52</sup>). It can be used to produce new working<br>hypothesis, and assess statistical evidence for effect on co-abundance from both observational and<br>experimental data. hypothesis, and assess statistical evidence for effect on co-abundance for effect on co-abundance from both observations.<br>experimental data. experimental data.

## Methods

*Milieu Interieur gut microbiome data*<br>The Milieu Intérieur Consortium is a population-based cohort initiated in September, 2012<sup>43</sup>. It The Milieu Intérieur Consortium is a population-based cohort initiated in September, 2012<sup>--</sup>. It nprises 1,000 healthy volunteers, all recruited in the suburban Rennes area (Ille-et-Vilaine, tagne, France), with a 1:1 se Bretagne, France), with a 1:1 sex ratio (500 males, 500 females) and an equal distribution across 5 decades (20 to <30 y, 30 to <40 y, 40 to <50 y, 50 to <60 y). The primary objectives of the MI Consortium are to define t Bretagne, France), France, Fr Consortium are to define the naturally occurring variability in a healthy population's immune<br>phenotypes and to characterize genetic, environmental and clinical factors driving this variability. The<br>cohort collected a bro Phenotypes and to characterize genetic, environmental and clinical factors driving this variability. The<br>cohort collected a broad range of variables, including genetic, genomic, and environmental data, on<br>most participants phenotypes and a broad range of variables, including genetic, genomic, and environmental data, on most participants. On their first visit the volunteers were also asked to fill in an extended form about socio-demographic, most participants. On their first visit the volunteers were also asked to fill in an extended form about<br>socio-demographic, lifestyle and family health history, all recorded in an electronic case report form<br>(eCRF). Gut mi socio-demographic, lifestyle and family health history, all recorded in an electronic case report form (eCRF). Gut microbiota composition was obtained from shotgun metagenomics sequencing, and taxonomic levels were reconst (eCRF). Gut microbiota composition was obtained from shotgun metagenomics sequencing, and<br>taxonomic levels were reconstructed by summing the normalized abundances within a branch at a<br>given level (Fig. 2), resulting in a (eCRF). Gut microbiota composition was obtained from shotgun metagenomics sequencing, and<br>taxonomic levels were reconstructed by summing the normalized abundances within a branch at a<br>given level (Fig. 2), resulting in a

## Covariance method

the level (Fig. 2), resulting in a total of 13,446 unique bacterial species. Further description of the<br>data generation are provided in **Supplementary Notes.**<br>Covariance method<br>Variability in the correlation between two s **Covariance method**<br> **Covariance Covariance method**<br>
Variability in the correlation between two standard<br>
through the element-wise product of those outcomes.<br>  $Y_1$  and  $Y_2$  is expressed as  $\rho_{Y_1Y_2} = cov(Y_1, Y_2)/(\sigma_{Y_1}\sigma_{Y_2})$ ,<br>
standardized outcomes an ough the element-wise product of those outcomes. The Pearson correlation coefficient between<br>and  $Y_2$  is expressed as  $\rho_{Y_1Y_2} = cov(Y_1, Y_2)/(\sigma_{Y_1}\sigma_{Y_2})$ , with  $cov(Y_1, Y_2) = \mathbb{E}[Y_1Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2]$ . For<br>ndardized  $Y_1$  and  $Y_2$  is expressed as  $\rho_{Y_1Y_2} = cov(Y_1, Y_2) / (\sigma_{Y_1} \sigma_{Y_2})$ , with  $cov(Y_1, Y_2) = \mathbb{E}[Y_1Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2]$ . For standardized outcomes and a sample size N, it can be re-expressed as the average of the elemen standardized outcomes and a sample size N, it can be re-expressed as the average of the element-<br>vise product across individuals:  $\rho_{Y_1Y_2} = (\sum_{i=1...N} Y_{1i}Y_{2i})/N$ . It follows that the effect of a predictor X<br>on  $cor(Y_1, Y_$ sample size N, it can b<br>  $: v \times v = (\sum_{i=1}^{n} N_i Y_i)^T$ reasonable assumptions, this test is independent of mean and variance effect<sup>42</sup>.

on  $cor(Y_1, Y_2)$  can be tested using a standard least-squares regression framework where *X* is treated<br>as a predictor and the product  $Y_1Y_2$  as the outcome. One can easily demonstrate that, under<br>reasonable assumptions, as a predictor and the product  $Y_1Y_2$  as the outcome. One can easily demonstrate that, under<br>reasonable assumptions, this test is independent of mean and variance effect<sup>42</sup>.<br>Extending the method to more than two outcom reasonable assumptions, this test is independent of mean and variance effect<sup>42</sup>.<br>
Extending the method to more than two outcomes can be done through the following three<br>
steps: i) starting with *K* centered outcomes  $Y_1$ External point of the method outcomes  $Y_1, ..., Y_K$ , all the pairwise products are computed:<br>  $= Y_i Y_j$  for  $i \in [\![1, K]\!]$  and  $i < j$ ; ii) The  $P_{ij}$  products are then mapped to the quantiles of a<br>
rmal distribution using an inver  $P_{ij} = Y_i Y_j$  for  $i \in [\![1, K]\!]$  and  $i < j$ ; ii) The  $P_{ij}$  products are then mapped to the quantiles of a<br>normal distribution using an inverse-rank normal transformation; iii) To reduce the dimension of<br>product matrix, P is t formal distribution using an inverse-rank normal transformation; iii) To reduce the dimension of<br>product matrix, P is then projected in a reduced latent space of dimension  $p \ll \frac{k(k-1)}{2}$  using the<br>Principal Components An product matrix, *P* is then projected in a reduced latent space of dimension  $p \ll \frac{k(k-1)}{2}$  using the<br>Principal Components Analysis transformation:  $PC_r = Q_r = \sum_{i=0}^{K} \sum_{j>i}^{K} \lambda_{ij}^{(r)} Y_i Y_j$ ; iv) The resulting<br>Principal Co product matrix, *P* is then projected in a reduced latent space of dimension  $p \ll \frac{\kappa(\kappa-1)}{2}$  using the<br>Principal Components Analysis transformation:  $PC_r = Q_r = \sum_{i=0}^K \sum_{j>i}^K \lambda_{ij}^{(r)} Y_i Y_j$ ; iv) The resulting<br>Principal Principal Components Analysis transformation:  $PC_r = Q_r = \sum_{i=0}^{n} \sum_{j>i}^{n} \lambda_{ij}^{V}$ <br>Principal Components (PC) are then mapped to the quantiles of a norm<br>inverse-rank transformation, and scaled. This gives, for N considered  $\sum_{i=0}^{R} \sum_{j>i} \lambda_{ij}^{V_i} Y_i Y_j$ ; iv) The resulting<br>of a normal distribution using an<br>isidered individuals, a matrix **Q** of<br>nation of each step is available in<sup>42</sup>.<br>itrix **C**, which can be continuous,<br>or and the covariance  $i=0$   $\sum_{j>i}^{n}$ inverse-rank transformation, and scaled. This gives, for *N* considered individuals, a matrix **Q** of<br>dimension  $N \times p$  that we can use for the test. A detailed explanation of each step is available in<sup>42</sup>.<br>Finally, given a dimension  $N \times p$  that we can use for the test. A detailed explanation of each step is available in<sup>42</sup>.<br>Finally, given a scaled predictor  $X$  and scaled covariates matrix  $C$ , which can be continuous, categorical or binar

dimension  $N \times p$  that we can use for the test. A detailed explanation of each step is available in<sup>-2</sup>.<br>Finally, given a scaled predictor  $X$  and scaled covariates matrix **C**, which can be continuous,<br>categorical or binar Finally, given a standard multivariate model:  $Q \sim X + C$ .<br>
When applied to taxa, we varied the number of principal components used in MANOCCA from<br>
two to one hundred but limited the number of PC analysed for each predictor conducted using a standard multivariate model:  $Q \sim X + C$ .<br>
When applied to taxa, we varied the number of principal components used in MANOCCA from<br>
two to one hundred but limited the number of PC analysed for each predicto When applied to taxa, we varied the number of princit<br>two to one hundred but limited the number of PC an<br>guidelines provided in<sup>42</sup>, and used a stringent multiple tes<br>the various number of PCs considered. Additional detai b to one hundred but limited the number of PC analysed for each predictor based on the delines provided in<sup>42</sup>, and used a stringent multiple testing significance threshold to account for various number of PCs considered. guidelines provided in<sup>42</sup>, and used a stringent multiple testing significance threshold to account for<br>the various number of PCs considered. Additional details are provided in **Supplementary Notes** and<br>**Figures S1-2**.<br>**Co** guidelines provided in<sup>22</sup>, and used a stringent multiple testing significance threshold to account for<br>the various number of PCs considered. Additional details are provided in **Supplementary Notes** and<br>**Figures S1-2**.<br>**Co** 

Figures S1-2.<br> **Contribution of taxa to covariance association signal**<br>
All the steps in the derivation of the statistical test are linear operations, which means that the<br>
contribution of features contributing to the MANO Figures 31-2.<br>Contribution<br>All the ste<br>contribution n<br>|
| **Contribution of taxa to covariance association signal**<br>All the steps in the derivation of the statistical test are linear operations, which means that the All the steps in the derivative is the statistical test are linear operation, make intended the intended test<br>tribution of features contributing to the MANOCCA association signal can be summed. Two types contribution of features contributing to the MANOCCA association signal can be summed. Two types

the covariance contribution assigned to each single taxon. The contribution of a given pair of taxa *i* and *j* to the covariance signal,  $\phi(P_{ij})$ , is defined as the square of the PCA loadings multiplied by the univariate and *j* to the covariance signal,  $\phi(P_{ij})$ , is defined as the square of the PCA loadings multiplied by the univariate association coefficient  $\hat{\beta}^2$  of the corresponding principal components with the considered predict univariate association coefficient  $\hat{\beta}^2$  of the corresponding principal components with the considered predictor:  $\phi(P_{ij}) = \sum_{r=1}^p \hat{\beta}_r^2 (\lambda_{ij}^{(r)})^2$ , where  $\lambda_{ij}^{(r)}$  is the loading of the *ij* pair of taxa for PC univariate association coefficient  $\beta^2$  of the corresponding principal components with the considered<br>predictor:  $\phi(P_{ij}) = \sum_{r=1}^p \hat{\beta}_r^2 (\lambda_{ij}^{(r)})^2$ , where  $\lambda_{ij}^{(r)}$  is the loading of the *ij* pair of taxa for PC *r* . The total number of PCs included in the analysis. The single taxa contribution,  $\psi(Y_i)$ , can be derived<br>by summing its contributions across all pairs:  $\psi(Y_i) = \sum_{j=1}^{K-1} \phi(P_{ij})$ , with  $j \neq i$ , and K is the total<br>number of  $\mu_{ij}$ ) =  $\Delta_{r=1}$ Pr  $\chi_{ij}$ <br>ber of PCs included<br>ts contributions act  $r_{r=1}^{\mu} \beta_r^2$  $\binom{r}{r}$ , where  $\lambda^{(r)}$ 丿<br>::  $Y_i$ ) =  $\sum_{j=1}^{n-1} \phi(P_{ij})$ 

## Environmental association screening using MANOCCA

the total number of pairs.<br> **Environmental association screening using MANOCCA**<br>
We applied MANOCCA to identify environmental factors associated with a change in covariances<br>
between taxa at the family, genus and species l and  $K$  is the total<br>and  $K$  is the total<br>ange in covariances<br>are asked to fill in a<br>le, and vaccination **Environmental association screening using MA**<br>We applied MANOCCA to identify environment<br>between taxa at the family, genus and species level<br>questionnaire of 44 pages, covering multiple panel<br>history. We selected the most or the total set of the total set of the total solutions of the total sphires, lifestyle, and vaccination<br>sphics, lifestyle, and vaccination<br>ading to the selection of 102 Environmental<br>
We applied M<br>
between taxa at t<br>
questionnaire of<br>
history. We sele<br>
environmental fa |
|
|
|
| We applied Manusdon States and Species level. Milieu Interieur volunteers were asked to fill in a stionnaire of 44 pages, covering multiple panels such as demographics, lifestyle, and vaccination cory. We selected the most questionnaire of 44 pages, covering multiple panels such as demographics, lifestyle, and vaccination<br>history. We selected the most relevant panels for the study, leading to the selection of 102<br>environmental factors. Among mistory. We selected the most relevant panels for the study, leading to the selection of 102<br>environmental factors. Among them was included diet information collected as part of the Nutrinet<br>study<sup>53</sup>: the top three factor http://web.arthout.com/selected the most included diet information collected as part of the Nutrinet<br>study<sup>53</sup>: the top three factors from the Nutrinet factors analysis, and the Nutrinet profiles were<br>binarized to yes/no. study<sup>53</sup>: the top three factors from the Nutrinet factors analysis, and the Nutrinet profiles were binarized to yes/no. We filtered out variables with more than half of the sample size in missing values, or a binary predi binary or a binary predictor with frequencies smaller than 5%. For categorical predictors displaying highly skewed distributions, outliers, defined as value three more standard deviation away from the mean were merged wit highly skewed distributions, outliers, defined as value three more standard deviation away from the<br>mean were merged with a lower occurring category. A total of 80 environmental factors remained<br>for analysis. After filteri be a mean were merged with a lower occurring category. A total of 80 environmental factors remained<br>for analysis. After filtering, we ended up with a cohort of 938 individuals with complete shotgun<br>sequencing, age, sex and mean analysis. After filtering, we ended up with a cohort of 938 individuals with complete shotgun<br>sequencing, age, sex and body mass index (BMI) data. For the genus and family levels, we kept taxa<br>abundant in at least 5% For analysis. After the sex and body mass index (BMI) data. For the genus and family levels, we kept taxa abundant in at least 5% of the cohort, leading to a drop from 1,192 genera to 718 genera, and 216 families to 151 fa shundant in at least 5% of the cohort, leading to a drop from 1,192 genera to 718 genera, and 216<br>families to 151 families. At the specie level, to avoid having too many species with regard to the<br>sample size, we set the t Families to 151 families. At the specie level, to avoid having too many species with regard to the<br>sample size, we set the threshold to 40% of the cohort leading to a drop from 3,885 species to 675<br>species.<br>**Comparison wit** 

Families to 151 species.<br>Species.<br>Comparison with Manova and alpha diversity<br>For comparison purposes, we considered three alternative multivariate methods: a standard<br>MANOVA and the alpha diversity, using the Shannon and S species.<br> **Comparison with Manova and alpha diversity**<br>
For comparison purposes, we considered three alternative multivariate methods: a standard<br>
MANOVA and the alpha diversity, using the Shannon and Simpson indexes. The **Compar**<br>For c<br>MANOV,<br>methodo<br>were ma<br>taxa rela しゅうしょう しゅうしゅう しゅうしょう しゅうしゅつ しゅうしゅう しゅうしょう しゅうしょう しゅうしゃ しゅうしゃ いちのとり しゅうしゃ しゅうしゃ しゅうしゃ しゅうしゃ **Comparison with Manova and alpha diversity**<br>For comparison purposes, we considered three alternative multivariate methods: a standard MANOVA and the alpha diversity, using the Shannon and Simpson indexes. The screening<br>methodology was the same as the one used for MANOCCA, though some pre-processing adjustments<br>were made to match the expected assumptions methodology was the same as the one used for MANOCCA, though some pre-processing adjustments<br>were made to match the expected assumptions of each method. The MANOVA was applied to the<br>taxa relative abundances from a given methodology was the same the same prioring of each method. The MANOVA was applied to the taxa relative abundances from a given phylogenic level, which was processed following standards from the literature<sup>54</sup>: proportion taxa relative abundances from a given phylogenic level, which was processed following standards<br>from the literature<sup>54</sup>: proportion followed by arcsin root transformation followed by a scaling. With<br>**Y** the matrix of resu from the literature<sup>54</sup>: proportion followed by arcsin root transformation followed by a scaling. With **Y** the matrix of resulting taxa, *X* the considered predictor, and **C** a matrix of covariates. We applied the Wilk's from the literature<sup>---</sup>: proportion followed by arcsin root transformation followed by a scaling. With **Y** the matrix of resulting taxa, *X* the considered predictor, and **C** a matrix of covariates. We applied the Wilk's size, we could compute a P-value for the statistic : det( $\tilde{Y}^T\tilde{Y} - N\hat{\beta}\hat{\beta}^T$ )/det ( $\tilde{Y}^T\tilde{Y}$ ) ~ $F(p, N-p-1)$ .

the Wilk's lambda test :  $\mathbf{Y} \sim X + \mathbf{C}$ , and in more details with  $\tilde{\mathbf{Y}} = \mathbf{Y} - \beta_c C$  the residual matrix after adjustment from the covariates,  $\hat{\beta} = (X^T X)^{-1} (X^T \tilde{Y})$  the regression coefficient and N the samp adjustment from the covariates,  $\beta = (X^T X)^{-1} (X^T Y)$  the regression coefficient and  $N$  the sample<br>size, we could compute a *P*-value for the statistic : det $(\tilde{Y}^T \tilde{Y} - N \beta \beta^T)/\det (\tilde{Y}^T \tilde{Y}) \sim F(p, N - p - 1)$ .<br>For the a size, we could compute a *P*-value for the statistic : det( $Y'Y - N\beta\beta'$ )/det ( $Y'Y$ ) ~ $F(p, N-p-1)$ .<br>
For the alpha diversity indexes, the raw abundances were used to the corresponding metric :<br>  $\alpha_{Shannon} = \sum_{i=1}^{N} \frac{x_i}{\sum_{j=1$  $F_{\text{anmon}} = \sum_{i=1}^{N} \frac{x_i}{\sum_{j=1}^{N} x_j} \log \left( \frac{x_i}{\sum_{j=1}^{N} x_j} \right)$  and  $\alpha_{Simpson} = 1 - \sum_{i=1}^{N} \frac{x_i(x_i-1)}{(\sum_{j=1}^{N} x_j)((\sum_{j=1}^{N} x_j)-1)}$ . The resulting  $\alpha$  is tested in a standard univariate linear regression adjusted for the  $\alpha_{Shannon} = \sum_{i=1}^{N} \frac{x_i}{\sum_{j=1}^{N}}$ was tested in a standa  $\sum_{j=1}^{n} x_j$ l<sup>umv</sup>  $\frac{1}{\sqrt{N}}$  $\sum_{j=1}^{n} x_j$  $\sum_{i=1}^{N} \frac{x_i}{\sum_{j=1}^{N} x_j} \log \left( \frac{x_i}{\sum_{j=1}^{N} x_j} \right)$  and  $\alpha_{Simpson} = 1 - \sum_{i=1}^{N} \frac{x_i(x_i - x_i)}{(\sum_{j=1}^{N} x_j)(\sum_{j=1}^{N} x_j)}$ <br>a standard univariate linear regression adjusted for the covaria<br>s assessed using a Wald test to th  $(\sum_{j=1}^{N} x_j)((\sum_{j=1}^{N} x_j)-1)$  $i'$  . The resulting = was tested in a standard univariate linear regression adjusted for the covariates:  $\alpha \sim \delta_X X + \delta_C C$ . The effect of X was assessed using a Wald test to the  $\hat{\delta}_X$ . effect of  $\lambda$  was assessed using a wald test to the  $\theta$  $\overline{a}$ 

**Deriving the covariance network**<br>Networks of variation in the covariance were built using the top 1,000 co-abunding pairs derived The MANOCCA test. In representing the network, we included three parameters: the total<br>mber of connections (qualified through the node size), the actual pairwise taxa connection (edges<br>the graph), and the direction of hos number of connections (qualified through the node size), the actual pairwise taxa connection (edges<br>in the graph), and the direction of host factor effect on the covariance (decrease or increase of co-<br>abundance). For eac number of connections (qualities in the node of contractions (decrease or increase of co-<br>abundance). For each pair  $Y_i Y_j$  adjusted for covariates  $C : \widetilde{Y_i Y_j} = Y_i Y_j - (C^T C)^{-1} C^T Y_i Y_j C$ , the<br>direction of effect was derived abundance). For each pair  $Y_i Y_j$  adjusted for covariates  $C : Y_i Y_j = Y_i Y_j - C$ <br>direction of effect was derived using the sign of the regression coefficient f<br> $\beta_{ij} = (X^T X)^{-1} X Y_i Y_j$ . For shared pairs with mixed direction of effe  $C^T C)^{-1} C^T Y_i Y_j C$ , the<br>or the predictor X:<br>was colored in black.<br>the association with<br>aap from *matplotlib*<br>al predictor colours. direction of effect was derived using the sign of the regression coefficient for the predictor  $X : \beta_{ij} = (X^T X)^{-1} X Y_i Y_j$ . For shared pairs with mixed direction of effect, the edge was colored in black.<br>To facilitate the re  $\beta_{ij} = (X^T X)^{-1} X Y_i Y_j$ . For shared pairs with mixed direction of effect, the edge was colored in black.<br>To facilitate the reading of the network, we coloured the node conditional on the association with<br>each of the four p

To facilitate the four predictors of interest, or shared among them. The 'viridis' cmap from *matplotlib*<br>was used as colour scheme, with each shared taxa being a combination of original predictor colours.<br>**Comparison with** Each of the four predictors of interest, of shared among them. The Vindis' emap from *interprotity*<br>was used as colour scheme, with each shared taxa being a combination of original predictor colours.<br>**Comparison with othe Comparison with other network-based approaches**<br>For comparison purposes, we derived a permutation-based network inference approach fo<br>binary predictor, which we used to validate the MANOCCA network. In brief, we derive t しょくにゅう For comparison purposes, we derived a permutation-based network inference approach for ary predictor, which we used to validate the MANOCCA network. In brief, we derive the pairwise<br>variance matrix for each categorical value, and then derive the empirical distribution of the<br>relation under the null by simul covariance matrix for each categorical value, and then derive the empirical distribution of the<br>correlation under the null by simulating  $N_{permutations}$  covariances after shuffling the abundances of<br>a bacterial taxa for each indi correlation under the null by simulating  $N_{permutations}$  covariances after shuffling the abundances of<br>a bacterial taxa for each individual. Using a fixed detection threshold, we can then select pairs of taxa<br>with extreme covaria a bacterial taxa for each individual. Using a fixed detection threshold, we can then select pairs of taxa<br>with extreme covariances. Since we are interested in variability of the covariance, we only keep the<br>pairs uniquely a bacterial taxa for existing a finite state in variability of the covariance, we only keep the<br>pairs uniquely detected across all values of the given predictor. When applied to compare results<br>from the sex and smoking an pairs uniquely detected across all values of the given predictor. When applied to compare results<br>from the sex and smoking analyses, we ran 100,000 permutations, and retrieved the unique pairs<br>detected in either group (wom pairs uniquely accretion analyses, we ran 100,000 permutations, and retrieved the unique pairs<br>detected in either group (women vs men, and non-smoker vs ever smoke). We also ran the SparCC<sup>49</sup><br>correlation analysis, as thi detected in either group (women vs men, and non-smoker vs ever smoke). We also ran the SparCC<sup>49</sup><br>correlation analysis, as this approach is commonly used and performed relatively well in a review of<br>existing approaches<sup>41</sup> existing approaches<sup>41</sup>. We ran SparCC using the recommended parameters, deriving the *P*-values<br>using 1,000 permutations, on both the simulated data (Fig. S5) and the real data (Fig. S6). Further<br>descriptions of both appr

existing approaches<sup>-\*</sup>. We ran SparCC using the recommended parameters, deriving the *P*-values<br>using 1,000 permutations, on both the simulated data (**Fig. S5**) and the real data (**Fig. S6**). Further<br>descriptions of both Using 1,000 permutations, on both the simulated data (Fig. 35) and the real data (Fig. 36). Further<br>descriptions of both approaches are provided in the **Supplementary Notes**.<br>Using co-abundance for prediction purposes<br>We Using co-abundance for prediction purposes<br>We assessed the performances of a predictive model based on cover-<br>implementation of a predictive model follows the standard used for multiv<br>given outcome A to be predicted, the ;<br>)<br>e<br>( Using co-abundance for prediction purposes<br>We assessed the performances of a predictive model based on covariance across taxa. The blementation of a predictive model follows the standard used for multivariate linear model. For a<br>en outcome A to be predicted, the estimated coefficients between A and  $PC_{i=1...L}^{(train)}$  obtained in a<br>ning dataset from MANOC given outcome A to be predicted, the estimated coefficients between A and  $PC_{i=1...L}^{(train)}$  obtained in a<br>training dataset from MANOCCA,  $\hat{\beta} = (\hat{\beta}_1 ... \hat{\beta}_L)$  are projected on the principal component from an<br>independent tes given outcome A to be predicted, the estimated coefficients between A and  $PC_{i=1...L}^{C_{i}}$  obtained in a<br>training dataset from MANOCCA,  $\hat{\beta} = (\hat{\beta}_1 ... \hat{\beta}_L)$  are projected on the principal component from an<br>independent te  $\frac{1}{1}$  (cruciii) training dataset from MANOCCA,  $\mathbf{p} = (p_1 ... p_L)$  are projected on the principal component from an<br>independent test dataset and summed up to form a predictive score  $S = \sum_{i=1}^{L} \hat{\beta}_i P C_i^{(test)}$ . Note that<br>the dimensionality  $\frac{1}{p}$  to form to the proof.<br>Digital to the proof of the proof independent test dataset and summed up to form a predictive score  $3 - \sum_{i=1}^n p_i v_i$ . Wote that<br>the dimensionality of the covariance data and the principal component analysis (PCA) step make the<br>implementation slightly mor  $_{i=1}^{\nu}\,\beta_i\,PC_i$ <br>ysis (PCA<br>nts deriv  $\mu_{i=1}$   $\beta_i$   $PC_i$ <sup>(cese)</sup> implementation slightly more complex. In particular, the principal components derived on the same<br>variables for two independent samples might not always match, with structure in the data being<br>capture by different compone implementation slightly more complement particular, the principal complete complete in the data being<br>capture by different components. To avoid this issue, PCA is not applied in the test data. Instead,<br> $PC_i^{(test)}$  are derive variables for two independent samples might not always minimal properties in the test data. Instead,<br>  $PC_i^{(test)}$  are derived by projecting out the loadings from the training sets:  $PC_i^{(test)} = \sum_{i=1}^{L} \lambda_i^{(i)} X_j$ ,<br>
where  $\lambda_j^{($  $PC_i^{(test)}$  are derived by projecting out the loadings from the training sets:  $PC_i^{(test)} = \sum_i^L \lambda_j^{(i)} X_j$ ,<br>where  $\lambda_j^{(i)}$  is the loading of variable  $PC_i^{(train)}$  for product of taxa *i* obtained in the train data. It also<br>implies  $PC_i^{(test)}$  are derived by projecting out the loadings from the training sets:  $PC_i^{(test)} = \sum_i A_j^{(i)} X_j$ <br>where  $\lambda_j^{(i)}$  is the loading of variable  $PC_i^{(train)}$  for product of taxa *i* obtained in the train data. It also<br>implies that  $\zeta^{(test)}$  are derived by projecting out the loadings from the training sets:  $PC_i^{(test)} = \sum_i^L \lambda_i^{(i)}$  $\cdot^{\prime}$  is the loading of variable  $PC_{\cdot}^{\prime\prime\prime\prime\mu\prime\mu\prime}$ 

, り f \_ p l / e Implies that the test dataset should have the same dimension (i.e. approximately the same list of taxa) as the train dataset.<br>We applied this approach for the prediction of age, BMI, smoking and sex using taxa from the thr is the loading of variable  $PC_i$ <br>at the test dataset should have train dataset.<br>blied this approach for the p<br>set taxa levels (species, genut<br>other factors as covariate. Fo<br>wo independent sets: a trainining samples. We mea implies that the test dataset.<br>
We applied this approach for the prediction of age, BMI, smoking and sex using taxa from the<br>
three lowest taxa levels (species, genus and family), using a 30-fold cross-validation, and with we applied this appro<br>three lowest taxa levels<br>including other factors as<br>split into two independen<br>10% remaining samples. We approach the prediction of the prediction, and without<br>be lowest taxa levels (species, genus and family), using a 30-fold cross-validation, and without<br>luding other factors as covariate. For each of the 30 cross-validat three lowest taxa is the lowest taxa in the lowest taxa including other factors as covariate. For each of the 30 cross-validation, the dataset was randomly split into two independent sets: a training set including 90% of t including other factors as constants as covariant of the 30 cross-validation, we can be considered the split into two independent sets: a training set including 90% of the data and a test set including the 10% remaining sa split into the independent sets: a training set including 90% of the later and a test set including squared-<br>10% remaining samples. We measured the accuracy of the predictive model using squared-

predictive power of binary outcome. It equals the probability of correctly classifying a random<br>sample from the test data.<br>**Acknowledgment**<br>This research was supported by the Agence Nationale pour la Recherche (ANR-20-CE15  $(S, A^{(test)})$  , and using the area under the<br>
The AUC is a common metric to quantify the<br>
robability of correctly classifying a random<br>
get pour la Recherche (ANR-20-CE15-0012-01). correlation for continuous outcomes, derived as  $cor(S, A^{(test)})^2$ , and using the area under the predictive power of binary outcome. It equals the probability of correctly classifying a random<br>sample from the test data.<br>**Acknowledgment**<br>This research was supported by the Agence Nationale pour la Recherche (ANR-20-CE15

## Acknowledgment

prediction the test data.<br> **Acknowledgment**<br>
This research was supported by the Agence Nationale pour la Recherche (ANR-20-CE15-0012-01).<br>
This work has been conducted as part of the INCEPTION program (Investissement d'Ave **Acknowledgment**<br>This research was support<br>This work has been condu<br>ANR-16-CONV-0005). The I<br>pour la Recherche (ANR-10 |<br>s<br>s<br>F<br>J This work has been conducted as part of the INCEPTION program (Investissement d'Avenir grant<br>ANR-16-CONV-0005). The Milieu Interieur consortium was also supported by the Agence Nationale<br>pour la Recherche (ANR-10-LABX-69-0 This work has been conducted as part of the Interiormal program (interferent directions) and<br>ANR-16-CONV-0005). The Milieu Interieur consortium was also supported by the Agence Nationale<br>pour la Recherche (ANR-10-LABX-69-0

## Code availability

pour la Recherche (ANR-10-LABX-69-01).<br> **Code availability**<br>
All code is available in Python and R at: <u>https://gitlab.pasteur.fr/statistical-genetics/manocca</u><br>
The complete network with full annotation is available in htm **Code availability<br>All code is available in Python and R at: <u>ht</u><br>The complete network with full annotatic<br>genetics/manocca**  $\begin{array}{c} \begin{array}{c} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{array} \end{array}$ The complete network with full annotation is available in html at: https://gitlab.pasteur.fr/statistical-genetics/manocca<br> **References**<br>
1. Integrative, H.M.P.R.N.C. The Integrative Human Microbiome Project. *Nature* 569,

## References

- 
- 
- 
- 4. Saraswati, S. & Sitaraman, R. Aging and the human gut microbiota-from correlation to causality. Front **References<br>1.** Integrative,<br>2. Xu, C., Zhu,<br>3. O'Toole, P.V<br>4. Saraswati, S<br>*Microbiol* 5,  $\begin{array}{c} \n\begin{array}{ccc} \n\text{1} & \text{1} & \text{1$ 1. Integrative, H.M.P. IN.N.C. The Integrative Human Microbiolie Project: Nature 369, 641-648 (2019).<br>
2. Xu, C., Zhu, H. & Qiu, P. Aging progression of human gut microbiota. *BMC Microbiol* 19, 236 (2019).<br>
3. O'Toole, P. 2. Xu, C., Zhu, H. & Qiu, P. Aging progression of human gut microbiota. BMC Microbiol 19, 236 (2019).<br>
3. O'Toole, P.W. & Jeffery, I.B. Gut microbiota and aging. Science **350**, 1214-5 (2015).<br>
4. Saraswati, S. & Sitaraman,
- *Microbiol 5, 704 (2014).*<br>Zhang, X. *et al.* Sex- an<br>populations of different<br>Kim, Y.S., Unno, T., Kim,<br>48-60 (2020).<br>Min, Y. *et al.* Sex-specifi<br>2408 (2019).
- Microbiol 5, 764 (2014).<br>
Microbiol 5, 764 (2014).<br>
2hang, X. *et al.* Sex- and age-related trajectories of the adult human gut microbiota-shared across<br>
populations of different ethnicities. *Nature Aging* 1, 87-100 (2021 Eriang, X. et al. Sex- and age-related trajectories of the adult human gut microbiota shared across<br>populations of different ethnicities. Nature Aging 1, 87-100 (2021).<br>6. Kim, Y.S., Unno, T., Kim, B.Y. & Park, M.S. Sex Di populations of different ethnicities. *Nature Aging 1, 87-100 (2021)*.<br>Kim, Y.S., Unno, T., Kim, B.Y. & Park, M.S. Sex Differences in Gut M<br>48-60 (2020).<br>Min, Y. *et al.* Sex-specific association between gut microbiome and
- 6. Kim, Y.S., Omio, T., Kim, B.Y. & Fank, M.S. Sex Differences in Gut Microbiota. World J Mens Health 38,<br>48-60 (2020).<br>7. Min, Y. *et al.* Sex-specific association between gut microbiome and fat distribution. *Nat Commun* Min, Y. *et al.*<br>2408 (2019).<br>Dominianni, *(*<br>microbiome. *F*<br>Zahavi, L. *et (*<br>2785-2792 (20
- 
- 7. Min, Y. et al. Sex-specific association between gut microbionie and fat distribution. Nat Commun 10,<br>2408 (2019).<br>8. Dominianni, C. *et al.* Sex, body mass index, and dietary fiber intake influence the human gut<br>microbi 2008<br>Dominianni,<br>microbiome.<br>Zahavi, L. *et*<br>2785-2792 (2<br>Gui, X., Yang 8. Bommanni, C. et al. Sex, body mass malex, and dietary meet make innefice the human gut<br>microbiome. PLoS One 10, e0124599 (2015).<br>2. Zahavi, L. et al. Bacterial SNPs in the human gut microbiome associate with host BMI. N microbiome. PLos One 10, eor244339 (2013).<br>Zahavi, L. *et al.* Bacterial SNPs in the humai<br>2785-2792 (2023).<br>Gui, X., Yang, Z. & Li, M.D. Effect of Cigaret<br>*Physiol* 12, 673341 (2021).<br>Fan, J. *et al.* Cross-talks between
- 2785-2792 (2023).<br>
9. Zahavi, L. et al. Bacterial SNPs in the human gut microbiome associate with host BMI. Nat Med 29,<br>
9. Gui, X., Yang, Z. & Li, M.D. Effect of Cigarette Smoke on Gut Microbiota: State of Knowledge. Fron
- 
- Gui, X., Yang, Z. &<br>*Physiol* **12**, 673341<br>Fan, J. *et al.* Cros<br>randomization stuc<br>Antinozzi, M. *et a.*<br>Review. *Biomedicir* 10. Gui, X., Yang, Z. & Li, W.D. Lifect of Cigarette Shoke on Gut Microbiota. State of Knowledge. Tront<br>
Physiol 12, 673341 (2021).<br>
11. Fan, J. et al. Cross-talks between gut microbiota and tobacco smoking: a two-sample M Fan, J. *et al.* Cross-talks l<br>Fan, J. *et al.* Cross-talks l<br>randomization study. *BMC*<br>Antinozzi, M. *et al.* Cigare<br>Review. *Biomedicines* **10**(20<br>Xu, Z. & Knight, R. Dietar<br>(2015). 11. Fan, J. et al. Cross-talks between gut microbiota and tobacco smoking. a two-sample Mendelian<br>
randomization study. *BMC Med* **21**, 163 (2023).<br>
12. Antinozzi, M. *et al.* Cigarette Smoking and Human Gut Microbiota in randomization study. *BMC Med 21*, 103 (2023).<br>Antinozzi, M. *et al.* Cigarette Smoking and Hu<br>Review. *Biomedicines* **10**(2022).<br>Xu, Z. & Knight, R. Dietary effects on human<br>(2015).<br>Stege, P.B. *et al.* Impact of long-ter
- 12. Antinozzi, M. et al. Cigarette Smoking and Human Gut Microbiota in Healthy Adults: A Systematic<br>
Review. Biomedicines 10(2022).<br>
13. Xu, Z. & Knight, R. Dietary effects on human gut microbiome diversity. *Br J Nutr* 11 Review. *Biomedicines* 10(2022).<br>Ru, Z. & Knight, R. Dietary efferences.<br>Stege, P.B. *et al.* Impact of Icoppulation. *Sci Rep* 12, 1892 (20<br>Rinninella, E. *et al.* The role of d<br>**62-63**, 101828 (2023).
- Stege,<br>Stege,<br>populat<br>Rinnine<br>**62-63**, 1<br>Perler,<br>Diet and
- 13. Xu, Z. & Kinght, R. Dietary effects on human gut microbionie diversity. Br J Nutr 113 Supply, S1-5<br>
14. Stege, P.B. *et al.* Impact of long-term dietary habits on the human gut resistome in the Dutch<br>
population. Sci R 14. Stege, P.B. et al. Impact of long-term dietary habits on the human gut resistome in the Dutch<br>population. Sci Rep 12, 1892 (2022).<br>**15.** Rinninella, E. et al. The role of diet in shaping human gut microbiota. Best Prac population. Scriber 12, 1892 (2022).<br>Rinninella, E. *et al.* The role of diet in<br>**62-63**, 101828 (2023).<br>Perler, B.K., Friedman, E.S. & Wu, G<br>Diet and Human Health. Annu Rev Pl<br>Qin, Y. *et al.* Combined effects of hos
- 15. Rinninella, E. *et al.* The role of diet in shaping human gut microbiota. *Best Pract Res Clin Gastroenterol*<br> **62-63**, 101828 (2023).<br>
Perler, B.K., Friedman, E.S. & Wu, G.D. The Role of the Gut Microbiota in the Rela **62-63, 101628 (2023).**<br>Perler, B.K., Friedman,<br>Diet and Human Healt<br>Qin, Y. *et al.* Combined<br>in a single population c<br>Bonder, M.J. *et al*. The
- Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease Diet and Human Health. Annu Rev Physiol 85, 443-468 (2023).<br>Qin, Y. *et al.* Combined effects of host genetics and diet on hu<br>in a single population cohort. *Nat Genet* 54, 134-142 (2022).<br>Bonder, M.J. *et al.* The effect 17. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease<br>in a single population cohort. *Nat Genet* 54, 134-142 (2022).<br>18. Bonder, M.J. *et al.* The effect of host genetic
- in a single population cohort. Nat Genet 34, 134-142 (2022).<br>Bonder, M.J. *et al.* The effect of host genetics on the gut micr 18. Bonder, M.J. et al. The effect of host genetics on the gut microbiome. Nat Genet 48, 1407-1412 (2016).

- 
- 19. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy<br>
19. Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing human gut<br>
19. Turp
- 
- cohort. Nat Genet 48, 1413-1417 (2016).<br>Kurilshikov, A. *et al.* Large-scale associ<br>microbiome composition. Nat Genet 53, 1<br>Lopera-Maya, E.A. *et al.* Effect of host &<br>Dutch Microbiome Project. Nat Genet 54,<br>Ruhlemann, M.C 20. Kurilshikov, A. et al. Large-scale association analyses identify host factors inhactenig human gut<br>microbiome composition. Nat Genet 53, 156-165 (2021).<br>21. Lopera-Maya, E.A. et al. Effect of host genetics on the gut m microbiome composition. Nat Genet 33, 156-165 (2021).<br>Lopera-Maya, E.A. *et al.* Effect of host genetics on the<br>Dutch Microbiome Project. Nat Genet 54, 143-151 (2022)<br>Ruhlemann, M.C. *et al.* Genome-wide association sti<br>in 21. Lopera-Maya, E.A. *et al.* Effect of host genetics on the gut microbiome in 7,738 participants of the<br>
22. Ruhlemann, M.C. *et al.* Genome-wide association study in 8,956 German individuals identifies<br>
influence of ABO Butch Microbiome Project. Nut Genet 34, 143-131 (2022).<br>Ruhlemann, M.C. *et al.* Genome-wide association student<br>Influence of ABO histo-blood groups on gut microbiome. N<br>Hou, K. *et al.* Microbiota in health and diseases.
- 
- 22. Ruhlemann, W.C. et *ul.* Genome-wide association study in 8,956 German multiplates influence of ABO histo-blood groups on gut microbiome. *Nature Genetics* 53, 147-+ (2021).<br>23. Hou, K. *et al.* Microbiota in health an influence of ABO histo-blood groups on gut microbiome. Nature demetites 33, 147-1 (2021).<br>Hou, K. *et al.* Microbiota in health and diseases. *Signal Transduct Target Ther 7*, 135 (2022).<br>Daniel, N., Lecuyer, E. & Chassain
- 
- 23. Hou, K. et al. Microbiota in health and diseases. Signal Transduct Target Ther 7, 135 (2022).<br>24. Daniel, N., Lecuyer, E. & Chassaing, B. Host/microbiota interactions in health and disease<br>mucosal microbiology! *Mucosa* mucosal microbiology! *Mucosal Immunol* **14**, 1006-1016 (2021).<br>25. Vijay, A. & Valdes, A.M. Role of the gut microbiome in chronic diseases: a narrative review. *Eur J Clin<br>26. Awany, D. <i>et al.* Host and Microbiome Genome Mucosal Inicrobiology! Mucosal Immanol 14, 1000-1010 (2021).<br>Vijay, A. & Valdes, A.M. Role of the gut microbiome in chronic (<br>Nutr 76, 489-501 (2022).<br>Awany, D. *et al.* Host and Microbiome Genome-Wide Ass<br>Challenges. Fron 25. Vijay, A. & Valdes, A.M. Role of the gut microbiome in chronic diseases: a narrative review. Eur J Clin<br>26. Awany, D. *et al.* Host and Microbiome Genome-Wide Association Studies: Current State and<br>27. Sanna, S., Kuril Nutr 76, 489-501 (2022).<br>Awany, D. *et al.* Host<br>Challenges. Front Genet 9<br>Sanna, S., Kurilshikov, A.,<br>studying effects of host g<br>Davenport, E.R. *et al.* Ge<br>e0140301 (2015).
- 26. Awany, D. et al. Host and Microbiome Genome-Wide Association Studies. Current State and<br>Challenges. Front Genet 9, 637 (2018).<br>27. Sanna, S., Kurilshikov, A., van der Graaf, A., Fu, J. & Zhernakova, A. Challenges and f
- 
- Challenges. Front Genet 9, 637 (2016).<br>Sanna, S., Kurilshikov, A., van der Graat<br>studying effects of host genetics on the<br>Davenport, E.R. *et al.* Genome-Wide A<br>e0140301 (2015).<br>Degraders, Metabolites, and Predicted 27. Studying effects of host genetics on the gut microbiome. *Nat Genet* 54, 100-106 (2022).<br>28. Davenport, E.R. *et al.* Genome-Wide Association Studies of the Human Gut Microbiota. *PLoS One* 10,<br>29. Raimondi, S. *et al.* Studying effects of host genetics on the gut microbiome. Nat Genet 54, 100-100 (2022).<br>Davenport, E.R. *et al.* Genome-Wide Association Studies of the Human Gut Microbiota<br>e0140301 (2015).<br>Raimondi, S. *et al.* Multivariat
- 28. Bavenport, E.R. et al. Genome-Wide Association Stadies of the Human Gut Microbiota. 7 209 One 10,<br>
29. Raimondi, S. *et al.* Multivariate Analysis in Microbiome Description: Correlation of Human Gut Protein<br>
Degraders, e0140301 (2015).<br>
Raimondi, S. *et al.* Multivariate Analysis in Microbiome Description: Correlation of Human Gut Protein<br>
Degraders, Metabolites, and Predicted Metabolic Functions. *Front Microbiol* **12**, 723479 (2021).<br> 29. Raimondi, S. et al. Multivariate Analysis in Microbiolne Description: Correlation of Human Gut Protein<br>Degraders, Metabolites, and Predicted Metabolic Functions. *Front Microbiol* 12, 723479 (2021).<br>30. Le Cao, K.A. *e* Degraders, Metabolites, and Predicted Metabolic Functions. Front Microbiol 12, 723479 (2021).<br>Le Cao, K.A. *et al.* MixMC: A Multivariate Statistical Framework to Gain Insight into Mi<br>Communities. PLoS One 11, e0160169 (20
- 
- 30. Le Cao, K.A. et al. MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial<br>Communities. PLoS One 11, e0160169 (2016).<br>31. Banerjee, K., Chen, J. & Zhan, X. Adaptive and powerful microbiome multivari Communities. PLOS One 11, e0160169 (2016).<br>Banerjee, K., Chen, J. & Zhan, X. Adaptive and<br>via feature selection. NAR Genom Bioinform 4,<br>Faust, K. & Raes, J. Microbial interactions: fr<br>(2012).<br>Kumar, M., Ji, B., Zengler, K. 32. Baner, R. & Raes, J. Microbial interactions: from networks to models. Nat Rev Microbiol 10, 538-50<br>32. Baust, K. & Raes, J. Microbial interactions: from networks to models. Nat Rev Microbiol 10, 538-50<br>33. Kumar, M., J
- 
- via reature selection. NAR Genom Biomjorm 4, Iqab120 (2022).<br>Faust, K. & Raes, J. Microbial interactions: from networks to<br>(2012).<br>Kumar, M., Ji, B., Zengler, K. & Nielsen, J. Modelling approac<br>Microbiol 4, 1253-1267 (2019 32. Faust, K. & Raes, J. Microbial interactions: from networks to models. Nat Rev Microbiol 10, 338-50<br>(2012).<br>33. Kumar, M., Ji, B., Zengler, K. & Nielsen, J. Modelling approaches for studying the microbiome. Nat<br>34. Schm ,<br>Kumar,<br>*Microbi*<br>Schmidt<br>**172**, 11<br>Chen, L<br>and obe 33. Kumar, M., 31, B., Zengler, K. & Nielsen, J. Modelling approaches for studying the microbiome. Nat<br>
Microbiol 4, 1253-1267 (2019).<br>
34. Schmidt, T.S.B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Microbiol 4, 1253-1267 (2015).<br>Schmidt, T.S.B., Raes, J. & Bork<br>172, 1198-1215 (2018).<br>Chen, L. *et al.* Gut microbial co<br>and obesity. Nat Commun 11, 4<br>Berry, D. & Widder, S. Deciph<br>occurrence networks. Front Mic
- 34. Schmidt, T.S.B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. Cell<br> **172**, 1198-1215 (2018).<br>
Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowe 172, 1198-1215 (2018).<br>Chen, L. *et al.* Gut micr<br>and obesity. *Nat Commi*<br>Berry, D. & Widder, S.<br>occurrence networks. Fl<br>Banerjee, S., Schlaeppi,<br>and functioning. *Nat Re*
- 33. Chen, L. et al. Gut microbial co-abundance networks show specificity in inflammatory bowel disease<br>and obesity. Nat Commun 11, 4018 (2020).<br>36. Berry, D. & Widder, S. Deciphering microbial interactions and detecting ke and obesity. Nat Commun 11, 4018 (2020).<br>Berry, D. & Widder, S. Deciphering micro<br>occurrence networks. *Front Microbiol* 5, 21!<br>Banerjee, S., Schlaeppi, K. & van der Heijde<br>and functioning. *Nat Rev Microbiol* 16, 567-<br>Wu,
- Banerjee, S., Schlaeppi, K. & van der Heijden, M.G.A.<br>Banerjee, S., Schlaeppi, K. & van der Heijden, M.G.A.<br>and functioning. *Nat Rev Microbiol* **16**, 567-576 (2011<br>Wu, G., Zhao, N., Zhang, C., Lam, Y.Y. & Zhao,<br>microbiome
- 38. Mu, G., Zhao, N., Zhang, C., Lam, Y.Y. & Zhao, L. Guild-based analysis for understanding gut<br>38. Wu, G., Zhao, N., Zhang, C., Lam, Y.Y. & Zhao, L. Guild-based analysis for understanding gut<br>39. Jackson, M.A. *et al.* D and functioning. Nut Rev Microbiol 16, 567-576 (2016).<br>Wu, G., Zhao, N., Zhang, C., Lam, Y.Y. & Zhao, L.<br>microbiome in human health and diseases. *Genome Me*<br>Jackson, M.A. *et al.* Detection of stable community si<br>networks
- 39. Jackson, M.A. *et al.* Detection of stable community structures within gut microbiota co-occurrence
- microbiome in human health and diseases. *Genome Med* 13, 22 (2021).<br>39. Jackson, M.A. *et al.* Detection of stable community structures within gut microbiota co-occurrence<br>networks from different human populations. *Peerl*
- microbiome in human health and diseases. Genome Med 13, 22 (2021).<br>Jackson, M.A. *et al.* Detection of stable community structures within<br>networks from different human populations. *Peerl* **6**, e4303 (2018).<br>Bayer, G., Gan 39. Jackson, M.A. *et al.* Detection of stable community structures within gut microbiota co-occurrence<br>networks from different human populations. *Peerl* **6**, e4303 (2018).<br>8ayer, G., Ganobis, C.M., Allen-Vercoe, E. & Phi networks from different human populations. Peer, **G**, e4303 (2016).<br>Bayer, G., Ganobis, C.M., Allen-Vercoe, E. & Philpott, D.J. Def<br>promising tools to understand and combat disease. *Microbes Infect*<br>Weiss, S. *et al.* Cor promising tools to understand and combat disease. *Microbes inject 23*, 104816 (2021).<br>Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in<br>precision. *ISME J* **10**, 1669-81 (2016).<br>Boe
- 
- 43. Thomas, S. *et al.* The Milieu Interieur study an integrative approach for study of human<br>immunological variance. *Clin Immunol* 157, 277-93 (2015).<br>44. Lozupone, C.A. *et al.* Meta-analyses of studies of the human m 41. Weiss, S. et al. Correlation detection strategies in iniciobial data sets vary widely in sensitivity and<br>precision. ISME J 10, 1669-81 (2016).<br>42. Boetto, C. et al. A multivariate outcome test of covariance. *bioRxiv*, precision. *ISME* J 10, 1009-81 (2016).<br>Boetto, C. *et al.* A multivariate outcor<br>Thomas, S. *et al.* The Milieu Inte<br>immunological variance. *Clin Immuno*<br>Lozupone, C.A. *et al.* Meta-analyses<br>(2013).<br>Tam, A. *et al.* Eff 42. Boetto, C. et al. A multivariate outcome test of covariance. bionary, 2023.09.20.390234 (2023).<br>43. Thomas, S. *et al.* The Milieu Interieur study - an integrative approach for study of<br>immunological variance. *Clin Im*
- 43. Thomas, S. et al. The Willieu Interieur study an integrative approach for study of human<br>immunological variance. Clin Immunol 157, 277-93 (2015).<br>44. Lozupone, C.A. *et al.* Meta-analyses of studies of the human micr Infinanciogical variance. C*immuniol* 157, 277-93 (2015).<br>Lozupone, C.A. *et al.* Meta-analyses of studies of the hu<br>(2013).<br>Tam, A. *et al.* Effects of sex and chronic cigarette smoke<br>*PLoS One* **15**, e0230932 (2020).<br>Ant
- 44. Lozupone, C.A. et al. Meta-analyses of studies of the human microbiotal denome Res 23, 1704-14<br>
45. Tam, A. et al. Effects of sex and chronic cigarette smoke exposure on the mouse cecal microbiome.<br> *PLoS One* **15**, e0 (2022).<br>Tam, A.<br>*PLoS Or*<br>Antone<br>populat<br>Meehar<br>family c
- PLOS One 13, e0230932 (2020).<br>Antonello, G. *et al.* Smoking a<br>population. Sci Rep 13, 18904 (2<br>Meehan, C.J. & Beiko, R.G. A pf<br>family of digestive tract-associa<sup>.</sup><br>Vacca, M. *et al.* The Controversi
- 45. Tam, A. et al. Effects of sex and chronic eigenctic smoke exposure on the mouse cecar microbiome.<br>
46. Antonello, G. *et al.* Smoking and salivary microbiota: a cross-sectional analysis of an Italian alpine<br>
147. Meeha 46. Antonello, G. et al. Shoking and salivary includiotel. a cross-sectional analysis of an Italian alphie<br>population. Sci Rep 13, 18904 (2023).<br>47. Meehan, C.J. & Beiko, R.G. A phylogenomic view of ecological specializati population. Scritep 13, 18904 (2023).<br>Meehan, C.J. & Beiko, R.G. A phyloge<br>family of digestive tract-associated ba<br>Vacca, M. *et al.* The Controversial Role 17. Framily of digestive tract-associated bacteria. *Genome Biol Evol* **6**, 703-13 (2014).<br>48. Vacca, M. *et al.* The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms* **8**(2020).
- family of digestive tract-associated bacteria. Genome Biol Evol 6, 703-13 (2014).<br>Vacca, M. *et al.* The Controversial Role of Human Gut Lachnospiraceae. *Microorg*<br>Macca, M. *et al.* The Controversial Role of Human Gut La 48. Vacca, M. et al. The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms* 8(2020).

- 
- 
- 49. Friedman, J. & Alm, E.J. Inferring Correlation Networks from Genomic Survey Data. 7209<br>
Computational Biology 8, e1002687 (2012).<br>
50. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. & Egozcue, J.J. Microbiome Dataset Computational Biology 8, e1002667 (2012).<br>Gloor, G.B., Macklaim, J.M., Pawlowsky<br>Compositional: And This Is Not Optional. *Frc*<br>Marcos-Zambrano, L.J. *et al.* Applications<br>Review on Feature Selection, Biomarker<br>*Microbiol* Compositional: And This Is Not Optional. *Front Microbiol* **8**, 2224 (2017).<br>
51. Marcos-Zambrano, L.J. *et al.* Applications of Machine Learning in Human Microbiome Studies: A<br>
Review on Feature Selection, Biomarker Ident Compositional: And This Is Not Optional: *Front Microbiol* 8, 2224 (2017).<br>Marcos-Zambrano, L.J. *et al.* Applications of Machine Learning in Hi<br>Review on Feature Selection, Biomarker Identification, Disease Prec<br>*Microbio* Fraction Scalings C.J. et al. Applications of Machine Learning in Human Microbiolne Statistics. A<br>Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Front<br>Microbiol 12, 634511 (2021).<br>
- Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Tront<br>Microbiol 12, 634511 (2021).<br>Ho, P.Y., Nguyen, T.H., Sanchez, J.M., Defelice, B.C. & Huang, K.C. Resource competition predicts<br> Microbiol 12, 034311 (2021).<br>Ho, P.Y., Nguyen, T.H., Sanc<br>assembly of gut bacterial com<br>(!!! INVALID CITATION !!! {}).<br>Sokal, R.R. & Rohlf, F.J. *Biome*<br>Freeman, 1981).
- 
- 53. Assembly of gut bacterial communities in vitro. Nat Microbiol 9, 1036-1048 (2024).<br>
53. (!!! INVALID CITATION !!! {}).<br>
54. Sokal, R.R. & Rohlf, F.J. *Biometry: The Principles and Practice of Statistics in Biological R* assembly of gut bacterial communities in vitro. Nat Microbiol 9, 1036-1048 (2024).<br>(!!! INVALID CITATION !!! {}).<br>Sokal, R.R. & Rohlf, F.J. *Biometry: The Principles and Practice of Statistics in Biologi*<br>Freeman, 1981). 54. Sokal, R.R. & Rohlf, F.J. *Biomi*<br>Freeman, 1981).<br>**Example 2** 54. Sokal, R.R. & Rohlf, F.J. Biometry: The Principles and Practice of Statistics in Biological Research, (W. H.<br>Freeman, 1981).  $\frac{1}{2}$

## Figures

Figure 1. Co-abundance testing principle<br>Panel a) illustrates the targeted mechanism at the individual level. The equilibrium of a given Panel a) illustrates the targeted mechanism at the individual level. The equilibrium of a given microbiota ecosystem can change conditionally on the environment, impacting not only the marginal abundance, but also the co-abundance of species. Under environment E1, the available local resources foster a complementary system involving species  $\alpha$ ,  $\beta$ , and  $\omega$ . Under environment E2, the available local resources foster a complementary system involving species  $\beta$ , and  $\delta$ . Panel b) illustrates an application scenario at the population level, where health features (biological, clinical illustrates an application scenario at the population level, where health features (biological, clinical or environmental variables) are tested for association with the covariance matrix of the microbiome measured in the same individuals. When the feature of interest is binary (e.g. exposed vs<br>unexposed), the approach consists in testing for statistical differences in the observed microbiome unexposed), the approach consists in testing for statistical differences in the observed microbiome covariance derived in exposed and unexposed participants separately. When the feature of interest is continuous, the same principle applies.



Figure 2. Milieu Interieur microbiota composition<br>Panel a) shows the relative abundance of taxa across the seven taxonomic levels (Kingdom, Phylum, Panel a) shows the relative abundance of taxa across the seven taxonomic levels (Kingdom, Phylum, Class, Order, Family, Genus, Species). Panel b) shows the covariance matrices at the Family, Genus, and Species levels derived from all 938 patients.



Figure 3. Screening for environmental and clinical factors<br>Association screening between 80 environmental and clinical factors from the Milieu Interieur cohort and the covariance of taxa at the species (a), genus (b) and family (c) taxonomic levels. Each panel display the - $log 10(P)$  of each predictor. Results from MANOCCA are based on the optimal number of display the -log10(P) of each predictor. Results from MANOCCA are based on the optimal number of principal components. Results from MANOCCA are compared against three alternative approaches: a standard MANOVA, and alpha diversity tests based on both Simpson and Shannon metrics. The red dash line indicates the stringent Bonferroni correction threshold accounting for all predictors and sets of PCs tested.



**Figure 4. Cross and inter-family interactions for age, sex, smoking and BMI**<br>We used the top 5% pairs of genera with the largest contribution to the covariance test to investigate whether variability in co-abundance involved genera from the same family or from different families. Panel (a) presents the relative abundance of the family from these top genera in the whole cohort Panel (a) presents the relative abundance of the family from these top genera in the whole cohort (grey bar), and the number of interactions observed for genera from each of these families. Panel (b) illustrates the difference between cross-family and within family interactions. Detailed co-abundance pairs are presented for age(b), sex(c), smoking(d) and BMI(e). The Y axis represents the families of the top subset of genera involved, and that together explain up to 50% of the signal. The X axis shows the distribution of families for the associated genera, defined as a count per family. The length of the bar indicates how many pairs are involved for each top family. Colours were assigned only for the top families, and white blocs correspond to the unlisted categories.



**Figure 5 Co-abundance network influenced by age, sex, smoking and BMI**<br>For the top four associated features from the MANOCCA (age, sex, BMI and smoking), we extracted For the top 1000 contributing pairs of genera out of the 259,560 total products and derived the direction<br>of effect of each predictor on the pair of co-abundance. We plotted the Venn diagram of shared pairs<br>between each fe of effect of each predictor on the pair of co-abundance. We plotted the Venn diagram of shared pairs<br>between each feature in a) and the overlap in taxa in b). In c) we show the distribution of direction of<br>effects per pred of effects per predictor, and for the age – smoking and sex – BMI intersections. We then used the pairs of features to derive a network of the changes in correlation with regard to each predictor. The node size, representi effects per predictor, and for the age – smoking and sex – BMI intersections. We then used the pairs<br>of features to derive a network of the changes in correlation with regard to each predictor. The node<br>size, representing of features to derive a network of the changes in correlation with regard to each predictor. The node<br>size, representing a genus, is proportional to its number of contributions with other genera, and<br>edges link the top con size, representing a genus, is proportional to its number of contributions with other genera, and<br>edges link the top contributing pairs. The edge colors indicates the direction of effect with green<br>indicating that an incre size, representing a genus, is proportional to the colors indicates the direction of effect with green indicating that an increase of the predictor drives an increase in co-abundance, red shows that an increase of the pre eduction of the predictor drives an increase in co-abundance, red shows that an increase of the predictor drives a reduction in co-abundance and black indicates a mixed direction of effect for the overlapping predictors.



Figure 6. Predictive power of covariance-based models<br>We estimated the predictive power of co-abundance models (blue/green boxplots) as compared to The standard mean-based multivariate model (red/orange boxplots). For each of the four most associated features (age, sex, smoking and BMI) and the three lower taxonomic levels (species, genus, family), we derived the pre associated features (age, sex, smoking and BMI) and the three lower taxonomic levels (species, genus, family), we derived the prediction accuracy using the squared-correlation  $(R^2)$  for continuous features and the area u Experies Species Centus Constants Centus Centus<br>
Species Centus Cen genus, ramily), we derived the prediction accuracy using the squared-correlation (R<br>features and the area under the receiver operating curve (AUC) for binary feature. 1<br>done using a 30-fold cross validation, with the score The analysis was<br>
Infe analysis was<br>
Ind the R2 being<br>
pdels was tested<br>
T done using a 30-fold cross validation, with the score being train in 90% of the data, and the R2 being<br>derived in the remaining 10%. The significance of the difference between the two models was tested<br>using a two-sample



AUC linear model = AUC MANOCCA