gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota

Pascal Andreu, Victòria; Augustijn, Hannah E; Chen, Lianmin; Zhernakova, Alexandra; Fu, Jingyuan; Fischbach, Michael A; Dodd, Dylan; Medema, Marnix H

Published in:
Nature Biotechnology

DOI:
10.1038/s41587-023-01675-1

Publication date:
2023

nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01675-1

gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota

Received: 2 March 2021

Accepted: 10 January 2023

Published online: 13 February 2023

Check for updates

Victòria Pascal Andreu[1], Hannah E. Augustijn[1,2,10], Lianmin Chen[2,3,4,5,10], Alexandra Zhernakova[2], Jingyuan Fu[2,3], Michael A. Fischbach[6,7,8] ✉, Dylan Dodd[7,9] ✉ & Marnix H. Medema[1] ✉

The gut microbiota produce hundreds of small molecules, many of which modulate host physiology. Although efforts have been made to identify biosynthetic genes for secondary metabolites, the chemical output of the gut microbiome consists predominantly of primary metabolites. Here we introduce the gutSMASH algorithm for identification of primary metabolic gene clusters, and we used it to systematically profile gut microbiome metabolism, identifying 19,890 gene clusters in 4,240 high-quality microbial genomes. We found marked differences in pathway distribution among phyla, reflecting distinct strategies for energy capture. These data explain taxonomic differences in short-chain fatty acid production and suggest a characteristic metabolic niche for each taxon. Analysis of 1,135 individuals from a Dutch population-based cohort shows that the level of microbiome-derived metabolites in plasma and feces is almost completely uncorrelated with the metagenomic abundance of corresponding metabolic genes, indicating a crucial role for pathway-specific gene regulation and metabolite flux. This work is a starting point for understanding differences in how bacterial taxa contribute to the chemistry of the microbiome.

The pathways encoding the production of microbial metabolites are often physically clustered in the genome, in regions known as metabolic gene clusters (MGCs). Current tools for computational prediction of metabolic pathways focus on gene clusters for natural product biosynthesis[1] or generic primary metabolism[2,3]. Here, we introduce an algorithm, called gutSMASH, to profile known and predicted novel specialized primary MGCs from the gut microbiome, which we define as gene clusters encoding primary metabolic pathways that are taxon-specific, niche-defining and important for (host–)microbiome interactions. We used this tool to perform a systematic analysis of primary MGCs in bacterial strains from the gut microbiome, and we identified the prevalence and abundance of each of these pathways across a large population-based cohort as well as a clinical cohort. Although gutSMASH has been built to specifically predict MGCs from anaerobic human gut bacteria, this tool can also be applied to microbial communities that inhabit other (animal) body sites.

[1]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. [2]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [3]Department of Pediatrics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [4]Changzhou Medical Center, Nanjing Medical University, Changzhou, China. [5]Department of Cardiology, Nanjing Medical University, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China. [6]Department of Bioengineering, Stanford University, Stanford, CA, USA. [7]Department of Microbiology and Immunology, Stanford University, Stanford, CA, USA. [8]Chan Zuckerberg Biohub, San Francisco, CA, USA. [9]Department of Pathology, Stanford University, Stanford, CA, USA. [10]These authors contributed equally: Hannah E. Augustijn, Lianmin Chen.
✉e-mail: fischbach@fischbachgroup.org; ddodd2@stanford.edu; marnix.medema@wur.nl

https://doi.org/10.1038/s41587-023-01675-1

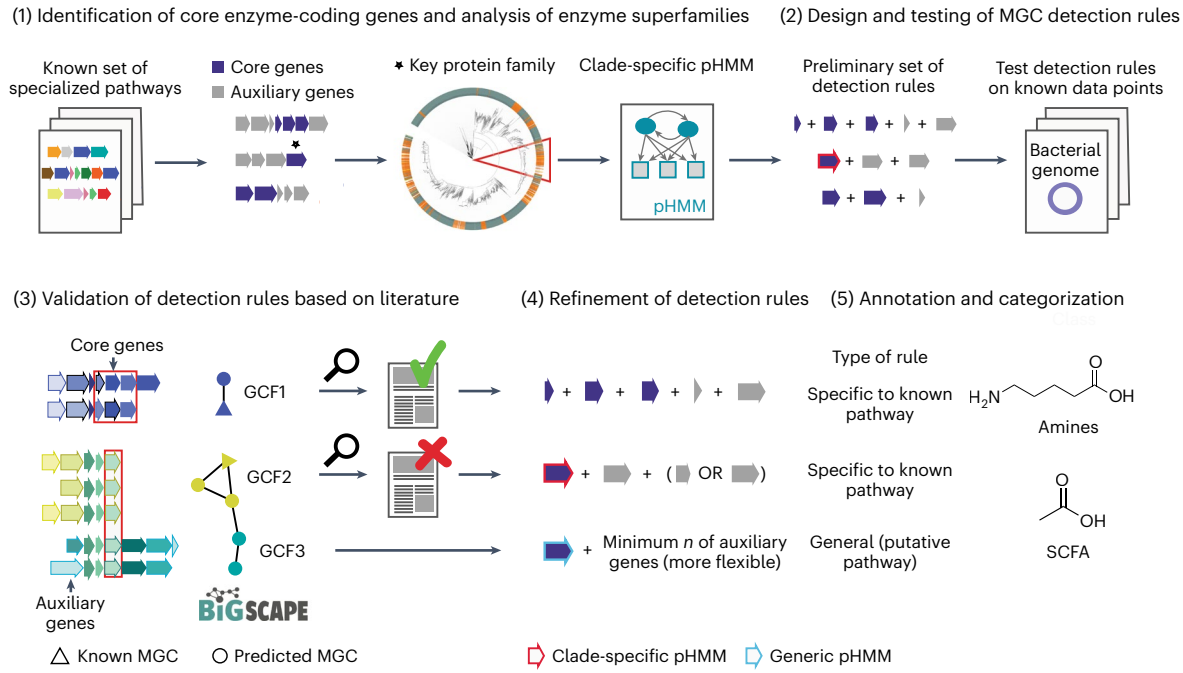


Fig. 1 | Development and design of detection rules for gutSMASH. (1) A set of known and characterized MGC-encoded pathways were curated from the literature. Protein domains were identified across all MGCs, and core enzymatic domains were manually identified. For enzymatic domains belonging to broad multifunctional enzyme families, protein superfamily phylogenies were built to create clade-specific pHMMs. (2) These domains were incorporated in the initial detection rules. The detection rules were run on a test set, and all the MGCs predicted by the same rule were grouped together and (3) run through BiG-SCAPE, which grouped the MGCs into gene cluster families (GCFs). (4) Based on a literature analysis of GCF members, detection rules were manually fine-tuned to either include or exclude MGC architectures that were either related to specialized primary metabolism or not. (5) Finally, fine-tuned detection rules were annotated and categorized into different MGC classes based on their metabolic end products.

Algorithms that identify physically clustered genes have become a mainstay of bacterial pathway identification[4–6]. Taking into account the conserved physical clustering of genes prevents false-positive hits based on sequence similarity alone. This principle has been widely applied in the field of natural product biosynthesis—for example, in antiSMASH[1], which predicts biosynthetic gene clusters (BGCs) by detecting physically clustered protein domains using profile hidden Markov models (pHMMs). In the present study, we tailored this gene cluster detection framework to detect MGCs involved in primary metabolism and bioenergetics.

Results

As a starting point, we constructed a dataset of 51 primary metabolic pathways from the gut microbiome with biochemical or genetic literature support (including MGCs as well as pathways encoded by a single genes) and identified core enzymes (that is, required for pathway function) to serve as a signature for the detection rules (Fig. 1, Supplementary Table 1 and Methods). To more accurately predict MGCs of interest, we performed three computational procedures. First, for core enzymes belonging to 12 of the protein superfamilies that are known to catalyze diverse types of reactions and were most commonly found across a wide range of pathways, we constructed phylogenies and used them to create clade-specific pHMMs to detect specific subfamilies (Supplementary Information results: 'Phylogenetic analysis of protein superfamilies to identify pathway-specific clades'). Second, we designed pathway-specific rules for each MGC type in our dataset (Methods). These rules were validated and optimized by detailed manual visual inspection and analysis of MGC sequence similarity networks made using BiG-SCAPE[7], generated from gutSMASH results on a set of 1,621 microbial genomes (online data: https://gutsmash.bioinformatics.nl/help.html#Validation; Supplementary Information results: 'Validation of gutSMASH detection rules by evaluating their predictive performance' and Supplementary Tables 2 and 3). Third, despite the fact that most specialized primary metabolic pathways are encoded in MGCs, there are also single-protein pathways that are in charge of the secretion of key specialized primary metabolites in the gut microbial ecosystem, such as serine dehydratase, which produces ammonia and pyruvate from serine[8]. For this reason, we also built ten clade-specific pHMMs to detect these (Methods: 'Assessing single-protein pathway abundance within representative human gut bacteria'). The above procedures led to the design of a set of detection rules included in the gutSMASH framework to identify both known and putative MGCs that are potentially relevant for metabolite-mediated microbiome-associated phenotypes and also assess the presence/absence patterns of single-protein pathways across microbial genera by using custom pHMMs (not included in the gutSMASH detection rule set). Although obtaining a precise estimate of precision and recall of the gutSMASH algorithm is infeasible due to the absence of large-scale experimentally verified MGCs from diverse taxa, additional manual validation on a dataset of 18 experimentally verified homologs of gutSMASH-detected MGCs, as well as on a dataset of 42 MGCs from five model organisms from different phyla, showed no false negatives or false positives (Supplementary Information results, 'Validation of gutSMASH detection rules by evaluating their predictive performance').

To profile the metabolic capacity of strains from the human gut microbiome, we selected a set of 4,240 unique high-quality reference genomes consisting of 1,520 genomes from the Culturable Genome Reference (CGR) collection[9], 2,308 genomes from the Microbial Reference Genomes collection of the Human Microbiome Project (HMP) consortium[10] and 414 additional genomes from the class Clostridia to account for their metabolic versatility[11] (Supplementary Table 4). We refrained from including metagenome-assembled genomes in this analysis as they often lack the taxon-specific genomic islands[12] on which many specialistic metabolic functions are encoded. In total, gutSMASH

predicted 19,890 MGCs across these genomes that are clear homologs of MGCs for our set of known pathway types (Methods: 'Evaluating the functional potential of the human microbiome using gutSMASH').

The combined results of the gutSMASH MGC scanning and the single-protein pHMM detection across the three reference collections provide unique insights into the metabolic traits encoded by the genomes of human gut bacteria. Although some genera harbor a small set of highly conserved pathways, (for example, *Akkermansia* and *Faecalibacterium*), other genera contain much larger interspecies differences (Fig. 2a). The genus *Clostridium* displays remarkable metabolic versatility, with 43 distinct MGC-encoded metabolic pathways present across members of this genus (Fig. 2a); this corroborates earlier results by Viera-Silva et al.[13], who showed high dissimilarity of metabolic module repertoires in Clostridia. Clostridial strains that are indistinguishable by 16S sequencing often harbor distinct gene cluster ensembles (Supplementary Fig. 1), suggesting that specialization in primary metabolism leads to functional differentiation even among closely related strains. *Clostridium* is a clear outlier: by comparison, the next most numerous sets of metabolic pathways are found within the Enterobacteriaceae (for example, *Salmonella*, *Escherichia*, *Enterobacter and Klebsiella*) with 22–25 metabolic pathways. Intriguingly, many of the metabolic pathways encoded by *Clostridium* and members of the Enterobacteriaceae are non-overlapping (with 23/43 *Clostridium* pathways not being identified among Enterobacteriaceae), highlighting the distinct metabolic strategies that these microbes employ within the gut (Fig. 2a). The *Bacteroides*, Actinobacteria (*Eggerthella* and *Collinsella*) and Verrucomicrobia (*Akkermansia*) harbor a more restricted set of primary metabolic pathways, likely reflecting versatility in upstream components of their metabolism (that is, glycan foraging and other forms of substrate utilization).

Our results provide insights into the metabolic strategies that microbes use to produce short-chain fatty acids (SCFAs). As expected, butyrate production is found mainly in certain Firmicutes and Fusobacteria; however, some *Alistipes* sp. within the Bacteroidetes phylum have genes for the acetate-to-butyrate pathway (Fig. 2a). This is consistent with previous reports that *Alistipes* sp. produce small amounts of this compound[14]. On the other hand, propionate production is largely confined to (and conserved in) the Bacteroidetes. However, the phylogenetic distribution of pathways that generate acetate—the most concentrated molecule produced in the gut[15]—has not yet been described. Two pathways for the conversion of pyruvate to acetate—pyruvate formate-lyase (PFL) (pyruvate to acetate/formate) and pyruvate:ferredoxin oxidoreductase (PFOR)—are widely distributed across microbial strains from diverse phyla (Fig. 2b). Two observations suggest that these two pathways are the most prolific source of acetate in the gut. First, some strains known to produce large quantities of acetate rely entirely on one or both of the pathways. Second, each one uses pyruvate as a substrate, consistent with a model in which these pathways are the primary conduit through which carbohydrate-derived carbon is converted to acetate. Additional taxon-specific pathways for acetate include the CO_2-to-acetate pathway and the glycine-to-acetate pathway (each specific to a subset of Firmicutes), as well as the choline and ethanolamine utilization pathways (widespread among Enterobacteriaceae and each found in different clades of Firmicutes) (Fig. 2a).

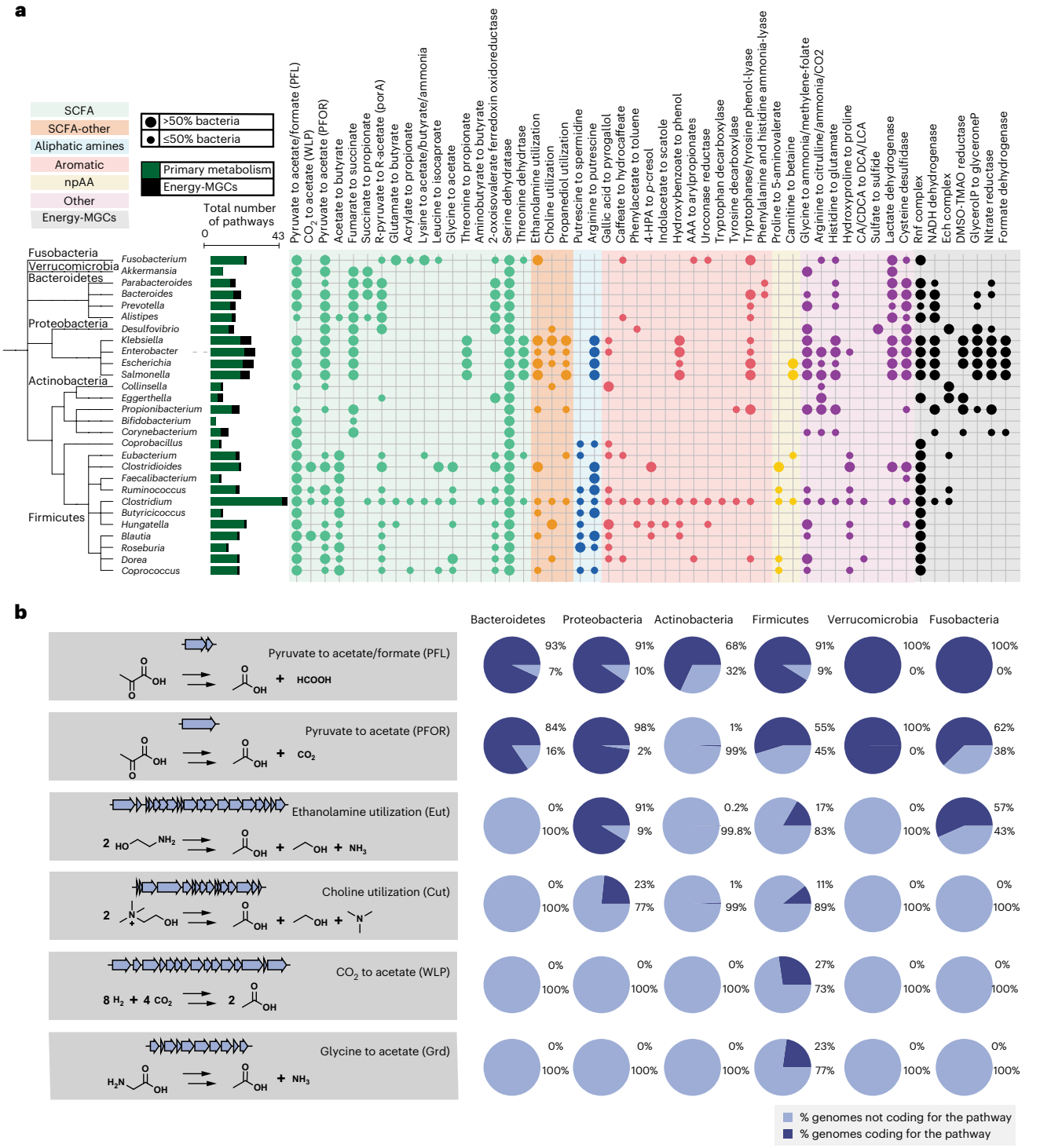
Our results demonstrate a striking difference in mechanisms for energy capture by three of the major bacterial genera in the gut: *Bacteroides*, *Escherichia* and *Clostridium*. When growing aerobically with glucose, *E. coli* generates most of its energy by channeling electrons through membrane-bound cytochromes using oxygen as the terminal electron acceptor (Fig. 2c). However, oxygen is limiting in the gut. Under anaerobic conditions, bacteria from the genus *Escherichia* employ alternate terminal electron acceptors, such as nitrate, dimethyl sulfoxide (DMSO), trimethylamine N-oxide (TMAO) and fumarate, by substituting alternate terminal reductases into their electron transport system (Fig. 2c). However, in the healthy gut, these alternate electron acceptors are either absent or available in limited amounts, likely explaining why these facultative anaerobes represent a small proportion of the healthy microbiome[16]. In contrast to the diversity of terminal reductases used by the *Escherichia*, *Bacteroides* genomes encode only fumarate reductase (Fig. 2c). They use a unique pathway, carboxylating phosphoenolpyruvate (PEP) to form fumarate, which they use as a terminal electron acceptor to run an anaerobic electron transport chain involving NADH dehydrogenase and fumarate reductase, ultimately forming propionate. Thus, the metabolic strategy employed by *Bacteroides* ensures a steady stream of electron acceptors to fuel their metabolism. Clostridia do not utilize similar mechanisms for energy capture as members of the genera *Escherichia* and *Bacteroides*. Recent analyses suggest that they use the Rhodobacter nitrogen fixation-like (Rnf) complex for generating a proton motive force[17,18]. Several pathways encoded by the genomes of *Clostridium* (for example, acetate to butyrate, aromatic amino acids (AAAs) to arylpropionates and leucine to isocaproate) (Fig. 2a) consist of an electron bifurcating acyl-CoA dehydrogenase enzyme. This complex bifurcates electrons from NADH to the low potential electron carrier ferredoxin, which can then donate electrons to the Rnf complex, which functions as a proton or sodium pump, generating an ion motive force. Although much still is to be learned about Clostridial metabolism, our findings suggest that their metabolism operates at a different scale of the redox tower compared to *Bacteroides* and Enterobacteriaceae, using low potential electron carriers to fuel their metabolism.

Next, we set out to determine the prevalence and abundance of each pathway in a cohort of human samples. We used BiG-MAP[19] to profile the relative abundance of each MGC class across 1,135 metagenomes from the population-based LifeLines DEEP cohort[20], by mapping metagenomic reads against a collection of 5,655 non-redundant MGCs detected in our set of reference genomes (Fig. 3a,b, Extended Data Fig. 1 and Supplementary Fig. 2). Some pathways, such as CO_2 to acetate (acetogenesis) and butyrate production from acetate or glutamate, as well as polyamine-forming pathways, were found in >99% of microbiomes. Others, such as 1,2-propanediol utilization and *p*-cresol production, both associated with negative effects on gut health[21,22], were observed at detectable levels in only 75% and 53% of the samples, respectively. In terms of abundance, it is striking that, for example, the bile-acid-induced (*bai*) operon for the formation of the secondary bile acids deoxycholic acid and lithocholic acid, which

Fig. 2 | Distribution of known pathways across most representative genera in the human gut. a, Circles represent the absence/presence of known pathways in each genus. Larger circles indicate cases in which more than 50% of the genomes for a genus encode the pathway, whereas smaller circles indicate cases in which 50% or fewer of the genomes encode it. Colored ranges indicate a categorization of MGCs by chemical class of their product, in which npAA represents non-proteinogenic amino acids and SCFA represents short-chain fatty acids. Taxonomic assignments were applied using the Genome Taxonomy Database release 95[32]. The tree was generated using phyloT (https://phylot.biobyte.de/) and visualized using iTOL[33]. Raw data are available in Supplementary Table 5. **b**, Distribution of the main acetate synthesis pathways at phylum level. Some of the pathways are ubiquitous across the five major phyla (for example, pyruvate to acetate/formate (PFL)), whereas others are found only in Firmicutes (CO_2 to acetate (WLP)). Raw data for the pie charts are available in Supplementary Table 6. Genes and gene clusters depicted are representatives from *Bacteroides thetaiotaomicron* (PFL and PFOR), *Salmonella enterica* (Eut), *Clostridium sporogenes* (Cut), *Clostridium difficile* (WLP) and *Clostridium sticklandii* (Grd). **c**, Bioenergetic strategies in *Escherichia* that has a variety of alternate electron acceptors to choose from compared to *Bacteroides* and *Clostridium*. CA, cholic acid; CDCA, chenodeoxycholic acid; Cut, choline use; DCA, deoxycholic acid; Eut, ethanolamine use; Grd, glycine reductase; Hyd, hydrogenase; LCA, lithocholic acid; Ndh, NADH dehydrogenase.

https://doi.org/10.1038/s41587-023-01675-1

has been characterized from very low-abundance *Clostridium scindens* strains[23], was still shown to be present in relatively high abundance across a subset of individuals. Analysis of the mapped reads showed that the vast majority of these mapped to a homologous MGC from the genus *Dorea* instead (Supplementary Fig. 2), for which the physiological relevance remains to be established. Although two of the three



https://doi.org/10.1038/s41587-023-01675-1

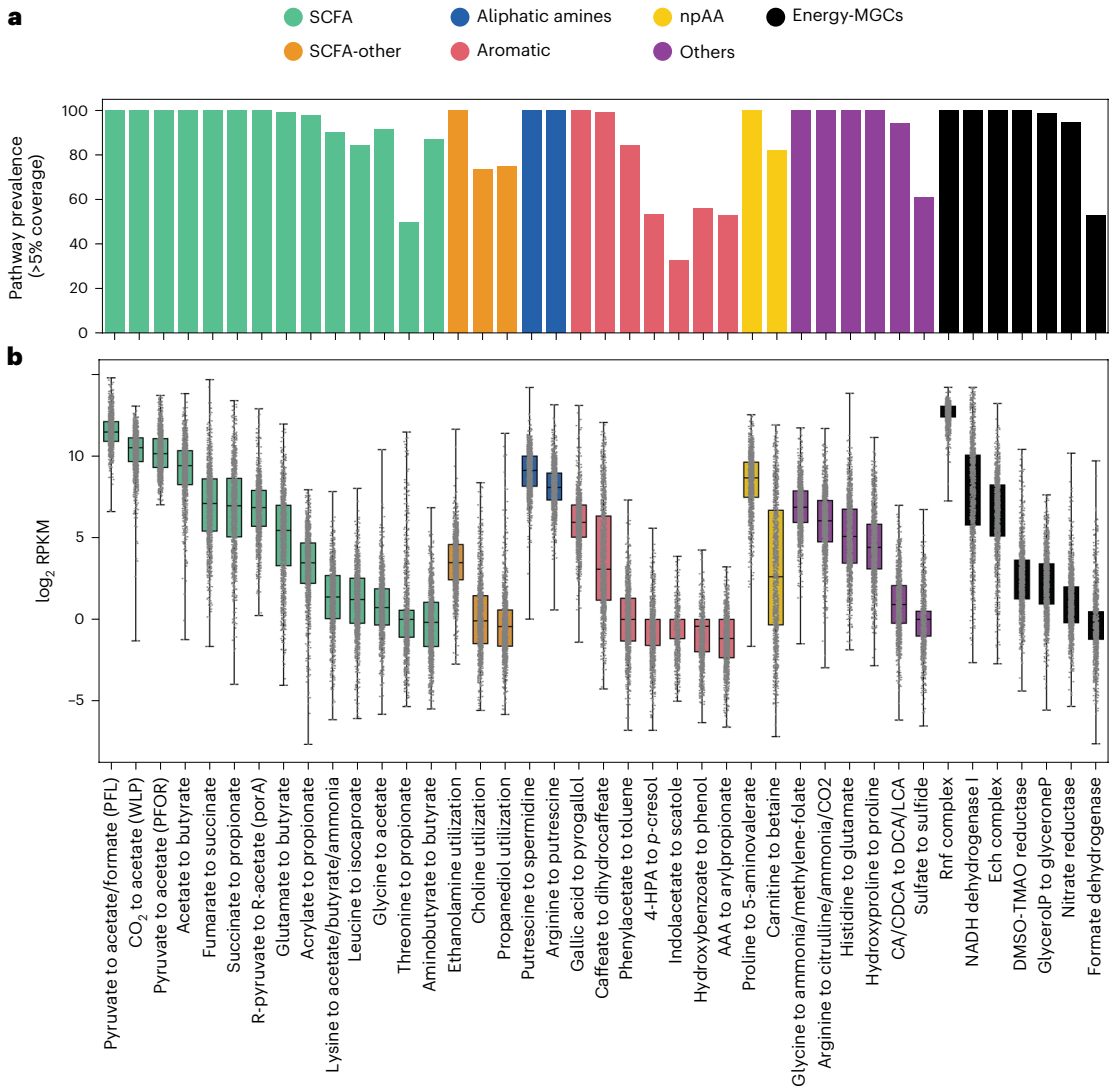


Fig. 3 | Prevalence and abundance of specialized primary metabolic pathways across 1,135 human microbiome samples. a, Prevalence of each of the 41 known MGC-encoded pathway classes across all microbiomes, measured as the percentage of samples in which core enzyme-coding genes of at least one reference MGC belonging to a given class were covered by metagenomic reads across >5% of their sequence length. This cutoff was kept low to avoid false negatives due to limited sequencing depth for low-abundance taxa (raw data available in Supplementary Table 7). **b**, Distributions of $\log_2$ RPKM relative abundance values of all 41 known pathway classes, categorized by product class, across all LifeLines DEEP metagenomes (n = 1,135; raw count data available in Supplementary Table 8). All samples are represented by a dot in the box plot, representing the $\log_2$ RPKM value for a given sample. The box limits indicate the quartiles of the dataset; the whiskers extend to 1.5× the interquartile range; and the center line denotes the median.

acetate-forming pathways (PFL and PFOR) were consistently found at high abundance levels, the abundance of all butyrate-forming pathways is highly variable across individuals, with a ~13-fold difference between lower and upper quartiles in the abundance distribution of the glutamate-to-butyrate pathway and a >130-fold difference between the 10th percentile and the 90th percentile.

The wide variability in the metagenome abundance of each pathway raises the question of whether metagenomic abundance of a pathway correlates with the level of its small-molecule product in the host. To address this question, we systematically compared the level of each pathway with the quantity of the corresponding metabolite as determined by plasma metabolomics. We found a striking lack of correlation between pathway and metabolite levels (r ranging from −0.04 to 0.24; Fig. 4a and Extended Data Fig. 2); also when abundances of multiple MGC types with the same end products were summed, correlations remained low (Supplementary Table 9). These data indicate that gene abundances in metagenomes are not (on their own) a useful predictor of plasma metabolic outputs. This lack of correlation may be due to several factors, such as dietary differences, transcriptional regulation linked to substrate availability, varying dynamics of diffusion and import of the metabolites out of the lumen into the host, secondary fermenters that degrade the end product of some of the gutSMASH-predicted pathways and the existence of pathways with similar substrate/product profiles that are yet unknown. To assess the effect of nutrient import, we also quantified metagenomic pathway abundance correlation with available fecal metabolomic data for SCFA from the LifeLines DEEP cohort, and this showed similarly low correlations (ranging from −0.16 to 0.11; Supplementary Table 10). We also mapped reads from 81 samples from the integrative Human Microbiome Project (iHMP), including 41 patients with Crohn's disease (CD), 17 patients with ulcerative colitis (UC) and 23 healthy individuals (Supplementary Table 11), to the same set of gutSMASH MGCs (Fig. 4b,c). Correlating pathway abundance levels on these samples further confirmed this pattern, with overall correlations ranging from

https://doi.org/10.1038/s41587-023-01675-1

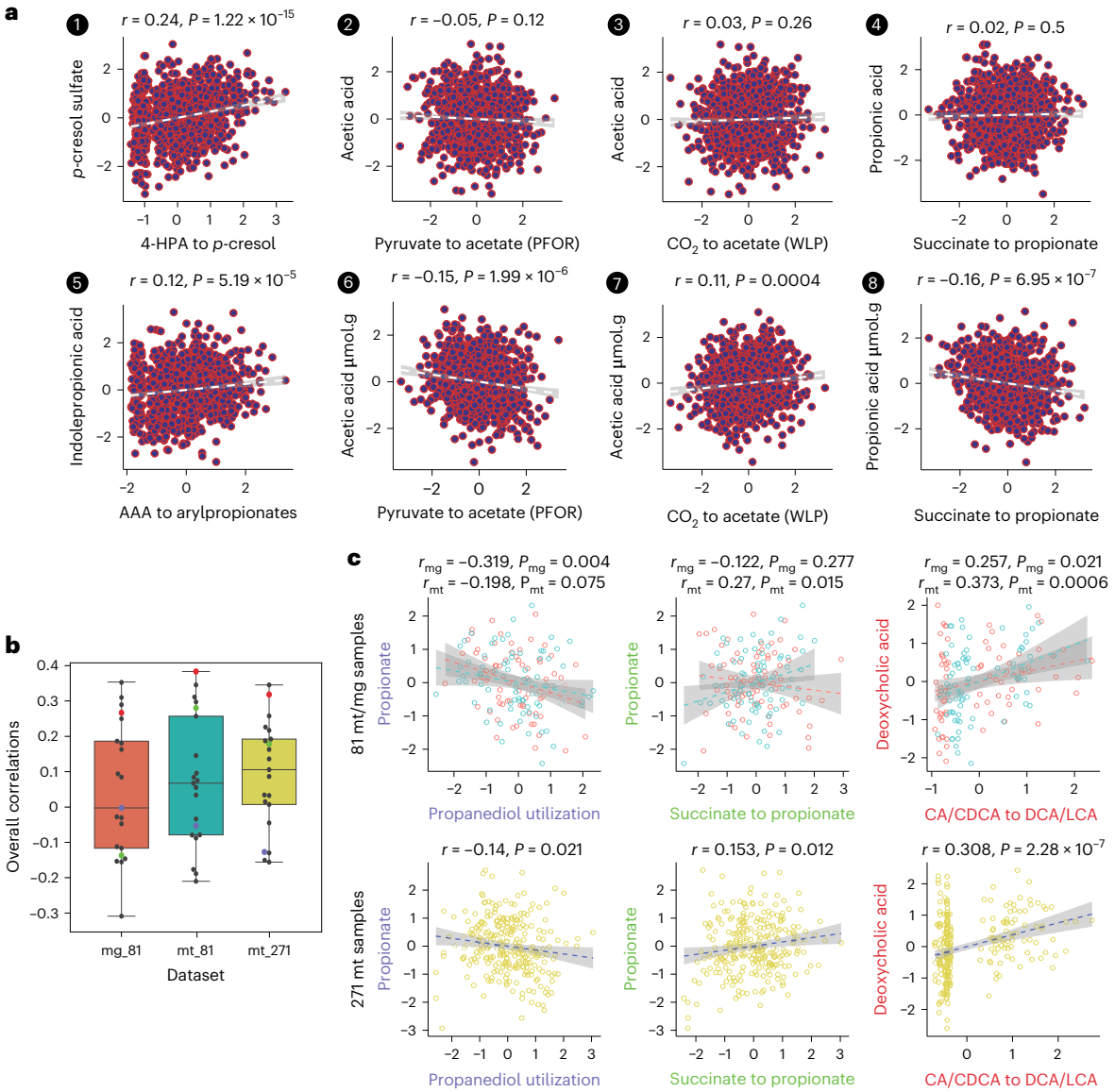


Fig. 4 | Pathway correlations with metabolomic data. a, Limited correlation of genetic pathway abundance with abundance of metabolites in blood plasma (correlation plots 1–5) and feces (correlation plots 6–8) from the LifeLines DEEP cohort (n = 1,055). The correlation plots 2–4 and 6–8 correspond to pathway association with plasma and fecal levels of the same SCFAs, respectively. The x axis indicates abundance of pathways, and the y axis indicates abundance levels of metabolites in plasma or feces. The gray line shows the best linear fit, with 95% confidence interval. Spearman correlation (two-sided) was used to check the relationship between pathway abundances and metabolite levels after adjusting for age, sex and read depth. The rank-based Spearman correlation coefficient and empirical P value are also shown. Spearman correlation (two-sided) is used to check the relationship between pathway abundances and metabolite levels after adjusting for age, sex and read depth. **b**, Overall correlation box plots between gutSMASH-predicted pathways and the iHMP data considering the 81 samples with paired metagenome/metabolome/metatranscriptome data when considering the metagenome/metabolome correlations (mg_81, red) and the metatranscriptome/metabolome correlations (mt_81, turquoise), as well as correlations for the 271 samples with metatranscriptome/metabolome data (mt_271, yellow). Individual data points are shown in the dot plot. The box limits indicate the quartiles of the dataset; the whiskers extend to 1.5× the interquartile range; and the center line denotes the median. **c**, Correlation (Spearman, two-sided) plots for three specific pathways within each dataset, with the mg_81 and mt_81 datasets being shown above in red/turquoise and the mt_271 dataset being shown below in yellow. For each pathway, a different color was used for the axis labels: purple for propanediol utilization, green for succinate to propionate and red for CA/CDCA to DCA/LCA; the corresponding data point in the box plot in **b** was colored accordingly.

−0.32 to 0.34 (Fig. 4b). Correlations did increase when splitting samples by disease status, ranging from −0.50 to 0.52 for healthy samples, −0.51 to 0.53 for patients with UC and −0.37 to 0.42 for patients with CD (Supplementary Table 11), suggesting that large-scale physiological differences (for example, differences in absolute microbial abundance) among human subjects are prominent confounding factors. Overall, our findings have important implications for analyses that make metabolic inferences from gene abundances[24] or the abundances of individual strains[25]. We speculate that a more detailed understanding of the influence of diet; differences in gene regulation; characteristic pathway flux (turnovers per unit time per protein copy), which may also be affected by secondary fermenters; and pharmacokinetic characteristics (for example, absorption, distribution, metabolism and excretion) could ultimately enable the prediction of metabolite abundance from metagenome abundance. Indeed, when we compared mapping of metatranscriptomic reads for the 81 iHMP samples (mixed phenotypes)

for which paired metagenomic/metatranscriptomic/metabolomic data were available, we already observed slightly higher correlations (ranging from −0.22 to 0.37; Fig. 4b), although the difference with the metagenomic data from the same samples was not statistically significant (Cochran's Q test coefficient ranging from 0.0013 to 0.983). At a false discovery rate (FDR) of <0.1, whereas we observed five significant associations between gutSMASH pathways and their corresponding metabolites for samples with paired metagenome/metabolome data, we observed six significant associations for samples with paired metatranscriptome/metabolome data (Supplementary Table 11). The correlations from a larger set of 271 iHMP metatranscriptomic/metabolomic samples, from which complete metadata were available (Supplementary Table 12), also seemed to show slightly stronger signals compared to the metagenome/metabolome data, with overall correlations ranging from −0.17 to 0.34 (Fig. 4b), although no direct comparison could be made in the absence of metagenome data. When split out across the three phenotypes, correlations ranged from −0.28 to 0.38 for healthy, −0.27 to 0.27 for UC and −0.24 to 0.42 for CD and yielded ten significant associations between pathway expression values and their metabolites. The correlations across these datasets varied quite a lot depending on the pathway (Fig. 4a), suggesting that the expression of some pathways is more specifically predictive for metabolite abundances. For instance, strong correlations were found between the CA/CDCA to DCA/LCA (*bai* operon) pathway with deoxycholic acid, possibly due to the fact that it is a taxonomically restricted pathway without known alternative pathways leading to the same products. In contrast, some pathways showed low or even slightly negative correlations, which may be explained by, for example, pathway competition, diffusion/transport differences or consumption by other bacteria. Overall, systematic detection of the relevant genes and gene clusters by gutSMASH provides a technological foundation for future studies to study how various factors influence microbial metabolite production and accumulation in the lumen as well as in plasma, by allowing mapping of metatranscriptomic data to these accurately defined and categorized sets of genomic loci across a wide range of conditions. Measuring absolute microbial abundance across samples will likely greatly help in this as well[26].

Discussion

The gutSMASH software constitutes a comprehensive automated tool designed to identify niche-defining primary metabolic pathways from genome sequences or metagenomic contigs. Even a full-fledged metabolic network reconstruction software such as PathwayTools[27] (which uses the extensive MetaCyc database[28]) lacks detection capabilities for two out of the 41 MGC-encoded pathways detected by gutSMASH (Supplementary Table 13). We also assessed the overlap of pathways between gutSMASH and GenomeProperties[29], and only five out of the 41 MGC-encoded pathways can be systematically annotated using the latter (Supplementary Table 13). Moreover, the identification of MGCs provides considerably increased confidence that putative detected homologs for a given pathway are truly working together. Downstream, detected MGCs can be used as input for read-based tools such as HUMAnN[30] or BiG-MAP[19] to measure abundance or expression levels of the encoded pathways. On top of these functionalities, the gutSMASH framework also facilitates identifying new (that is, uncharacterized) pathways in the microbiome. To this end, we designed an additional set of rules, referred as general rules in Fig. 1, to detect primary MGCs of unknown function that harbor at least one of the following key enzymes: Fe-S flavoenzymes[31], glycyl-radical enzymes, 2-hydroxyglutaryl-CoA-dehydratase-related enzymes and/or enzymes involved in oxidative decarboxylation. After running gutSMASH on the 4,240 microbial genomes and pulling out the putative MGCs (Supplementary Methods: 'Analysis of distant homologs and putative MGCs from CGR, HMP and Clostridioides datasets'), we found 12,256 putative MGCs from 760 different species, that, after redundancy filtering at 90% sequence similarity, were classified into 932 GCFs. Within these, we manually prioritized a range of gene clusters with unprecedented enzyme-coding gene content highlighted in Extended Data Figs. 3 and 4 (Supplementary Information results: 'Analysis of putative clusters and distant homologs: relevant candidates to study further'). These putative MGCs can be a potential source to discover new pathways and metabolites. Thus, gutSMASH can be a valuable tool in the field of enzyme/pathway discovery, to link metabolites to gene clusters and to identify genes responsible for microbiome-associated phenotypes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01675-1.

References

1. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
2. Karp, P. D. et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20**, 1085–1093 (2019).
3. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
4. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
5. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes—a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
6. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571 (2021).
7. Navarro-Muñoz, J. C. et al. A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
8. Kitamoto, S. et al. Dietary L-serine confers a competitive fitness advantage to Enterobacteriaceae in the inflamed gut. *Nat. Microbiol* **5**, 116–125 (2020).
9. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
10. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
11. Tracy, B. P., Jones, S. W., Fast, A. G., Indurthi, D. C. & Papoutsakis, E. T. Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. *Curr. Opin. Biotechnol.* **23**, 364–381 (2012).
12. Maguire, F. et al. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb. Genom.* **6**, mgen000436 (2020).
13. Vieira-Silva, S. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol* **1**, 16088 (2016).
14. Rautio, M et al. Reclassification of *Bacteroides putredinis* (Weinberg et al., 1937) in a new genus *Alistipes* gen. nov., as *Alistipes putredinis* comb. Nov., and description of *Alistipes finegoldii* sp. Nov., from human sources. *Syst. Appl. Microbiol.* **26**, 182–188 (2003).
15. Cummings, J. H., Pomare, E. W., Branch, W. J., Naylor, C. P. & Macfarlane, G. T. Short chain fatty acids in human large intestine, portal, hepatic and venous blood. *Gut* **28**, 1221–1227 (1987).
16. Jones, S. A. et al. Anaerobic respiration of *Escherichia coli* in the mouse intestine. *Infect. Immun.* **79**, 4218–4226 (2011).

17. Tremblay, P. L., Zhang, T., Dar, S. A., Leang, C. & Lovley, D. R. The Rnf complex of *Clostridium ljungdahlii* is a proton-translocating ferredoxin:NAD^+ oxidoreductase essential for autotrophic growth. *mBio* **4**, e00406–e00412 (2012).
18. Liu, Y. et al. *Clostridium sporogenes* uses reductive Stickland metabolism in the gut to generate ATP and produce circulating metabolites. *Nat. Microbiol* **7**, 695–706 (2022).
19. Andreu, V. P. et al. BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. *mSystems* **6**, e0093721 (2021).
20. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
21. Faber, F. et al. Respiration of microbiota-derived 1,2-propanediol drives *Salmonella* expansion during colitis. *PLoS Pathog.* **13**, e1006129 (2017).
22. Andriamihaja, M. et al. The deleterious metabolic and genotoxic effects of the bacterial metabolite *p*-cresol on colonic epithelial cells. *Free Radic. Biol. Med.* **85**, 219–227 (2015).
23. Funabashi, M. et al. A metabolic pathway for bile acid dehydroxylation by the gut microbiome. *Nature* **582**, 566–570 (2020).
24. Mallick, H. et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* **10**, 3136 (2019).
25. Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).
26. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
27. Karp, P. D. et al. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioform.* **22**, 109–126 (2021).
28. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes—a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).
29. Richardson, L. J. et al. Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes. *Nucleic Acids Res.* **47**, D564–D572 (2019).
30. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
31. Pascal Andreu, V., Fischbach, M. A. & Medema, M. H. Computational genomic discovery of diverse gene clusters harbouring Fe-S flavoenzymes in anaerobic gut microbiota. *Microb. Genom.* **6**, e000373 (2020).
32. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
33. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

Methods

gutSMASH is a Python-based pipeline that was built from antiSMASH version 5.0 source code. The latest command line version is freely available and can be downloaded and installed at https://github.com/victoriapascal/gutsmash/tree/gutsmash.

Finding pathway signatures for known and characterized MGCs

To create a new set of detection rules, 41 known and characterized MGCs were gathered from the literature and used as positive controls. The protein sequences of these MGCs were searched using hmmscan (HMMER suit version 3.1b2, February 2015; http://hmmer.org/). From the resulting pHMM hits, auxiliary and core domains were manually identified for each pathway, to ultimately determine the pathway signature and specify it in the corresponding detection rule. To discern and more precisely identify key enzymes of interest sharing a keystone domain, we used custom-made pHMMs following a procedure described in the Supplementary Methods: 'Toward a more robust MGC identification by building new HMM profiles'. Altogether, the knowledge on the core enzyme-coding genes and the newly built pHMMs helped to construct a preliminary set of detection rules to predict known pathways.

New HMM profiles for robust MGC identification

Certain core domains are shared across diverse pathways, including the PFL-like domain and the HGD-D domain. In total, 13 keystone domains were found to be ubiquitous in multiple pathways (Supplementary Table 14). Hence, to increase gutSMASH precision and discern between enzyme subfamilies of interest, 12 protein superfamily phylogenies were constructed by aligning the protein sequences harboring the domain of interest from the MGC collection (Methods: 'Exploring the yet unknown metabolic diversity by creating general detection rules'; for an example, see Supplementary Fig. 3); the respective reference proteome[34] at a 15% or 35% co-membership threshold (the latter only for the domains Gly_radical and Acyl-CoA_dh_1); and any experimentally characterized UniProt representatives. After aligning the sequences with hmmalign[35], approximately maximum-likelihood phylogenetic trees using FastTree 2.1 (ref. [36]) were inferred to further annotate the tree with iTOL[32]. Thus, from the desired and functionally relevant clades, specific pHMMs were built by extracting the amino acid sequence of the clade-specific proteins, aligning them with Clustal Omega[37], trimming the edges of the multiple sequence alignment using Jalview[38], re-aligning all the sequences with Clustal Omega and finally building a pHMM using hmmbuild (HMMER suite version 3.1b2, February 2015; http://hmmer.org/). Subsequently, for all the newly created pHMMs, sensitivity was assessed using ten-fold jackknife cross-validation. Each clade was divided randomly into training and testing sets. The protein sequences from the training set were aligned using Clustal Omega and used to create a pHMM. Next, the protein sequences of the test set were hmmscanned (HMMER suit version 3.1b2, February 2015; http://hmmer.org/) against the newly built testing pHMMs. When a sequence scored positively for multiple domains in the same region, only the domain with a higher bit score was picked out. Sensitivity then accounted for the number of sequences positively associated with the correct pHMM out of the total number of sequences in the testing set. The same procedure was repeated ten times. The pHMMs with a true-positive rate higher than 0.85 across the ten rounds were included in the detection rules. In total, 43 newly built pHMMs were included in the corresponding detection rules (Supplementary Table 15). Moreover, a pHMM to capture succinate dehydrogenase/fumarate reductase was built by aligning ten protein sequences of such enzymes and building the model from this alignment using a hmmbuild. To also competitively score similar Pfam domains, Hhsearch pre-computed results obtained from the Pfam FTP (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/) were parsed and included in the gutSMASH code.

Testing and validating detection rules for known pathways

To evaluate the performance of the preliminary set of detection rules, a total of 1,621 bacterial genomes, including 1,520 genomes from the CGR collection[9] and 101 manually selected genomes from the most representative bacterial genera in the human gut, were used as input for gutSMASH (Supplementary Table 3). The predicted MGCs were classified based on the detection rule that they were predicted from, to later run BiG-SCAPE on each sub-collection. The resulting networks were screened individually to evaluate the taxonomic and architectural diversity, to assess if any architectural variant or taxon (based on literature) was missing from the MGC pool or was incorrectly predicted by the detection rule. Hence, this procedure ultimately helped to tweak the detection rules to predict true homologs of the known pathways (Supplementary Results and Supplementary Table 16). After two iterations of fine-tuning and testing, all detection rules were performing as intended and constituted the new set of detection rules of gutSMASH version 1.0.

gutSMASH customized databases and output visualization

The antiSMASH version 5.0 source code was further tailored to meet gutSMASH functionality. The 32,144 predicted MGCs obtained from running gutSMASH on the CGR, HMP and Clostridiales collections (Methods: 'Evaluating functional potential of gut bacteria using gutSMASH') were used to create the ClusterBlast database. In a similar way, 59 positive controls carrying the known pathways (from which we created the specific-to-known pathway detection rules) were used to create the KnownClusterBlast database. These databases facilitate comparative gene cluster analysis using BLAST[39]. Thus, they allow assessing how broadly distributed an MGC is across bacteria (in the case of ClusterBlast) or evaluating the similarity between the predicted MGC and a known and functionally characterized MGC (when using KnownClusterBlast).

Another functionality of antiSMASH is to classify coding genes based on the domains into five major functional categories: core biosynthetic, additional biosynthetic, transport-related, regulatory, resistance and other, using the pmCOG (primary metabolism Clusters of Orthologous Groups) tool, which is embedded in antiSMASH (there, originally named smCOG for secondary metabolism Clusters of Orthologous Groups). Thus, the pHMM library pmCOG was updated to include relevant domains found in specialized primary metabolism. Also, two other important functional categories were added: electron transport-related genes and encapsulation genes.

Exploring unknown metabolic diversity using general rules

With the objective of creating general detection rules to predict putative MGCs, a similar approach used to screen the surrounding genes around a Fe-S flavoenzyme coding gene was used[31]. Some of the representative known pathways share proteins with biochemically similar functions; these include, for instance, PFL-like enzymes that are found in the threonine-to-propionate pathway, the choline utilization pathway and the pyruvate-to-acetate pathways. To cover a large amount of sequence diversity, we created a database that included 11,000 complete genomes and 98,886 draft genomes available in GenBank (in February 2017) to use clusterTools[40], a software to find remote homologs of known MGCs. As input, a subset of the known pathways used to design the detection rules for known pathways was used as input (Supplementary Table 17). The output of several iterated clusterTools searches were grouped to acquire a collection of over 29,000 clusters. For visualization and manual scoring purposes, MultiGeneBlast[41] was run using the clusterTools output as input. Thus, MGCs harboring at least half of the genes from the query gene list and with a cumulative BLAST score higher than 1,000 were included in the MGC collection. To filter out redundant sequences, we used MMseqs2 (ref. [42]) at a 95% similarity cutoff. From the resulting network of 1,599 groups, a maximum of one random representative plus singletons were picked

creating a 'non-redundant' set of almost 3,200 clusters. This collection was screened for gene clusters harboring the *baiCD* or *baiH* coding gene (Oxidored_FMN and Pyr_redox_2), pyruvate formate-lyase (PFL-like or Gly_radical), pyruvate ferredoxin (POR, POR_N or PFOR_II), thiamine pyrophosphate enzyme (TPP_enzyme_C) and 2-hydroxyglutaryl-CoA dehydratase (HGD-D), each of which are keystone domains in charge of important anaerobic reactions. This helped create general detection rules, by identifying which other enzyme-coding Pfam domains are found around these 'anchor' domains in flanking regions; this was systematically analyzed per gene cluster family to make sure that the general rules captured all major families of homologous MGCs of interest. Also, when validating the specific-to-known pathway detection rules, whenever a specific rule predicted interesting MGCs that were variants of the representative pathway with likely differing functions, a general rule was created out of the specific one by loosening up the Pfam requirements. The full list of general rules can be found in Supplementary Table 18.

Assessing single-protein pathway abundance

To include single-protein pathways in our analysis to assess the overall abundance of specialized primary metabolic pathways, ten enzyme families were selected for downstream analysis. Following the same procedure as described in Methods: 'Toward a more robust MGC identification by building new HMM profiles', protein phylogenies were built for each protein superfamily. Similarly, from the pathway-specific monophyletic clades, we built new pHMMs. A bitscore threshold for each newly built pHMM was calibrated to identify with high confidence proteins belonging to the same functional clades. To this end, the protein sequences that composed the superfamily phylogeny were subjected to an hmmsearch run with the new pHMM. The bitscore reported by hmmsearch for the most distantly related protein within the pathway-specific clade was chosen as the threshold for that specific pHMM. Next, the protein sequences from the CGR, HMP and Clostridiales collections (further information in Methods: 'Evaluating the functional potential of the human microbiome using gutSMASH') were scanned using the newly built pHMMs. Finally, the hmmsearch output tables for each pHMM were parsed so that the proteins with a bitscore equal to or higher than the chosen threshold were deemed hits. In those cases in which the single-protein sequence codes for two Pfam domains—as, for instance, the serine dehydratase (SDH_alpha and SDH_beta)—one of the Pfam domains was selected to create a protein phylogeny to further build a clade-specific pHMM, in this case SDH_alpha. Then, the protein sequences from the three collections were subjected to hmmsearch runs with both the clade-specific pHMM and the other co-occurring Pfam domain (in this case SDH_beta). The sequences that harbor both the specific pHMM at the chosen threshold and the co-occurring domain with an *e* value $\leq 10^{-5}$ were deemed hits.

Evaluating the functional potential of the human microbiome

To evaluate the metabolic potential of the human microbiome, gutSMASH was run on three different genome collections: (1) the CGR collection, with 1,520 CGR genomes deposited under PRJNA482748; (2) the HMP reference genomes, with 2,146 HMP bacterial genomes downloaded in September 2019 from https://www.hmpdacc.org/hmp/catalog/grid.php?dataset=genomic; and (3) 414 Clostridiales complete genomes under taxonomy ID 186802. The genomic FASTA sequence of these genomes was used as input for gutSMASH, which used Prodigal[43] to annotate genes across all of them in a consistent way. Moreover, to assess which MGC belonged to known pathways, the KnownClusterBlast (Supplementary Methods: 'gutSMASH customized databases and output visualization') option was enabled. Thus, from the KnownClusterBlast output, the predicted regions were classified as known when the following two requirements were met: (1) an overall pathway similarity of at least 50% and at least half of the genes with a minimum protein sequence similarity of 40% and (2) an overall similarity of 60% and half of the genes with protein sequence similarity higher than 30%. However, in order not to penalize MGCs with similar domain profiles but substantially larger sizes, the requirements to be considered 'known' slightly changed for the KnownClusterBlast MGCs longer than 17 coding genes. In this case, the same requirements as described above were used, but, instead of considering candidates with at least half of the coding genes having either 30% or 40% minimum sequence identity, one third of the genes were required to be present with the same minimum sequence identity. This was the case for the ethanolamine utilization operon, the *bai* operon characterized from *C. scindens* ATCC35704 (CA/CDCA to DCA/LCA pathway), the acetyl-CoA pathway (CO_2 to acetate (Wood–Ljungdahl pathway (WLP))), the tetrathionate-to-thiosulfate pathway and the NADH dehydrogenase I complex. Thus, all the MGCs that did not satisfy these conditions were classified as putative MGCs. The phylogenetic tree in Fig. 2 was generated using phyloT version 2 (https://phylot.biobyte.de/). The Genome Taxonomy Database[32] was used to assign the taxonomy to the genomes of the three collections (when present), and those taxonomic identifiers were the ones used for the subsequent pathway absence/presence analysis. Finally, the tree was annotated using iTOL[33].

Analysis of distant homologs and putative MGCs

The putative MGCs predicted from the CGR, HMP and Clostridiales genome collections were selected following the definition of 'known' and 'putative' gene clusters stated in Methods: 'Evaluating the functional potential of the human microbiome using gutSMASH'. To account for redundant MGCs, protein sequences extracted from all gene clusters were subjected to a redundancy filtering of 90% sequence similarity using MMseqs2. From the resulting clustering, two random representatives were chosen from each group, including the singletons. The resulting non-redundant collection of 3,040 putative MGCs was used as input for BiG-SCAPE using the default thresholds. The network in Extended Data Fig. 3 was constructed and annotated using Cytoscape version 3.0 (ref. [44]).

Mapping metagenomics reads from cohort samples to MGCs

The HMP-, CGR- and Clostridiales-predicted MGCs were used as input for BiG-MAP[19], a tool that assesses gene cluster abundance or expression across metagenomics or metatranscriptomics data, respectively, by mapping the genomic reads onto the gene cluster sequences. The BiG-MAP family module grouped the 32,146 MGCs into 5,655 GCFs. Next, the reads of 1,135 participants of the population-based cohort LifeLines-DEEP[45] (quality filtered using KneadData version 0.7.2) were mapped onto the resulting 5,655 Mash[46] representative MGCs by using the BiG-MAP.map module. (All LifeLines participants signed an informed consent form before sample collection. The ethics review board of the University Medical Center Groningen approved the study [reference no. M12.113965].) To assess the abundance of known pathways, the RPKM values from the known MGCs (following the definition of 'known' stated in Methods: 'Evaluating the functional potential of the human microbiome using gutSMASH') were pulled out. The RPKM values of all the MGCs predicted by the same detection rule were merged. The pathway abundance (RPKM) was computed by dividing the gene clusters in 2-kb-sized bins and assessing the lower quartile number of reads mapping the 2-kb bins for each gene cluster and sample. In contrast, a pathway was annotated as present in a sample when reads from that sample were found to be mapping to at least 5% of the core region of that MGC. This threshold was kept low to enable detection of MGCs from low-abundant microbes and avoid false negatives due to limited sequencing depth. The lowest percentage identity of reads mapped to MGCs was 78% at the nucleotide level, which instilled confidence that finding multiple reads mapping to different locations within a MGC provides sufficient evidence for its presence in a sample. The pathway prevalence was also computed using 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80% core coverage thresholds (Extended Data Fig. 1), and results for

increasing thresholds were consistent with gradual loss of detection capability for pathways known to be associated with low-abundance bacteria, such as the AAA-to-arylpropionate pathway (AAA reductive branch). To also take into account metatranscriptomics and fecal metabolome data, the 81 paired metagenomes, metatranscriptomes and metabolomes from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB) study[47] were similarly analyzed using BiG-MAP. In this case, to speed up calculation, only the bacterial genomes whose gutSMASH run predicted at least one 'known' (following the definition as described above) MGC were used as input for the BiG-MAP.family module. The same 'known' MGC collection as described above was used. In total, for the BiG-MAP.family module, 1,764 gutSMASH runs were used, which included 8,109 gene clusters that were downsized to 6,301. This reduced MGC reference collection was then used by the BiG-MAP.map module that aligned the metagenomic and metatranscriptomics to the reference collection independently. The same procedure was followed for the 271 samples with paired metatranscriptome and metabolome data.

Correlating pathway abundance with metabolite concentrations

To evaluate the correlation between the gene cluster abundance and metabolite concentrations, the masses of seven metabolites derived from several gutSMASH-predicted gene clusters could be found in the mass spectrometry (MS) data of the plasma measured in LifeLines DEEP[20,45]. Untargeted metabolomics profiling was done using flow-injection time-of-flight mass spectrometry (FI-MS) as described by Chen et al.[48]. These metabolites included acetic acid, indolepropionic acid, isovaleric acid, *p*-cresol, *p*-cresol sulfate, phenylacetic acid and propionic acid (Fig. 4 and Extended Data Fig. 2). Both metabolite and pathway abundance (RPKM counts) were inverse-rank-transformed, and the linear regression was applied to adjust covariates, including age, sex and metagenomic sequencing depth (only for pathway abundance). Metabolite and pathway abundance residuals from the linear regression model were then used to perform the Spearman correlation test. Finally, the Benjamini–Hochberg method was applied to control for FDR. The RPKM counts of the gutSMASH-predicted pathways involved in the synthesis of SCFAs were correlated in the same manner with the fecal SCFA MS data also collected from the LifeLines DEEP cohort. Specifically, the SCFAs measured in the fecal metabolomes were acetate, propionate, butyrate and caproate.

To further assess the relationship between MGC abundance/expression with fecal metabolite concentration, the data derived from analyzing the IBDMDB data with BiG-MAP were used similarly as the Lifelines DEEP data. In this case, gene cluster abundance and expression values were correlated with the following fecal metabolites: betaine, butyrate, deoxycholate, glutamate, hydrocinnamate, indole-3-propionate, lithocholate, *p*-hydroxyphenylacetate, phenylacetate, proline, propionate, putrescine, spermidine, succinate and TMAO. Correlations were made for each individual subgroup in the dataset that included CD, UC and healthy samples.

Reporting summary

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

Data availability

The LifeLines DEEP cohort raw metagenomic sequencing data, metabolome data and human phenotypes (that is, age and sex) used for the analysis presented in this study are available at the European Genome-phenome Archive under accession EGAS00001001704. Taxonomic assignments of bacteria were performed according to Genome Taxonomy Database release 95 (https://gtdb.ecogenomic.org/). Lists of accessions of genome assemblies used are available in Supplementary Tables 3 and 4. iHMP multi-omics data were downloaded from https://ibdmdb.org. Raw sequence data of the iHMP are available from the National Center for Biotechnology Information's Sequence Read Archive via BioProject PRJNA398089; metatranscriptome data are available through Gene Expression Omnibus series accession number GSE111889; and metabolomics data are available at the Metabolomics Workbench (http://www.metabolomicsworkbench.org; Project ID PR000639). Source data are provided with this paper.

Code availability

The gutSMASH source code is available freely under an open-source AGPL-3.0 license from https://github.com/victoriapascal/gutsmash/.

References

34. Chen, C. et al. Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS ONE* **6**, e18910 (2011).
35. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
36. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
37. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
38. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
39. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
40. de los Santos, E. L. C. & Challis, G. L. clusterTools: proximity searches for functional elements to identify putative biosynthetic gene clusters. Preprint at https://www.biorxiv.org/content/10.1101/119214v2 (2017).
41. Medema, M. H., Takano, E. & Breitling, R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol. Biol. Evol.* **30**, 1218–1223 (2013).
42. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
43. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
44. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
45. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
46. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
47. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
48. Chen, L. et al. Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome. *Nat. Med.* **28**, 2333–2343 (2022).

Acknowledgements

This work was supported by the Chan Zuckerberg Biohub (M.A.F.); DARPA awards HR0011-15-C-0084 and HR0112020030 (M.A.F.); National Institutes of Health (NIH) awards R01 DK101674, DP1 DK113598 and P01 HL147823 (to M.A.F.); the Leducq Foundation; and a European Research Council (ERC) Starting Grant (948770-DECIPHER to M.H.M.). A.Z. is supported by ERC Starting Grant 715772; Netherlands Organization for Scientific Research NWO-VIDI grant 016.178.056;

Netherlands Heart Foundation CVON grant 2018-27; and NWO Gravitation grant ExposomeNL 024.004.017. J.F. is supported by the ERC Consolidator Grant (grant agreement no. 101001678); NWO-VICI grant VI.C.202.022; Dutch Heart Foundation IN-CONTROL (CVON2018-27); the Netherlands Organ-on-Chip Initiative; and the NWO Gravitation Project (024.003.001), funded by the Ministry of Education, Culture and Science of the government of The Netherlands. L.C. is supported by a Foundation de Cock-Hadders grant (20:20-13) and a joint fellowship from the University Medical Centre Groningen and the China Scholarship Council (CSC201708320268). D.D. was supported by NIH awards K08 DK110335, R35 GM142873 and R01 AT011396.

Author contributions

M.A.F. and M.H.M. initially conceived the project, with modifications and extensions introduced on the advice of V.P.A., A.Z., J.F. and D.D. The gutSMASH software was developed and used to analyze genomic data by V.P.A., with input from M.H.M., D.D. and M.A.F. Analysis of metagenomic and metatranscriptomics data was performed by H.E.A., V.P.A. and L.C. Correlations with metabolomic data were performed by L.C. M.H.M., D.D. and M.A.F. coordinated and supervised the study as a whole, and A.Z. and J.F. coordinated and supervised analysis of cohort data. All authors contributed to data interpretation. V.P.A., M.A.F., D.D. and M.H.M. drafted the initial manuscript, with input from the other authors. All authors read and contributed to the final manuscript.

Competing interests

M.A.F. is a co-founder and director of Federation Bio, a co-founder of Revolution Medicines and a member of the scientific advisory board of NGM Biopharmaceuticals. D.D. is a co-founder of Federation Bio. M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. The remaining authors declare no competing interests.

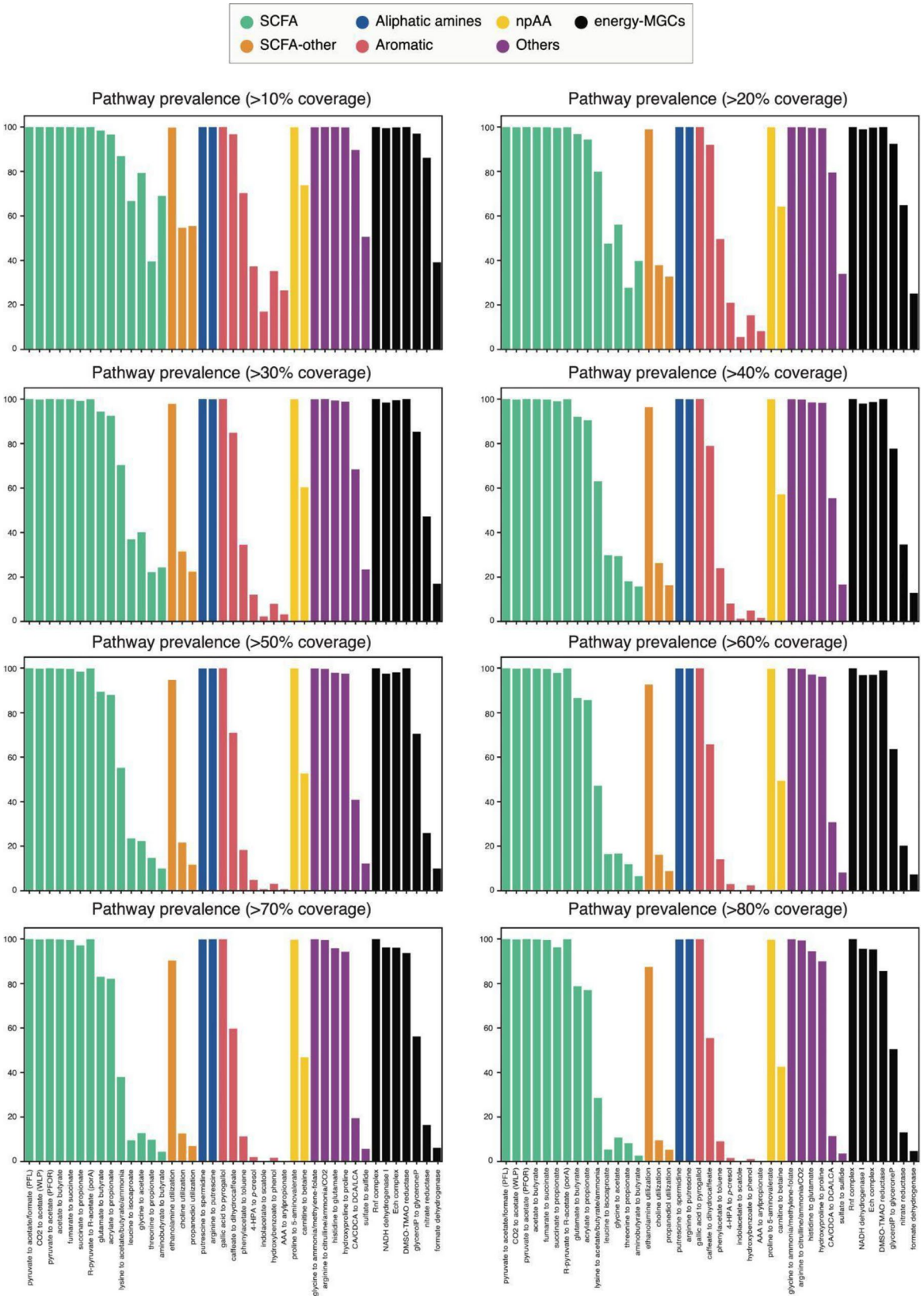
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-023-01675-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01675-1.

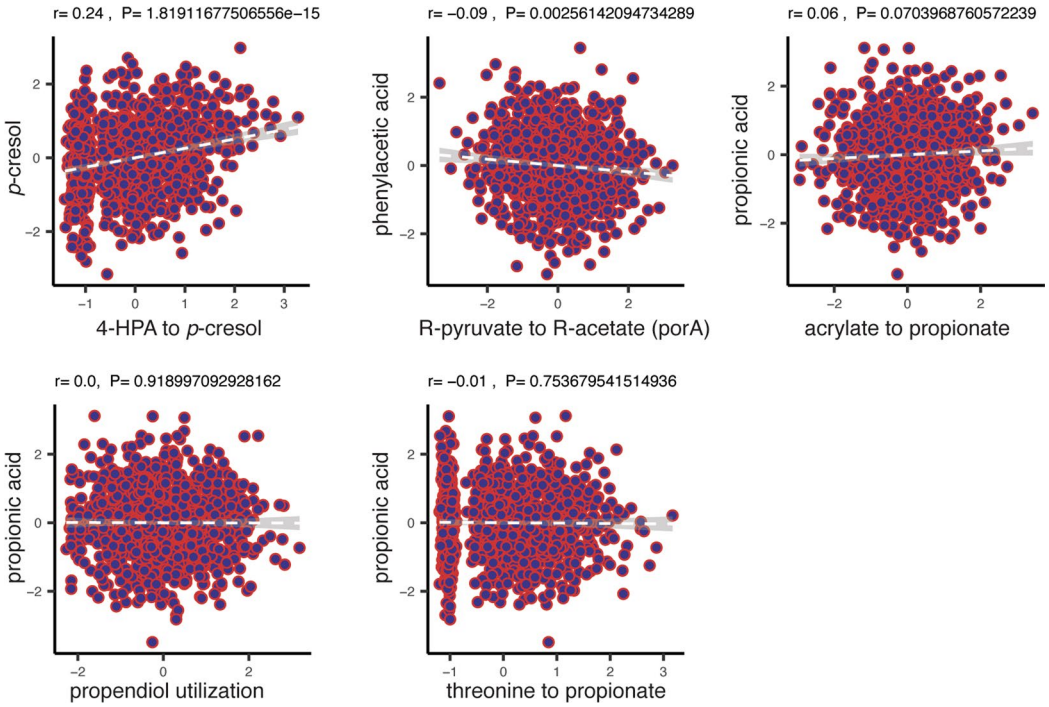
Correspondence and requests for materials should be addressed to Michael A. Fischbach, Dylan Dodd or Marnix H. Medema.

https://doi.org/10.1038/s41587-023-01675-1



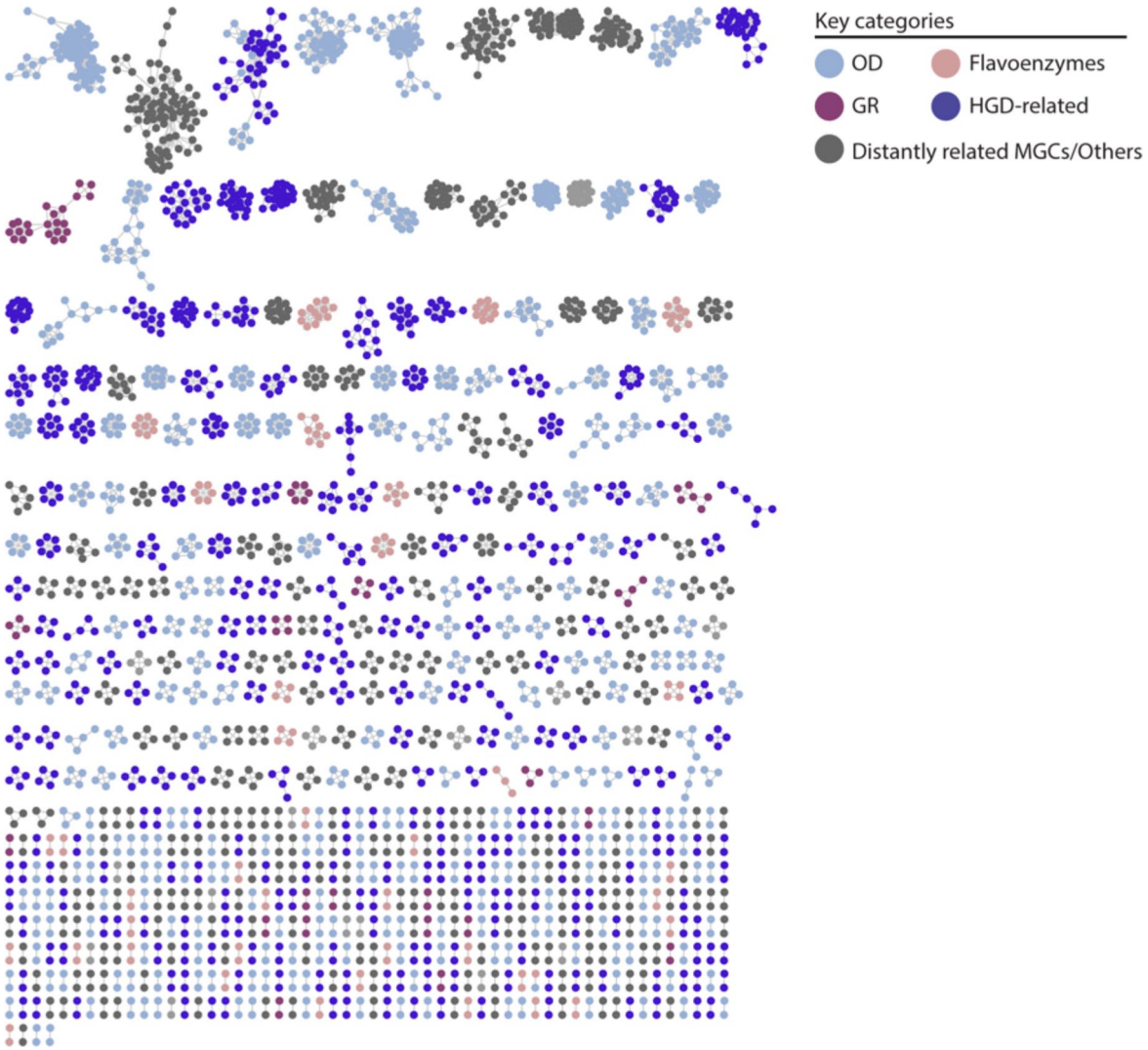
Extended Data Fig. 1 | Pathway prevalence using different core coverage thresholds. Pathway prevalence was computed by assessing the number of reads (per sample) mapping to known gene clusters at a certain core coverage cutoff. The figure illustrates how the pathway prevalence gradually changes when increasing the core coverage cutoff from 10 to 80%.

https://doi.org/10.1038/s41587-023-01675-1



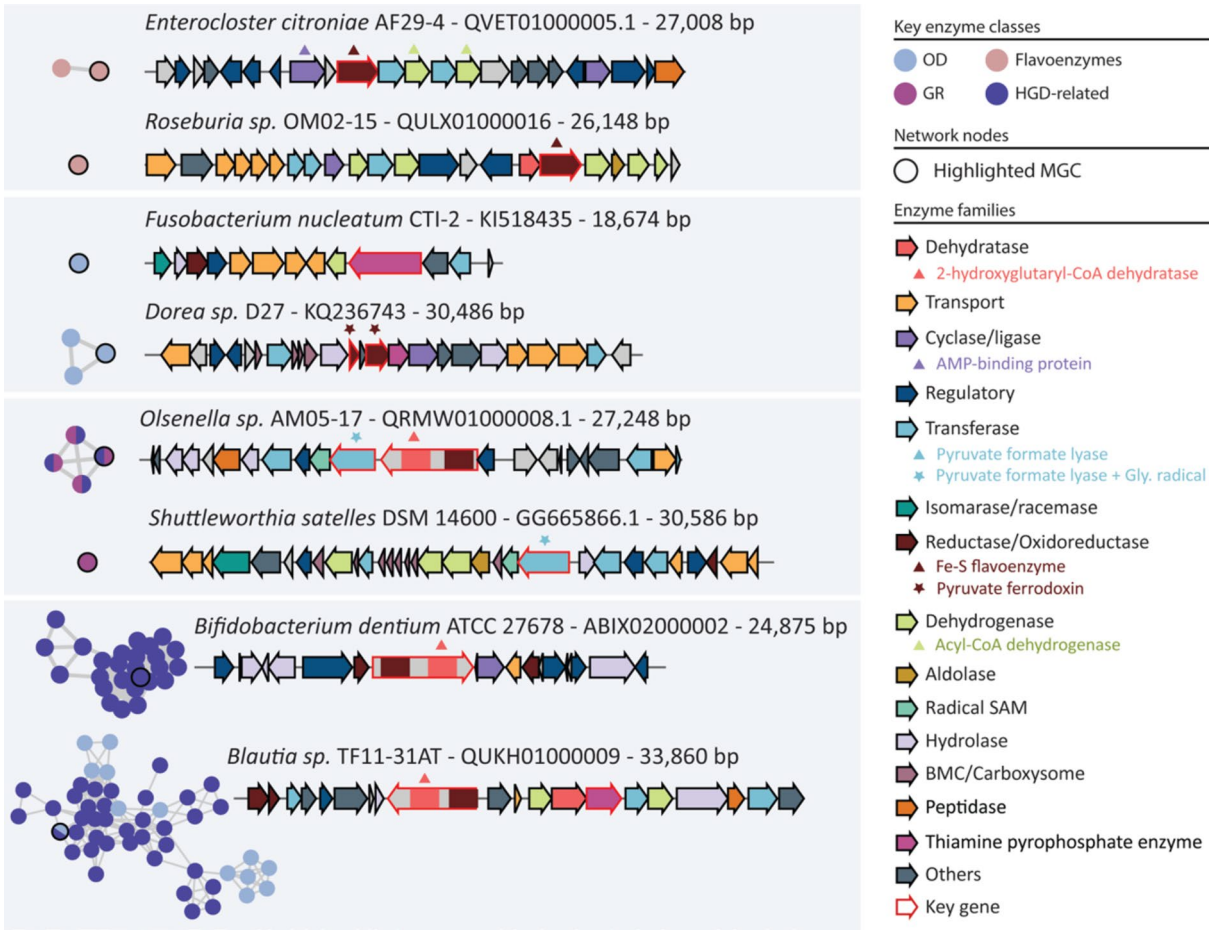
Extended Data Fig. 2 | Limited correlation of genetic pathway abundance with metabolites abundance in blood plasma. This figure shows correlation plots for additional metabolites not shown in Fig. 4a. Spearman correlation (two sided with rho and empirical *P* value are reported) is used to check the relationship between pathway abundances and metabolite levels after adjusting for age, sex and read depth. $n = 1054$ biologically independent samples.

https://doi.org/10.1038/s41587-023-01675-1



Extended Data Fig. 3 | Network of putative non-redundant MGCs predicted by gutSMASH. From all the unknown predicted MGCs, a redundancy filtering of 0.9 sequence similarity was applied using MMseqs2. From each cluster, two representatives were picked, and all representatives were used as input for BiG-SCAPE using the default cutoffs. The network contains 2,921 nodes and 7,474 edges. The MGCs have been classified into four different categories based on the key enzyme classes they code for. The GR (glycyl-radical) category is composed of MGCs that include pyruvate formate-lyase (PFL-like) and/or glycyl radical (Gly_radical), OD (oxidative decarboxylation) involves MGCs with at least one of the following Pfam domains: pyruvate ferredoxin/flavodoxin oxidoreductase (POR), pyruvate flavodoxin/ferredoxin oxidoreductase, thiamine diP-bdg (POR_N), pyruvate:ferredoxin oxidoreductase core domain II (PFOR_II) and thiamine pyrophosphate enzyme, C-terminal TPP binding domain (TPP_enzyme_C). The flavoenzymes category is a combination of MGCs harbouring at least one of the custom-made BaiCD and BaiH pHMMs. HGD-D-related MGCs, as the name states, include enzymes matching any of the 2-hydroxyglutaryl-CoA dehydratase, D-component (HGD-D)-related pHMM domains.

https://doi.org/10.1038/s41587-023-01675-1



Extended Data Fig. 4 | Subset of unknown MGCs predicted by gutSMASH manually picked. The network/nodes present in the left side of the figure represent the subnetwork extracted from the complete network in Extended Data Fig. 3. The arrows have been coloured-coded based on the Pfam domains found in the protein-coding sequences and the functional annotations of these proteins.

nature research

Corresponding author(s): Michael Fischbach, Dylan Dodd, Marnix Medema

Last updated by author(s): 2023/01/02

Reporting Summary

Nature Research 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 Research 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	No software was used for data collection.
Data analysis	Sequence similarity networks were constructed using BiG-SCAPE v1.0 and annotated in Cytoscape v3.0. Read mapping to metabolic gene clusters was performed using BiG-MAP v1.0, which makes use of Bowtie v2.1.0 for mapping. Redundancy of gene clusters was controlled using clustering with MMseqs v2. Phylogenetic trees were constructed using FastTree v2.1 and visualized using iTOL v5, based on alignments generated using Clustal Omega v1.2.2 and curated in JalView v2. Profile Hidden Markov models were built and used to search sequence data using HMMER v3.1b2. Homologous gene clusters were identified using clusterTools v1 and MultiGeneBlast v1.1.14. Raw metatranscriptome/ metagenome reads were filtered using kneadData (v0.4.6.1). The gutSMASH software written for this study is available from https://github.com/victoriapascal/gutsmash.

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The LifeLines DEEP cohort raw metagenomic sequencing data, metabolome data and human phenotypes (i.e. age and sex) used for the analysis presented in this study are available at the European Genome-phenome Archive under accession EGAS00001001704. Taxonomic assignments of bacteria were performed according

to the Genome Taxonomy Database release 95 (https://gtdb.ecogenomic.org/). Lists of accessions of genome assemblies used are available in Tables S3 and S4. iHMP multi-omics data were downloaded from https://ibdmdb.org. Raw sequence data of the iHMP are also available from the NCBI sequence read archive (SRA) via BioProject PRJNA398089, metatranscriptome data through GEO Series accession number GSE111889, and metabolomics data at the Metabolomics Workbench (http://www.metabolomicsworkbench.org; Project ID PR000639).

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	This study included 1,135 individuals from the population-based LifeLines-DEEP cohort with fecal metagenomic sequencing data available. 1054 of them also have plasma metabolomics available. In order to ensure the analysis power, the study includes as much as subjects as possible. Thus no sample size calculation was performed.
Data exclusions	Samples with low quality of metagenomics sequencing were excluded.
Replication	There was no direct replication in independent cohorts. However, indirect replication and comparison were performed. For example, we compared associations between plasma and fecal levels of the same metabolites. We also compared the associations between metagenomics from the LLD cohort and metatranscriptomics data of the iHMP cohort.
Randomization	This is human cohort-based analysis. The sample collection, sequencing and metabolomics profiling were performed in a random order. No extra randomization was done for this study. We included age, sex and read depth of sequencing data as covariates in our correlation analyses.
Blinding	This study is a human cohort based, observational study. Thus no blinding was performed.

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
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The study has included the population-based LifeLines-Deep DEEP (n=1,135), with 58.20% female. The mean age (SD) of participants was 45.04 (13.60) years and their mean (SD) BMI was 25.26 (4.18);
Recruitment	The LifeLines-DEEP cohort was a random subset of the population-based Lifelines cohort. The recruitment of participants was through the LifeLines organization. Eligible participants were invited to participate in the LifeLines Cohort Study through their GP. A large number of GPs within the northern three provinces of The Netherlands (Friesland, Groningen and Drenthe) were involved and invited all their patients between the ages of 25 and 50 years, unless the participating GP considered the patient not eligible based on the following criteria: severe psychiatric or physical illness; limited life expectancy (<5 years);insufficient knowledge of the Dutch language to complete a Dutch questionnaire. Participants were asked to indicate whether their family members, such as partners, parents, parents-in-law and children would also be willing to participate in the study. If so, permission was asked to send them an invitation to participate. Children could only participate if one of their parents was a participant. In addition, inhabitants of the northern provinces could also register themselves via the LifeLines website.

Ethics oversight	All participants signed an informed consent form prior to sample collection. Institutional ethics review board (IRB) approval was available for the Lifelines DEEP (ref. M12.113965).

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