nature microbiology

LETTERS

PUBLISHED: 28 NOVEMBER 2016 | VOLUME: 2 | ARTICLE NUMBER: 16221

Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome

Antoine M. Snijders[1†], Sasha A. Langley[1†], Young-Mo Kim[2†], Colin J. Brislawn[2], Cecilia Noecker[3], Erika M. Zink[2], Sarah J. Fansler[2], Cameron P. Casey[2], Darla R. Miller[4], Yurong Huang[1], Gary H. Karpen[1,5], Susan E. Celniker[6], James B. Brown[6], Elhanan Borenstein[3,7,8], Janet K. Jansson[2*], Thomas O. Metz[2*] and Jian-Hua Mao[1*]

Although the gut microbiome plays important roles in host physiology, health and disease[1], we lack understanding of the complex interplay between host genetics and early life environment on the microbial and metabolic composition of the gut. We used the genetically diverse Collaborative Cross mouse system[2] to discover that early life history impacts the microbiome composition, whereas dietary changes have only a moderate effect. By contrast, the gut metabolome was shaped mostly by diet, with specific non-dietary metabolites explained by microbial metabolism. Quantitative trait analysis identified mouse genetic trait loci (QTL) that impact the abundances of specific microbes. Human orthologues of genes in the mouse QTL are implicated in gastrointestinal cancer. Additionally, genes located in mouse QTL for Lactobacillales abundance are implicated in arthritis, rheumatic disease and diabetes. Furthermore, Lactobacillales abundance was predictive of higher host T-helper cell counts, suggesting an important link between Lactobacillales and host adaptive immunity.

To decipher the respective contributions of host genetics, early life history and diet on the gut microbiome we leveraged 30 independent, genetically distinct Collaborative Cross (CC) mouse strains (Fig. 1a and Supplementary Table 1), a large multi-parental panel of recombinant inbred strains with defined single nucleotide polymorphisms (SNPs) that captures ~90% of the known variation in laboratory mice. Sixteen strains were maintained in a specific pathogen-free (SPF) facility (Built Environment 1, BE1), and 14 additional strains were maintained in a barrier facility that screens for additional infectious agents, including *Pasteurella pneumotropica* and *Helicobacter* (Built Environment 2, BE2). Mice were fed the same water and food sources at both locations. Faecal samples were collected at 12 weeks of age (Fig. 1a) and the gut microbiome composition was characterized by sequencing 16S rRNA genes (V4 hypervariable region; Supplementary Table 2).

Unsupervised hierarchical clustering of the 300 most abundant operational taxonomic units (OTUs) revealed two main clusters, each associated with a specific BE (Supplementary Fig. 1a), indicating a strong effect of BE on microbiome composition. We observed differences in the relative abundances of specific microbial families (Fig. 1b–d and Supplementary Table 3). Specifically, there were higher relative abundances of Alcaligenaceae, Verrucomicrobiacae, Erysipelotrichaceae and Deferribacteracea and lower relative abundances of Clostridiales in BE1 compared to BE2. The consistency of BE-specific microbial signatures across genetically diverse CC strains strongly suggests that the BE influence on the gut microbe composition is at least in part independent of genetic background.

Mice from the same 30 CC strains were then transferred to a third SPF facility (BE3) to investigate the stability of the gut microbiome in response to a new environment where all mouse strains were subjected to the same conditions of husbandry. Faecal samples were collected from mice at 2, 4, 6 and 8 weeks after arrival at BE3, and the faecal microbiome was profiled by 16S sequencing (Fig. 1a). Principal coordinate analysis (PCoA) using Bray–Curtis distance revealed that the microbiome was stable and remained largely defined by the source BE, even after 8 weeks in BE3 (Fig. 1e and Supplementary Fig. 1b). To further assess the persistence of the source building effect on the microbiome, we performed 16S rRNA gene sequencing of faecal samples from eight CC strains born at BE3 (that is, second generation). PCoA confirmed that mice born at BE3 maintained their parents' source building microbial signature (Fig. 1e). We conclude that the gut microbiome, at least partially shaped by early life history, is persistent and shared between parents and their offspring, even when challenged with a new environment. Future studies need to be conducted to investigate the stability of the source building microbial signature across multiple generations.

One of our main aims was to determine the influence of host genetics on the gut microbiome. Hierarchical clustering of OTUs revealed that the majority of samples collected from the same strains of mice at different time points clustered together, suggesting that host genetics plays a role in determining the gut microbial composition (Supplementary Fig. 2). To identify genetic loci associated with specific OTUs, we performed independent quantitative trait loci (QTL) analyses by interrogating 50,107 SNPs across the genome (Supplementary Table 4). This analysis identified 169 joint QTL intervals that were significantly associated with the abundances of ten or more OTUs ($-\log_{10}(P \text{ value}) > 6$) (Fig. 2a and

[1]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [2]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. [3]Department of Genome Sciences, University of Washington, Seattle, Washington 98105, USA. [4]Systems Genetics Core Facility, Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [5]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [6]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [7]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA. [8]Santa Fe Institute, Santa Fe, New Mexico 87501, USA. †These authors contributed equally to this work. *e-mail: janet.jansson@pnnl.gov; thomas.metz@pnnl.gov; jhmao@lbl.gov

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

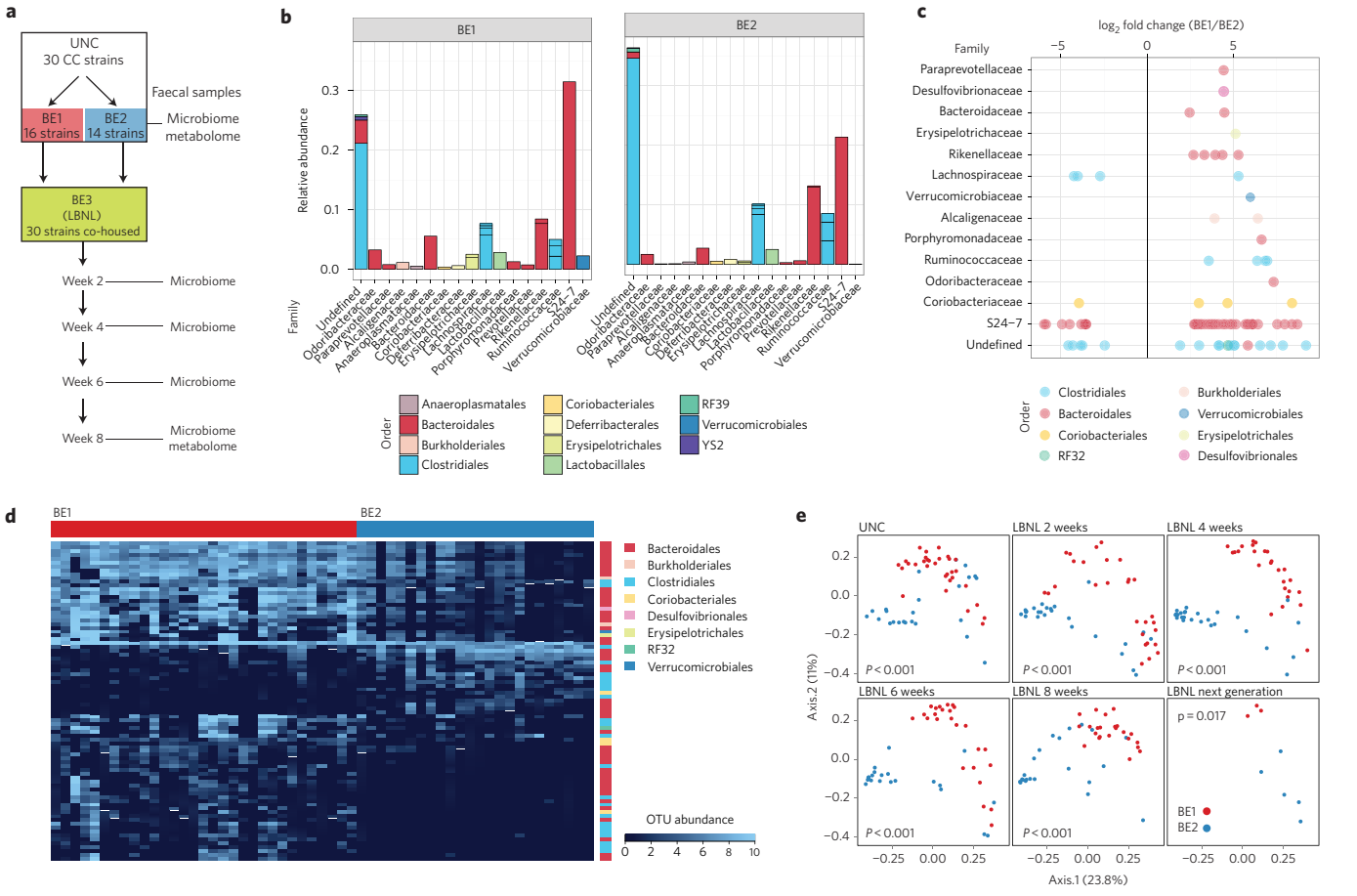

Figure 1 | Early life environment determines gut microbiome structure. a, Schematic of the study design. **b**, Normalized relative abundance of the most common genera in the two built environments (BEs), coloured by order and separated at family level. **c**, Differentially abundant faecal OTUs between animal facility BEs (BE1 versus BE2 at UNC; $\alpha = 0.01$) **d**, Heatmap of differential abundance of taxa between BE1 and BE2 across 30 mouse strains. **e**, The distinct microbiome established at birth (BE1 and BE2) is sustained after transfer to a novel environment (BE3) and passed to the second generation born in BE3. Samples are colour coded by BE at UNC (red, BE1; blue, BE2) in this multidimensional scaling ordination of Bray–Curtis distances between normalized samples. The ADONIS test was used to assess the statistical significance of clustering based on BE ($P < 0.001$, $R^2 = 0.09617$ blocking by time point); the built environment BE effects were also significant when tested separately at each time point (all $P < 0.02$).

Supplementary Table 5) and revealed a complex host genetic architecture of the gut microbiome composition (Supplementary Fig. 3). These genetic linkages were predominantly driven by the most abundant representative in the OTU data—a member of the Clostridiales (family unknown, Fig. 2a, green track). Abundances of other bacterial families were also controlled by multiple genetic loci (Fig. 2 and Supplementary Table 6). Interestingly, the major histocompatibility complex (MHC) locus on chromosome 17 was significantly ($P < 0.0001$) linked to the abundance of Lactobacillaceae (Fig. 2b,c), consistent with an earlier report showing that MHC variation shapes microbial communities in the mouse gut, in particular the genus *Lactobacillus*[3]. Our findings also support earlier reports showing an association of the gut microbiota with host genetic variations[4–7]. However, on leveraging the CC mice we identified over a hundred novel genetic loci that impact the gut microbiome.

To investigate the association of the gut microbiome with host phenotypes and behaviour, we measured body weight, rotarod performance and immune cell abundance in the mice. Random forest analysis indicated that Lactobacillaceae abundance was predictive of T cell counts in peripheral blood (Fig. 3a; adjusted $P = 0.02$), driven predominantly by T-helper cell levels (adjusted $P = 0.00087$) but not T-suppressor cell levels (Fig. 3a). Modest associations were found for B-cell counts, body weight and rotarod performance (Supplementary Fig. 4). These results are consistent with reports that (1) *Lactobacillus* consumption is associated with an increase in CD4 counts in patients with HIV[8,9], (2) *Lactobacilli* can regulate behaviour in mice[10] and (3) *Lactobacilli* can serve as natural enhancers of cellular immune responses[11]. Our results, using a non-targeted approach to assess the microbiome, extend these findings by demonstrating that only *Lactobacilli* have statistically significant associations with T cell counts, emphasizing the importance of *Lactobacilli* for health of the host.

The QTL that were specifically associated with Lactobacillaceae abundance in mice displayed significant enrichment for genes implicated in autoimmune disorders such as diabetes and arthritis (Fig. 3c). Examples of candidate genes in QTL linking *Lactobacillus* QTL with human phenotypes (Fig. 2b) include *Prospero Homeobox 1* (*Prox1*) on chromosome 1 (associated with type 2 diabetes, obesity and fasting glucose levels), *Catenin Alpha 3* (*Ctnna3*) on chromosome 10 (associated with serum pyroglutamine metabolite levels and arrhythmogenic right ventricular dysplasia, familial 13) and *Insulin Like Growth Factor 2 MRNA Binding Protein 2* (*Igf2bp2*), *Transformer 2 Beta Homolog* (*Tra2b*) and *ST6 Beta-Galactoside Alpha-2,6-Sialyltransferase 1* (*St6gal1*) on chromosome 16 (associated with type 2 diabetes and colon adenocarcinoma and colorectal cancer)[12–16]. These results suggest an important role for host regulation of Lactobacillaceae abundance in health and disease and are concordant with recommendations for

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

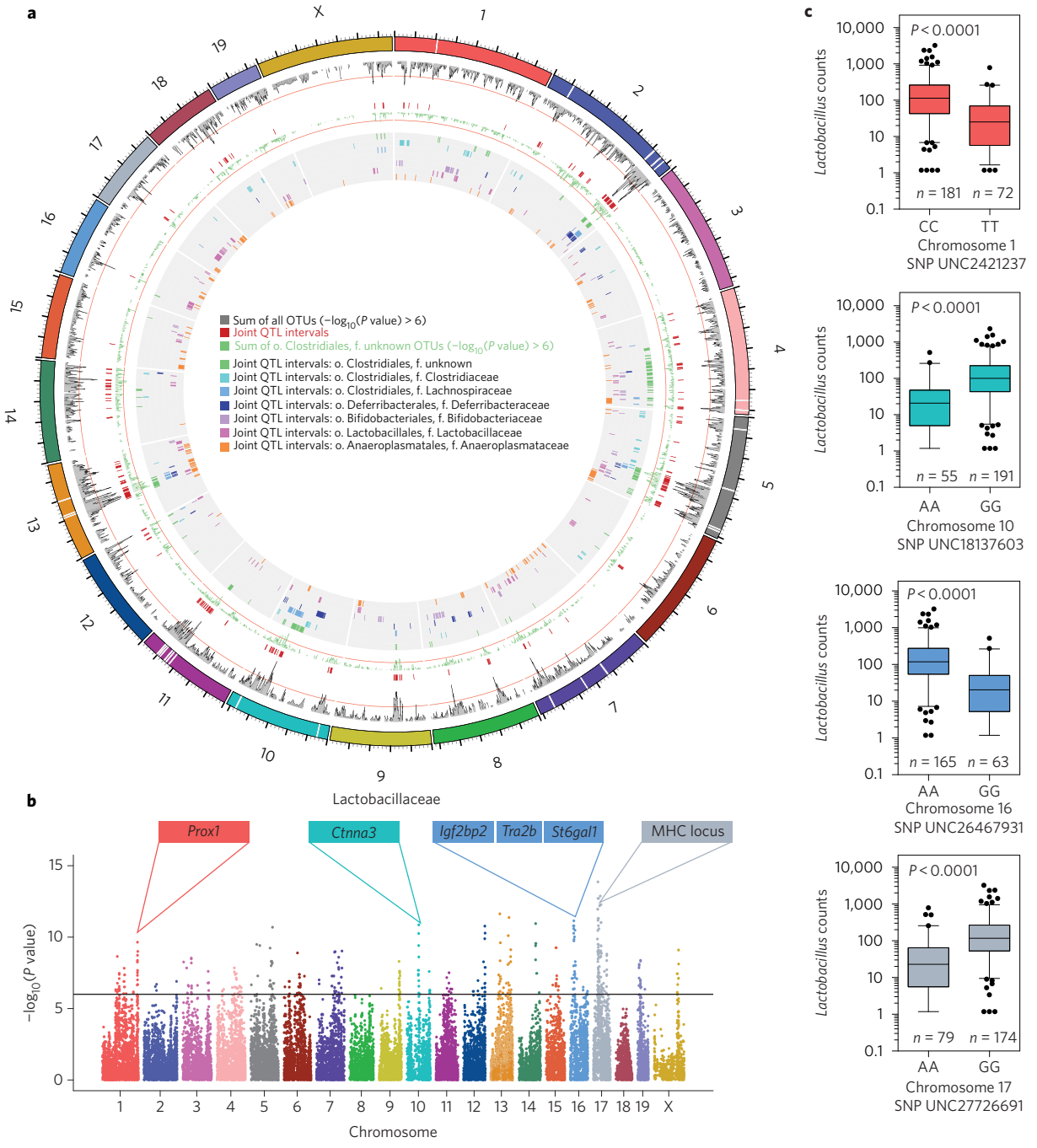

Figure 2 | A GWAS identifies host genetic loci that impact gut microbiome composition and abundances. a, Genomic architecture of QTL for gut microbiome composition. The outer layer shows chromosome location (each chromosome is uniquely coloured and labelled; major tick marks within each chromosome arm correspond to 25 Mb). The second layer (grey) shows the number of OTUs at each SNP that reach QTL significance (Mann–Whitney *U* test, $P \leq 1 \times 10^{-6}$). The third layer (red) shows the genomic intervals based on QTL significance for ≥10 OTUs. The fourth layer (green) shows the number of OTUs at each SNP that reach QTL significance for Clostridiales (family (f.) unknown), with the red line indicating 10 OTUs. The 5th–11th layers represent genomic intervals based on QTL significance for Clostridiales f. unknown, Clostridiaceae and Lachnospiraceae, Deferribacterales f. Deferribacteraceae, Bifidobacteriales f. Bifidobacteriaceae, Lactobacillales f. Lactobacillaceae and Anaeroplasmatales f. Anaeroplasmataceae, respectively. **b**, Manhattan plot of the GWAS analysis. OTUs are merged at the family level for Lactobacillaceae with the *x* axis showing genomic location and the *y* axis showing the association level. The $-\log_{10}(P \text{ value})$ is shown for 20,199 SNPs ordered by genomic position. The horizontal black line indicates the QTL significance threshold at $-\log_{10}(P \text{ value}) = 6$. Candidate genes located in representative QTL are listed above the plot. **c**, SNP-specific association with Lactobacillaceae abundance for examples on chromosomes 1 (191,968,724 bp), 10 (66,380,845 bp), 16 (22,739,388 bp) and 17 (27,818,729 bp) (whiskers represent 5th and 95th percentiles).

the use of probiotics containing *Lactobacillus* species as adjunctive therapies for the treatment of rheumatoid arthritis and diabetes[17,18].

Candidate genes located within the boundaries of the 169 identified QTL (2,699 genes; Supplementary Table 5) were analysed to assess additional human relevance. We found that genes controlling the abundance of specific members of the microbiome were significantly enriched in human gastrointestinal cancer ($1.02 \times 10^{-7} < P < 3.17 \times 10^{-19}$), inflammatory responses ($1.75 \times 10^{-3} < P < 7.15 \times 10^{-6}$)

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

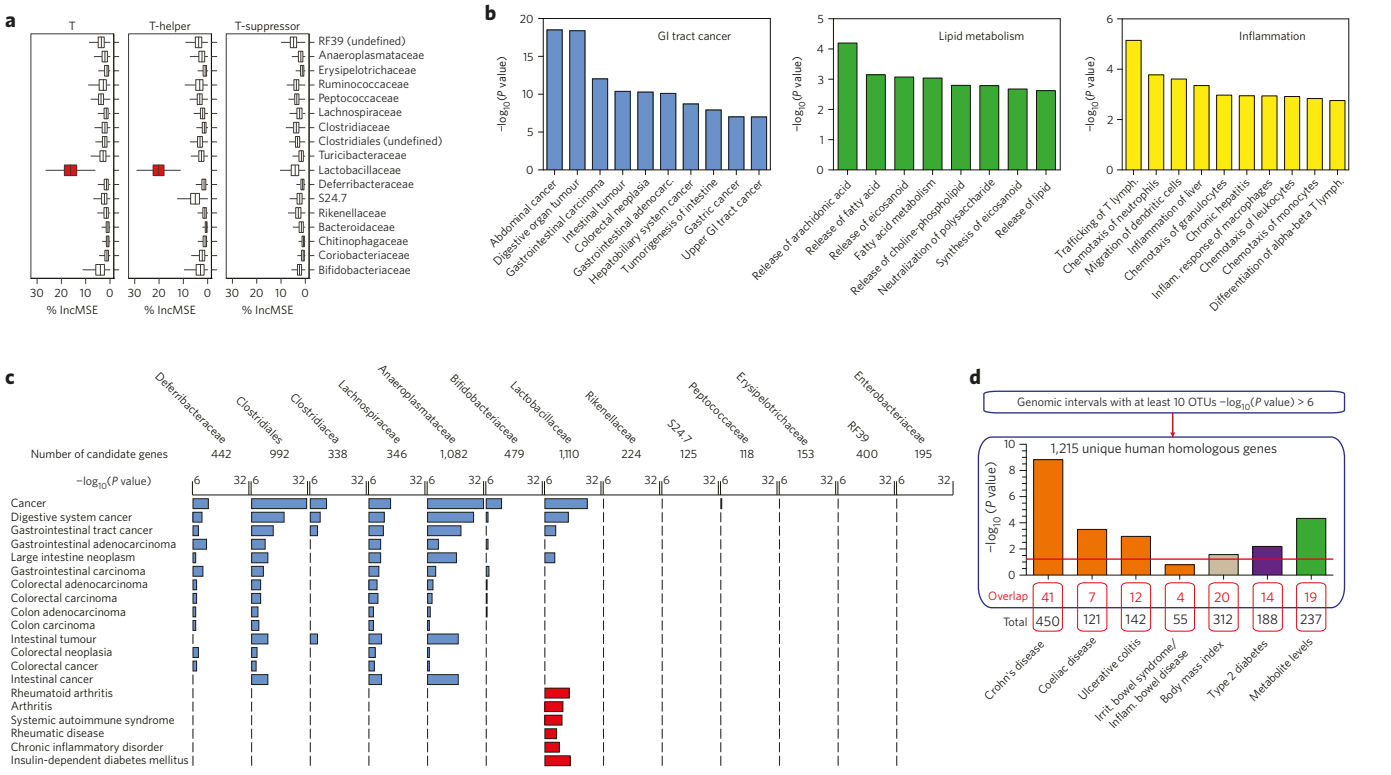

Figure 3 | Association of microbial abundance with host phenotypes and their implications for human disease. a, Random forest analysis to assess the association between microbial abundance at the family level and mouse peripheral blood T ($CD3^+/CD45R^-$), T-helper ($CD3^+/CD45R^-/CD4^+/CD8^-$) and T-suppressor ($CD3^+/CD45R^-/CD4^-/CD8^+$) cell counts. Significant associations are indicated in red ($P < 0.05$). **b**, Human homologues of candidate genes within joint QTL intervals defined in Fig. 2a (third layer) are significantly enriched for genes implicated in gastrointestinal (GI) tract cancer, lipid metabolism and immune system functions (ingenuity pathway analysis, IPA). **c**, Human homologues of candidate genes in QTL for Deferribacteraceae, Clostridiales f. unknown, Clostridiacea, Anaeroplasmataceae, Bifidobacteriaceae, Lachnospiraceae and Lactobacillaceae are significantly enriched for GI tract cancer (blue bars indicate $-\log_{10}(P \text{ value}) > 6$), while remaining families do not show significant enrichment. Candidate genes in QTL for Lactobacillaceae are also significantly enriched for arthritis, diabetes and chronic inflammatory disorder (red bars). **d**, Human homologues of candidate genes within joint QTL intervals defined in Fig. 2a (third layer, 1,215 unique human homologues) showed significant overlap with GWASs for Crohn's disease (41 genes), coeliac disease (7 genes), ulcerative colitis (12 genes), body mass index (20 genes), type 2 diabetes (14 genes) and metabolite levels in blood serum or cerebral spinal fluid (19 genes). Human gut-related diseases are indicated in orange. Horizontal red line indicates significance threshold for overlap at $-\log_{10}(P \text{ value}) > 1.3$.

and lipid metabolism ($2.38 \times 10^{-3} < P < 6.35 \times 10^{-5}$) (Fig. 3b), providing further evidence for the involvement of both host genetics and the gut microbiome in health and disease. Genome-wide association (GWA) analysis was used to identify the genetic loci associated with abundance of microbial taxa at the family level. We found 13 of the taxa associated with QTL that contained >100 genes (Supplementary Table 7), of which 7 were significantly enriched for human genes implicated in gastrointestinal tract cancer (Fig. 3c). Comparing the mouse genes to a previously compiled list of human disease-related genes identified by GWA studies (GWAS)[19], a significant overlap was observed for genes associated with Crohn's disease, coeliac disease, ulcerative colitis and type 2 diabetes (Fig. 3d and Supplementary Table 8). We conclude that candidate mouse genes within the loci identified as controlling microbiome abundance exhibit significant overlap with human genes previously linked to disease states, suggesting that the microbiome may contribute to their aetiology.

Investigation of the faecal metabolite composition allowed us to determine the influence of early life environment and diet on the gut metabolome. For these analyses we focused on 24 CC strains that were housed in BE1 and BE2 (fed Diet 2, Labdiet Prolab 3500) and in BE3 (fed Diet 1, Labdiet Picolab 5053). Although the two diets have similar macronutrient compositions, the metabolite profiles are quite distinct (Fig. 4a, lower panel; Supplementary Table 9). Extracts from the stool samples were analysed by gas chromatography–mass spectrometry (GC–MS) and metabolites were identified by comparison to a reference library containing mass spectral and retention index information for over 850 metabolites[20]. A total of 122 unique metabolites were identified, including amino acids, sterols, mono- and disaccharides, glycolytic and tricarboxylic acid cycle intermediates, short- and long-chain fatty acids, and products of microbial metabolism. An additional 110 peaks were detected but not identified (Supplementary Table 10).

The metabolites significantly clustered by diet, with differences in relative abundances of proteinogenic amino acids, mono- and disaccharides, sterols and fatty acids driving the separation (Fig. 4a,b and Supplementary Fig. 5a). To validate that the gut metabolome is primarily influenced by diet, four CC strains were maintained on Diet 1 for one week, then on Diet 2 for one week, before switching back to Diet 1 for an additional week. Fresh faecal samples were collected at the end of each week for metabolome profiling (Supplementary Table 11). Although only subtle changes were observed in microbial abundance (Supplementary Fig. 5b; $P = 0.273$), there was a major and reversible shift in the metabolome profile that coincided with dietary changes (Fig. 4c), demonstrating that the metabolome profile is largely influenced by diet.

We used a metabolic modelling-based framework, MIMOSA[21], to identify metabolites whose variation across samples is explained by variation in the metabolic potential of the microbiome, based on

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

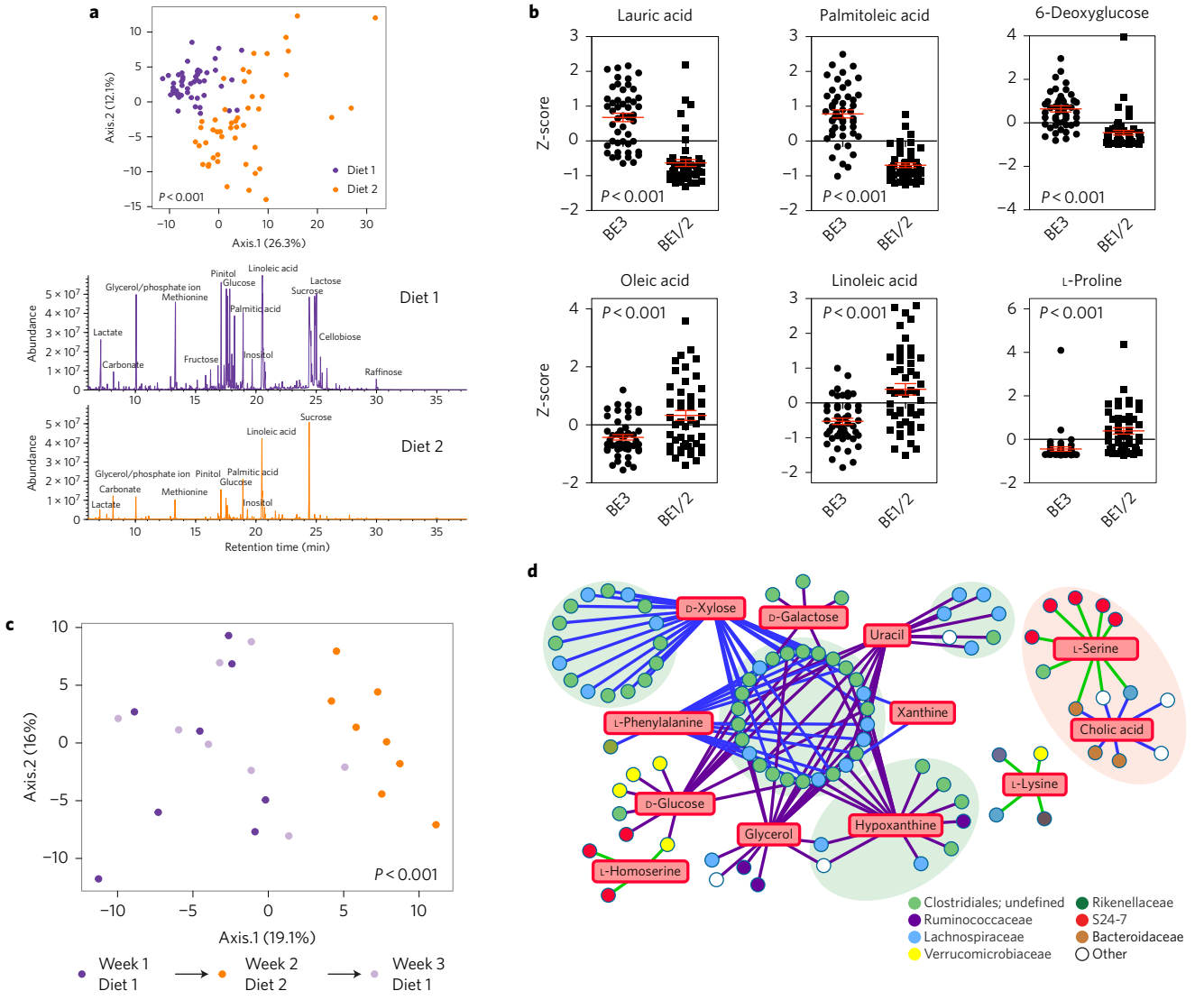

Figure 4 | Dietary and microbial influences on metabolite profiles. a, Top: diet is the main contributor to metabolite profiles. Bottom: GC–MS chromatograms of diet 1 and diet 2. PCoA of metabolite profiles were measured in faecal samples of 24 CC strains maintained on two different diets ($P < 0.001$, $R^2 = 0.16597$, ADONIS). **b**, Relative abundance of select metabolites in faecal samples from individual mice fed on different diets. Error bars indicate mean ± s.e.m. **c**, Diet is a primary contributor to metabolite profiles and correlates strongly with principal coordinate 1. Metabolite profiles were measured in faecal samples of four CC strains (males and females were analysed separately for each strain) maintained for one week on standard chow (Labdiet Picolab 5053; diet 1) or one week on autoclaved chow (Labdiet Prolab 3500; diet 2), followed by one week on standard chow ($P < 0.001$, $R^2 = 0.14237$ between two diets, ADONIS). **d**, Metabolic modelling-based taxonomic contributors to metabolite variation for mice on the autoclaved Labdiet Prolab 3500 chow (BE1 and BE2). Individual OTUs shown (circles; coloured at the family level) are those whose metabolic capacity and variation across samples are consistent with the metabolic potential of the entire community and with measured variation in the linked metabolites (squares). Green and orange clouds behind OTU sub-networks indicate Clostridiales and Bacteroidales enrichment. Edge colour indicates whether a given OTU potentially impacts a certain metabolite variation via synthesis (blue edges), degradation (green edges) or both (purple edges).

differences in species composition and estimated gene composition. By applying MIMOSA to the pooled set of metabolome samples from both diets, we found that variation in dietary metabolites (compounds detected in chow pellets by metabolomics) was poorly explained by microbial community composition (Supplementary Tables 9 and 12). However, the variation in a high proportion of non-dietary metabolites (47.6%; 10 out of 21 metabolites not detected in chow) was consistent with predicted community metabolic potential (CMP), suggesting a substantial role for microbial metabolism in metabolite synthesis and/or degradation (Supplementary Fig. 6a). Specifically, the observed variation in many gut metabolites was consistent with the predicted CMP, including hypoxanthine, L-homoserine, 5-hydroxyindoleacetate and cholate (Supplementary Figs 6 and 7). More metabolites varied consistently with predicted CMP in samples from the nutritionally simpler Diet 2, suggesting that the microbiome may have a larger and more direct impact on the faecal metabolome in this context. The predicted CMP was driven by the metabolic potential of a diverse set of taxa, including OTUs from the phyla Firmicutes, Bacteroidetes and Actinobacteria (Fig. 4d and Supplementary Figs 7 and 8). Interestingly, the measured concentrations of several metabolites present in one or both diets were negatively correlated with predicted CMP (mostly on the basis of microbial degradation enzymes; Supplementary Fig. 6b), indicating that food containing these metabolites could drive the expansion of microbes that use them efficiently. These findings highlight the combined impacts of diet and microbiome composition on the gut metabolome and the complex interactions between them.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Our studies using the CC mouse cohort and an integrated, systematic analysis paradigm revealed how gut microbiome composition and function are shaped by interactions between host genotype, early life environment and diet, and identified several host genetic loci that regulate microbial abundance. Using multivariate analysis we quantified the relative influence of environment and genetics on microbial abundance and determined that genetics plays a larger role than environment (Supplementary Fig. 9). This study provides a foundation for future investigations of how reciprocal interactions between host genotype, environmental factors, gut microbiome and metabolome compositions contribute to a wide spectrum of mammalian traits and disease susceptibility.

Methods

Mouse husbandry and faecal sample collection. Mice were obtained from the Systems Genetics Core Facility at the University of North Carolina (UNC)[22]. Before their relocation to UNC, CC lines were generated and bred at Tel Aviv University in Israel[23], Geniad in Australia[24] and Oak Ridge National Laboratory in the USA[25]. All studies were performed on young adult mice (age 9–15 weeks). For each of 30 strains (for strain information and number of replicate samples see Supplementary Table 1), two males and two females were housed separately and maintained on PicoLab Rodent Diet 20 (5053). The number of CC strains used is sufficient to detect genetic association. The investigators were not blinded in the analysis of the phenotypes because the correct genotype of CC mice was needed to perform genotype–phenotype and phenotype–phenotype association analysis. Mice from different strains were always housed in different cages. We observed a subtle change in microbial composition in samples collected 16 h after a cage change compared to <2 h. However, to collect sufficient mouse faecal material for combined microbiome and metabolomic analysis, all faecal samples were consistently collected from each cage, avoiding areas clearly contaminated with urine, 16 h after cage change at 2, 4, 6 and 8 weeks after arrival at Lawrence Berkeley National Laboratory (LBNL). All animal procedures were approved by the UNC Chapel Hill or LBNL Institutional Animal Care and Use Committees.

Faecal samples were stored at −80 °C for downstream metabolite and microbial analyses. Faecal samples from different strains were collected in the same way to avoid collection and storage biases. Genotyping data for CC mice were obtained from UNC (http://csbio.unc.edu/CCstatus/index.py).

Faecal samples were collected from a different cohort of genetically identical young adult mice at UNC Chapel Hill (maintained on Labdiet Prolab 3500) to determine the effect of environment on the faecal microbiome and metabolome. Faecal samples were then manually homogenized on ice with a micropestle, 0.25 g was used for DNA isolation, 0.05 g for metabolite extraction and the remainder stored at −80 °C.

Microbiome analyses. Genomic DNA was extracted from 0.25 g of the homogenized faecal samples using the PowerSoil DNA Isolation Kit (http://www.mobio.com/) according to the manufacturer's instructions. PCR amplification of the V4 region of the 16S rRNA gene was performed using the protocol developed by the Earth Microbiome Project (http://press.igsb.anl.gov/earthmicrobiome/empstandard-protocols/16s/) and described in ref. 26 using updated primers described in ref. 27. Amplicons were sequenced on an Illumina MiSeq using the 150 base pair (bp) MiSeq Reagent Kit v2 (http://www.illumina.com/) according to the manufacturer's instructions.

QIIME 1.9.1 was used to join, quality filter and demultiplex libraries from three MiSeq runs[28,29]. VSEARCH 1.1.3 was used to dereplicate, sort by abundance, remove single reads and then to cluster at 97% similarity. VSEARCH was also used to check these clusters for chimaeras and construct an abundance table by mapping labelled reads to chimaera-checked clusters[30–32]. Taxonomy was assigned to the centroid of each cluster using the Qiime script assign_taxonomy.py and the Greengenes database. The centroids were aligned to Greengenes with PyNast and a phylogenetic tree was constructed using FastTree[33–35].

Statistical analysis and visualization were performed in R using the packages Phyloseq, DESeq2 and ggplot2. Both Bray–Curtis distances and UniFrac distances were used to compare microbial communities[36–39].

QTL mapping. 16S data from 30 strains (253 samples) were used in the analysis. OTUs showing significant differences in abundance (t-test $P < 0.01$ or DESeq2 adjusted $P < 0.01$) based on source building were filtered from the data, leaving 644 OTUs (of a total of 3,786). These OTUs represented 15% of total sequencing data. Genetic association was assessed for each OTU separately, and OTUs were merged at the family level. Genotype data for 77,597 SNPs were obtained from the UNC Systems Genetics Core website (http://csbio.unc.edu/CCstatus/index.py) and filtered for minor allele frequency >4 of the 30 CC strains, leaving 50,107 SNPs. At each SNP, normalized OTU counts from CC samples were assigned to their respective alleles. We then used the Mann–Whitney U test[40] to test the significance of associations between OTU abundance and allele classes at each SNP. We used permutation to ascertain the significance of our results on an individual OTU basis to obtain a nonparametric estimate of the false discovery rate (FDR), as follows. For data combined at the family level, 15 OTUs in the upper and lower quintiles of P value sums across all SNPs (a proxy for signal of genetic association) and 15 OTUs with the lowest sums of P values, we performed 1,000 permutations of strain identifiers and then computed the same statistic at each SNP (Supplementary Fig. 10). This confirmed that a cutoff of $-\log_{10}(P \text{ value}) > 6$ was a conservative threshold with a genome-wide FDR of <1%.

QTL were defined by merging SNPs with $-\log_{10}(P \text{ value}) > 6$ within 1 Mb into multi-SNP intervals. Those with only one SNP were removed, and the remaining QTL boundaries were extended to the adjacent neighbouring SNPs. Putative candidate genes were defined as those genes (gencode.vM7[41]) partially overlapping with or contained within a QTL locus. The list of candidate genes was analysed using Ingenuity Pathway Analysis, converted into human homologues using the MGI homology resources[42] (downloaded October 2015) and compared to human GWAS downloaded from www.ebi.ac.uk/gwas (ref. 19). The significance of overlap between mouse and human candidate genes was calculated using ConceptGen (http://conceptgen.ncibi.org/core/conceptGen/index.jsp). Visualization of genetic association and QTL was performed in R using ggplot2 and ggbio and with Circos[39,43,44].

OTU association with host phenotypes. Whole blood was collected into ethylenediaminetetraacetic acid-coated tubes at 12 weeks of age in a cohort of 267 mice across 16 CC strains. Complete blood cell counts were acquired using a HemaVet950FS. Lymphocyte subpopulations were identified by fluorescence-activated cell sorting (FACS) using cell- specific markers for B cells, T cells, T-helper and T-suppressor cells. Antibodies (BD Biosciences) used for this analysis were rat anti-mouse CD3-PE, rat anti-mouse CC45R/B220 PerCP, rat anti-mouse CD8a antibody APC and rat anti-mouse CD4 antibody Alexa 488. The percentages of cells in blood were determined on a BD FACS Calibur (Becton Dickinson) and data were analysed with FlowJo software (Tree Star). Body weight and rotarod performance were measured as described previously[45] at 10 weeks of age for a cohort of 365 mice across 16 CC strains.

We modelled data collected in mouse strains as statistically exchangeable to enable analysis in cases where we collected phenotypic and normalized 16S data on mice from the same strain, but not the same mice. Random pairs of phenotypic and normalized 16S data (combined at the family level) were sampled 1,000 times and each subjected to random forest regression analysis (microbial abundances as predictors, phenotype as the response vector). Analysis was performed using the R randomForest implementation[46] (ntree = 1,000, all other parameters set to default). It is necessary to resample 1,000 times to model variance associated with random pairing under the exchangeability model. We then generated null distributions ($n = 1{,}000$) by permutation of strain identifiers (after sampling to reproduce paired data). For each taxonomic family, observed and null importance measures (%IncMSE: % increase in mean squared error) were compared to determine significance. P values were computed as the natural nonparametric estimate of the likelihood of the observed distribution under the permuted distribution. Specifically, for each observed score $u_i \in Q$ and the null distribution Q, we computed the rank of $u_i \in Q$, denoted $r_{Q,i}$ and the empirical quantile under the null $p_i = (1/n)r_{Q,i}$ where $n = 1{,}001$, then our final P value is given by $P - \text{value} = \sum_{i=1}^{n} (1/n)p_i$. Note that $P_i = (1/n)r_{Q,i}$. This P value has the desirable property that it is bounded below by the sample size simulated for the null (and of course bounded above by 1). Given that it is nonparametric, it is conservative. Tukey boxplots of %IncMSE were generated using the default method in ggplot2 (ref. 39).

To estimate the proportion of OTU variation explained by genetics and source BE, SNPs were selected where sufficient statistical power exists for modelling (with comparable allelic frequencies for source BE1 and BE2 ($0.4 \leq$ fraction of allele frequency in BE1 and BE2 ≤ 0.6) from joint QTL intervals). SNPs with ambiguous or heterozygous genotypes for any strain were filtered. For each interval, a representative SNP with the lowest sum of Mann–Whitney U P values (across OTUs) was selected. OTU counts were then modelled as a linear function of SNP genotype (105 SNPs) and source BE using the glm() function in R. Data were subsampled (leaving out 20% of the data) and the model was fitted 100 times. For each OTU, mean percent deviance explained by BE and combined SNPs was reported.

Extraction of metabolites from faecal homogenates. Metabolites were extracted from mouse faecal samples using a methanol/sonication method (for strain information and number of replicate samples see Supplementary Table 1)[47]. Briefly, portions of the homogenized samples were weighed and extracted with cold (−20 °C) methanol proportionally (1 ml solvent added per 100 mg homogenate) in a microcentrifuge tube. The average weight of the homogenized faecal samples was 69.3 ± 26.3 mg (mean ± standard deviation, s.d.) and the methanol extracts contained the same theoretical concentration of metabolites. A 100 µl volume of each methanol extract was transferred to glass vials and dried in a speed-vac concentrator (Labconco CentriVap Benchtop Vacuum Concentrator). Dried metabolite extracts were chemically derivatized using a modified version of the protocol used to create FiehnLib[20]. Briefly, dried metabolite extracts were dried again to remove any residual water if they had been stored at −80 °C. To protect carbonyl

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

groups and reduce the number of tautomeric isomers, 20 µl of methoxyamine in pyridine (30 mg ml^{-1}) was added to each sample, followed by vortexing for 30 s and incubation at 37 °C with generous shaking (1,000 r.p.m.) for 90 min. At this point, the sample vials were inverted once to capture any condensation of solvent at the cap surface, followed by a brief centrifugation at 1,000*g* for 1 min. To derivatize hydroxyl and amine groups to trimethylsilylated (TMS) forms, 80 µl of *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) were then added to each vial, followed by vortexing for 10 s and incubation at 37 °C with shaking (1,000 r.p.m.) for 30 min. Again, the sample vials were inverted once, followed by centrifugation at 1,000*g* for 5 min. The samples were allowed to cool to room temperature and analysed the same day.

An Agilent GC 7890A coupled with a single quadrupole MSD 5975C (Agilent Technologies) was used and the samples were blocked and analysed in random order for each experiment. An HP-5MS column (30 m × 0.25 mm × 0.25 µm; Agilent Technologies) was used for untargeted metabolomics analyses. The sample injection mode was splitless and 1 µl of each sample was injected. The injection port temperature was held at 250 °C throughout the analysis. The GC oven was held at 60 °C for 1 min after injection and the temperature was then increased to 325 °C by 10 °C min^{-1}, followed by a 5 min hold at 325 °C (ref. 48). The helium gas flow rates for each experiment were determined by the Agilent Retention Time Locking function based on analysis of deuterated myristic acid and were in the range of 0.45–0.5 ml min^{-1}. Data were collected over the mass range 50–550 *m/z*. A mixture of fatty acid methyl esters (FAMEs) (C8–C28) was analysed once per day together with the samples for retention index alignment purposes during subsequent data analysis.

GC–MS raw data files were processed using the Metabolite Detector software, version 2.5 beta (ref. 49). Briefly, Agilent .D files were converted to netCDF format using Agilent Chemstation, followed by conversion to binary files using Metabolite Detector. Retention indices (RIs) of detected metabolites were calculated based on analysis of the FAMEs mixture, followed by their chromatographic alignment across all analyses after deconvolution. Metabolites were initially identified by matching experimental spectra to an augmented version of FiehnLib[20] (that is, the Agilent Fiehn Metabolomics Retention Time Locked (RTL) Library, containing spectra and validated retention indices for over 700 metabolites), using a Metabolite Detector match probability threshold of 0.6 (combined retention index and spectral probability). All metabolite identifications were manually validated to reduce deconvolution errors during automated data-processing and to eliminate false identifications. We propose that this approach results in a metabolite identification confidence of Level 1.5 (Level 1 is highest, Level 4 is lowest), according to the guidelines recommended by the Metabolomics Standards Initiative Chemical Analysis Working Group of the Metabolomics Society[50]. The library used to identify metabolites was generated by an external laboratory, but this library contains both retention indices and mass spectra from analyses of authentic chemical standards and our analyses were performed using methods identical to those used to create the library. The NIST 14 GC–MS library was also used to cross-validate the spectral matching scores obtained using the Agilent library and to provide identifications of unmatched metabolites (Level 2 identifications). The three most abundant fragment ions in the spectra of each identified metabolite were automatically determined by Metabolite Detector and their summed abundances were integrated across the GC elution profile; fragment ions due to trimethylsilylation (that is, *m/z* 73 and 147) were excluded from the determination of metabolite abundance. A matrix of identified metabolites, unidentified metabolite features (characterized by mass spectra and retention indices and assigned as 'unknown'; Level 4 identifications) and their abundances was created for subsequent data analysis. Features resulting from GC column bleeding were removed from the data matrices before further data processing and analysis.

Metabolic modelling-based taxonomic and metabolomic integration. We produced a closed-reference OTU table using VSEARCH to align reads from all 77 samples with both sequencing and metabolomics data to the preclustered Greengenes database. We rarefied the OTU table to 4,000 reads and used it as input to MIMOSA (http://elbo.gs.washington.edu/software_MIMOSA.html), a framework for integrating taxonomic and metabolomic microbiome data[21]. MIMOSA uses genomic data, metabolic information and taxonomic composition to predict the community-wide biosynthetic and degradation potential for each metabolite in each sample and identifies metabolites whose variation across samples is consistent with (and can be explained by) variation in this predicted metabolic potential. Metagenome content was inferred for each sample using PICRUSt[51] and normalized using MUSiCC[52]. From these data, a community-wide metabolic model was constructed for each sample and community metabolic potential (CMP) scores were calculated, representing the relative capacity of the predicted community enzyme content in that sample to synthesize or degrade each metabolite. We then compared variation in these scores across samples to variation in measured metabolite concentrations using a rank-based Mantel test, to identify metabolites for which variation in concentration across samples is positively correlated (consistent) with variation in community metabolism (as predicted by the CMP scores), using a local FDR *q*-value less than 0.01 as the significance threshold. We similarly identified metabolites for which variation in concentration across samples is negatively correlated (contrasting) with CMP scores, with the same significance threshold. To identify potential contributing OTUs for each metabolite, we calculated the Pearson correlation between the CMP scores obtained for a given metabolite across samples using the entire community and the CMP scores generated based on each species by itself (that is, recalculating the metagenome content and CMP scores based solely on the abundance of this species). OTUs for which this correlation coefficient for a given metabolite was greater than 0.5 were classified as potential contributing OTUs for that metabolite. Additional details about this computational framework have been described previously[21].

Data availability. Sequence data are available at the Qiita management platform (https://qiita.ucsd.edu/study/description/10500). Scripts to replicate this analysis are available at https://github.com/pnnl/jansson_snijders_collaborative_cross. All raw GC–MS data are available via the MetaboLights metabolomics data repository (http://www.ebi.ac.uk/metabolights/MTBLS345, ID MTBLS345).

Received 23 August 2016; accepted 07 October 2016; published 28 November 2016

References

1. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148,** 1258–1270 (2012).
2. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190,** 389–401 (2012).
3. Kubinak, J. L. *et al.* MHC variation sculpts individualized microbial communities that control susceptibility to enteric infection. *Nat. Commun.* **6,** 8642 (2015).
4. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159,** 789–799 (2014).
5. McKnite, A. M. *et al.* Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS ONE* **7,** e39191 (2012).
6. Benson, A. K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl Acad. Sci. USA* **107,** 18933–18938 (2010).
7. Benson, A. K. Host genetic architecture and the landscape of microbiome composition: humans weigh in. *Genome Biol.* **16,** 203 (2015).
8. Anukam, K. C., Osazuwa, E. O., Osadolor, H. B., Bruce, A. W. & Reid, G. Yogurt containing probiotic *Lactobacillus rhamnosus* GR-1 and *L. reuteri* RC-14 helps resolve moderate diarrhea and increases CD4 count in HIV/AIDS patients. *J. Clin. Gastroenterol.* **42,** 239–243 (2008).
9. Trois, L., Cardoso, E. M. & Miura, E. Use of probiotics in HIV-infected children: a randomized double-blind controlled study. *J. Trop. Pediatr.* **54,** 19–24 (2008).
10. Bravo, J. A. *et al.* Ingestion of *Lactobacillus* strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proc. Natl Acad. Sci. USA* **108,** 16050–16055 (2011).
11. Mohamadzadeh, M. *et al. Lactobacilli* activate human dendritic cells that skew T cells toward T helper 1 polarization. *Proc. Natl Acad. Sci. USA* **102,** 2880–2885 (2005).
12. Replication, D. I. G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46,** 234–244 (2014).
13. Gong, Y. *et al.* PROX1 gene variant is associated with fasting glucose change after antihypertensive treatment. *Pharmacotherapy* **34,** 123–130 (2014).
14. Yu, B. *et al.* Genome-wide association study of a heart failure related metabolomic profile among African Americans in the Atherosclerosis Risk in Communities (ARIC) study. *Genet. Epidemiol.* **37,** 840–845 (2013).
15. Kim, H. J. *et al.* Combined linkage and association analyses identify a novel locus for obesity near *PROX1* in Asians. *Obesity* **21,** 2405–2412 (2013).
16. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44,** 659–669 (2012).
17. Alipour, B. *et al.* Effects of *Lactobacillus casei* supplementation on disease activity and inflammatory cytokines in rheumatoid arthritis patients: a randomized double-blind clinical trial. *Int. J. Rheum. Dis.* **17,** 519–527 (2014).
18. Bordalo Tonucci, L. *et al.* Clinical application of probiotics in diabetes mellitus: therapeutics and new perspectives. *Crit. Rev. Food Sci. Nutr.* http://dx.doi.org/10.1080/10408398.2014.934438 (2015).
19. Hindorff, L. *et al. A Catalog of Published Genome-Wide Association Studies*; http://www.ebi.ac.uk/gwas
20. Kind, T. *et al.* Fiehnlib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **81,** 10038–10048 (2009).
21. Noecker, C. *et al.* Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* **1,** e00013-15 (2016).
22. Welsh, C. E. *et al.* Status and access to the Collaborative Cross population. *Mamm. Genome.* **23,** 706–712 (2012).
23. Iraqi, F. A., Churchill, G. & Mott, R. The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort. *Mamm. Genome* **19,** 379–381 (2008).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

24. Morahan, G., Balmer, L. & Monley, D. Establishment of 'The Gene Mine': a resource for rapid identification of complex trait genes. *Mamm. Genome.* **19,** 390–393 (2008).
25. Chesler, E. J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm. Genome.* **19,** 382–389 (2008).
26. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
27. Walters, W. *et al.* Improved bacterial 16S rRNA gene (V4 and V4–5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *mSystems* **1,** e00009-15 (2015).
28. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).
29. Aronesty, E. *ea-utils: Command-Line Tools for Processing Biological Sequencing Data* (Expression Analysis, 2011); https://github.com/ExpressionAnalysis/ea-utils
30. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
31. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
32. Rognes, T., Flouri, T. & Mahe, F. *vsearch: VSEARCH Version 1.1.3* (2015); https://zenodo.org/record/16153#.VwwcqxMrKuM
33. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6,** 610–618 (2012).
34. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26,** 266–267 (2010).
35. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5,** e9490 (2010).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
37. Lozupone, C. & Knight, R. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
38. McMurdie, P. J. & Holmes, S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8,** e61217 (2013).
39. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2010).
40. R-Core-Team. R: *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2016); http://www.R-project.org/
41. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mamm. Genome* **26,** 366–378 (2015).
42. Eppig, J. T. *et al.* The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.* **43,** D726–D736 (2015).
43. Yin, T., Cook, D. & Lawrence, M. Ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13,** R77 (2012).
44. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).
45. Mao, J. H. *et al.* Identification of genetic factors that modify motor performance and body weight using Collaborative Cross mice. *Sci. Rep.* **5,** 16247 (2015).
46. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2,** 18–22 (2002).
47. Walker, A. *et al.* Importance of sulfur-containing metabolites in discriminating fecal extracts between normal and type-2 diabetic mice. *J. Proteome Res.* **13,** 4220–4231 (2014).
48. Kim, Y. M. *et al.* Salmonella modulates metabolism during growth under conditions that induce expression of virulence genes. *Mol. Biosyst.* **9,** 1522–1534 (2013).
49. Hiller, K. *et al.* Metabolitedetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal. Chem.* **81,** 3429–3439 (2009).
50. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3,** 211–221 (2007).
51. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31,** 814–821 (2013).
52. Manor, O. & Borenstein, E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol.* **16,** 27 (2015).

Acknowledgements

The authors thank S.E. Cates, N.N. Robinson and G.D. Shaw in the Systems Genetics Core at UNC for technical assistance and M.H. Stoiber for helpful discussions, especially regarding statistical analysis. This work was primarily supported by funding from the Office of Naval Research under ONR contract N0001415IP00021 (J.J., J.H.M. and A.M.S.). Additional support was provided by the Low Dose Scientific Focus Area, Office of Biological and Environmental Research, US Department of Energy (G.K., J.H.M. and A.M.S.) and the Lawrence Berkeley National Laboratory Directed Research and Development (LDRD) program funding under the Microbes to Biomes (M2B) initiative (S.C., B.B., G.K., J.H.M. and A.M.S.). C.N. was supported by an NSF IGERT DGE-1258485 fellowship and in part by New Innovator Award DP2 AT007802-01 to E.B. Partial support was also provided under the Microbiomes in Transition (MinT) Initiative as part of the Laboratory Directed Research and Development Program at PNNL. Metabolomic measurements were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the US DOE OBER and located at PNNL in Richland, Washington. PNNL and LBNL are multi-program national laboratories operated by Battelle for the DOE under contract DE-AC05-76RLO 1830 and the University of California for the DOE under contract DE AC02-05CH11231, respectively.

Author contributions

A.M.S., J.-H.M. and J.K.J. conceived and designed the study. A.M.S. and J.-H.M. performed the mouse experiments, acquired the data, performed data analysis, interpreted results and co-wrote the manuscript. S.A.L. performed data analysis, interpreted results and co-wrote the manuscript. T.O.M. and Y.-M.K. performed metabolome data analysis, interpreted results and co-wrote the manuscript. C.J.B. performed microbiome data analysis and interpreted results. C.N. performed metabolic modelling-based taxonomic and metabolomics integration. E.M.Z. prepared microbiome samples and performed GC–MS-based metabolomics analysis. S.J.F. carried out microbiome sequencing. C.P.C. performed metabolome data analysis and interpreted results. D.R.M. acquired data. Y.H. performed *in vivo* experiments and collected data. G.H.K. and S.E.C. interpreted results and co-wrote the manuscript. J.B.B. supervised the integrative data analysis, interpreted results and co-wrote the manuscript. E.B. performed data analysis, interpreted results and co-wrote the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.K.J., T.O.M. and J.H.M.

How to cite this article: Snijders, A. M. *et al.* Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. *Nat. Microbiol.* **2,** 16221 (2017).

Competing interests

The authors declare no competing financial interests

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.