nature communications

Article

https://doi.org/10.1038/s41467-024-53832-y

Identification of gut microbiome features associated with host metabolic health in a large population-based cohort

Received: 7 June 2024

Accepted: 21 October 2024

Published online: 29 October 2024

Check for updates

Ayya Keshet [1,2] & Eran Segal [1,2]

The complex relationship between the gut microbiome and host metabolic health has been an emerging research area. Several recent studies have highlighted the potential effects of the microbiome's diversity, composition and metabolic production capabilities on Body Mass Index (BMI), liver health, glucose homeostasis and Type-2 Diabetes (T2D). The majority of these studies were constrained by relatively small cohorts, mostly focusing on individuals with metabolic disorders, limiting a comprehensive understanding of the microbiome's role in metabolic health. Leveraging a large-scale, comprehensive cohort of nearly 9000 individuals, measured using Continuous Glucose Monitoring (CGM), Dual-energy X-ray absorptiometry (DXA) scan and liver Ultrasound (US) we examined the functional profile of the gut microbiome, and its relation to 38 metabolic health measures. We identified 145 unique bacterial pathways significantly correlated with metabolic health measures, with 86.9% of these showing significant associations with more than one metabolic health measure. Furthermore, 87,678 unique bacterial gene families were found to be significantly associated with at least one metabolic health measure. Notably, "key" bacterial pathways such as purine ribonucleosides degradation and anaerobic energy metabolism demonstrated multiple robust associations across various metabolic health measures, highlighting their potential roles in regulating metabolic processes. Our results remained largely unchanged after adjustments for nutritional habits and for BMI they were replicated in a geographically independent cohort. These insights pave the way for future research and potentially the development of microbiome-targeted interventions to enhance metabolic health.

The human gut microbiome, identified as a key aspect in host health, has been linked with various conditions including inflammatory bowel disease[1], cancer[2], and cardiovascular health[3]. Its emerging connection with metabolic health, however, is drawing particular attention for its potential to unravel the complexities of several metabolic disorders. In recent years, the significance of its contribution to host metabolic health is increasingly recognized; changes in gut microbiome diversity, and the abundance of specific species were shown to associate and be used to identify T2D[4]; gut microbiome species were linked to the biosynthesis of branched-chain amino acids (BCAAs), which are elevated in individuals with insulin resistance[5]; studies utilizing technological and computational advances devised metagenome-wide association studies revealing functional shifts in bacterial genes in individuals with T2D[6],

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. e-mail: eran.segal@weizmann.ac.il

Table 1 | Baseline characteristics of the study population

Characteristic, mean (STD) or counts %	Male *n* = 4081 (46.1%)	Female *n* = 4778 (53.9%)	All *n* = 8859
Age	51.2 (7.8)	52.3 (7.8)	51.8 (7.8)
Body composition			
BMI (kg/m^2)	26.5 (3.8)	25.6 (4.4)	26.0 (4.1)
Total fat mass (g)	22,958.3 (7992.0)	25,744.5 (8479.0)	24,436.9 (8369.8)
Scanned VAT mass (g)	1180.5 (729.6)	637.6 (452.4)	892.5 (657.2)
Android tissue fat (%)	0.3 (0.1)	0.4 (0.1)	0.4 (0.1)
Liver ultrasound			
Liver sound speed (m/s)	1548.5 (27.8)	1540.7 (27.5)	1544.4 (27.9)
Liver viscosity (Pa.s)	1.7 (0.2)	1.7 (0.3)	1.7 (0.3)
Liver elasticity (kPa)	5.4 (1.0)	4.6 (0.9)	5.0 (1.0)
Liver attenuation (dB/cm/MHz)	0.4 (0.1)	0.4 (0.1)	0.4 (0.1)
Select CGM-derived measures			
eA1C	5.0 (0.4)	4.9 (0.4)	5.0 (0.4)
SD	15.3 (4.5)	15.1 (4.1)	15.2 (4.3)
HBGI	0.2 (0.7)	0.1 (0.3)	0.2 (0.5)
LBGI	2.2 (1.9)	2.5 (1.9)	2.4 (1.9)
MAGE	37.8 (11.9)	37.6 (10.8)	37.7 (11.3)
MODD	13.3 (4.1)	12.7 (3.5)	13.0 (3.8)

BMI body mass index, *VAT* visceral adipose tissue, *eA1C* estimated A1C, *SD* standard deviation, *HBGI* high blood glucose index, *LBGI* low blood glucose index, *MAGE* mean amplitude of glycemic excursions, *MODD* mean of daily differences.

and associations between single-nucleotide-polymorphisms (SNPs) in the gut microbiome and host BMI[7].

Alongside research discoveries, the characterization of metabolic health is constantly expanding with the use of diverse tools. Continuous glucose monitoring (CGM) devices[8] which continuously measure interstitial glucose levels enable a comprehensive understanding of an individual's glucose homeostasis, fluctuations over time and potential disruptions in glucose metabolism. Dual-energy X-ray absorptiometry (DXA) scans are used as simple noninvasive methods to accurately measure body composition factors such as visceral fat and abdominal fat[9]. Liver US is used to assess liver fibrosis and the progression of non-alcoholic fatty liver disease (NAFLD)[10]. Use of these tools to characterize and investigate metabolic health revealed further connections between gut microbiome and metabolic health. A study integrating CGM and gut microbiome showed that glycemic response to identical meals is highly variable among healthy individuals, and that gut microbiome improves prediction of glycemic response to meals[11]. A recent study found positive associations between species from the *Firmicutes* and *Proteobacteria* phylums and mean glucose as measured by CGM[12]. Researchers recently showed that the gut microbiome is crucial for hepatic gluconeogenesis, pointing to specific changes in amino acids that are important for controlling blood sugar levels[13], and that gut microbiome composition of individuals with NAFLD differs from that of healthy controls[14]. Several studies have shown that the gut microbiome alpha-diversity is decreased in individuals with overweight, with increased levels of the *Firmicutes* phylum[15], which was also shown to associate with differences in visceral adipose tissue (VAT) volume[16].

Despite the significant strides made in understanding the gut microbiome's influence on metabolic health, our knowledge remains fragmented. While studies have highlighted associations between the microbiome and metabolic factors such as glucose homeostasis and obesity[17], a comprehensive understanding of the functional mechanisms by which the microbiome influences these processes is still evolving. Recent research has begun to shed light on the metabolic pathways influenced by the gut microbiome, revealing potential links to energy metabolism, inflammation, and insulin sensitivity. Several studies have established the relation between gut microbiome composition and diversity to short-chain fatty acids (SCFAs) production, which in turn regulates several metabolic pathways, is involved in obesity, insulin resistance and T2D[18]. Gut microbiome dysbiosis has also been linked to alterations in bile acid (BA) metabolism, which is crucial for liver health[19]. In addition, research has identified gut bacteria associated with insulin resistance and insulin sensitivity, and demonstrated that these bacteria improve host insulin resistance in mouse models[20]. Recent research has increasingly focused on the functional aspects of the gut microbiome. A study demonstrated that adherence to a Mediterranean-style diet is linked with specific functional components of the gut microbiome, potentially mediating its protective effect against cardiometabolic diseases[21]. A more recent study found an elevated abundance of functions related to bacterial cellular metabolism, particularly favoring glycolysis, in patients with type 2 diabetes[22].

While previous studies, some conducted in well-phenotyped cohorts, and investigated both the composition and functional aspects of the gut microbiome[21,22], provided valuable insights, we are still far from understanding the role of the gut microbiome in host metabolic health. Previous research shows that large cohort sizes are crucial for accurately capturing associations between the microbiome and host phenotypes[23]. Expanding research to large, predominantly healthy populations can enhance the generalizability of findings and contribute to a more comprehensive understanding of the microbiome's role in metabolic health. In this study, we aim to bridge some of these gaps by leveraging a unique cohort collected as part of the Human Phenotype Project (HPP)—a large scale, prospective, longitudinal study[24], with 8859 non-diabetic adults aged 40–70, measured through CGM, DXA scan, liver US and gut microbiome metagenomic sequencing. We analyzed the functional profile of the gut microbiome, as the relative abundance of bacterial gene families and pathways obtained using HUMAnN[25], and their associations with 38 metabolic health measures. Associations were studied in a framework similar to that of genome-wide association studies (GWAS) with adaptations to metagenomic data, uncovering bacterial gene families and pathways in the gut microbiome potentially involved in host glucose homeostasis and metabolic health.

Results

Study population and design

A total of 8859 participants from the HPP[24] who were measured with CGM, DXA scan, liver US and gut microbiome metagenomic sequencing were analyzed—4778 (53.9%) women and 4081 (46.1%) men, with an average age of 51.8 (7.8) years and average BMI of 26.0 (4.1) (kg/m^2). Thirty-eight metabolic health measures were defined as the outcomes of interest (see "Metabolic health measures" in Methods). Table 1 summarizes the main characteristics of the participants, with select CGM-derived measures (see Table S1 for the complete characteristics with all CGM measures, and "Study population" in Methods for the complete description of the HPP).

Gut microbiome metagenomic sequencing data was functionally profiled with HUMAnN[25] and MetaPhlAn[26] to obtain bacterial gene families and pathways abundances (see "Microbiome sample collection and processing" and "Metagenomic reads mapping and functional profiling" in Methods). The 38 metabolic measures of interest were divided into 10 categories: Body composition including BMI and measures from DXA scan, Liver US, and 8 categories of CGM-derived measures (see "Metabolic health measures" in Methods).

Associations between each bacterial gene family and bacterial pathway, with each metabolic health measure were assessed using linear regression models, adjusted for age and sex, in a similar way to genome-wide association studies (GWAS). Associations were

https://doi.org/10.1038/s41467-024-53832-y

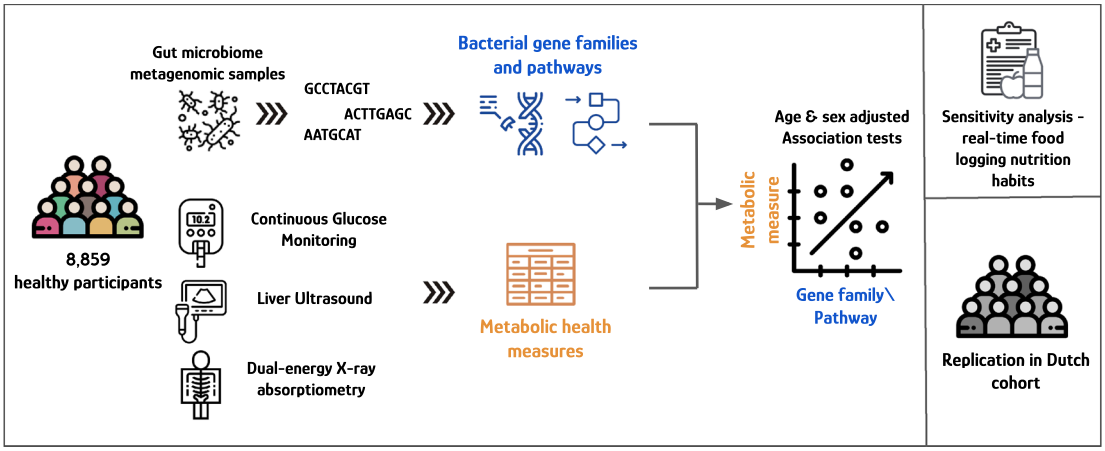

Fig. 1 | Study overview. Illustration of the study design.

considered significant if found to have a Bonferroni-corrected *p*-value < 0.05 (see "Bacterial gene families/pathways−metabolic measures associations" in Methods). Overview of the study design is illustrated in Fig. 1.

Bacterial gene families and pathways associate with metabolic health measures

Overall, 565,398 unique bacterial gene families and 346 unique bacterial pathways were associated with 38 metabolic health measures. We found 87,678 (15.5%) unique bacterial gene families significantly associated with metabolic health measures in 284,759 different correlations, and 145 (41.9%) unique bacterial pathways significantly associated with metabolic health measures in 862 different correlations. Most significant associations were negative−both for the bacterial gene families (239,077 (83.96%)) and the bacterial pathways (621 (72.0%)). From the 145 bacterial pathways that were significantly associated with at least one metabolic health measure, most pathways (126 (86.9%)) were significantly associated with two or more metabolic health measures. Similarly from the 87,678 bacterial gene families, the majority of the bacterial gene families (64,199 (73.22%)) were associated with two or more metabolic health measures.

The body composition category showed the highest numbers of significant associations, both to bacterial gene families and bacterial pathways (Fig. 2). 21,871 unique bacterial gene families and 14 unique bacterial pathways were significantly correlated with at least one metabolic health measure from this category, and showed no significant associations with any measures from other categories. Most bacterial gene families and pathways were correlated with more than one measure in this category, yet some unique associations were found. Android fat tissue (%), the measure with the highest number of bacterial gene families associations, had 12,093 bacterial gene families uniquely associated solely with it and no other metabolic health measure that was tested. BMI had the highest number of bacterial pathways associated with it, and 10 bacterial pathways were uniquely correlated solely with it. All measures in this category showed a similar pattern, with the majority of association, both for bacterial gene families and bacterial pathways, being negative associations.

Diverse association patterns were observed across other categories. Liver viscosity and liver elasticity[10] showed no significant associations with any bacterial gene family or pathway. Liver sound speed demonstrated predominantly positive significant associations with both bacterial gene families and bacterial pathways. In contrast, liver attenuation, which describes the reduction in ultrasound beam intensity as it traverses the liver and has been linked to clinical liver state indicators[27], exhibited almost only negative significant associations with bacterial gene families and bacterial pathways.

CGM-derived measures showed variable association patterns. Measures of short-term variability; SD within series (SDwsh)[28], Standard Deviation of the Rate of Change (SD.Roc)[29] and Mean Absolute Glucose (MAG)[30], were associated with a high number of bacterial gene families and pathways compared to other categories of CGM-derived measures (Fig. 2). Similar to body composition most associations with these short-term variability measures were negative. Measures of hyperglycemia−% above 180, HBGI[31] and Hyper index[32]−showed the lowest numbers of significant associations with bacterial gene families and pathways. eA1C[33] and GMI[34], both transformations of the mean-glucose measured from CGM, showed similar association patterns, and were found to be significantly associated with 3396 bacterial gene families and 14 pathways.

Short-term variability, between-day variability, and within-day variability measures from CGM showed similar association patterns within their category, in terms of the number of significant associations and the correlation direction. Overall variability measures showed a differing pattern where, for example, the IQR (interquartile range) was significantly associated with 2967 bacterial gene families and 6 pathways, while the CV (coefficient of variation) was only significantly associated with 25 bacterial gene families and 3 pathways (Fig. 2).

Despite most associations across body composition, Liver US and CGM-derived measures in the different categories being negative, for measures of hypoglycemia−% below 70, LBGI[31] and Hypo index[32]−most associations to bacterial gene families were positive, as well as for ADRR[35] and IGC[32] (Fig. 2). The Average Daily Risk Range (ADRR) is calculated to quantify the daily glucose variability and risk, integrating both hypoglycemic and hyperglycemic excursions into a single metric that reflects overall glycemic risk. It is derived by averaging the daily risk scores, which assess the likelihood and severity of glucose deviations from the normal range. The Index of Glycemic Control (IGC) provides a composite measure that combines average glucose levels, variability, and episodes of hyper- and hypoglycemia, giving a holistic view of glucose control efficacy and stability.

To address high correlations between bacterial gene families, and identify independent bacterial gene families associated with the examined metabolic health measures, two approaches were implemented. Upfront gene filtering was employed prior to the association analysis to obtain uncorrelated representative gene families (see "Upfront gene filtering" in Methods), resulting in 55,925 pre-filtered bacterial gene families that were then associated with metabolic health measures. Clumping, commonly used in GWAS analyses (see "Gene clumping" in Methods), was employed post association analysis, to each metabolic health measure individually, to obtain a subset of independent associations. Both approaches resulted in similar

https://doi.org/10.1038/s41467-024-53832-y

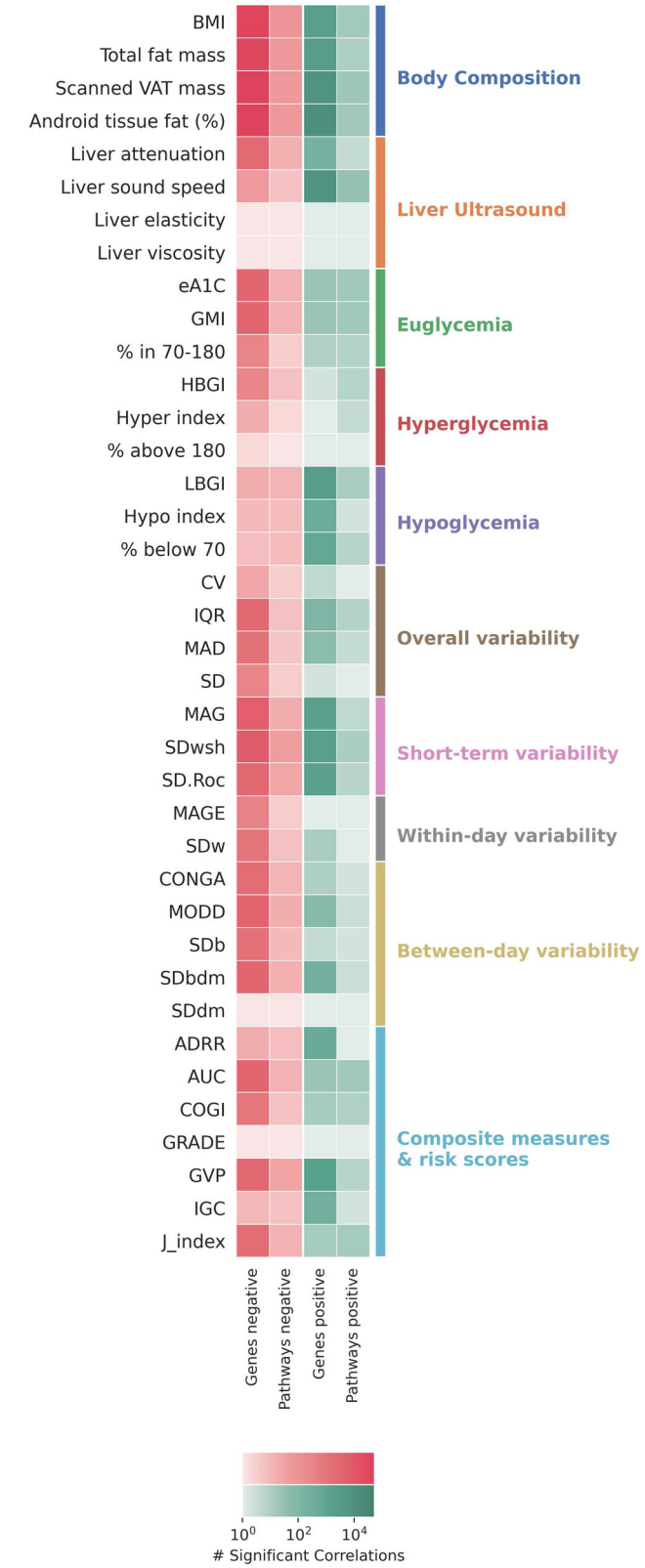

Fig. 2 | Number of significantly associated bacterial gene families and pathways with metabolic health measures. Heatmap showing the number of positive and negative significant associations between each metabolic health measure and bacterial gene families and pathways. Associations were obtained using two-sided linear regression adjusted for age and sex, and corrected for multiple comparisons using the Bonferroni method. Associations with Bonferroni-corrected p-value < 0.05 were considered significant. Colored bar on the right shows the category assigned for each metabolic health measure.

numbers of significant uncorrelated bacterial gene families associated with metabolic health measures (Figs. S1 and S2). In both, Android fat tissue (%) was the measure with the highest number of independent bacterial gene families associations (1342 from clumping, 2217 from pre-filtering), followed by Scanned visceral adipose tissue (VAT) mass (1317 from clumping, 2013 from pre-filtering) (Figs. S1 and S2). While measures of short-term CGM variability, had hundreds or tens of independent associations with bacterial gene families, for example SD.wsh (148 from clumping, 140 from pre-filtering) and SD.Roc (87 from clumping, 78 from pre-filtering), measures of overall variability, within-day variability, and between-day variability were found to have only several to tens of independent associations. The IQR and MAD were associated with 2967 and 1231 bacterial gene families accordingly. Clumping identified only 14 and 11 independent clumps of bacterial gene families (Fig. S1), and pre-filtering identified 45 and 18 uncorrelated bacterial gene families (Fig. S2).

"Key" associated bacterial gene families and pathways

Aiming to narrow our focus and pinpoint crucial bacterial gene families and pathways, we devised a methodology to select "key" bacterial gene families and pathways—prioritizing entities based on their significance across multiple metabolic health measures and the robustness of their associations, defined by their Bonferroni-corrected p-values (see "Selecting "key" associated bacterial gene families and pathways" in Methods). 5 bacterial pathways and 5 independent bacterial gene families were selected as "key" entities, for each group of metabolic health measures: Body composition, Liver US and all CGM-derived measures. This process results in 11 unique bacterial pathways and 15 unique bacterial gene families that were chosen as "key" entities. Figure 3 displays a heatmap of the correlations of these "key" entities with the metabolic health measures. Tables S2 and S3 show for each "key" bacterial pathway and gene family the metabolic health groups it was chosen for.

The 11 "key" pathways clustered to 3 groups, with distinct correlation patterns to the tested metabolic health measures (Fig. 3A, Table S2). Cluster I, showing strong positive correlations to most metabolic health measures, with negative correlations to hypoglycemia measures derived from CGM, included 2 pathways with roles in the biosynthesis of amino acids; superpathway of L-aspartate and L-asparagine biosynthesis (ASPASN-PWY) and L-methionine biosynthesis IV (PWY-7977). High ratio of asparagine to aspartate was shown to be associated with an increased risk for T2D[36], and aspartate was significantly associated with decreased insulin secretion, elevation of fasting glucose levels and increased risk of T2D[37]. Methionine metabolism have been previously linked to metabolic syndrome and related diseases, and elevated serum S-adenosylmethionine (SAM) was found in individuals with NAFLD[38]. In addition, Methionine levels have been found to significantly associate with Matsuda insulin sensitivity index in a long-term follow-up study of ~5000 Finnish men[37]. The third pathway in this cluster was PWY0-1586, responsible for peptidoglycan maturation.

Pathways in cluster II and III showed an opposite correlation pattern to Cluster I with negative correlations with body composition and most CGM-derived, Liver US measures, and positive correlations with hypoglycemia measures. Cluster III showed stronger correlation, with higher correlation coefficients compared to Cluster II. Cluster II includes NAD salvage pathway II (PNC IV cycle) (PWY-7761) and purine ribonucleosides degradation (PWY0-1296). NAD+ metabolism plays a key role in insulin sensitivity and is, at times, disrupted by obesity and age[39,40]. Elevating NAD+ levels, using nicotinamide mononucleotide (NMN), a key NAD+ intermediate, was shown to correct metabolic disturbances and ameliorate glucose intolerance and insulin resistance in T2D mice[41]. Purine dysregulation has been tied to several metabolic diseases, including gout and metabolic syndrome[42], which are often

https://doi.org/10.1038/s41467-024-53832-y

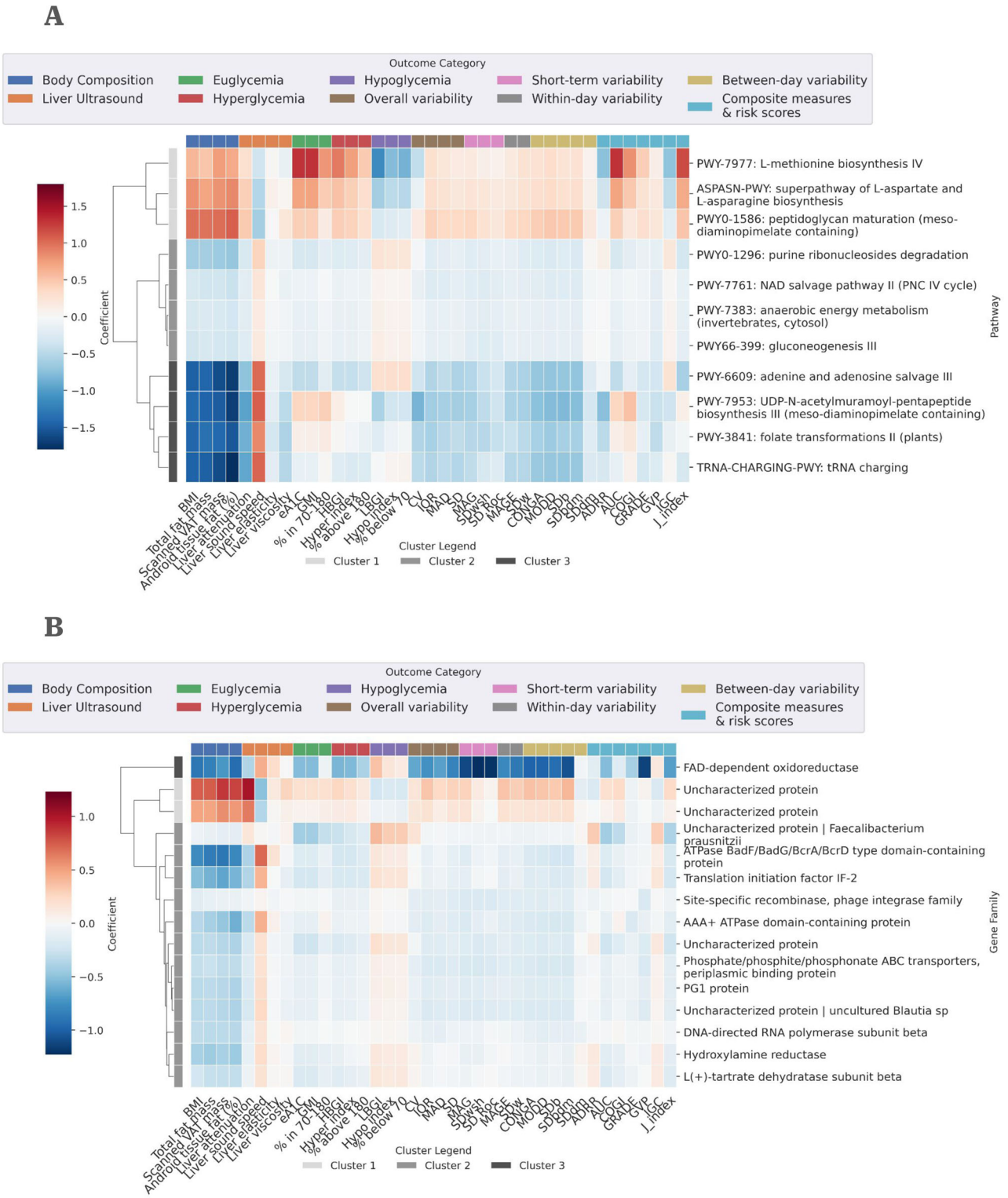

Fig. 3 | Heatmap of "key" bacterial pathways and gene families. Heatmaps showing the correlation coefficient of the **A** 11 "key" bacterial pathways and **B** 15 "key" clumped bacterial gene families with the highest number of significantly associated metabolic health measures and lowest *p*-value ranking across groups (see "Selecting "key" associated bacterial gene families and pathways" in Methods). Associations were performed using two-sided linear regression adjusted for age and sex. Multiple comparisons were corrected using the Bonferroni method, and associations with Bonferroni-corrected *p*-value < 0.05 were considered significant.

accompanied by insulin resistance. The purine degradation pathway was shown to be enriched in the gut microbiome of individuals with obesity in a study of a small Korean population[43]. The two additional pathways in this cluster—gluconeogenesis III and anaerobic energy metabolism (invertebrates, cytosol)—have known, critical roles in energy and glucose metabolism, yet to the best of our knowledge, no current evidence connect them directly to any of the examined host metabolic health measures.

Cluster III, with the same correlation pattern as Cluster II yet with stronger associations, included PWY-6609: adenine and adenosine salvage III. This pathway is part of the purine salvage pathway, which recycles adenine and adenosine. The adenosine system has been highlighted for its critical role in regulation of insulin and glucose homeostasis in individuals with T2D, and its receptor system was associated with development or progression of diabetes mellitus, with specific focus on T2D[44]. Another pathway in this cluster is PWY-3841: folate transformations II. Folate depletion has been previously reported to potentially cause oxidative stress in the liver, leading to the development of NAFLD[45]. Folate deficiency has also been tied to obesity, and lower levels of folate have been observed in individuals with obesity[46]. The two remaining pathways in Cluster III are TRNA-CHARGING-PWY: tRNA Charging, a pathway essential for protein synthesis, enabling the attachment of amino acids to their corresponding tRNA molecules, and PWY-7953: UDP-N-acetylmuramoyl-pentapeptide biosynthesis III, a pathway involved in the synthesis of bacterial cell wall peptidoglycan, which is crucial for bacterial growth and survival.

"Key" clumped bacterial gene families showed two correlation patterns, similar to those observed in the "key" bacterial pathways (Fig. 3B). "Key" bacterial gene families were annotated using the UniRef database[47]. Cluster I, showing mostly positive correlations with metabolic health measures, included 2 bacterial gene families, chosen by their associations with the metabolic health measures in the Liver US group (Table S3), both uncharacterized proteins. Cluster III, containing a single bacterial gene family—FAD-dependent oxidoreductase, showed an opposite correlation pattern to Cluster I, with this bacterial gene family associating negatively with almost all metabolic health measures. Cluster II of the "key" bacterial gene families includes several protein families involved in energy metabolism, such as ATPase BadF/BadG/BcrA/BcrD Type Domain-Containing protein, AAA+ ATPase domain-containing protein and Phosphate/phosphite/phosphonate ABC transporters and periplasmic binding protein. "Key" pre-filtered bacterial gene families showed slightly stronger associations to metabolic health measures than those of the "key" clumped gene families (Figs. 3B and S3), with a single correlation pattern of mostly negative correlations and positive correlations only to hypoglycemia measures derived from CGM (Fig. S3). Cluster II encompasses a broad spectrum of gene families (Table S4), including those with recognized enzymatic roles in metabolism such as Pyridoxamine 5′-phosphate oxidase family protein[48] and Aminotransferase, which previous research highlighted as potentially involved in the pathogenesis of metabolic syndrome[49]. Several of the "key" bacterial gene families in the clumping analysis and pre-filtering analysis were uncharacterized proteins sourced from the species *Faecalibacterium prausnitzii* (Figs. 3B and S3); one of the most abundant bacterial species in the colon of healthy humans. Changes in *Faecalibacterium prausnitzii* abundances were related to various intestinal and metabolic diseases[50].

Community-level contributions to bacterial pathways

The "key" bacterial pathways identified in the previous analysis represent specific functionalities associated with metabolic health measures. These functionalities could result from two different biological scenarios: (1) a community-level functionality where multiple microbial species or genera contribute to an overall functional capacity, or (2) a scenario where a single microbial species or genus predominantly encodes the functionality. To distinguish between these possibilities, we utilized the HUMAnN generated abundance contributions from specific organisms[25] to the abundance of bacterial pathways. For each bacterial pathway, we identified the contributing genera and species (see "Metagenomic reads mapping and functional profiling" in Methods) and examined the distribution of their contributions in terms of RPKs.

Of the 346 bacterial pathways associated with metabolic health measures, 74 (21.4%) were attributed to unclassified bacteria, while the majority, 272 (78.6%), were linked to known bacterial species and genera. The number of contributing genera and species varied across bacterial pathways (Fig. S4), with an average of 13.02 (±20.03) species and 7.4 (±10.05) genera contributing to the abundance of each bacterial pathway. Notably, most pathways (231 (66.8%)) had their abundances contributed by multiple species.

The "key" bacterial pathways in Cluster I exhibited a diverse contribution from multiple genera, with 10–24 genera, and 22–53 species contributing to each bacterial pathway's abundance (Figs. S5 and S6). This suggests a community-level functionality where several genera and species are collectively responsible for the metabolic functions associated with these pathways. Similarly, the pathways in Cluster III also showed contributions from a wide range of genera and species, with 16–33 genera and 40–65 species contributing to each pathway's abundance (Figs. S9 and S10). This pattern further supports the concept of community-level functionality for these "key" bacterial pathways. In contrast, the "key" bacterial pathways in Cluster II were predominantly associated with unclassified bacteria (Figs. S7 and S8), meaning we currently lack sufficient taxonomic resolution to attribute these functions to specific known genera or species. As a result, it remains unclear whether these pathways are driven by a single microbial taxon or reflect a broader community-level functionality. Only one pathway in this cluster, PWY0-1296, had contributions from 19 recognized genera and 23 recognized species (Figs. S7 and S8).

The variable distributions of the number of contributing genera and species, and the diversity of genera and species contributing to the "key" pathways suggests that these likely represent community-level functionalities associated with metabolic health measures. In addition, the presence of unclassified contributions for some pathways highlights the importance of analyzing gut microbiome data at the functional level, as there are still under-characterized taxonomic groups that could lead to a loss of valuable information.

Sensitivity analysis—nutritional habits from real-time food intake logging

Nutritional habits potentially affect host metabolic health measures such as body composition, glucose control, as measured by CGM[51], and gut microbiome composition and function[52]. To address this, we integrated information from real-time food intake logging, reported by participants during their CGM connection, and repeated all association analyses adjusting for 5 additional nutritional summaries; Healthy Food Diversity (HFD) Index[53], median total caloric intake, and median amount of calories (as percentage) consumed from carbohydrates, proteins and lipids (see "Nutritional summaries as sensitivity analysis" in Methods). 862 significant bacterial pathway associations and 284,759 significant bacterial gene families associations with metabolic health measures were discovered in our main analysis. Of these, 539 (62.53%) bacterial pathway associations and 193,927 (68.1%) bacterial gene families associations replicated when adjusting for nutritional summaries. All replicated associations replicated with the same correlation direction.

Android tissue fat (%) retained the highest number of associations with bacterial gene families, with 53,186 significant associations replicating 80.2% of the significant associations found in the main analysis. BMI retained the highest number of associations with bacterial pathways, with 99 significant associations replicating 98.1% of the significant associations found in the main analysis. Body composition metabolic health measures in general largely preserved their significant correlations, as well as liver attenuation and liver sound speed from liver US, as illustrated in Fig. S11. Euglycemia measures derived from CGM also showed similar correlations after adjusting for nutritional summaries. The noticeable decrease in significant associations for overall, short-term, and within-day variability measures after

https://doi.org/10.1038/s41467-024-53832-y

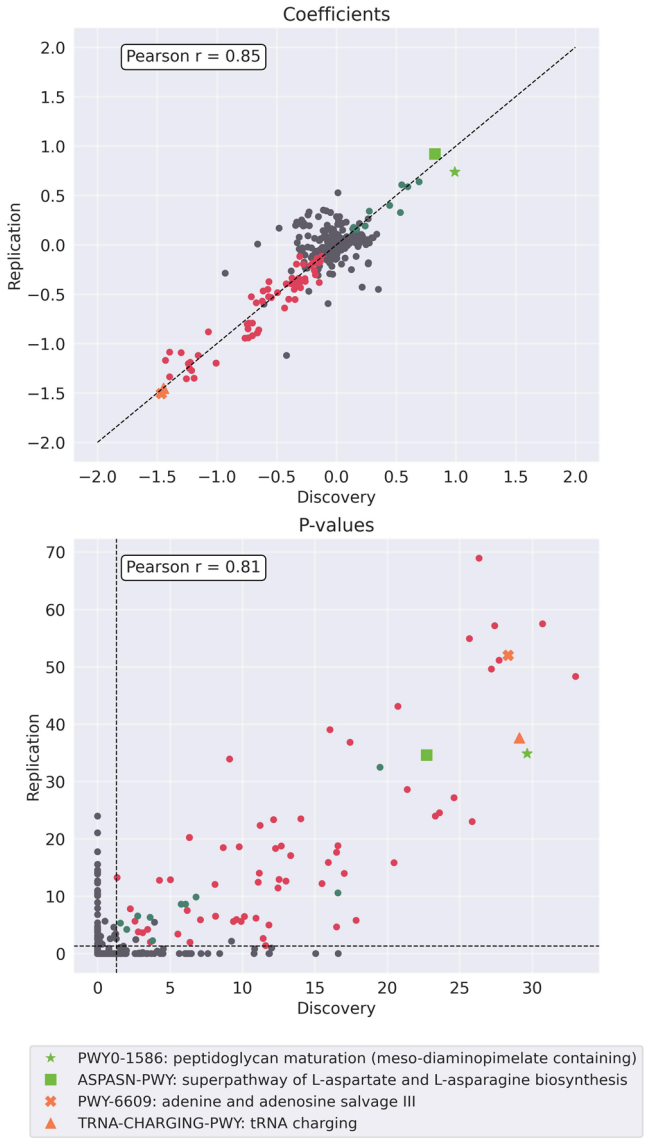

Fig. 4 | Bacterial pathways correlate with BMI replicate in an independent cohort. Coefficients (top) and *P*-values (−log scale) (bottom) of the associations between bacterial pathways and BMI, in the discovery vs. the replication cohort. Associations were performed using two-sided linear regression adjusted for age and sex. Multiple comparisons were corrected using the Bonferroni method, and associations with Bonferroni-corrected p-value < 0.05 were considered significant. Colored dots present bacterial pathways with significant associations in both cohorts, and with either a negative (red) or positive (green) correlation coefficient in both cohorts. Light green (Square and Asterix) mark the top two bacterial pathways with positive coefficients in both cohorts, Orange (Triangle and X) mark the top two bacterial pathways with negative coefficients in both cohorts.

adjusting for nutritional summaries aligns with prior findings that highlighted the correlations between nutritional summaries and CGM variability measures[54] (Fig. 2, Fig. S11). MAGE, a within-day glucose variability measure known to be affected by nutritional habits, such as carbohydrate consumption[55,56], had 352 bacterial gene families and 3 pathways associated with it prior to the adjustment to nutritional habits. Following the adjustment, no significant associations were found. MODD on the other hand, a day to day variability measure[57], had 4438 and 19 significant associations with bacterial gene families and pathways in the main analysis, and 436 (9.8%) bacterial gene families and 5 (26.3%) pathways associations remained significant when adjusting to nutritional summaries, suggesting these are indeed associations of the bacterial functional levels with MODD, not mediated nor confounded by diet.

Repeating the analysis of "key" pathways with the adjusted associations for nutritional summaries, we observed that 7 of the 11 "key" pathways remained unchanged, and showed similar correlation patterns to the main analysis (Fig. 3A, Fig. S12). The 13 "key" pathways after adjusting to nutritional summaries, were also clustered to 3 groups, with similar correlation patterns as the 11 "key" pathways in the main analysis (Tables S2 and S5). Six additional "key" bacterial pathways were introduced in the sensitivity analysis. PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II, PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis and PYRIDNUCSYN-PWY: NAD de novo biosynthesis I (from aspartate), in cluster I, showing mostly positive correlations to metabolic health measures. Bacterial pathway PWY-6122 was previously found to be enriched in individuals with hyperuricemia (HUA) and high levels of liver enzymes[58]. HUA, characterized by elevated uric acid levels in the blood, is linked to both insulin resistance[59] and diabetes[60]. In cluster II, two bacterial pathways revealed in the sensitivity analysis were ARGSYN-PWY: L-arginine biosynthesis I (via L-ornithine) and ARGSYNBSUN-PWY: L-arginine biosynthesis II (acetyl cycle), both related to the biosynthesis of arginine and showing mostly negative correlations to metabolic health measures. Previous works suggest that L-arginine may have potential to prevent and/or relieve type 2 diabetes via restoring insulin sensitivity in vivo[61]. The last "key" pathway revealed in the sensitivity analysis was in cluster III, PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing), showing mostly negative correlations to metabolic health measures.

Associations with BMI replicate in an independent cohort

Utilizing a cohort of 8205 Dutch individuals, with gut microbiome metagenomic samples and information on age, sex, and BMI[62] we sought to assess the replicability of our results in an independent cohort. 8205 individuals, of them 3451 (42.1%) men and 4754 (57.9%) women were analyzed, with an average age of 48.4 (14.8) years and an average BMI of 25.6 (4.4) (kg/m^2) (Table S6). Metagenomic samples were processed in the same manner as in the discovery cohort, and association models were run only against BMI as the only metabolic health measure available in the replication cohort (see "Replication in independent cohort" in Methods).

128 pathways and 66,697 bacterial gene families were significantly correlated with BMI in the discovery cohort. 83 (64.84%) pathways were also significantly correlated with BMI in the replication cohort, where 78 (93.9%) replicated with the same correlation direction. 33,302 (49.9%) bacterial gene families were significantly replicated in the replication cohort, where 33,294 (99.9%) replicated with the same correlation direction.

Figure 4 displays the association coefficients and p-values of bacterial pathways, from the discovery and the replication cohorts, illustrating a consistent pattern in the distribution of significantly correlated pathways, both negative and positive. Importantly, the pathways that exhibited the strongest correlations—identified by the highest association coefficients from the significantly associated bacterial pathways —were consistent across both cohorts (Fig. 4). The top two bacterial pathways with the strongest positive correlations with BMI in both cohorts were PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing) and ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis. These pathways, also identified as "key" pathways in our main analysis, demonstrated multiple significant correlations with metabolic health measures in the discovery cohort. The top two pathways that exhibited the strongest negative correlations with BMI in both cohorts were PWY-6609: adenine and adenosine salvage III, and TRNA-CHARGING-PWY: tRNA charging, both also selected as "key" bacterial pathways in the discovery cohort analysis, as well as the sensitivity analysis adjusting for

nutritional summaries. The similar patterns of correlations coefficients and *p*-values were also found for bacterial gene families, as depicted in Fig. S13.

Discussion

In recent years the human gut microbiome has emerged as a relevant factor in many host metabolic disorders, including obesity, T2D, non-alcoholic fatty liver and cardiometabolic diseases[17]. Microbiome composition, richness and functionality has been found to differ in individuals with metabolic disorders[17,63]. This led to the attempts to harness the microbiome as a therapeutic target; probiotic, prebiotic and synbiotic supplements were examined to improve glucose control measures as fasting glucose and HbA1C in individuals with T2D and prediabetes[64]: high fiber diet demonstrated protective impact on the gut ecosystem which in turn improved glucose homeostasis in individuals with T2D[65]: fecal microbiota transplantation (FMT) was tested as a therapeutic for several metabolic diseases as obesity, insulin resistance, and metabolic syndrome[66].

There are several mechanisms by which the gut microbiome possibly affects host metabolic health; Damage to the gut barrier and increased gut permeability results in increased translocation of bacterial endotoxins, mainly lipopolysaccharide (LPS), which might worsen glucose homeostasis via increased metabolic endotoxemia[67]: Gut bacteria are involved in the metabolism of various metabolites that play a role in host metabolic health, among them SCFAs, associated with several metabolic diseases and BCAAs, which irregularities in their metabolism contribute to obesity and T2D[68]: The gut microbiota regulates BA homeostasis through different bio-transformation, and recent studies support the role of BA in NAFLD ang glucose control[67,69]. Yet, despite the advances in both research and clinical work - many gaps remain in our knowledge on the functional bacterial pathways in which the gut microbiome affects host metabolic health[17].

In this study, we aimed to bridge this gap by using a large-scale unique cohort of more than 9000 individuals, measured with gut microbiome metagenomic sequences, DXA scans, liver US and CGM[24]. Analyzing the gut microbiome on a functional level of bacterial gene families and pathways, associating them with clinically validated CGM-derived metrics, body composition and measures from liver US, we uncovered 87,678 unique bacterial gene families and 145 unique bacterial pathways significantly associated with metabolic health measures. The majority of the bacterial gene families' significant associations were to body composition such as Android fat tissue (%) and BMI, with 21,871 unique bacterial gene families associated with at least one metabolic health measure of body composition, and no metabolic health measure from other categories of liver US and CGM-derived measures. 86.9% of bacterial pathways were associated with 2 or more metabolic health measures. Interestingly, most of the associations identified were negative, indicating an inverse relationship between certain bacterial functions and metabolic health measures of the host. This pattern is somewhat aligned with previous findings, where gut microbiome diversity and gene count were inversely associated with metabolic health measures such as obesity[70], insulin resistance and dyslipidemia[71]. This high percent of negative associations might point to unexplored microbial-host interactions impacting the host metabolic health, and further investigations are necessary to fully understand these.

Several "key" bacterial gene families, with multiple robust associations across metabolic health measures, were annotated as uncharacterized proteins, yet were sourced from species with known relations to host metabolic health—such as *Faecalibacterium prausnitzii (F.prausnitzii)* and *Blautia*. *F.prausnitzii* is known for its protective effects against metabolic disorders[50], and changes in its abundance have been linked to obesity and T2D[50], and a recent study also found several SNPs in an energy metabolism coding region of *F.prausnitzii* associated with BMI in healthy individuals[7]. Different species of *Blautia* have been tied to Visceral Fat Accumulation (VFA)[72], and indicators of impaired lipid and glucose metabolism[73]. The fact that these "key" bacterial genes are uncharacterized proteins, and are currently unknown, stresses the need to further study and annotate microbial genes and functions. "Key" bacterial pathways identified were involved in the metabolism of several amino-acids and metabolites previously tied with obesity, insulin sensitivity and T2D such as aspartate, asparagine, methionine, leucine, NAD+, 5-aminoimidazole ribonucleotide and purine[36,37,58–60,74]. Only a few of these bacterial pathways were shown to be associated with metabolic disorders[58], and for most this is the first time, to the best of our knowledge, that a direct connection is shown between them and metabolic health measures such as glucose control measured by CGM and measures from liver US. Taxonomy contributions analysis revealed variability in the number of species and genera contributing to the abundance of bacterial pathways analyzed, with a mean of 13.02 (±20.03) species and 7.4 (±10.05) genera contributing to the abundance of each bacterial pathway. Several diverse genera contributed to the abundance of the majority (8 of 11) of the "key" bacterial pathways, while the remaining "key" pathways (3 of 11) all originated from unclassified genera, possibly due to lack of taxonomic resolution. These findings support the assumption that the associations identified present community-level functionalities that associate with host metabolic health, and further stress the importance of moving towards functional-based analysis of the microbiome[75].

Sensitivity analysis revealed that many associations remained significant when adjusting for nutritional summaries obtained from real-time food loggings. Specifically, associations with body composition measures and liver US measures were almost not affected and most remained significant. CGM derived measures of variability, which were previously shown to be correlated with nutrition[54], showed a decrease in the number of significant associations. Seven of the 11 "key" bacterial pathways were still selected as "key" bacterial pathways following adjustment for nutritional summaries. These 7 replicated "key" bacterial pathways were also significantly associated with BMI, in a technically and geographically independent cohort[62], along with 64.8% of the bacterial pathways from the discovery cohort. These findings suggest that the associations detected here point to potential functional relations between the gut microbiome and host metabolic health measures, not affected by dietary habits, and environmental factors as they replicated in a population with differing background and environment. Since the replication cohort only contained information on BMI, only associations to BMI were tested for replication, and future research will be needed to examine the replication of associations with other metabolic health measures.

While our study provides valuable insights into the interactions between the gut microbiome and host metabolic health, several limitations must be acknowledged. One significant limitation is the estimation of bacterial functional level based on DNA sequencing. Despite the advances in both tools and databases, functional estimations from metagenomic data are still limited by the accuracy and completeness of reference databases[76]. The reliance on reference genomes can introduce biases, as these databases may not capture the full diversity of the microbiome, particularly for less well-studied or novel bacterial species. Another limitation is the uniformity of our study population, which is a relatively healthy one. Since BMI being the only metabolic health measure available in our replication cohort, the additional associations revealed in this study will need further validation in diverse populations to ensure their generalizability. Future examination of the associations depicted here in a population including individuals with obesity or diabetes might reveal stronger signals. Lastly, as any computational analysis of the associations of gut microbiome features with host phenotypes, further investigation into the causal relationship of the results will be needed.

As evidence on the importance of the gut microbiome in host health, and multiple metabolic outcomes such as obesity, liver disease, and diabetes are accumulating[17,67], the need to elucidate the functional pathways through which the gut microbiome is involved in these conditions is becoming increasingly critical. Understanding these pathways offers the potential to develop targeted microbiome interventions that could ameliorate or prevent these metabolic disorders[77]. Findings revealed in this study, discovered on a large relatively healthy population and showed generalizability under sensitivity analysis and in an independent cohort, could serve as a starting point for future research to further clarify the connections between the gut microbiome and host health, and develop microbiome-targeted treatments to improve host metabolic health.

Methods

All participants signed an informed consent form upon arrival to the research site. All identifying details of the participants were removed prior to the computational analysis. The HPP cohort study is conducted according to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of the Weizmann Institute of Science.

Study population

The cohort analyzed in this study was collected as part of the Human Phenotype Project (HPP). A complete description of the inclusion and exclusion criteria to the study, the measures obtained and the measurement techniques can be found in the original paper describing the HPP[24]. The study includes individuals between the age of 40–70 years old, recruited to the study between January 2019 and May 2023. Data collected at baseline includes self-reported sex, medical history, lifestyle and nutritional habits, vital signs, anthropometrics, metagenomic sequencing of the gut microbiome, blood tests results, electrocardiography, carotid ultrasound, liver US, dual-energy X-ray absorptiometry (DXA) scan and retinal imaging. Continuous measurements include glucose levels using a CGM device for 2 weeks, along with real-time food intake loggins and sleep monitoring by a home sleep apnea test device for 3 nights. In this study we included participants with both CGM and gut-microbiome data, excluding participants who reported taking diabetes-related medications (medications starting with A10 in ATC code), which resulted in 8859 participants, of them 4778 (53.9%) women and 4081 (46.1%) men, with an average age of 51.8 (7.8) years and average BMI of 26.01 (4.1).

Metabolic health measures

Participants recruited to the HPP were connected to a FreeStyle Libre Pro Flash continuous glucose monitoring (FSL-CGM) system for two weeks. Data from the first and last days of the CGM connection were removed prior to the analyses performed here, to maximize data accuracy. 30 CGM-derived measures of glucose control and glucose variability were calculated using the iglu R package[78]. A short description of the calculated measures can be found in Table S7. CGM-derived measures were divided into 7 categories: Euglycemia, Hyperglycemia, Hypoglycemia, Overall variability, Short-term variability, Within-day variability, Between-day variability, and Composite measures & risk scores.

BMI measured at baseline visit along with fat mass, scanned VAT mass and android tissue fat percent obtained using DXA scan were also analyzed in the category of Body composition, and liver measures obtained from ultrasound tests—liver attenuation, viscosity, elasticity, and sound speed, were examined in the category of liver US. Overall, 38 measures of metabolic health were included in this study.

Microbiome sample collection and processing

Microbiome sampling was done using an OMNIgene·GUT (OMR-200, DNA Genotek) stool collection kit, which has the advantage of maintaining DNA integrity in typical ambient temperature fluctuations. Each participant was given a kit and was requested to collect a fecal sample at home. The collected samples were transferred at room temperature to our participant reception center at Weizmann Institute of Science, where they were documented and frozen at −20 °C immediately. Then, samples were transferred in a cooler to our research facilities where they were stored at −20 °C until DNA extraction was performed. Laboratory work was done in the Segal laboratory at the Weizmann Institute of Science. Metagenomic DNA was purified using PowerMag Microbial DNA Isolation Kit (MO BIO Laboratories, 27200-4) optimized for the Tecan automated platform. Libraries for next-generation sequencing were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, E7775) and sequenced on a NovaSeq sequencing platform (Illumina). Sequencing was performed with a 100-bp single-end reads kit and a depth of 10 million reads per sample, using Illumina unique dual sequencing indexes (IDT–Syntezza Bioscience). DNA purification, library preparation, and sequencing were performed in batches of 384 samples. A standard microbial community (ZymoBIOMICS Gut Microbiome Standard, D6331) was inserted into each batch for quality control. No batch corrections were performed. We filtered metagenomic reads containing Illumina adapters and low-quality reads and trimmed low-quality read edges. We detected host DNA by mapping reads to the human genome using Bowtie 2[79] with inclusive parameters and removed those reads.

Metagenomic reads mapping and functional profiling

Processed metagenomic reads were mapped to a genome reference set, representing bacterial species from the human gut microbiome. The reference set used, alongside the methods for its taxonomic annotation are described in detail in Levitan et al.[80]. To determine the relative abundance of species in samples the URA algorithm was used[23], using genomic sequences that are unique to single species in the reference set to determine the existence of each bacterial species in each sample. Relative abundance was clipped at a minimum of 0.0001, meaning this is the smallest possible value of relative abundance for any species.

Metagenomic gene and pathway abundance data were obtained using HUMAnN v3.6.1[25] and MetaPhlAn v4.0.6[26] utilizing the Jan21 MetaPhlAn database, and the UniRef database[47]. Genes and pathways detection thresholds were determined based on the first percentile of all positive, non-zero abundance values across the dataset, derived from HUMAnN3 outputs. *UNMAPPED* reads and *UNINTEGRATED* reads were excluded from threshold determinations for genes and pathways, accordingly. Zero abundance values were adjusted to half the detection threshold value. Abundances were sum-normalized, followed by scaling by a factor of 1,000,000 to achieve parts per million and a log10 transformation.

HUMAnN abundance contributions from specific individual organisms were used to analyze contributing genera and species to the abundance of bacterial gene families and pathways. RPKs were summed over all species belonging to each genus. A genus or species were determined as contributing to the abundance of a bacterial gene family or pathway when it contributed any reads (RPKs > 0) in at least 500 samples.

Upfront gene filtering

Upfront gene filtering was applied to identify uncorrelated bacterial gene families prior to running the associations with metabolic health measures. Bacterial gene families were randomly divided into groups of up to 5000. Within each group, pairwise Spearman correlations were calculated, and hierarchical clustering was applied to the resulting distance matrix using average linkage clustering with a correlation threshold of 0.3. For each cluster, the gene with the highest average abundance was selected as the representative. This process was

repeated three times to further refine the gene representatives and minimize feature redundancy, resulting in 12 groups of ~5000 bacterial gene representatives in each group. To derive a final set of representatives, each pair from the 12 groups was re-processed by calculating Spearman correlations and applying hierarchical clustering with the same correlation threshold of 0.3. The final list of filtered bacterial gene families was defined as the unique set of representatives from all pairs of the 12 groups. The weighted mean correlation and pooled standard deviation were computed based on the correlations within each group pair. The final list of bacterial genes included 55,925 bacterial genes, with a weighted mean correlation of 0.01 and a pooled standard deviation of 0.03.

Bacterial gene families/pathways−metabolic measures associations

Bacterial gene/pathway associations with each metabolic health measure were performed using linear regression with statsmodels OLS[81], adjusting for age and sex. For each model, we only included participants that had no missing values in any of the covariates or the metabolic measure. Metabolic health measures were cleaned for outliers prior to model running. For each metabolic measure we first identified the fraction of the data that includes 95% of the values within the smallest range. Using this data we calculated the mean and std of the metabolic measure distribution, removed measures more than 8 SD away from the mean, and clipped measures more than 5 SD away from the mean. Models were only run for bacterial genes/pathways that were detected in at least 500 samples, based on the bacterial genes/pathways detection thresholds calculated from the data (see "Metagenomic reads mapping and functional profiling"). Bonferroni correction was applied separately for bacterial gene families and pathways, considering all metabolic health measures collectively within each category. Specifically, the correction was performed on the entire set of p-values from the analysis of bacterial pathways across all outcomes, and this process was repeated independently for bacterial gene families. Correlations with Bonferroni-corrected p-values < 0.05 were considered statistically significant.

Gene clumping

Clumping was applied to identify independent genes associated with each metabolic health measure. For each measure, we compiled a list of bacterial gene families significantly associated with it, by a Bonferroni-corrected p-value < 0.05. These bacterial gene families were ranked by their p-values. Beginning with the top-ranked bacterial gene family, we added it to our final list of independent bacterial gene families for the metabolic measure. Subsequently, we evaluated the remaining bacterial gene families for correlation with this top-ranked bacterial gene family, using Spearman's correlation. Bacterial gene families exhibiting a correlation coefficient of 0.3 or higher, with a significance level below 0.05, were clustered with the leading bacterial gene family and excluded from further consideration. This iterative process continued until no bacterial gene families remained in the initial list, ensuring each bacterial gene family in the final list represented an independent association with the metabolic health measure.

Selecting "key" associated bacterial gene families and pathways

To focus on a subgroup of bacterial gene families and pathways, we devised a process to identify "key" bacterial gene families and pathways from those who were found to significantly correlate to metabolic health measures. For each metabolic health measure, we ranked microbial pathways that were significantly correlated to it by their Bonferroni-corrected p-value. Then, we divided our metabolic health measures into three groups, by their categories; Body composition, Liver US and all CGM-derived measures. Each pathway was then represented by its sum of rankings, along with the number of metabolic health measures it was significantly correlated to in each group. Top 5 "key" pathways were chosen for each group of metabolic health measures−those with the highest number of metabolic measures, and the lowest ranking. For the bacterial gene families, a similar process was repeated both for the pre-filtered genes, and for each of the top 5 (lowest p-value) clumped clusters, to ensure the "key" bacterial gene families will be uncorrelated. Clumped bacterial gene families were ranked within each cluster for each metabolic health measure by their Bonferroni-corrected p-value, and one "key" gene family, with the highest number of metabolic health measures in each group and lowest summed ranking, was chosen for each clumped cluster.

Nutritional summaries as sensitivity analysis

To analyze the potential confounding of nutritional habits on the associations between bacterial gene families/pathways and metabolic health measures, we repeated the association analysis, adjusting for summaries of nutritional habits. Using data from continuous real-time food intake logging that was reported by participants while wearing the CGM, we created summaries of nutritional habits. To assure quality control of the data we first removed individual loggings of food items containing more than 5000 calories, and days in which less than 500 calories were reported. Following, we aggregated loggings from the same date to create a daily summary of the amount of calories consumed from carbohydrates, proteins, and lipids, and the total calories (kcal) consumed. For each day we calculated the percent of calories consumed from carbohydrates, proteins, and lipids, and summarized the median of these over all days, along with the median total caloric intake over all days. In addition, the Healthy Food Diversity (HFD) Index, which measures dietary variety and quality[53] was calculated by determining the proportion of each food in the diet, multiplying by health factors, and computing the Berry Index. The HFD Index was normalized by the maximum index value in the study population. With the nutritional summaries calculated for each participant, we repeated the association process described in the "Bacterial gene families/pathways−metabolic measures associations" section, adjusting each OLS model for age and sex as well as all nutritional summaries; HFD Index, median daily caloric intake, and median daily percent of caloric intake from carbohydrates proteins and lipids.

Replication in independent cohort

Replication of the results in an independent cohort was done using metagenomic samples obtained from Gacesa et al.[62]. We only included individuals who had microbiome samples, and complete information on age, sex, and BMI, which resulted in 8205 participants, of them 4754 (57.9%) women and 3451 (42.1%) men, with an average age of 48.4 (14.8) years and average BMI of 25.6 (4.4). To process gut microbiome samples in the same manner as the discovery cohort we used only one of the paired-end reads, and truncated reads at 75 bp. Metagenomic gene and pathway abundance data were then obtained in the same process as the discovery cohort (see Metagenomic read mapping and functional profiling). As only BMI information was available for the replication cohort, gene/pathway associations were tested for this measure only, in the same process as in the discovery cohort (see Bacterial gene families/pathways−metabolic measures associations). Bonferroni correction was performed separately for bacterial gene families and pathways, across all models, and correlations with Bonferroni-corrected p-value < 0.05 were considered statistically significant. Comparisons to the discovery cohort were made based on the Bonferroni-corrected p-values of all models of the BMI measure alone, to align with the statistical analysis of a single measure in the replication cohort.

Reporting summary

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

Data availability

Data in this paper are part of the Human Phenotype Project (HPP). The raw metagenomic data and basic phenotypes (age, sex, and BMI) used in this study are available at the European Genome-phenome Archive (https://ega-archive.org/) under accession EGAS00001007204. The other data are accessible to researchers from universities and other research institutions at https://humanphenotypeproject.org/home. Interested bona fide researchers should contact info@pheno.ai to obtain instructions for accessing the data.

Code availability

Analyses in this study were performed using publicly available Python libraries, as detailed in the "Methods" section.

References

1. Manichanh, C. et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205–211 (2006).
2. Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A. & Wargo, J. A. The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. *Cancer Cell* **33**, 570–580 (2018).
3. Tang, W. H. W. et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **368**, 1575–1584 (2013).
4. Li, Q. et al. Implication of the gut microbiome composition of type 2 diabetic patients from northern China. *Sci. Rep.* **10**, 5450 (2020).
5. Pedersen, H. K. et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
6. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
7. Zahavi, L. et al. Bacterial SNPs in the human gut microbiome associate with host BMI. *Nat. Med.* **29**, 2785–2792 (2023).
8. Rodbard, D. Continuous glucose monitoring: a review of successes, challenges, and opportunities. *Diabetes Technol. Ther.* **18**, S3–S13 (2016).
9. Kaul, S. et al. Dual-energy X-ray absorptiometry for quantification of visceral fat. *Obesity* **20**, 1313–1318 (2012).
10. Yoneda, M., Honda, Y., Nogami, A., Imajo, K. & Nakajima, A. Advances in ultrasound elastography for nonalcoholic fatty liver disease. *J. Med. Ultrason.* **47**, 521–533 (2020).
11. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
12. Arnoriaga-Rodríguez, M. et al. Gut microbiota composition and functionality are associated with REM sleep duration and continuous glucose levels. *J. Clin. Endocrinol. Metab.* **108**, 2931–2939 (2023).
13. Krisko, T. I. et al. Dissociation of adaptive thermogenesis from glucose homeostasis in microbiome-deficient mice. *Cell Metab.* **31**, 592–604.e9 (2020).
14. Wang, R. et al. Gut microbiome, liver immunology, and liver diseases. *Cell. Mol. Immunol.* **18**, 4–17 (2021).
15. Komodromou, I. et al. Exploring the dynamic relationship between the gut microbiome and body composition across the human lifespan: a systematic review. *Nutrients* **16**, 660 (2024).
16. U-Din, M. et al. Characteristics of abdominal visceral adipose tissue, metabolic health and the gut microbiome in adults. *J. Clin. Endocrinol. Metab.* **109**, 680–690 (2024).
17. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).
18. Portincasa, P. et al. Gut microbiota and short chain fatty acids: implications in glucose homeostasis. *Int. J. Mol. Sci.* **23**, 1105 (2022).
19. Milosevic, I. et al. Gut-liver axis, gut microbiota, and its modulation in the management of liver diseases: a review of the literature. *Int. J. Mol. Sci.* **20**, 395 (2019).
20. Takeuchi, T. et al. Gut microbial carbohydrate metabolism contributes to insulin resistance. *Nature* **621**, 389–395 (2023).
21. 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).
22. Mei, Z. et al. Strain-specific gut microbial signatures in type 2 diabetes identified in a cross-cohort analysis of 8,117 metagenomes. *Nat. Med.* **30**, 2265–2276 (2024).
23. Rothschild, D. et al. An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *PLoS ONE* **17**, e0265756 (2022).
24. Shilo, S. et al. 10 K: a large-scale prospective longitudinal study in Israel. *Eur. J. Epidemiol.* **36**, 1187–1194 (2021).
25. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
26. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
27. Joy, D., Thava, V. R. & Scott, B. B. Diagnosis of fatty liver disease: is biopsy necessary? *Eur. J. Gastroenterol. Hepatol.* **15**, 539–543 (2003).
28. Rodbard, D. New and improved methods to characterize glycemic variability using continuous glucose monitoring. *Diabetes Technol. Ther.* **11**, 551–565 (2009).
29. Clarke, W. & Kovatchev, B. Statistical tools to analyze continuous glucose monitor data. *Diabetes Technol. Ther.* **11**, S45–S54 (2009).
30. Hermanides, J. et al. Glucose variability is associated with intensive care unit mortality. *Crit. Care Med.* **38**, 838–842 (2010).
31. Kovatchev, B. P., Straume, M., Cox, D. J. & Farhy, L. S. Risk analysis of blood glucose data: a quantitative approach to optimizing the control of insulin dependent diabetes. *J. Theor. Med.* **3**, 1–10 (2000).
32. Rodbard, D. Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control. *Diabetes Technol. Ther.* **11**, S55–S67 (2009).
33. Nathan, D. M. et al. Translating the A1C assay into estimated average glucose values. *Diabetes Care* **31**, 1473–1478 (2008).
34. Bergenstal, R. M. et al. Glucose management indicator (GMI): a new term for estimating A1C from continuous glucose monitoring. *Diabetes Care* **41**, 2275–2280 (2018).
35. Kovatchev, B. P., Otto, E., Cox, D., Gonder-Frederick, L. & Clarke, W. Evaluation of a new measure of blood glucose variability in diabetes. *Diabetes Care* **29**, 2433–2438 (2006).
36. Luo, H.-H., Feng, X.-F., Yang, X.-L., Hou, R.-Q. & Fang, Z.-Z. Interactive effects of asparagine and aspartate homeostasis with sex and age for the risk of type 2 diabetes risk. *Biol. Sex. Differ.* **11**, 58 (2020).
37. Vangipurapu, J., Stancáková, A., Smith, U., Kuusisto, J. & Laakso, M. Nine amino acids are associated with decreased insulin secretion and elevated glucose levels in a 7.4-year follow-up study of 5,181 Finnish men. *Diabetes* **68**, 1353–1358 (2019).
38. Tang, Y. et al. Association of serum methionine metabolites with non-alcoholic fatty liver disease: a cross-sectional study. *Nutr. Metab.* **19**, 21 (2022).
39. Srivastava, S. Emerging therapeutic roles for NAD(+) metabolism in mitochondrial and age-related disorders. *Clin. Transl. Med.* **5**, 25 (2016).
40. Hong, W., Mo, F., Zhang, Z., Huang, M. & Wei, X. Nicotinamide mononucleotide: a promising molecule for therapy of diverse diseases by targeting NAD+ metabolism. *Front. Cell Dev. Biol.* **8**, 246 (2020).
41. Yoshino, J., Mills, K. F., Yoon, M. J. & Imai, S. Nicotinamide mononucleotide, a key NAD(+) intermediate, treats the pathophysiology of diet- and age-induced diabetes in mice. *Cell Metab.* **14**, 528–536 (2011).

42. Kim, S.-K. Interrelationship of uric acid, gout, and metabolic syndrome: focus on hypertension, cardiovascular disease, and insulin resistance. *J. Rheum. Dis.* **25**, 19 (2018).
43. Kim, G. et al. Bifidobacterial carbohydrate/nucleoside metabolism enhances oxidative phosphorylation in white adipose tissue to protect against diet-induced obesity. *Microbiome* **10**, 188 (2022).
44. Peleli, M. & Carlstrom, M. Adenosine signaling in diabetes mellitus and associated cardiovascular and renal complications. *Mol. Asp. Med.* **55**, 62–74 (2017).
45. Radziejewska, A., Muzsik, A., Milagro, F. I., Martínez, J. A. & Chmurzynska, A. One-carbon metabolism and nonalcoholic fatty liver disease: the crosstalk between nutrients, microbiota, and genetics. *Lifestyle Genom.* **13**, 53–63 (2020).
46. Köse, S., Sözlü, S., Bölükbaşi, H., Ünsal, N. & Gezmen-Karadağ, M. Obesity is associated with folate metabolism. *Int. J. Vitam. Nutr. Res.* **90**, 353–364 (2020).
47. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
48. di Salvo, M. L., Safo, M. K., Musayev, F. N., Bossa, F. & Schirch, V. Structure and mechanism of *Escherichia coli* pyridoxine 5′-phosphate oxidase. *Biochim. Biophys. Acta (BBA) - Proteins Proteom.* **1647**, 76–82 (2003).
49. Sookoian, S. & Pirola, C. J. Alanine and aspartate aminotransferase and glutamine-cycling pathway: their roles in pathogenesis of metabolic syndrome. *World J. Gastroenterol.* **18**, 3775–3781 (2012).
50. Leylabadlo, H. E. et al. The critical role of *Faecalibacterium prausnitzii* in human health: an overview. *Microb. Pathog.* **149**, 104344 (2020).
51. Rein, M. et al. Effects of personalized diets by prediction of glycemic responses on glycemic control and metabolic health in newly diagnosed T2DM: a randomized dietary intervention pilot trial. *BMC Med.* **20**, 56 (2022).
52. Fu, J., Zheng, Y., Gao, Y. & Xu, W. Dietary fiber intake and gut microbiota in human health. *Microorganisms* **10**, 2507 (2022).
53. Vadiveloo, M., Dixon, L. B., Mijanovich, T., Elbel, B. & Parekh, N. Development and evaluation of the US Healthy Food Diversity index. *Br. J. Nutr.* **112**, 1562–1574 (2014).
54. Keshet, A. et al. CGMap: characterizing continuous glucose monitor data in thousands of non-diabetic individuals. *Cell Metab.* **35**, 758–769.e3 (2023).
55. Blaychfeld-Magnazi, M., Reshef, N., Zornitzki, T., Madar, Z. & Knobler, H. The effect of a low-carbohydrate high-fat diet and ethnicity on daily glucose profile in type 2 diabetes determined by continuous glucose monitoring. *Eur. J. Nutr.* **59**, 1929–1936 (2020).
56. Lin, Y.-H. et al. Impact of carbohydrate on glucose variability in patients with type 1 diabetes assessed through professional continuous glucose monitoring: a retrospective study. *Diabetes Ther.* **10**, 2289–2304 (2019).
57. Service, F. J. et al. Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes* **19**, 644–655 (1970).
58. Sheng, S. et al. Structural and functional alterations of gut microbiota in males with hyperuricemia and high levels of liver enzymes. *Front Med.* **8**, 779994 (2021).
59. Vuorinen-Markkola, H. & Yki-Järvinen, H. Hyperuricemia and insulin resistance. *J. Clin. Endocrinol. Metab.* **78**, 25–29 (1994).
60. Li, C., Hsieh, M.-C. & Chang, S.-J. Metabolic syndrome, diabetes, and hyperuricemia. *Curr. Opin. Rheumatol.* **25**, 210–216 (2013).
61. Hu, S. et al. L-Arginine modulates glucose and lipid metabolism in obesity and diabetes. *Curr. Protein Pept. Sci.* **18**, 599–608 (2017).
62. Gacesa, R. et al. Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).
63. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
64. Wang, Z. et al. Effects of probiotic/prebiotic/synbiotic supplementation on blood glucose profiles: a systematic review and meta-analysis of randomized controlled trials. *Public Health* **210**, 149–159 (2022).
65. Chen, L. et al. High-fiber diet ameliorates gut microbiota, serum metabolism and emotional mood in type 2 diabetes patients. *Front. Cell. Infect. Microbiol.* **13**, 1069954 (2023).
66. Brandt, L. J. & Aroniadis, O. C. An overview of fecal microbiota transplantation: techniques, indications, and outcomes. *Gastrointest. Endosc.* **78**, 240–249 (2013).
67. Howard, E. J., Lam, T. K. T. & Duca, F. A. The gut microbiome: connecting diet, glucose homeostasis, and disease. *Annu. Rev. Med.* **73**, 469–481 (2022).
68. Li, N., Cen, Z., Zhao, Z., Li, Z. & Chen, S. BCAA dysmetabolism in the host and gut microbiome, a key player in the development of obesity and T2DM. *Med. Microecol.* **16**, 100078 (2023).
69. Wang, C. et al. Role of bile acids in dysbiosis and treatment of nonalcoholic fatty liver disease. *Mediators Inflamm.* **2019**, 7659509 (2019).
70. Pinart, M. et al. Gut microbiome composition in obese and non-obese persons: a systematic review and meta-analysis. *Nutrients* **14**, 12 (2021).
71. Lee, C. J., Sears, C. L. & Maruthur, N. Gut microbiome and its role in obesity and insulin resistance. *Ann. N. Y. Acad. Sci.* **1461**, 37–52 (2020).
72. Ozato, N. et al. Two *Blautia* species associated with visceral fat accumulation: a one-year longitudinal study. *Biology* **11**, 318 (2022).
73. Lippert, K. et al. Gut microbiota dysbiosis associated with glucose metabolism disorders and the metabolic syndrome in older adults. *Benef. Microbes* **8**, 545–556 (2017).
74. Nie, C., He, T., Zhang, W., Zhang, G. & Ma, X. Branched chain amino acids: beyond nutrition metabolism. *Int. J. Mol. Sci.* **19**, 954 (2018).
75. Fahy, S., O'Connor, J. A., Sleator, R. D. & Lucey, B. From species to genes: a new diagnostic paradigm. *Antibiotics* **13**, 661 (2024).
76. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front. Genet.* **8**, 23 (2017).
77. Marchesi, J. R. et al. The gut microbiota and host health: a new clinical frontier. *Gut* **65**, 330–339 (2016).
78. Broll, S. et al. Interpreting blood GLUcose data with R package iglu. *PLoS ONE* **16**, e0248560 (2021).
79. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
80. Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M. & Segal, E. An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nat. Commun.* **13**, 3863 (2022).
81. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* 92–96 (SciPy, 2010).

Acknowledgements

We thank Tal Bauman, Liron Zahavi, Saar Shoer, Tal Shor, Adina Weinberger and members of the Segal group for fruitful discussions. E.S. is supported by the Crown Human Genome Center; the Larson Charitable Foundation New Scientist Fund; the Else Kröner Fresenius Foundation; the White Rose International Foundation; the Ben B. and Joyce E. Eisenberg Foundation; the Nissenbaum Family; Marcos Pinheiro de Andrade and Vanessa Buchheim; Lady Michelle Michels; Aliza Moussaieff; and grants funded by the Minerva Foundation, with funding from the Federal German Ministry for Education and Research and by the European Research Council and the Israel Science Foundation. A.K. is partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center.

Author contributions

A.K. conceived the project, designed and conducted the analyses, interpreted the results and wrote the manuscript. E.S. directed and supervised the project.

Competing interests

A.K. and E.S. are paid consultants to Pheno.AI, Ltd.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53832-y.

Correspondence and requests for materials should be addressed to Eran Segal.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.