nature metabolism

Article

https://doi.org/10.1038/s42255-025-01220-1

Metagenomic estimation of dietary intake from human stool

Received: 14 February 2024

Accepted: 16 January 2025

Published online: 18 February 2025

Check for updates

Christian Diener[1,2] ✉, Hannah D. Holscher[3], Klara Filek[1], Karen D. Corbin[4], Christine Moissl-Eichinger[1,5] & Sean M. Gibbons[2,6,7,8] ✉

Dietary intake is tightly coupled to gut microbiota composition, human metabolism and the incidence of virtually all major chronic diseases. Dietary and nutrient intake are usually assessed using self-reporting methods, including dietary questionnaires and food records, which suffer from reporting biases and require strong compliance from study participants. Here, we present Metagenomic Estimation of Dietary Intake (MEDI): a method for quantifying food-derived DNA in human faecal metagenomes. We show that DNA-containing food components can be reliably detected in stool-derived metagenomic data, even when present at low abundances (more than ten reads). We show how MEDI dietary intake profiles can be converted into detailed metabolic representations of nutrient intake. MEDI identifies the onset of solid food consumption in infants, shows significant agreement with food frequency questionnaire responses in an adult population and shows agreement with food and nutrient intake in two controlled-feeding studies. Finally, we identify specific dietary features associated with metabolic syndrome in a large clinical cohort without dietary records, providing a proof-of-concept for detailed tracking of individual-specific, health-relevant dietary patterns without the need for questionnaires.

Dietary intake and nutrition are key determinants of human growth and development, metabolic health and chronic disease risk[1–3]. Diet also shapes the composition of the human gut microbiota[4], and in turn, the effects of diet on the host can be influenced by the ecology of the gut[5]. Across an individual's lifespan, dietary intake patterns can either alleviate or exacerbate a wide range of disease conditions, including cardiovascular disease, diabetes, Alzheimer's disease and cancer[6–9].

Accurate tracking of dietary intake, including the quantification of dietary metabolites, nutrients and energy content, is critical to understanding phenotypic heterogeneity in human cohort studies. However, diet tracking is often hampered by challenges in obtaining high-quality, unbiased data[10]. In many cross-sectional studies, individual dietary intake data are obtained from questionnaires, which need to strike a balance between granularity and ease of use. The most common methodology used for assessing dietary intake, the food frequency questionnaire (FFQ), asks participants for coarse-grain information on habitual dietary patterns for a set of common food categories[11]. More detailed information can be obtained from dietary records, in which study participants fill out food diaries, sometimes with clinician support, for a pre-specified time frame[12]. Dietary records provide more detailed quantification of food intake, including time of day and cooking methods, and can be used to generate more detailed mappings to macronutrient and micronutrient intake; however, these methods are often dependent on proprietary platforms for nutrient analysis and

[1]Diagnostic and Research Institute of Hygiene, Microbiology and Environmental Medicine, Medical University of Graz, Graz, Austria. [2]Institute for Systems Biology, Seattle, WA, USA. [3]Department of Food Science and Human Nutrition, University of Illinois Urbana-Champaign, Urbana, IL, USA. [4]AdventHealth Translational Research Institute, Orlando, FL, USA. [5]BioTechMed Graz, Graz, Austria. [6]Department of Bioengineering, University of Washington, Seattle, WA, USA. [7]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [8]eScience Institute, University of Washington, Seattle, WA, USA. ✉e-mail: christian.diener@medunigraz.at; sgibbons@isbscience.org

strong participant compliance, making it difficult to compare dietary intake data across cohorts and to maintain participant compliance in longitudinal studies[13]. Standardized questionnaires, or common food databases, often do not reflect diverse human populations, leading to inequalities whereby the diets of minority and underrepresented populations are not accurately represented in existing questionnaires or recall surveys[14]. Finally, dietary questionnaires rely on the ability of study participants to accurately and without bias recall their own food intake, which is known to be fraught and has led to some debate on the utility of self-reported dietary data[15–17]. Therefore, there is a demand for approaches that can quantify dietary and nutrient intake patterns without the need for FFQs or recall surveys.

One survey-free technique for coarsely assessing diet quality is the quantification of major diet-influenced analytes in human plasma or serum. This approach is used widely in clinical settings, in which regular measurement of blood glucose, cholesterol and other lipid levels are the current standard of care[18,19]. However, these clinical chemistries represent a limited breadth of diet-relevant features, which could be expanded on by using targeted or untargeted metabolomics of blood, saliva, urine or faecal samples. Another approach has been to capture images of meals (for example, with a smartphone) and apply machine learning to these images to track dietary intake[20,21]. Image tracking and physical sensors have proven to be challenging approaches, requiring large training databases, showing a limited ability to estimate portion size and relying on a fairly high degree of participant compliance[22,23].

Molecular 'omics-based approaches to diet tracking have the potential to increase the sensitivity and resolution of intake assessments while reducing the compliance burden for participants. Owing to the complexities of absorption and metabolism of many of the compounds found in the diet, blood metabolomics currently provide limited information on food intake. To overcome this limitation, recent work has focussed on constructing curated databases of food-specific mass-spectrometry spectra that can be used to identify the presence of specific foods in faecal, urine or blood metabolomes[24,25]. These reference spectra can identify specific food components with high accuracy but provide limited information about the overall abundance of a food item. An alternative approach is to quantify dietary intake using residual food-derived DNA in stool. Prior studies have shown that plant intake patterns can be quantified in human stool by targeting plant-specific marker genes for amplicon sequencing[26,27]. Although effective, these targeted methods require additional sample processing before sequencing. Metagenomic shotgun sequencing (MGS) of faecal DNA, on the other hand, is a common data type in human microbiome research that could potentially be used to capture dietary intake information[28,29]. Leveraging MGS data directly for diet tracking would require no additional sample processing steps. However, there are several challenges to detecting food-derived DNA in MGS data, which has delayed the implementation of metagenomic-based diet tracking.

Faecal MGS data are commonly used to quantify the taxonomic and functional composition of bacterial, archaeal, viral and fungal communities in the gut, where genes can be identified with de novo methods[30,31]. However, although we expect food-derived DNA to be present in stool, the low frequency of these reads relative to microbial-derived and host-derived reads makes de novo gene prediction from these sequences intractable[32]. Alternatively, individual reads can be mapped to large databases of reference sequences, using efficient hashing schemes, to generate read-specific taxonomic annotations[33–35]. However, quantification of dietary intake through reference-based approaches is hampered by the increased genomic complexity of eukaryotes and by the lack of dedicated food genome databases. Furthermore, these reference mapping approaches are prone to false-positive assignments, requiring the inclusion of decoy genomes in the database from other organisms that are known to be present in the sample, like host-associated reads and reads coming from the microbiota[36,37]. Finally, even if we could quantify the taxonomic composition of food-associated reads, we currently lack the ability to automatically translate this taxonomic information into nutrient content.

Here, we aimed to overcome these limitations by building a comprehensive food genome database for annotating food-associated reads in human stool, along with high-resolution mappings between food items and dietary metabolite profiles. We paired the constructed database with a fast, scalable and decoy-aware mapping strategy and validated its performance using simulated and in vivo data from infants and adults, showing that we can reliably quantify certain dietary and nutrient intake components from faecal MGS data. Finally, we demonstrated that our MGS-based dietary assessments were strongly associated with variation in metabolic health in a large European cohort.

Results

Linking food genomes to nutrient information

Although databases that map food items to nutrient content exist, such as FoodData Central (FDC; https://fdc.nal.usda.gov) and FOODB (www.foodb.ca), none of these databases are directly linked to genomic data from the plants, animals and fungi present in the human diet. Most food databases contain both compound or mixture foods (foods containing several individual components, such as a pizza or a muffin) and single-organism foods (such as cucumber or chicken); only the latter can be uniquely mapped to a specific genome. Of the 992 foods present in FOODB, 619 represent such single-organism foods and can be mapped to the National Center for Biotechnology Information (NCBI) Taxonomy Database, and we aimed to obtain genomic assemblies for as many of these single-organism foods as possible. Food items were mapped through NCBI taxonomy IDs in a multi-tiered approach (Fig. 1a). Food items were first matched to RefSeq genomes at the species level and then at the genus level if no match could be found for the respective species, yielding a set of 459 foods mapping to 331 unique genome assemblies (Fig. 1b,c). This was followed by a search in the full NCBI Nucleotide Database at the species and genus level to obtain partial assemblies for food items without a full reference assembly. A total of 98 partial assemblies representing 102 additional foods could be identified in this way (Fig. 1b,c), resulting in a final database containing 429 genomes and genomic assemblies representing 561 food items and their associated strains (91% of all single-organism foods in the FOODB). The observed redundancy, in which several food items matched the same genomic assembly, was caused by the presence of either several strains of the same food species in the database (for example, the *Brassica oleracea* group) or different preparations derived from the same organism (for example, orange juice vs orange slices). Additionally, many of the compound foods in FOODB (for example, pizza) were covered by a combination of single-organism foods contained in the MEDI database (for example, tomato, wheat, cow, pig, etc.).

Macronutrient, energy and specific metabolite contents were retained separately for each food item and preparation type, even when matching the same taxon, to allow for manual selection of preparations or strains after mapping. The resulting database contained a total of 489 billion base pairs (bp) covering all major phyla of common food components, along with their nutrient and metabolite composition (Fig. 1d). The majority of genomic data came from Phylum Streptophyta, which includes most of the common plant-based foods, followed by Phylum Chordata, which comprises most of the animal-derived foods (Fig. 1b,d). Phylogenetic distance, quantified by average nucleotide identity, was associated with the relative protein and carbohydrate content of the food items (Fig. 1d; PERMANOVA $P = 0.001$, $R^2 = 0.06$ and 0.05, respectively), revealing that there is a weak phylogenetic association with macronutrient content. In particular, macronutrient composition varied only by 10–20 g per 100 g within 90–95% average nucleotide identity, suggesting that nutrient composition is similar for species within the same genus (Extended Data Fig. 1a).

https://doi.org/10.1038/s42255-025-01220-1

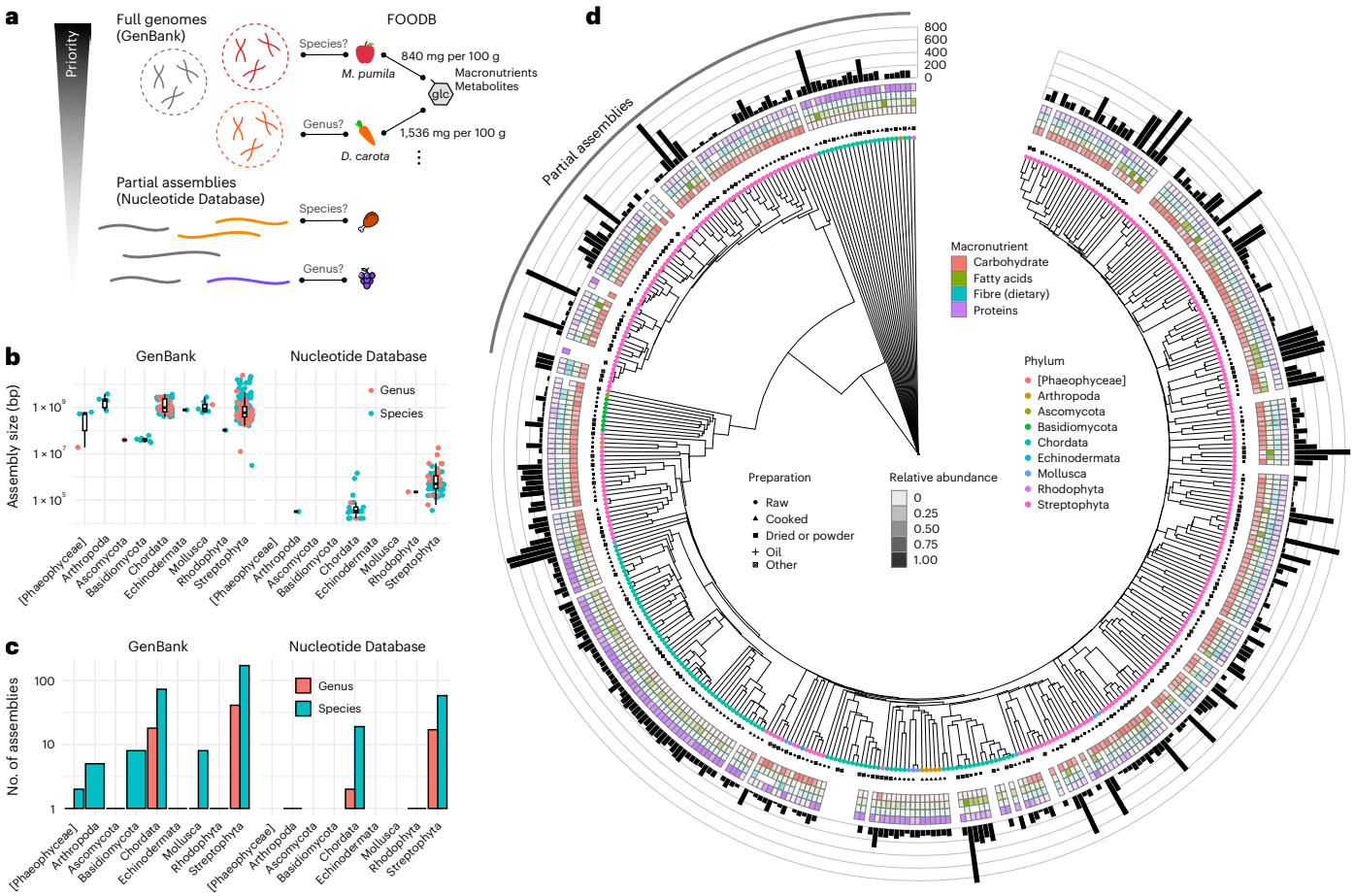


Fig. 1 | Constructing a metagenomic food database. a, Illustration of the search strategy used to map food items to assemblies and their connection to nutrient content. **b**, Assembly size for the identified food-related organisms. Titles denote the database yielding the hit (GenBank, complete genomes; Nucleotide Database, partial assemblies). Boxplots show 25%, 50% and 75% quantiles; the centre denotes the median and whiskers extend to the smallest and largest data points within 1.5 interquartile ranges. **c**, Number of food organisms matched and the respective taxonomic rank where the match was found. **d**, Phylogenetic tree of the identified food organism assemblies, generated using UPGMA on estimated average nucleotide identity (estimated using MASH). Coloured circles denote the phylum, symbols indicate the dominant (that is, the most common, least-processed in FOODB) food preparation type, filled rectangles show macronutrient composition per 100 g of biomass and black bars show the energy content of individual food-assembly pairings per 100 g of biomass.

Decoy-aware, efficient mapping of food-derived sequences

The size of our food genome database exceeded commonly used databases for the classification of bacteria, archaea and viral genomes by at least fivefold. Therefore, we developed a computationally efficient mapping strategy, which we termed MEDI. MEDI is based on the Kraken 2 mapping scheme, to ensure scalability to very large datasets (Extended Data Fig. 1b)[33]. Kraken 2 uses a fast *k*-mer hash to identify the least-common ancestor (LCA) for each single read, which can be paired with a Bayesian redistribution approach that passes read counts down the phylogenetic tree (Bracken)[38]. Given that the majority of genetic material in human-associated stool microbiome samples is probably from bacteria or the host, there is a high chance of false-positive identification of background DNA as food components. To avoid this issue, we opted for a decoy-aware approach, in which the *k*-mer hash also included additional background genomes belonging to bacteria, archaea, viruses, common plasmids and the human genome[39]. We combined this with an additional post-classification filtering step (before Bracken redistribution) that removed individual reads with inconsistent *k*-mer classification patterns that included taxa from distant clades (see Fig. 2a and Methods). The resulting abundance estimates were then used to quantify the food items present in a sample. MEDI food quantification is based on relative read abundances, without correcting for genome size, which is correlated with but not the same as the relative taxonomic abundance (number of genome copies) or relative biomass[40]. Correcting for genome size would be ideal, but most food-derived DNA is heavily degraded by the time it reaches the large intestine. In addition, the MEDI database combines partial assemblies and sequence matching at the genus rank, which obscures species-level genome size variation. Fortunately, for organisms with known genome size information, we observed only subtle correlation ($r = 0.04$) between genome size and relative read abundances in real-world datasets (Extended Data Fig. 1c). Therefore, we moved forward with using the uncorrected relative food abundances to derive the macronutrient and metabolic compound composition in a given sample (standardized to a 100 g portion of a mixture of the respective food items), whereby the relative abundance of a given food item was used as a proxy for relative biomass (see Methods). If multiple unique food items or preparation types matched a single genomic assembly, the average nutrient and metabolite profile was used. This provided an estimate of nutrient composition within a sample based solely on single-organism food DNA.

MEDI was tested on an artificial ground-truth dataset using simulated reads. In brief, we generated an average abundance profile of the decoy organisms present in faecal samples from a healthy population of 365 individuals from the integrative human microbiome project (iHMP), drawing subsamples from this background distribution to

https://doi.org/10.1038/s42255-025-01220-1

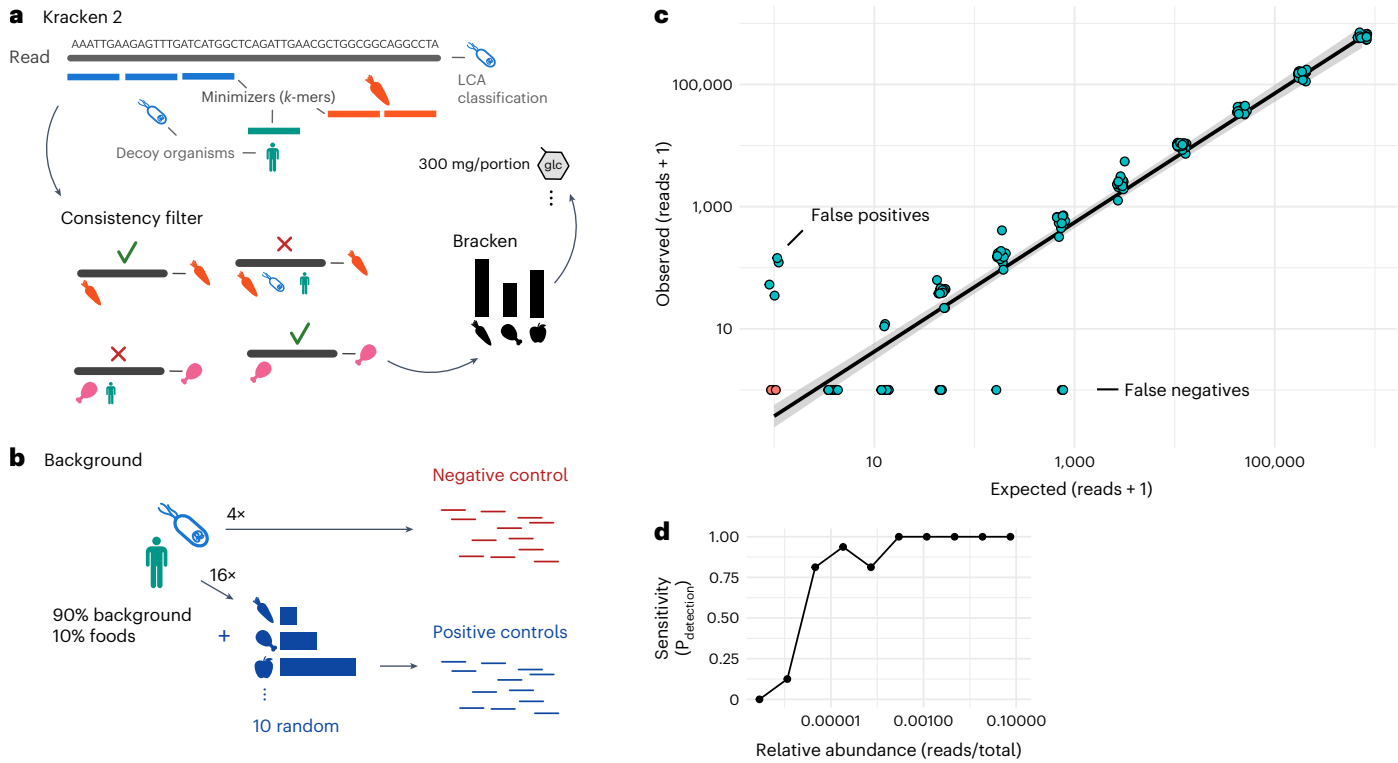


Fig. 2 | Food genome quantification on simulated ground-truth data. **a**, Illustration of the mapping and filtering strategy used by MEDI. Individual *k*-mer assignments (LCA classifications) were used to assign consistency scores to reads and to filter reads with discordant mappings. **b**, Sampling strategy for the ground-truth data. All samples contain at least 90% background of an average bacteria, archaea and host background. Positive samples contain simulated reads from ten random food assemblies with exponentially increasing abundances. **c**, Quantification performance across simulated negative and positive controls. Points denoting a detected food item in a single sample are slightly jittered on the *x* axis to resolve overlaps. The black line denotes a linear regression fit (mean relationship between ground truth and observed) and the grey area is the 95% confidence interval around that mean. Fill colour denotes negative (red) or positive samples (blue). False-positive organisms are generally connected to organisms within the same taxonomic family. **d**, Probability of detecting a true-positive food item in a sample as a function of relative food item abundance (that is, detection power).

simulate different decoy communities[28]. Positive-control samples (that is, samples containing a dietary signal) were generated by injecting 10% food reads from ten randomly chosen food items into each individual sample (Fig. 2b). Given the high prevalence (90% of total abundance) and richness of decoy organisms in the simulated dataset, one would expect methods that incorrectly map background taxa to foods to perform poorly. The ten reference foods added to each sample were logarithmically staggered in abundance within each sample, creating a relative abundance range of 0.00003–7.5% across food items. Four background samples without any food reads added were used as negative controls. This simulated dataset allowed us to assess the prevalence of true positives, false positives, true negatives and false negatives as well as to quantify the taxonomic specificity and detection limits of our approach. MEDI was able to identify and quantify diet-derived sequences in all simulated samples (Fig. 2c; mean $R^2 = 0.96$ (0.91–0.99), $P < 10 \times 10^{-6}$ for all positive samples). None of the ten million reads in each of the food-negative samples were classified as food-derived. The false-positive rate was slightly higher in the food-positive samples, in which we observed four false-positive classifications across all samples. Misidentified food items were generally from the same genus as a true-positive food item that was also present in a sample, which did not alter quantification accuracy for the true-positive food items (Fig. 2c). Despite the strong filtering to prevent false positives, MEDI was highly sensitive, providing >80% power for detecting a food item with a relative abundance as low as 0.001% (ten reads per million; Fig. 2d). In summary, MEDI was able to distinguish and quantify sequence reads from food items in simulated metagenomic samples, with a negligible rate of cross-domain mismatching from the gut microbiome or the host.

MEDI estimates correspond to data from controlled-feeding studies

MEDI estimates the abundance of food-derived DNA and uses this data to provide a prediction of dietary nutrient content from faecal metagenomes. To evaluate whether these estimates correspond to daily food intake patterns, we applied MEDI to metagenomic sequencing data from two controlled-feeding studies (Fig. 3a). Both studies were selected to have defined dietary intervention and supervised feeding to ensure that study participants consumed defined meals throughout the study duration. Thus, available dietary intake data corresponds to the ground truth of food intake in both studies.

In a previous study[41], termed the MBD study here, participants either consumed a prototypical Western diet or a microbiome enhancer diet (MBD) in a randomized crossover design ($n = 17$). The MBD was enriched in high-fibre and digestion-resistant foods that were more likely to survive passage into the large intestine[41]. MEDI estimates of food intake showed significant differences in beta diversity between diets (Fig. 3b; PERMANOVA $R^2 = 0.1$, $P = 0.007$). The relative abundance of metagenomic reads assigned to foods was almost sixfold higher in the MBD intervention than in the Western diet intervention (Fig. 3c; 0.035% vs 0.0062%, Mann–Whitney *U*-test $P = 0.0007$). Additionally, MEDI identified specific enrichment of several known components of the MBD diet relative to the Western diet, including flax seeds, quinoa, oats, spinach, rye, barley and strawberry (Extended Data Fig. 2a).

In a different study[42], termed the PATH study here, participants ($n = 48$) were provided daily meals that contained 90% of the same ingredients and had matched macronutrient content across study arms: the intervention group ($n = 28$), which received one large avocado

https://doi.org/10.1038/s42255-025-01220-1

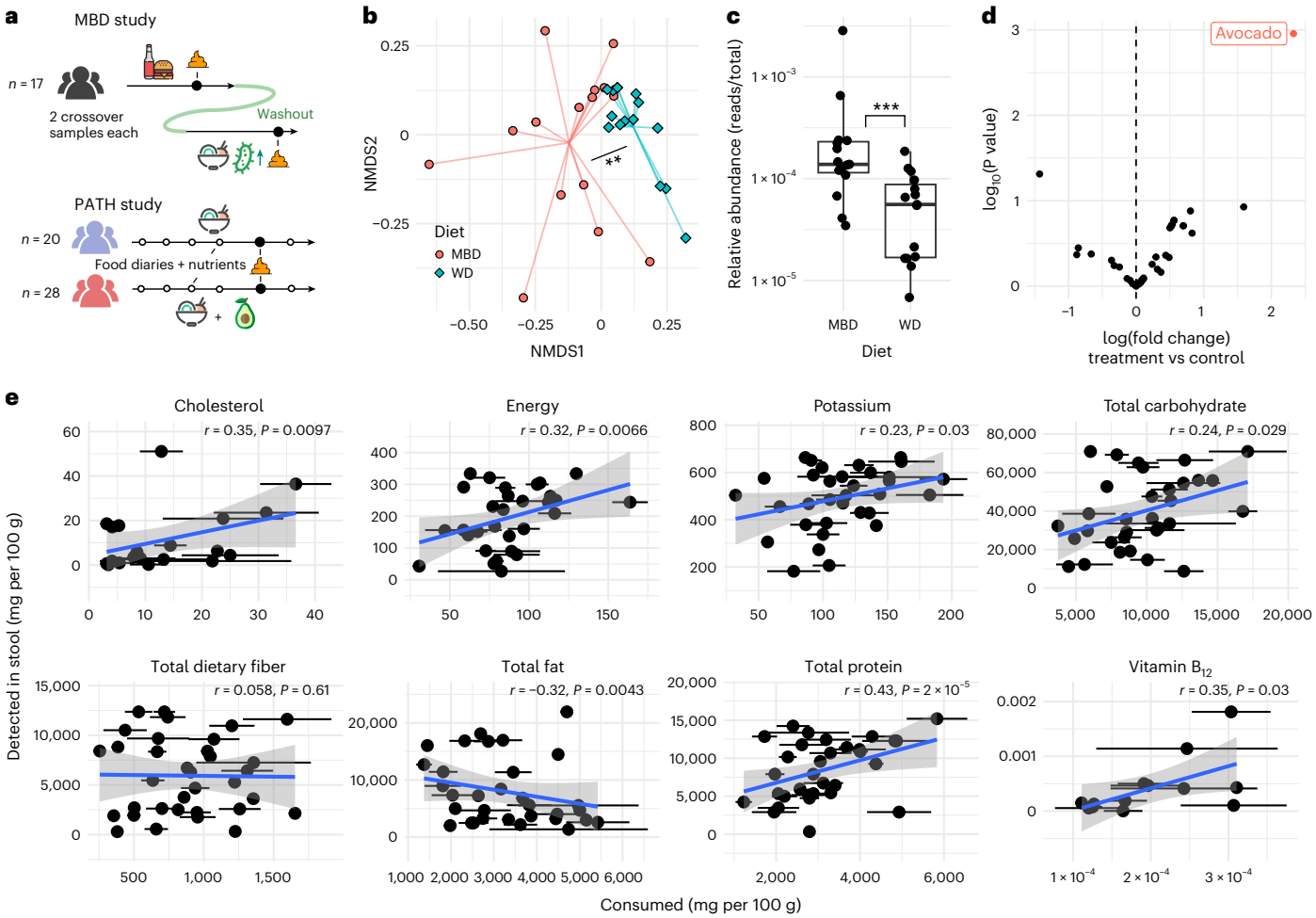


Fig. 3 | MEDI recapitulates data from controlled-feeding studies. a, Outline and cohort sizes of the controlled-feeding studies used. **b**, Non-metric multidimensional scaling of MEDI food abundance beta diversity (Bray–Curtis distance) for the MBD study ($n = 30$, only samples with detected food (30 out of 34)). Individual lines connect each sample with the group centroid. Colours denote diet group (WD, Western diet; MBD, microbiome enhancer diet). Asterisks denote significance from a PERMANOVA (**$P = 0.005$). **c**, Relative abundance of foods (food reads / total reads) for all samples with detected foods in the MBD study ($n = 30$ metagenomes from $n = 17$ individuals, each subjected to both diets). Boxplots show 25%, 50% and 75% quantiles; the centre denotes the median and whiskers extend to the smallest and largest data points within 1.5 interquartile ranges. Asterisks denote significance under a two-sided Mann–Whitney U-test (***$P = 0.0007$). **d**, Volcano plot for differential abundance analysis of food abundances in the PATH study. Each point denotes a food species detected by MEDI. Red colour denotes food item with an FDR-adjusted $P < 0.05$ limma-voom regression of read counts vs intervention group ($n = 48$). **e**, MEDI predictions from faecal DNA (y axis) and nutrient consumption obtained from food diaries (x axis) in a controlled-feeding study (PATH), in which the dietary intake recorded in the daily food record precedes the stool sample by at least 48 h. Each point denotes a single individual. For the food diaries, points represent means over all measured intake amounts; error bars, s.e.m. (s.d. / sqrt(n)), normalized to a 100 g portion (all samples within the offset, 38 individuals with 124 food record diary entries). For the MEDI data, x-coordinate points represent estimates of intake based on weighting nutrient profiles of food items by food item relative abundance and assuming a 100 g portion. Blue lines denote regression slopes and grey areas represent 95% confidence intervals. Annotations denote correlation r and P value from a two-sided Pearson product-moment correlation test.

daily, and the control group ($n = 20$), which received daily meals devoid of avocado[42]. Differential abundance analysis of the 73 food items detected by MEDI across samples in the study identified avocado as the sole food item that differed in abundance across the study groups (Fig. 3d; 2.3-fold change, false discovery rate (FDR)-corrected $q = 0.04$). The detailed daily food diaries from the PATH study also allowed us to compare MEDI estimates of nutrient composition in faecal samples to overall intake data. We evaluated energy content and major macronutrients (protein, carbohydrate, fat and fibre) as well as a set of micronutrients (potassium, vitamin B_{12}) and cholesterol. We only observed agreement between MEDI estimates and intake data when faecal samples were obtained 24–48 h after dietary intake (Extended Data Fig. 2b), which is consistent with previous estimates of an average transit time of 1–2 days[43]. Here, MEDI estimates agreed with food diary data for energy, protein, carbohydrate, potassium, cholesterol and vitamin B_{12} intake (Fig. 3e). No agreement was observed for total dietary fibre and total fat intake. However, MEDI-inferred fibre intake was significantly correlated with soluble fibre intake and the consumption of grains (Extended Data Fig. 2c). This disagreement in total fibre and fat content may be a result of biases introduced by food processing, whereby many processed foods are depleted in complex fibres (for example, white bread or white rice) that would be present in the whole food, and many refined fats (for example, vegetable oils) are depleted in source-organism DNA[44,45]. In particular, fat content estimated by MEDI was negatively correlated with consumed fat in the dietary record data ($r = -0.32$, $P = 0.0043$), suggesting a trend in which individuals who consume more fat from whole foods (for example, eating whole avocados or olives) consume less fats overall, with the higher-fat consumers deriving more fat from processed or refined foods (for example, avocado oil or olive oil). Overall, we see strong

https://doi.org/10.1038/s42255-025-01220-1

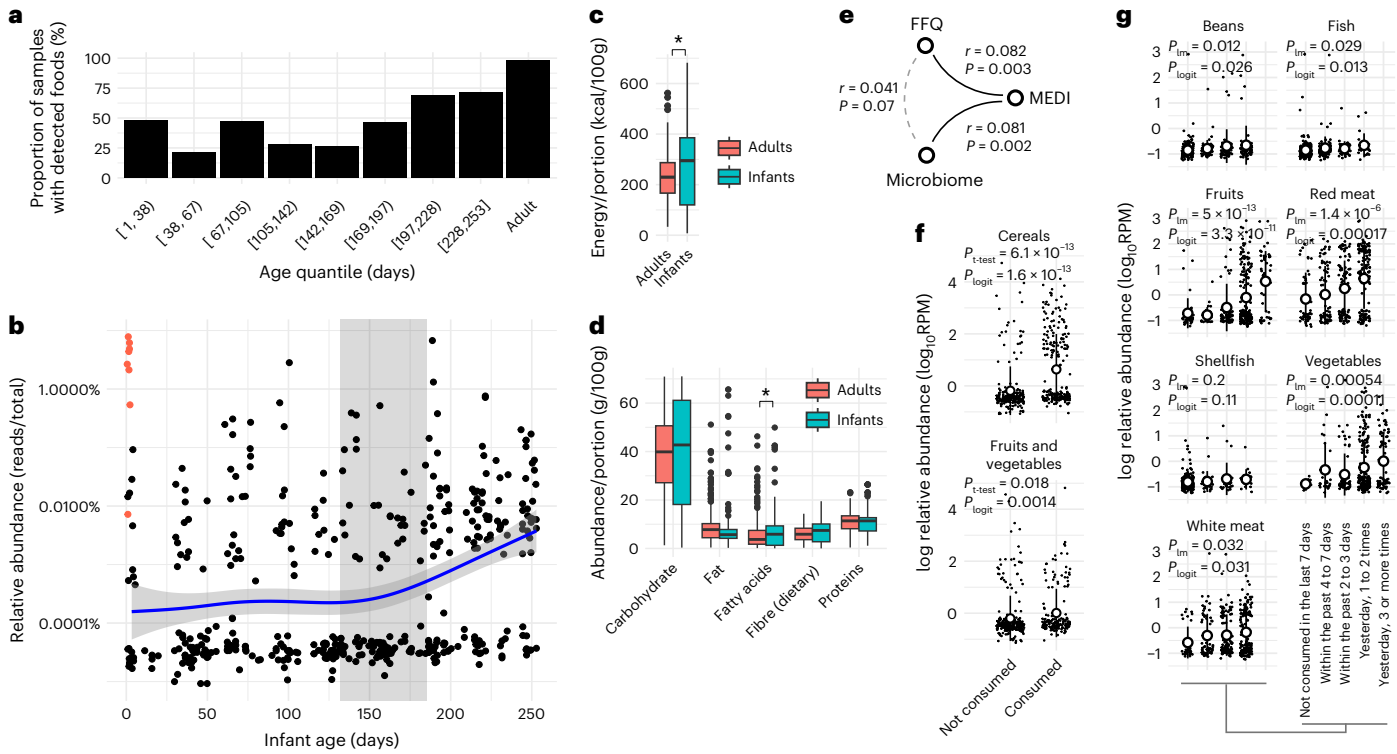


Fig. 4 | MEDI food abundances across infants and adults. a, Fraction of samples with at least one detected food read across different age groups. **b**, Relative abundance of food-derived reads in a cohort of 447 infants. The blue line denotes the smoothing spline of the observed reads; the light blue area denotes the 95% confidence interval of the mean spline curve. Orange dots denote samples with less than 95% overall abundance mapped to bacteria (that is, low bacterial biomass). Grey shaded area denotes the interquartile area of the onset of solid food intake across infants. **c**, Energy content per standardized portion size (100 g) per sample in adults and infants. Shown are only samples with detected food items ($n = 196$ for infants and $n = 359$ for adults). Asterisk denotes significance under a Welch t-test: $^*P = 0.024$. **d**, Macronutrient content per standardized portion size in infants and adults. Shown are only samples with detected food items ($n = 196$ for infants and $n = 359$ for adults). Asterisk denotes significance under a two-sided Welch t-test: $^*P = 0.015$. In **c** and **d**, boxplots show 25%, 50% and 75% quantiles; the centre denotes the median and whiskers extend to the smallest and largest data points within 1.5 interquartile ranges. **e**, One-sided Mantel permutation test statistics for beta diversity agreement between MEDI-predicted food abundances, FFQs and microbial species abundances (Bray–Curtis distances; see Methods). Correlation between pairwise distance measures is indicated by r; Mantel test P value is shown. **f**, Comparison of relative food group abundances with paired diet frequency questionnaire data from infants. RPM, reads per million. Circles denote the mean; error bars, s.d. ($n = 447$). $P_{t\text{-test}}$ indicates the P value of a two-sided Welch t-test of log-transformed relative abundances; P_{logit} denotes the P value of a logistic regression of food occurrence against food frequency strata. Axis labels are common across both plots in this panel. **g**, Comparison of MEDI-predicted relative food group abundances with diet frequency questionnaires in adults. Circles denote the mean; error bars, s.d. (only samples with paired FFQs, $n = 361$), P_{lm} indicates the ANOVA P value of a regression of log-transformed relative abundances; P_{logit} denotes the P value of a logistic regression of food occurrence against food frequency strata. Axis labels are common across all plots in this panel.

agreement between validated nutrient intake and MEDI estimates for a number of dietary features and poor agreement for others. As one might expect, MEDI appears to be optimal for detecting nutrient intake from whole foods that are resistant to digestion in the stomach and proximal gut and is sub-optimal for detecting nutrient intake from processed foods.

Applying MEDI to infant and adult stool metagenomes

To assess the frequency of food-derived reads across different stages of life, we quantified food abundances and contents in faecal samples using MEDI across two larger human datasets from infants and adults. Infant MGS data were obtained from a previously published cohort from St. Louis, USA, describing 447 longitudinal faecal samples from 60 infants of 1–253 days of age[46]. Adult reference samples were obtained from 351 healthy individuals (mean age of 31 years) within the iHMP project, as this subcohort includes standardized FFQs[47].

As expected, we generally observed a lower prevalence of food-derived reads in infant stool than in adult stool. Food-derived genomic material could be detected in less than 50% of infant faecal samples up to the onset of solid food intake (around day 160), when the prevalence of samples with detected food reads increased steadily (see Fig. 4a). This presence–absence pattern was correlated with age (beta = 0.17, logistic regression $P = 2 \times 10^{-5}$) but not feeding type (breast-fed or formula, logistic regression $P = 0.9$). By contrast, food-derived genomic material was detected in 98% of adult samples (359 out of 365). The total relative abundances of detected food items varied across infant samples and increased with age, independent of bacterial biomass and delivery mode (repeated-measures mixed-effects model, log(beta) = 0.13, $P = 0.0002$; see Methods), with a notable increase at the onset of solid food consumption (Fig. 4b). A somewhat higher relative abundance of foods could be observed in the first 3 days of life, but this signal was accompanied by a smaller fraction of bacterial reads (especially in infants delivered by caesarean section; see orange points in Fig. 4b), which increases the sensitivity for food detection. Bacterial biomass remained stable after the first 3 days of life in infants and in adults (usually representing more than 99% of metagenomic reads; Extended Data Fig. 3). In contrast to total bacterial relative abundance, which differed very little across samples (Extended Data Fig. 3), total food-read relative abundance varied over four orders of magnitude in both infants and adults (0.0004–1.3% of total reads in adults and 0.0004–7.8% in infants). In summary, although the relative metagenomic abundances of bacterial or human reads are

generally quite stable, food-read relative abundances are much more variable across infants and adults.

Mapping the nutrient and metabolite composition of the identified foods to a standardized portion allowed us to compare nutrient composition between infants and adults. Energy content per portion (kcal per 100 g) was positively correlated with infant age (Pearson test $r = 0.33$, $P < 3 \times 10^{-6}$) in concordance with an increased calorie demand for growth across the infant lifespan and with an increased reliance on solid foods with infant age. Similarly, the average energy density of the MEDI-inferred diets was slightly higher in infants than in adults (Fig. 4c; 255 kcal per 100 g vs 229 kcal per 100 g, Welch t-test $P = 0.02$). Macronutrient composition was similar between infants and adults and mirrored common nutritional recommendations[48]. Compared to infants, MEDI-inferred adult diets contained fewer fatty acids per standardized 100 g portion (Fig. 4d; 5.6 g per 100 g vs 7 g per 100 g, Welch t-test $P = 0.02$).

To assess the overall association of MEDI predictions with food frequency data and gut microbiome composition, we compared beta diversity between MEDI predictions, FFQ data and microbial species abundances in the iHMP cohort with Mantel permutation tests (Fig. 4e; based on Bray–Curtis distances; see Methods). Beta diversity estimates of MEDI-inferred food abundance were associated with FFQ beta diversity estimated from iHMP ($r = 0.082$, $P = 0.001$) as well as microbiome beta diversity ($r = 0.081$, $P = 0.001$). FFQ beta diversity, however, was not significantly associated with microbiome beta diversity ($r = 0.041$, $P = 0.074$), suggesting that MEDI estimates of dietary intake show a stronger association with microbial community composition in the human gut than FFQ data.

MEDI-inferred dietary intake was concordant with FFQ data from both infants and adults. Reported consumption of fruits, vegetables and cereals led to increased prevalence (logistic regression) and abundance (linear regression) of food-derived reads in infants (Fig. 4f). Within the adult iHMP cohort, food frequency patterns were captured by MEDI estimates for both prevalence and abundance of several food categories that could be mapped to FOODB food groups or subgroups (Fig. 4g), with the exception of shellfish, which generally showed relative abundances beneath the MEDI detection limit (that is, less than ten reads per sample on average; Fig. 2d).

In summary, MEDI was able to accurately capture many important aspects of dietary and nutritional intake across infants and adults.

Applying MEDI to a study of metabolic syndrome

To illustrate the ability of MEDI to identify dietary patterns associated with health and disease states, we performed a cross-sectional study to identify dietary features associated with metabolic syndrome in the absence of available dietary questionnaire data. Here, we leveraged a subcohort of 533 individuals with paired faecal samples and metabolic health information from the METACARDIS study[49]. The selected cohort consisted of 274 healthy individuals (healthy cohort in METACARDIS) and 259 individuals with varying clinical manifestations of metabolic syndrome, split into 134 individuals receiving medication (metabolically matched cohort (MMC) in METACARDIS) and 125 untreated individuals (untreated metabolically matched cohort (UMMC) in METACARDIS). MEDI identified a median of 1,687 food-derived reads per sample (0–575,020; Fig. 5a). Wheat, hibiscus, cocoa, pork, oats and flax were the most commonly detected food items in this cohort, accompanied by many food items that were only detected in small subsets of the people (Fig. 5a). Similarly, MEDI-inferred macronutrient and metabolite composition varied substantially across individuals, with a set of highly prevalent compounds detected across the cohort and other sets of metabolites only observed in small subgroups of individuals (Extended Data Fig. 4). Nutritional profiles could be clustered into a smaller subspace of carbohydrate and protein content, with a tendency towards higher energy content in high-protein–low-carbohydrate diets (Fig. 5b). However, neither protein nor carbohydrate content were associated with metabolic syndrome (Welch t-test of healthy cohort vs MMC and UMMC, both $P > 0.05$).

We next ran a systematic differential abundance analysis to identify dietary features associated with metabolic syndrome states (MMC or UMMC) compared to healthy individuals. We found that faecal samples from individuals with metabolic syndrome contained 121% more pork and 69% more chicken than samples from healthy controls, which in turn showed greater abundances of apple, pineapple and tomato DNA (Fig. 5c; limma-voom regressions, all FDR-corrected $P < 0.05$). At a broader taxonomic scale, metabolic syndrome was associated with a lower abundance of Streptophyta, which includes most plants found in the human diet, and a slight but non-significant increased abundance of Chordata, which include many animal-based foods (Fig. 5d). These results are consistent with prior studies that have identified higher consumption of animal products, such as pork, and lower consumption of fruits and vegetables as risk factors for metabolic syndrome and cardiovascular disease[50,51].

Several MEDI-inferred macronutrient and metabolite abundances (per standardized 100 g portion; Extended Data Fig. 5) were associated with metabolic syndrome or health in this cohort (Fig. 5e). Beta-lactose and cholesterol abundances were slightly higher in individuals with metabolic syndrome (25% and 15% increase, respectively; FDR-corrected $P = 0.003$ and 0.025), as were several fatty acids including arachidonic and vaccenic acids. MEDI-inferred diets from healthy individuals contained higher abundances of sugars (in particular, fructose), myo-inositol and ellagic acid. Although it seems counterintuitive that higher sugar consumption protects against metabolic syndrome, we note that MEDI does not quantify added sugars and is limited to naturally occurring sugars that, for the most part, are coming from fruits. This result provides a further caveat to interpreting MEDI estimates, which appear to be generally effective at inferring whole food intake but not at inferring processed food intake. In summary, MEDI was able to identify dietary patterns known to be associated with metabolic syndrome in a questionnaire-free setting, directly from stool MGS data.

Discussion

Obtaining accurate and unbiased estimates of dietary intake from large cross-sectional and longitudinal human cohorts is a fundamental challenge that has yet to be resolved. Here, we introduce MEDI, a semi-quantitative method for assessing dietary and nutritional intake directly from human stool DNA. MGS of stool DNA is the current gold standard in quantifying the taxonomic and functional composition of the human gut microbiome, and MEDI makes it possible to derive additional dietary intake information from this widely available data type. MEDI will allow for the extraction of dietary information from the hundreds of thousands of human stool metagenomic samples that have been deposited in public databases.

Dietary patterns are a major determinant of microbiome composition and a strong confounder of human cohort studies. We showed that MEDI estimates can recover dietary intake patterns in two controlled-feeding studies and that MEDI-inferred nutrient profiles show strong agreement with questionnaire-based nutrient profiles for a set of common macronutrients and micronutrients. However, the sensitivity of MEDI estimates was dependent on both sequencing depth and the temporal offset between food records and stool sampling. MEDI estimates tended to be sparse for lower-abundance food items. Specifically, we were often underpowered to detect lower-abundance foods (less than ten reads per million; Figs. 2d and 4g), and MEDI-inferred dietary intake was most resonant with food record data from more than 24–48 h before stool sampling (Extended Data Fig. 2b). Therefore, targeting higher metagenomic sequencing depths per sample (for example, >30 million reads, based on MEDI thresholds and observed food abundances; see Methods and Fig. 4g), repeated samplings of the same individual and accounting for intestinal transit times are

https://doi.org/10.1038/s42255-025-01220-1

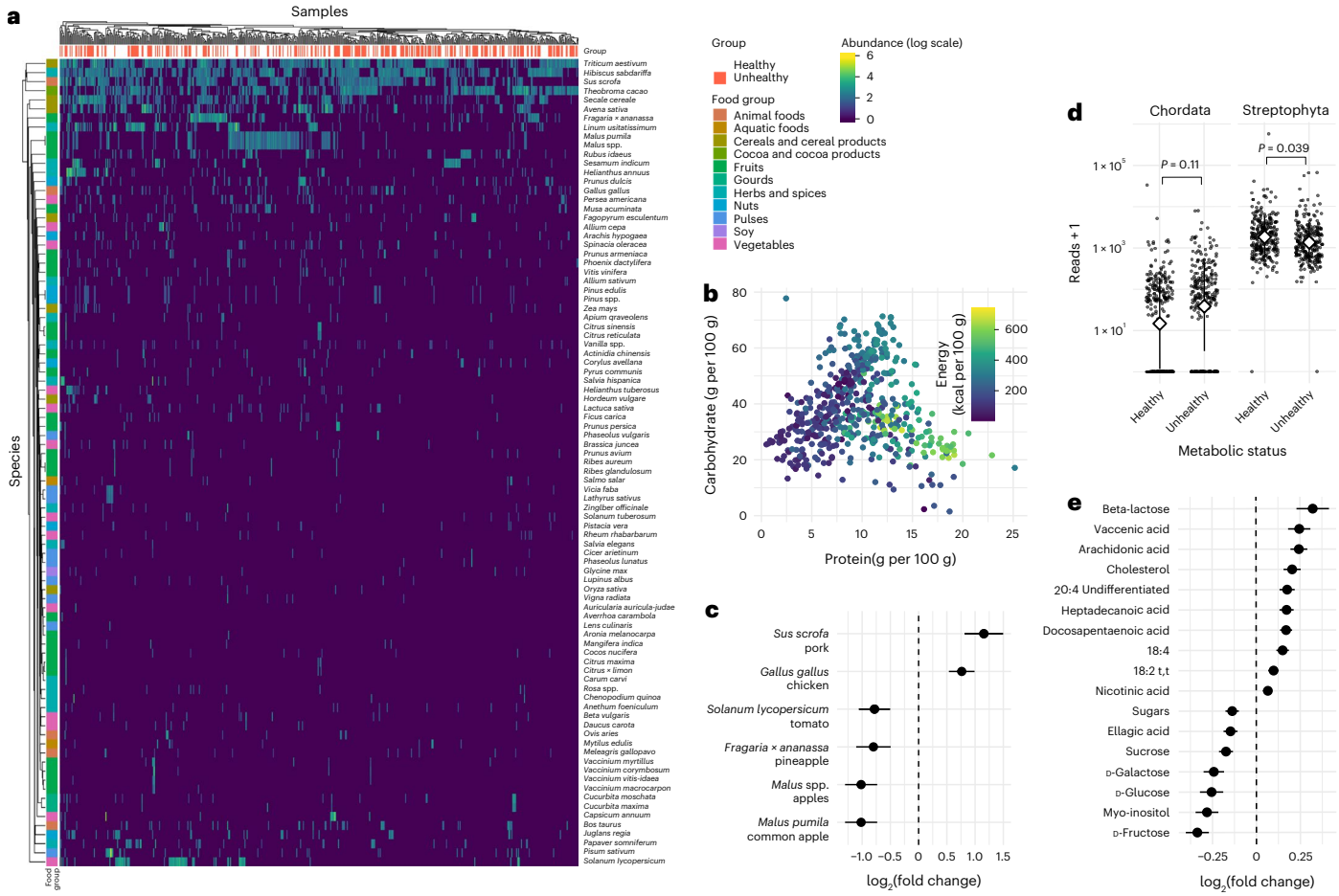


Fig. 5 | MEDI dietary intake estimates were associated with metabolic health. **a**, MEDI-detected food abundances across a cohort of 533 metabolically healthy and unhealthy individuals from the METACARDIS cohort. Fill colours denote abundance ($\log_{10}$(reads + 1)). Column annotations denote metabolic health status from the original METACARDIS cohort. Row annotations denote the major food groups from FOODB. **b**, Relationship between protein and carbohydrate abundances for all samples. Fill colour denotes energy content. **c**, Food-derived organisms with a significant association with metabolic health (FDR-corrected $P < 0.05$ in a limma-voom regression of read counts vs metabolic health status). Bars denote standard errors of the $\log_2$(fold change) ($n = 533$). Common food names are indicated below species. **d**, Food-derived phyla associated with metabolic health. FDR-corrected limma-voom P values are shown above. **e**, Food-derived compounds associated with metabolic health (FDR-corrected $P < 0.05$ in a linear regression of log abundance vs metabolic health status). Bars denote standard errors of $\log_2$(fold change) ($n = 533$). In **c** and **e**, positive log(fold changes) denote increased abundances in metabolically unhealthy individuals and negative log(fold changes) denote species more abundant in healthy individuals. Raw and corrected P values for **c** and **e** can be found in the Source data.

strategies that should optimize the performance of MEDI on human cohorts.

The total relative abundances of food-derived reads varied greatly across infants and adults, ranging from 0–1.5% of the total sequencing reads in adults and 0–8% in infants. This stands in stark contrast to bacterial or host-derived (human) reads, whose relative abundances do not vary much across individuals after the first few weeks of life (Extended Data Fig. 3). This is probably a consequence of the dynamic and variable nature of food items passing through the human body, and we postulate that DNA from whole foods, especially from plants, is more likely to survive passage through the gut than DNA from ultra-processed foods. For example, the low prevalence of food DNA in infant samples could be caused by breast milk intake (that is, we do not consider human DNA as a component of the diet) and the consumption of infant formula (that is, highly processed food). Despite this inherent variation, the relative abundance of food-derived DNA was strongly associated with age in infants and mirrored the transition to solid foods. Food DNA was found in most adult stool metagenomes (>98%) and recapitulated FFQ data (Fig. 4). However, FFQs only allowed for the quantification of broad food groups, whereas MEDI could identify individual species and distinguish between differences in food prevalence (consumption frequency) and abundance (relative amount consumed).

Additionally, MEDI allows for the mapping of food items to nutrient and metabolite intake, given a representative portion. These estimates of nutritional intake stood in good agreement with general nutrition recommendations and were able to identify anticipated shifts in nutrient consumption between infants and adults[52,53]. We saw strong agreement between MEDI estimates and nutrient intake data from a controlled-feeding study for total energy, protein, carbohydrate, potassium, cholesterol and vitamin B_{12} intake (Fig. 3). Finally, MEDI-inferred nutritional intake was consistent with dietary markers of metabolic syndrome, even in the absence of questionnaires or other dietary metadata (Fig. 5).

MEDI may also provide insight into the consumption of certain highly processed foods, owing to the presence of DNA that is probably derived from common non-caloric bulking agents added to these foods, like cotton-derived cellulose and pine wood pulp[54,55]. Specifically, hibiscus (cotton plants are closely related to hibiscus) and pine tree DNA were among the most prevalent food components identified in MEDI-inferred diets from the METACARDIS cohort. Although

hibiscus flowers and pine nuts are also present in the diet (for example, tea and pesto, respectively), cross-mapping of reads from cotton and pine wood pulp can occur. In addition to bulking agents, taxa like hibiscus can be turned into natural dyes that are added to certain processed foods, like meat products, alcoholic beverages or soft drinks[56]. Indeed, we found significant associations between *Hibiscus* abundance and the frequency of consuming processed meats, alcoholic beverages and soft drinks (Extended Data Fig. 6), suggesting that MEDI might also be used to identify or track certain food additives.

Limitations

Despite the promise of faecal DNA-based dietary tracking, there are certain limitations to these approaches that merit discussion. MEDI quantifies residual food DNA in human stool, which is probably biased towards whole foods and fails to capture certain aspects of dietary intake. For example, we postulate that many commonly consumed food items, including highly processed foods or supplements, do not leave a residual DNA signal in stool. This bias extends to the subsequent nutrient mappings, which do not account for common dietary ingredients, like added sugars or cooking oils. This is reflected in the observed positive association between MEDI-derived sugar abundances and better metabolic health, as MEDI-predicted sugars are largely derived from fruit[57,58]. However, there may be merit in disentangling sugars from whole foods, such as fruits and vegetables, from processed sugar. Most processed sugar is probably absorbed in the small intestine. Metabolites associated with faecal DNA may have survived passage through the upper gut and are more representative of the nutrient environment in the colon, which is particularly relevant to the metabolic activity of our commensal microbiota.

MEDI does not inherently account for differences in preparation types of foods when predicting dietary nutrient content. By default, MEDI calculates the mean nutrient content across all preparation types for a given food item in FOODB. However, MEDI does allow users to manually annotate the preparation types for food-derived DNA in a sample when calculating nutrient composition if users have access to preparation type information.

Another limitation is that existing food databases are biased towards the diets of largely white, affluent populations in Europe and America, often lacking foods consumed in indigenous societies, in non-industrialized countries or within minority populations in industrialized countries[59,60]. This is part of a larger issue that, on a global scale, many food items are poorly characterized[14]. As more food-related organisms are sequenced and their nutrient contents are quantified, MEDI inferences will become more and more accurate across diverse human populations. Finally, many of the limitations outlined above also extend to dietary questionnaires.

Applying MEDI to metagenomic data generated from individual food items, homogenized meals, saliva and stool from the same individuals may help us to better understand DNA degradation dynamics throughout the food preparation and digestion processes. Furthermore, combining MEDI estimates with metabolomics or amplicon-based approaches may help to resolve some of the limitations discussed above. Finally, additional decoy genomes of common model organisms, like the mouse genome, could be added to the MEDI database to extend the method beyond human stool. A better understanding of the biases introduced into faecal DNA-based diet tracking from food processing, cooking and digestion will serve to improve the quantitative accuracy of our method.

Conclusion

We have developed a data-driven methodology for estimating dietary and nutritional intake from food-derived DNA in human faecal metagenomes, called MEDI. Although dietary questionnaires are still the gold standard of dietary intake assessment, MEDI provides a secondary alternative for approximating dietary intake that can be readily applied to the large treasure trove of existing human stool MGS data for which dietary information is not currently available. By leveraging a common data type that is regularly collected to investigate the composition of the human gut microbiota, MEDI provides a value addition to any past, present or future metagenomic study for which rough estimates of dietary intake would prove useful. We believe that MEDI will be a valuable tool for nutritionists, epidemiologists, anthropologists, clinicians and microbiome researchers.

Methods

Food genome database and mapping hash construction

CSV files describing FOODB (v.1.0) were downloaded and all food items that mapped to the NCBI Taxonomy Database were extracted. NCBI taxonomy IDs were mapped onto canonical ranks using taxonkit (v.0.15.0) and searched for in NCBI GenBank first, followed by a search of the NCBI Nucleotide Database. Manifest files listing all available assemblies were downloaded from NCBI GenBank and each species ID from in the prior step was searched for in the assembly table. All food-derived NCBI taxonomic IDs without a species-level match were then matched at the genus level. Whenever there were multiple potential matches, they were ranked in preference by submission date, preferring most recent to oldest, and by RefSeq quality, preferring 'reference genome' over 'representative genome'. Finally, within this ordering, additional ties were cleared by the assembly type, preferring complete genomes to chromosomes, followed by contigs. All food-derived taxa without a match to NCBI GenBank were then searched in the full NCBI Nucleotide Database for any partial genomic assembly or genes of at least 10,000 bp in length, returning the longest contiguous sequences and up to a maximum of 500 records. Similarly, NCBI Nucleotide Database searches were first performed on the species rank, followed by a search on the genus rank, for all non-identified taxa. All matched assemblies were then downloaded in parallel using the NCBI eFetch API, annotated with the corresponding NCBI taxonomic ID in Kraken annotation format, and compressed. Average nucleotide identity was estimated using the SOURMASH (v.4.8.2) 'compare –ani' command.

Food information, macronutrient composition, energy content and detailed metabolite abundances were extracted from the FOODB data for each matched taxon. Macronutrients and energy content were extracted from the 'Nutrient' source type in FOODB and metabolite abundances from the 'Compound' source type. Here, the standardized contents from the FOODB were used as a basis and converted to mg per 100 g by a manual mapping of all unique abundance units in the FOODB to the appropriate scaling factor. To ensure an accurate reflection of nutrient profiles, manual curation of the FOODB was performed. For this, common food items were cross-checked by comparing their macronutrient composition to what was found in FDC (https://fdc.nal.usda.gov). Most macronutrients were in line with FDC data, but energy content and cholesterol content did not agree. Standardized energy content exceeded the theoretical maximum of 900 kcal per 100 g for several food items in FOODB. However, the original content ('orig_content' entry in FOODB) was in the correct range and did agree with FDC data. Additional validation of energy content was performed by calculating energy values based on the Atwater method from the macronutrient content of proteins, carbohydrates and fat. Calculated energy and energy values from the original content entries showed a Pearson correlation of 0.98 (Extended Data Fig. 6a). Additionally, the histogram of cholesterol abundances in the FOODB showed a separated group of unreasonably high cholesterol contents for some foods (Extended Data Fig. 6b), often exceeding 1 g per 100 g. Given that unreasonably high values were usually off by a factor of 1,000.0, we assumed that this was caused by a mixup of µg and mg during data entry. Adjusting cholesterol values from the high group by dividing by a factor of 1,000 led to agreement with the cholesterol content in the FDC data and was used as standardized content in the derived MEDI database.

A new Kraken 2 database was built using the same initial taxonomy dump as used in the previous ID matching with the *kraken2-build* command using Kraken v.2.1.3. This database was first filled with decoy sequences, comprising all complete genomes from bacteria, archaea, viruses, plasmids, common vector contaminants (UniVec core) and the human reference genome (GRCh3838), sourced from NCBI GenBank. The raw hash table of food-related genomic sequences and decoy genomes was approximately 420 GB in size. The database was then indexed with Bracken for 100 bp and 150 bp reads. The self-classification step required to calculate the *k*-mer distributions for Bracken was performed manually as a separate step and used the same memory mapping and caching strategy as described below for the mapping.

Metagenomic data download and preprocessing

Raw MGS data were downloaded from the National Institutes of Health (NIH) Sequence Read Archive (SRA) for all datasets using the SRA download pipeline from https://github.com/gibbons-lab/pipelines. Preprocessing was performed using FASTP[61], trimming the first five bases at the 5′ end and using a sliding window trimming on the 3′ end with a cutoff of a quality score of 20. Reads shorter than 50 bp after trimming were discarded from the analysis. Quality control summaries were inspected using MultiQC[62], verifying that samples were free of any remaining sequencing adaptors and that the insert size in paired-end data was correct.

Metagenomic mapping and read counting

The MEDI pipeline was implemented as a set of Nextflow workflows[63] and is available at https://github.com/gibbons-lab/medi. For all mapping performed in this publication, Kraken 2 was run with memory mapping turned on to parallelize database loading across all running mapping processes. This allowed for instantaneous parallelized reading of the large mapping index and led to much lower amortized computation time for individual processes, as subsequent sample processing would use a cached version of the hash (see Extended Data Fig. 1). To extend this strategy to high-performance computing clusters, in which samples are usually processed on different compute nodes, we implemented a batch strategy whereby several hundred samples were collected into a single batch that was guaranteed to run on the same compute node. Using six CPUs, this resulted in classification rates of around 300,000 reads per second on the local high-performance computing cluster using SLURM (MedBioNode at the Medical University of Graz).

Kraken 2 was run with a default confidence cutoff of 0.3 to ensure sufficient specificity in LCA calculation. To improve mapping accuracy, we also combined this with an additional post-mapping filter that worked specifically on the canonical ranks of reads and the individual *k*-mers classified within each read (Fig. 2a). Here, reads were filtered based on cutoffs for consistency, mapping entropy and multiplicity. Consistency denotes the fraction of *k*-mer-level taxonomy assignments that are contained in the final read classification. Hence, let $S_i = \{s_r\}_i$ be the set of taxonomic identifiers assigned to read i for all ranks r and let $K_i = \{k_j\}$ be the LCA taxonomy assignments for each classified *k*-mer j in read i. The consistency of read i is then given by $C_i = |\{k_j \in S_i\}|/|K_i|$. A consistency of 1 means that all individual *k*-mer assignments fall onto a single taxonomic path, whereas a low consistency means that many individual *k*-mer assignments lie on conflicting branches of the phylogenetic tree. Multiplicity is the number of unique *k*-mers assignments at the same rank r as the read assignment, $M_i = |\{k_j | rank(k_j) = r\}|$ Finally, mapping entropy is the Shannon index of the *k*-mer assignments on the same rank as the final read assignment. Thus, if $p_r(k)$ is the relative frequency of taxon k at rank r (relative abundance of *k*-mer classifications within the read), then the mapping entropy is given as $-\sum_i p_{ir}(k) \log p_{ir}(k)$. Scoring and filtering methods were implemented in the *architeuthis* software (https://github.com/cdiener/architeuthis) in the Go programming language (https://golang.org). After Kraken 2 read-level classification, we removed all reads with consistency of <0.95, entropy of >0.1 and multiplicity of >4. This typically removed about 10% of the classified reads from Kraken 2. Final abundance estimation was performed with Bracken, using a minimum clade abundance of ten reads to avoid redistribution to taxa with very low occurrence.

MBD study

The MBD study was a randomized crossover controlled study in which participants consumed two diets with a >14-day washout in between the two diet periods (NCT02939703). The full methods and results have been previously published[41,64]. In brief, diets were prepared in a metabolic kitchen to precisely match each study participant's measured energy expenditure. The diets were matched in total macronutrients. The MBD was designed to maximize dietary substrates that reach the colon by being high in fibre, high in resistant starch, rich in whole foods and limited in processed foods. The Western diet was the opposite, essentially 'starving' the colonic microbes. Diets were consumed outpatient for 11 days and inpatient for 11 days. Adherence to provided diets was monitored and was ~100%. Faecal samples were collected while participants were domiciled in our metabolic ward and were processed under an anaerobic hood within 1 h of production. The metagenomic sequences were derived from samples that were composited over 6 days of controlled feeding. Raw sequencing data were downloaded as FASTQ data as described above from the SRA projects PRJNA913183 and PRJNA947193. A total of 17 individuals completed the crossover study, providing samples for the two dietary interventions for each individual. When metagenomic samples existed for a single individual and dietary intervention combination, we chose the sample with the largest sequencing depth. This yielded the final dataset of 34 samples[41]. Beta diversity was calculated using Bray–Curtis distances on the relative abundance data of food items for all samples with at least one read mapped to food items. Non-metric multidimensional scaling of the Bray–Curtis distances was performed using the 'metaMDS' function from the 'vegan' R package (https://cran.r-project.org/package=vegan). Association with dietary intervention groups was assessed with PERMANOVA using the 'adonis2' function from the 'vegan' package with 1,000 random permutations of the data. Differences in total faecal food abundances were assessed by comparison of the total relative abundance of food items in both diet groups using a Mann–Whitney *U*-test for all samples that had at least one detected food read (30 out of 34 samples).

PATH study

Raw sequencing data in FASTQ format, metadata and dietary record data were provided by the study authors based on what was reported in the original study[42]. FASTQ files were preprocessed and analysed using MEDI as before. Data from baseline faecal samples were not used here. Differential abundance for food items was assessed using limma-voom log-normal regression on the food species read counts. FDR was controlled using the Benjamini–Hochberg method ($q < 0.05$).

The temporal offset between food diary entry and faecal samples was calculated from the faecal sampling timepoint (including time down to minutes) and the food diary entry date that corresponded to the consumption time (date only). A reference time of 0:00 was used for the food diary entry, yielding the largest possible offset between a pair of a single faecal sample and food diary entry for a single individual. A positive offset denoted a faecal sample obtained after food diary entry and a negative offset denoted faecal samples obtained before food entry. A set of common macronutrients and some representative micronutrients were selected, and their IDs were mapped manually between MEDI IDs and the identifiers used in the food diary nutrient data. Standard errors for each data point were calculated as s.d. / sqrt(n). Agreements between MEDI predictions (single value for

each individual obtained from the respective faecal sample) and mean nutrient abundances from food diaries were assessed using Pearson moment-product correlation. Analyses for specific offset groups were run by grouping offsets into quantile groups of <0 h, 0–48 h, 48–96 h and >96 h, yielding similar sample sizes for each group, and assessing correlations using means from only the specific quantile group (see Extended Data Fig. 2).

Infant metagenomic time series

Raw sequencing data were downloaded and processed from the NIH SRA project PRJNA473126 as described above. Preprocessed FASTQ files were then analysed with the MEDI pipeline, described earlier. Diet summaries from the metadata provided on SRA were used to identify infant–timepoint pairs, when a specific food group was consumed. The onset of solid food consumption was the first timepoint for each infant, when solid foods like fruits, meat, cereal or sweets were listed in the diet summary. Associations of total food reads against age were run in a linear mixed-effects model using $\log_{10}$(reads + 1 / total reads) as the dependent variable and infant age as a fixed effect, with random intercepts and slopes for individual infants. The delivery mode (categorical: vaginal birth vs caesarean) and abundance of bacterial reads ($\log_{10}$(bacteria reads + 1 / total reads)) were used as covariates. Statistical significance for individual covariates was obtained using Satterthwaite's degrees of freedom method in the 'lmerTest' R package (https://cran.r-project.org/package=lmerTest). To evaluate associations with FFQs, food reads were first summed for all food items in the vegetable and fruit food groups in the FOODB to yield a 'fruits and vegetables' abundance and separately for the cereal FOODB food group to obtain read counts for cereals. Significance for abundances was then assessed by a Welch *t*-test of relative abundances (group food reads / total reads) between infants who consumed the specific food group and those who did not. Significance for prevalence was assessed by logistic regression on indicator variables being set to one if the food group was detected in the sample (groups reads > 0) or to zero otherwise, with the consumption group from FFQs as a categorical covariate.

iHMP Project inflammatory bowel disease data

A list of individual run IDs was obtained from the NIH SRA bioproject PRJNA398089, keeping data from only healthy controls. Metadata for the samples were downloaded from https://ibdmdb.org, keeping only those raw FASTQ files with a match to the metadata. Samples were processed as described above and analysed with the MEDI pipeline. All iHMP raw sequencing files were also processed by an in-house metagenomics pipeline that uses Kraken 2 and Bracken with the default database (not containing foods) (https://github.com/gibbons-lab/pipelines). Bracken-estimated species abundances were averaged across all samples, and only taxa with at least 100 reads on average were kept. This yielded an average abundance profile for non-food-associated taxa in the human gut, representing the full diversity of the 365 individuals. Macronutrient profiles were provided by MEDI as described before in units of mg per 100 g and converted to g per 100 g. Energy content was provided as kcal per 100 g. For comparisons with infant data, extracted macronutrients and energy abundance were tested with a Welch *t*-test across groups.

To evaluate associations with FFQs, coarse food groups in the iHMP metadata were first manually mapped to an appropriate set of food groups in FOODB. No overlap in food items between different food groups was observed. The full mapping rules can be found in the source code repository in the 'ihmp.rmd' R markdown notebook. Food reads were first summed for all food items in the tested food group to obtain read counts for cereals. Consumption frequencies in the iHMP metadata were translated to mean intake frequencies. For instance, 'within the past 2 to 3 days' would yield a frequency of 1 / (2.5 days) and 'yesterday, three or more times' would yield 3 / (1 day). Significance for abundances was then assessed by a linear regression of $\log_{10}$ relative abundances with pseudocounts (group food reads + 1 / total reads) with the obtained FFQ consumption frequencies as a covariate. Significance for prevalence was assessed by logistic regression on indicator variables being set to one if the food group was detected in the sample (groups reads > 0) or to zero otherwise, with the consumption frequency from FFQs as a continuous covariate.

For beta diversity analysis, bacteria and archaea species abundances were extracted from the Bracken output (S_counts.csv file returned by MEDI) and converted to relative abundances by dividing individual species read counts in each sample by the sum of *Bacteria* and *Archaea* reads in each sample. We opted not to normalize by total reads counts here to not introduce bias, as this would also contain food reads, thus yielding smaller relative microbiome abundances for samples with a high proportion of food reads. Species were filtered for those with at least ten reads on average across all samples, and pairwise Bray–Curtis distances between samples were calculated. For FFQ distances, food frequencies in 1 / day as derived above were used to calculate Bray–Curtis distances between samples, yielding a minimum distance when two individuals consumed the same food groups with the same frequency. MEDI food abundances were first transformed to relative abundances by dividing by the sum of food reads in each sample with a similar reasoning as for microbiome data. Food items were filtered to those appearing in at least 10% of all iHMP samples to limit the influence of missing observations. This yielded a total of 16 common food items, which was in the same range as the food groups covered by the FFQ data (19 groups). Bray–Curtis distances were calculated between all sample pairs to yield the MEDI beta diversity data. Associations between pairwise beta diversity measures were then assessed with Mantel tests, using 1,000 random permutations and the Pearson product-moment correlation.

METACARDIS data

Raw sequencing data were downloaded and processed from the NIH SRA project PRJEB37249 as described above and analysed with MEDI; metadata were obtained from Supplemental Table 14 of the associated publication[49]. From the 888 processed samples, 533 matched either healthy controls, MMC or UMMC groups (where MMC and UMMC groups individuals with metabolic syndrome). Differential abundance for food items was assessed using limma-voom log-normal regression on the taxon read counts. Only food items that appeared in more than 10% of all samples with an average of ten reads or higher were tested. FDR was controlled using the Benjamini–Hochberg method ($q < 0.05$). Phyla differential abundance testing was performed in the same manner after summing food reads on the phylum rank. Differential abundance for food content was also assessed using limma, with linear regression on log-transformed food abundances with an added pseudo-abundance of 1×10^{-6} to avoid taking the logarithm of zero abundances ($\log_{10}$(standardized abundance + 1×10^{-6})). Only nutrients and compounds that appeared in more than 10% of all samples with an average abundance of 1×10^{-6} mg per 100 g or higher were tested. FDR was again controlled using the Benjamini–Hochberg method ($q < 0.05$).

Statistics and reproducibility

In this study, we performed analyses exclusively on data from previously conducted studies and therefore sample sizes were not determined a priori. Effect sizes are similar to what was observed in previous studies[47,49]. C.D., K.F. and S.M.G. were blinded to the food items contained in the MBD diet until after the analyses were conducted. Thus, inference of the food components of the MBD diet was conducted without knowledge of the ground truth. Sensitivity was assessed using the deeply sequenced iHMP cohort. Some low-abundance foods, such as beans and fish, were quantified with abundances as low as 0.3 reads per million, even in the high consumption frequency groups (Fig. 4g). Given that MEDI does require at least ten reads to call a food genome as present in a sample, this suggests a minimum sequencing depth of around 30 million reads to reliably quantify those food groups. No

data were excluded from the analyses. We generally chose statistical tests that were either non-parametric or robust to violations of data distribution. However, data distributions were not formally tested.

Simulated ground-truth data

Reference genomes for all bacteria and archaea in the decoy data used in the Kraken 2 mapping database were downloaded along with a 1000 Genomes Project human genome assembly[65]. This human genome assembly was used instead of the one in the Kraken 2 database to test for the effect of having human sequence fragments that may not have been represented in the decoy database. A background relative abundance distribution for bacterial, archaeal and viral taxa was established using the iHMP healthy cohort, as described above. The respective NCBI taxonomy IDs for each species in the average abundance profile were used to match organisms to reference genomes from the NCBI RefSeq database and download the assemblies in FASTA format. Reads of 150 bp length were sampled with DWGSIM (https://github.com/nh13/DWGSIM), using a uniformly decreasing (5′ to 3′) error rate of 0.001–0.005 for the forward reads and 0.05–0.01 for the reverse reads to a final depth of ten million paired-end reads per sample. For this process, reads were sampled from each individual reference genome to a final n of $r_i \times 10{,}000{,}000$, where r_i is the relative abundance of taxon i in the background abundance profile. Four negative-control samples were generated by sampling from the background only without introducing any food reads. Positive samples were generated by repeatedly choosing ten random food species and adding one million reads to nine million reads of the background (10% final food-read abundance) using the food genomic sequences downloaded during database construction and read sampling as described above. Individual food abundances were staggered by a natural log level for each of the ten food items, yielding ten log levels of abundance variation, ranging from less than ten to more than half a million reads across the ten food species. The simulation of food-positive samples was repeated 16 times (16 random sets of ten foods). After sampling, reads were shuffled and quantified using the MEDI mapping pipeline described before. The false-positive mapping rate was quantified as the fraction of reads that were assigned to a food item not present in the sample. Mapping accuracy was quantified as the Pearson product-moment correlation of expected log-transformed relative read abundance +1 versus the observed log-transformed read abundance +1.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data for specific food items are available at https://foodb.ca. Individual matched genomic assemblies can be downloaded from GenBank or the Nucleotide Database and are listed at https://github.com/Gibbons-Lab/medi-paper/blob/main/db/data/manifest.csv. Metagenomic sequencing data for the studied cohorts are available on the NCBI SRA under accession numbers PRJNA473126 (infants), PRJNA398089 (iHMP), PRJEB37249 (METACARDIS), PRJNA947193 (MBD) and PRJNA1198318 (PATH). Source data are provided with this paper.

Code availability

All intermediate data files, metadata and analysis code have been uploaded to GitHub (https://github.com/Gibbons-Lab/medi-paper). The MEDI software package is available on GitHub (https://github.com/Gibbons-Lab/medi).

References

1. Harding, J. E., Cormack, B. E., Alexander, T., Alsweiler, J. M. & Bloomfield, F. H. Advances in nutrition of the newborn infant. *Lancet* **389**, 1660–1668 (2017).
2. de Ridder, D., Kroese, F., Evers, C., Adriaanse, M. & Gillebaart, M. Healthy diet: health impact, prevalence, correlates, and interventions. *Psychol. Health* **32**, 907–941 (2017).
3. Clark, M., Hill, J. & Tilman, D. The diet, health, and environment trilemma. *Annu. Rev. Environ. Resour.* **43**, 109–134 (2018).
4. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
5. Wang, D. D. et al. The gut microbiome modulates the protective association between a Mediterranean diet and cardiometabolic disease risk. *Nat. Med.* **27**, 333–343 (2021).
6. Gu, Y., Nieves, J. W., Stern, Y., Luchsinger, J. A. & Scarmeas, N. Food combination and Alzheimer disease risk: a protective diet. *Arch. Neurol.* **67**, 699–706 (2010).
7. Mente, A. et al. Diet, cardiovascular disease, and mortality in 80 countries. *Eur. Heart J.* **44**, 2560–2579 (2023).
8. Magkos, F., Hjorth, M. F. & Astrup, A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. *Nat. Rev. Endocrinol.* **16**, 545–555 (2020).
9. Key, T. J., Allen, N. E., Spencer, E. A. & Travis, R. C. The effect of diet on risk of cancer. *Lancet* **360**, 861–868 (2002).
10. Ludwig, D. S., Ebbeling, C. B. & Heymsfield, S. B. Improving the quality of dietary research. *JAMA* **322**, 1549–1550 (2019).
11. Molag, M. L. et al. Design characteristics of food frequency questionnaires in relation to their validity. *Am. J. Epidemiol.* **166**, 1468–1478 (2007).
12. Timon, C. M. et al. A review of the design and validation of web- and computer-based 24-h dietary recall tools. *Nutr. Res. Rev.* **29**, 268–280 (2016).
13. Conway, J. M., Ingwersen, L. A. & Moshfegh, A. J. Accuracy of dietary recall using the USDA five-step multiple-pass method in men: an observational validation study. *J. Am. Diet. Assoc.* **104**, 595–603 (2004).
14. Abu-Saad, K., Shahar, D. R., Vardi, H. & Fraser, D. Importance of ethnic foods as predictors of and contributors to nutrient intake levels in a minority population. *Eur. J. Clin. Nutr.* **64**, S88–S94 (2010).
15. Mozaffarian, D. & Forouhi, N. G. Dietary guidelines and health—Is nutrition science up to the task? *Brit. Med. J.* **360**, k822 (2018).
16. Taubes, G. Epidemiology faces its limits. *Science* **269**, 164–169 (1995).
17. Young, S. S. & Karr, A. Deming, data and observational studies. *Signif. (Oxf.)* **8**, 116–120 (2011).
18. Sturgeon, C. M. et al. National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers. *Clin. Chem.* **54**, e11–e79 (2008).
19. Mundi, S. et al. Endothelial permeability, LDL deposition, and cardiovascular risk factors—a review. *Cardiovasc. Res.* **114**, 35–52 (2018).
20. Zuppinger, C. et al. Performance of the digital dietary assessment tool MyFoodRepo. *Nutrients* **14**, 635 (2022).
21. Mohanty, S. P. et al. The food recognition benchmark: using deep learning to recognize food in images. *Front. Nutr.* **9**, 875143 (2022).
22. Mortazavi, B. J. & Gutierrez-Osuna, R. A review of digital innovations for diet monitoring and precision nutrition. *J. Diabetes Sci. Technol.* **17**, 217–223 (2023).
23. Hassannejad, H. et al. Automatic diet monitoring: a review of computer vision and wearable sensor-based methods. *Int. J. Food Sci. Nutr.* **68**, 656–670 (2017).
24. West, K. A., Schmid, R., Gauglitz, J. M., Wang, M. & Dorrestein, P. C. foodMASST a mass spectrometry search tool for foods and beverages. *NPJ Sci. Food* **6**, 22 (2022).
25. Dorrestein, P. Metabolomics technologies for defining diet influences on brain metabolome and in Alzheimer's disease. *Alzheimers Dement.* **18**, e067277 (2022).

26. Petrone, B. L. et al. Diversity of plant DNA in stool is linked to dietary quality, age, and household income. *Proc. Natl Acad. Sci. USA* **120**, e2304441120 (2023).
27. Deagle, B. E., Thomas, A. C., Shaffer, A. K., Trites, A. W. & Jarman, S. N. Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: Which counts count? *Mol. Ecol. Resour.* **13**, 620–633 (2013).
28. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
29. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
30. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
31. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
32. Brent, M. R. How does eukaryotic gene prediction work? *Nat. Biotechnol.* **25**, 883–885 (2007).
33. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
34. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. *BMC Genomics* **16**, 236 (2015).
35. Shen, W. et al. KMCP: accurate metagenomic profiling of both prokaryotic and viral populations by pseudo-mapping. *Bioinformatics* **39**, btac845 (2023).
36. Gihawi, A. et al. Major data analysis errors invalidate cancer microbiome findings. *Mbio* **14**, e0160723 (2023).
37. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique *k*-mer counts. *Genome Biol.* **19**, 198 (2018).
38. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
39. Srivastava, A. et al. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.* **21**, 239 (2020).
40. Sun, Z. et al. Challenges in benchmarking metagenomic profilers. *Nat. Methods* **18**, 618–626 (2021).
41. Corbin, K. D. et al. Host–diet–gut microbiome interactions influence human energy balance: a randomized clinical trial. *Nat. Commun.* **14**, 3161 (2023).
42. Thompson, S. V. et al. Avocado consumption alters gastrointestinal bacteria abundance and microbial metabolite concentrations among adults with overweight or obesity: a randomized controlled trial. *J. Nutr.* **151**, 753–762 (2021).
43. Asnicar, F. et al. Original research: blue poo: impact of gut transit time on the gut microbiome using a novel marker. *Gut* **70**, 1665 (2021).
44. Duan, Y., Pi, Y., Li, C. & Jiang, K. An optimized procedure for detection of genetically modified DNA in refined vegetable oils. *Food Sci. Biotechnol.* **30**, 129–135 (2021).
45. Scollo, F. et al. Absolute quantification of olive oil DNA by droplet digital-PCR (ddPCR): comparison of isolation and amplification methodologies. *Food Chem.* **213**, 388–394 (2016).
46. Baumann-Dudenhoeffer, A. M., D'Souza, A. W., Tarr, P. I., Warner, B. B. & Dantas, G. Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nat. Med.* **24**, 1822–1829 (2018).
47. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
48. Manore, M. M. Exercise and the Institute of Medicine recommendations for nutrition. *Curr. Sports Med. Rep.* **4**, 193–198 (2005).
49. Fromentin, S. et al. Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat. Med.* **28**, 303–314 (2022).
50. Thomas, M. S., Calle, M. & Fernandez, M. L. Healthy plant-based diets improve dyslipidemias, insulin resistance, and inflammation in metabolic syndrome. A narrative review. *Adv. Nutr.* **14**, 44–54 (2023).
51. Neuenschwander, M. et al. Substitution of animal-based with plant-based foods on cardiometabolic health and all-cause mortality: a systematic review and meta-analysis of prospective studies. *BMC Medicine* **21**, 404 (2023).
52. Embleton, N. D. Optimal protein and energy intakes in preterm infants. *Early Hum. Dev.* **83**, 831–837 (2007).
53. Uauy, R., Mena, P. & Valenzuela, A. Essential fatty acids as determinants of lipid requirements in infants, children and adults. *Eur. J. Clin. Nutr.* **53**, S66–S77 (1999).
54. Neis, F. A., de Costa, F., de Araújo, A. T. Jr., Fett, J. P. & Fett-Neto, A. G. Multiple industrial uses of non-wood pine products. *Ind. Crops Prod.* **130**, 248–258 (2019).
55. Wallick, D. Cellulose polymers in microencapsulation of food additives. In *Microencapsulation in the Food Industry* (eds Gaonkar A. et al.) 181–193 (Elsevier, 2014).
56. Li, N., Simon, J. E. & Wu, Q. Development of a scalable, high-anthocyanin and low-acidity natural red food colorant from *Hibiscus sabdariffa* L. *Food Chem.* **461**, 140782 (2024).
57. Ruxton, C. H. S., Gardner, E. J. & McNulty, H. M. Is sugar consumption detrimental to health? A review of the evidence 1995–2006. *Crit. Rev. Food Sci. Nutr.* **50**, 1–19 (2010).
58. Crovetto, M. et al. Effect of healthy and unhealthy habits on obesity: a multicentric study. *Nutrition* **54**, 7–11 (2018).
59. Gibbons, S. M. et al. Perspective: leveraging the gut microbiota to predict personalized responses to dietary, prebiotic, and probiotic interventions. *Adv. Nutr.* **13**, 1450–1461 (2022).
60. Lovegrove, J. A., Hodson, L., Sharma, S. & Lanham-New S. A. *Nutrition Research Methodologies* (John Wiley & Sons, 2015).
61. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
62. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
63. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
64. Corbin, K. D. et al. Integrative and quantitative bioenergetics: design of a study to assess the impact of the gut microbiome on host energy balance. *Contemp. Clin. Trials Commun.* **19**, 100646 (2020).
65. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

Acknowledgements

Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the NIH under award number R01DK133468 (to S.M.G.) and by a Global Grants for Gut Health Award from Nature Portfolio and Yakult (to S.M.G.). This research was funded in part by the Austrian Science Fund (FWF): grant Cluster of Excellence CoE7 (to C.D. and C.M.-E.) and SFB ImmunoMetabolism 10.55776/F8300 (to C.M.-E.). Computational resources for this work were provided by the MedBioNode High-Performance Computing cluster at the Medical University of Graz. H.D.H. acknowledges funding for the PATH study from the Foundation for Food and Agriculture Research (FFAR) New Innovator Award and Hass Avocado Board.

Author contributions

C.D. and S.M.G. conceived of the study. C.D. wrote and tested the software. C.D., K.F. and S.M.G. performed analyses. H.D.H., K.D.C., C.M.-E. and S.M.G. provided datasets and resources. C.D. wrote the initial draft of the paper. C.D. and S.M.G. provided supervision. All authors contributed to writing and revising the paper.

Competing interests

The authors report no financial or non-financial competing interests relevant to the work presented in this paper. S.M.G. received funding from a Global Grants for Gut Health Award from Nature Portfolio and Yakult. However, the funders were not involved in conducting the research, drafting the paper or reviewing the work.

Additional information

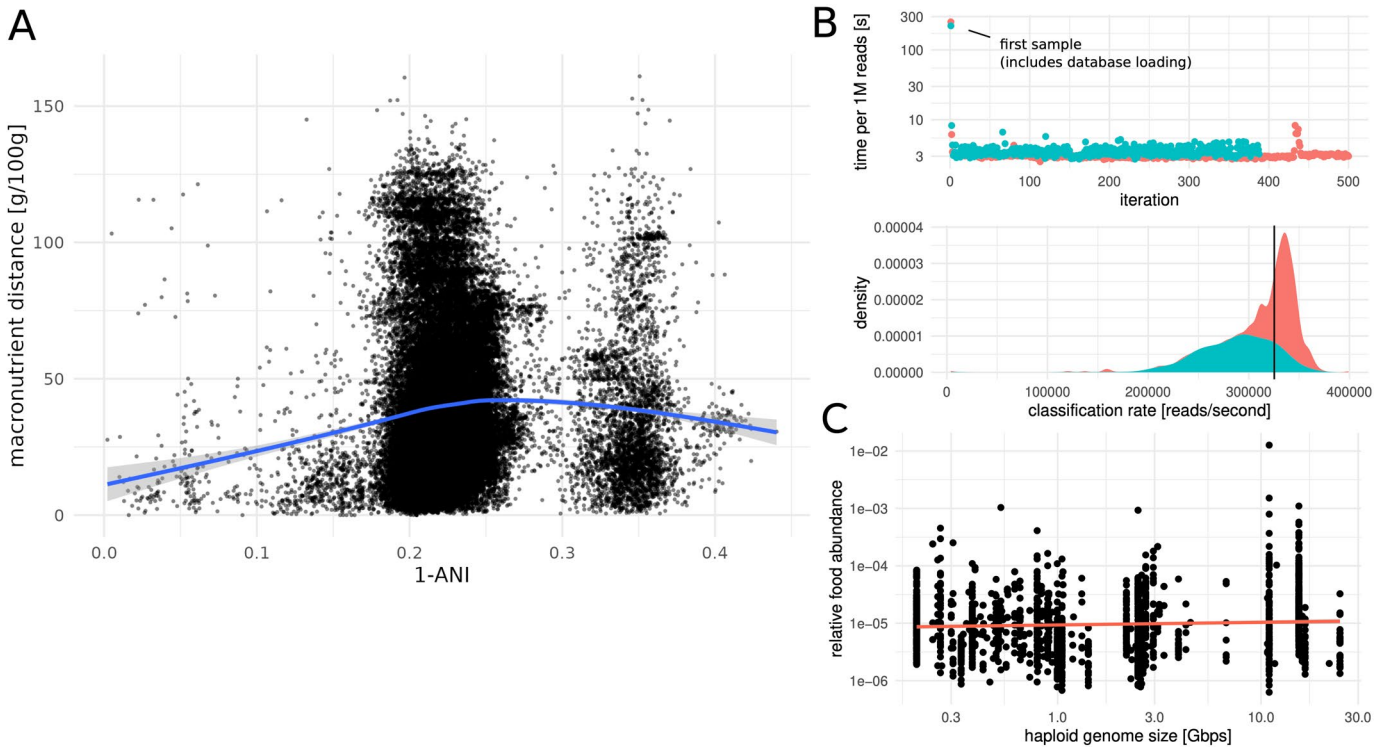
Extended data is available for this paper at https://doi.org/10.1038/s42255-025-01220-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42255-025-01220-1.

Correspondence and requests for materials should be addressed to Christian Diener or Sean M. Gibbons.

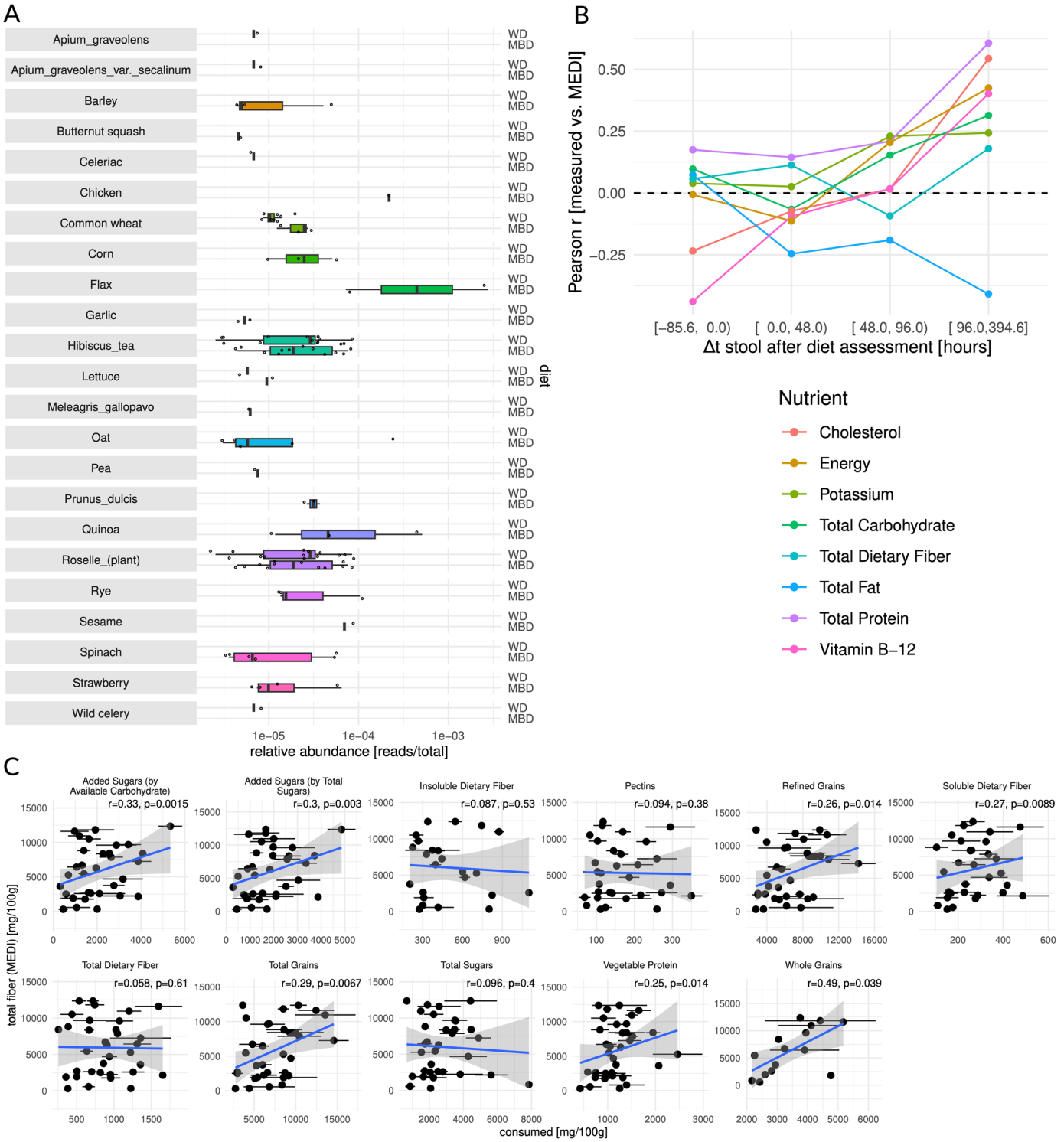
Peer review information *Nature Metabolism* thanks Lars Dragsted and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Yanina-Yasmin Pesch, in collaboration with the *Nature Metabolism* team.

https://doi.org/10.1038/s42255-025-01220-1



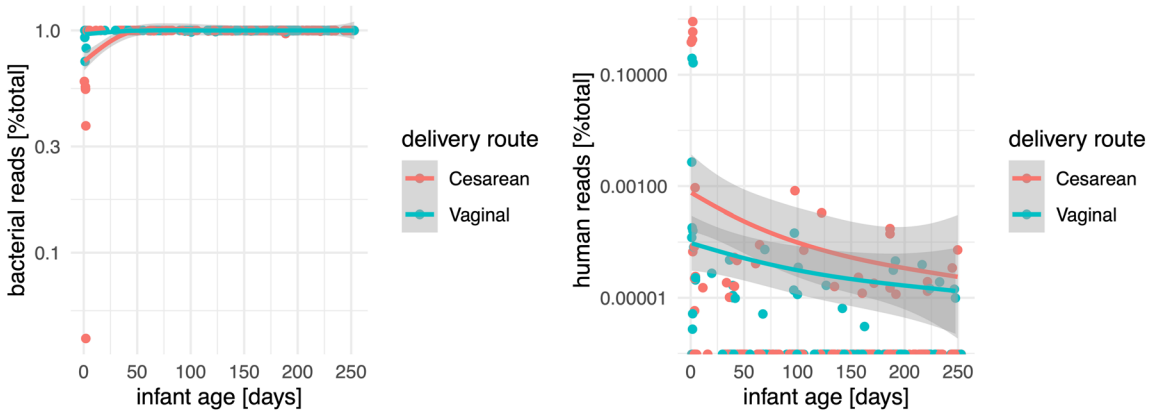
Extended Data Fig. 1 | MEDI benchmarks. (**a**) Genomic distance (1 - ANI) vs. macronutrient distance (euclidean, in g/100 g). The blue line denotes a smooth spline regression and shaded area denotes the 95% confidence interval of the mean spline regression. (**b**) Benchmark of cached and batched processing using MEDI (6 CPUs per process, see Methods). 888 samples were divided into two batches of 500 and 388 FASTQ files and processes separately in parallel. Each point denotes a single FASTQ file and colors denote the batch. Vertical line denotes median classification rate. (**c**) Relationship between (haploid) genome/assembly size and food abundance in the iHMP data set. Shown are only genomes/assemblies with at least 1 million basepairs.

https://doi.org/10.1038/s42255-025-01220-1



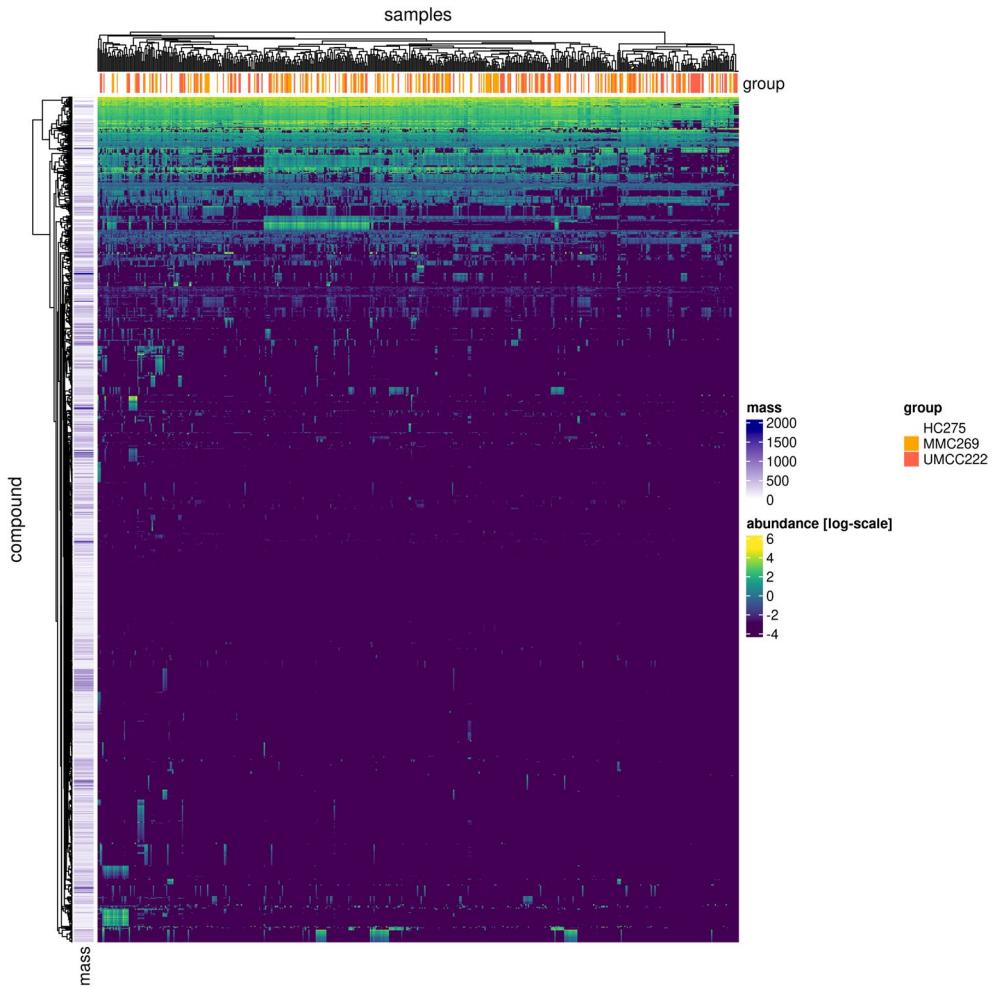
Extended Data Fig. 2 | Foods and nutrients in controlled feeding studies. (**a**) Food abundances in the MBD cohort by diet group (n = 30). Boxplots show 25%, 50%, and 75% quantiles.The center denotes the median and whiskers extend to the smallest and largest data points within 1.5 interquartile ranges. (**b**) Correlation between MEDI estimates and ground truth for varying fecal samples/food diary entry offsets. (**c**) MEDI predictions of total fiber content from fecal DNA (y-axis) and nutrient consumption of sugars, fibers and grains obtained from food diaries (x-axis) in a controlled-feeding study (PATH), where the dietary intake recorded in the daily food record precede the stool sample by at least 48 h. Each point denotes a single individual. For the food diaries, points represent means over all measured intake amounts and error bars denote the standard error of the mean (sd/sqrt(n)), normalized to a 100 g portion (all samples within the offset, 38 individuals with 124 food record diary entries). For the MEDI data, points x-coordinate represent point estimates of intake based on weighting nutrient profiles of food items by food item relative abundance and assuming a 100 g portion. Blue lines denote regression slopes and gray areas represent 95% confidence intervals. Annotations denote correlation coefficient (*r*) and p-value (p) from a Pearson product-moment correlation test.

https://doi.org/10.1038/s42255-025-01220-1



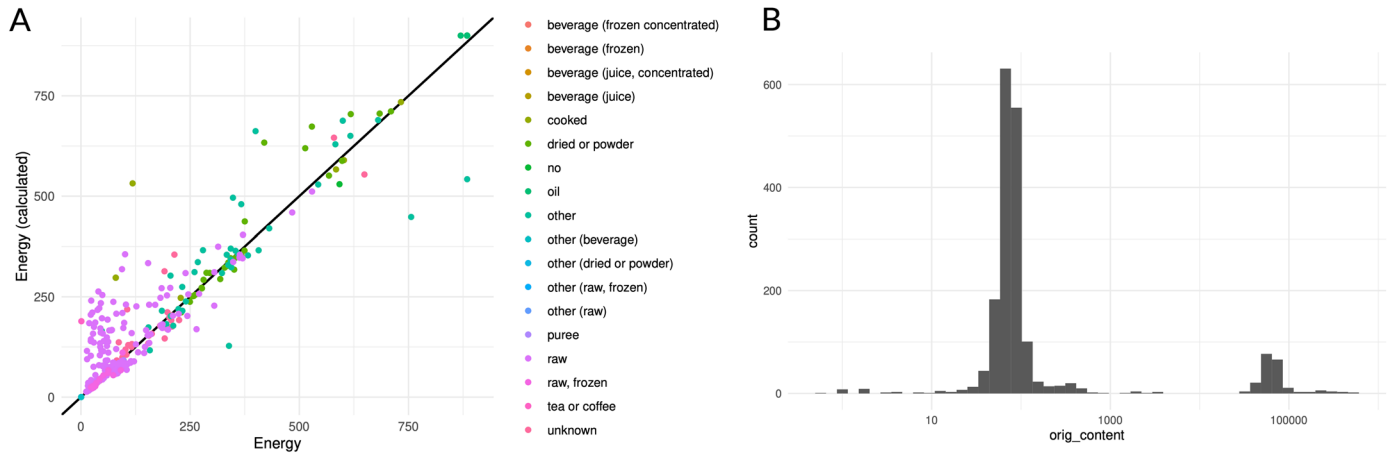
Extended Data Fig. 3 | Non-food reads in infant samples. Relative abundance of bacterial and human reads across infant timeseries, colored by delivery route. Lines denotes a smooth spline regression and shaded areas denotes the 95% confidence interval of the spline regression.

https://doi.org/10.1038/s42255-025-01220-1



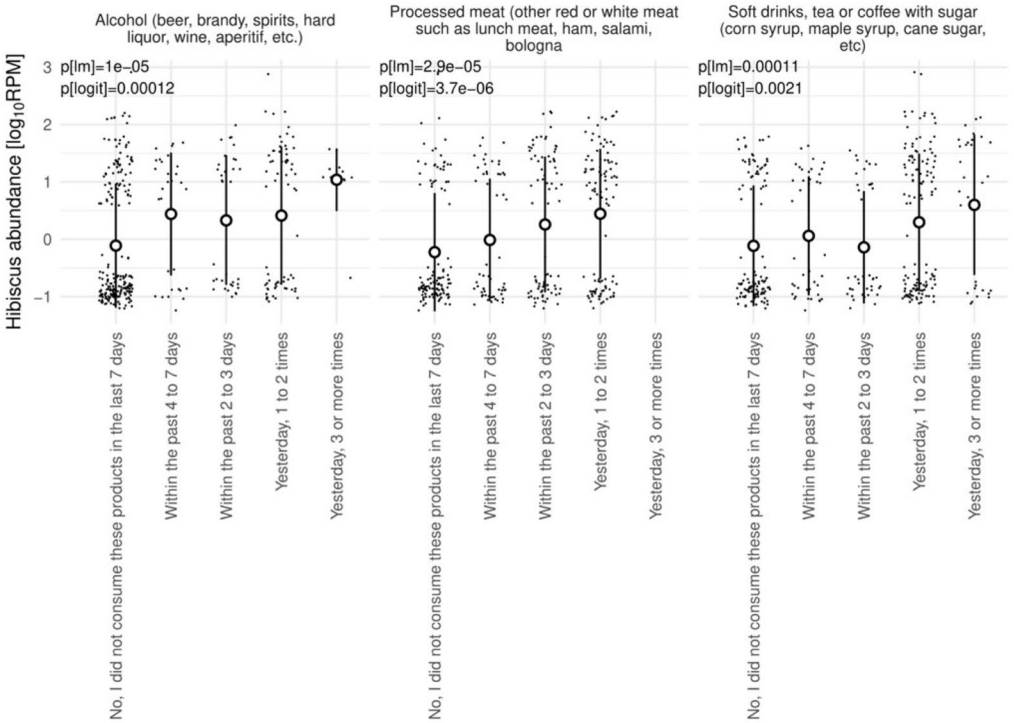
Extended Data Fig. 4 | MEDI dietary intake estimates were associated with metabolic health. Abundances per 100 g portion for 1703 compounds across a cohort of 533 metabolically healthy and unhealthy individuals from the METACARDIS cohort. Fill colors denote abundance per standard portion (mg/100 g). Column annotations denote metabolic health status from the original METACARDIS cohort (HC - healthy cohort, MMC - IHD metabolically matched cohort, UMMC - untreated metabolically matched cohort). Here, MMC and UMMC denote disease-free but metabolically unhealthy groups. Row annotations denote the monomer mass of the compound (in g/mol).

https://doi.org/10.1038/s42255-025-01220-1



Extended Data Fig. 5 | Curation of FOODB data. (**a**) Original content (x-axis) vs. energy content calculated by the Adwater method based on macronutrient content (Pearson $r = 0.94$, two-sided product-moment correlation test $p < 2.2e\text{-}16$). Colors denote detailed unique preparation types in the FOODB. (**b**) Cholesterol abundances across foods in the FOODB before adjustment.

https://doi.org/10.1038/s42255-025-01220-1



Extended Data Fig. 6 | Hibiscus associations. Significant associations between food frequency questionnaires (FFQs) and *Hibiscus* genus abundance in the iHMP cohort (see Methods, n = 361). Associations were run for all 19 FFQ questions. Circles denote the mean and error bar denote standard deviation. p[lm] indicates the ANOVA p-value of a regression of log-transformed relative abundances and p[logit] denotes the p-value of a logistic regression of food occurrence against food frequency strata. Axis labels are common across all plots within this panel. Shown are only food groups with a Bonferroni-adjusted p(lm) < 0.05.

nature portfolio

Corresponding author(s): Christian Diener
Sean M Gibbons

Last updated by author(s): 12/16/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Raw metagenomic shotgun sequencing data was downloaded from the NIH Sequence Read Archive for all data sets using the SRA download pipeline from https://github.com/gibbons-lab/pipelines.
Data analysis	The MEDI 1.0 pipeline was implemented as a set of Nextflow workflows and is available at https://github.com/gibbons-lab/medi which uses DWGSIM 0.1.14, FASTP 0.23.4, KRAKEN2 2.1.3, BRACKEN 2.6.0, R 4.3.0, and SOURMASH 4.8.6. Mapping post-filtering was implemented in architeuthis 0.2.0 available from https://github.com/cdiener/architeuthis. Additional analyses and figures were generated with RStudio 2023.12.1, R 4.3.0, vegan 2.6, LIMMA 3.62.1 and are available as notebooks from https://github.com/gibbons-lab/medi-paper.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data for specific food items is available on https://foodb.ca/downloads. Validation data for nutrient content was extracted USDA Food Data Central and is available at https://fdc.nal.usda.gov/. Individual matched genomic assemblies can be downloaded from Genbank or the Nucleotide database and are listed at https://github.com/Gibbons-Lab/medi-paper/blob/main/db/data/manifest.csv. Metagenomic sequencing data for the studied cohorts are available on the NCBI SRA under accession numbers PRJNA473126 (infants), PRJNA398089 (iHMP), PRJEB37249 (METACARDIS), PRJNA947193 (MBD), and PRJNAXXXXXX (PATH).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Information about the sex of individuals is available in the study-specific metadata at https://github.com/gibbons-lab/medi-paper. As this study investigated the association between food-derived DNA in human stool with phenotypes, we did not stratify or adjust for sex. For infant time series, mixed effects models were run with individual infants as random effects, thus accounting for sex-specific differences.
Reporting on race, ethnicity, or other socially relevant groupings	Information about ethnicity and race is available in the study-specific metadata at https://github.com/gibbons-lab/medi-paper, where it was provided in the original study. We did not study associations of ethnicity or race with human phenotypes in this study.
Population characteristics	Cohort characteristics were the following (mean age and standard deviation, sex F:M, number of individuals): MBD: 30.8 years ± 1.9, 8:9, 17 / PATH: 34.3 years ± 5.4, 47:47, 94 / infants: 134.2 days ± 74.6, 225:222, 447 / iHMP healthy: 31.1 years + 19.9, 146:218, 364 / METACARDIS: 54.6 years ± 13.5, 292:241, 533 .
Recruitment	No participants were recruited in this study. Only public data was used.
Ethics oversight	This study only used publicly available and anonymized data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes for individual cohorts were 34 (MBD, 17 individuals with diet cross-over), 94 (PATH), 447 (infant times series), 364 (iHMP cohort), and 533 (METACARDIS) and are also listed in Table S2. All used data came from previously published studies. MEDI is not a cross-sectional method, thus, cross-sectional power was not assessed. Detection power was assessed with simulated ground truth data and 80% detection power was achieved with 10-100 true positive reads.
Data exclusions	We did not exclude data during the course of the analysis. Prior to the analyses, samples in the iHMP cohort with missing metadata could not be included.
Replication	Mapping accuracy and detection power was evaluated with simulated ground truth sequencing data covering 160 different food items and 200 million reads. Inferred dietary composition was also validated with controlled feeding studies and paired food frequency questionnaires, where available. Food detection and quantification of major macronutrients was successful. Dietary fibers predictions did not significantly associate with 24h food diaries.
Randomization	As this was not an intervention study, randomization did not apply here.
Blinding	As this was not an intervention study, blinding did not apply here.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☒ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.