ARTICLES
https://doi.org/10.1038/s41564-017-0084-4
nature microbiology

Metatranscriptome of human faecal microbial communities in a cohort of adult men

Galeb S. Abu-Ali[1,2], Raaj S. Mehta[3,4], Jason Lloyd-Price[1,2], Himel Mallick[1,2], Tobyn Branck[1,5], Kerry L. Ivey[6,7], David A. Drew[3,4], Casey DuLong[1], Eric Rimm[6,8], Jacques Izard[9], Andrew T. Chan[2,3,4,8]* and Curtis Huttenhower[1,2]*

The gut microbiome is intimately related to human health, but it is not yet known which functional activities are driven by specific microorganisms' ecological configurations or transcription. We report a large-scale investigation of 372 human faecal metatranscriptomes and 929 metagenomes from a subset of 308 men in the Health Professionals Follow-Up Study. We identified a metatranscriptomic 'core' universally transcribed over time and across participants, often by different microorganisms. In contrast to the housekeeping functions enriched in this core, a 'variable' metatranscriptome included specialized pathways that were differentially expressed both across participants and among microorganisms. Finally, longitudinal metagenomic profiles allowed ecological interaction network reconstruction, which remained stable over the six-month timespan, as did strain tracking within and between participants. These results provide an initial characterization of human faecal microbial ecology into core, subject-specific, microorganism-specific and temporally variable transcription, and they differentiate metagenomically versus metatranscriptomically informative aspects of the human faecal microbiome.

Alterations in the human gut microbiome have been implicated in a wide range of complex, chronic conditions, including inflammatory bowel disease (IBD), obesity, diabetes mellitus, cancer and cardiovascular disease[1,2]. There is an appreciable body of work on the metagenomic potential of faecal communities[3–6], yet little information is available regarding transcriptional activity of the microbiome. The metatranscriptome represents a link between the metagenome and community phenotype, and surveying its molecular activity is important to understanding the functional ecology of the human gut microbiome.

Metatranscriptomics has most commonly been applied to ecological profiles of environmental microbial populations. For instance, deep sequencing of marine bacterioplankton RNA established transcript inventories, uncovered gene expression trends among metabolic generalists and specialists, and identified patterns of substrate use and elemental cycling in the ocean ecosystem[7]. Early human faecal metatranscriptomics suggested subject-specific relationships between microbiome transcripts and gene copy number, differing across biological functions[8]. Our own pilot work with the cohort studied here introduced protocols for integrating metatranscriptomic sampling into large-scale epidemiological studies, demonstrating that metatranscriptional profiles are less variable than faecal taxonomic profiles but more individualized than metagenomic function[9].

However, previous studies have not surveyed the human faecal metatranscriptome at sufficient scale to identify areas in which it is uniquely informative relative to the underlying metagenome. It is not yet clear, for example, which human conditions are associated with specific microorganisms in the gut, versus their metagenomic functional profiles, versus metatranscriptomic activity[10]. Culture-independent faecal microbial transcription has been used in only a few cases to date to identify causal mechanisms in health outcomes, such as strain-specific variation in *Eggerthella lenta* expression influencing the efficacy of cardiac therapy[11]. Integrated metagenomics and metatranscriptomics have also been used in molecular diagnostics for cancer risk and transplant rejection[12]. Correspondingly, it remains to be determined what short-term or long-term health-linked exposures can be assessed using faecal metagenomes or metatranscriptomes in prospective cohorts.

To address these gaps, we interrogated the human faecal metagenome and metatranscriptome in 929 and 372 samples, respectively, collected at up to four time points each from 308 healthy senior men participating in the Men's Lifestyle Validation Study (MLVS), nested within the Health Professionals Follow-up Study (HPFS)[13]. This article provides an overview of the molecular biology and microbial ecology of these communities; a companion paper[14] investigates the stability of features in such data for human population epidemiology. We differentiated 'core' versus variably transcribed functions, assigned them to specific microorganisms, and assessed differences between subjects cross-sectionally and within subjects over time. Finally, ecological co-occurrence and strain diversity were assessed metagenomically, both remaining strikingly stable over time, and the latter comparing near-identically between this and an independent population from the Human Microbiome Project (HMP). Together, these findings therefore provide an in-depth large-scale exploration of the human faecal metatranscriptome in a species-specific context.

[1]Biostatistics Department, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [2]The Broad Institute, Cambridge, MA, USA. [3]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [4]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [5]U.S. Army Natick Soldier Systems Center in Natick, Natick, MA, USA. [6]Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [7]South Australian Health and Medical Research Institute, Infection and Immunity Theme, School of Medicine, Flinders University, Adelaide, South Australia, Australia. [8]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [9]University of Nebraska, Lincoln, Lincoln, NE, USA. Galeb S. Abu-Ali, Raaj S. Mehta, Andrew T. Chan and Curtis Huttenhower contributed equally to this work. *e-mail: achan@mgh.harvard.edu; chuttenh@hsph.harvard.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Results

Meta'omic taxonomic and functional profiling. We generated taxonomic and functional profiles of 308 participants' stool microbiome samples at up to four time points each from DNA ($n=929$) and RNA ($n=372$) reads using MetaPhlAn2 (ref. [15]) and HUMAnN2 (ref. [16]) (Fig. 1; Methods). From the DNA reads, a total of 468 microbial species were detected, with individual samples containing 72 ± 13 (mean $\pm$ s.d.) species. HUMAnN2 identified 1,569,171 unique UniRef90 gene families in metagenomes and 602,896 in metatranscriptomes (Supplementary Table 1). Overall, 75.3% of all DNA reads and 64.1% of all RNA reads were assignable to UniRef90 gene families by HUMAnN2; of these, 54.8% and 58.1% UniRef90 gene families possessed functional characterization, respectively, and 10.7% and 13.2% of characterized gene families were assignable to MetaCyc pathways[17]. Intriguingly, an average of 69% and 85.4% UniRef90 relative abundances for metagenomes and metatranscriptomes, respectively, were attributable to gene families lacking biochemical characterization. Stool sample collection, sequence data generation and quality control are described in Methods.

Prior to investigating the metatranscriptome, we compared the metagenome-based taxonomic profile of this older cohort to previous population studies (Supplementary Fig. 1), as the mean age

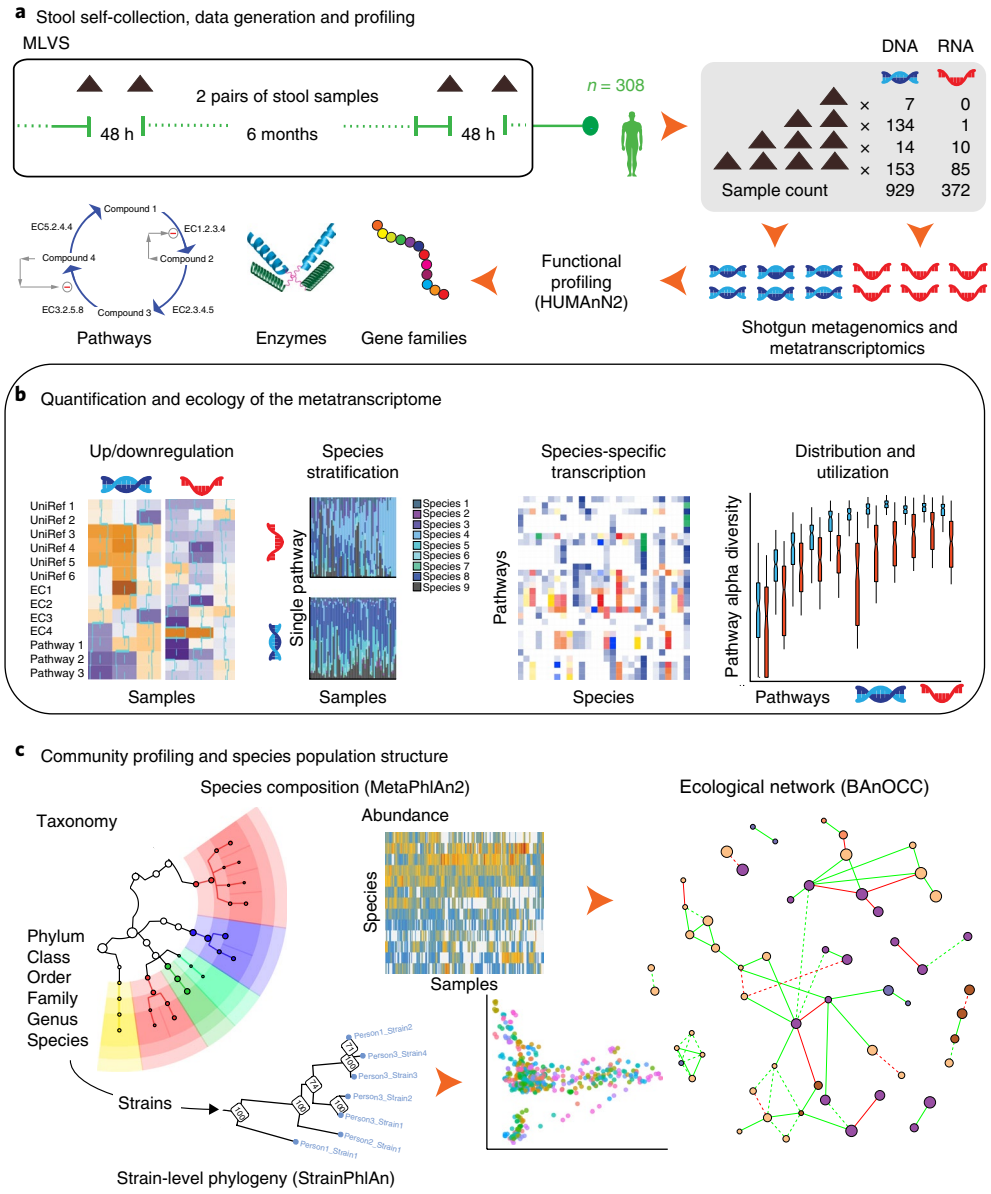

Fig. 1 | Metatranscriptomic and metagenomic taxonomic and functional profile of a prospective human cohort. a, Participants ($n=308$) from the MLVS, embedded within the HPFS prospective cohort[13], provided a target of four stool samples each. These stool samples were self-collected in 2 pairs, 6 months apart, with each pair spanning 2–3 days. This yielded 929 total metagenomes and 372 metatranscriptomes, sequenced using previously published protocols[9] and functionally profiled using HUMAnN2 (ref. [16]). **b**, To estimate gene family, enzyme class and pathway relative transcription, RNA abundances were normalized to corresponding DNA abundances. We then evaluated 'core' (prevalently transcribed) and variable transcriptional elements, in addition to the ecological and phylogenetic diversity of metatranscription and carriage of functional elements among species. EC, Enzyme Commission number for enzymes. **c**, Taxonomic profiles were determined using MetaPhlAn2 (ref. [15]) from both DNA and RNA data (for RNA viruses). These were also used for ecological interaction network reconstruction[26] with BAnOCC[45] (http://huttenhower.sph.harvard.edu/banocc) and for strain tracking with StrainPhlAn[29].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

of our participants was 69 ± 6 years. As in earlier studies with comparable populations and protocols[18,19], and in contrast to younger cohorts with different sample handling methodology[3], Firmicutes were generally prevalent and abundant, in contrast to previous studies of comparable but smaller populations such as ELDERMET[19,20] (Supplementary Fig. 2). In addition, a small number of both DNA and RNA viruses were quantified confidently by MetaPhlAn2, which is probably an underestimate of the gut virome diversity, as our extraction protocol did not enrich for virus-like particles. Although gut viral ecology is more difficult to analyse than that of the bacteriome owing to inadequate viral reference sequences[21], these methods allow for some incidental analysis of DNA phage and RNA plant viruses in human faecal metagenomes and metatranscriptomes (see Supplementary Information).

As a first indication of important differences between faecal metagenomic and metatranscriptomic profiles, each sample's metatranscriptome contained, on average, 5.4% of the total pool of gene families observed across the data set; metagenomes averaged >10% of the pool (Supplementary Table 1). This indicates that, as in a single organism's genome, only a subset of faecal functional potential is active under the circumstances captured by a typical sample. Technical factors had a minor role in this, as, although RNA was sequenced at slightly shallower depth (Supplementary Data set 1), rarefaction indicates that metagenomic taxa, functions and metatranscriptomic functions were well saturated at these depths (Supplementary Fig. 3). Biological factors seemed to dominate, as transcribed elements should typically be at most those also observed metagenomically. Ecologically, this is also in agreement with our previous observation that the metatranscriptome is more variable than the metagenome[9].

Core and variable faecal metatranscriptomes differ from the metagenome. To identify important pathways expressed (and not just metagenomically encoded) by microorganisms in the human gut, we delineated 'core' and 'variable' portions of the faecal metatranscriptome (Fig. 2, Supplementary Data sets 2 and 3). The former was defined as a set of prevalently transcribed pathways that was robust to sequencing depth and specific prevalence threshold (Supplementary Fig. 4). From 289 pathways with detectable transcription in at least two samples, 81 (28%) were core by this definition, 48 of which had a mean DNA-normalized transcript abundance of >1 when transcribed (see Methods for quantification of metatranscriptional activity). This was remarkably smaller than a similarly defined core metagenome in this cohort: from 407 pathways with detectable DNA in at least two samples, 182 (45%) were similarly prevalent, even though there were almost three-times more metagenomes than metatranscriptomes. It is noteworthy that neither GC content nor open reading frame (ORF) length had effects on transcription ratios (see Supplementary Information and Supplementary Fig. 5). This suggests that gene expression rather than gene abundance underlies qualitative and quantitative differences among metatranscriptomes.

Unlike the core metagenome, which includes various host-adapted microbial community features[3], the core metatranscriptome was enriched mainly for housekeeping functions. Nineteen nucleotide biosynthesis pathways, 19 glycolysis and carbohydrate metabolism pathways and 15 amino acid biosynthesis pathways accounted for the majority of core metatranscribed pathways. Note that as annotated in MetaCyc, glycolysis represents an umbrella term that also includes anaerobic fermentation, not an indicator of aerobic respiration. Glycolysis in this sense had the highest transcript abundance (8.30 ± 5.69 (mean ± s.d.)) and, together with three nucleotide metabolism pathways, was over-transcribed (relative to DNA abundance) in virtually all metatranscriptome samples (Fig. 2a,b and Supplementary Fig. 6).

Other notable core metatranscriptome pathways included the non-oxidative pentose phosphate cycle (to support nucleic acid synthesis), breakdown of carboxylates, synthesis of co-factors (folate, flavin, pantothenate and coeznyme A (CoA)), unsaturated fatty acids, and microbial recycling of uric acid (the end product of purine breakdown) to salvage nitrogen. Conversely, cell wall peptidoglycan and phospholipids, and amino acid synthesis were pathways with the lowest transcript abundance in the core metatranscriptome. Interestingly, the core metatranscriptome also included synthesis of $preQ_0$ (prequeosine-0) and queosine, pleiotropic bacterial metabolites[22] that the human host salvages for regulating translational efficiency and fidelity[23].

In contrast to the limited core set of metatranscriptomic pathways, the variable metatranscriptome comprised 95 functionally diverse pathways that were well detected in DNA but below detection in at least half of RNA samples (Fig. 2c); 36 of these had no detectable RNA in more than two-thirds of the metatranscriptome samples. The bulk of these pathways are involved in biosynthesis of various amino acids, long fatty acids, terpenoids, polyamines, co-factors (NAD, haem and tetrapyrolle) and the (p)ppGpp alarmone[24]. A smaller number are pathways involved in the degradation of various alcohols, sugars, formaldehyde and sulfate reduction. Finally, 26 pathways were below detection in most DNA (and matching RNA) samples. However, when transcribed, some of these pathways had the highest transcript abundance, for example, methanogenesis and factor 420 biosynthesis (Fig. 2d and Supplementary Fig. 6f), suggesting that metagenomically rare but overtranscribed pathways may be uniquely responsible for subject-specific gut microbial bioactivity.

Faecal microbiome pathways are transcribed by a limited subset of microorganisms encoding them metagenomically. Using paired metagenomic and metatranscriptomic functional profiles, we next assessed the relationship between faecal microorganisms that tend to carry pathways metagenomically versus those that express them metatranscriptomically (Fig. 3 and Supplementary Fig. 7). At a global level, these two aspects of functional diversity corresponded (in large part as only microorganisms carrying a pathway can express it) with three main groups of pathways that dominate the functional diversity of these samples. First, housekeeping functions (nucleotide, carbohydrate and amino acid metabolism, among others) were carried by almost all stool species (high contributional alpha diversity); they were also actively transcribed by abundant and prevalent microorganisms, predominantly *Bacteroides*, *Eubacterium* and *Ruminococcus* spp. (Fig. 3a, bottom cluster). A second pathway cluster (Fig. 3a, top left) with modest contributional alpha diversity was dominated by *Faecalibacterium prausnitzii*, a particularly prevalent organism in this cohort that, when abundant, tended to contribute the majority of all pathways it encodes (synthesis of various amino acids, non-oxidative pentose phosphate pathway and putrescine synthesis, among other pathways). The third, low diversity, cluster (Fig. 3a, top right) was encoded by limited numbers of opportunists, for example, *Escherichia coli*, *Sutterella*, *Enterobacter* and *Enterococcus* spp. This cluster included pathways such as synthesis of the enterobactin siderophore, lipid A and sulfate reduction. Pyruvate fermentation to acetate and lactate, synthesis of glycogen and tetrapyrrole were examples of pathways with fairly even metagenomic and metatranscriptomic species contributions. Thus, in summary, the same microorganisms were principal contributors to RNA and DNA abundances in roughly one-third of pathways. However, for the majority of pathways, overtranscription was observed from members of either the Bacteroidetes (mainly for pathways contributing to active growth in the gut) or the Proteobacteria (for the subset of generalists pathways upregulated in faeces) relative to the baseline level of metagenomic carriage (for example, by the Firmicutes).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

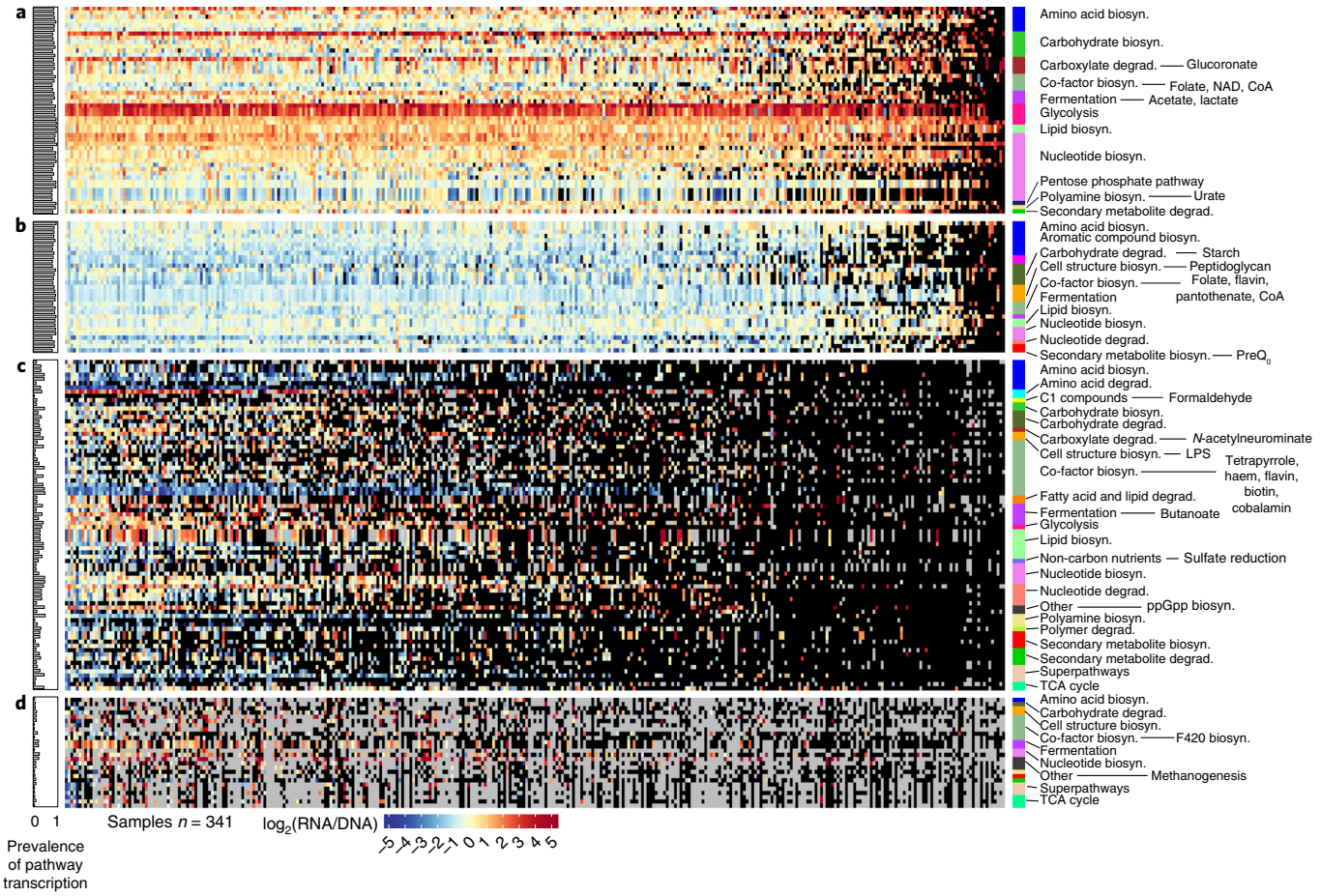

Fig. 2 | Core and variable metatranscriptomes of the stool microbiome. DNA-normalized transcript abundances for 239 gut microbiome pathways with detectable RNA in >10 of the 341 metatranscriptomes, collected from 96 MLVS participants. Samples (columns) were sorted left to right based on decreasing number of transcribed pathways per sample. **a**, Core metatranscriptome pathways (transcribed in >80% of samples) with RNA/DNA transcription ratio of >1. **b**, Low-expression core metatranscriptome pathways with transcript abundance detectable in >80% of samples but an RNA/DNA ratio of <1. **c**,**d**, Variably metatranscribed pathways detected in DNA but below detection in at least half of RNA samples (**c**), and variably metatranscribed pathways below detection in DNA (and matching RNA) in 30–80% of the 341 samples (**d**). Several pathways representative of functional categories are annotated, and the complete annotation of all pathway names and definitions are in Supplementary Fig. 6a–d. Thirty-eight pathways that did not fall into either of the four sections based on these criteria are in Supplementary Fig. 6e. The distribution range of pathways with the overall 30 highest and 30 lowest mean DNA-normalized transcript abundances among the 341 metatranscriptome samples are in Supplementary Fig. 6f. The grey colour represents pathways that were below detection in both DNA and RNA in a given sample; the black colour represents pathways that were detected in DNA but below detection in RNA. biosyn., biosynthesis; degrad., degradation; ppGpp, guanosine tetraphosphate; TCA, tricarboxylic acid.

These three major patterns in the functional structure of the faecal metatranscriptome are explained in greater detail by the pathways that are most commonly expressed by abundant microorganisms. First, all species shared enrichments for housekeeping functions (for example, nucleotide synthesis and fermentation, among others) (Fig. 3b, left columns). A set of anaerobic biosynthesis pathways were mainly expressed by Firmicutes from the upper left cluster of the ordination (Fig. 3b, middle columns). Transcription of cell structure, secondary metabolites and co-factor synthesis pathways was characteristic of *Bacteroides* spp. (Fig. 3b, right columns). Finally, the Enterobacteriaceae were not abundant in most subjects, so few of their pathways were prevalent enough to appear in those selected for visualization, but the subset of their large pangenome upregulated in faeces overlapped to a degree with the anaerobic metabolism expressed by the Firmicutes (Fig. 3b, middle columns).

Many pathways are transcribed by few organisms per community, even when broadly encoded metagenomically. Finally, we noted that per-microorganism pathway expression is often very different among individual hosts than is metagenomic pathway carriage, indicating that these two molecular measurements may have distinct applications in human population studies (Fig. 4). Overall, metagenomic richness (the number of contributing species) generally exceeded metatranscriptomic richness, as expected. As above, a subset of pathways were both broadly encoded and expressed (high diversity), with relatively few differences between individuals, and these were again mainly housekeeping functions. However, at the other end of the spectrum, many pathways of greater biochemical interest were transcribed by only one or a few species, even when diversely carried metagenomically. An extreme example of low transcriptional diversity pathways was methanogenesis, encoded and transcribed solely by *Methanobrevibacter smithii*; this class of metatranscriptomic functions may indicate keystone processes and their associated microorganisms in the gut.

To better understand this phenomenon, and to expand on our previous observation of basal transcription in a substantial fraction of gut pathways[9], we next identified pathways for which the transcriptional activities of microorganisms mirrored their

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

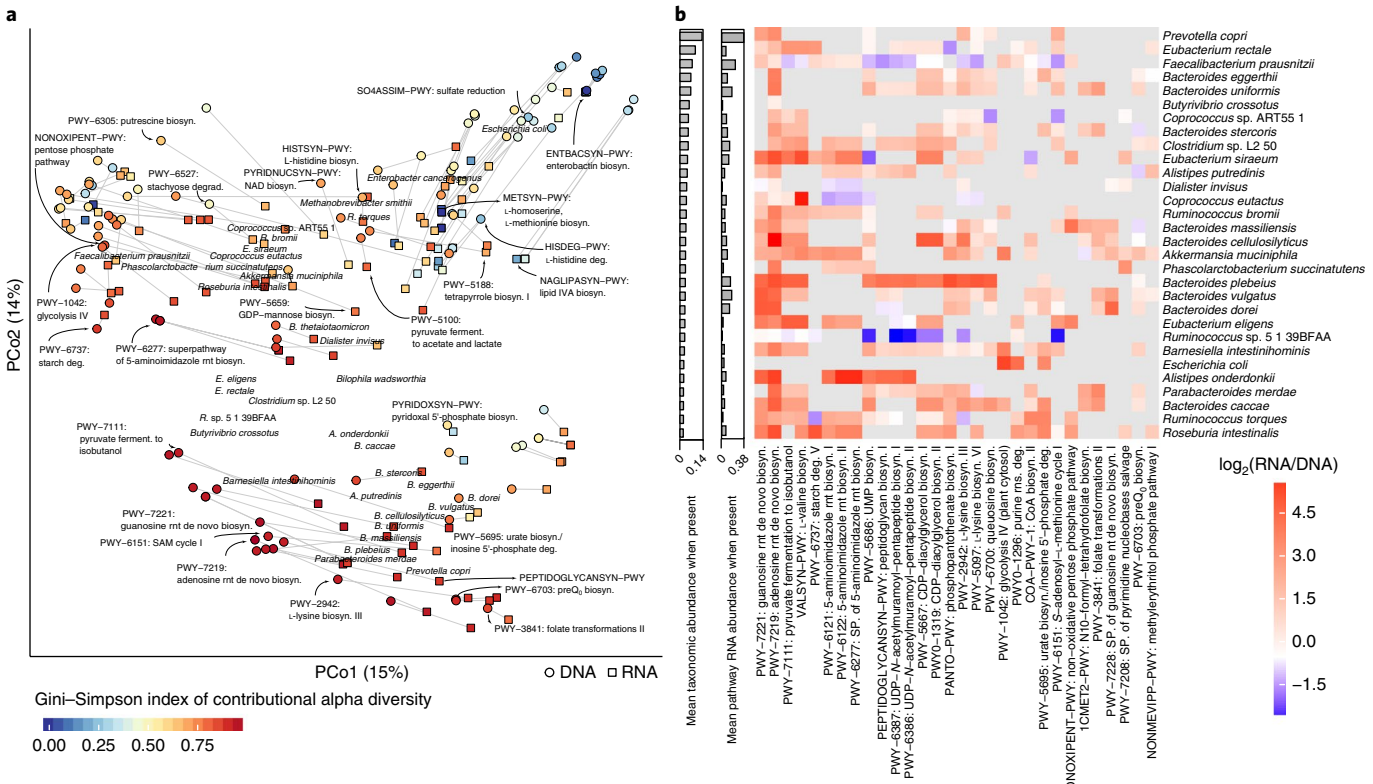

Fig. 3 | The gut metatranscriptome is personalized and broadly taxonomically distributed. a, Structure of the stool metagenome and metatranscriptome as contributed by diverse species. Principal coordinates (PCoA) analysis of pathways with microbial species' contributions to their DNA and RNA abundances using Bray–Curtis dissimilarity, with a bi-plot overlay indicating centroids of abundant species' contributions. Thus, each pathway is denoted by two points, summarizing the organisms contributing them metagenomically (averaged over 913 samples from 307 participants) and metatranscriptomically (341 samples from 96 participants), and a subset of examples are labelled. The resulting joint ordination indicates broad agreement between species carrying (metagenomically) and expressing (metatranscriptomically) groups of pathways in the faecal microbiome. **b**, Transcription ratios of 30 pathways that were most prevalently transcribed among the top 30 species, using the same data sets as in **a**. Pathways for which DNA or RNA were not detected in a given species are grey. A given pathway–species combination in the heatmap represents the transcript abundance averaged over all samples that measured a non-zero RNA/DNA ratio for that species. Only pathway–species combinations in at least 5 samples (from a total of 341) were considered. Columns in the heatmap were ordered based on average linkage clustering on a Euclidean distance matrix of $\log_2$ pathway transcription ratios. biosyn., biosynthesis; deg., degradation; ferment., fermentation; rns, ribonucleoside; rnt, ribonucleotide; SAM, S-adenosyl methionine; SP., super pathway. Complete pathway and taxa annotations and abundances can be found in Supplementary Data sets 2 and 3.

metagenomic carriage (Fig. 4b). For this, we used the weighted mean of the Spearman correlation between taxonomically stratified metagenomic and metatranscriptomic profiles (see Methods). To emphasize the correlation of the principal species encoding each pathway, correlation coefficients were weighted by the mean metagenomic potential of the species. High values indicated that species-level RNA abundances closely follow DNA abundances, whereas low values indicated departure from basal expression. In general, carbohydrate metabolism and nucleotide biosynthesis particularly tended to be enriched for high correlations, whereas amino acid biosynthesis was enriched among low correlations. Low correlations indicate more context-specific, variable expression of the pathway, consistent with the generally amino-acid-rich environment of the gut. Co-factor biosynthesis pathways were roughly in the middle of this range, perhaps due to the diversity of compounds produced with these functions and their variable availability in the gut. For example, pantothenate is found in most foods and is easily salvaged from the gut environment, whereas folate transformations are critical for one-carbon (C1) metabolism[25], making this pathway more widely transcribed when present in species.

Notably, pathways with similar metagenomic contributions were not necessarily similarly transcribed (Fig. 4c,d). Core pathways were both metagenomically and metatranscriptomically diverse, carried and expressed by many organisms per community (Fig. 4d and Supplementary Fig. 8). However, typical variable pathways, even when broadly distributed metagenomically, were often transcribed by one or few organisms per individual. Transcribing organisms were often neither the most abundant nor the same species across individuals (Fig. 4c). Similar patterns were observed for L-isoleucine biosynthesis III (PWY-5103), L-tryptophan biosynthesis (TRPSYN-PWY), degradation of various sugars (stachyose (PWY-6527), sucrose (PWY-621), rhamnose (RHAMCAT–PWY)), non-oxidative pentose phosphate pathway, $preQ_0$ biosynthesis (PWY–6703), and others (Supplementary Fig. 8). This finding highlights another aspect of inter-individual diversity in the microbiome: different microorganisms may activate shared pathways among individuals, with as-yet-unknown functional specializations and consequences.

Inter-microbial species interactions are stable in the stool ecosystem of older men. To leverage this cohort's taxonomic profiles independently of their metatranscriptomes, we also inferred metagenomic microbial interaction networks[26] using the Bayesian Analysis of Compositional Covariance (BAnOCC) framework (see Methods; Fig. 5 and Supplementary Fig. 9). We generated one network per time point, allowing the stability of co-occurrence and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

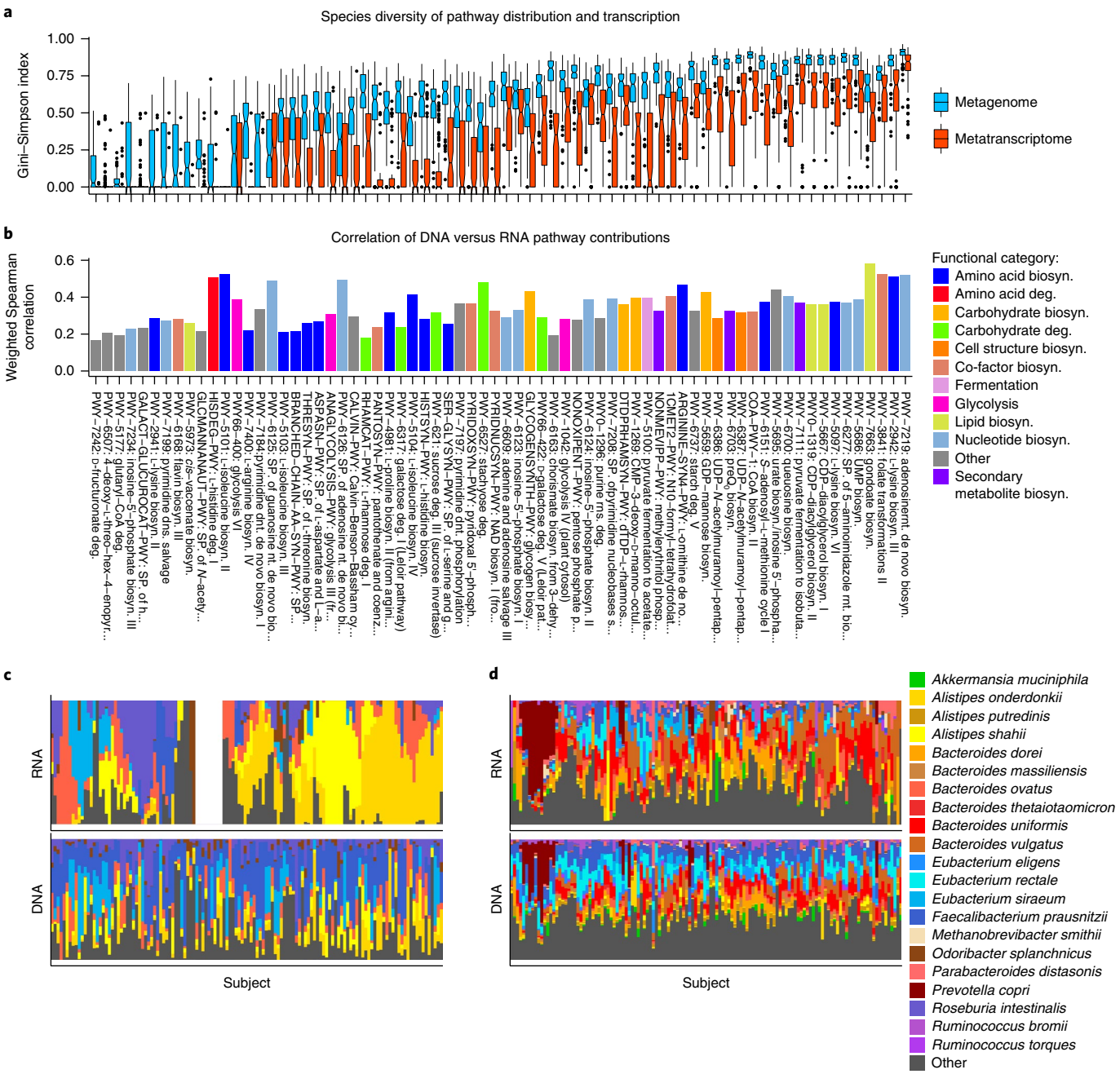

Fig. 4 | Transcriptional landscape of the stool microbiome. a, Distributions of alpha diversity (Gini–Simpson index) for the species-specific metagenomic and metatranscriptomic contributions to each pathway, for 70 non-redundant pathways with the highest community-level RNA abundances, averaged across 341 metatranscriptomes from 96 participants. Pathways were sorted by the sum of the median metagenomic alpha diversity and the weighted Spearman correlation from **b**. Boxplot whiskers represent 1.5 times the interquartile range from the first and third quartiles. **b**, Concordance of metagenomic potential with metatranscriptomic activity (metagenome-weighted mean of per-species Spearman correlations; see Methods). Metatranscriptomic diversity is, as expected, consistently lower than metagenomes, with pathways carried by only a few organisms that are also more differentially transcribed; dns, deoxyribonucleosides. See caption of Fig. 3 for definitions. **c**,**d**, Metagenomic potential (bottom) and metatranscriptomic activity (top), for example, contribute pathways with differing ecological structure, specifically GDP-mannose biosynthesis (PWY-5659) (**c**) and adenosine ribonucleotide de novo biosynthesis (PWY-7219) (**d**). Abundances were normalized within each pathway for 189 subject–week pairs, from 96 participants. Subjects (columns) were ordered to emphasize blocks of subjects with similar metatranscriptomic profiles (see Methods). The top 8 (**c**) and top 15 (**d**) species, in terms of their mean metatranscriptomic contribution to the pathways in **c** and **d,** are shown for clarity. Examples show transcriptional ecologies that either differ strikingly from **c** or generally mirror **d** in their metagenomic diversity. Pathway definitions on the *x* axis of panel **b** were truncated for space; full definitions can be found in Supplementary Data sets 2 and 3, as well as on www.metacyc.org by searching for pathway names, for example PWY-7219.

co-exclusion relationships to be determined. Negative associations between *Bacteroides* spp. and Ruminococcaceae or Prevotellaceae observed previously[26] were not recapitulated in this population, and Firmicutes species predominantly co-occurred with other members of the phylum. This is consistent with the observation that co-occurrence/co-exclusion among microbial community members

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

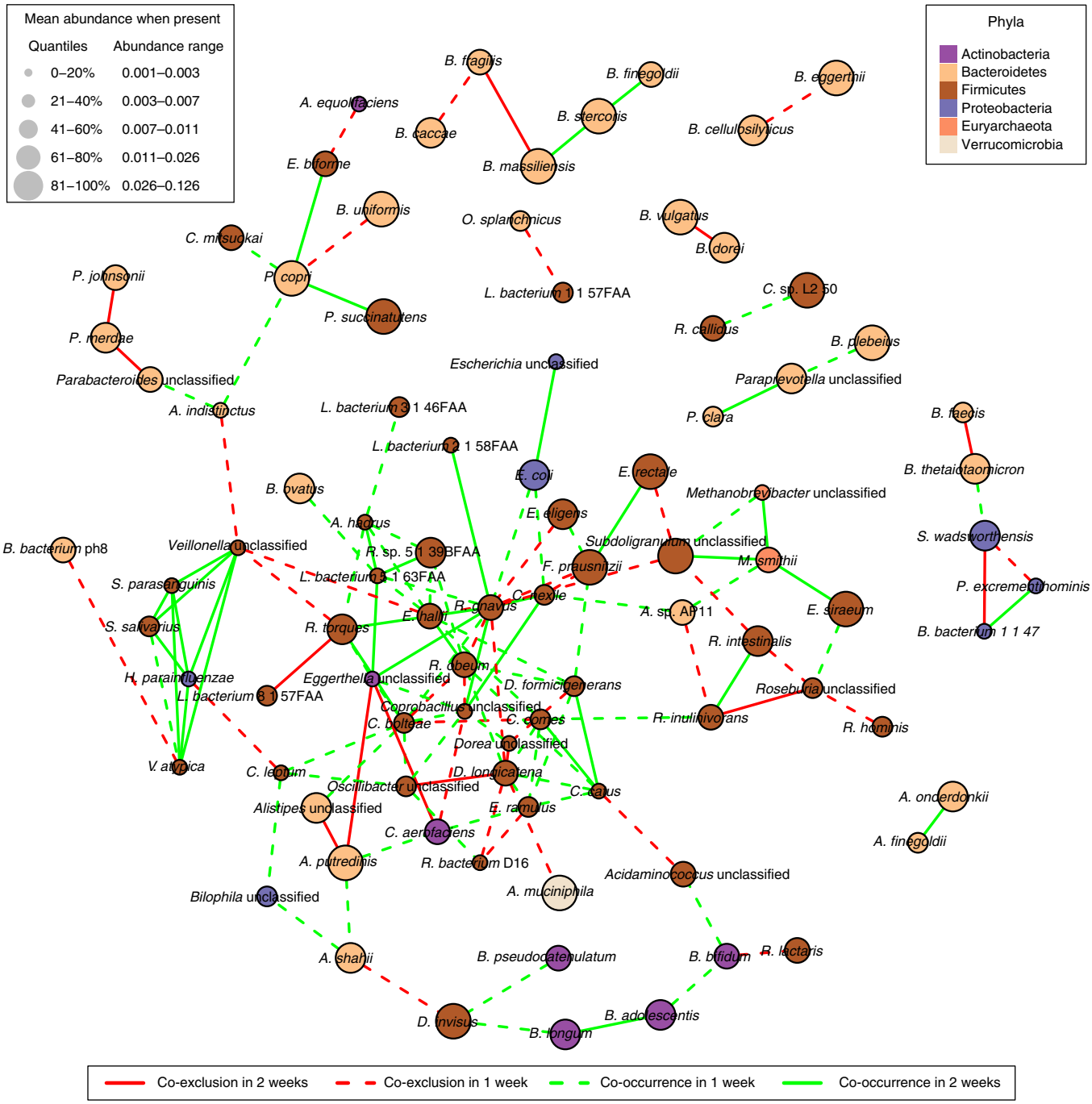

Fig. 5 | Ecological interactions in the gut microbiome. Significant co-variation and co-exclusion relationships among 104 species in 913 stool metagenomes from 307 MLVS participants. Each node represents a species and edges correspond to significant interactions inferred by BAnOCC (see Methods). Stool microbiome taxonomic profiles were averaged within each subject for the first and second collection pairs (separated by 6 months). Interactions in at least one time point are included here. No alternating associations (positive at one time and negative in another) were detected. Ninety-five per cent credible interval criteria were used to assess significance and only estimated absolute correlations with effect sizes of ≥0.15 are reported. Networks for individual time points are shown in Supplementary Fig. 9. Complete taxonomic profiles for all stool metagenomes can be found in Supplementary Table 3.

is not based solely on phylogenetic relatedness, but also on how microorganisms complement each other functionally[27].

Genetic divergence patterns of stool-associated bacterial strains are species-specific and preserved among host populations. Finally, we assessed strain-level variation in stool bacterial populations using StrainPhlAn (see Methods) and provided a comparison of species population structure between the MLVS and the HMP[3,28] cohorts (Fig. 6). Twenty-one species had sufficient genomic coverage for reliable strain identification from metagenomes in both cohorts. *Eubacterium siraeum*, for example, demonstrated discrete strain clustering indicative of a clonal population, with strains from the same individual remaining near-identical over the sampling period. Conversely, nucleotide variation among *F. prausnitzii* strains did not reveal a discernible pattern, and was characterized by the highest median and widest range of nucleotide substitution rates,

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

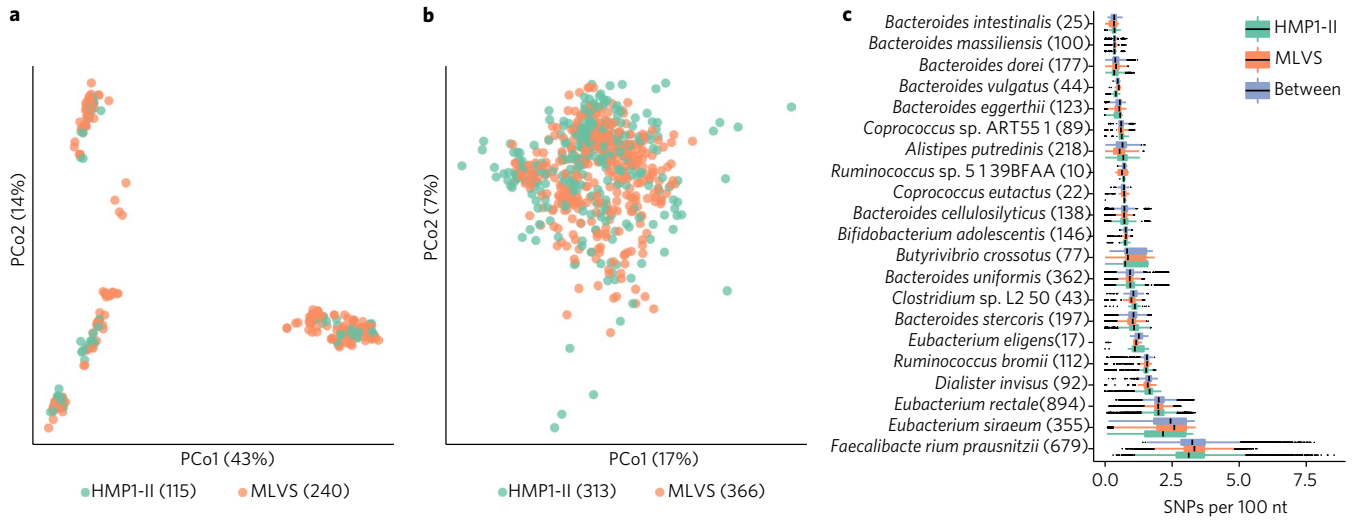

Fig. 6 | Species-specific patterns of evolutionary divergence within species preserved across cohorts. a,**b**, Panels show strain-level diversity within *Eubacterium siraeum* (**a**) and *F. prausnitzii* (**b**). Each point represents one sample's strain, ordinated by PCoA of sequence dissimilarity (Kimura two-parameter distance). **c**, Pairwise nucleotide substitution rates within and between cohorts for 21 out of 30 species in Fig. 3 with sufficient prevalence in both cohorts for informative comparison. The rate of nucleotide divergence between the MLVS and HMP cohorts are summarized with the 'Between' boxplots for each species. Lines represent median values, points denote outliers outside 1.5 times the interquartile range. All numbers in parentheses are sample counts in which indicated strains were above the limit of detection, from a total of 913 MLVS stool metagenomes and 553 HMP stool metagenomes (from 253 male and female HMP participants) that were analysed with StrainPhlAn. SNPs, single-nucleotide polymorphisms.

consistent with extreme genomic diversity whereby each strain can be substantially distinct from another. Several species, including *Bacteroides stercoris*, *Bacteroides uniformis* and *Butyrivibrio crossotus*, presented an intermediate structure with weak clonal propagation of a potential outgroup (Supplementary Fig. 10).

Remarkably, nucleotide substitution rates were near-identical between the two unrelated cohorts (Fig. 6c). The average ratio of between-cohort to within-microorganism divergence was 1.0 ± 0.03 (mean $\pm$ s.d.), indicating that bacterial population structure is consistent across host populations. Previous studies have shown that human gut microorganisms can diverge between geographically and genetically distinct host populations[29]. However, the existence of such closely related microbial strains between independent North American cohorts with very different age ranges (18–40 versus 65–81 yr) suggests that specific microbial strains (or related strain groups) may be a useful, stable feature to assay during epidemiological studies.

Discussion

The present study has provided an overview of the faecal metatranscriptome in a prospective, large-scale cohort of elderly males; identified core and variably transcribed pathways; delineated how these differ from metagenomic functional potential; and ascribed them to specific contributing organisms. Finally, paired metagenomes in this study also allowed species-specific ecological interaction networks to be reconstructed, which proved stable over time, as did strain tracking within species. This stability, in combination with the commonality of strain-level microbial population structures between cohorts, suggests that they might represent particularly effective measurement targets for the microbiome in population studies. Together, these findings extend our earlier pilot study in eight individuals[9] and accompany epidemiological work[14] to integrate metatranscriptomics into population studies.

It is evident from our results that the metatranscriptome is more temporally dynamic, context-sensitive and species-specific than the metagenome. The observed incongruence between species abundance and transcriptional activity agrees with an earlier study of taxa–function relationships in the human faecal metatranscriptome[30]. This is expected, as transcription is a highly variable process even among cells of the same species under steady-state conditions[31]. However, heterogeneity among species' transcriptional contributions in otherwise similar metagenomes may be influenced by many factors, including nutrient availability, preferential utilization or xenobiotics; temporal differences in environmental sensing that stagger response to stimuli among species; and metabolic dependence that drives cross-feeding of intermediate metabolites through community members[32,33]. This extends the typical model of transcriptional behaviour from individual microbial or metazoan cell populations to niche ecology.

A feature of the metatranscriptome, which is critical for human microbiome population studies, was the propensity of different organisms to appear as primary transcribers of pathways between individuals. In relating metagenomic features to the metatranscriptome, we observed only 44% of the 'core' metagenomic potential (81 transcribed pathways out of 182 prevalent metagenomically) to be transcribed in the cohort. In combination with previous studies of 'core' metagenomics[18,34,35], this suggests a functional ecological model in which a prevalent metagenome encoding substantial redundancy is distributed among many microorganisms per individual, with the microorganisms containing this core varying among hosts. Transcription at any one time or in any given environment is then typically dominated by one or a few members. This model of microbial ecology would be analogous to silencing versus upregulation of distinct portions of the human genome among cell types, which also consists of a 'core' underlying DNA genome with long-term (epigenetics, instead of phylogeny) and short-term (transcriptional) regulatory mechanisms[36].

In addition to these insights into metatranscriptional activity, stool taxonomic profiles in the MLVS cohort remained diverse and stable, which is in agreement with previous studies of the microbiome in elderly individuals[20]. However, few profiles of this unique ecosystem have yet been generated, and those that do exist tend to use widely varied technical characteristics and study design, prohibiting direct comparisons. ELDERMET[37], for example, posited

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

a trend towards Bacteroidetes dominance in ageing, but this replicated neither across technologies within this cohort nor in our MLVS data. As MLVS data are drawn from within the broader HPFS, for which several decades of dietary and environmental data are available, we anticipate that future studies focusing on these detailed metadata will further detail microbial links to lifestyle and nutrition.

A challenge going forward will thus be to identify epidemiological contexts in which metatranscriptomic features are specifically informative, and the appropriate ways in which to measure them. This might include, for example, tests for health outcomes that are predicted uniquely by metagenomic activities. These may, based on this study's results, be functionally consistent but contributed by different microorganisms across individuals, even when not differentially represented in underlying metagenomes. It also remains to associate metatranscriptomic responses with detailed information on immediate lifestyle exposures, such as recent diet, to determine temporal responses of the metatranscriptome to key environmental perturbations. Ultimately, if there are health outcomes for which causal molecular mechanisms are uniquely detectable in the faecal metatranscriptome, its functional profile will need to be better characterized and integrated as a measurement in human epidemiological population studies.

Methods

MLVS cohort, stool sample collection, shotgun sequencing and quality control. The HPFS is a prospective cohort study aimed at investigating the determinants of men's health, into which 51,529 US men aged 40–75 yr were recruited in 1986 and subsequently followed biennially[13]. For this analysis, we used data from a sub-study of the HPFS, the MLVS, in which 308 participants provided up to two pairs each of self-collected stool samples from consecutive bowel movements, during 2012. The second pair of samples was collected approximately 6 months after the first. The median time between consecutive bowel movements for a pair of samples was 48 h; collection dates are in Supplementary Table 2. At the time of collection, the age of participants ranged between 65 and 81 yr. Cohort details, sample collection and immediate ex situ conservation of metagenomic and metatranscriptomic components, laboratory handling, and paired-end (100 × 100 nucleotides (nt)) shotgun sequencing of RNA and DNA are detailed in the companion manuscript[14] and in our pilot study[9], respectively. Study protocol 22067–102 titled "Men's Lifestyle Validation Study and Microbiome Correlation" was approved by the Harvard Chan School of Public Health Institutional Review Board, and informed consent was obtained from all participants.

The gut microbiome, as captured by stool, was sampled from 308 male participants (aged 65–81 yr) within the MLVS sub-cohort of the HPFS (Fig. 1). Each participant provided up to two pairs of self-collected stool samples from consecutive bowel movements; with the second pair of samples collected approximately 6 months after the first. DNA was extracted from all 929 resulting samples, in addition to RNA from a subset of 372 samples spanning 96 participants. Illumina HiSeq sequencing yielded a total of 4.5 Tera nt of paired-end reads (100 × 100 nt). This included an average of 3.8 Gnt ± 1.5 Gnt (mean ± s.d. Giga nt) before quality filtering (see below) and 1.9 Gnt ± 0.7 Gnt afterward per metagenome, and 3.0 Gnt ± 2.4 Gnt and 1.3 Gnt ± 1.0 Gnt before and after quality control for metatranscriptomes, respectively. Forty-one samples (16 DNA and 25 RNA) had <1 million reads after quality filtering and were excluded from further analysis. Thus, the final data sets analysed comprised 913 metagenomes and 347 metatranscriptomes.

Taxonomic and functional profiling of metagenomic and metatranscriptomic samples. Sequence reads were passed through the KneadData v0.3 quality control pipeline (http://huttenhower.sph.harvard.edu/kneaddata), which incorporates the Trimmomatic[38] and BMTagger[39] filtering and decontamination algorithms to remove low-quality read bases (thresholding Phred quality score at <20) and remove reads of human origin, respectively. Trimmed non-human reads shorter than 70 nt were discarded. Taxonomic profiling was performed using the MetaPhlAn2 classifier[15], which relies on approximately 1 million clade-specific marker genes derived from 17,000 microbial genomes (corresponding to >7,500 bacterial, viral, archaeal and eukaryotic species) to unambiguously classify metagenomic reads to taxonomies and yield relative abundances of taxa identified in the sample. We quality-controlled taxonomic profiles by requiring at least 10% of clade-specific markers to recruit at least 1 read per kilobase (RPK) for inclusion in subsequent analyses. In addition to DNA, RNA (complementary DNA (cDNA)) reads were also analysed with MetaPhlAn2 to quantify RNA viruses.

Metagenomes and metatranscriptomes were functionally profiled using HUMAnN2 (ref. [16]) to quantify genes and pathways (http://huttenhower.sph.harvard.edu/humann2). Briefly, for each sample, taxonomic profiling is used to identify detectable organisms. Reads are recruited to sample-specific pangenomes, including all gene families in any detected microorganisms using Bowtie2 (ref. [40]). Unmapped reads are aligned against UniRef90 (ref. [41]) using DIAMOND translated search[42]. Hits are counted per gene family and normalized for length and alignment quality. For calculating abundances from reads that map to more than one reference sequence, search hits are weighted by significance (alignment quality, gene length and gene coverage). UniRef90 abundances from both the nucleotide and protein levels were then (1) mapped to level 4 EC nomenclature, and (2) combined into structured pathways from MetaCyc[17]. We used the MinPath[43] and gap filling options in HUMAnN2 version 0.8.0. For the purposes of functional profiling, each read can be (1) mapped to a specific organism's characterized gene family in one or more known pathways, (2) mapped to a characterized protein family (without assignment to a specific organism), (3) mapped to an uncharacterized gene family (not in any pathway), or (4) not mapped to any gene family. HUMAnN2 refers to these as (1) species-specific, (2) unclassified, (3) unintegrated, and (4) unmapped reads, respectively. Per sample breakdown of HUMAnN2 mapping categories (that is, mapped, unclassified, unintegrated and unmapped RPKs) are provided in Supplementary Data set 5. Reads that mapped only at the amino acid level are not used when calculating specific taxa's functional contributions. Instead, only reads that mapped (unambiguously) at the nucleotide level are included in these totals.

Quantification of metatranscriptomic functional activity. Metatranscriptomic functional activity was assessed in the 341 samples with both RNA and DNA data in a manner not unlike two-channel microarrays using RNA/DNA ratios (see RNA/DNA normalization below). Owing to the compositionality of RNA and DNA measurements, the resulting ratio is relative to the mean transcript abundance of the entire microbial community. That is, a ratio of 1 implies that the pathway is transcribed at the mean transcription abundance of all pathways in the microbial community. These quotients of RNA/DNA feature abundances allowed unbiased comparison of transcript abundances of metagenomic features between samples and also provided a comparative index of over/undertranscription (relative to DNA copy number) within individual microbiome samples. Pathways that had <1 RPK of either RNA or DNA were treated as not detected in the analyses.

RNA/DNA normalization. Metatranscriptomic features (that is, RNA abundances of genes, enzymes and pathways) were normalized to corresponding metagenomic features to obtain an estimate of the mean transcription abundance λ as follows:

$$\lambda_{f,i} = \frac{R_{f,i} H\left(R_{f,i} - t\right)}{\sum_i R_{f,i}} \times \frac{\sum_i D_{f,i}}{D_{f,i} H\left(D_{f,i} - t\right)}$$

$R_{f,i}$ and $D_{f,i}$ are the counts, in RPK, of feature f in sample i, for the metatranscriptome and metagenome respectively. $H(x)$ is a unit step function with threshold x, and t is the detection threshold, here set to 1 RPK. This threshold ensures that RNA and DNA abundances that are confidently quantified (>1 RPK) are included, reducing the effect of increased noise due to genes and transcripts with low sequencing coverage. Summary statistics and analyses were only performed where both the numerator and denominator are measurable, although such gene families were tracked in RNA or DNA alone, respectively. Note that a value $\lambda_{f,i} = 1$ does not imply that the feature has an equal number of copies in RNA as it does in DNA. Rather, it implies that the mean transcription abundance is equal to the global mean transcription abundance of all organisms in the community in sample i. Thus, $\lambda_{f,i} > 1$ implies a transcript abundance above the community mean, which may be different in different samples.

EC dispersion. Co-expression of functionally related ECs was quantified by the mean variance of the standardized EC expression log-ratios for the set of ECs contributing to a given pathway. Specifically:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P|-1}\sum_{c\in P}(z_{c,i}-z_{\bullet,i})^2,\ z_{c,i}=\frac{x_{c,i}-x_{\bullet,i}}{\sqrt{\mathrm{Var}[x_{c,i}]}}$$

where $x_{c,i}$ is the expression log-ratio for EC c in sample i, P is a given pathway, $\langle\bullet\rangle$ represents the mean, and $\mathrm{Var}[\bullet]$ represents the variance. When the expression of functionally related ECs is not related (that is, uncorrelated), then this value is expected to be 1. Lower values indicate the presence of co-expression of ECs, with 0 indicating a perfect relationship.

Species-specific meta'omic concordance. Concordance between species-level metagenomic and metatranscriptomic pathway abundances was assessed using the mean of the Spearman correlation, weighted by the mean metagenomic

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

contribution of each species to the overall pathway abundance. After averaging across multiple samples per subject, we calculated:

$$\mathrm{WSpear}(p) = \frac{\sum_s \left[\mathrm{Spearman}\left(d_{p,s,\bullet}, r_{p,s,\bullet}\right) \sum_i d_{p,s,i}\right]}{\sum_s \sum_i d_{p,s,i}}$$

where $d_{p,s,i}$ and $r_{p,s,i}$ are the relative abundances of pathway p, contributed by species s in sample i in DNA and RNA, respectively. Spearman is the Spearman correlation between two vectors, where ties are given the mean rank of the tied values, and defined to be 0 when either vector has no variance. This weighting downweights the concordance for species that do not contribute much of the pathway's abundance, mitigating the uncertainty inherent in estimating transcript abundances of low-abundance species and genes.

Microbial ecology networks. Ecological covariation was assessed using BAnOCC, a Bayesian model for detecting significant pairwise associations in compositional data[26] (http://huttenhower.sph.harvard.edu/banocc). Briefly, BAnOCC models the sequence generation process using a log-normal distribution on unobserved absolute counts, and constrains the associated correlation matrix through a sparsity-inducing prior. For posterior inference, we use the 95% credible interval, that is, a correlation estimate is considered significant if the corresponding 95% credible interval excludes zero. We estimated ecological networks for the two sampling time points independently, from within-subject means of species abundance profiles, and then overlaid the networks to assess similarities and differences between networks over the length of the sampling period.

Inference of strain-level population structure. Strain profiling was carried out using StrainPhlAn v1.0 (ref. [29]) (http://segatalab.cibio.unitn.it/tools/strainphlan). Briefly, after mapping reads to MetaPhlAn2 species-specific markers for sufficiently abundant species in each sample, a per-sample consensus sequence is built for each marker. For each species, these are concatenated, aligned, and variants identified relative to reference. Here, pairwise evolutionary distances were calculated from these variant alignments, with the Kimura two-parameter distance[44] for ordination analysis using R packages vegan and ggplot2.

Structure of the stool metagenome and metatranscriptome as contributed by diverse species. The input for the joint ordination of pathways and species (Fig. 3a) was the pathway genomic and transcriptional abundance of known taxonomic provenance only, averaged over 913 metagenomes and 341 metatranscriptomes. Species contributing $<1.0\times10^{-7}$ metagenomic relative abundance to a pathway in <5% of pathways were removed, and the same criteria were used to remove metagenomic pathways found in species. Out of 339 species that were found to contribute to 219 pathways, 223 (65.8%) species and 109 (34.2%) pathways satisfied the filtering criteria. The metatranscriptome pathway by species matrix was subset to the same 109 pathways as for metagenomes, which were contributed by 194 species, and merged with the DNA pathway by species matrices into a single input matrix.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Sequence data have been deposited in the Sequence Read Archive under BioProject accession PRJNA354235. Data from the HPFS, including metadata not included in the current manuscript but collected as part of the MLVS, can be obtained through written application. As per standard controlled access procedure, applications to use HPFS resources will be reviewed by our External Collaborators Committee for scientific aims, evaluation of the fit of the data for the proposed methodology, and verification that the proposed use meets the guidelines of the Ethics and Governance Framework and the consent that was provided by the participants. Investigators wishing to use the HFPS/MLVS cohort data are asked to submit a brief (2 pages) description of the proposed project ('letter of intent') to Eric Rimm, HPFS Director (erimm@hsph.harvard.edu).

Received: 27 March 2017; Accepted: 23 November 2017;
Published online: 15 January 2018

References

1. O'Doherty, K. C., Virani, A. & Wilcox, E. S. The human microbiome and public health: social and ethical considerations. *Am. J. Public Health* **106**, 414–420 (2016).
2. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroen.* **31**, 69–75 (2015).
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
4. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
5. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
6. Korpela, K. et al. Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. *Nat. Commun.* **7**, 10410 (2016).
7. Satinsky, B. M. et al. Microspatial gene expression patterns in the Amazon River plume. *Proc. Natl Acad. Sci. USA* **111**, 11085–11090 (2014).
8. Turnbaugh, P. J. et al. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl Acad. Sci. USA* **107**, 7503–7508 (2010).
9. Franzosa, E. A. et al. Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111**, E2329–E2338 (2014).
10. Segata, N. et al. Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**, 666 (2013).
11. Haiser, H. J. et al. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* **341**, 295–298 (2013).
12. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
13. Chan, A. T. et al. Aspirin dose and duration of use and risk of colorectal cancer in men. *Gastroenterology* **134**, 21–28 (2008).
14. Mehta, R. et al. Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* (in press).
15. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
16. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
17. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
18. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
19. Claesson, M. J. et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178–184 (2012).
20. Claesson, M. J. et al. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl Acad. Sci. USA* **108**, 4586–4591 (2011).
21. Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157**, 142–150 (2014).
22. McCarty, R. M. & Bandarian, V. Biosynthesis of pyrrolopyrimidines. *Bioorg. Chem.* **43**, 15–25 (2012).
23. Vinayak, M. & Pathak, C. Queuosine modification of tRNA: its divergent role in cellular machinery. *Biosci. Rep.* **30**, 135–148 (2009).
24. Hauryliuk, V., Atkinson, G. C., Murakami, K. S., Tenson, T. & Gerdes, K. Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nat. Rev. Microbiol.* **13**, 298–309 (2015).
25. Chistoserdova, L., Kalyuzhnaya, M. G. & Lidstrom, M. E. The expanding world of methylotrophic metabolism. *Annu. Rev. Microbiol.* **63**, 477–499 (2009).
26. Faust, K. et al. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
27. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl Acad. Sci. USA* **110**, 12804–12809 (2013).
28. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
29. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
30. Gosalbes, M. J. et al. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE* **6**, e17447 (2011).
31. Sanchez, A. & Golding, I. Genetic determinants and cellular constraints in noisy gene expression. *Science* **342**, 1188–1193 (2013).
32. Pande, S. et al. Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J.* **8**, 953–962 (2014).
33. D'Souza, G. & Kost, C. Experimental evolution of metabolic dependency in bacteria. *PLoS Genet.* **12**, e1006364 (2016).
34. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
35. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
36. Pimentel, D. Population regulation and genetic feedback. *Science* **159**, 1432–1437 (1968).
37. O'Toole, P. W. & Jeffery, I. B. Gut microbiota and aging. *Science* **350**, 1214–1215 (2015).
38. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
39. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

41. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
42. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
43. Ye, Y. & Doak, T. G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* **5**, e1000465 (2009).
44. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
45 Schwager, E., Mallick, H., Ventz, S. & Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLoS Comput. Biol.* **13**, e1005852 (2017).

Acknowledgements

We thank the participants in the MLVS and the HMP who graciously contributed to this research. This work was supported by funding from STARR Cancer Consortium Award no. I7-A714 to C.H., NCI R01CA202704 (A.T.C., C.H. and J.I.), NIDDK DK098311 (A.T.C.), and NIDDK U54DE023798 (C.H.). J.I. is further supported by Nebraska Tobacco Settlement Biomedical Research Development Funds. K.L.I. is supported by the National Health and Medical Research Council. Components of the Men's Lifestyle Validation Study were supported by NCI U01CA152904 and UM1 CA167552. R.S.M. is supported by a Howard Hughes Medical Institute Fellowship Award. We are also grateful for initial pilot funding provided by B. Wu and E. Larsen. A.T.C. is a Stuart and Suzanne Steele MGH Research Scholar.

Author contributions

Study design and management were by J.I., A.T.C. and C.H. Sample collection and data generation were performed by K.L.I., D.A.D., C.D., E.R. and J.I. Data analysis was conducted by G.S.A.-A., R.S.M., J.L.-P., H.M. and T.B. Manuscript preparation and writing were conducted by G.S.A.-A., R.S.M., J.L.-P., H.M., D.A.D., J.I., A.T.C. and C.H.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-017-0084-4.

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

Correspondence and requests for materials should be addressed to A.T.C. or C.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

nature research

Corresponding author(s): Andrew Chan & Curtis Huttenhower

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

The cohort was pre-existing, detailed in citation 13{PMID:18005960} of the manuscript.
The Health Professionals Follow-up Study (HPFS) is an ongoing prospective cohort study which began in 1986 among 51,529 US male podiatrists, dentists, osteopathic physicians, veterinarians, pharmacists, and optometrists aged 40 to 75 years at enrollment. In this study, participants have returned questionnaires every two to four years with greater than 90% follow-up to provide information about lifestyle and dietary factors, medication use, and diagnoses of colorectal cancer and other diseases. The Men's Lifestyle Validation Study (MLVS) was established among 700 men aged 52 to 81 years (median 69 years) nested within HPFS who had completed the 2010 food frequency questionnaire and previously provided a blood sample.

2. Data exclusions

Describe any data exclusions.

Men with coronary artery disease, stroke/TIA, cancer (except squamous or basal cell skin cancer), or major neurological disease (ALS, Alzheimer's, Parkinson's, epilepsy, MS) were excluded.
Sequencing data was excluded if samples had insufficient input material after DNA/RNA extraction or with insufficient sequencing depth (<2 million reads) after sequencing.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Sample collection and data generation was validated in a previous study (Franzosa et al., Relating the metatranscriptome and metagenome of the human gut, PNAS 2014). All code for the taxonomic and functional analysis has been validated and is publicly available (see Methods section for details).

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Patients were consented prior to a screening colonoscopy, which separated them into confirmed IBD patients and non-IBD controls.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not used for the data collection and analysis.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Sequence reads were processed with the KneadData v0.5.1 quality control (QC) pipeline. Taxonomic profiling was performed using the MetaPhlAn2 classifier. Functional profiling of metagenomes and metatranscriptomes was performed using HUMAnN2. Details are provided in the methods section of the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Not applicable.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Not applicable.
b. Describe the method of cell line authentication used.	Not applicable.
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Not applicable.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No research animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The Men's Lifestyle Validation Study (MLVS) was established among 700 men aged 52 to 81 years (median 69 years) nested within HPFS who had completed the 2010 food frequency questionnaire and previously provided a blood sample.