Alteration of Gut Microbiome in Patients With Schizophrenia Indicates Links Between Bacterial Tyrosine Biosynthesis and Cognitive Dysfunction

Supplement 1

Table S1: Phenotype characteristics of study participants

	Schizophrenia cases	Dysmetabolic controls	Healthy controls	$P_{Dysmetabolic}$	$P_{Healthy}$
N (% women)	132 (55.3%)	132 (55.3%)	132 (56.8%)	1	1
Age (y)	41 (12)	56 (8)	39.8 (12)	<0.001	0.668
BMI (kg/m^2)	35.0 (6.2)	33.7 (4.5)	23.8 (3.8)	0.071	<0.001
Waist circumference (cfm)	116 (14)	110 (12)	82 (11)	0.001	<0.001
Systolic BP (mmHg)	129 (16)	144 (19)	123 (18)	<0.001	0.024
Diastolic BP (mmHg)	82 (10)	88 (10)	76 (11)	<0.001	<0.001
HbA1c (mmol/mol)	39 (10)	38 (4)	34 (3.)	0.75	<0.001
P-Cholesterol (mmol/L)	5.0 (1.1)	5.5 (0.9)	4.8 (1.1)	0.003	0.200
P-LDL (mmol/L)	3.0 (1.1)	3.4 (0.8)	3.2 (1.0)	0.001	0.33
P-HDL (mmol/L)	1.23 (0.37)	1.37 (0.39)	1.45 (0.41)	0.010	<0.001
P-TG (mmol/L))	2.09 (1.65)	1.46 (0.68)	1.10 (0.62)	<0.001	<0.001
BACS	235.9 (51.7)				
SAPS	1.8 (1.5)				
SANS	2.2 (1.2)				
GAF	46.0 (7.2)				
Diagnosis (%)	F.20: 120 (90.9%), F.25: 10 (7.5%), F.062: 1 (0.8%), F.22: 1 (0.8%)				
Disease duration (y)	26.3 (8.5)				

Data is presented as mean (SD) or number of individuals (%). P-values are from Wilcoxon rank sum tests (continuous variables) or Fisher's exact tests (categorical variables) comparing schizophrenia cases to healthy ($P_{Healthy}$) and dysmetababolic ($P_{Dysmetabolic}$). *BACS: Brief Assessment of Cognition Score. GAF: Global Assessment of Functioning. SAPS: Scale for the Assessment of Positive Symptoms. SANS: Scale for Assessment of Negative Symptoms. All values for mentioned scores are given in arbitrary units. BMI: Body mass index. BP: Blood pressure. HbA1c: Haemoglobin A1c. P-HDL: Plasma high-density lipoprotein. P-LDL: Plasma low-density lipoprotein. P-TG: Plasma triglyceride.*

Figure S1: Overview of medication taken by participants.

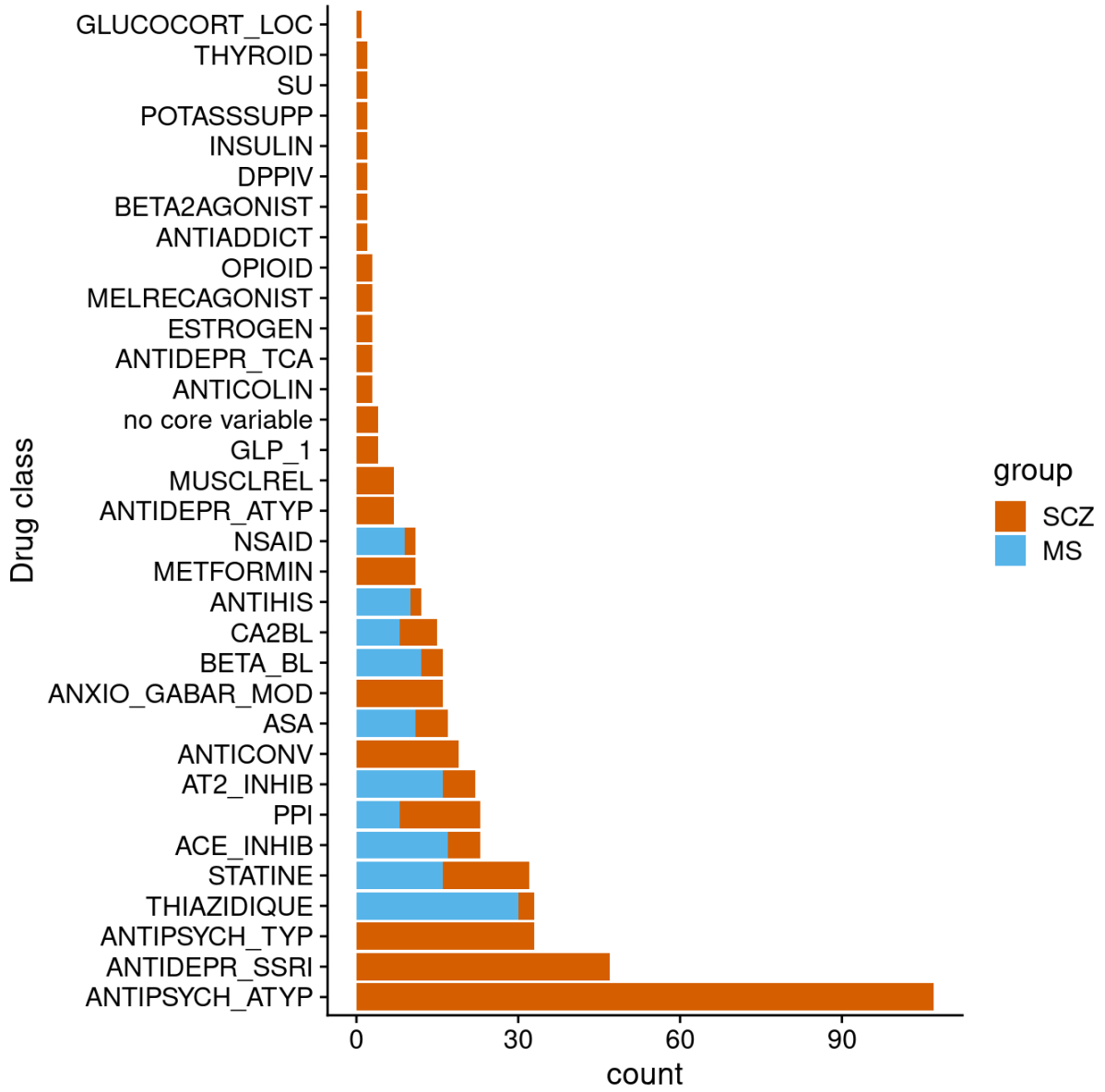

Figure S1: Number of individuals taking a class of medication. Drugs were grouped according to their Anatomical Therapeutic Chemical (ATC) code.

Figure S2: Contrast in Shannon index (alpha-diversity)

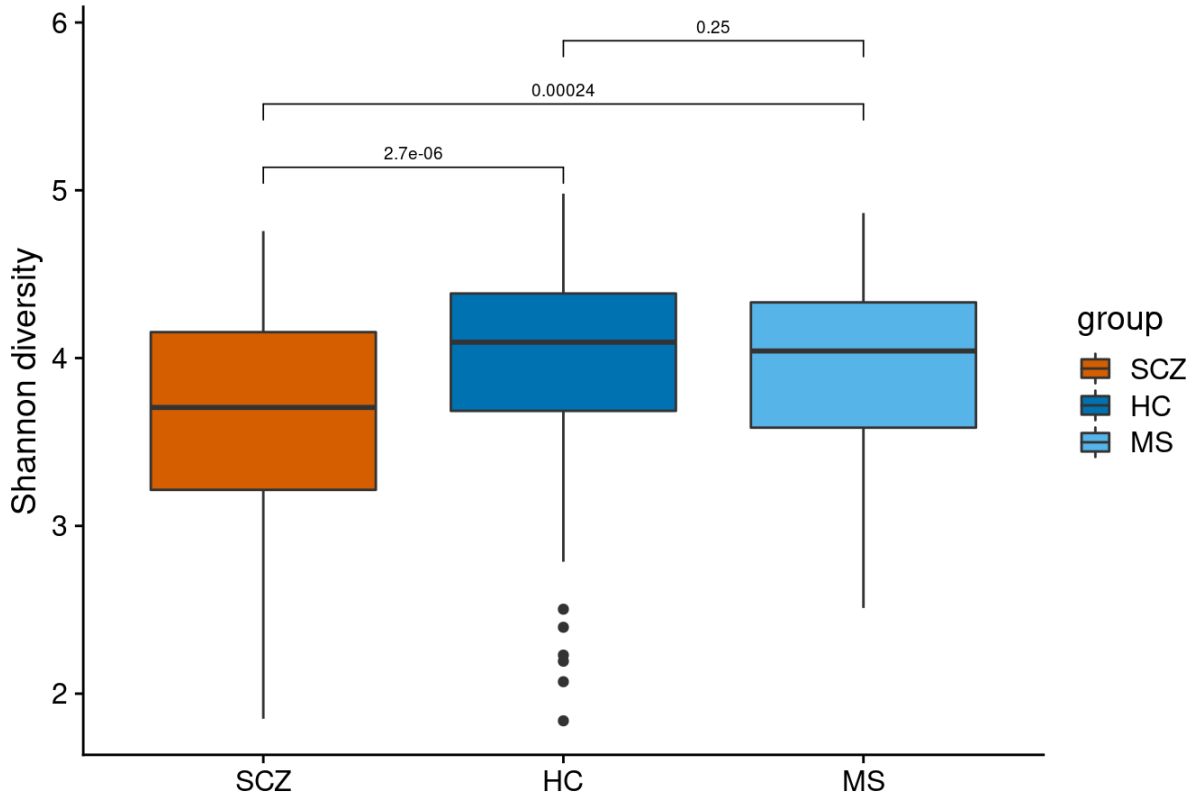

Figure S2: Difference in Shannon index between patients with SCZ, HCs and MS. P-values from Wilcoxon tests are displayed. Boxes represent the median and interquartile ranges (IQRs) between the first and third quartiles; whiskers represent the lowest or highest values within 1.5 times IQR from the first or third quartiles. *HC = Healthy Controls; MS = dys-metabolic controls with Metabolic Syndrome; SCZ = patients with schizophrenia.*

Figure S3: MGS-based classification of SCZ patients and controls

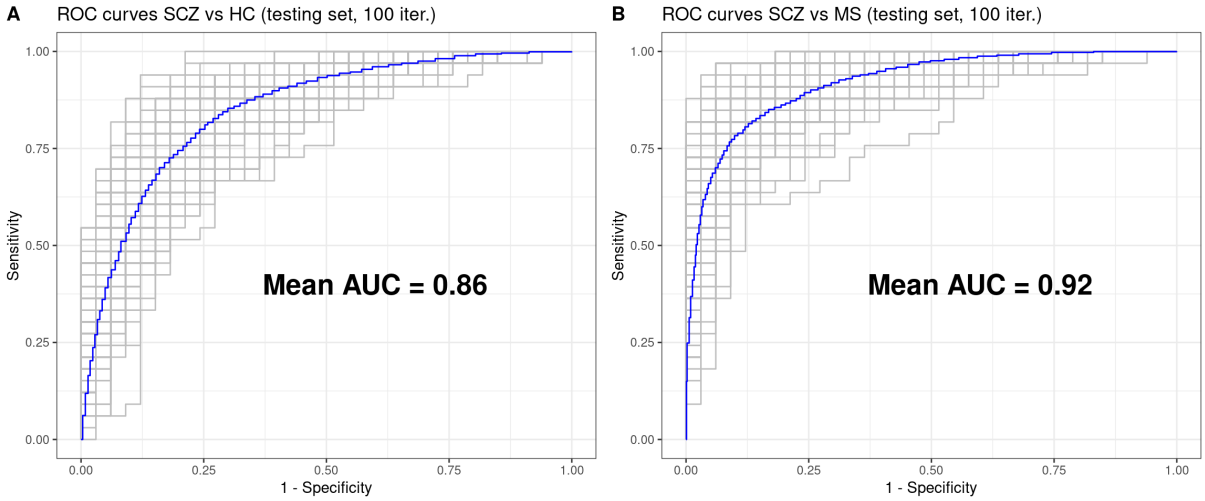

Figure S3: ROC curves of the MGS-based elastic net classifications between SCZ patients and **(A)** HC or **(B)** MS. *HC = Healthy Controls; MS = dys-metabolic controls with Metabolic Syndrome; SCZ = patients with schizophrenia; MGS = Metagenomic Species.*

Figure S4: Taxonomic contrast between schizophrenia cases and healthy controls

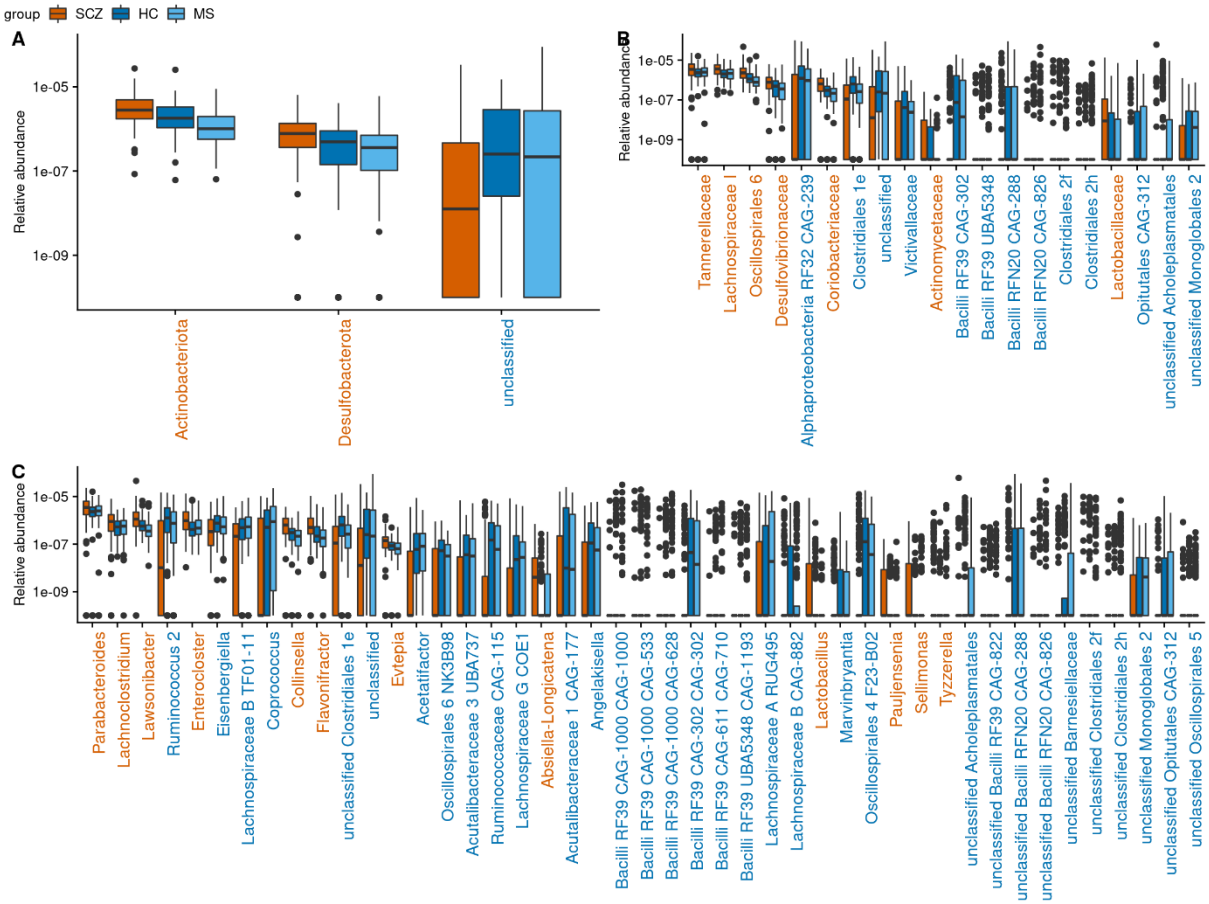

Figure S4: Distribution of contrasted taxonomy at (A) phylum level, (B) family level and (C) genus level. Red-labeled and blue-labeled taxonomy are enriched and depleted in schizophrenia cases, respectively. *HC = Healthy Controls; MS = dys-metabolic controls with Metabolic Syndrome; SCZ = patients with schizophrenia; MGS = Metagenomic Species.*

Figure S5: Beta-diversity and schizophrenia diagnosis.

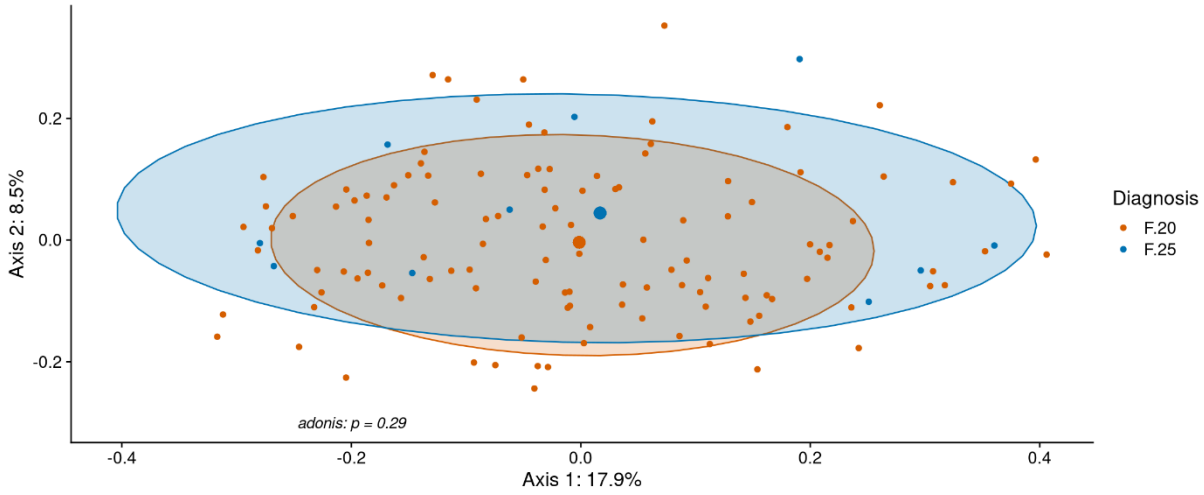

Figure S5: Principal Coordinates Analysis ordination of the Bray-Curtis dissimilarity matrix computed on the MGS abundance from patients with SCZ diagnosed as F.20 or F.25. P-value associated with the PERMANOVA analysis between these two groups is displayed.

Figure S6: Modelling of Brief Assessment of Cognition Scores (BACS)

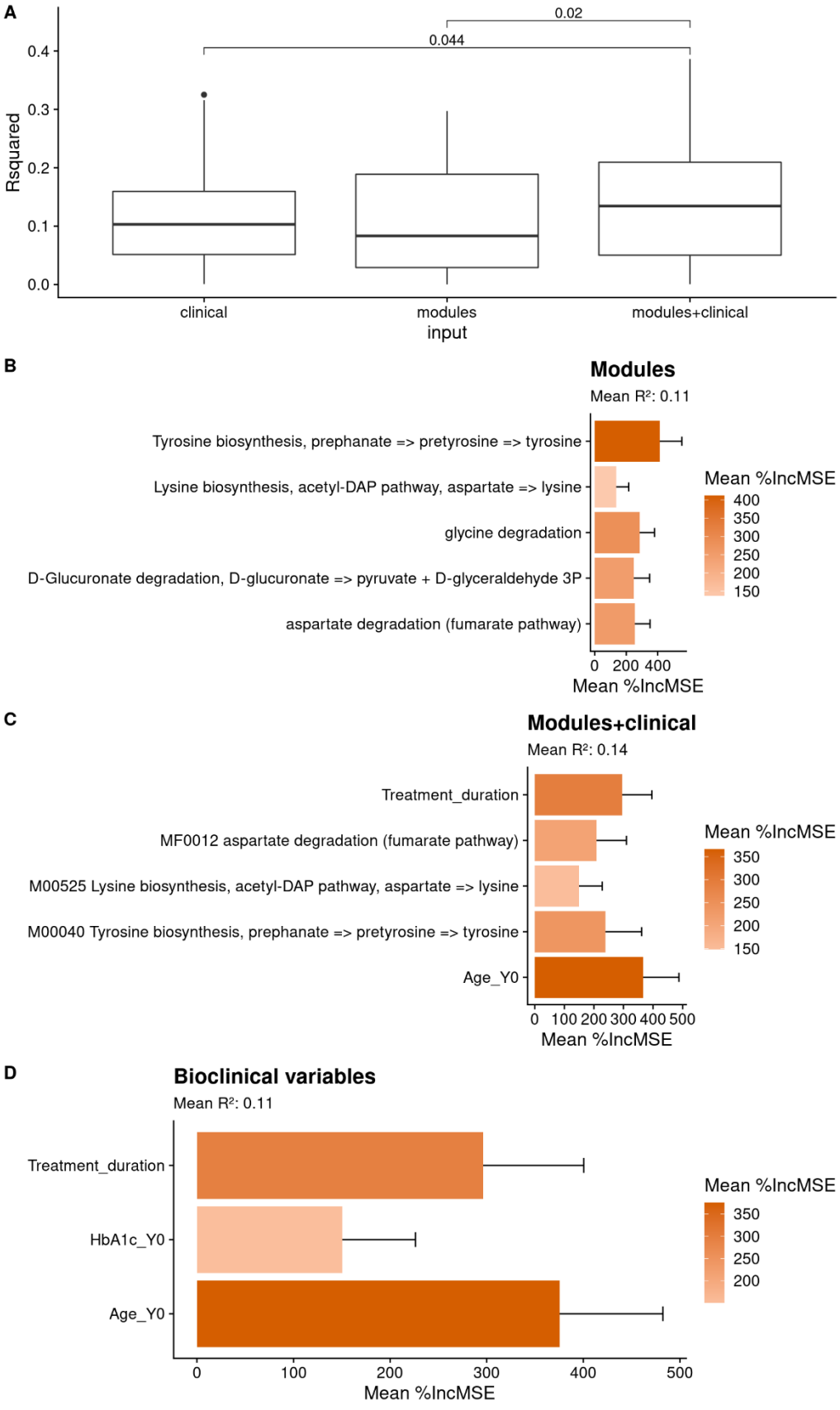

Figure S6: (A) R^2 distribution of BACS modelling using two types of explanatory variables: metagenome-predicted functional modules or a combination of functional modules and bioclinical variables. For each group of explanatory variables, 100 models were run after resampling of the training and testing set. Performance (R^2) was assessed on the testing set. **(B-C)** Features that were selected for BACS regression at least 50 times out of 100 resampling of training and testing sets when considering **(B)** only functional modules or **(C)** functional modules and clinical variables as explanatory variables. *%IncMSE: increase in mean-squared error*.

Supplementary Methods

Clinical examinations and sample collection

Participants (cases with schizophrenia (SCZ) and normal-weight healthy controls (HC)) were examined in the morning following an overnight fast. Body weight was recorded on an electronic scale in light indoor clothing or underwear. Height was measured using a wall-mounted stadiometer. Body-mass index (BMI) was calculated as weight in kg divided by the square of height in meters. Waist circumference was measured midway between the iliac crest and the lower costal margin using a non-expandable measuring tape. Blood pressure was recorded as the mean of triplicate measurements following a 5-minute rest using an automated sphygmomanometer.

Blood samples were taken by puncture of the antecubital vein in the morning after an overnight fast. Plasma triglyceride (TG), total cholesterol (TC), and high-density lipoprotein (HDL) were analysed on a Vitros 5600 system (CV 14.6%, 11.6%, and 17.0%, respectively). Very-low-density lipoprotein (VLDL) was calculated as $VLDL = 0.45 \times TG$. Low-density lipoprotein was calculated as $LDL = TC - HDL - VLDL$. Non-HDL was calculated as $Non\ HDL = TC - HDL$. Haemoglobin A1c (HbA1c) was analysed by high-performance liquid chromatography (HPLC) on a TOSOH G8 system (Tosoh Bioscience, San Francisco, CA USA, CV 7.2%). Information about medication taken by cases with schizophrenia was obtained from case record prescription and interview (Figure S1). Stool samples were collected by the participants at home following standardized procedures (International Human Microbiome Standards, IHMS (1)) including immediate freezing at −18°C. Samples were transported to the laboratory using an insulating coller bag or styrofoam boxe containing cooling elements or dry ice. At the laboratory, samples were stored at −80°C until DNA extraction.

DNA extraction of stool samples and shotgun sequencing

DNA extraction from aliquots of fecal samples obtained from cases with SCZ and HC was performed following IHMS SOP P7 V2 (1). DNA was quantitated using Qubit Fluorometric Quantitation (ThermoFisher Scientific, Waltham, US) and qualified using DNA size profiling on a Fragment Analyzer (Agilent Technologies, Santa Clara, US). Three µg of high molecular weight DNA (>10 kbp) was used to build the library. Shearing of DNA into fragments of approximately 150 bp was performed using an ultrasonicator (Covaris, Woburn, US) and DNA fragment library construction was performed using the Ion Plus Fragment Library and Ion Xpress Barcode Adapters Kits (ThermoFisher Scientific, Waltham, US). Purified and amplified DNA fragment libraries were sequenced using the Ion Proton Sequencer (ThermoFisher Scientific, Waltham, US), generating 21.9 million reads ± 2.6 of 150 bp (in average) per sample. DNA from fecal samples of controls with the Metabolic Syndrome (MS) from the Metahit project was previously extracted and sequenced as described (2). To match the single-read sequencing of SCZ and HC, reverse reads from the MS samples were removed.

Microbial gene count table

To create the gene count table, the METEOR software was used (3): first, reads were filtered for low-quality by AlienTrimmer (4). Reads that aligned to the human genome (identity > 95%) were also discarded. Remaining reads were trimmed to 75 bases and mapped to the Integrated Gut Catalogue 2 (5,6) (IGC2), comprising 10.4 million of genes, using Bowtie2 (7). The unique mapped reads (reads mapped to a unique gene in the catalogue) were attributed to their corresponding genes. The shared reads (reads that mapped with the same alignment score to multiple genes in the catalogue) were attributed according to the ratio of their unique mapping counts of the captured genes. The resulting count table was further

processed using the R package *MetaOMineR* v1.31 (2). It was downsized to 18 million high-quality reads (considering mapped and unmapped reads) to take into account differences in sequencing. Then the downsized matrix was normalized for gene length and transformed into a frequency matrix (freads per kilobase and per million reads mapped, FPKM normalization). Since SCZ gut microbiota has been found to be enriched in species from the oral cavity (8), the same process was repeated on an oral microbiota catalogue of 8.4 million genes (9).

Metagenomic Species (MGS) profiles

The IGC2 and the oral catalogues were organized into 1990 and 853 Metagenomic Species (MGS, cluster of co-abundant genes), respectively, using MSPminer (6,9,10). After removing duplicated MGS (ie, MGS present in both catalogues), we were left with 2,741 MGS. Taxonomical annotation of MGS was performed using an in-house pipeline. First, all genes were aligned on public databases (ncbi, wgs (11)) using Blast (12). An MGS was annotated with the lowest taxonomical rank (from species to superkingdom) that brought consensus in at least 50% of its genes. To avoid misleading annotations due to error in databases, for each gene the 20 first hits were considered. Relative abundance of an MGS was computed as the mean abundance of its 100 'marker' genes (that is, the genes that correlate the most altogether). If less than 10% of 'marker' genes were seen in a sample, the abundance of the MGS was set to 0. Relative abundances at higher taxonomical ranks were computed as the sum of the MGS that belong to a given taxa. MGS count was assessed as the number of MGS present in a sample (that is, whose abundance is strictly positive).

Microbiome functional potential

Three databases were used to estimate gene functional potential: Kyoto Encyclopedia of Genes and Genomes (KEGG) (13); eggNOG (14); and TIGRFAM (15). Genes from the IGC2 and the oral catalogues were mapped with diamond (16) onto KEGG orthologs (KO) from the KEGG database (version 8.9). Each gene was assigned to the best-ranked KO among hits with e-value $< 10^{-5}$ and a bit score > 60. The same procedure was used with eggNOG (version 3.0). The gene catalogues were searched against TIGRFAM profiles (version 15.0) using HMMER 3.2.1 (17). Then we assessed presence of KEGG modules, Gut-Metabolic Modules (GMMs) (18) and Gut-Brain Modules (GBMs) (19) for each MGS and each sample. A functional module consists in an ensemble of KOs (or NOGs, or TIGRFAMs). We considered a functional module to be present in a pair MGS/sample if at least 90% of its components were present in the genes of the MGS and detected in the sample. Finally, we measured the potential of a module in a sample by summing abundances of all MGS found to carry this module in the sample.

Software pipeline for drug-aware univariate biomarker analysis

To assess to what extent observed differences between SZC and HC subjects in microbiome feature abundance are confounded, in the sense of, being consequences of other (treatment or risk factor) variables different between the groups more so than characteristic of SZC itself, we additionally employed the post-hoc filtering approach implemented in the R package metadeconfoundR (20) that was devised within the MetaCardis consortium (21). It functions in two steps. In the first, all associations between -omics features and the set of independent variables (disease status, drug treatment status, and risk markers including age, smoking status and BMI) are determined under nonparametric statistics (MWU or Spearman tests, adjusted for multiple -omics features tested using the Benjamini-Hochberg method). For each feature significantly (FDR < 0.1) associated with disease status (SZC vs HC), it is checked whether it has significant associations with any potential confounder. If not, it is considered trivially unconfounded (NC - Not Confounded). If at least one covariate also has significant association with the feature, then for each such covariate a post-hoc test for confounding is

applied. This test takes the form of a nested linear model comparisons (likelihood ratio test for P-values), where the dependent variable is the feature (X), and the independent variables are the disease status (A) and the tested covariate (B) versus a model containing only the covariate (B), thus testing whether disease status (A) adds explanatory power beyond the covariate (B). If this holds (LRT $P < 0.05$) for all covariates (B), then disease status is strictly deconfounded (SD) with regards to its effect on feature X; it cannot be reduced to any confounding factor. For each covariate (B) where significance is lost, a complementary modelling test is performed of the complementary model pairs - predicting X as a function of (A) and (B) versus a model containing (A) alone, thus testing whether the covariate (B) in turn is equally reducible to (A). If for at least one such covariate (B), (B) has independent effect (LRT $P < 0.05$) on top of (A), then the feature X is considered confounded by (B). However, if in none of the pairwise tests, the original significance holds, then (A) and (B) are considered so correlated that their relative influence cannot be disentangled. We consider these cases laxly deconfounded (LD), in the sense that for these cases clear confounding influence cannot be concluded, but also not ruled out. The R package was applied to the present dataset considering medication status either as binary variables or as normalized dosages.

Supplemental References

1. Dore, J., Ehrlich, S.D., Levenez, F., Pelletier, E., Alberti, A., Bertrand, L., Bork, P., Costea, P.I., Sunagawa, S., Guarner F, Manichanh, C., Santiago, A., Zhao, L., Shen, J., Zhang, C., Versalovic, J., Luna, R.A., Petrosino, J., Yang, H., Li, S., Wang J, Allen-Vercoe, E., Gloor, G., Singh B (2015): International Human Microbiome Standards. Retrieved from http://www.human-microbiome.org/

2. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, *et al.* (2013): Richness of human gut microbiome correlates with metabolic markers. *Nature* 500: 541–546.

3. Pons N, Batto J-M, Kennedy S, Almeida M, Boumezbeur F, Moumen B, *et al.* (2010): METEOR -a plateform for quantitative metagenomic profiling of complex ecosystems. Retrieved from https://forgemia.inra.fr/metagenopolis/meteor

4. Criscuolo A, Brisse S (2013): AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* 102: 500–506.

5. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, *et al.* (2017): Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol* 18: 142.

6. Plaza Onate F, Pons N, Gauthier F, Almeida M, Ehrlich SD, Le Chatelier E (2021): Updated Metagenomic Species Pan-genomes (MSPs) of the human gastrointestinal microbiota. Portail Data INRAE. https://doi.org/10.15454/FLANUP

7. Langmead B, Salzberg S (2013): Bowtie2. *Nat Methods* 9: 357–359.

8. Zhu F, Ju Y, Wang W, Wang Q, Guo R, Ma Q, *et al.* (2020): Metagenome-wide association of gut microbiome features for schizophrenia. *Nat Commun* 11: 1612.

9. Le Chatelier E, Almeida M, Plaza Oñate F, Pons N, Gauthier F, Ghozlane A, *et al.* (2021): A catalog of genes and species of the human oral microbiota. Portail Data INRAE. https://doi.org/10.15454/WQ4UTV

10. Plaza Oñate F, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, *et al.* (2019): MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* 35: 1544–1552.

11. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, *et al.* (2019): Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 47: D23–D28.

12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990): Altschul_1990_5424.pdf. *Journal of Molecular Biology*, vol. 215. pp 403–410.

13. Qi M, Wang R, Jing B, Jian F, Ning C, Zhang L (2016): Prevalence and multilocus genotyping of Cryptosporidium andersoni in dairy cattle and He cattle in Xinjiang, China. *Infect Genet Evol* 44: 313–317.

14. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, *et al.* (2016): EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44: D286–D293.

15. Haft DH (2001): TIGRFAMs: a protein family resource for the functional identification of

proteins. *Nucleic Acids Res* 29: 41–43.

16. Buchfink B, Xie C, Huson DH (2015): Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12: 59–60.

17. Eddy S (1998): HMMER user's guide: biological sequence analysis using prole hidden Markov models.

18. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, *et al.* (2016): Species-function relationships shape ecological properties of the human gut microbiome. *Nat Microbiol* 1: 16088.

19. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, *et al.* (2019): The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat Microbiol* 4: 623–632.

20. TillBirkner (2021): TillBirkner/metadeconfoundR: MetadeconfoundR Release for Documentation of the MetaDrugs Analysis as Part of the MetaCardis Consortium. Zenodo. Retrieved from https://github.com/TillBirkner/metadeconfoundR_V0.1.5_doc

21. Forslund SK, Chakaroun R, Zimmermann-Kogadeeva M, Markó L, Aron-Wisnewsky J, Nielsen T, *et al.* (2021): Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature*. https://doi.org/10.1038/s41586-021-04177-9