

Supplementary Material

Contents

Supplementary Notes.....	1
Milieu Interieur gut microbiome data	1
Impact of dimension reduction in MANOCCA on association test.	1
Compositional data	2
Existing network-based approaches	2
Supplementary Figures	3
Figure S1. Milieu Interieur microbiota composition and MANOCCA calibration	5
Figure S2. Environmental screening	7
Figure S3. Covariance versus mean effect	8
Figure S4. Methods comparison and compositionality	10
References	10

Supplementary Notes

Supplementary Note 1: Milieu Interieur gut microbiome data

Stool samples were collected during two visits. For participants that had two samples, we only kept the sample from the first visit. Shotgun metagenomic was applied to derive the bacterial profiles up to the specie level. Detailed description of the data generation can be found in Byrd et al¹. In brief, stool specimens were collected in a double-lined sealable bag containing a GENbag Anaer atmosphere generator (Aerocult; Biomerieux). Then fresh samples were aliquoted into cryotubes and stored at -80°C. Stool aliquots were shipped to the CRO Diversigen for DNA extraction and shotgun metagenomic sequencing using an Illumina HiSeq 2500. In the end, 21 trillion raw paired-end reads from 1,359 samples from 946 of the donors were obtained. To process the reads, Illumina TruSeq adapters were trimmed with Trimmomatic v0.36²; low-quality and low-complexity reads were removed with prinseq-lite 0.20.4³; and Bowtie2 v2.1.0⁴ was used to remove reads mapping to PhiX or the PacBio human genome. After processing, there were on average 13.9 ± 2.9 million reads per sample. Of an initial 1,000 recruited donors, 44 were excluded from this analysis because of lack of consent for sharing their data outside of the MI consortium. An additional 10 donors were excluded because of technical issues in the extraction and sequencing steps (e.g., low DNA extraction yield), resulting in a sample size for the shotgun dataset of 946 donors. The final quantification was made using a Kraken-GTDB database⁵ on June 25, 2019. In the end, 23,505 genomes representing 13,446 unique bacterial species were downloaded and formatted into a Kraken2 database⁶.

Supplementary Note 2: Impact of dimension reduction in MANOCCA on association test.

When applied to taxa, we varied the number of principal components used in MANOCCA from two to one hundred but limited the number of PC analysed for each predictor based on the guidelines⁷, and used a stringent multiple testing significance threshold to account for the various number of PCs considered. The optimal number of principal components –corresponding to the smallest observed *P* value– varied substantially, but leans towards small values with a median of 11 PCs over the 80 variables analysed (**Fig. S2d**). However, it is substantially higher for three out of the four variables identified with maximum signal obtained at the genus level for 62, 83, 30 and 3

43 PCs for age, sex, smoking and BMI, respectively. This high number of PCs suggests that the association with the co-
44 abundance involves a fairly large number of taxa.

45 46 **Supplementary Note 3: Compositional data**

47 Due to high-throughput sequencing (HTS) technologies, microbiome data is compositional in nature⁸. The counts
48 given by 16S or shotgun sequencing depends on the sequencing depth and can vary substantially across samples.
49 It can therefore not be related to the absolute count of bacteria in the input sample, meaning that only the relative
50 abundance of microbial taxa can be used. This implies that every sample sums to a constant total. As pointed in
51 previous studies^{9,10}, this constraint can lead to issues such as spurious correlations, non-independence, and biased
52 estimates when using traditional statistical methods. Typically, transformations like the centered log-ratio (CLR)
53 transformation is utilized to address these challenges but brings other issues such as addressing 0 abundance
54 counts¹¹.

55 In this study, we applied our covariance-based test directly on the relative abundance. Although future work
56 might investigate alternative metric, not modelling the compositional component should not impact the accuracy
57 of our results. First, as showed in the simulation from **Figure S4i-n**, the negative bias in correlation estimates
58 discussed in Gloor et al⁹ appears when confronted with a small amount of OTUs in the model, but becomes
59 negligible when reaching the sample size and number of OTU analyzed in the present study. Second, the proposed
60 approach is not qualifying the correlation between taxa per se, but instead estimates changes in covariance
61 conditional on a given variable of interest. This is a critical point. Indeed, even though some bias remains, it will
62 only constitute a constant offset. Assuming the host factors considered in the present study are not associated with
63 the total count of reads across individuals, this bias should have no impact on estimating the relationship between
64 factors and taxa co-abundance.

65 66 **Supplementary Note 4: Existing network-based approaches**

67 There is no gold-standard method to investigate factors associated with taxa co-abundance, and existing
68 methods are known to be sensitive to parameter tuning¹². Existing approaches are typically based on the evaluation
69 of networks co-occurrence and consists in three steps: i) estimating the pairwise correlation across taxa, ii)
70 dichotomizing the correlation metric based on a given threshold in order to define a sparse network, and iii)
71 comparing the inferred network based on descriptive statistics such the number of edges and nodes across various
72 conditions. This approach is only applicable to binary (e.g. a disease status) or categorical predictors (e.g. multiple
73 cohorts) with few categories, as it requires building a network for each group considered. Moreover, comparison
74 between networks derived from different approaches is not straightforward, and typically rely on *ad hoc* empirical
75 *P*-value derived through permutation, or test of heterogeneity applied at the taxa level¹³. Finally, not all methods
76 allow for adjustment for potential covariates making the interpretation of the relative contribution of each factor
77 difficult.

78 For comparison purposes, we used SparCC to identify the taxa displaying variability in co-abundance conditional
79 on binary predictors. SparCC¹⁴ was developed to estimate correlations from high-throughput sequencing microbial
80 communities data, and especially to address the compositional aspects of such data. It is based on the work from
81 Aitchison¹⁵ on the estimation of the variance of the log ratio of two quantities of OTUs i and j : $t_{ij} =$
82 $Var[\log(x_i/x_j)]$. Since $t_{ij} = Var[\log(x_i)] + Var[\log(x_j)] - 2Cov[\log(x_i), \log(x_j)]$, the quantity t_{ij} can be
83 interpreted in relation to the basis abundance's variances. SparCC relies on the assumption that the true correlation
84 network is sparse and estimates the basis variance iteratively, computing all correlations and removing extreme
85 correlations, which drives the sparse correlation assumption. Second, the log ratio cannot be derived for OUT that
86 have an abundance of zero¹¹. SparCC addresses this issue by applying a pseudo-count to all null occurrences of
87 OTUs, assuming that all OTUs have a minimal occurrence in every sample. SparCC has proven very useful for the
88 specific purpose of characterizing correlations. Its performance to identify predictors associated with variability in
89 co-abundances is less clear¹⁶. In practice, an empirical *P*-value needs to be derived, which requires to run N_{simu}
90 times SparCC, which can be a strong computational burden and subject to noise if N_{simu} is too small. Furthermore,
91 SparCC can only be applied to categorical predictors, strongly limiting its applicability. Should the significant
92 correlations be in the union, difference or intersection of correlations detected for each predictor values.

93 We also derived a naïve empirical permutation correlation test, which mimics an approach presented in a review
94 of existing methods¹². Here, we derived an original correlation value, and then shuffled the data N_s times to derive
95 an empirical distribution of correlations and derive an empirical P -value as $P = \min(\#above, \#below)/N_s$. This
96 approach displays severe limitations, the main one being the noise detections that requires an extreme number of
97 permutations to eliminate.

98 In the case of MANOCCA, the usage of a linear framework allows to bypass any empirical simulation by using
99 parametric statistics. Additionally, MANOCCA tests for a progressive change in correlations, allowing to take into
100 account information from ordinal and continuous predictors. The resulting features are directly linked to changes
101 in the covariance structure of the outcome. Finally, in comparison with the literature this approach allows to adjust
102 for unwanted confounding variables in the model.

103

104 **Supplementary Figures**

105

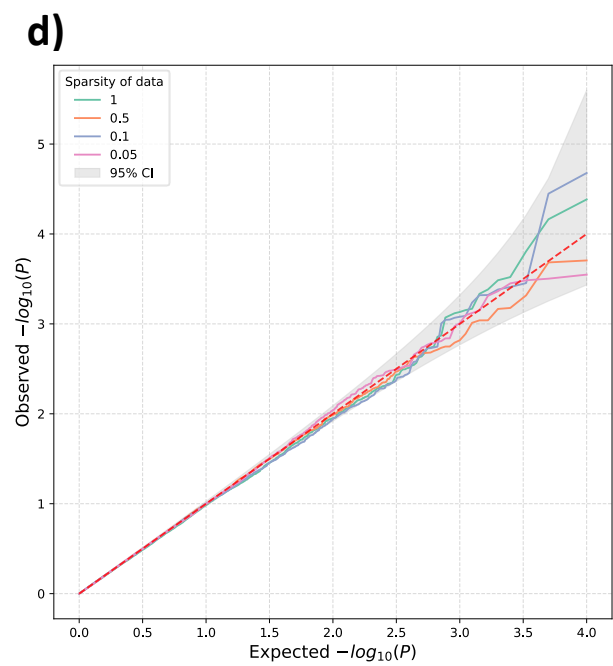
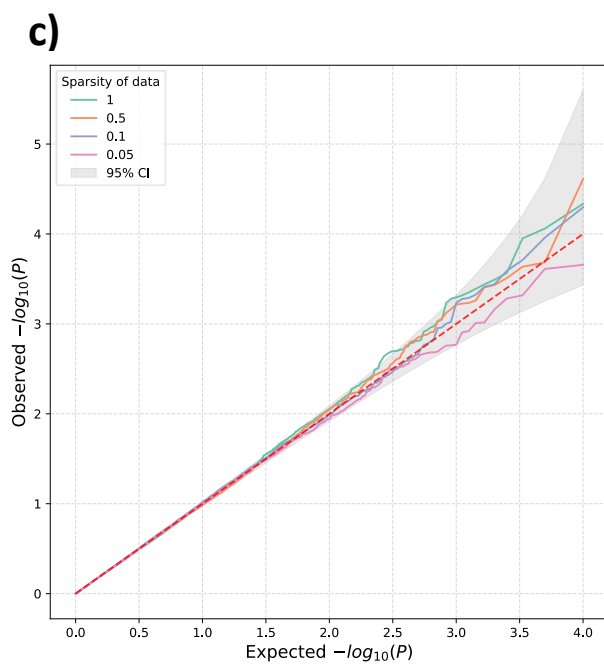
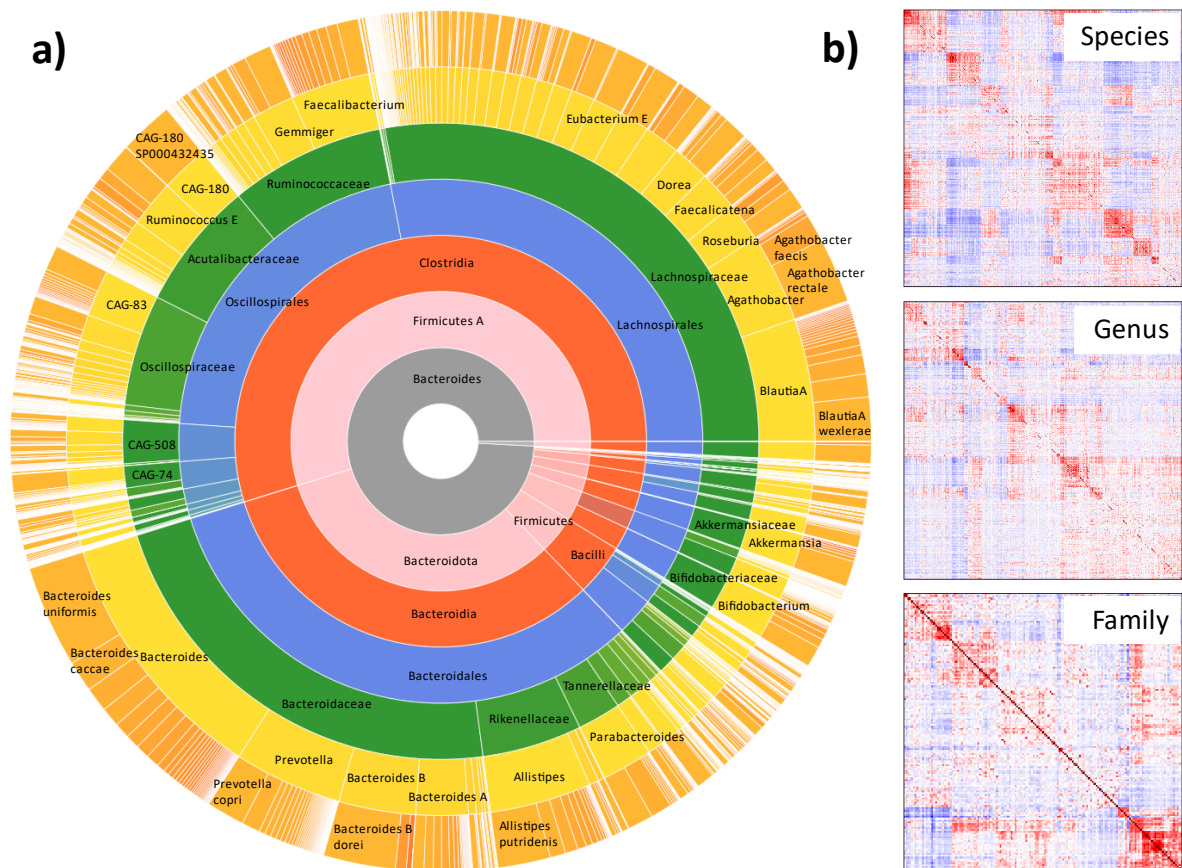
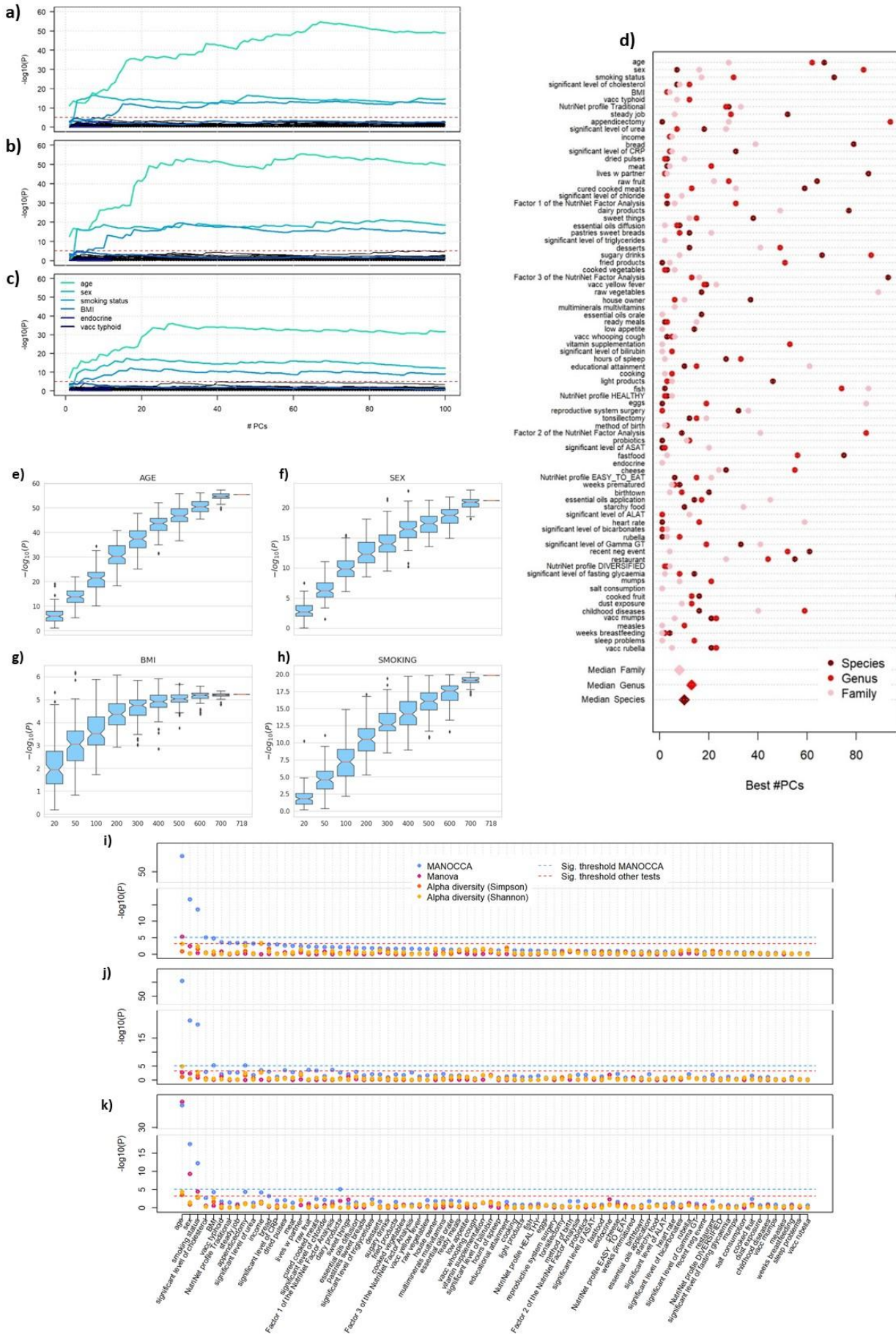


Figure S1. Milieu Interieur microbiota composition and MANOCCA calibration

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. Panels c) and d) display calibration plots with increasing data sparsity when considering a normally distributed predictor (c) or binary predictor of frequency 0.4 (d). QQplots were generated using a simulated multivariate normal distribution of sample size 1000 and with 100 outcomes. The MANOCCA was performed over 10000 simulations on the PCA transformed matrix of products, keeping the 50 top PCs for the analysis. Outcome sparsity was incrementally simulated using a multinomial distribution of frequency 1 (no sparsity), 0.5, 0.1, 0.05 (minimum threshold use in this analysis).



119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141

Figure S2. Environmental screening

Applying MANOCCA when the number of products is larger than the sample size requires reducing the dimension of the outcome data. Here we used principal component analysis (PCA) applied to the matrix of products, including only the top principal components (PCs) explaining the largest amount of total variance. We investigated the power of MANOCCA when varying the number of PCs kept in the model. The panels display the $-\log_{10}(P)$ from the MANOCCA for each of the 80 predictors considered, as a function of the number of principal components selected, varying from 1 to 100 for the species (a), genus (b) and family (c) levels. The most associated predictors are highlighted in blue gradient. For each of the 80 environmental predictors considered, we recorded the best number of principal components (PCs) maximizing the association signal from MANOCCA in panel d). Results are presented for the species (dark red), genus (red) and family (pink) levels. At the bottom we recorded for each taxonomic level the median of the number of PCs kept across all predictors. To assess the sensitivity of the co-abundance signal observed, we re-ran our analysis on random subsets of 20 to 700 genus taxa (from a total of 718 genera). For each sample size, we sampled 100 subsets and derived the P -value from MANOCCA. The four panels e-h) show boxplots across the series of 100 subsets for the four host factors of interest: age (e), sex (f), BMI (g), and smoking (h). Finally, the association screening between 80 environmental and clinical factors from the Milieu Interieur cohort and the covariance of taxa at the species (i), genus (j) and family (k) taxonomic levels. Each panel display the $-\log_{10}(P)$ 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.

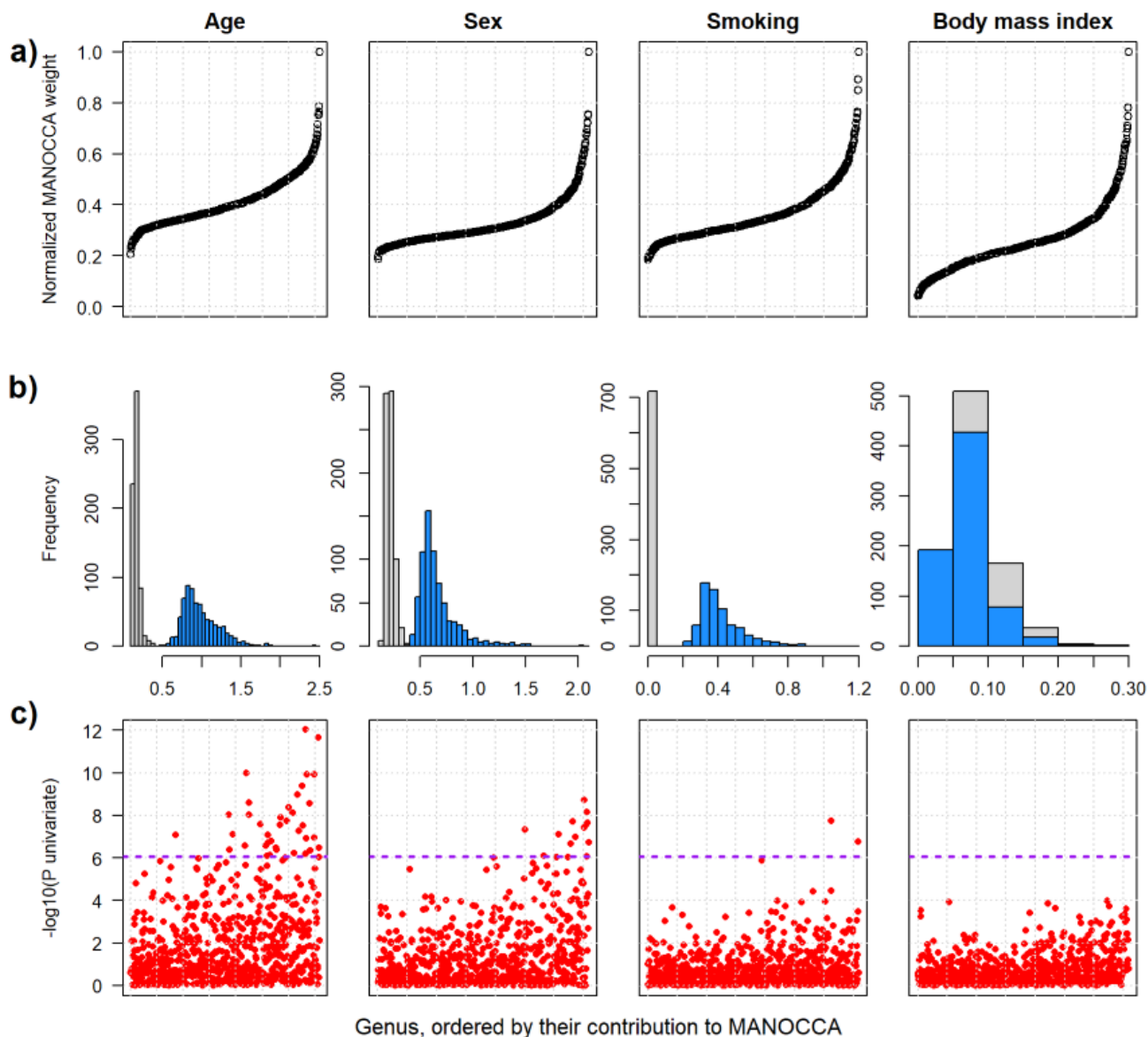
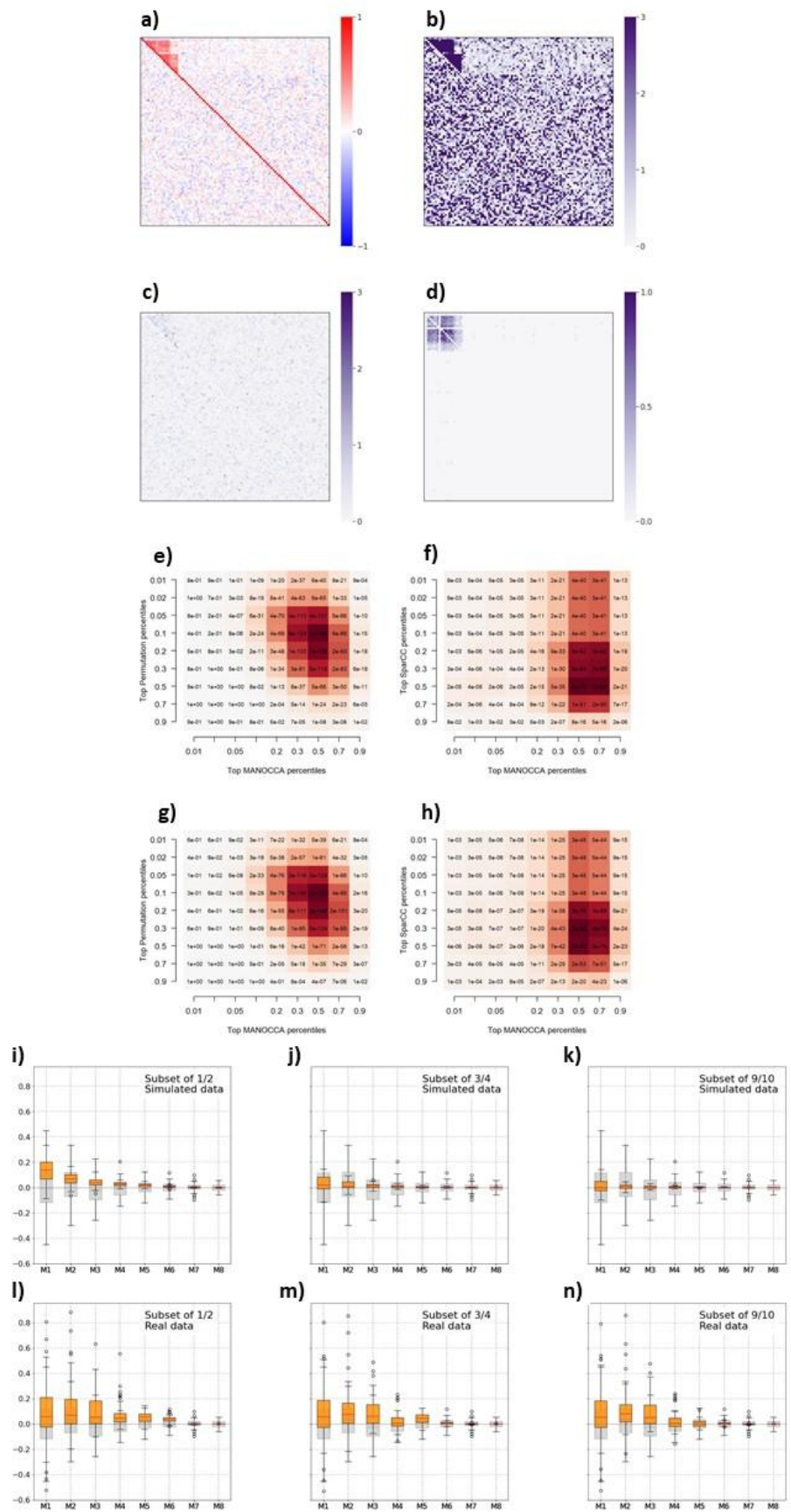


Figure S3. Covariance versus mean effect

We derived the contribution to the covariance signal from each taxon for the top four associated host variables: age, sex, BMI and smoking status. We compared those contributions to a univariate screening of mean abundance for the same taxa. Panel a) displays the normalized contribution across all the 721 genus taxa for MANOCCA. Panel b) displays the distribution of the contribution (blue), as compared to a null distribution obtained after shuffling the predictor considered (grey). Panel c) displays the $-\log_{10}(P\text{-value})$ from the univariate linear regression for the same taxa. The dotted blue line displays the Bonferroni threshold for the univariate tests.



159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192

Figure S4. Methods comparison and compositionality

We ran a simulation study where data where 100 variables were drawn two multivariate normal distribution with correlation matrices C_1 and C_2 , respectively. Correlation matrix C_1 was simulated from random noise, and the second correlation matrix C_2 was a copy of C_1 but the upper left square of 20 variables was modified to induce strong correlations. Both matrices are presented in the lower and upper triangles of panel a), respectively. We generated a dataset including 400 samples drawn using C_1 , and 600 samples drawn using C_2 . We ran a permutation-based approach, SparCC and MANOCCA on the simulated dataset. Panel b) and c) show in the lower triangle the $-\log_{10}(Pvalue)$ for each cell on data generated using C_1 and using C_2 in the upper triangle. Panel d) displays the top 190 normalized MANOCCA features derived using the simulated data as outcome and the binary predictor 0 if data was generated using C_1 and 1 if using C_2 . We estimated the significance of the overlap between networks of impacted co-abundances derived using the MANOCCA's contributing products, and two alternative approaches, an ad hoc permutation-based approach and SparCC. Because the two latter approaches are only applicable to binary predictor, we conducted our comparison only on sex (panels e and f)) and using a binary version of smoking (never/ever smoking) (panels g and h)). We applied all methods at the genus level, comparing a total of 257,403 pairs of taxa. Pairs were ranked based on their MANOCCA's contribution, and the P-value for the permutation-based approach and SparCC. Pairs of genera belonging to the top percentiles, varied from 0.01 to 0.9, were extracted for each approach, and the significance of the overlap between the two set was derived using a Binomial test. For both predictors and both approaches, the most significant overlap was observed when using the top 50% contributing pairs of MANOCCA and the top 10% associated pairs from the permutation-based approach. Finally, we assessed the extent of spurious correlations due to the compositional component of microbiome data across series of 50 replicates with various sample size and number of variables. We considered eight settings with [N,M], where N is the sample size and M the number of variables: [20,10] (M1), [50,20] (M2), [100,30](M3), [200,40](M4), [400,60](M5), [600,100](M6), [800,300](M7), [900,700](M8). For each setting we derived the mean difference between correlations computed on the whole dataset and correlations computed on a subset of 1/2, 3/4, 9/10 of the variables. In grey we display the basis expected randomness of correlations between two normal distributions with no compositional aspect. Panels i-k) display results with data simulated using a uniform distribution between 10 and 10000 and for increase proportions of subsets. Panel l-n) display results on the real data using the 718 genera.

References

1. Byrd, A.L. *et al.* Gut microbiome stability and dynamics in healthy donors and patients with non-gastrointestinal cancers. *J Exp Med* **218**(2021).
2. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).
3. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-4 (2011).
4. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
5. Parks, D.H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004 (2018).
6. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46 (2014).
7. Boetto, C. *et al.* A multivariate outcome test of covariance. *bioRxiv*, 2023.09.20.558234 (2023).
8. Quinn, T.P., Erb, I., Richardson, M.F. & Crowley, T.M. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34**, 2870-2878 (2018).

9. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. & Egozcue, J.J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**, 2224 (2017).
10. Greenacre, M., Martinez-Alvaro, M. & Blasco, A. Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. *Front Microbiol* **12**, 727398 (2021).
11. Martín-Fernández, J.A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology* **35**, 253-278 (2003).
12. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* **10**, 1669-81 (2016).
13. Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat Commun* **11**, 4018 (2020).
14. Friedman, J. & Alm, E.J. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology* **8**, e1002687 (2012).
15. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)* **44**, 139-177 (1982).
16. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics* **20**, 329 (2019).