Supplementary Materials

Legend of the Supplementary Figures

Supplementary Figure 1. Heat map reporting the Z-score normalized weekly consumption of foods and drinks in each dietary group. The row order is the result of hierarchical clustering analysis. P-value was calculated by the Kruskal-Wallis method. ***p < 0.001; **p < 0.01; *p < 0.05. **B.** Box and violin plots reporting the estimated daily intake of specific nutrients in each dietary group. The intake was normalized on the total daily Kilocalories.

Supplementary Figure 2. A. Violin plots reporting the normalized expression level of three stool DEmiRNAs whose expression is significantly correlated with the years of non-omnivorous diet. The expression of these miRNAs is reported by dividing the subjects based on years of non-omnivorous diets (top) or based on age quartiles (bottom). P-value by Kruskal-Wallis test. **B-C.** Heat maps reporting the clustering analysis of the Spearman correlation results between stool miRNAs expression level and the estimated nutrient intake (**B**) or the detected microbial species (**C**). Only pairs of miRNA and nutrient/microbial species associated with an absolute rho greater than 0.3 in at least one comparison are reported.

Supplementary Figure 3. A. Bar plot showing the number of Human microRNA Disease Database (HMDD) annotations divided in four main disease classes. The color-code indicates whether a miRNA was up-regulated or down-regulated in the disease condition. On the right, the bar represents the log2FC computed in comparison between VN or VT with respect to omnivores (O) subjects. Asterisks represent the adjusted p-value from DESeq2 analysis. ***adj. p < 0.001; **adj. p < 0.01; *adj. p < 0.05. **B.** Scatter plot reporting the correlation between the median expression levels of the 49 stool DEmiRNAs considering the two groups of O subjects. **C.** Heat map reporting the log2FC expression of the comparisons VN and VT subjects with respect to the two groups of O subjects considered, respectively, for the discovery (disc.) and the validation (val.) analysis.

Supplementary Figure 4. A. Box plot reporting the relative abundances of bacterial species whose levels are significantly different in stool samples of subjects characterized by high or lower expression of miR-425-3p (top) and miR-638 (bottom).

Legend and figshare reference of the Supplementary Tables

Supplementary Table 1 (https://doi.org/10.6084/m9.figshare.14587596.v1). Alignment statistics and information of the small RNA-Seq data in stool (**A**) and plasma (**B**) samples and alignment statistics and information of the stool shotgun metagenomic data (**C**).

Supplementary Table 2. (https://doi.org/10.6084/m9.figshare.14587581.v2). Dietary and lifestyle information of the analysed subjects. For continuous covariates, the average and standard deviation are reported.

Supplementary Table 3. (https://doi.org/10.6084/m9.figshare.14587611.v2)**. A**) Expression levels and differential expression analysis of stool miRNAs; **B)** Results of the GLM analysis modeling stool DEmiRNA levels as a function of different subject covariates. **C**) Results of the differential expression analysis of the 49 stool DEmiRNAs considering different covariates in the DESeq2 model; **D**) Results of a correlation analysis between stool DEmiRNAs expression and subject age and years of non-omnivorous diet; **E**) HMDD annotations for the hairpin of stool DEmiRNAs; **F**) Differential expression analysis of stool DEmiRNAs considering an independent group of O subjects (O validation).

Supplementary Table 4. (https://doi.org/10.6084/m9.figshare.14587608.v2). **A**) Stool miRNAs belonging to specific correlation clusters of nutrients or bacterial species as reported in Supplementary Figure 2B-C. **B**) List of miRNA-nutrient associations supported by a significant Spearman correlation (adj. $p < 0.05$). Information of a GLM model computed for each pair is reported. Models adjusted by sex, age, and BMI.

Supplementary Table 5. (https://doi.org/10.6084/m9.figshare.14587599.v2). **A**) Expression levels and differential expression analysis of plasma miRNAs. **B**) Plasma levels of the DEmiRNAs identified in the analysis of stool samples. **C**) Correlation analysis between miRNA levels in stool and plasma samples.

Supplementary Table 6. (https://doi.org/10.6084/m9.figshare.14587584.v2). Analysis of xenomiR expression levels in stool (**A**) and in plasma samples (**B**). **C**) Set of RNA sequences

aligned with gma-MIR6300 using BLAST. **D**) Homology analysis between detected xenomiRs and human miRNAs

Supplementary Table 7. (https://doi.org/10.6084/m9.figshare.14587602.v1). **A**) List of functional processes enriched in targets of different groups of stool miRNAs. **B**) Validated targets of stool miRNA analysed in the functional enrichment analysis.

Supplementary Table 8. (https://doi.org/10.6084/m9.figshare.14818062.v2). **A)** Normalized expression levels of the stool DEmiRNAs in data from the TissueAtlas. The Z-score-transformed expression levels are also reported at the bottom. **B)** Epigenetic states computed at miRNA loci considering the data from the Roadmap Epigenomics Project. AT, active TSS; BT, Bivalent/Poised TSS; E, Enhancer; Q, Quiescent region; R, Repressed region; ST, Strong transcription; WT, Weak transcription; ZNF, Zinc-finger genes and repeated region.

Supplementary Methods

Study population recruitment and dietary information

A cohort of 141 individuals categorized as vegetarians (VT), vegans (VN), and omnivores (O) was recruited on a voluntary basis in Turin (Italy) between May 2017 and July 2019. Of the all potentially eligible subjects, 120 (72 women and 48 men) were included in the study because covering all the following selection criteria: VN, VT, and O dietary regime followed for more than one year, no consumption of antibiotics in the previous three months from sampling, no use of anti-inflammatory drugs in the month prior the sampling, no evidence of intestinal pathologies (Crohn's disease, chronic ulcerative colitis, bacterial overgrowth syndrome, constipation, celiac disease, Irritable Bowel Syndrome), and other pathologies (type I or type II diabetes, cardiovascular or cerebrovascular diseases, cancer, neurodegenerative disease, rheumatoid arthritis, allergies), no pregnancy and lactation. The study cohort included the same proportion of VN, VT, and O, matched for sex and age (**Figure 1A**). Three subjects made use of anti-hypertensive drugs and only one made use of a cholesterol-lowering medication. Anthropometric data (to calculate BMI) and abdominal circumference were collected for all subjects at the time of sampling. For each subject, the physical activity index was calculated based on the different sources of recreational, household, and occupational activities self-reported in the questionnaires, according to the EPIC study guidelines (https://dapa-toolkit.mrc.ac.uk/instrument/40). Each subject was

categorized into one of the four following groups: inactive, moderately inactive, moderately active, and active.

For an independent group of unmatched 45 omnivorous healthy individuals (31 women and 14 men), stool small RNA-Seq data (performed in the same laboratory and with the same protocol/pipeline of analysis) were also available for miRNA profiles validation.

The VT and VN volunteers were recruited with the collaboration of the Italian Society of Vegetarian Nutrition (http://www.scienzavegetariana.it/) or, like the group of O, among personal contacts and by a distribution of an informative leaflet. All participants signed informed written consent. The design of the study, the informed consent and protocols were approved by the local Ethics Committee on Colorectal_miRNA CEC2014 (Azienda Ospedaliera "SS. Antonio e Biagio e C. Arrigo" of Alessandria).

All recruited subjects filled in a validated self-administered food-frequency questionnaire (FFQ) assessing the usual diet, together with lifestyle and personal history data, in accordance with the European Prospective Investigation into Cancer and Nutrition EPIC study standards [1, 2].

The FFQ consisted of 248 questions concerning 188 different food items and included photos with two or three sample dishes of definite sizes or references to standard portion sizes. The composition in nutrients of individual food items was obtained from Italian food composition tables [3] and the average intake of macro and micronutrients for each volunteer was estimated.

The questionnaires were filled in on the web platform AcQUE, developed under the Node.js and Angular frameworks. The application uses modern, accessible, and usable web components to simplify the insertion of answers. It also supports the correct management of conditional questions, allowing the subjects to focus exclusively on questions of interest and making the compilation more fluid and less prone to errors. Finally, data can be exportable in a .csv file.

Collection of biological samples

Naturally evacuated fecal samples were obtained from all volunteers previously instructed to self-collect the specimen at home. Stool samples were collected in the nucleic acid collection and transport tubes with RNA/DNA stabilizing solution (Norgen Biotek Corp). Aliquots (200µl) of stool samples were stored at -80°C until RNA and DNA extraction [4].

Plasma and serum samples were collected according to standard phlebotomy procedures when volunteers brought the stool samples to the laboratory. Blood samples were collected

into ethylenediaminetetraacetic acid and serum clot activator tubes and immediately placed on ice.

Plasma tubes were centrifuged at ×1000g for 10 min at room temperature; once separated from the rest of the blood, plasma was distributed in cryovial tubes. One tube was immediately used for RNA extraction while the other aliquots were labeled and stored at −80°C. The time between sample procurement and storage was less than three hrs.

Serum tubes were centrifuged at ×4000g for 10 min at room temperature: one aliquot of serum (500µl) was used for the analyses of vitamin B12 and ferritin (in collaboration with a local private laboratory).

Extraction of total RNA from stool and plasma

Total RNA from stool was extracted using the Stool Total RNA Purification Kit (Norgen Biotek Corp) as described in Tarallo et al. [5].

Total RNA was extracted from 200µl of plasma with the miRNeasy plasma/serum mini kit (Qiagen) using the QiaCube extractor (Qiagen) following the manufacturer's instructions and as described in [4, 6].

RNA concentration was quantified by Qubit(R) microRNA Assay Kit (Invitrogen) according to the MIQE guidelines (http://miqe.gene-quantification.info/).

Extraction of total DNA from stool

The DNA extraction was performed with the QIAamp DNA stool MiniKit (QIAGEN, Hilden, Germany) according to the instructions of the manufacturer and as described in [5, 7]. The DNA quantification was performed with Qubit fluorometer (Qubit(R) DNA HS Assay Kit; Invitrogen).

Small RNA-sequencing (sRNA-Seq) and bioinformatic analyses

Library preparation for small RNA (sRNA) transcripts was performed with the NEBNext Multiplex Small RNA Library Prep Set for Illumina (Protocol E7330, New England BioLabs Inc., USA; New England BioLabs Inc., USA) as described in [4]. For each sample, 250 ng of RNA was used as starting material to prepare libraries. Each library was prepared with a unique indexed primer so that the libraries could all be pooled into one sequencing lane.

The obtained sequence libraries (24-samples multiplexed) were subjected to the Illumina sequencing pipeline, passing through clonal cluster generation on a single-read flow cell (Illumina Inc., USA) by bridge amplification on the cBot (TruSeq SR Cluster Kit v3-cBOT-

HS, Illumina Inc., USA) and 50 cycles sequencing-by-synthesis on the HiSeq 2000 (Illumina Inc., USA) (in collaboration with Genecore Facility at EMBL, Heidelberg, Germany).

Small RNA-sequencing (sRNA-Seq) analyses were performed using a previously described approach [5, 8]. Fastq files were quality checked using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads shorter than 14 nt were discarded. The reads passing the quality control were clipped from the adapter sequences using Cutadapt v1.10 by imposing a maximum error rate in terms of mismatches, insertions, and deletions equal to 0.15. Trimmed reads were mapped against human miRNA sequences from miRBase v22 [9]. The alignment was performed using the BWA algorithm v0.7.12 [10] with default settings. The number of reads obtained for each sample and the alignment statistics are reported in **Supplementary Table 1**.

Human miRNAs were annotated and quantified using two methods called the "knowledge-based" and "position-based" methods, as described in Tarallo et al. [5]. miRNAs whose assigned arms were derived from the "position-based" methodology were indicated in italic.

The results generated by the annotation and quantification methods were merged into a unique mature miRNA count matrix. In the case of mature miRNAs processed from different precursors but characterized by the same mature sequence, the read counts were summed. This procedure was applied following a manual analysis of the miRNA sequences annotated in miRBase.

Differential expression analysis was performed with DESeq2 R package v1.22.2 [11] using the likelihood ratio test (LRT) function. This function was selected to correct the analysis, including age and sex as covariates. The miRNAs without any reads in all the samples were excluded from the analysis. A miRNA was considered as detected in a specific dietary group if associated with a median number of normalized reads higher than 15. A miRNA was defined as differentially expressed (DEmiRNA) if associated with a Benjamini - Hochberg (BH)-adjusted p-value < 0.05 and supported by at least a median number of reads higher than 15 within at least one of the sample groups considered. Graphical representation of the analysis results was performed using the R packages UpSetR v1.4.0 [12], GGally v1.5.0, and ggplot2 v3.30.

The set of DEmiRNAs was overlapped with the annotation from Human microRNA Disease Database (HMDD) v3.2 [13], considering the annotations associated with each miRNA precursor. Among the database annotations, those related to gastrointestinal cancers, metabolic disorders, immune-related, or cardiovascular diseases were considered and grouped together.

Shotgun metagenomic and bioinformatic analyses

Libraries were prepared using the Nextera DNA Library Preparation kit (Illumina) and sequenced on an Illumina HiSeq platform as described in [7].

Host contamination was removed using the human sequence removal procedure from the Human Microbiome Project [14]. Raw reads were quality-trimmed (Phred score < 25) and reads shorter than 60 bp were discarded with the SolexaQA++ software [15]. The number of reads obtained for each sample is reported in **Supplementary Table 1**. Taxonomic profiling was carried out by using MetaPhlAn3 [16]. Functional profiles were obtained by HUMAnN 3.0 [17].

Statistical analyses

Covariate analysis. The analysis of the dietary and lifestyle covariates among subject groups was performed using the Chi-square test for categorical variables. For continuous covariates, Kruskal-Wallis and Wilcoxon Rank-Sum tests were performed. Clustering of weekly food/drink intake of each subject was performed using the hierarchical clustering function implemented in the ComplexHeatmap v2.0.0 R package [18]. The analysis was performed on Z-score-transformed data and using the Ward.D2 clustering method. Radar plot representation of the median weekly food/drink intake was performed by normalizing the median intake measured in each group by the maximum median value among the groups.

Correlation and regression analysis. Correlation analyses between DEmiRNA expression and food/drink or nutrient intake were performed using the Spearman method implemented in the cor R function. Multiple testing correction of the correlation p-values were performed using the BH method. Only correlations associated with an adjusted p-value<0.05 were considered as statistically significant.

The Generalized Linear Regression Model (GLM) was computed using the glm R function to assess the relevance of subject covariates in explaining the observed stool DEmiRNA levels. DEmiRNA expression levels were considered as the dependent variable of the model, while subjects age, sex, BMI, physical activity, waist circumference, and season of sample collection were considered as the independent variables of the model. A covariate was considered related to DEmiRNA expression if associated with a p-value <0.05. The relationship between DEmiRNA expression and the estimated intake of nutrients was evaluated computing a GLM adjusted for age, sex, and BMI, and estimated nutrient daily intake were considered as the independent variables of the model. The significance of the

GLM model was computed with respect to a null model using F-test implemented in the anova R function. Models associated with a p-value < 0.05 were considered statistically significant. The contribution of a nutrient to the model was considered significant if the associated p-value was lower than 0.05 and lower than those computed for the contribution to the model of sex, age, and BMI.

The graphical representation of the significant miRNA-nutrient interactions was performed using Cytoscape v3.8.0 [19]. Specifically, DEmiRNAs and nutrients were represented as network nodes with a node size proportional to the node degree. The network edges were colored based on the values of the Spearman rho, and their thickness was represented as proportional to the absolute correlation value.

Integration between miRNA expression, nutrient intake, and taxa

The R package mixOmics was employed for the integration of the three datasets (taxonomic profiles, miRNA expression profiles and dietary information), using the DIABLO model (Data Integration Analysis for Biomarker discovery using Latent cOmponents [20]. Datasets were integrated after normalization (*scale* R function) and removal of near-zero variables (*nearZeroVar* function of R package *caret*). Classification accuracy of the features identified by the DIABLO analysis was performed using a 10-fold cross-validation approach with the Random Forest classifier implemented in Weka v3.8.5 (https://www.cs.waikato.ac.nz/ml/weka/). The classification efficiency was evaluated in terms of Area Under the receiver operating characteristic Curve (AUC).

miRNA targets functional enrichment analysis

Functional enrichment analysis of miRNA target genes was performed using RBiomirGS v0.2.12 [21] with the miRNA-target gene annotations obtained from miRTarbase v7.0 [22] and miRecords [23]. The analysis was performed on gene sets retrieved from the Molecular Signatures Database v7.1 [24] and using: (i) DEmiRNAs coherently differentially expressed in VN and VT subjects with respect to O; (ii) miRNAs significantly correlated with specific nutrients; (iii); miRNAs significantly correlated with taxa; (iv) DEmiRNAs from the DIABLO analysis. The log2FC and adjusted p-value computed between the VN and O diets was used as input for the tool. The gene sets characterized by an adjusted p-value lower than 0.05 were considered as significantly enriched.

miRNA expression and epigenetic state analysis in intestinal tissues

The analysis of miRNA expression in intestinal tissue samples was performed considering the data from the TissueAtlas [25], including data of two colon and three small intestine samples. A miRNA was considered expressed in a sample if associated with a normalized expression level greater than 10. Each miRNA expression level was also converted to Z-score considering the mean and standard deviation computed across the 61 tissue samples of the TissueAtlas. The epigenetic status of miRNA coding genes was evaluated using the chromatin status defined by the Roadmap Epigenomics Project [26]. Specifically, the analysis was performed considering the 15 statuses (Core 15-state model) defined by integrating epigenetic data of duodenum, small intestine, colon, rectum mucosa, or colon smooth muscle samples. Only mature miRNAs with univocal genomic loci were considered for the analysis.

References

1. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondiere UR, Hemon B, Casagrande C, Vignat J *et al*: **European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection**. *Public Health Nutr* 2002, **5**(6B):1113-1124.
2. Pala V, Sieri S, Palli D, Salvini S, Berrino F, Bellegotti M, Frasca G, Tumino R, Sacerdote C, Fiorini L *et al*: **Diet in the Italian EPIC cohorts: presentation of data and methodological issues**. *Tumori* 2003, **89**(6):594-607.
3. Salvini S PM, Gnagnarella P, Maisonneuve P, Turrini A.: **Banca Dati di Composizione degli Alimenti per Studi Epidemiologici in Italia**. 1998.
4. Ferrero G, Cordero F, Tarallo S, Arigoni M, Riccardo F, Gallo G, Ronco G, Allasia M, Kulkarni N, Matullo G *et al*: **Small non-coding RNA profiling in human biofluids and surrogate tissues from healthy individuals: description of the diverse and most represented species**. *Oncotarget* 2018, **9**(3):3097-3111.
5. Tarallo S, Ferrero G, Gallo G, Francavilla A, Clerico G, Realis Luc A, Manghi P, Thomas AM, Vineis P, Segata N *et al*: **Altered Fecal Small RNA Profiles in Colorectal Cancer Reflect Gut Microbiome Composition in Stool Samples**. *mSystems* 2019, **4**(5).
6. Sabo AA, Birolo G, Naccarati A, Dragomir MP, Aneli S, Allione A, Oderda M, Allasia M, Gontero P, Sacerdote C *et al*: **Small Non-Coding RNA Profiling in Plasma Extracellular Vesicles of Bladder Cancer Patients by Next-Generation Sequencing: Expression Levels of miR-126-3p and piR-5936 Increase with Higher Histologic Grades**. *Cancers (Basel)* 2020, **12**(6).
7. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C *et al*: **Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation**. *Nat Med* 2019, **25**(4):667-678.
8. Kulkarni N, Alessandri L, Panero R, Arigoni M, Olivero M, Ferrero G, Cordero F, Beccuti M, Calogero RA: **Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines**. *BMC Bioinformatics* 2018, **19**(Suppl 10):349.

9. Kozomara A, Birgaoanu M, Griffiths-Jones S: **miRBase: from microRNA sequences to function**. *Nucleic Acids Res* 2019, **47**(D1):D155-D162.
10. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
11. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
12. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H: **UpSet: Visualization of Intersecting Sets**. *IEEE Trans Vis Comput Graph* 2014, **20**(12):1983-1992.
13. Huang Z, Liu L, Gao Y, Shi J, Cui Q, Li J, Zhou Y: **Benchmark of computational methods for predicting microRNA-disease associations**. *Genome Biol* 2019, **20**(1):202.
14. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project**. *Nature* 2007, **449**(7164):804-810.
15. Cox MP, Peterson DA, Biggs PJ: **SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data**. *BMC Bioinformatics* 2010, **11**:485.
16. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N: **MetaPhlAn2 for enhanced metagenomic taxonomic profiling**. *Nat Methods* 2015, **12**(10):902-903.
17. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N *et al*: **Species-level functional profiling of metagenomes and metatranscriptomes**. *Nat Methods* 2018, **15**(11):962-968.
18. Gu Z, Eils R, Schlesner M: **Complex heatmaps reveal patterns and correlations in multidimensional genomic data**. *Bioinformatics* 2016, **32**(18):2847-2849.
19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.
20. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Le Cao KA: **DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays**. *Bioinformatics* 2019, **35**(17):3055-3062.
21. Zhang J, Storey KB: **RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration**. *PeerJ* 2018, **6**:e4262.
22. Chou CH, Fu TC, Tsai HH, Hsu CC, Wang CH, Wang JS: **High-intensity interval training enhances mitochondrial bioenergetics of platelets in patients with heart failure**. *Int J Cardiol* 2019, **274**:214-220.
23. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T: **miRecords: an integrated resource for microRNA-target interactions**. *Nucleic Acids Res* 2009, **37**(Database issue):D105-110.
24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
25. Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C*, et al.* **Distribution of miRNA expression across human tissues.** *Nucleic Acids Res* 2016;**44**:3865-77.
26. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A*, et al.* **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015;**518**:317-30.