Article

Evaluating microbiome-directed fibre snacks in gnotobiotic mice and humans

https://doi.org/10.1038/s41586-021-03671-4

Received: 1 July 2020

Accepted: 25 May 2021

Published online: 23 June 2021

Check for updates

Omar Delannoy-Bruno[1,2], Chandani Desai[1,2], Arjun S. Raman[1,2,3], Robert Y. Chen[1,2], Matthew C. Hibberd[1,2,3], Jiye Cheng[1,2,3], Nathan Han[1,2], Juan J. Castillo[4], Garret Couture[4], Carlito B. Lebrilla[4], Ruteja A. Barve[5], Vincent Lombard[6], Bernard Henrissat[6,7], Semen A. Leyn[8], Dmitry A. Rodionov[8,9], Andrei L. Osterman[8], David K. Hayashi[10], Alexandra Meynier[10], Sophie Vinoy[10], Kyleigh Kirbach[11], Tara Wilmot[11], Andrew C. Heath[12], Samuel Klein[11], Michael J. Barratt[1,2,3] & Jeffrey I. Gordon[1,2,3]✉

Changing food preferences brought about by westernization that have deleterious health effects[1,2]—combined with myriad forces that are contributing to increased food insecurity—are catalysing efforts to identify more nutritious and affordable foods[3]. Consumption of dietary fibre can help to prevent cardiovascular disease, type 2 diabetes and obesity[4–6]. A substantial number of reports have explored the effects of dietary fibre on the gut microbial community[7–9]. However, the microbiome is complex, dynamic and exhibits considerable intra- and interpersonal variation in its composition and functions. The large number of potential interactions between the components of the microbiome makes it challenging to define the mechanisms by which food ingredients affect community properties. Here we address the question of how foods containing different fibre preparations can be designed to alter functions associated with specific components of the microbiome. Because a marked increase in snack consumption is associated with westernization, we formulated snack prototypes using plant fibres from different sustainable sources that targeted distinct features of the gut microbiomes of individuals with obesity when transplanted into gnotobiotic mice. We used these snacks to supplement controlled diets that were consumed by adult individuals with obesity or who were overweight. Fibre-specific changes in their microbiomes were linked to changes in their plasma proteomes indicative of an altered physiological state.

It was previously observed that when transplanting faecal microbiomes from twin pairs—of whom one was obese and the other was not—into germ-free mice, increased adiposity developed in the recipients of microbiomes from the twin members who were obese[10]. Cohousing mice shortly after they received these microbiomes prevented the cage mate who received the transplant from the twin with obesity from developing obesity and associated metabolic abnormalities; protection was associated with the invasion of members of Bacteroidales (notably several *Bacteroides*) from the transplanted gut community of the twin who was not obese into the gut community transplanted from the twin with obesity. Invasion did not occur when mice that had received these transplants were cohoused and consumed a fibre-deficient diet that was high in saturated fats and low in fruits and vegetables (HiSF–LoFV diet)[10]. A follow-up screen of 34 food-grade fibres was conducted in gnotobiotic mice that were fed the HiSF-LoFV diet and harboured a defined consortium of cultured gut bacterial strains from a donor who was not obese. The results identified fibre preparations that selectively increased the abundances of targeted *Bacteroides* that are under-represented in obesity-associated microbial communities[11]. Here we characterize the effects of dietary supplementation with different lead fibres emerging from this screen in gnotobiotic mice fed the HiSF–LoFV diet and colonized with faecal microbiomes from nine adult individuals with obesity. Snack prototypes containing one, two or four lead fibres were subsequently tested in two pilot controlled-diet studies in humans, in which participants consumed a diet that was similar in nutritional composition to the one used in the mouse studies. The effects of the fibre supplements on components of the gut microbiome (including carbohydrate-active enzymes (CAZymes) and various metabolic pathways) and host biology (defined by measuring changes in 1,305 plasma protein biomarkers and mediators of numerous physiological functions) were evaluated using computational approaches for feature selection. The results provide an illustration of translatability

[1]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, MO, USA. [2]Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St Louis, MO, USA. [3]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO, USA. [4]Department of Chemistry, University of California, Davis, CA, USA. [5]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. [6]Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique and Aix-Marseille Université, Marseille, France. [7]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia. [8]Infectious and Inflammatory Disease Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [9]A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. [10]Mondelēz Global LLC, Chicago, IL, USA. [11]Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [12]Department of Psychiatry, Washington University School of Medicine, St Louis, MO, USA. ✉e-mail: jgordon@wustl.edu

from gnotobiotic mice to the human population that these mice are designed to portray, and the intersection between the microbiome and nutrition.

Gnotobiotic mouse studies

Fibres isolated from the endosperm of peas, the vesicular pulp of oranges and the bran of barley each contain a diverse set of glycans. Arabinan and galacturonan are the most abundant glycans in pea fibre. Orange fibre also contains arabinan and galacturonan but, in contrast to pea fibre, is dominated by galacturonan. Barley bran contains mixed-linkage β-glucans[11] with arabinose and xylose represented in arabinoxylans (linear β-1,4-linked xylose with α-1,2- and α-1,3-linked arabinose decorations) (Extended Data Fig. 1a, b, Supplementary Table 1a).

Adult germ-free mice were each colonized with a faecal sample obtained from one of nine 32–41-year-old women with obesity (n = 6–10 mice per donor microbiome) (Supplementary Table 1b). Each mouse in each group was subjected to a diet oscillation protocol in which the base HiSF–LoFV diet was supplemented with 10% (w/w) of one of the three types of fibre (Extended Data Figs. 1c, d, 2a, Supplementary Table 1c, d). The unsupplemented HiSF–LoFV diet was given for 10 days between each period of fibre supplementation. We identified bacterial taxa by sequencing 16S rDNA genes, and other genes in the faecal microbiome by sequencing whole-community DNA (Supplementary Table 1e–i). Our microbiome annotation focused on genes represented in (1) the carbohydrate-active enzymes (CAZy) database[12] with an emphasis on glycoside hydrolases and polysaccharide lyases (Supplementary Table 2) and (2) metabolic pathways (microbial community (mc)SEED subsystems)[13]. We used higher-order singular value decomposition (HOSVD)—a technique from the fields of signal processing and machine-learning that has recently been shown to be useful in analysing multifeature time-series studies of microbial communities[14]—to identify changes specific to each fibre treatment (Extended Data Fig. 2b–e). Our analysis included 269 bacterial taxa (defined as V4-16S rDNA amplicon sequence variants (ASVs)), 318 genes encoding CAZymes and 81 mcSEED metabolic pathways that were identified in faecal samples collected from mice during the diet oscillation.

Extended Data Figures 2f, 3a show glycoside hydrolase and polysaccharide lyase genes with increases in abundances that were statistically significant (q value < 0.1, linear mixed-effects model) and rightmost projecting along tensor component 1 (Supplementary Table 3a–e). Consistent with the composition of pea fibre, these genes encode reported or predicted α-arabinofuranosidases, rhamnogalacturonan lyases, galacturonidases, xylanases and xylosidases, among others. mcSEED metabolic pathways with increases in abundance that were statistically significant (q value < 0.1; linear mixed-effects model) and with projections along tensor component 1 that were at the tails of the distribution (α < 0.1) are shown in Extended Data Fig. 3b. They included genes in mcSEED pathways involved in the use of arabinose and arabino-oligosaccharides (AOS), xylose and xylo-oligosaccharides (XOS), plus galacturonate, glucuronate and rhamnose and their respective oligosaccharides (Extended Data Figs. 3b, 4a, Supplementary Table 3f–j). Reflecting their similarities in polysaccharide composition, orange and pea fibre each produced increases in genes encoding CAZymes involved in the processing of arabinan and galacturonans. Although both fibre preparations contain cellulose, the magnitude of the increase in abundances of some β-glucanases and β-glucosidases (GH5_1, GH5_26, GH48, GH30_6, GH5_38 and GH9) was more pronounced with orange fibre (Extended Data Fig. 2f). As with pea fibre, there were statistically significant increases in the representation of mcSEED pathways involved in use of arabinose and AOS, xylose and XOS, as well as galacturonate, glucuronate and rhamnose and their respective oligosaccharides (Extended Data Fig. 3b). The consumption of barley bran led to a decrease in CAZymes involved in processing arabinan (with the exception of GH5_13, GH43_35 and GH43_12, which may recognize arabinose in arabinoxylans that are prominently represented in barley bran). By contrast, glucanases involved in β-glucan processing increased to a greater degree than with the other two fibre preparations; they include GH5_4 (a multifunctional subfamily with β-glucanase activity specific to xylo-glucan plus endo-β-1,4 xylanase activity) and GH5_46 (an endo-β-1,4 glucanase that rose only with barley bran) (Extended Data Fig. 2f).

Members of *Bacteroides* (*Bacteroides vulgatus*, *Bacteroides thetaiotaomicron*, *Bacteroides xylanisolvens*, *Bacteroides cellulosilyticus*, *Bacteroides nordii* and *Bacteroides ovatus*) plus *Blautia wexlerae* were the major drivers of configurational changes in the microbiota induced by pea fibre, as defined by the magnitude and statistical significance of increases in their relative abundances (q value < 0.1, linear mixed-effects model) and their rightmost projection along tensor component 1. Similar to pea fibre, *B. thetaiotaomicron* and *B. vulgatus* were the major responders to orange fibre (in addition to *B. nordii*, *B. xylanisolvens* and *B. cellulosilyticus*). *Bacteroides vulgatus* and *B. thetaiotaomicron* decreased with barley bran, whereas *Bacteroides uniformis*, *B. ovatus*, *B. xylanisolvens* and *Ruminococcus bicirculans* increased (Extended Data Figs. 4b, 5a, b, Supplementary Table 3k–o). Notably, the degree of interpersonal variation in the community response to all three fibres was less pronounced at the level of CAZymes and mcSEED metabolic pathways than in ASV feature space (P = 0.002 to <0.0001; Dunn's Kruskal–Wallis test of Bray–Curtis dissimilarity distances) (Extended Data Figs. 3a, b, 5b, Supplementary Table 1i).

Microbiome effects of fibre snacks in humans

To test the translatability of results obtained in gnotobiotic mice, we first performed an open-label, single-group-assignment, controlled-diet study that involved 13 participants (12 of whom completed the study) with obesity or who were overweight (body mass index of 27 kg m^{-2} to 34 kg m^{-2}), and a snack-food prototype containing pea fibre (Supplementary Table 4a). Participants (Supplementary Table 5a) followed a 45-day regimen (Fig. 1a) in which their normal diet was replaced with the equivalent of the HiSF–LoFV diet, in the form of specifically designed, pre-prepared meals (Supplementary Table 4b). One week after switching to this diet, participants began consuming a single 35-g snack containing 8.1 g of extruded pea fibre: a dose escalation followed, such that 7 days after beginning supplementation, 3 snacks per day were being consumed (1 per meal). This dose was continued for 21 days, after which time supplementation with snacks ceased. No adjustments in the amount of diet consumed were needed to maintain constant weight during or after the period of snack supplementation (Supplementary Table 5b, c). We then performed HOSVD of the representation of CAZymes, mcSEED metabolic pathways and ASVs (Supplementary Tables 6a, b, 7a–l).

Figure 1c and Extended Data Fig. 6a show glycoside hydrolases and polysaccharide lyases with responses that were conserved in study participants and in gnotobiotic mice fed pea fibre; they include statistically significant increases in the abundances of genes encoding GH43-subfamily arabinofuranosidases (GH43_2 and GH43_19), the β-glucanase–β-xylanase GH5_4, pectin and pectate lyases (PL1 and PL9), and the rhamnogalacturonan lyase PL11 (q value < 0.1, linear mixed-effects model). Figure 1d and Extended Data Fig. 6b document the mcSEED pathway responses that are shared between mice and humans, including an increase in the representation of the arabinose and AOS utilization pathway. Cross-correlation of faecal levels of monosaccharides and glycosidic linkages with discriminatory CAZyme gene abundances confirmed that the microbiomes of the study participants contain CAZymes that recognize and cleave 1,2-arabinofuranose and 1,3-arabinofuranose linkages found in the branches of pea fibre arabinan[11,15], leaving the backbone (1,5-arabinofuranose) to accumulate (Extended Data Fig. 7, Supplementary Table 8a–c, Supplementary Results).

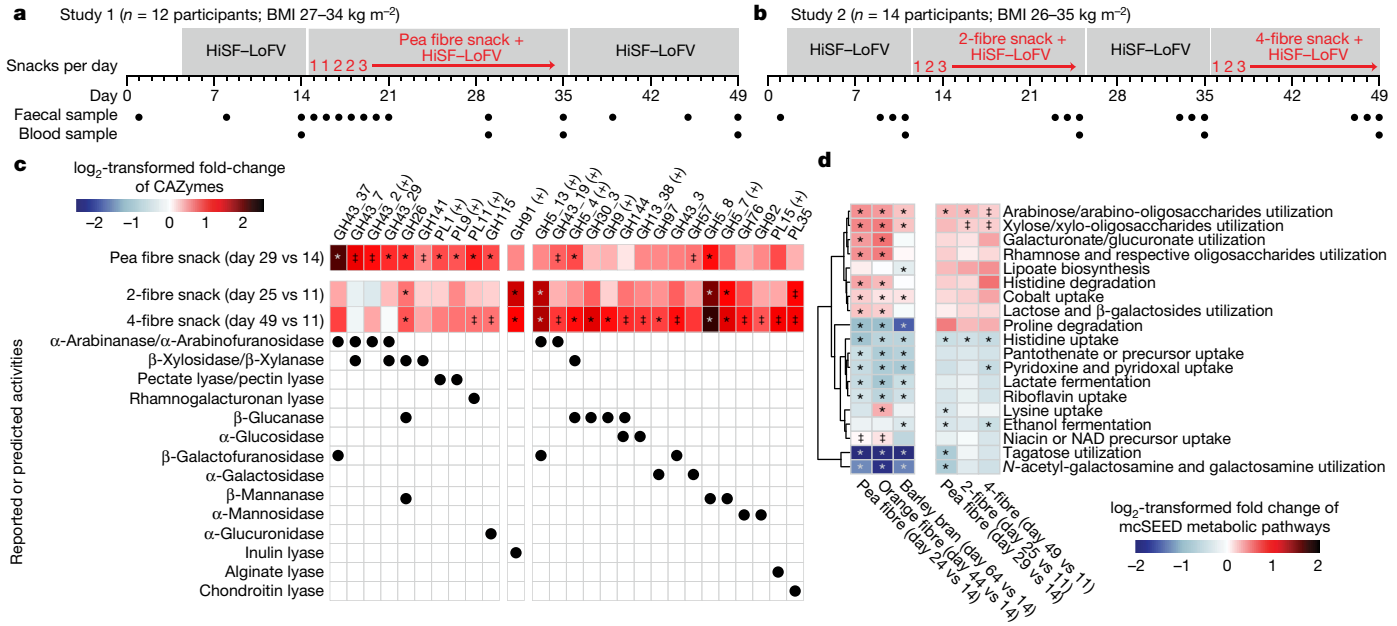


Fig. 1 | Controlled-diet study of the effects of fibre snack prototypes on the faecal microbiomes of individuals with obesity or who are overweight. **a**, **b**, Study designs. **c**, Heat map plotting discriminatory CAZymes for all three fibre snack treatments with a $\log_2$-transformed fold change in abundance that was statistically significant in the faecal microbiomes of participants during consumption of either of the fibre snacks relative to last day of the pretreatment phase. Data are averaged for all participants during each dietary intervention period ($n = 12$ and 14 participants for study 1 and 2, respectively; $n = 66$ faecal samples analysed). CAZymes marked with a + were also fibre-treatment discriminatory in the gnotobiotic mouse studies. **d**, Heat map of discriminatory mcSEED pathways represented in the microbiomes of gnotobiotic mice and human participants with changes in abundance that were statistically significant during at least one of the dietary fibre interventions. The heat map depicted on the left shows the grand mean of the $\log_2$-transformed fold change in abundances of mcSEED pathways in mice containing the gut communities of nine different human donors with obesity ($n = 6$ to 10 mice per donor microbiome; $n = 232$ faecal samples analysed). The heat map on the right shows the mean $\log_2$-transformed fold change in the abundances of mcSEED pathways in participants consuming the pea-fibre, two-fibre or four-fibre snacks ($n = 12$ and 14 for studies 1 and 2, respectively; $n = 66$ faecal samples analysed). The order of mcSEED pathways from top to bottom is based on hierarchical clustering (Euclidean distances). ‡q value < 0.1, *q value < 0.05 (linear mixed-effects model, false-discovery-rate-corrected).

Fourteen adult individuals with obesity or who were overweight (Supplementary Table 5a) completed a second open-label, single-group-assignment, controlled-diet study that tested two multifibre snack prototypes: one contained pea fibre and inulin (a β-2,1-linked fructose polymer that produced pronounced increases in *Bacteroides caccae* in a previous fibre screening experiment[11]) (10.1 g fibre per 30-g snack; 64% pea fibre and 36% inulin) and the other contained pea fibre, inulin, orange fibre and barley bran (10.5 g fibre per 30-g snack; 33% pea fibre, 36% inulin, 11% orange fibre and 20% barley bran) (Supplementary Table 4a). Nine individuals participated in both studies. The interval between the two studies for these individuals ranged from 63 to 96 days; their body weights were not significantly different between each study (−0.6 ± 3.3 kg; mean ± s.d.) (Supplementary Table 5a). All 14 participants followed a 48-day controlled-diet protocol in which their normal diet was replaced with the HiSF–LoFV diet (Fig. 1b). They consumed the two-fibre snack for 14 days (3 servings per day), returned to the unsupplemented HiSF–LoFV diet for 10 days, and subsequently switched to the four-fibre snack (3 servings per day) for 14 days. No caloric adjustments had to be made in the amount of HiSF–LoFV diet consumed to maintain a constant body weight in all participants (Supplementary Table 5b, c).

We conducted HOSVD of changes in the representation of CAZymes, mcSEED pathways and ASVs; data were normalized to the last day of the pre-intervention phase (day 11) (Supplementary Tables 6c, d, 7m–x). Figure 1c shows glycoside hydrolase and polysaccharide lyase genes with increases in abundances that were statistically significant (q value < 0.1, linear mixed-effects model) and that were the rightmost (90th percentile ($\alpha < 0.1$)) ranked feature projections along tensor component 1 (Supplementary Table 7y); 6 and 19 CAZymes satisfied these criteria for the two- and four-fibre snacks, respectively. An inulin lyase (GH91) was among the CAZymes that exhibited the greatest increases in participants consuming the two-fibre snack. Eight of the 19 CAZymes that exhibited significant responses in humans with the four-fibre snack also exhibited statistically significant increases in mice (Fig. 1c). Among these 19 CAZymes, the responses of 9 were uniquely associated with the four-fibre blend: GH30_3 (activity shown for fungal β-1,6-glucan[16]), GH9 (various β-glucan substrates, including xyloglucan[17] and mixed linkage β-1,3-1,4-glucans such as barley β-glucan[18]), GH144 (β-1,2-glucans[19]), GH13_38 (α-glucosidase), GH97 (various activities previously reported, including as an α-glucosidase[20] and α-galactosidase[21]), GH43_3 (activity previously reported with a synthetic 4-nitrophenyl-β-D-galactofuranoside substrate[22], but also probably on α-L-arabinofuranosides as previously documented for GH43_34[23]), GH76 (α-1,6-mannanase), GH92 (α-mannosidase) and PL15 (alginate lyase activity reported[24], although other functions cannot be excluded). The microbiome response to the four-fibre snack is therefore consistent, at least in part, with its inclusion of barley bran and orange fibre.

The two- and four-fibre snacks both produced statistically significant increases in the representation of mcSEED pathways involved in utilization of arabinose and AOS, and xylose and XOS (changes that we also documented in mice) (Fig. 1d, Supplementary Table 7z), as well as statistically significant increases in several *Bacteroides* (Extended Data Fig. 5c, Supplementary Table 7aa). In both of the human studies (as in the mouse study), there was a substantially greater degree of interpersonal variation in the pattern of change in the abundances of

ASVs compared to genes encoding CAZymes and the mcSEED pathways ($P < 0.0001$, Dunn's Kruskal–Wallis test of Bray–Curtis dissimilarity distances) (Extended Data Fig. 6a, b, Supplementary Table 6e, f). Together, these results provided evidence that the four-fibre blend produced the largest number of statistically significant CAZyme responses; they also raised the question of whether the different treatments were associated with qualitatively and quantitatively distinct host responses.

Plasma proteome responses

We used the analytic approach described in Extended Data Fig. 8 to relate CAZyme responses to each type of fibre snack with the plasma proteomic responses of each participant. In the first step, a cross-correlation matrix was created for each snack in which columns were measured plasma proteins (n = 1,170 and 1,205 for studies 1 and 2, respectively) (Supplementary Table 9a–c) representing biomarkers and regulators of a range of physiological functions, and rows were the identified fibre-responsive CAZymes. Each element of the matrix contained the Spearman correlation between plasma protein levels (Supplementary Table 10a–c) and CAZyme gene abundances over time for all participants. In the second step, we performed singular value decomposition on this matrix to identify the plasma proteins that were most highly correlated with fibre-responsive CAZymes. We focused on the first singular vector because it explained the highest percentage of the cross-correlation variance for responses to each of the snacks. Using a histogram of protein projections along the first singular vector, we selected those proteins that were located within the 10th and 90th percentiles of the distribution ($\alpha < 0.1$) (Extended Data Fig. 8). In the third step, we analysed the proteins in the 10th and the 90th percentile group separately using a knowledge generation algorithm (comprehensive multi-omics platform for biological interpretation (CompBio)). CompBio performs an automated extraction of knowledge from all PubMed abstracts that reference entities of interest using contextual language processing and a biological language dictionary that is not restricted to fixed pathway and ontology knowledge bases. CompBio uses conditional probability analysis to compute the statistical enrichment of contextually associated biological 'concepts' (for example, pathways) over those that occur by random sampling. Related concepts are built from the list of differentially represented proteins, and are further clustered into higher-level 'themes' (for example, biological processes). The statistical enrichment of concepts (and, subsequently, themes) is determined using a normalized enrichment score (Methods). The fourth step identified strong correlations (absolute Spearman rho >0.45) between singular value decomposition projections (first singular vector) of proteomic themes and CAZymes responsive to a given fibre treatment.

Supplementary Table 10d–f lists themes and their enrichment scores for each of the different fibre-snack formulations and provides a breakdown of all proteins and concepts in each theme for each treatment. These themes encompass a broad swath of processes related to cell signalling, energy homeostasis and metabolism, cell growth, development and differentiation, immune processes, extracellular matrix, collagen–bone biology, vascular biology, coagulation and thrombosis, and neuronal function. Extended Data Figure 9 shows examples of themes that were statistically significantly enriched for plasma proteins with abundances that strongly correlated with discriminatory CAZymes during consumption of the four-fibre snack as well as the specificity of these thematic responses compared to the other snack prototypes.

Extended Data Figure 10 plots the correlation values between (1) the 19 CAZymes that satisfied our threshold criteria for statistically significant change in their abundances across all participants consuming the four-fibre snack and (2) biological themes that were statistically significantly enriched for proteins in the 10th and 90th percentiles. Among the 9 CAZymes with significant changes in abundance that were unique to the four-fibre snack, two had Spearman rho values |≥0.45| with themes; the abundance of GH43_3 across study participants was correlated with changes in themes related to glucose metabolism, calcineurin signalling, AKT signalling, heat shock proteins and chaperones, and BCL-2 related apoptosis, whereas changes in GH13_38 were correlated with a theme related to kallikrein–kinin proteases (Supplementary Table 10i). Four CAZymes exhibited statistically significant increases in their abundances after exposure to both the pea fibre and four-fibre snacks, including GH5_4 (encompasses xyloglucan-specific glucanase and xylanase activities); with the exception of GH115, they were also statistically significantly increased with pea, orange and barley bran fibre consumption in gnotobiotic mice. Changes in GH5_4 in participants consuming the four-fibre or pea-fibre snacks were strongly correlated (Spearman rho ≥ |0.45|) with changes in proteomic themes related to immune processes (C-type lectin receptor signalling, neutrophil proteases, and antigen presentation–co-stimulation), plus IGF–IGFBP signalling, FGF signalling in synostosis and ErbB signalling pathways (Extended Data Fig. 10, Supplementary Table 10g–i).

This analysis was also able to identify subgroups of individuals on the basis of their CAZyme and proteomic responses to a given treatment type (Extended Data Fig. 11a–c). For example, participants in the four-fibre study can be subdivided into 2 groups on the basis of changes in 23 themes (including glucose metabolism), derived from proteins in the 10th percentile. The grouping of participants on the basis of these themes could not be attributed to differences in their consumption of the four-fibre snack prototype as judged by mass spectrometry-based measurements of a faecal biomarker of this orange-fibre-containing formulation (Extended Data Fig. 11d–f, Supplementary Table 8d–f, Supplementary Results).

Finally, we expanded our analysis to include plasma proteins that did not necessarily have statistically significant correlations with treatment-associated CAZymes. Extended Data Figure 12 and Supplementary Table 9d–f describe proteins with $\log_2$-transformed fold abundances that were significantly altered ($P < 0.05$, limma[25]) after treatment with the one-, two- or four-fibre snacks compared to the initial phase of unsupplemented HiSF–LoFV diet consumption. For example, non-receptor tyrosine-protein kinase (TYK2), which is a central transducer of signals elicited by several cytokines that control diverse cellular processes, increased with all three snack formulations. Mice that lack TYK2 develop obesity[26]. The chemokine CXCL1 was significantly increased after treatment with the four-fibre snack. In mice, increases in muscle and serum levels of CXCL1 attenuate diet-induced accumulation of fat, with associated upregulation of genes involved in fatty acid oxidation[27]. Neurexophilin 1 (NXPH1), which is a neuronal glycoprotein that binds to α-neurexin and regulates numerous neuronal processes, was also increased after treatment with the four-fibre snack; single-nucleotide polymorphisms in *NXPH1* have previously been associated with type 2 diabetes, blood pressure levels, and levels of cholesterol, triglyceride and C-reactive protein[28]. The four-fibre snack produced statistically significant reductions of fibroblast growth factor 2 (FGF2). Expression of this adipokine is increased in mice with obesity induced by a high-fat diet (in which case it exacerbates an obesogenic adipocyte inflammatory response through activation of the NLRP3 inflammasome)[29]. Together, our data indicate that the different snack prototypes elicit shared, as well as fibre-specific, responses that encompass plasma biomarkers and mediators of diverse physiological processes.

Discussion

Correlation of discriminatory CAZymes in the gut microbiomes of participants in our controlled-diet studies with discriminatory features of their plasma proteomes provided a way to operationally define responses to the different fibre snack prototypes, as well as a means for assessing the translatability of results from gnotobiotic mice to humans. Identifying treatment-discriminatory CAZymes can also

provide insights into the bioactive components of fibre preparations. However, much remains to be determined (through direct biochemical tests) about the substrate specificities of CAZymes, including those identified as 'fibre-responsive' in the current study. Moreover, some of the CAZyme associations we identify could be due to changes in the gut microbiome that are not directly related to breakdown of glycans in the snacks we tested (the so-called 'butterfly effect'[30]). This work needs to be extended to include analyses of the expression of various microbiome-encoded metabolic pathways, including—but not limited to—those involving CAZymes, and how their levels of expression correlate with plasma proteomic responses. The feature selection achieved with our analytic pipeline yielded a set of biomarkers that could be used to design and interpret necessary follow-on randomized controlled, proof-of-concept studies of lead fibre snack prototypes. A goal of these studies is to provide a rigorous scientific foundation for claims about how such products may affect different aspects of the physiology of consumers, who is most likely to benefit and the required dose. Finally, additional tests are needed of the translatability of findings from gnotobiotic mice to humans: specifically, tests of causality between changes in microbiome features and host biological responses, and the mechanisms that connect changes in the microbiome to these host responses. Investments in efforts such as these could provide one way to accelerate identification and testing of new, more nutritious food formulations at a time when many factors are operating to create large disparities in nutritional status among members of many societies.

Online content

Any methods, additional references, Nature Research 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/s41586-021-03671-4.

1. NCD Risk Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *Lancet* **387**, 1377–1396 (2016).
2. GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **393**, 1958–1972 (2019).
3. Willett, W. et al. Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. *Lancet* **393**, 447–492 (2019).
4. Hauner, H. et al. Evidence-based guideline of the German Nutrition Society: carbohydrate intake and prevention of nutrition-related diseases. *Ann. Nutr. Metab.* **60** (Suppl 1), 1–58 (2012).
5. Reynolds, A. et al. Carbohydrate quality and human health: a series of systematic reviews and meta-analyses. *Lancet* **393**, 434–445 (2019).
6. Zhao, L. et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **359**, 1151–1156 (2018).
7. Asnicar, F. et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).
8. Kovatcheva-Datchary, P. et al. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. *Cell Metab.* **22**, 971–982 (2015).
9. Sonnenburg, E. D. et al. Specificity of polysaccharide use in intestinal *Bacteroides* species determines diet-induced microbiota alterations. *Cell* **141**, 1241–1252 (2010).
10. Ridaura, V. K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
11. Patnode, M. L. et al. Interspecies competition impacts targeted manipulation of human gut bacteria by fiber-derived glycans. *Cell* **179**, 59–73.e13 (2019).
12. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
13. Overbeek, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).
14. Martino, C. et al. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39**, 165–168 (2021).
15. Wesener, D. A. et al. Microbiota functional activity biosensors for characterizing nutrient metabolism in vivo. *eLife* **10**, e64478 (2021).
16. Temple, M. J. et al. A *Bacteroidetes* locus dedicated to fungal 1,6-β-glucan degradation: Unique substrate conformation drives specificity of the key endo-1,6-β-glucanase. *J. Biol. Chem.* **292**, 10639–10650 (2017).
17. Larsbrink, J. et al. A discrete genetic locus confers xyloglucan metabolism in select human gut *Bacteroidetes*. *Nature* **506**, 498–502 (2014).
18. Schröder, C. et al. Characterization of a theme C glycoside hydrolase family 9 endo-beta-glucanase from a biogas reactor metagenome. *Protein J.* **37**, 454–460 (2018).
19. Shimizu, H. et al. Characterization and structural analysis of a novel exo-type enzyme acting on β-1,2-glucooligosaccharides from *Parabacteroides distasonis*. *Biochemistry* **57**, 3849–3860 (2018).
20. Li, W. et al. PspAG97A: a halophilic α-glucoside hydrolase with wide substrate specificity from glycoside hydrolase family 97. *J. Microbiol. Biotechnol.* **26**, 1933–1942 (2016).
21. Gloster, T. M., Turkenburg, J. P., Potts, J. R., Henrissat, B. & Davies, G. J. Divergence of catalytic mechanism within a glycosidase family provides insight into evolution of carbohydrate metabolism by human gut flora. *Chem. Biol.* **15**, 1058–1067 (2008).
22. Helbert, W. et al. Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. *Proc. Natl Acad. Sci. USA* **116**, 6063–6068 (2019).
23. Ndeh, D. et al. Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature* **544**, 65–70 (2017).
24. Hashimoto, W., Miyake, O., Ochiai, A. & Murata, K. Molecular identification of *Sphingomonas* sp. A1 alginate lyase (A1-IV') as a member of novel polysaccharide lyase family 15 and implications in alginate lyase evolution. *J. Biosci. Bioeng.* **99**, 48–54 (2005).
25. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
26. Derecka, M. et al. Tyk2 and Stat3 regulate brown adipose tissue differentiation and obesity. *Cell Metab.* **16**, 814–824 (2012).
27. Pedersen, L., Olsen, C. H., Pedersen, B. K. & Hojman, P. Muscle-derived expression of the chemokine CXCL1 attenuates diet-induced obesity and improves fatty acid oxidation in the muscle. *Am. J. Physiol. Endocrinol. Metab.* **302**, E831–E840 (2012).
28. Kraja, A. T. et al. Genetic analysis of 16 NMR-lipoprotein fractions in humans, the GOLDN study. *Lipids* **48**, 155–165 (2013).
29. ZhuGe, D. L., Javaid, H. M. A., Sahar, N. E., Zhao, Y. Z. & Huh, J. Y. Fibroblast growth factor 2 exacerbates inflammation in adipocytes through NLRP3 inflammasome activation. *Arch. Pharm. Res.* **43**, 1311–1324 (2020).
30. Michalak, L. et al. Microbiota-directed fibre activates both targeted and secondary metabolic shifts in the distal gut. *Nat. Commun.* **11**, 5773 (2020).

Methods

Blinding, randomization and power calculations

Before colonization with human faecal microbial communities, germ-free mice were randomized with respect to litter of origin but were matched by starting weight across treatment groups. The diet oscillation protocol allowed us to use each mouse as its own control. The number of mice studied per treatment group was based on our previous knowledge of the reproducibility of colonization of recipient germ-free mice with a given human donor microbiome and ref. [11], which characterized the effect size of the selected fibres on a defined consortium of cultured human gut bacterial strains, including members of *Bacteroides*. All data were generated from mouse biospecimens without knowledge of their treatment group of origin.

Human studies were single-group, open label studies and were therefore not blinded or randomized. As there was little precedent for controlled-diet studies of this type (examining the effects of single- and multi-fibre-supplemented snack prototypes on gut microbiome composition and/or host physiology), a formal power calculation was not possible. However, as each subject served as his or her own control with repeated microbiome sampling, we aimed to have a minimum of 12 participants complete each of these studies. Primary generation of data from human biospecimens was performed by individuals who were not involved in study design or interpretation of the results, without knowledge of the stage in the study from which the biospecimens were collected.

Gnotobiotic mouse studies

All mouse experiments were carried out using protocols approved by the Institutional Animal Care and Use Committee (IACUC) of Washington University in St Louis.

Diets. The HiSF–LoFV diet was milled to powder (D90 particle size, 980 μm), and mixed with powdered fibre preparations (10% (w/w)). Fibre content was defined for each preparation (Association of Official Agricultural Chemists (AOAC) 2009.01). Similarly, protein, fat, total carbohydrate, ash and water content were measured (protein AOAC 920.123; fat AOAC 933.05; ash AOAC 935.42; moisture AOAC 926.08; total carbohydrate (100 − (protein + fat + ash + moisture)). The powdered food mixtures were vacuum-packed in sterile plastic containers and sterilized by gamma irradiation (20–50 kilograys, Steris). Sterility was confirmed by culturing the diets under aerobic and anaerobic conditions (atmosphere, 75% N_2, 20% CO_2, 5% H_2) at 37 °C in TYG medium.

Diet oscillation studies. A 500-mg aliquot of pulverized frozen faecal sample, obtained from nine unrelated adult women with obesity in the Missouri Adolescent Female Twin Study (MOAFTS) cohort[31] (listed in Supplementary Table 1b), was diluted in 5 ml of reduced PBS (1× PBS supplemented with 0.1% resazurin (w/v), 0.05% L-cysteine–HCl) in an anaerobic Coy chamber (atmosphere, 75% N_2, 20% CO_2, 5% H_2). The sample was vortexed for 2 min at room temperature in 5 ml of 2-mm diameter autoclaved borosilicate glass beads to disrupt clumps of bacterial cells trapped within the faecal matrix. The resulting suspension was filtered through a sterile nylon mesh cell strainer (100-μm pore diameter; BD Falcon). The filtrate was then mixed with 5 ml of sterile PBS containing 0.1% resazurin (w/v), 0.05% L-cysteine–HCl and 30% (v/v) glycerol, transferred to a sterile glass crimped tube and stored at −80 °C until further use. Aliquots of the stored filtrate were transported in a frozen state to the gnotobiotic mouse facility. The outer surface of the tube was sterilized by a 30-min exposure to chlorine dioxide in the transfer sleeve attached to the gnotobiotic isolator. The tube was then introduced into the isolator.

To test the effects of different fibre preparations on uncultured human faecal microbial communities, germ-free male C57BL/6J mice were dually housed in plastic cages located in plastic flexible film gnotobiotic isolators (Class Biologically Clean). The gnotobiotic facility was maintained at 23 °C and 40% humidity under a strict 12-h light cycle (lights on at 06:00 h). Cages contained autoclaved paper 'shepherd shacks' to facilitate the natural nesting behaviours of mice and to provide environmental enrichment.

Germ-free mice were weaned onto and subsequently maintained on an autoclaved, low-fat, high-plant polysaccharide chow (Envigo, catalogue number 2018S) that was administered ad libitum. Four days before colonization, mice were switched to a diet low in saturated fats and high in fruits and vegetables (LoSF–HiFV) that was formulated on the basis of the National Health and Nutrition Examination Survey (NHANES) of US dietary practices[10]. A 300-μl aliquot of a clarified suspension of a given faecal sample was introduced into the stomachs of 12–16-week-old male mice using an oral gavage needle. Recipients were maintained in separate gnotobiotic isolators dedicated to mice colonized with the microbial community of the same human donor. Four days after gavage, mice were switched to the HiSF–LoFV diet[10].

Mice in the experimental groups completed a 64-day multiphase diet-oscillation feeding protocol. On day 4 after gavage of the human donor faecal community, mice were fed a pelleted version of the HiSF–LoFV diet for 10 days. Beginning on experimental day 14, mice were fed 20–30-g aliquots of a mixture that was made from the milled HiSF–LoFV diet supplemented with 10% (w/w) raw pea fibre (pea fibre EF 100; J. Rettenmaier & Söhne) and hydrated with 10–15 ml sterile water (mixing of the sterile powdered diet and sterile water occurred within the gnotobiotic isolator). The resulting dough-like mixture was pressed into a plastic feeding dish and placed on the cage floor for feeding ad libitum. Food supply was monitored daily. A freshly hydrated aliquot of the diet was supplied every 3 days to prevent food levels from dropping below roughly one third of the original volume. The HiSF–LoFV and pea fibre mixture was administered for 10 days, after which time mice were returned to the unsupplemented pelleted HiSF–LoFV diet for 10 days ('wash out period'). On day 34, mice were switched to a hydrated mixture made from the milled HiSF–LoFV diet supplemented with 10% (w/w) coarse orange fibre (CitriFi 100; Fibre Star). This mixture was administered for 10 days. Mice were then returned to the unsupplemented pelleted HiSF–LoFV diet for 10 days. On experimental day 54, mice began receiving 20–30-g aliquots of a hydrated diet and fibre mixture made from the powdered HiSF–LoFV diet supplemented with 10% (w/w) raw barley bran fibre (Barley Balance - concentrated (1–3) (1–4) β-glucan; PolyCell Technologies). The HiSF–LoFV and barley bran fibre mixture was administered for 10 days. All mice were euthanized without prior fasting on experimental day 64.

Bedding (Aspen Woodchips; Northeastern Products) was replaced after each 10-day diet oscillation period to prevent any leftover food or faecal matter from carrying over and being ingested during the subsequent diet phase. Fresh faecal samples were collected from each mouse into sterile cryo-resistant polypropylene tubes within seconds of being produced on day 4 after initial colonization while consuming the LoSF-HiFV diet, and on days 5 and 10 of each 10-day oscillation period. Samples were placed in liquid nitrogen 45–60 min after they were collected. Pre-colonization faecal samples were also collected to verify the germ-free status of mice (by culture and by bacterial V4-16S rDNA amplicon sequencing).

Human studies

Individuals provided written informed consent before participating in these studies, which were approved by the IRB at Washington University School of Medicine in St Louis. The first study (ClinicalTrials.gov NCT04159259) was performed between February and July 2019. The second study (ClinicalTrials.gov NCT04101344) was conducted between August and December 2019. The primary objective of these open-label, single-group-assignment studies was to test the effects of different fibre-supplemented snack prototypes on the gut microbiomes of participants who were overweight or obese consuming a

controlled background diet containing quantities of saturated fats in the upper tertile, and quantities of fruits and vegetables in the lower tertile (HiSF–LoFV) of consumption in the NHANES database[10]. Both studies were conducted at the Clinical Translational Research Unit at Washington University School of Medicine in St Louis.

Study 1 design. A total of 19 men and women with obesity or who were overweight (body mass index ≥ 25.0 and ≤ 35.0 kg m^{-2}), aged ≥18 and ≤60 years, were screened for potential participation in this study. Participants completed a comprehensive medical evaluation, including a medical history, physical examination, assessment of food preferences and aversions, and standard blood tests. A faecal sample was collected during the medical evaluation phase and used to determine whether *Bacteroides* species were present in their microbiota (*B. vulgatus*, *B. thetaiotaomicron*, *B. cellulosilyticus*, *B. uniformis* and/or *B. ovatus*); those whose faecal microbiota contained less than 0.1% relative abundance of *B. vulgatus* (defined by V4 16S rDNA amplicon sequencing), and less than 0.1% relative abundance of at least one of the other *Bacteroides* were excluded from the study. Additional exclusion criteria included: (1) history of previous bariatric surgery; (2) substantial organ system dysfunction (for example, diabetes, severe pulmonary, kidney, liver or cardiovascular disease); (3) history of inflammatory bowel disease; (4) pregnant or lactating; (5) use of medications known to affect the study outcome measures that could not be temporarily discontinued; (6) use during the month before screening of medications known to affect the composition of the gut microbiota (for example, antibiotics); (7) bowel movements <3 times per week; (8) vegans, vegetarians and those with lactose intolerance and/or severe allergies, aversions or sensitivities to foods and ingredients included in the prescribed meal plan; and (9) individuals who were not able to grant voluntary informed consent. Of 19 participants who were screened, 4 were excluded on the basis of the *Bacteroides* criterion and two were excluded on the basis of the screening assessment. Of the 13 individuals who were enrolled, 12 completed the study per protocol and 1 participant was excluded due to substantial non-compliance (Supplementary Table 5a). Side effects potentially related to the intervention were mild or moderate, and included reported transient 'abdominal bloating' (n = 3 participants), 'gastrointestinal cramps or upset stomach' (n = 3) and/or 'diarrhoea' (n = 2).

The study design is described in Fig. 1a. Participants who met the enrolment criteria were asked to maintain their usual eating habits for 4 days, with faecal samples collected at home on days 1, 2, 3 and 4. On day 5 through to day 14, participants consumed only HiSF–LoFV meals provided by the study team (described in 'Design, manufacture and distribution of HiSF–LoFV diets'). Faecal samples were collected at home on study days 6, 8, 10, 12 and 14 and a fasting blood sample was obtained on day 14. Starting on day 15, participants supplemented their HiSF–LoFV diet with the pea-fibre-containing snack, beginning with 1 bar per day on days 15 and 16 (with lunch), 2 bars per day on days 17 and 18 (one at lunch and one at dinner) and 3 bars per day from day 19 through day 35 (with breakfast, lunch and dinner). All at-home faecal samples during the ramp-up period (that is, days 15–20) were collected and, subsequently, every two days thereafter (that is, on study days 21, 23, 25, 27, 29, 31, 33 and 35), with fasting blood collected on days 29 and 35. After completing 17 consecutive days consuming 3 snacks per day (24.3 g of pea fibre), participants returned to the unsupplemented HiSF–LoFV diet for an additional 14 days, with faecal samples collected every two days (study days 37, 39, 41, 43, 45, 47 and 49) and a fasting blood sample on day 49. The primary outcome measures determined before and during fibre snack consumption (day 14 versus days 29–35) were; (1) changes in the plasma lipid profile, (2) changes to the gut microbiome, (3) changes in the plasma proteome, and (4) changes in HbA1c. Urine was collected but was not analysed.

Study 2 design. A total of 23 men and women with obesity or who were overweight (BMI ≥ 25.0 and ≤ 35.0 kg m^{-2}) and aged ≥18 and ≤60 years were screened for eligibility using the same exclusion criteria as study 1, with the exception that there was no pre-screen for the representation of *Bacteroides* species in faecal samples. Among the 23 individuals who completed screening, 19 were enrolled and 14 completed the study protocol. Four participants did not complete the intervention for personal reasons; one participant was withdrawn for failure to comply (Supplementary Table 5a). Potential intervention-related side effects were mild and included transient abdominal bloating (n = 1 participant), gastrointestinal cramping during dose escalation (n = 1) and 'constipation' (n = 1).

The study design is described in Fig. 1b. On day 1, participants provided a faecal sample before entering the controlled-diet phase. From day 2 to day11, participants consumed HiSF–LoFV meals provided by the study team in the form of packed-out meals and snacks; they collected at-home faecal samples on study days 5, 9,10 and 11. A fasting blood sample was obtained on day 11. On day 12, participants began supplementing their diet with the two-fibre snack prototype (1 snack serving per day on day 12 (with lunch), 2 per day on day 13 (one at lunch and one at dinner) and 3 per day on days 14 through 25 (with breakfast, lunch and dinner)). Participants collected all at-home faecal samples during the ramp-up period (that is, days 12–14) and, subsequently, at-home faecal samples on days 18, 23, 24 and 25. A fasting blood sample was collected on day 25 immediately before returning to the unsupplemented HiSF–LoFV diet (days 26–35). During this wash-out phase of the study, faecal samples were collected on days 28, 33, 34 and 35 and a fasting blood sample on day 35. Consumption of the four-fibre snack prototype began on day 36 (1 snack serving per day on day 36 (with lunch), 2 snacks per day on day 37 (one at lunch and one at dinner) and 3 snacks per day on days 38 through 49 (at breakfast, lunch and dinner)). Participants collected all at-home faecal samples during the ramp-up period (days 36–38) and, subsequently, at-home faecal samples on days 41, 47, 48 and 49. A fasting blood sample was obtained on day 49 (the last day of the study). The primary outcome measure was the effect of the fibre blend snacks on the gut microbiome as determined by analyses of faecal samples collected before supplementation (day 11) and end of each supplementation phase (day 25 and day 49, respectively). The secondary outcome was the effect of the fibre blends on the blood (plasma) proteome. Urine was collected but was not analysed.

Fibre snacks. Snack food prototypes were prepared by Mondelēz Global LLC, tested to confirm nutritional content as well as the absence of microbial contamination or pathogens, and then shipped to and stored at Washington University in St Louis. Participants received weekly shipments of the snack food prototypes. The composition of these prototypes is described in Supplementary Table 4a. Their organoleptic properties were designed on the basis of common US consumer preferences.

Design, manufacture and distribution of HiSF–LoFV diets. Participants consumed a diet composed of approximately 40% fat, 20% protein and 40% carbohydrate during the dietary intervention. The HiSF–LoFV diet is high in refined grains (white bread and pasta, bagels and corn cereals), added sugars (sugar-sweetened beverages, candies and desserts), vegetables sourced primarily from potatoes and tomatoes, and protein and fat derived from animals. Representative diets are described in Supplementary Table 4b.

Each participant's estimated energy requirements were calculated using the Mifflin St. Jeor equation[32] multiplied by an appropriate physical activity level. To ensure consistent intake of nutrients across all participants and to ensure weight stability, a registered dietitian designed a seven-day cycle menu specific to the participant's energy needs and instructed each participant to consume only foods prescribed by the study team during the dietary intervention. All food was provided in the form of packed-out meals and snacks prepared by the metabolic kitchen in the Clinical Translational Research Unit (CTRU) at Washington University School of Medicine.

Collection of clinical metadata. Participants were provided with electronic smart scales (BodyTrace) to enable weight monitoring between study visits. At enrolment, habitual dietary patterns were assessed using the National Cancer Institute Diet History Questionnaire III (DHQIII) food frequency questionnaire[33]. Participants visited the CTRU on a weekly basis to pick up packed-out meals (using insulated bags and rolling coolers), have their body weight measured, and any changes to their health and medications reviewed. During the study, participants recorded all food and beverage intake using a web-based food diary during all diet phases. An experienced study dietitian instructed study participants on how to complete the food records and reviewed these records with the participants at each study visit to ensure the accuracy of self-reported data. In addition, a member of the study team contacted participants regularly to (1) check on study progress, (2) discuss prescribed and non-prescribed foods and beverages consumed, (3) discuss weight changes and (4) ensure participants had sufficient faecal collection kits.

Preparation of blood samples. Fasting blood samples were obtained in the CTRU. Conventional blood chemistry tests were performed by the Clinical Laboratory Improvement Amendments (CLIA)-certified Core Laboratory for Clinical Studies (CLCS) at Washington University School of Medicine. To prepare plasma for SOMAscan proteomics analysis (SomaLogic), blood samples (10–20 ml) were aliquoted into EDTA-K2 treated tubes and centrifuged at 2,000*g* for 10 min at 4 °C. Following centrifugation, plasma was immediately transferred into cryo-resistant polypropylene tubes (0.5-ml aliquots) and stored at −80 °C before analysis according to manufacturer's recommendations.

Statistical analysis of clinical study outcomes

The effects of supplementation with each of the snack fibre prototypes on BMI and clinical chemistry parameters, including lipids and HbA1c, were assessed using a linear mixed-effects model that included a random-effect coefficient to control for differences in characteristics between individuals (Supplementary Table 5c). Supplementation-associated, statistically significant differences ($P \leq 0.05$) for each analyte were determined using analysis of variance (ANOVA), with false-discovery rate (FDR) correction (least-squares means of linear mixed-effects model). Analyses of fibre-supplementation-associated changes in the microbiome and the plasma proteome using HOSVD, cross-correlation singular value decomposition (CC-SVD) and linear models are described in detail in 'Shotgun sequencing and annotation of microbiomes', 'HOSVD', 'CC-SVD and CompBio' and 'Linear model analysis of SOMAscan data generated from human plasma samples'.

Faecal sample collection, processing and culture-independent analyses

Sample collection and processing. Participants collected faecal samples using small medically approved collection containers. Containers were labelled with a unique study identifier and the collection date and time. Participants were provided with a dedicated −20 °C freezer at the beginning of the study for temporary storage of faecal samples. Each sample was frozen immediately after collection. Every few days, frozen specimens were shipped with frozen gel packs, using prepaid express shipping boxes provided by the study team, to a biospecimen repository at Washington University in St Louis where they were stored at −80 °C until the time of processing.

Faecal samples were homogenized with a porcelain mortar (4 L) and pestle while submerged in liquid nitrogen and aliquots of each pulverized frozen sample were stored at −80 °C. DNA was extracted from each pulverized human faecal sample (about 50–100 mg) (or mouse faecal pellets (about 20–50 mg)) by first bead-beating (BioSpec Mini-beadbeater-96) for 4 min in 500 µl buffer A (200 mM NaCl, 200 mM Trizma base, 20 mM EDTA), 210 µl of 20% SDS, and 500 µl phenol:chloroform:isoamyl alcohol (25:24:1), plus 250 µl of 0.1-mm-diameter zirconium oxide beads and a 3.97-mm-diameter steel ball. Following centrifugation at 3,220*g* for 4 min, DNA was purified (QiaQuick 96 purification kit; Qiagen), eluted in 130 µl of 10 mM Tris-HCl pH 8.5 (buffer EB, Qiagen) and quantified (Quant-iT dsDNA broad range kit; Invitrogen). Purified DNA was stored at −20 °C before further processing.

16S rDNA amplicon sequencing and identification of ASVs. Purified DNA samples were adjusted to a concentration of 1 ng µl^{-1} and subjected to PCR using barcoded primers directed against variable region 4 of the bacterial 16S rRNA gene[34]. PCR amplification was performed using the following cycling conditions: denaturation (94 °C for 2 min), 26 cycles of 94 °C for 15 s, 50 °C for 30 s and 68 °C for 30 s, and then incubation at 68 °C for 2 min. Amplicons with sample-specific barcodes were quantified, pooled and sequenced (Illumina MiSeq instrument, paired-end 250-nt reads).

Paired-end reads were demultiplexed, trimmed to 200 nucleotides, merged and chimaeras removed using the 1.13.0 version of the DADA2 pipeline[35] in R (v.3.6.0)[36]. ASVs generated from DADA2 were aligned against the GreenGenes 2016 (v.13.8) reference database to 97% sequence identity, followed by taxonomic and species assignment with Ribosomal Database Project (RDP) (release 11.5) and SILVA (v.128). The resulting ASV table was filtered to include only ASVs with ≥0.1% relative abundance in at least five samples and rarefied to 15,000 reads per sample.

Supplementary Table 1e also shows results obtained with another approach to taxonomic assignment. In this procedure, each representative sequence was aligned (NCBI BLAST toolkit version 2.10.0) to a 16S rRNA gene reference database compiled by joining unique sequences from RDP version 11.5 and the NCBI 16S ribosomal RNA Project. Alignment results were sorted based on their percentage of sequence identity, with maximum values denoted as *M*. Hits were selected with identities in the range [*M*] to [*M*−(1−*M*)/*S*] in which *S* is a scaling parameter that controls the maximum number of taxonomic descriptors accepted for a multi-taxonomic assignment based on 16S rDNA sequence identity (in this study, set to 4)[37].

Shotgun sequencing and annotation of microbiomes. Purified DNA samples were adjusted to a concentration of 0.75 ng µl^{-1}. Sequencing libraries were generated from each DNA sample using the Nextera DNA Library Prep Kit (Illumina) with the reaction volume scaled down tenfold to 2.5 µl[38]. Samples were pooled and sequenced using an Illumina NextSeq 550 instrument in the case of all mouse samples ($10.7 \pm 0.6 \times 10^6$ paired-end 150-nucleotide-long reads per sample (mean ± s.d.)) and all human samples in study 2 ($12.8 \pm 1.2 \times 10^6$ paired-end 150-nucleotide-long reads per sample), while an Illumina NovaSeq 6000 instrument was used to sequence human samples collected in study 1 ($28.0 \pm 4.2 \times 10^6$ paired-end 150-nucleotide-long reads per sample).

After sequencing, reads were demultiplexed (bcl2fastq, Illumina), adaptor sequences were trimmed using cutadapt[39] and reads were quality-filtered with Sickle[40]. Human and mouse DNA sequences were identified and removed using Bowtie2[41] and either the *Homo sapiens* genome (hg19 build) or the *Mus musculus* C57BL/6J strain genome (UCSC mm10), depending on sample origin, before further processing. Microbial reads were assembled using IDBA-UD[42] (samples from gnotobiotic mice) or MEGAHIT[43] (samples from human studies) and annotated with prokka[44]. Counts for each open reading frame (ORF) were generated by mapping quality-controlled, paired-end reads generated from each sample to the corresponding assembled sample contigs. Duplicate reads (optical- and PCR-generated) were identified and removed from mapped data using the Picard MarkDuplicates tool (v.2.9.3). Alignments were processed to generate count

data (featureCounts; Subread v.1.5.3 package)[45] for each ORF in each sample and normalized (transcripts per kilobase per million reads) in R (v.3.6.0).

Functional profiles for each faecal microbiome were generated by assigning microbiome-encoded proteins to 81 microbial community SEED[13] (mcSEED) metabolic pathways (subsystems) or pathway modules that capture core metabolism of nutrients or metabolites in four major categories (amino acids, sugars, fermentation products and vitamins) projected over about 2,600 reference bacterial genomes. To do so, translated protein sequences from each faecal DNA sample assembly were queried against representative protein sequences from the mcSEED subsystems or pathway modules using DIAMOND[46] (threshold ≥ 80% identity). In the case of multiple high per cent identity alignments, proteins were assigned the annotation of the best hit found in mcSEED.

CAZyme annotations were performed for the full set of ORFs identified in each faecal sample. Amino acid sequences were analysed using a custom script, the purpose of which is to accommodate the modular structure of CAZymes (which often carry a variable number of ancillary modules in addition to their catalytic domain). The script first compares each sequence to the full-length sequences stored in the CAZy database (download date 21 April 2020) using BlastP (version 2.3.0+)[47]. Query sequences with 100% coverage, >50% amino acid sequence identity and *E*-value ≤ 10^{-6} with a sequence already in the CAZy database are assigned to the same family or subfamily (or the same families if the found sequence contains more than one CAZy module). All remaining sequences were subject to a follow-on second similarity search that involved two parallel steps: (1) a BlastP comparison of the sequence of interest against a library composed of the individual (isolated) modules (that is, where the individual modules (catalytic or ancillary) have been 'isolated' as opposed to full-length sequences in the CAZy database that can contain several modules); and (2) a HMMER3[48] search against a curated collection of hidden Markov models based on each of the CAZy module families (and subfamilies where they have been defined, for example, for GH5, GH13, GH16, GH30 and GH43). The script then assigns the sequences to the corresponding family or families (and subfamily or subfamilies) that satisfy the following two conditions: (1) a BlastP *E*-value <10^{-4} and (2) a hmmhitstart ≤ 0.05 or hmmhitend ≥ 0.95 (the hidden Markov model search helps to resolve problems of overlap between modules that are sometimes observed when using only BlastP). For this second similarity-search strategy, CAZyme assignments were made only when the two methods indicated the same subject in the CAZy database with >90% alignment overlap and an *E*-value <10^{-4} for all CAZyme families (except for lytic polysaccharide monooxygenases (LPMOs); threshold 10^{-25}). Carbohydrate esterases and auxiliary activities (aside from LPMOs) were not analysed. The above thresholds were designed to eliminate, as much as possible, assignment of query sequences to oxidoreductases or esterases not specific for carbohydrates and to produce annotation results more consistent with the manual procedure used to add and update proteins contained in the CAZy database.

To generate CAZyme gene family and subfamily abundance tables amenable to downstream statistical analyses, per-sample abundance data were aggregated using a custom script in R. Genes encoding all non-CAZyme ORFs were removed. The abundances of genes annotated with multiple CAZyme families and subfamilies were propagated to each individual family or subfamily member, and then abundances were summed across all corresponding CAZyme families within each faecal sample.

HOSVD

HOSVD is used when an input matrix has more than two degrees of freedom. Mathematically, these types of matrices are called 'tensors.' Unlike singular value decomposition (SVD), HOSVD is not a technique with an analytical solution—that is, a tensor of rank *N* cannot be written as a product of *N*+1 tensors as in the case with SVD. As a consequence, several methods of approximation exist to deconstruct higher-order tensors for feature-reduction purposes. Canonical polyadic decomposition deconstructs a tensor into a sum of rank-1 tensors (arrays) related to each other through a core tensor. Extended Data Figure 2b shows the result of canonical polyadic decomposition on a three-dimensional tensor *O*. The core tensor, *G*, is a three-dimensional tensor comprised of only diagonal elements, each of which specifies the amount of variance carried by a tensor component analogous to the singular values computed by SVD. Each tensor component relates the rows, columns and third-dimensional entries of *O*. The number of tensor components is determined by creating a tensor that is randomized with respect to the rows, columns and third-dimensional entries and performing canonical polyadic decomposition over 100 trials. The randomization process scrambles the correlations between each dimension of the tensor; therefore, the resulting canonical polyadic decomposition reflects a random distribution of tensor component values. We used the alternating least squares algorithm for the canonical polyadic decomposition (CP-ALS) to iteratively improve the matrix factorization. The lowest tensor component with a variance that was above that of tensor component 1 of the scrambled tensor defined the number of tensor components we considered. HOSVD was performed using MATLAB (version 2019b).

Aptamer-based characterization of the plasma proteome

Levels of proteins were quantified in a 50-µl aliquot of plasma using the SOMAscan 1.3K Proteomic Assay plasma/serum kit (SomaLogic). Procedures used for quality-control filtering and analysis of differential protein abundances have previously been described[49]. In brief, microarrays were scanned with an Agilent SureScan instrument at 5-µm resolution and the Cy3 fluorescence readout was quantified. Raw fluorescence signal values from each SOMAmer were processed using standardization procedures recommended by the manufacturer (that is, datasets were normalized to remove hybridization variation within a run followed by median normalization across all samples to remove other assay biases). The final .adat file was $\log_2$-transformed, quantile-normalized and then filtered to remove non-human SOMAmers.

CC-SVD and CompBio

CC-SVD begins by computing the cross-correlation matrix between two feature types (microbial CAZymes and host plasma proteins). Given two matrices of dimensions $N^{m \times n}$ (with elements $N_{i,j}$) and $P^{m \times p}$ (with elements $P_{k,l}$) in which *m* is the number of samples and *n* and *p* are the number of features of each feature type, a cross-correlation matrix is calculated by taking each feature in the $m \times n$ matrix *N* and correlating them with each feature in the $m \times p$ matrix *P* (step 1 of Extended Data Fig. 8). The resulting matrix is a $n \times p$ cross-correlation matrix $A^{n \times p}$, in which each element $A_{j,l}$ contains the correlation between feature $N_{1:m,j}$ from the first matrix and feature $P_{1:m,l}$ from the second matrix (these starting matrices contain abundance information, whereas the resulting cross-correlation matrix contains correlations between features). Next, SVD is used to decompose the cross-correlation matrix *A* into left and right singular matrices which contain left (*U*) and right (*V*) singular vectors, respectively; the left singular vectors correspond to the features of *N* and the right singular vectors correspond to the features of *P* (step 2 of Extended Data Fig. 8). A singular vector represents a module of cross-correlated features with a unique correlation profile, and the projections of each feature onto a singular vector represents the module membership of that feature (for example, how similar the correlation profile of a feature is to the correlation profile of the overall module). To define a module, a user-defined threshold truncates the leading and trailing tails of the distribution of projections along a singular vector, and the features above and below the truncation are considered module members; in this case, modules were defined as those proteins with projections along the tails (10th and 90th percentiles)

of the distribution ($\alpha < 0.1$) of the right singular vector 1 (SV1). SVD determines a projection for all features along each singular vector, providing a continuous measure of module membership. The number of singular vectors that should be considered is determined using a random-matrix approximation that has previously been described[50]. The continuous nature of projection values enabled us to rank-order proteins by their projections along SV1.

A literature analysis was then performed using CompBio V2.0 (available from PercayAI; https://www.percayai.com); our goal was to identify relevant processes and pathways represented by modules of plasma proteins derived from CC-SVD analysis (10th and 90th percentiles of projections along SV1) that were highly correlated with CAZyme genes that were significantly increased by fibre snack consumption (step 3 of Extended Data Fig. 8). As previously described[51–54], CompBio uses an automated extraction of knowledge from all PubMed abstracts that reference entities of interest (Entrez Gene Symbol of input plasma proteins or their synonyms) by using contextual language processing and a biological language dictionary that is not restricted to fixed pathway and ontology knowledge bases. Conditional probability analysis was used to compute the statistical enrichment of contextually associated biological concepts (for example, pathways) over those that occur by random sampling. These related concepts, built from the list of highly correlated entities that were used as input, were further clustered into higher-level themes (for example, biological processes, cell types and structures and so on).

Within CompBio, scoring of a given input entity, concept and overall theme enrichment was accomplished using a multi-component function referred to as the normalized enrichment score (NES), commonly used in gene set enrichment analysis[55]. For our analysis, NES used an empirical *P* value derived from several thousand random lists of proteins, each containing a similar number of proteins to those in the 10th- and 90th-percentile groups. NES for concepts was first calculated and the values were then used to compute a NES score for their associated themes:

$$\mathrm{NES}_{\text{concept from 10th or 90th percentile)}} = -\log(P_{\text{concept random}}) \times (\text{enrichment score}_{\text{concept from 10th or 90th percentile}}/\text{mean enrichment score}_{\text{concept random}})$$

$$\mathrm{NES}_{\text{theme 1}} = \mathrm{NES}_{\text{concept 1}} + \mathrm{NES}_{\text{concept 2}} \ldots \mathrm{NES}_{\text{concept } n} \text{ from 10th or 90th percentile bin} = -\log(P_{\text{concept1, 2 ... } n}) \times (\text{enrichment score}_{\text{concept from 10th or 90th percentile}}/\text{mean enrichment score}_{\text{concept random}})$$

A two-step approach was used for annotation of themes within CompBio. For each theme, the most enriched concepts were first mapped to pathways and processes in common knowledge bases such as Gene Ontology, KEGG and Reactome pathways. Second, as CompBio is not restricted to these knowledge bases in its biological discovery process, themes were further annotated by subject area experts, which—together with the initial mapping to existing knowledge bases—resulted in a summary description of the biology described by assimilation of the most enriched concepts in each theme.

We performed an SVD analysis on the $\log_2$-transformed fold change of plasma protein profiles enriched in each biological theme (for each treatment and all participants). Proteins belonging to a theme were selected, and SVD was performed on the $m \times p$ protein abundance matrix in which m is the number of samples and p is the number of proteins in the biological theme. SV1 projections for each biological theme were cross-correlated with CAZymes with abundances that were significantly increased by the given treatment (step 4 of Extended Data Fig. 8). The resulting cross-correlation matrix, which contains Spearman rho values as elements, was plotted as a heat map using the ggplot package (v.3.3.2)[56] in R (v.3.6.0).

Linear model analysis of SOMAscan data generated from human plasma samples

Differentially expressed plasma proteins that passed quality-control criteria (1,205 for human study 1 and 1,170 for human study 2) were identified using limma (v.3.42.2)[25] in R (v.3.6.0). Plasma protein datasets generated from each snack food intervention were analysed separately. In brief, samples collected on day 14 (pre-intervention phase) and day 29 (after two weeks of consuming the pea-fibre snack) from participants in study 1, and samples collected on day 11 (pre-intervention phase) and day 25 (2 weeks of consumption of the two-fibre snack), and on day 11 (pre-intervention phase) and day 49 (two weeks of consumption of the four-fibre snack) from participants in study 2 were analysed using a linear least-squares model (design = model.matrix(~ 0 + time point + participant ID)). Plasma proteins were selected if their differences in abundance between the two time points were statistically significant ($P < 0.05$) for a given fibre snack treatment.

Mass spectrometry-based carbohydrate analysis of fibre preparations, diets and faecal samples

Monosaccharide and linkage analysis of fibre preparations. Previously described methods[11] were used to define the carbohydrate composition of the pea, orange and barley bran fibre preparations. Following a pre-hydrolysis step (incubation in concentrated sulfuric acid (72%) for 30 min at 30 °C to release glucose from cellulose), the fibres were hydrolysed by treatment with 1 M sulfuric acid for 6 h at 100 °C. Individual neutral sugars were analysed by gas chromatography as their alditol acetate derivatives[57,58]. The meta-hydroxydiphenyl colorimetric acid method was used to measure uronic acid (as galacturonic acid)[59,60]; sodium tetraborate was used to differentiate glucuronic acid from galacturonic acid[61]. Methylation of galacturonic acid (in pectin) was estimated according a previous publication[62].

Linkage analysis of fibres followed previously detailed procedures[63] with minor modifications that allowed for discrimination of galactose, galacturonic acid, and methyl-esterified galacturonic acid. In brief, reduction of carboxymethyl ester groups of uronic acids was performed with $NaBD_4$ and imidazole-HCl, followed by activation of carboxylic acid groups with carbodiimide and a second reduction with imidazole-HCl, $NaBH_4$ (D/H) and $NaBD_4$ (D/D). Samples were dialysed, freeze-dried and then solubilized in DMSO before methylation of the accessible hydroxyl groups of reduced polysaccharides using iodomethane. Acid hydrolysis with trifluoroacetic acid and a subsequent reduction of partially methylated sugars with $NaBD_4$ was performed. Finally, samples were acetylated, extracted as partially methylated alditol acetates into dichloromethane, and analysed by gas chromatography–mass spectrometry[64].

Monosaccharide and linkage analysis of mouse diets and faecal biospecimens. For homogenization of HiSF–LoFV diet with and without fibre supplementation, a 10 mg ml^{-1} stock solution was prepared from frozen starting material. Pre-weighed mouse and human faecal samples were diluted tenfold in nanopure water (Thermo Fisher) and homogenized overnight. Samples were lyophilised to complete dryness and diluted to create a stock solution (10 mg ml^{-1} in water). Stock solutions were bullet-blended using 1.4-mm stainless steel beads followed by incubation at 100 °C for 1 h. Finally, samples were subjected to another round of bullet blending and aliquots were taken for monosaccharide and linkage analysis.

Methods for monosaccharide analysis of diets and faecal samples were adapted from a previous publication[65]. In brief, three 10-µl aliquots were taken from each bullet-blended 'stock', transferred to a 96-well plate and subjected to acid hydrolysis (4 M trifluoroacetic acid for 1 h at 121 °C). The reaction was quenched with 855 µl of ice-cold nanopure water. Hydrolysed samples were derivatized with 1-phenyl-3-methyl-5-pyrazolone (PMP) according to previously described conditions[66]. Samples and 14 monosaccharide standards (0.001–100 µg ml^{-1}) were reacted in 0.2 M PMP (prepared in methanol) and 28% NH_4OH at 70 °C for 30 min. Derivatized glycosides were then dried to completion (vacuum centrifuge) and reconstituted in nanopure water. Excess PMP was removed (chloroform extraction) and a 1-µl

aliquot of the aqueous layer was injected into an Agilent 1290 Infinity II ultrahigh-performance liquid chromatography (UHPLC) system coupled to an Agilent 6495A triple quadrupole mass spectrometer under dynamic multiple reaction monitoring (dMRM) mode. Monosaccharides were quantified using calibration curves generated with external standards.

The procedure for linkage analysis was adapted from previously described protocols[67,68]. In short, three replicate 5-µl aliquots of each bullet-blended stock solution were incubated in saturated NaOH and iodomethane (in DMSO) to achieve methylation of free hydroxyl groups. Excess NaOH and DMSO were removed by extraction with dichloromethane and water. Permethylated samples were subsequently hydrolysed and derivatized (using the same procedure employed for monosaccharide analysis). Derivatized samples were subjected for UHPLC–dMRM mass spectrometry. Glycosidic linkages present in samples were identified using a pool of oligosaccharide standards and a comprehensive linkage library that has previously been described[67,68].

A cross-correlation matrix was computed between the $\log_2$-transformed fold-change of statistically significant pea-fibre-responsive CAZymes and faecal monosaccharides and glycosidic linkages in participants on days 29 and 35 (during consumption of pea fibre snacks) and days 45 and 49 (post-intervention phase) normalized to day 14 (pre-intervention phase). A cross-correlation matrix generated from this analysis with Spearman rho values as elements is plotted as a heat map in Extended Data Fig. 7.

Liquid chromatography–quadrupole time-of-flight mass spectrometry identification of a faecal biomarker of orange fibre consumption. Methods for preparing samples and performing liquid chromatography–quadrupole time-of-flight mass spectrometry LC–QTOF MS using an Agilent 1290 LC system coupled to an Agilent 6545 Q-TOF mass spectrometer are detailed in a previous publication[69]. Five µl of each prepared faecal sample for positive ESI ionization were injected into a BEH C18 column (2.1 × 150 mm, 1.7 µm, Waters) that was heated to 35 °C. The mobile phase was 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B). The following gradient was applied at a flow rate of 0.3 ml min^{-1} over 14 min; 95% A/5% B to 100% B, followed by 3 min at 100% B. Analysis of faecal samples revealed a biomarker (m/z 274.1442) of orange fibre consumption.

Reporting summary

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

Data availability

V4-16S rRNA sequences in raw format before post-processing and data analysis, plus shotgun sequencing datasets generated from faecal DNA, have been deposited at the European Nucleotide Archive under study accession PRJEB38148. Raw proteomic datasets generated from the aptamer-based 1.3K SomaLogic platform have been deposited in the European Genome-Phenome Archive (EGA) under accession IDs EGAD00010002133 (pea fibre study) and EGAD00010002132 (fibre blends study). Any other relevant data are available from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

Code for HOSVD (CP-ALS plus randomization code) and CC-SVD is available via Zenodo (https://doi.org/10.5281/zenodo.4767887).

31. Bucholz, K. K., Heath, A. C. & Madden, P. A. Transitions in drinking in adolescent females: evidence from the Missouri adolescent female twin study. *Alcohol. Clin. Exp. Res.* **24**, 914–923 (2000).
32. Mifflin, M. D. et al. A new predictive equation for resting energy expenditure in healthy individuals. *Am. J. Clin. Nutr.* **51**, 241–247 (1990).
33. Subar, A. F. et al. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am. J. Epidemiol.* **154**, 1089–1099 (2001).
34. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA* **108** (Suppl 1), 4516–4522 (2011).
35. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
36. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
37. Di Luccia, B. et al. Combined prebiotic and microbial intervention improves oral cholera vaccination responses in a mouse model of childhood undernutrition. *Cell Host Microbe* **27**, 899–908.e5 (2020).
38. Baym, M. et al. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE* **10**, e0128036 (2015).
39. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJ.* **17**, 10–12 (2011
40. Joshi, N. A. & Fass, J. N. *Sickle: A Sliding-Window, Adaptive, Quality-based Trimming tool for FastQ Files (Version 1.33) Software* (2011).
41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
42. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
43. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
44. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
45. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
46. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
47. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
48. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
49. Chen, R. Y. et al. Duodenal microbiota in stunted undernourished children with enteropathy. *N. Engl. J. Med.* **383**, 321–333 (2020).
50. Plerou, V. et al. Random matrix approach to cross correlations in financial data. *Phys. Rev. E* **65**, 066126 (2002).
51. Winkler, E. S. et al. Human neutralizing antibodies against SARS-CoV-2 require intact Fc effector functions for optimal therapeutic protection. *Cell* **184**, 1804–1820.e16 (2021).
52. Zou, W. et al. Ablation of fat cells in adult mice induces massive bone gain. *Cell Metab.* **32**, 801–813.e6 (2020).
53. Adamo, L. et al. Proteomic signatures of heart failure in relation to left ventricular ejection fraction. *J. Am. Coll. Cardiol.* **76**, 1982–1994 (2020).
54. Tsingas, M. et al. *Sox9* deletion causes severe intervertebral disc degeneration characterized by apoptosis, matrix remodeling, and compartment-specific transcriptomic changes. *Matrix Biol.* **94**, 110–133 (2020).
55. Joly, J. H., Lowry, W. E. & Graham, N. A. Differential gene set enrichment analysis: a statistical approach to quantify the relative enrichment of two gene sets. *Bioinformatics* **36**, 5247–5254 (2020).
56. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
57. Blakeney, A. B., Harris, P. J., Henry, R. J. & Stone, B. A. A simple and rapid preparation of alditol acetates for monosaccharide analysis. *Carbohydr. Res.* **113**, 291–299 (1983).
58. Englyst, H. N. & Cummings, J. H. Improved method for measurement of dietary fiber as non-starch polysaccharides in plant foods. *J. Assoc. Off. Anal. Chem.* **71**, 808–814 (1988).
59. Blumenkrantz, N. & Asboe-Hansen, G. New method for quantitative determination of uronic acids. *Anal. Biochem.* **54**, 484–489 (1973).
60. Thibault, J.-F. Automatisation du dosage des substances pectiques par la méthode au métahydroxydiphényle. *Lebensm. Wiss. Technol.* **12**, 247–251 (1979).
61. Filisetti-Cozzi, T. M. C. C. & Carpita, N. C. Measurement of uronic acids without interference from neutral sugars. *Anal. Biochem.* **197**, 157–162 (1991).
62. Levigne, S., Thomas, M., Ralet, M.-C., Quemener, B. & Thibault, J.-F. Determination of the degrees of methylation and acetylation of pectins using a C18 column and internal standards. *Food Hydrocoll.* **16**, 547–550 (2002).
63. Pettolino, F. A., Walsh, C., Fincher, G. B. & Bacic, A. Determining the polysaccharide composition of plant cell walls. *Nat. Protoc.* **7**, 1590–1607 (2012).
64. Buffetto, F. et al. The deconstruction of pectic rhamnogalacturonan I unmasks the occurrence of a novel arabinogalactan oligosaccharide epitope. *Plant Cell Physiol.* **56**, 2181–2196 (2015).
65. Amicucci, M. J. et al. A rapid-throughput adaptable method for determining the monosaccharide composition of polysaccharides. *Int. J. Mass Spectrom.* **438**, 22–28 (2019).
66. Xu, G., Amicucci, M. J., Cheng, Z., Galermo, A. G. & Lebrilla, C. B. Revisiting monosaccharide analysis - quantitation of a comprehensive set of monosaccharides using dynamic multiple reaction monitoring. *Analyst* **143**, 200–207 (2018).
67. Galermo, A. G. et al. Liquid chromatography-tandem mass spectrometry approach for determining glycosidic linkages. *Anal. Chem.* **90**, 13073–13080 (2018).
68. Galermo, A. G., Nandita, E., Castillo, J. J., Amicucci, M. J. & Lebrilla, C. B. Development of an extensive linkage library for characterization of carbohydrates. *Anal. Chem.* **91**, 13022–13031 (2019).
69. Cowardin, C. A. et al. Mechanisms by which sialylated milk oligosaccharides impact bone biology in a gnotobiotic mouse model of infant undernutrition. *Proc. Natl Acad. Sci. USA* **116**, 11988–11996 (2019).

Acknowledgements We thank M. Karlsson and D. O'Donnell for their assistance with gnotobiotic mouse husbandry; S. Marion for her role in collecting clinical metadata and faecal samples from members of the MOAFTs study cohort with obesity, and the two human studies; S. Waller for design of the controlled diet study menus and oversight of the metabolic kitchen; S. Torbitzky for her assistance with coordination and planning of the human studies; M. Patnode for his input about the selection of lead fibres; S. Le Gall, L. Saulnier and B. Laillet for carbohydrate analysis of pea, orange and barley bran fibre preparations; G. Cesbron Lavau and M. Okoniewska for formulation and scale-up production of snack prototypes for human studies; S. Abdel-Hamid for production of the HiSF–LoFV diets given to gnotobiotic mice; L. Dimartino for analysis of the fibre ingredients in these diets; and L. Kyro for assistance with figure illustrations. Technical support was provided by S. Deng, J. Serugo, J. Lelwala-Guruge, K. Ahsan, S. Bale, J. Veitinger, J. Forman and S. Karlsson (archiving and processing mouse and human biospecimens), M. Meier, J. Hoisington-López and M. Crosby (bacterial V4-16S rDNA amplicon and faecal microbiome shotgun sequencing), T. Juehne, A. Lutz and J. Yu (generating SOMAscan datasets), and R. Head and C. Storer (assistance with CompBio analyses). This work was funded by grants from the NIH (DK078669, DK70977, UL1-TR002345) and from Mondelēz Global LLC. Controlled diet studies of human participants were overseen by members of the Clinical Science Research Core of the Nutrition Obesity Research Center (NORC), which is supported by NIH grant P30 DK056341. Plasma proteomic data sets were generated by the Genome Technology Access Center at Washington University School of Medicine, which is supported in part by NIH Grants P30 CA91842 and UL1TR002345. O.D.-B. received support from NIH R25GM103757, T32GM007067 and T32HL130357 as a pre-doctoral trainee. J.I.G. is the recipient of a Thought Leader Award from Agilent Technologies.

Author contributions O.D.-B. and J.I.G. designed the gnotobiotic mouse studies. A.C.H. oversaw collection of faecal samples from human donors with obesity used to colonize germ-free mice. O.D.-B. and N.H. performed mouse studies. M.J.B., S.K., O.D.-B. and J.I.G. designed the human studies together with D.K.H., A.M. and S.V., who oversaw the design, manufacture and quality-control analysis of the fibre snack prototypes used in the two human studies. A.M. and S.V. organized carbohydrate and glycosidic linkage composition analysis of fibre preparations. Controlled diet studies of human participants were overseen by S.K. together with K.K. and T.W. J.J.C., G.C. and C.B.L. conducted mass spectrometric assays of mouse diets and faecal samples. J.C. performed LC–QTOF–MS analyses of human faecal samples collected from participants consuming the two- and four-fibre-containing snacks. O.D.-B. oversaw the archiving and processing of mouse and human biospecimens and generated the 16S rDNA and shotgun sequencing datasets from these samples. M.C.H. and C.D. implemented the metagenomic assembly and annotation pipeline. D.A.R., S.A.L. and A.O. performed mcSEED pathway reconstructions of faecal microbiomes, and V.L. and B.H. provided CAZyme annotations. A.S.R. developed the HOSVD and R.Y.C. the CC-SVD analytic platforms that were applied to datasets generated from mice and humans. O.D.-B. and R.A.B. performed CompBio analyses of plasma proteome datasets generated from human studies. O.D.-B., C.D., M.J.B. and J.I.G. analysed the data. O.D.-B. and J.I.G. wrote the paper with assistance provided by the co-authors.

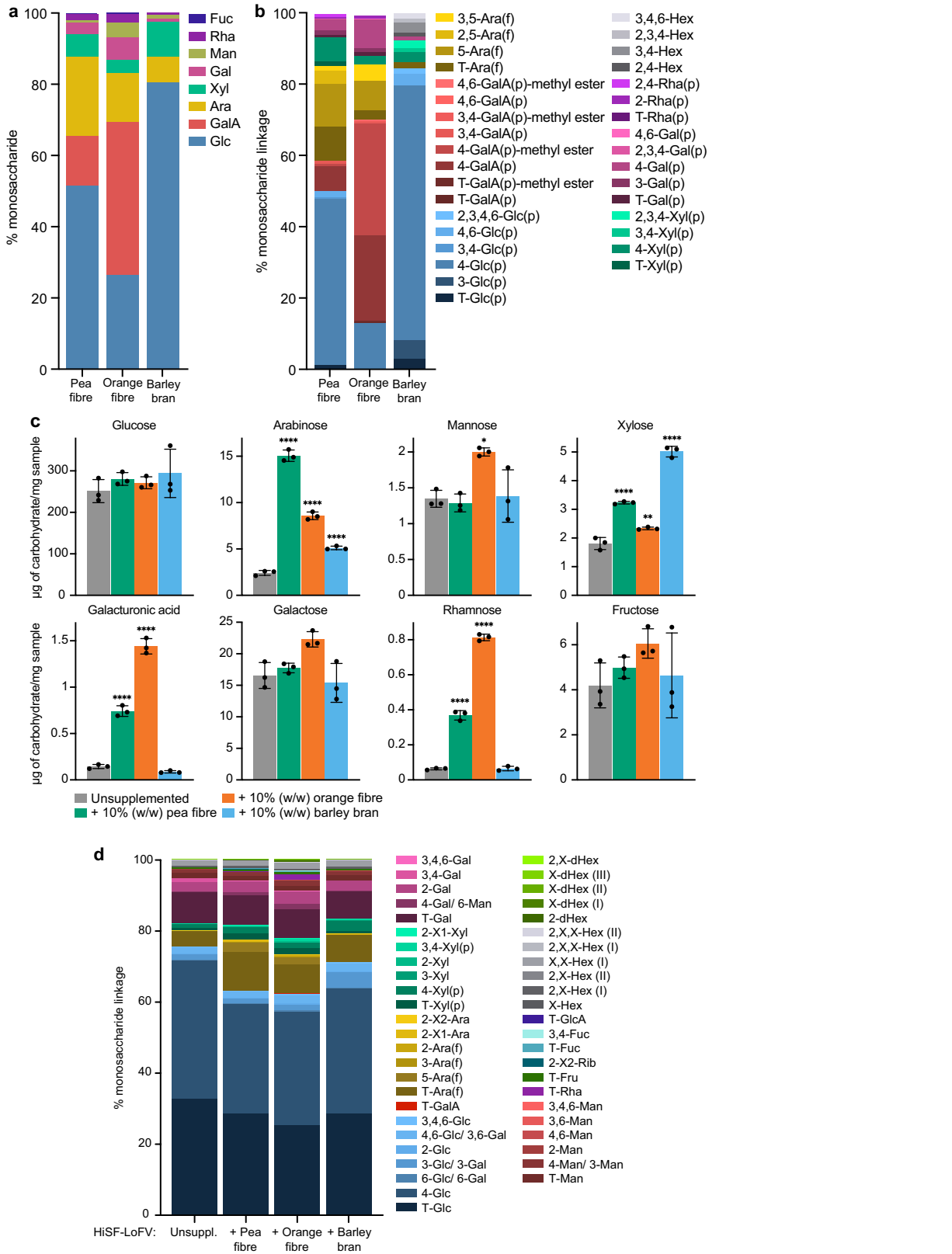
Competing interests J.I.G. is a co-founder of Matatu, Inc., a company characterizing the role of diet-by-microbiota interactions in animal health. A.O. and D.R. are co-founders of Phenobiome Inc., a company pursuing development of computational tools for predictive phenotype profiling of microbial communities. C.B.L. is a co-founder of Evolve Biosystems, interVenn Bio and BCD Bioscience, companies involved in the characterization of glycans and developing carbohydrate applications for human health. D.K.H., A.M. and S.V. are employees of Mondelēz Global LLC, a multinational company engaged in production of snack foods. The remaining authors declare that they have no competing financial interests. A patent application related to the fibre-snack formulations described in this Article has been filed and published (WO 2021/016129).

Additional information

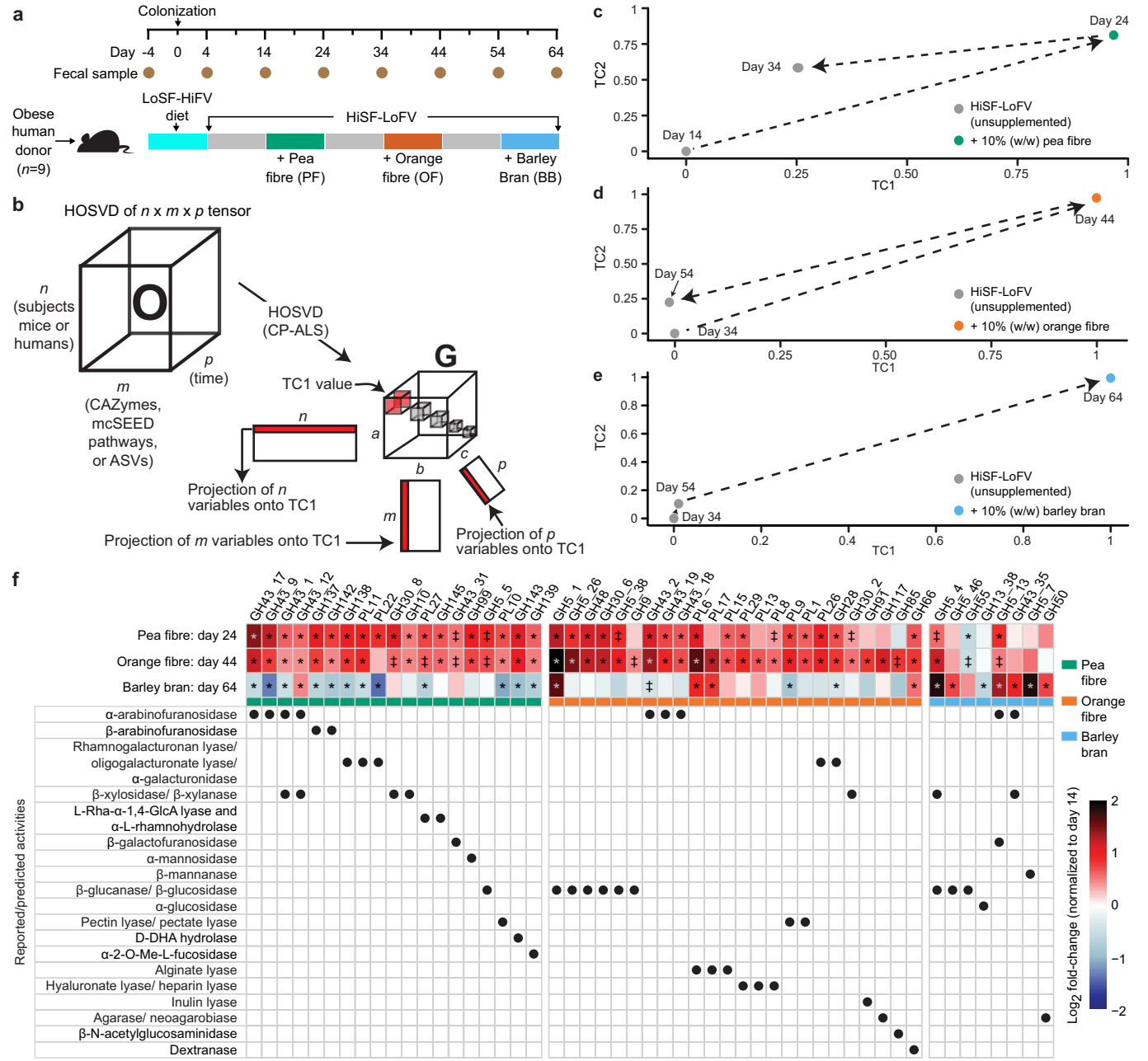
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03671-4.

Correspondence and requests for materials should be addressed to J.I.G.

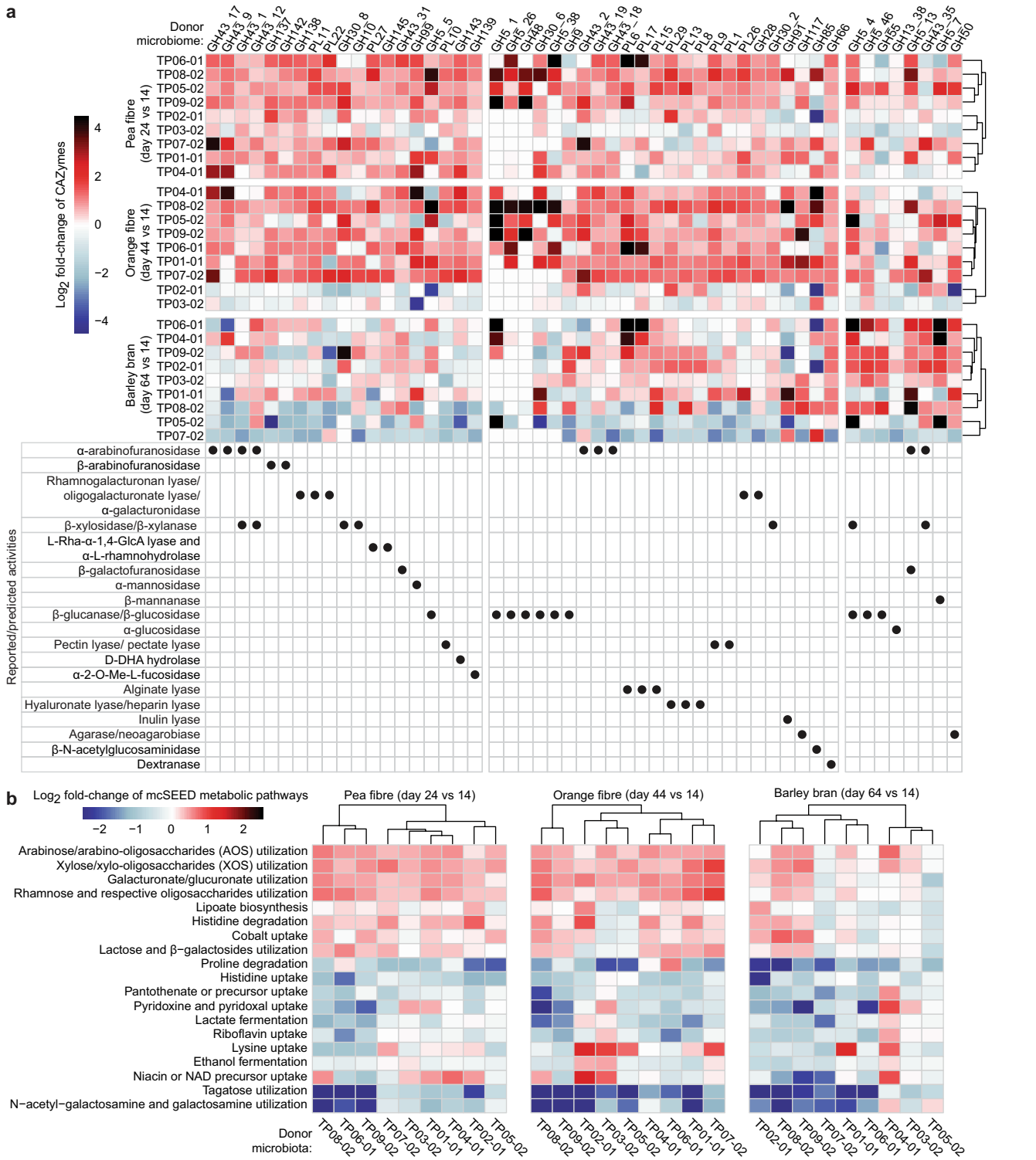
Peer review information *Nature* thanks David Bolam, Eran Elinav and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



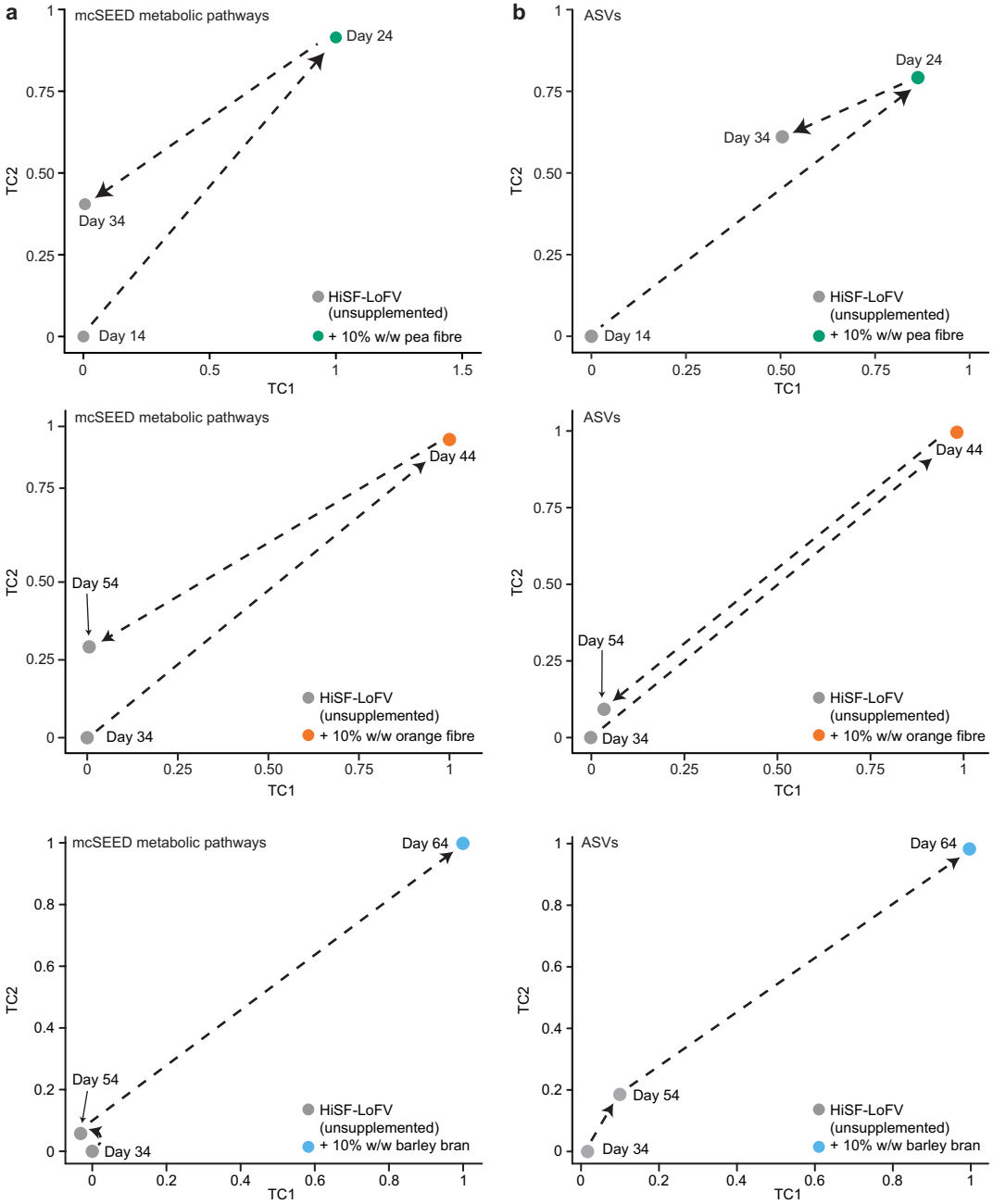
Extended Data Fig. 1 | Monosaccharide content and glycosidic linkages present in the fibre preparations, and in the unsupplemented and fibre-supplemented HiSF–LoFV diets fed to gnotobiotic mice. **a**, **b**, Monosaccharides and linkages, some represented by their methylated monosaccharide derivatives, in the fibre preparations. Stacked bars represent the mean for technical replicates ($n = 3$) for each glycosyl linkage determination. **c**, Monosaccharides in the unsupplemented and fibre-supplemented HiSF–LoFV diets. Bars represent the mean ± s.d. for technical replicates ($n = 3$). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$ compared to the unsupplemented diet phase (one-way ANOVA with Holm–Šídák multiple comparison correction). **d**, Linkages, represented by their methylated monosaccharide derivatives in the unsupplemented and fibre-supplemented HiSF–LoFV diets. Stacked bars represent the mean for technical replicates ($n = 3$) for each glycosidic linkage determination. Glc, glucose; Gal, galactose; GalA, galacturonic acid; GlcA, glucuronic acid; Ara, arabinose; Xyl, xylose; Man, mannose; Fru, fructose; Fuc, fucose; Rha, rhamnose; Rib, ribose; Hex, hexose; dHex, deoxyhexose; T, terminal; f, furanose; p, pyranose; X, undefined linkage.



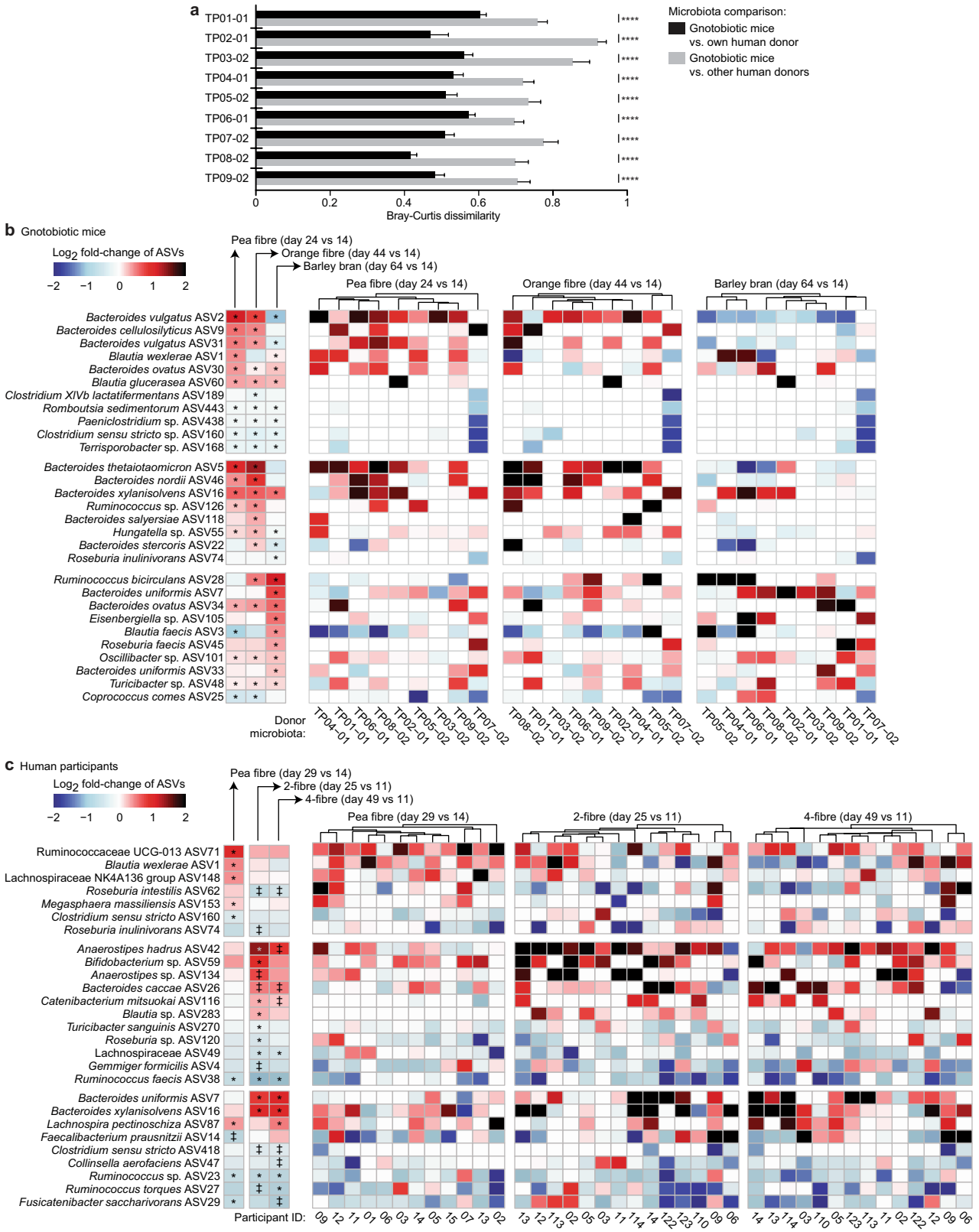
Extended Data Fig. 2 | The effects of dietary fibres in gnotobiotic mice fed a HiSF-LoFV diet and colonized with faecal microbial communities of nine human donors with obesity. a, Experimental design. **b**, HOSVD. Three-dimensional matrices are termed 'tensors'. A tensor (O) with dimensions n, m and p, in which n represents subjects (mice or humans), m represents features (CAZymes, mcSEED metabolic pathways and ASVs) and p represents time, can be analysed by HOSVD in which a 'core tensor' (G) is created—a tensor in which the only non-zero values are along the diagonal (boxes shown in G). Each box represents a 'tensor component' (TC). Each tensor component relates the variation between each axis of the O tensor. Additionally, three new matrices are created that are related to each other through each tensor component; for example, 'Projection of variables onto TC1' indicates that variation defined by tensor component 1 is defined by variation across the first row of n, the first column of m and the first column of p. **c**–**e**, HOSVD applied to CAZymes in faecal microbiomes of mice colonized with microbial communities of nine human donors with obesity (n = 348 faecal samples analysed) during each of the three dietary fibre interventions in the diet oscillation experiment. **f**, Heat map of discriminatory CAZymes with $\log_2$-tranformed fold changes in abundance that were defined as statistically significant during at least one dietary intervention. The grand mean of the data is shown for mice containing the nine different human donor microbiomes sampled at the indicated time points and normalized to day 14 values (n = 6–10 mice per group; n = 232 faecal samples analysed). The order of CAZymes from left to right is based on their function (rows below) and magnitude of their change within and across fibre treatments. ‡q value < 0.10; *q value < 0.05 (linear mixed-effects model, FDR-corrected).



Extended Data Fig. 3 | Responses of CAZymes and mcSEED metabolic pathways identified by HOSVD analysis as discriminatory for fibre snack consumption in gnotobiotic mice colonized with the faecal communities of nine human donors with obesity. a, **b**, Heat map of discriminatory CAZymes and mcSEED metabolic pathways with changes in abundance that were statistically significant during at least one dietary intervention. Data are averaged for mice containing a given human donor microbiota ($n = 6$ to 10 mice per group; $n = 232$ faecal samples analysed) sampled at the indicated time points and normalized to day-14 values. The order of CAZymes from left to right of the heat map in **a** follows the same order as in Extended Data Fig. 2f, and the order of mcSEED pathways from top to bottom in **b** follows the same order as shown in Fig. 1d. Hierarchical clustering (Euclidean distances) of CAZyme and mcSEED metabolic pathway profiles was used to group donor microbiomes.

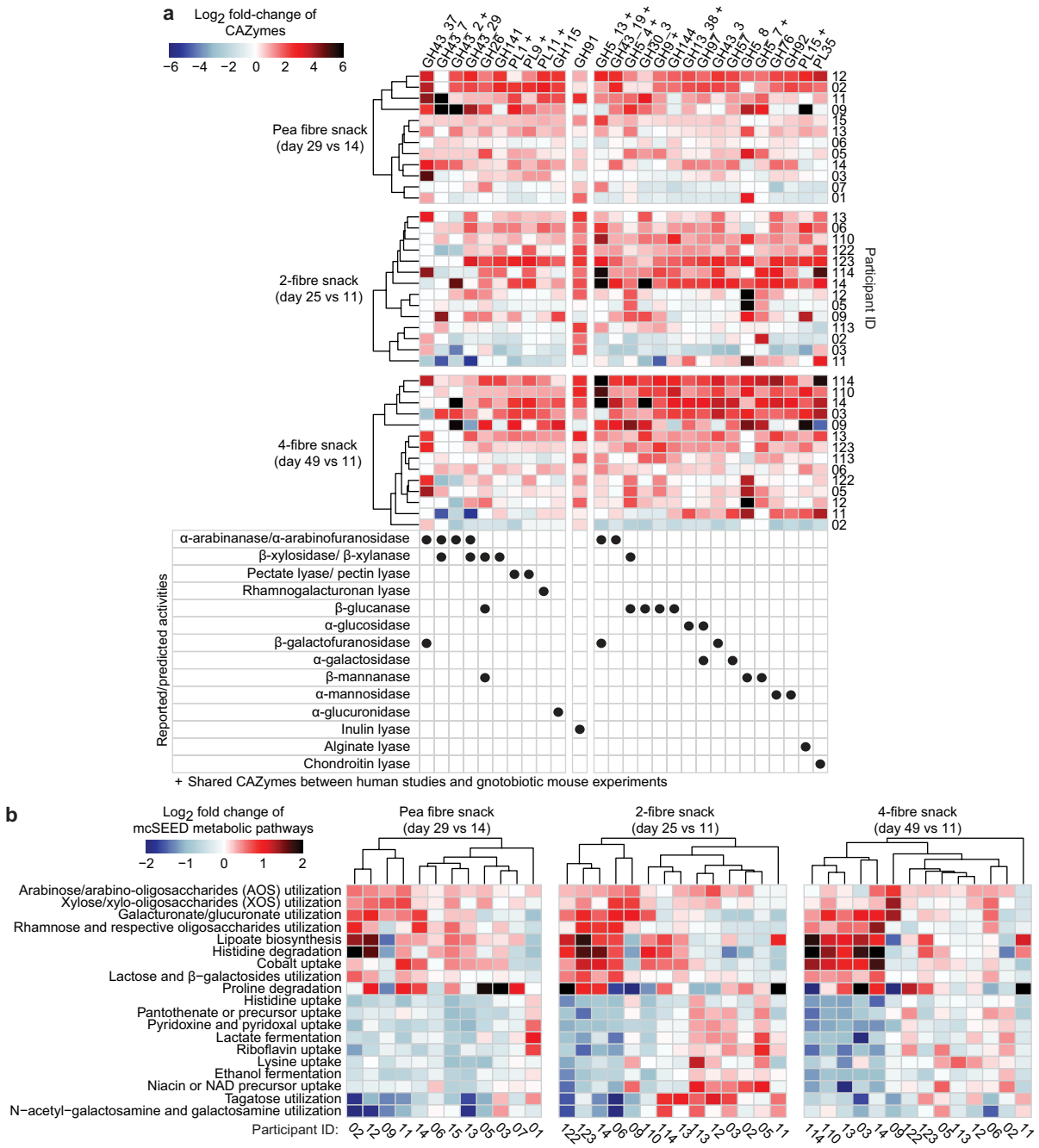


Extended Data Fig. 4 | HOSVD applied to mcSEED metabolic pathway and ASV datasets generated from the faecal microbial communities of mice containing microbial communities of human donors with obesity during the pea, orange and barley bran fibre phases of the diet oscillation. **a**, Microbiome configurations as defined by the representation of mcSEED metabolic pathways on tensor component 1 and tensor component 2 during pea fibre, orange fibre and barley bran phase of diet oscillation. **b**, Projections of microbiota configuration as defined by the representation of bacterial taxa (ASVs) on tensor component 1 and tensor component 2 during pea fibre, orange fibre and barley bran phases of the diet oscillation protocol ($n = 9$ human microbiomes; $n = 6$–10 mouse recipients of each human microbiome; $n = 348$ faecal samples analysed for data presented in **a**, **b**).

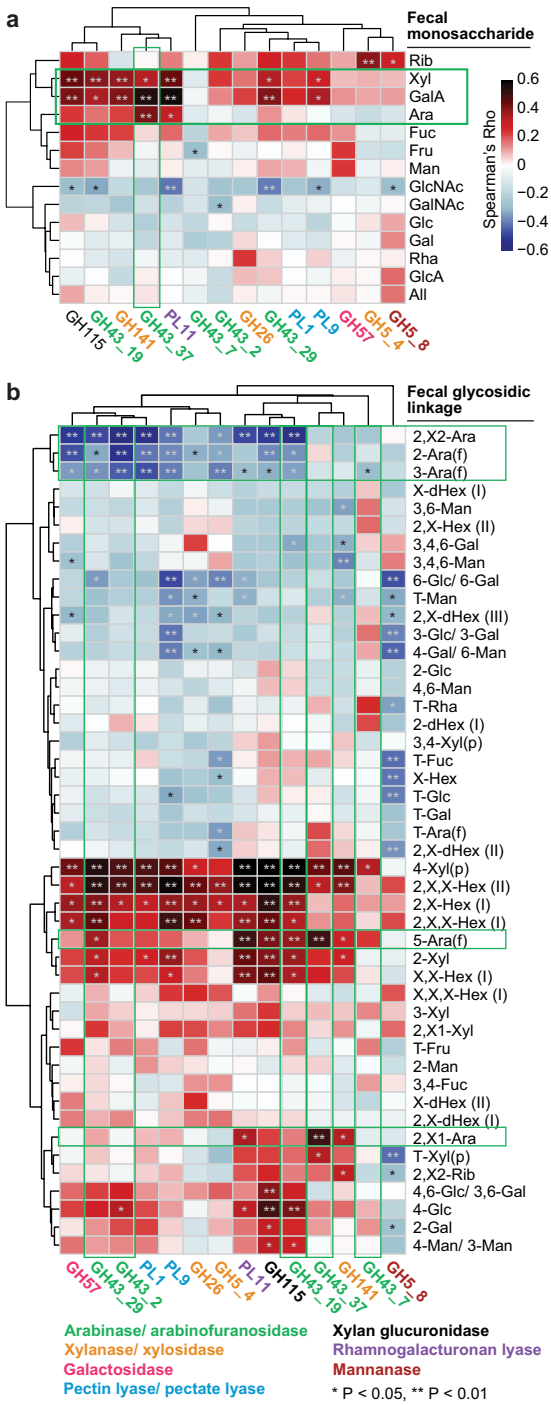


Extended Data Fig. 5 | See next page for caption.

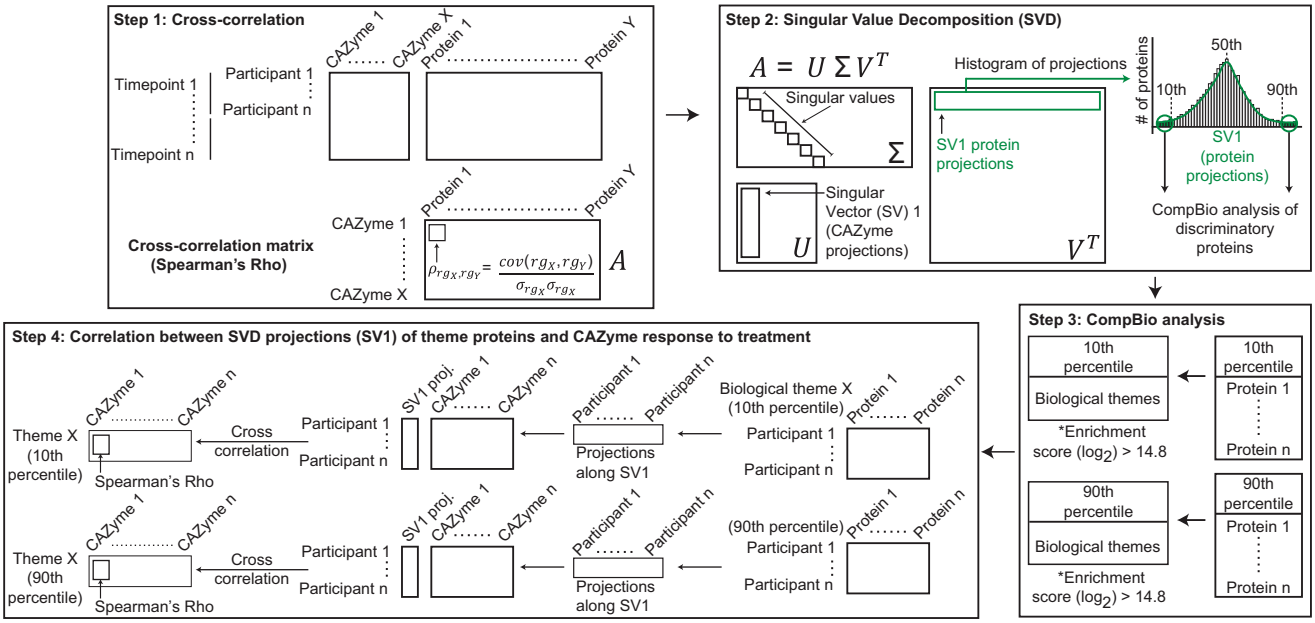
Extended Data Fig. 5 | Responses of bacterial taxa (ASVs) identified by HOSVD as discriminatory for dietary fibre consumption in gnotobiotic mice colonized with human donor microbiota, and in human participants enrolled in the controlled-diet studies. a, Bray–Curtis dissimilarity distances calculated from the ASV content of communities sampled at all time points (days 4, 9, 14, 19, 24, 29, 34, 39, 44, 49, 54, 59 and 64 after colonization) from a given group of recipient mice compared to the ASV content of their corresponding human donor community before transplantation (Bray–Curtis distances calculated from ASV abundances in faecal samples collected from each group of mice (n = 6–10 mice; n = 752 faecal samples in total) compared to the abundances of these ASVs in each of their corresponding nine human-donor faecal communities). ****P value < 0.0001 (one-way ANOVA, Šídák's correction). **b**, Heat map of statistically significant $\log_2$-transformed fold changes in the abundances of discriminatory ASVs in gnotobiotic mice during at least one of the fibre interventions. The heat map on the left shows the grand mean for data obtained from all groups of mice and the heat map on the right shows averaged data for mice containing a given donor microbiota (n = 6 to 10 mice per group; n = 232 faecal samples obtained at the indicated time points with data normalized to day-14 values). Hierarchical clustering (Euclidean distances) of ASV profiles was used to group donor microbiota with similar responses to each fibre supplement. **c**, ASVs with $\log_2$-transformed fold changes in abundance that were statistically significant in human participants after at least one of the fibre snack interventions. Left, mean values for participants enrolled in each study; the three panels to the right show changes in ASV abundances in individual participants after consumption of each of the fibre snacks. Data are normalized to pretreatment time points, that is, day 14 (study 1) and day 11 (study 2) (n = 12 and 14 participants for study 1 and 2, respectively, n = 66 faecal samples analysed). Hierarchical clustering (Euclidean distances) of ASV profiles was used to group participants with similar responses to a given fibre snack. ‡q value < 0.1, *q value < 0.05 (linear mixed-effects model, FDR-corrected).



Extended Data Fig. 6 | Identification by HOSVD of fibre-snack-discriminatory CAZymes and mcSEED metabolic pathways in human participants in the controlled-diet studies. a, **b**, CAZymes and mcSEED metabolic pathways with $\log_2$-transformed fold changes in abundance were statistically significant during at least one of the three fibre snack interventions. Data are shown for each participant after consumption of each fibre snack type and are normalized to pretreatment time points, that is, day 14 (study 1) and day 11 (study 2) ($n = 12$ and 14 participants for study 1 and 2, respectively, $n = 66$ faecal samples analysed). Hierarchical clustering (Euclidean distances) of CAZyme and mcSEED metabolic pathway profiles was used to group participants with similar responses to each fibre snack type. CAZymes marked with a + were also fibre-treatment discriminatory in the gnotobiotic mouse studies.

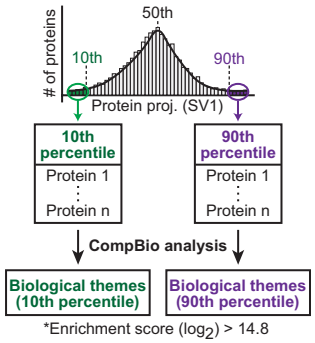


Extended Data Fig. 7 | Spearman-rank cross-correlation analysis of the representation of CAZymes, monosaccharides and glycosidic linkages in the faecal communities of participants consuming the pea-fibre snack prototype. a, **b**, Correlations between the $\log_2$-transformed fold change of statistically significant, HOSVD-defined discriminatory CAZyme gene abundances (matched by time and participant) to the $\log_2$-transformed fold change in levels of monosaccharides and glycosidic linkages at days 25 and 35 (fibre snack consumption), and days 45 and 49 (post-intervention phase) normalized to day 14 (pre-intervention phase; $n = 60$ faecal samples analysed; 12 participants). Green boxes in **a** highlight a statistically significant positive correlation between GH43_37 (arabinofuranosidase) and arabinose, a prominent monosaccharide component of pea fibre. In **b**, evidence is provided that participant microbiomes contain CAZymes that cleave multiple branches of pea fibre arabinan, resulting in accumulation of its 1,5-arabinofuranose backbone in faeces. Further details are in Supplementary Results. $^*P < 0.05$; $^{**}P < 0.01$. Glucose (Glc), galacturonic acid (GalA), arabinose (Ara), xylose (Xyl), galactose (Gal), mannose (Man), rhamnose (Rha), fucose (Fuc), fructose (Fru), glucuronic acid (GlcA), *N*-acetylglucosamine (GlcNAc), *N*-acetylgalactosamine (GalNAc), allose (All), ribose (Rib), hexose (Hex), deoxyhexose (dHex); terminal (T), pyranose (p), furanose (f), undefined linkage (X).



Extended Data Fig. 8 | Schematic of the analytic pipeline for identifying associations between changes in the plasma proteome and CAZyme responses after fibre snack consumption. Step 1 shows cross-correlation analysis between plasma proteins and discriminatory CAZymes with changes in abundance that were statistically significant. Step 2 shows SVD analysis of Spearman's rho values of the cross-correlation matrix. Proteins with projections along SV1 are plotted in a histogram to identify those proteins most correlated with discriminatory CAZymes (those within the 10th and 90th percentile, $\alpha < 0.1$). Step 3 represents a CompBio-based analysis of groups of proteins with SV1 projections within the 10th and 90th percentiles. Biological themes enriched in proteins binned in the 10th and 90th percentiles are generated (threshold cutoff for enrichment score ($\log_2$) > 14.8). Step 4 is a SVD analysis of protein profiles within each biological theme for all participants, followed by cross-correlation between SVD projections (SV1) of themed proteins and discriminatory CAZyme responses to treatment.

a



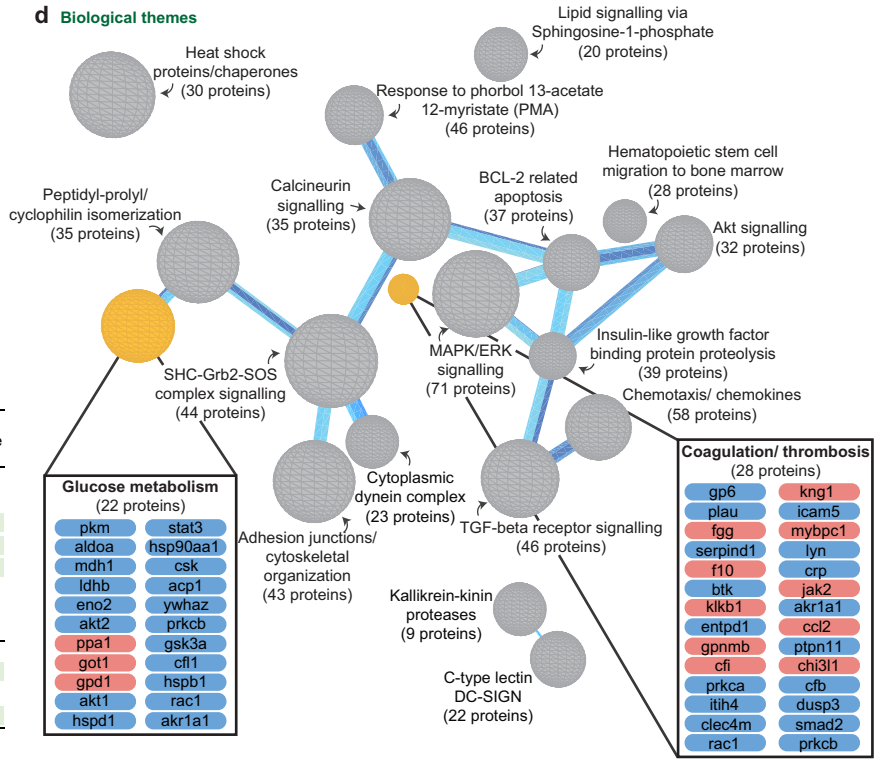
b

Biological themes (10th percentile)	Enrichment score (log_2) 4-fibre snack	2-fibre snack	Pea fibre snack
Signalling pathways			
SHC-Grb2-SOS complex signalling	17.6		
MAPK/ERK signalling	17.5	16.2	15.1
TGF-beta receptor signalling	15.8	15.6	16.2
Calcineurin signalling	15.8		14.9
Lipid signalling via Sphingosine-1-phosphate (S1P)	15.6		
AKT signalling	15.5		
Inflammation/ Coagulation			
Chemotaxis/ chemokines	15.7	16.7	15.0
Kallikrein-kinin proteases	15.6		
Coagulation/ thrombosis	14.9	15.1	15.3
Energy homeostasis			
Glucose metabolism	15.6	15.4	
Other processes related to cell growth/ division/ development			
Heat shock proteins/ chaperones	16.5	15.8	
Adhesion junctions/ cytoskeletal	16.1	14.9	15.2
Peptidyl-prolyl/ cyclophilin	15.8		
Hematopoietic stem cell migration to bone marrow	15.5		14.9
Response to phorbol 13-acetate 12-myristate (PMA)	15.5		
BCL-2 related apoptosis	15.5		15.6
Insulin-like growth factor binding protein proteolysis	15.2		

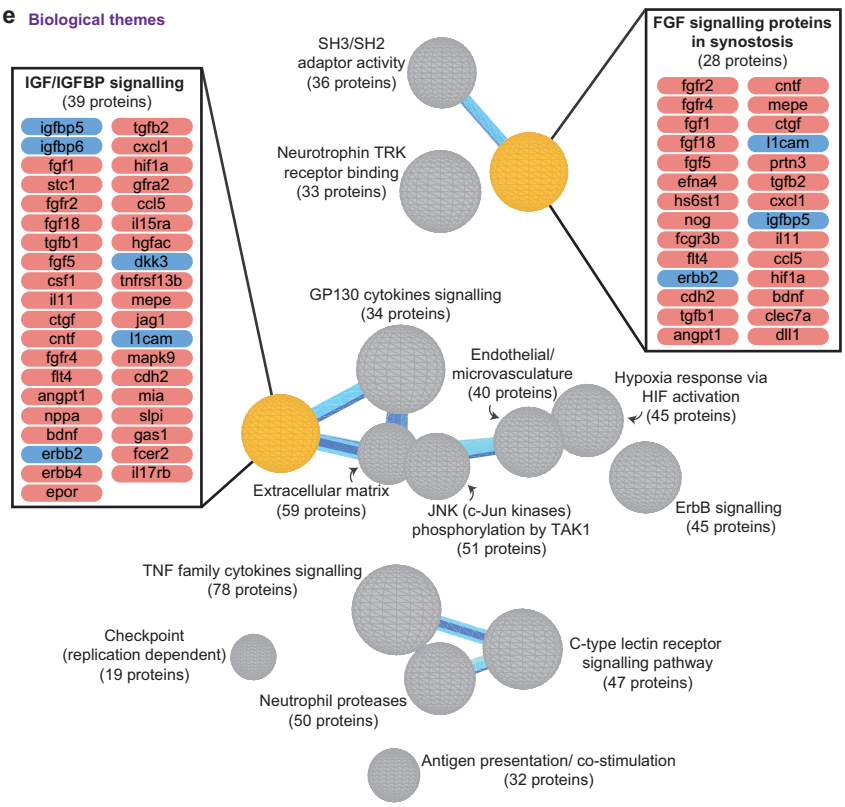
c

Biological themes (90th percentile)	Enrichment score (log_2) 4-fibre snack	2-fibre snack	Pea fibre snack
Immune processes			
TNF family cytokines signalling	16.7		
GP130 cytokines signalling	16.4	16.1	
C-type lectin receptor signalling	16.3		15.1
Neutrophil proteases	16.0	15.8	
Antigen presentation/co-stimulation	15.5	15.0	15.3
Signalling pathways			
ErbB signalling	16.7	17.1	16.9
IGF/IGFBP signalling	16.1	16.5	15.3
SH3/SH2 adaptor activity	16.1		
JNK phosphorylation/ activation mediated by TAK1	15.7		
ECM/ Collagen/ Bone			
FGF signalling in synostosis	16.5	16.5	15.8
Extracellular matrix (ECM)	15.7	16.1	
Angiogenesis/ vascularization			
Endothelial/ microvasculature	15.8	16.2	15.6
Hypoxia response - HIF activation	15.7		
Growth/ development/ differentiation			
Checkpoint (replication dependent)	15.6		
Neuronal function			
Neurotrophin TRK receptor binding	16.2		

d Biological themes

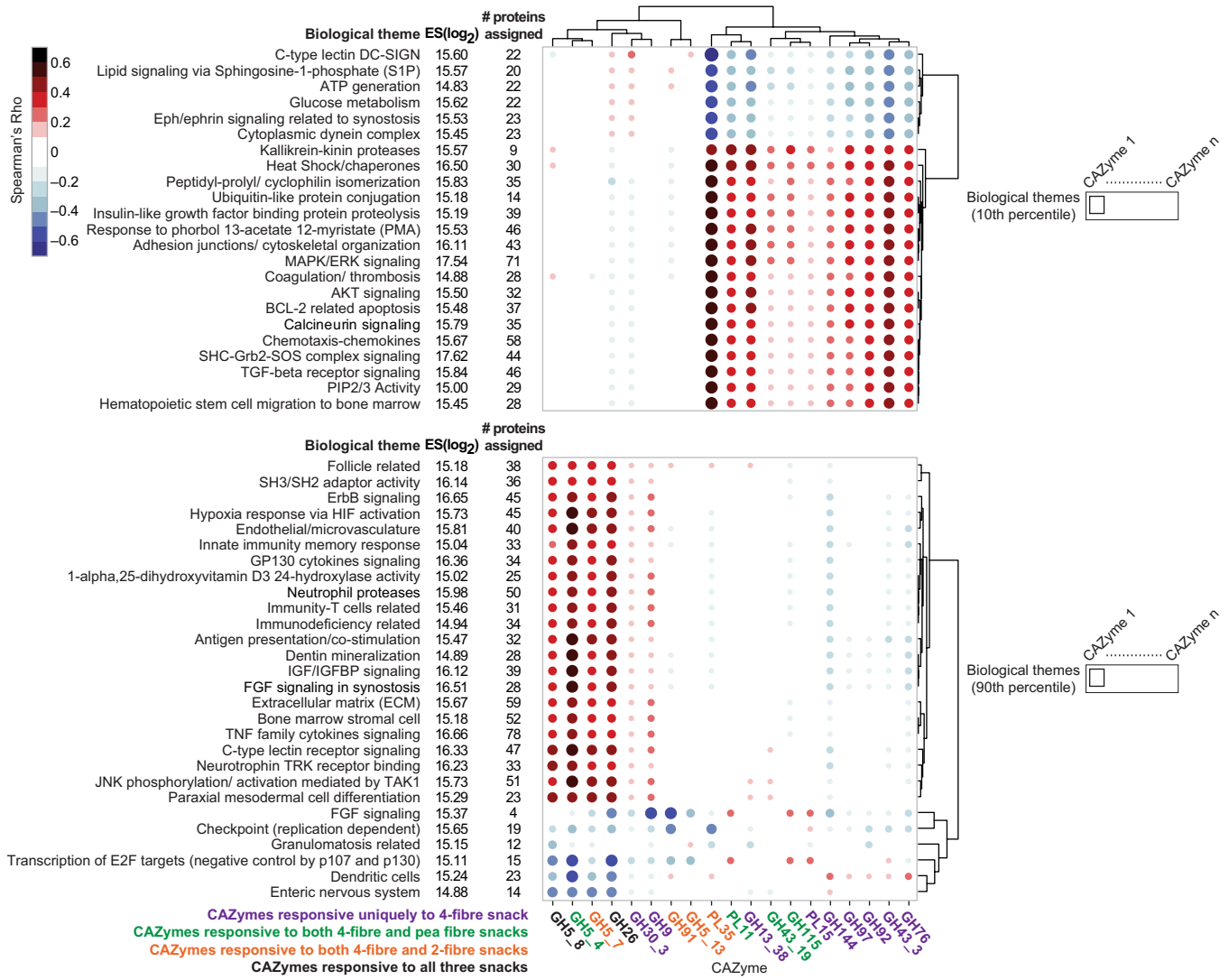


e Biological themes

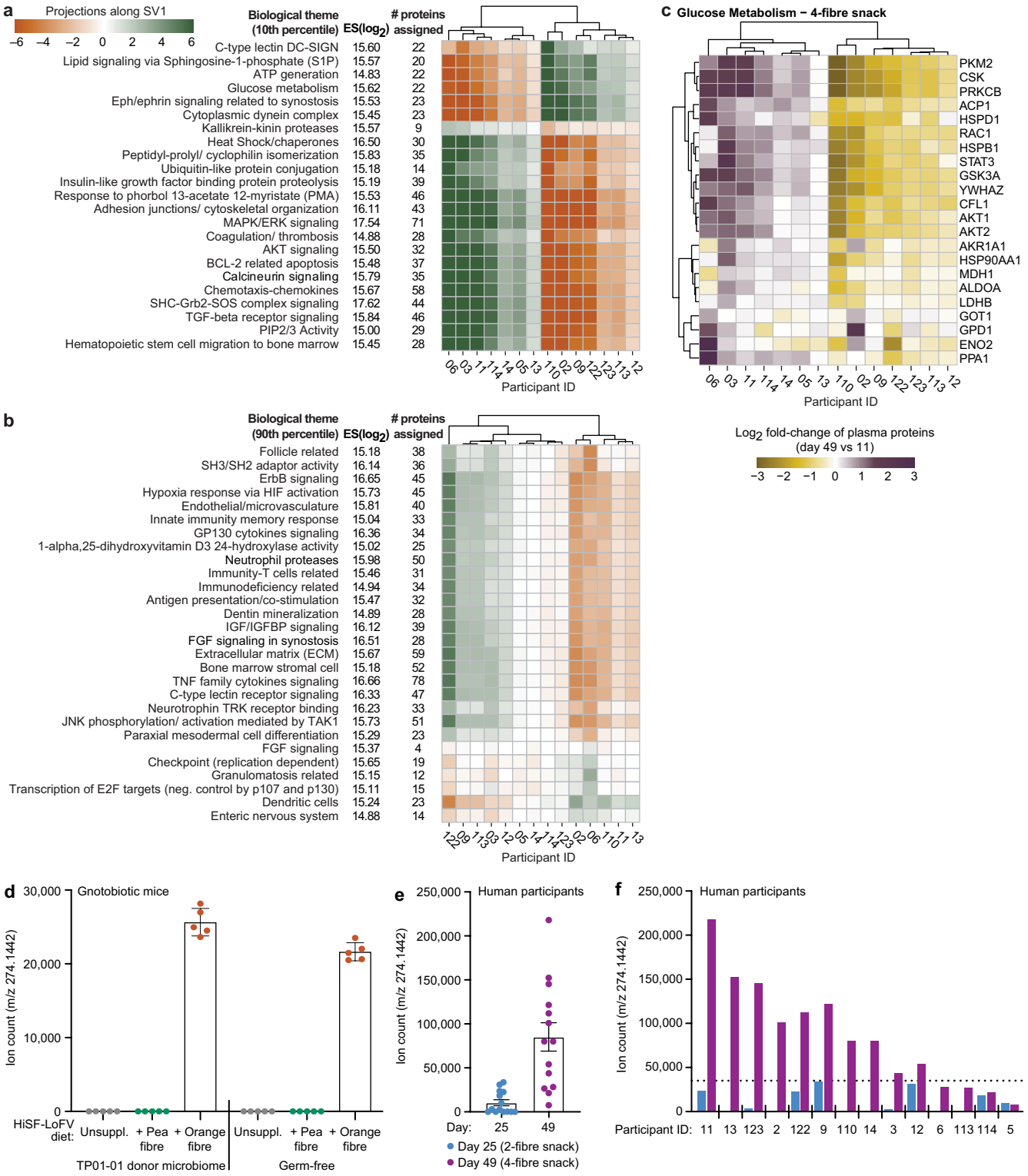


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | CAZyme-associated plasma proteome responses to consumption of the four-fibre snack prototype. a–**c**, Contextual language processing literature analysis (CompBio) of proteins with abundances that were significantly correlated with treatment-discriminatory CAZymes in participants consuming the four-fibre snack. The analysis procedure is summarized in Extended Data Fig. 8. Treatment-responsive proteins, identified by CC-SVD as having projections at the extremes of SV1 (10th and 90th percentiles of the distribution), are grouped into biological themes identified by CompBio, on the basis of a conditional probability analysis, as being significantly enriched for contextually associated biological concepts (processes or pathways) over those that occur by random sampling of the literature. Themes with enrichment scores (log_2-transformed) > 14.8 in the plasma proteomes of participants who consumed the four-fibre snack are shown in **b**, **c** (Supplementary Table 10d–f provides a comprehensive list of themes associated with this and the other fibre snacks). **d**, **e**, Biological themes based on proteins positioned in 10th and 90th percentiles (**d** and **e**, respectively) are portrayed as spheres. The size of a sphere is related to its enrichment score in the plasma proteome after consumption of the four-fibre snack. The thickness of the blue lines connecting themes signifies the number of proteins shared between them. Component proteins of exemplary themes (orange spheres) are listed in boxes and coloured by their median log_2-transformed fold change in response to consumption of the snack (blue, decrease; red, increase).

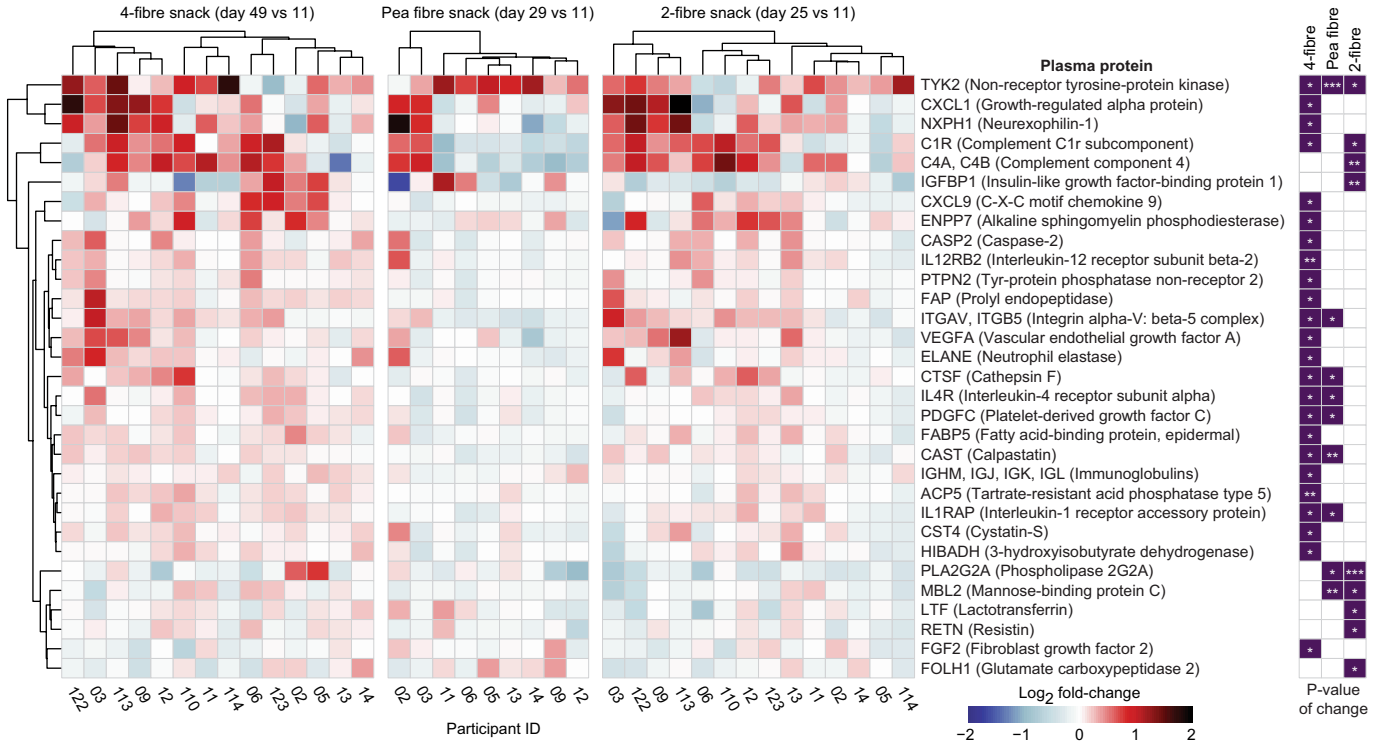


Extended Data Fig. 10 | Connecting host responses defined by plasma proteomic features to microbiome responses defined by CAZyme features in participants consuming the four-fibre snack. CC-SVD analysis of the plasma proteome with significantly changed discriminatory CAZymes in all participants after consuming the four-fibre snack. Two distinct groups of proteins with significant correlations with CAZymes are shown; one group with SV1 projections situated in the 10th percentile (top) and the other in the 90th percentile (bottom). CompBio analysis revealed biological themes that were significantly correlated with these CAZymes. Themes, their enrichment scores (log_2-transformed), the number of proteins comprising each theme and the cross-correlation (Spearman's rho) values between SV1 projections of themes and discriminatory CAZymes are presented. Each circle represents the correlation between a biological theme and a fibre-responsive CAZyme, with larger and darker circles indicating stronger correlations (positive correlations are coloured in red; negative correlations are coloured in blue).



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Individual responses of the plasma proteome of participants consuming the four-fibre snack prototype. a, **b**, Heat maps plotting the projections on SV1 of changes in the representation of biological themes during consumption of the four-fibre snack. **c**, Heat map plotting the $\log_2$-transformed fold change in the levels of plasma proteins enriched in the glucose metabolism theme. Data for each participant are shown, normalized to the last day of the pretreatment phase on day 11. The four-fibre snack produced the greatest reduction in HOMA-IR among the three different snacks tested (Supplementary Table 5c). However, this reduction did not achieve statistical significance ($P = 0.078$, linear mixed-effects model) after the short period of snack consumption in this study. **d**–**f**, LC-QTOF-MS analysis of a biomarker of orange fibre consumption present in faecal samples obtained from gnotobiotic mice and humans. **d**, Comparison of levels of the m/z 274.1442 analyte in colonized and germ-free mice fed the unsupplemented, orange-fibre-supplemented or pea-fibre-supplemented HiSF–LoFV diet for 10 days. The analyte is detectable only when orange fibre is consumed and is not dependent upon on the human donor microbiome for its generation. Bars represent mean values ± s.d. for biological replicates ($n = 5$ mice per group). **d**, **e**, Comparisons of levels of the analyte in faecal samples obtained from participants in human study 2 on days 25 and 49 when they were consuming the maximum dose of the two-fibre (pea and inulin) and four-fibre (pea fibre, inulin, orange fibre plus barley bran) snack food prototypes. The bar graph in **e** represents mean values ± s.d. for technical replicates ($n = 14$ participants). The difference documented between participants consuming the 2- versus 4-fibre snacks is statistically significant ($P = 0.0004$, paired two-tailed t-test). The horizontal dashed line in **f** denotes a baseline value operationally defined as the highest level of detection of the analyte in participants consuming the two-fibre snack food prototype lacking orange fibre.



Extended Data Fig. 12 | Plasma proteins with statistically significant changes in their abundances as a function of fibre treatment type and participant. Heat map plotting the $\log_2$-transformed fold change in the abundances of plasma proteins in participants consuming the indicated fibre snack prototype. Data from the nine participants in study 1 (pea-fibre snack) who were also enrolled in study 2 (two- and four-fibre snacks) are shown. Changes in protein levels are referenced to their abundances on the last day of the pretreatment phase (day 14 and day 11 in study 1 and study 2, respectively) ($n = 66$ blood plasma samples analysed). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ (linear model, limma[25]).

nature research

Corresponding author(s): Jeffrey I. Gordon

Last updated by author(s): May 20, 2021

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	DADA2 (v. 1.13.0) for pre-processing of V4-16S rRNA amplicon reads; GreenGenes 2016 (v. 13.8), RDP 16 (release 11.5), SILVA (v. 128), and NCBI BLAST toolkit (v. 2.10.0) for alignment and taxonomic assignment of amplicon sequence variants (ASVs) generated using DADA2. For demultiplexed shotgun sequencing reads, Cutadapt (v. 1.9.1) was used for trimming of adapters; Sickle (v. 1.33) for quality filtering; Bowtie2 (v. 2.3.5) for removal of human and mouse DNA sequences; IDBA-UD (v. 1.1.3) and MEGAHIT (v. 1.1.4) for assemblies; and Prokka (v. 1.13.7) for annotation. Duplicate reads were removed [Picard MarkDuplicates tool (v. 2.9.3)]. Count data of mapped reads were generated with featureCounts (Subread v. 1.5.3). DIAMOND (v. 0.9.34) for best-hit annotations with mcSEED subsystems/pathway modules. Blastp (v. 2.3.0+) and HMMER3 (v. 3.3) were used for assignment of Prokka-annotated open reading frames (ORFs) to CAZyme families using the CAZy database. R (v. 3.6.0) and MATLAB (v. 2019b) was employed for data analysis. Changes in the plasma proteomes as a function of fiber snack consumption were defined using a knowledge generation engine (CompBio V2.0, PercayAI Inc., St. Louis, MO) as well as limma (v.3.42.2). The resulting cross-correlation matrix, which contains Spearman Rho values as elements, was plotted as a heatmap using the ggplot package (v. 3.3.2) in R (v. 3.6.0). Code for HOSVD (CP-ALS plus randomization code) and CC-SVD is available on Zenodo (https://doi.org/10.5281/zenodo.476788). Packages within R for data analysis can be found in Methods. The mcSEED database is accessible at www.theseed.org.

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

V4-16S rRNA sequences in raw format prior to post-processing and data analysis, plus shotgun sequencing datasets generated from faecal DNA, have been deposited at the European Nucleotide Archive under study accession PRJEB38148. Raw proteomic datasets generated from the aptamer-based 1.3K SomaLogic platform have been deposited in the European Genome-Phenome Archive (EGA) under accession IDs EGAD00010002133 (Pea Fiber study) and EGAD00010002132 (Fiber Blends study). There are no restrictions on data availability.

Field-specific reporting

☒ 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	No formal power calculation was performed for the mouse or human studies. The number of animals studied per treatment group was based on our previous knowledge of the reproducibility of colonization of recipient germ-free mice with a given human donor microbiome and our cited Patnode et al. (2019) reference which characterized the effect size of the selected fibers on a defined consortium of cultured human gut bacterial strains, including members of Bacteroides. Each animal in each treatment group served as its own control in our diet oscillation protocol (see section on Replication below). The two human studies we report were not randomized controlled trials. Instead, they were open label single arm studies in which all participants received the same fibre-based food supplement(s), in the same order, for the same periods of time while they consumed a high saturated fat, low fruits and vegetable (HiSF-LoFV) diet. Each subject served as his/her own control. Since there was little precedent for controlled diet studies of this type examining the effects of single- and multi-fibre supplemented snack prototypes on gut microbiome composition and/or host physiology, a formal power calculation was not possible. However, since each subject served as his or her own control with repeated microbiome sampling, we aimed to have a minimum of 12 participants complete each of these studies.
Data exclusions	For gnotobiotic mouse studies: We prioritized samples were both V4-16S rDNA amplicon and whole community shotgun sequencing datasets had been generated: faecal samples collected from mice on experimental days 4, 9, 19, 29, 39, 49, 59 were excluded because shotgun sequencing datasets were not obtained from these samples. DNA sequencing datasets generated from mice colonized with the faecal microbial community of human donor TP01-01 and monotonously fed the HiSF-LoFV diet for the duration of the experiment were not included in our HOSVD analyses because we focused on characterizing the effects of fibre supplementation of the HiSF-LoFV diet with either of the three plant-derived fibers in the diet oscillation study (the latter is described in detail in the Methods section). For human studies: We focused on DNA sequencing datasets obtained from samples collected at those time points where all participants produced a faecal sample. We excluded samples collected on days when participants were on a free diet (beginning of each study); we were focused on characterizing changes in the gut microbiome as a function of fiber snack supplementation of the controlled HiSF-LoFV diet and not the changes that occur as they transitioned from their free diet to the control diet.
Replication	Six to 10 germ-free mice were colonized with a given donor's microbiome (n=9 different obese donors). Replication was successful.
Randomization	Before colonization with human faecal microbial communities, germ-free mice were randomized with respect to litter of origin but were matched by starting weight across treatment groups. The diet oscillation protocol allowed us to use each mouse as its own control. The human studies were single group, open label trials and therefore were not randomized.
Blinding	Mouse experiments: all data were generated from samples without knowledge of their treatment group of origin. The human studies were single group, open label trials and therefore were not blinded. Primary generation of data from human biospecimens was performed by individuals who were not involved in study design or interpretation of the results, without knowledge of the stage in the study from which the biospecimens were collected.

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

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Germ-free male C57BL/6J mice (12-16-weeks-old). Animals were dually housed in plastic cages located in plastic flexible film gnotobiotic isolators. The gnotobiotic facility was maintained at 23 °C and 40% humidity under a 12 h light cycle (lights on at 0600 h).
Wild animals	No wild animals were used in the study
Field-collected samples	No field collected samples were used in the study
Ethics oversight	All mouse experiments were carried out using protocols approved by the Institutional Animal Care and Use Committee (IACUC) of Washington University in St. Louis.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Overweight and obese men and women (BMI ≥25.0 and ≤35.0 kg/m2), aged ≥18 and ≤60 years were enrolled in the two, single-arm pilot studies. Exclusion criteria included: (i) previous bariatric surgery; (ii) significant organ system dysfunction (e.g., diabetes, severe pulmonary, kidney, liver or cardiovascular disease); (iii) cancer or cancer that has been in remission for <5 years; (iv) major psychiatric illness; (v) inflammatory gastrointestinal disease; (vi) pregnant or lactating women; (viii) use of medications that are known to affect the study outcome measures and that cannot be temporarily discontinued for this study; (ix) use of medications known to affect the composition of the gut microbiota within the last 30 days (most notably antibiotics); (x) bowel movements <3 times per week; (xi) vegans, vegetarians, those with lactose intolerance and/or severe allergies/ aversions/sensitivities to foods and ingredients included in the prescribed meal plan; (xii) persons who are not able to grant voluntary informed consent; and (xii) persons who are unable or unwilling to follow the study protocol or who, for any reason, the research team considers not an appropriate candidate for this study, including non-compliance with screening appointments or study visits. Women of child-bearing age were required to have a pregnancy test; if positive, they were excluded.
Recruitment	The Washington University Center for Human Nutrition maintains a database of ~4500 individuals who had previously been consented for other studies. This database was queried to identify individuals who met our inclusion criteria. A member of the research team then telephoned potential participants to tell them about the research study and determine whether they would be interested in participating. If a potential participant expressed interest, a member of the research team obtained verbal consent and then conducted a telephone screen. If interest/eligibility was confirmed, the research team member mailed or e-mailed the individual a consent document and scheduled a screening/first study visit. At this visit, study procedures and details of the consent document were fully explained; any questions were addressed and written informed consent was then obtained. Nine of the 14 participants enrolled in Study 2 had participated in Study 1.
Ethics oversight	Washington University in St. Louis IRB (FWA00002284)

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

Clinical data

Policy information about clinical studies

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

Clinical trial registration	ClinicalTrials.gov Identifiers: NCT04159259, NCT04101344
Study protocol	Available
Data collection	Study 1 was performed between February and July, 2019. Study 2 was conducted between August and December, 2019. All study participants were recruited from the St. Louis Missouri metropolitan area. Faecal samples were collected at home by participants and shipped in a frozen state to Washington University School of Medicine. Body weight was measured daily using smart

scales. Self-reported gastrointestinal symptoms and medication histories were obtained. Phlebotomy was performed at the study center for plasma proteomic analysis and measurement of standard clinical chemistry analytes.

Outcomes

1. Changes in the faecal microbiomes (representation of microbial genes encoding CAZymes and components of mcSEED metabolic pathways) and faecal microbiota (representation of bacterial taxa) in study participants before, during and after transitioning from a HiSF-LoFV diet to the same diet supplemented with one of three fiber snack prototypes for a period of 10-21 days.

2. Changes in host biological state (defined by changes in the plasma proteome) of study participants before, during and after transitioning from a HiSF-LoFV diet, to the same diet supplemented with one of three fiber snack prototypes, for a period of 10–21 days. A total of 1305 plasma proteins were quantified using an aptamer-based assay (SOMAscan).

3. Relationship between fiber-specific changes in the abundances of genes in participants' microbiome [notably, levels of genes encoding carbohydrate active enzymes (CAZymes) and components of various metabolic pathways] and alterations in the plasma proteome; these relationships were determined using cross-correlation singular value decomposition (CC-SVD) and a knowledge generation algorithm (CompBio).