JEM Journal of Experimental Medicine

TECHNICAL ADVANCES AND RESOURCES

Gut microbiome stability and dynamics in healthy donors and patients with non-gastrointestinal cancers

Allyson L. Byrd[1], Menghan Liu[1,2], Kei E. Fujimura[1], Svetlana Lyalina[3], Deepti R. Nagarkar[1], Bruno Charbit[4], Jacob Bergstedt[5], Etienne Patin[5], Oliver J. Harrison[6], Lluís Quintana-Murci[5,7], Ira Mellman[1], Darragh Duffy[4,8], Matthew L. Albert[9], and the Milieu Intérieur Consortium

As microbial therapeutics are increasingly being tested in diverse patient populations, it is essential to understand the host and environmental factors influencing the microbiome. Through analysis of 1,359 gut microbiome samples from 946 healthy donors of the Milieu Intérieur cohort, we detail how microbiome composition is associated with host factors, lifestyle parameters, and disease states. Using a genome-based taxonomy, we found biological sex was the strongest driver of community composition. Additionally, bacterial populations shift across decades of life (age 20–69), with Bacteroidota species consistently increased with age while Actinobacteriota species, including *Bifidobacterium*, decreased. Longitudinal sampling revealed that short-term stability exceeds interindividual differences. By accounting for these factors, we defined global shifts in the microbiomes of patients with non-gastrointestinal tumors compared with healthy donors. Together, these results demonstrated that the microbiome displays predictable variations as a function of sex, age, and disease state. These variations must be considered when designing microbiome-targeted therapies or interpreting differences thought to be linked to pathophysiology or therapeutic response.

Introduction

Microbial therapeutics, including fecal microbiota transplants (FMTs), bacterial consortia, and probiotics, are increasingly being tested in patients with *Clostridium difficile* infections and other gastrointestinal (GI) disorders (Allegretti et al., 2019), including inflammatory bowel disease (IBD) and, more recently, non-GI indications such as autism (Kang et al., 2019) and cancer (Mullard, 2018). In parallel to microbial therapeutics, microbial signatures are being evaluated as a novel class of biomarkers, applied for stratification of efficacy and safety in clinical trials across multiple indications (Ananthakrishnan et al., 2017; Dubin et al., 2016). Notably, this rapid increase in microbial therapeutics and biomarkers demands a rigorous reevaluation of the factors influencing an individual's personal gut microbiome over time. Such understanding is essential for optimizing clinical trials with any microbial component. For example, without a complete understanding of the factors influencing the gut microbiome in health and disease, we cannot determine whether the optimal FMT should be sourced from a patient who previously responded to a therapy or a healthy donor who is matched for age and sex.

In this paper, we present a comprehensive assessment of the gut microbiome of 946 well-defined healthy French donors from the Milieu Intérieur (MI) Consortium, with 1,359 shotgun metagenomic samples. Designed to study the genetic and environmental factors underlying immunological variance between individuals, the MI Consortium comprises 500 women and 500 men evenly stratified across five decades of life, from 20 to 69 yr of age, for whom extensive metadata, including demographic variables, serological measures, dietary information, and systemic immune profiles, are available and easily accessible (Patin

[1]Department of Cancer Immunology, Genentech Inc., San Francisco, CA; [2]Sackler Institute of Graduate Biomedical Sciences, New York University School of Medicine, New York, NY; [3]Personalized Health Care, Genentech Inc., San Francisco, CA; [4]Cytometry and Biomarkers UTechS, CRT, Institut Pasteur, Paris, France; [5]Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, Centre National de la Recherche Scientifique, Paris, France; [6]Center for Fundamental Immunology, Benaroya Research Institute, Seattle, WA; [7]Collège de France, Paris, France; [8]Translational Immunology Lab, Institut Pasteur, Paris, France; [9]Insitro, South San Francisco, CA.

D. Duffy, M.L. Albert, and L. Quintana-Murci are co-coordinators of the Milieu Intérieur Consortium.

The Milieu Intérieur Consortium is composed of the following team leaders: Laurent Abel, Andres Alcover, Hugues Aschard, Kalla Astrom, Philippe Bousso, Pierre Bruhns, Ana Cumano, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Jacques Fellay, Ivo Gomperts-Boneca, Milena Hasan, Serge Hercberg, Olivier Lantz, Hugo Mouquet, Etienne Patin, Sandra Pellegrini, Stanislas Pol, Antonio Rausell, Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Frédéric Tangy, Antoine Toubert, Mathilde Touvier, Marie-Noëlle Ungeheuer, Darragh Duffy, Matthew L. Albert, and Lluis Quintana-Murci. Correspondence to Allyson L. Byrd: byrd.allyson@gene.com; Darragh Duffy: darragh.duffy@pasteur.fr.

Rockefeller University Press
J. Exp. Med. 2020 Vol. 218 No. 1 e20200606
https://doi.org/10.1084/jem.20200606

et al., 2018; Thomas et al., 2015). Integrating these data with those from cancer patients, we demonstrate clear evidence for altered microbial communities in cancer patients across multiple non-GI indications.

To build on the findings of several landmark microbiome studies (Falony et al., 2016; Human Microbiome Project Consortium, 2012; Zhernakova et al., 2016), many of which relied on an older reference library for taxonomic classification of microbial sequence reads (Truong et al., 2015), we leveraged an expanded set of reference genomes with a novel taxonomy that corrects many misclassifications in public databases to discover new biological insights, particularly around age and sex (Parks et al., 2018). Notably, an independent dataset was used for replication of many of the findings (Zeevi et al., 2015). Study of short-term longitudinal samplings from half the donors found that individuals are more similar to themselves over time compared with others (Costello et al., 2009; Flores et al., 2014; Mehta et al., 2018). However, the degree of stability between individuals was quite variable and was influenced by lifestyle factors as well as baseline composition.

Overall, the aims of the study are threefold. First, we introduce a new microbiome analysis approach that uses an expanded set of reference genomes with a novel taxonomy to discover new, statistically robust insights into host/bacteria biology that will enable personalized medicine approaches for microbial therapeutics and biomarkers. Second, we provide the rich metadata and 1,000-plus deep shotgun metagenomic samples described here as a resource on which future microbiome studies can test and build new computational tools, as well as be compared against disease cohorts. Finally, while demonstrating the utility of this resource as a control population, we define global shifts in the gut microbiomes of patients with non-GI tumors compared with healthy donors.

Results

The Genome Taxonomy Database (GTDB) improves taxonomic resolution of *k*-mer–based approaches

Historically, microbial sequencing efforts focused predominantly on a small number of organisms, often causes of nosocomial infections (Fig. S1, A and B). By contrast, reference databases, including National Center for Biotechnology Information (NCBI) GenBank, are increasingly populated with genomic information of commensal microbes (Browne et al., 2016; Forster et al., 2019; Poyet et al., 2019; Zou et al., 2019). As genome reference databases expand, historical, microbiology-based taxonomic assignments do not reflect population-level relationships inferred from genome sequencing. This is particularly problematic for *k*-mer–based analyses, which use sequence similarity between closely related genomes to infer which taxa are present (Nasko et al., 2018).

To overcome these issues, in lieu of the traditional NCBI taxonomy, we generated a custom reference database of 23,505 RefSeq genomes with GTDB taxonomies (see Materials and methods and Data S1, table 1). Briefly, GTDB is a bacterial taxonomy based on a concatenated protein phylogeny in which polyphyletic groups were removed and taxonomic ranks were normalized on the basis of relative evolutionary divergence (Parks et al., 2018). The impact of this procedure was particularly prominent for species of the genus *Clostridium*, which were split into 121 unique genera spanning 29 families (Parks et al., 2018). This could be especially meaningful for analysis of gut microbiome samples, as *Clostridium* species are prevalent community members and often emerge in association studies.

The RefSeq sequences and taxonomic tree from the GTDB, including its naming conventions, were used to build a reference database for the *k*-mer–based program Kraken2 (Wood and Salzberg, 2014) and read-reassignment step Bracken2 (Lu et al., 2017). This custom Kraken2/GTDB pipeline was applied to 1,359 quality-controlled samples from 946 MI donors (Fig. S1, C–F; Data S1, tables 2 and 3) and compared using both the marker gene–based tool Metaphlan2 (Truong et al., 2015) and Kraken2, with the same 23,505 reference genomes using their original NCBI taxonomies (Fig. S2). Consistently, more bacterial taxa were identified per sample with Kraken2 than Metaphlan2, a result of the updated reference database and higher sensitivity of this *k*-mer–based approach (Fig. S2, A–C; McIntyre et al., 2017). Between the two Kraken databases (GTDB and NCBI), richness varied depending on how taxa were redistributed by GTDB. For example, GTDB split 2,397 NCBI genera into 3,205, while it collapsed 18,795 NCBI species into 13,446 (Fig. S2, A and D). Despite finer-level differences, the overall distribution of phyla across the three approaches was similar (Fig. S2 E), indicating that Kraken2/GTDB pipeline results would be consistent with previous analyses. As such, a combination of *k*-mer–based read assignment and genome-based taxonomy allows higher-resolution analysis of shotgun metagenomic samples.

Variable gut microbiomes in a restricted geographical region

To complement our optimized taxa-based approach and further use the resolution afforded by shotgun metagenomic sequencing, we applied HUMANn2 to identify the functional potential of microbial pathways present in the MI samples (Franzosa et al., 2018; Data S1, table 4). Using both the Kraken2/GTDB and HUMANn2 pipelines, we identified a broad range of diversity across the 946 individuals in this geographically restricted cohort of healthy French adults. This diversity was observed in terms of metabolic pathway richness (282 ± 40, mean ± SD), species richness (248 ± 32), and Shannon diversity (3.7 ± 0.35), which accounts for both richness and evenness (Data S1, table 2). Across donors, our GTDB pipeline confirmed Firmicutes and Bacteroidota (formerly Bacteroidetes) as the most abundant phyla in the gut, but enabled distinction among the original Firmicutes phyla, which was further divided in the GTDB into 12 distinct categories: Firmicutes, Firmicutes_A, Firmicutes_B, … Firmicutes_K (Data S1, table 1). Notably, throughout the GTDB, the group containing type material (if known) kept the original unsuffixed name. Of those, seven were present in this cohort, with Firmicutes_A the most abundant, followed by Firmicutes and Firmicutes_C (Fig. 1 A and Data S1, table 3), highlighting the finer granularity, even at the phylum level, provided by GTDB-based taxonomic calls. Subsequent application of the Bray–Curtis (BC) distance metric, a means to assess species presence/absence in addition to relative abundance across donors, demonstrated that samples fell along a gradient defined by the relative abundances of Firmicutes_A and Bacteroidota, with lesser contributions from Actinobacteriota and Firmicutes (first dimension of multidimensional scaling [MDS] projection; Fig. 1 B).

Using this extensively characterized cohort, we explored how 154 metadata variables (Fig. S3, A–C; and Data S1, tables 5 and 6), including 42 laboratory measurements and 43 dietary variables, contributed to overall bacterial community composition. We identified 52 variables (34% of total) that associated with Kraken2-GTDB species (permutational multivariate ANOVA [PERMANOVA] test false discovery rate [FDR] < 0.05); 51 of which replicated at the genera level (PERMANOVA test FDR < 0.1; Fig. 1 C and Data S1, table 7). The top contributors were age and sex, with lesser contributions from diet, such as consumption of raw fruit and cooked and cured meats, as well as frequency of fast food consumption, in line with previous reports of 16S rRNA analyses from this cohort (Partula et al., 2019; Scepanovic et al., 2019). Notably, sex and age were associated with 24 and 44 of the other metadata variables, respectively, which confounds our ability to dissociate the individual effects of these variables on microbial community composition (Fig. S3, D and E). In total, these factors explained <10% of population variability, indicating that the majority of variance in community composition remains unexplained. Drawbacks of this analysis are the absence of Bristol stool score, a measure of stool consistency, and levels of chromogranin A, a protein secreted by enteroendocrine cells, the factors most associated with community composition in previous European cohorts (Falony et al., 2016; Zhernakova et al., 2016). Although genetic data were also available for these donors, they were not considered here based on previous analyses that the effects of host genetics on microbiome are minimal in this (Scepanovic et al., 2019) and other (Rothschild et al., 2018) cohorts, in part owing to the small population sizes by genome-wide association study standards (Goodrich et al., 2017).

In this healthy cohort, medication usage was low, with only 28% of individuals (n = 266) taking medication of any kind. Notably, donors were excluded if they used antibiotics in the 3 mo preceding enrollment. Of all medications, only oral contraception was taken by >10% of participants (n = 111). In premenopausal women, oral contraception was taken by 36% (110/303) and explained 0.005% of the variance (P = 0.04). In contrast, relatively common medications, β-blockers and proton pump inhibitors, were taken by only seven and four individuals, respectively. Despite this, medication usage was a significant, albeit minor, contributor (R^2 = 0.003) to microbial community composition, highlighting how xenobiotics can and do influence the gut microbiome (Jackson et al., 2018; Maier et al., 2018).

MI bacterial profiles are comparable to those from Israeli donors

To determine whether MI bacterial profiles were unique to this population or comparable with other non-European healthy cohorts, we ran our Kraken2/GTDB pipeline on 1,159 samples from 851 Israeli donors originally published by Zeevi et al. (2015) (Data S1, table 8), for whom age, sex, and body mass index (BMI) were provided (Fig. 2, A and B). After accounting for sequencing depth (mean read count: MI, 13.9 ± 2.9 million; Zeevi, 13.5 ± 7 million), we found that richness across taxonomic levels was consistently elevated in the MI samples, even though the percentage of unmapped reads was comparable (MI 42.2% versus Zeevi 39.6%; Fig. 2, C–E). More specifically, we identified on average 24 more species in samples from the MI donors than from the Zeevi cohort. In addition to potential technical and lifestyle reasons, this discrepancy could reflect the stricter inclusion and exclusion criteria, and thus the greater overall health, of the MI donors (Thomas et al., 2015; Zeevi et al., 2015).

On the whole, community composition, including relative taxa abundances and β diversity, was consistent across both cohorts (Fig. 2, F–H). Notably however, the contributions of age and sex to community composition were almost two times greater in MI than Zeevi (age: R^2 = 0.0087 versus 0.0038; sex: R^2 = 0.011 versus 0.0066; Fig. 2 I), highlighting how stratification of age and sex in the MI cohort provided enhanced statistical power to identify new correlations (Fig. 2, A and B; Zeevi et al., 2015). Despite technical differences, as well as geographic and cultural distinctions between these cohorts, our findings demonstrate a comparable makeup of the gut microbiome. This allowed us to use the Zeevi samples as a replication cohort to demonstrate the reproducibility of our findings in MI.

Prevotella species are more abundant in male donors

Given that sex and age were the variables most strongly associated with bacterial community composition in healthy individuals, we leveraged the statistical power of the MI cohort to explore which taxa were differentially abundant between sexes and across decades of life. To identify bacteria differentially abundant between the 473 females and 473 males, we conducted DESeq2 analysis using age and BMI as covariates (Love et al., 2014) on 485 abundant species (prevalence >5% and mean relative abundance >0.01%; Data S1, table 9). Of the 71 differentially abundant species (FDR < 0.05), 5 were more abundant in females, while 18 were more abundant in males, with $\log_2$ fold change >1 (Fig. 3 A). In total, 11 of 32 prevalent *Prevotella* species were more abundant in males than females, corresponding to a greater overall richness or number of unique *Prevotella* species in males (Fig. 3, B and C). Similarly, in the Zeevi cohort, five species of *Prevotella* were more abundant in males (Fig. 3 B and Data S1, table 10; Zeevi et al., 2015). Notably, even when a species was significantly differentially abundant between sexes in only one cohort, the direction of this trend was also consistent in the other, indicating that higher *Prevotella* abundance in males compared with females is a biological phenomenon consistent across multiple species and populations. This information increases the granularity of trends presented in two previous studies, one that identified *Bacteroides-Prevotella* as broadly more abundant in males than females based on 16S rRNA-targeted oligonucleotide probes (Mueller et al., 2006), and another that found males were three times more likely to have an enterotype consisting of fewer *Bacteroides* and higher *Prevotella* (Ding and Schloss, 2014). Although the factors driving preferential colonization of *Prevotella* in males are unknown, from these data we could generate hypotheses surrounding the roles of gonadal

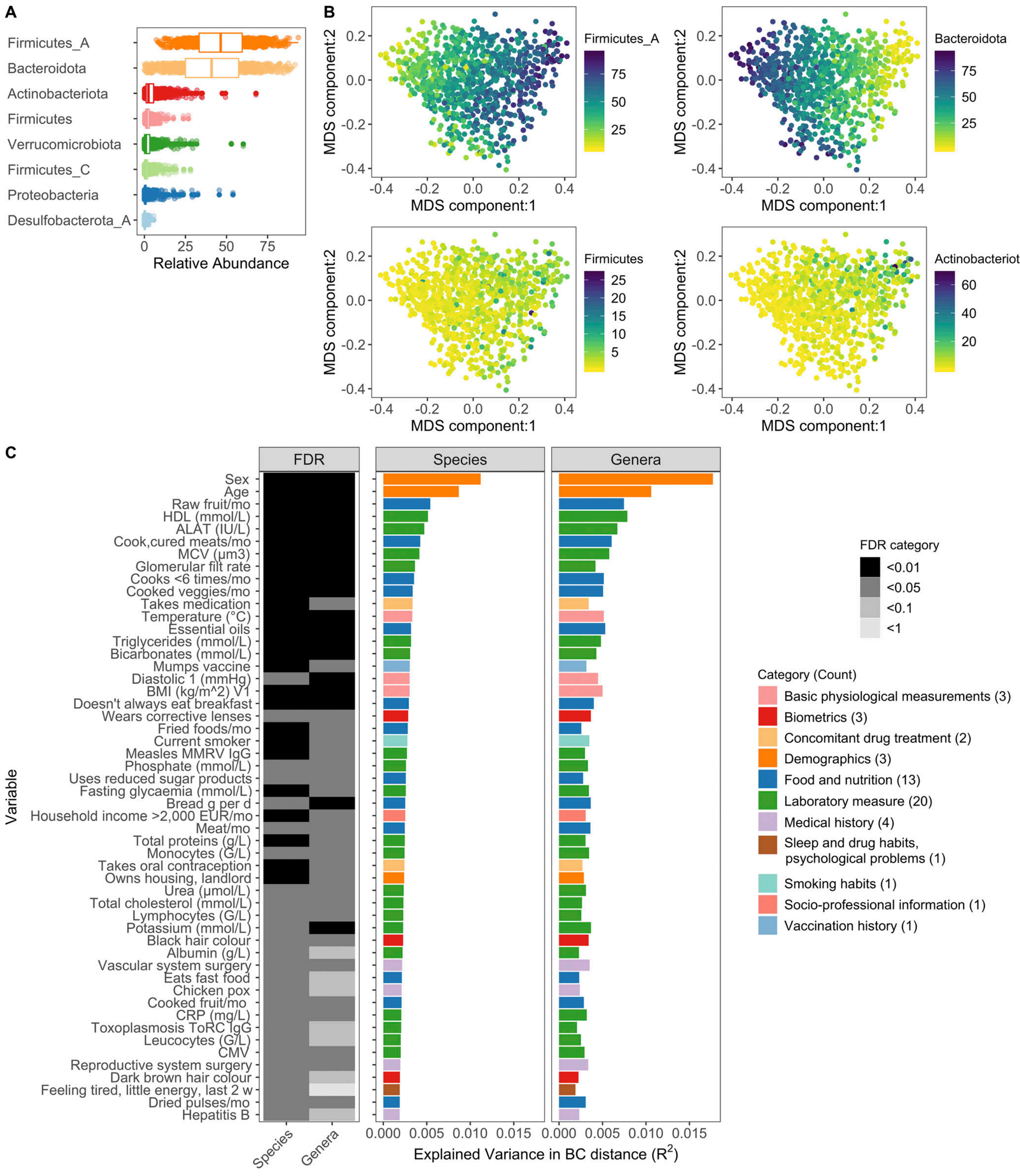


Figure 1. **Interindividual variation of bacterial composition is associated with many factors. (A)** Boxplots of the top eight phyla. Each dot corresponds to one donor. Firmicutes_A, Firmicutes_C, and Firmicutes were split into unique phyla by the GTDB. **(B)** MDS plots of BC distance of bacterial species composition. Ordination was primarily driven by the top two phyla Firmicutes_A and Bacteroidota. Each dot corresponds to one donor while color indicates relative abundance of each phyla. **(C)** In total, 52 factors (Benjamini–Hochberg FDR < 0.05) were associated with interindividual variation of the gut microbiome. The bar plots indicate the amount of interindividual variance explained by each factor for the species and genera level BC distance. Variables are ordered by the percentage variance explained in Kraken species. Colors of the bars correspond to the broad metadata category. The rectangles to the left indicate the statistical strength, as measured by FDR, of the association. For each variable, samples with NA values were excluded. For all these analyses, one sample per donor was used: when available, V1; if not V2. See also Fig. S1, Fig. S2, Fig. S3, and Data S1, tables 1–7. CMV, cytomegalovirus; CRP, C-reactive protein; EUR, Euro; HDL, high-density lipoprotein; MCV, mean corpuscular volume; MMRV, measles, mumps, rubella, and varicella vaccine.

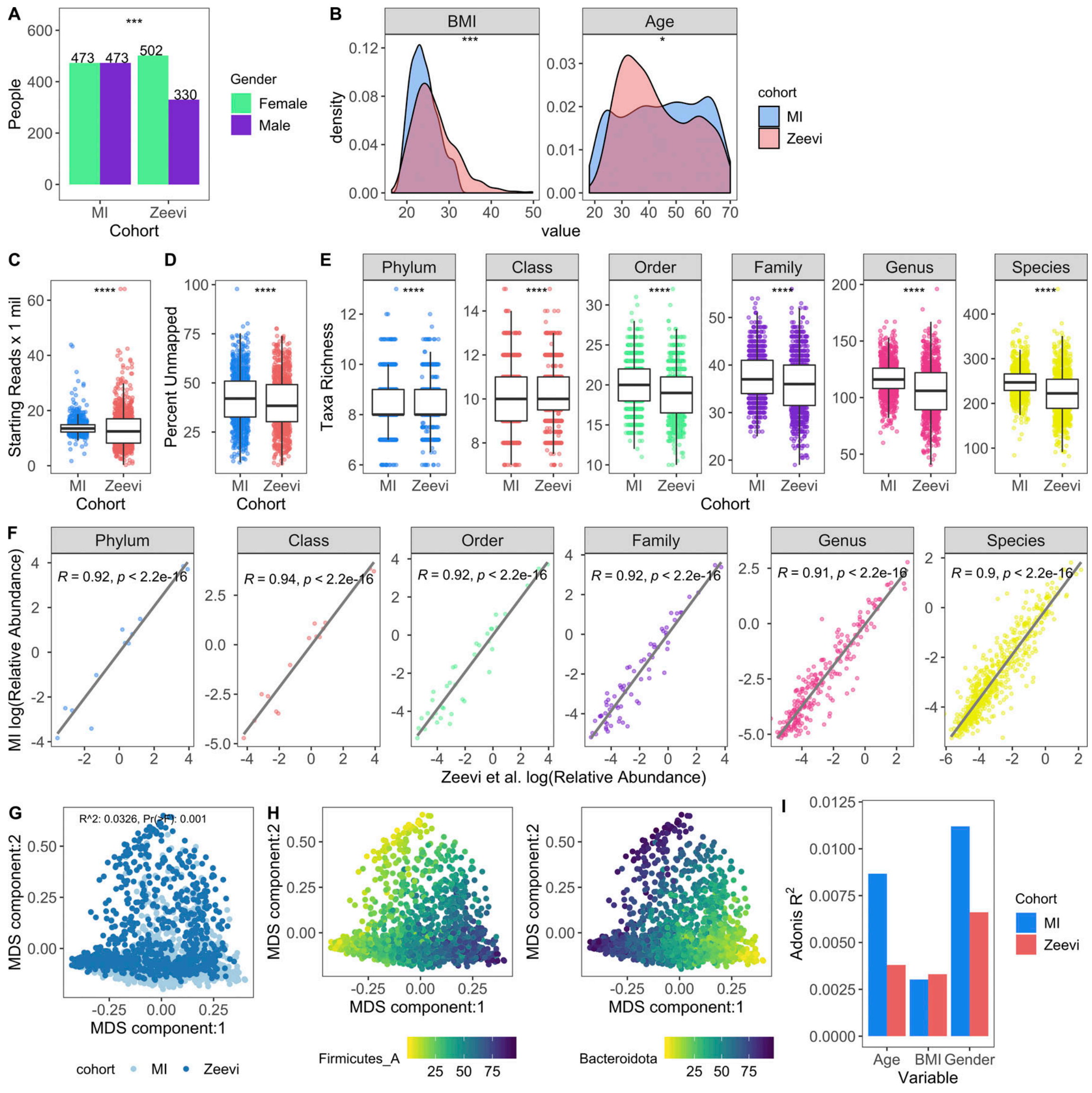


Figure 2. **Microbial profiles of MI donors in comparison to those from Zeevi et al. (2015). (A)** Bars indicate the number of females and males in both cohorts; P value by Fisher's exact test; ***, P < 0.001. **(B)** Density plots show distribution of BMI and age in both cohorts; P value by Wilcoxon rank sum; *, P < 0.05; ***, P < 0.001. **(C)** Bar plots compare sequencing depth across cohorts. **(D)** Bar plots show percentage of reads unmapped after the Kraken2-GTDB pipeline. **(E)** Boxplots show richness across taxonomic levels. Each dot corresponds to one donor. **(C–E)** P values by Wilcoxon rank sum; ****, P < 0.0001. **(F)** Association of relative abundances of each taxon across taxonomic levels with Spearman correlation. Each dot corresponds to one taxon. Only taxa present in >5% of either cohort were considered. Log relative abundances are shown. **(G)** MDS ordination plot of all samples in both the MI and Zeevi cohorts. Each dot corresponds to one donor and is colored by the cohort. R^2 indicates the amount of interindividual variation (calculated with BC) explained by the cohort and was calculated with the PERMANOVA test adonis. **(H)** MDS ordination plot of samples in both the MI and Zeevi cohorts. Each dot corresponds to one donor and is colored by the relative abundance of the phyla. **(I)** Bars indicate the amount of interindividual variation (calculated with BC) explained by each of the variables in each of the cohorts. See also Data S1, table 8.

hormones and microbial community composition. In support of this hypothesis, in a longitudinal study of the oral microbiome, serum levels of testosterone in boys and estradiol and progesterone in girls were positively correlated with levels of *Prevotella intermedia* (Nakagawa et al., 1994).

When considering 364 prevalent metabolic pathways (prevalence >5%), we identified 65 (FDR < 0.05) that were differentially abundant between the sexes (Data S1, table 11). Of those, the pathway CRNFORCAT-PWY: creatinine degradation I was the most strongly enriched in men (Fig. 3 D). Biologically, this is consistent with men having higher blood levels of creatinine (Fig. 3 E). Across both sexes, but not in each individually, circulating creatinine levels were significantly associated with the abundance of the CRNFORCAT-PWY pathway (both sexes: Spearman ρ = 0.087, P = 0.0014; men: ρ = 0.028, P = 0.47; women: ρ = −0.047, P = 0.22). Of species positively associated with this pathway, many were more abundant in males, including the top species *Holdemanella biformis*, which was also correlated with circulating creatinine levels (Spearman ρ = 0.13, P = 4.1 × 10^{-5}; Fig. 3, F and G; and Data S1, table 11). Overall, this

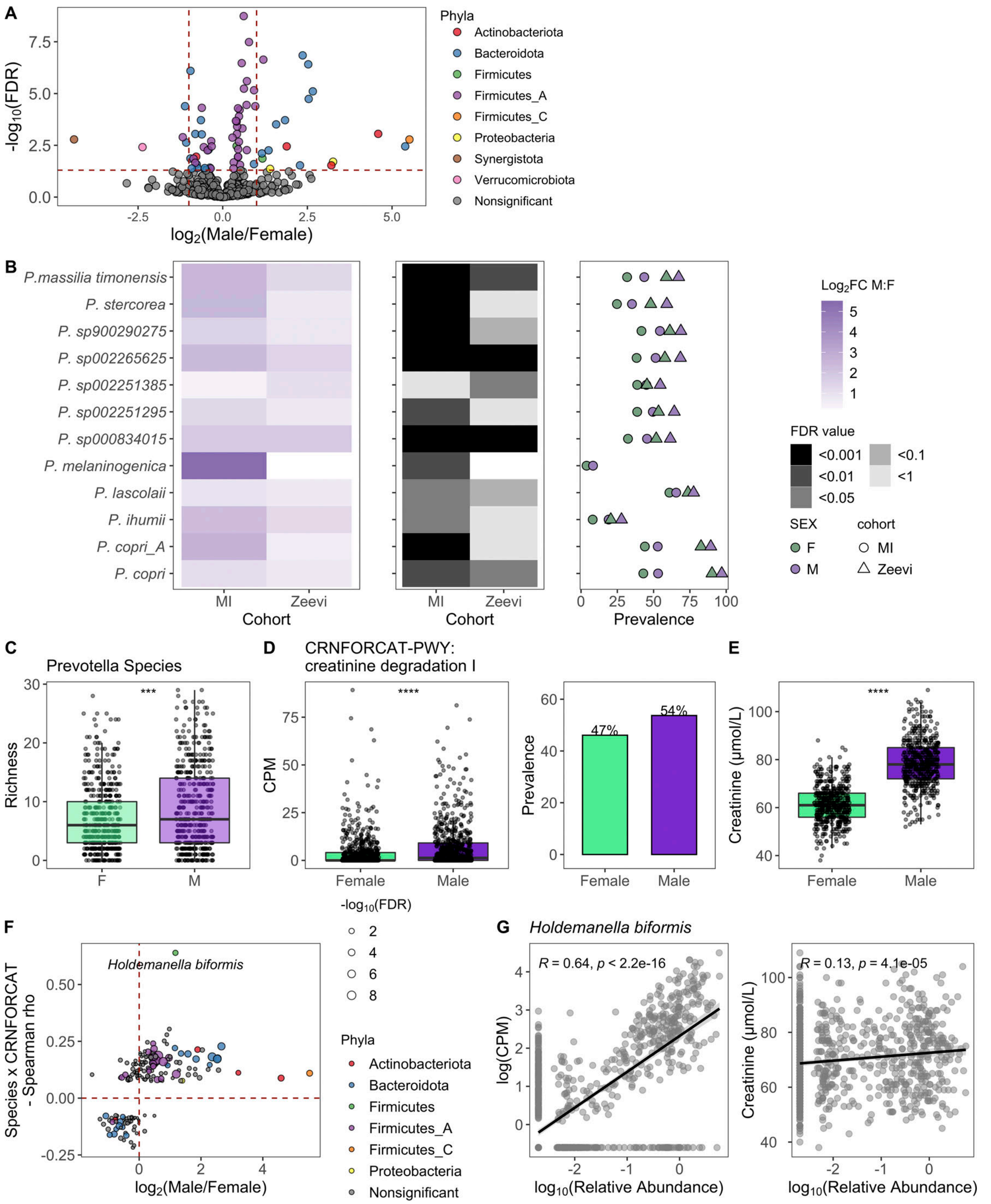


Figure 3. **Taxa, particularly *Prevotella*, were differentially present between males and females. (A)** Volcano plot of 485 abundant species, of which 71 were differentially abundant between males and females based on DESeq2 (FDR < 0.05). Each species is colored by its taxonomic phyla. **(B)** 12 *Prevotella* species were more abundant in males consistently across cohorts. Left: Log_2 fold change (log_2FC) of species in males versus females. Middle panel indicates the FDR value. Notably, even when a species was significant in only one cohort, the direction was consistent in the other. Right: Prevalence of each species. Color indicates the sex, while shape indicates the cohort. **(C)** Boxplots show richness of *Prevotella* species in males and females. **(D)** Abundance measure in counts per million (CPM) and prevalence of the pathway CRNFORCAT-PWY: creatinine degradation I in males and females. **(E)** Blood creatinine levels in males and females. **(C–E)** P values by Wilcoxon rank sum; ***, P < 0.001, ****, P < 0.0001. **(F)** In the scatter plot, dots represent the 189 species that were significantly correlated with CRNFORCAT-PWY. Of those, species that were significantly differentially abundant between males and females are colored by their phyla label. Species not associated with sex are colored gray. Size of the point corresponds to the $-log_{10}$(FDR) of the species by sex correlation. **(G)** Scatter plots show the correlation of *Holdemanella biformis* relative abundance and CPM of CRNFORCAT (left) or circulating creatinine levels (right). Trend lines show 95% confidence intervals and were modeled with lm. Statistics based on Spearman correlation. See also Data S1, tables 9–11.

exemplifies how adaptation to use available nutrients may influence microbiome composition.

The gut microbiome is dynamic across decades of life

It is well characterized that the composition of the gut microbiome differs dramatically between newborns and adults, with the neonatal microbiome transitioning to a more adult-like state upon consumption of solid food and cessation of breastfeeding (Bäckhed et al., 2015; Stewart et al., 2018). However, how the gut microbiome changes throughout adult life has been primarily studied in smaller cohorts by culturing or 16S rRNA analyses (An et al., 2018). The design of the MI cohort provides a unique opportunity to explore how in the absence of underlying disease the gut microbiome is dynamic across the adult decades (20–69 yr old).

In total, we found that 40% of abundant species (192/485) were differently abundant by donor age (Spearman correlations of age by taxa relative abundance, FDR < 0.05); consistent results were identified with linear models taking into account sex and BMI (Data S1, table 12). Notably, two of the top five phyla (Bacteroidota and Actinobacteriota) experienced shifts in relative abundance across the decades (Fig. 4, A and B). Transitions were most pronounced around 40–50 yr old, a time span when many people experience the preclinical stages of chronic diseases, and women begin to experience hormonal changes associated with onset of menopause; average age in this cohort 50 ± 4.2 yr. Across phyla, the correlations with age were conserved across sexes and cohorts (Fig. 4, C and D). For example, relative abundances of Bacteroidota species were primarily increased with age, while Actinobacteriota, including 15 species of *Bifidobacterium*, were decreased with age (Fig. 4 B). This gradual decline of *Bifidobacterium* was true in terms of both relative abundance as well as presence/absence of individual species (Fig. 5 A). In fact, the overall richness or number of unique *Bifidobacterium* species present in an individual steadily declines throughout life (Fig. 5 B). Notably, many *Collinsella* species that were positively correlated with *Bifidobacterium* were also decreased with age (Figs. 4 B and 5 C).

The decline in *Bifidobacterium* prominence with age is particularly interesting in light of *Bifidobacterium* being the dominant bacteria in many newborns, gradually decreasing as infants cease breastfeeding (Stewart et al., 2018). The association of *Bifidobacterium* and old age indicates that the loss of *Bifidobacterium* occurs not only in infants, but continues throughout adulthood (An et al., 2018; Biagi et al., 2010, 2016; Kato et al., 2017; Mueller et al., 2006). Similar to our findings associating *Prevotella* and sex (Fig. 3 B), we built on previous lower-resolution findings to reveal that the trend was consistent across numerous species within the genera and across cohorts (Fig. 5 D and Data S1, table 13), highlighting how the phenomenon is intrinsic to this species. Notably, the only exception, *Bifidobacterium animalis*, is a common probiotic-associated strain (Turroni et al., 2009), rather than a persistent colonizer. Moving forward, comparative genomic analyses between these different species could reveal features associated with colonization in older adults.

We then focused our attention on 364 prevalent microbial pathways (prevalence >5%) and identified 108 that correlated with age (Data S1, table 14), of which 31 were increased and 77 were decreased (FDR < 0.05), including several lactose and galactose degradation pathways (Fig. 5 E). Consistent with the previous results, these pathways were strongly correlated with species in the *Bifidobacterium*, *Collinsella*, and *Blautia* genera (Fig. 5 G and Data S1, table 14). Lower levels of lactose/galactose degradation may explain increased lactose intolerance in older adults and presents a possible opportunity for microbial therapeutic intervention (Gingold-Belfer et al., 2020; Savaiano et al., 2013). Notably in this cohort, the abundance of these pathways was not associated with consumption of dairy products, e.g., milk, cheese, and yogurt (Spearman ρ > 0.3).

Other pathways associated with age were related to L-histidine. In this case, pathways for L-histidine biosynthesis were decreased with age, while those for degradation were increased (Fig. 5 F). Concordantly, the biosynthesis pathway was positively correlated with species decreased with age, while degradation pathways were correlated with species increased with age (Fig. 5 G and Data S1, table 14). In total, these results indicate that gut L-histidine levels may be decreased in older adults, which could lead to an altered immune state, as L-histidine metabolites have been demonstrated to influence colonic inflammation (Gao et al., 2017). Overall, understanding the multitude of microbial correlations with age is incredibly important for appreciating the microbial shifts observed in diseases affecting older individuals.

Short-term stability is variable across donors

To complement our cross-sectional study of the microbiome across the decades, we leveraged longitudinal sampling of roughly half the cohort ($n = 413$) to study short-term (17 ± 3.3 d) dynamics within an individual in the absence of antibiotic exposure. By comparing species BC distances within and between individuals, we found that in the short term, intraindividual differences were less than the interindividual ones (Fig. 6 A). This is consistent with previously published findings (Costello et al., 2009; Flores et al., 2014; Mehta et al., 2018) and also reflected the analysis of relative abundance and presence/absence of metabolic pathways (Fig. 6 A). While within an individual species and pathway, stabilities were highly correlated (Spearman $\rho = 0.75$, $P < 2.2 \times 10^{-16}$), differences between donors were less dramatic at the pathway level, reflecting the more conserved nature of annotated metabolic pathways versus species profiles across individuals (Human Microbiome Project Consortium, 2012).

Although stability was the norm, the degree of species stability (quantified as 1 – BC distance; Mehta et al., 2018) was variable across the 413 donors (Fig. 6, B and C), and as such we investigated which microbial and metadata features underlie this personalized stability trait. Using Spearman correlations, we identified relative abundances of the phyla Firmicutes and Firmicutes_A as enriched at baseline in the less stable donors, while Bacteroidota was higher in donors with greater species stability over time (Fig. 6 D and Data S1, table 15; Flores et al., 2014). Notably, similar trends also were observed when analyzing pathway stability (Fig. 6 E and Data S1, table 15). This is consistent with observations that spore-forming bacteria,

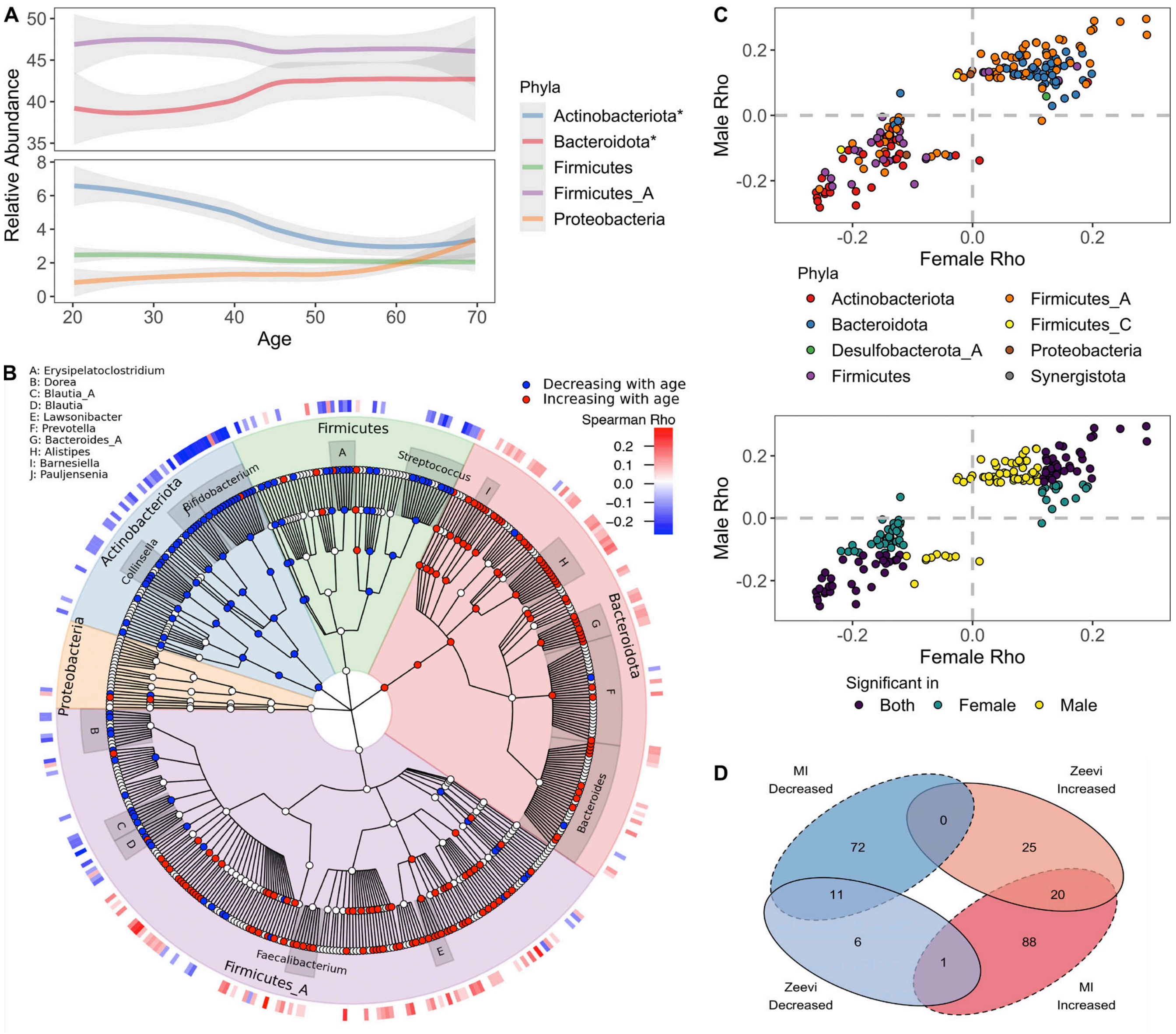


Figure 4. **Bacterial profiles are dynamic across decades of life. (A)** Abundance of the five most abundant phyla across decades of life. Curves show 95% confidence intervals and were modeled with LOESS regression. Stars in the figure legend indicate those phyla statistically associated with age. **(B)** GraPhlAn taxonomic tree of the 454 species in the top five phyla found at significant prevalence and abundance across all donors. Relative abundance of taxa in red were decreased with age, while those in blue increased. Association between taxa relative abundance and age was determined by Spearman correlation (FDR < 0.05). The heatmap on the outer ring indicates the strength of the correlation (Spearman ρ). Genera with at least four species associated with age are labeled. **(C)** Scatter plots compare the Spearman ρ values of bacteria ~ age for males and females, one point = one species. In the bottom plot, points are colored based on whether the correlation was significant (FDR < 0.05) in males, females, or both. In the top plot, points are colored based on the phyla designation of the species. **(D)** Venn diagram comparing the bacterial species statistically (FDR < 0.05) associated with age in the MI and Zeevi et al. (2015) cohorts. See also Data S1, tables 12 and 13.

including many Firmicutes species, are intrinsically less stable (Kearney et al., 2018). Using the generalized linear model (GLM), we compared for the first time how stability is influenced by an extensive list of metadata variables. This revealed that BMI and circulating triglyceride levels were negatively associated with stability, while conversely, consumption of sweet items (e.g., chocolate, sweets, honey, and jam) and raw fruit were positively associated (Fig. 6 F and Data S1, table 16), concordant with diet being a key determinant of human gut microbiome variation (Johnson et al., 2019). While the previous results were only marginally significant (P < 0.05), consistent with previous findings (Flores et al., 2014; Mehta et al., 2018) and ecological theory (McCann, 2000; Schindler et al., 2010; Tilman, 1999), baseline species Shannon diversity was positively associated with stability (FDR = 0.014; Fig. 6 F), i.e., individuals with more diverse communities were more resilient to change than individuals with lower diversity.

We then applied the BC metric to calculate stability (1 – BC) of individual species and pathways (Data S1, tables 17 and 18; Faith et al., 2013; Franzosa et al., 2015). For both species and pathways,

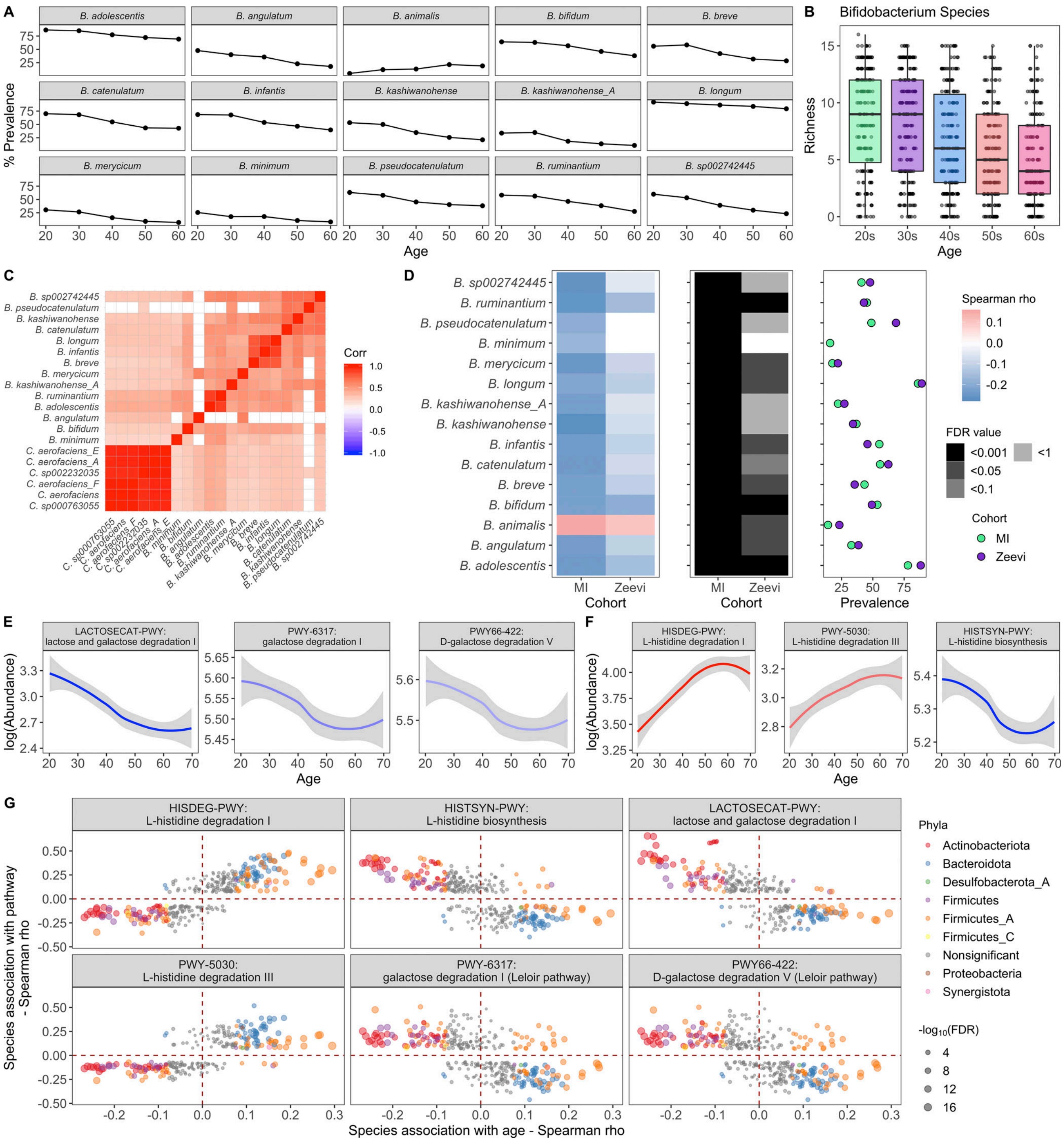
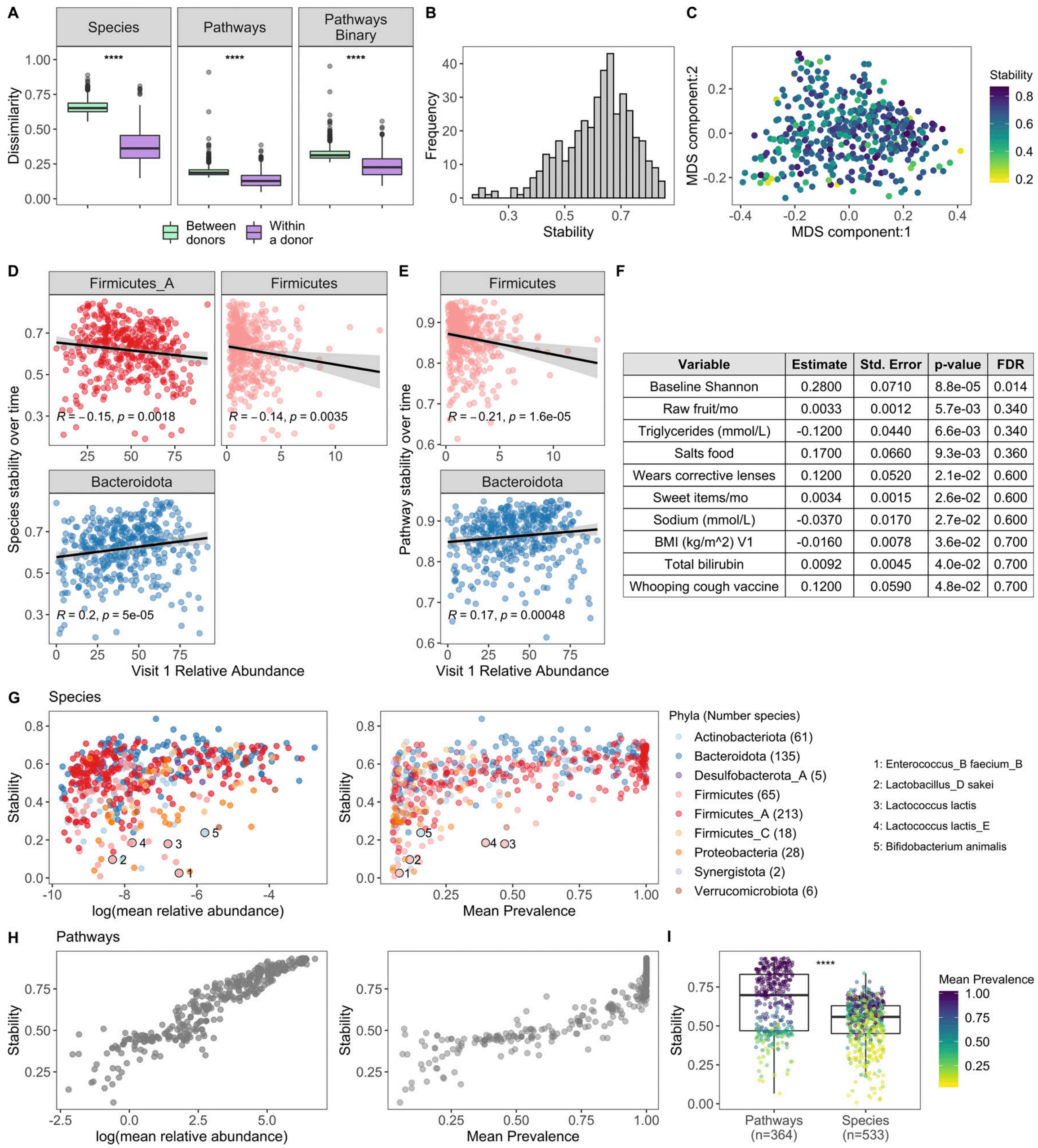


Figure 5. ***Bifidobacterium* species and *Bifidobacterium*-associated pathways decrease with age.** **(A)** Lines indicate the percent prevalence of 15 *Bifidobacterium* species across the different decades. Only *Bifidobacterium* species significantly associated with age are shown. **(B)** Boxplots show richness of *Bifidobacterium* species in donors grouped by their age. **(C)** Correlation plot of *Bifidobacteria* species (prevalence >30%) and their cocorrelated *Collinsella* species (Spearman ρ > 0.4). **(D)** 15 *Bifidobacterium* species were associated with age across cohorts. Left: Correlation of species with age (Spearman ρ). Middle: FDR value. Notably, even when a species was significant in only one cohort, the trend was consistent in the other. Right: Prevalence of each species. Color indicates the cohort. **(E)** Abundance (counts per million [CPM]) of lactose/galactose degradation pathways decreased with age. Curves show 95% confidence intervals and were modeled with LOESS regression. **(F)** Abundance (CPM) of L-histidine degradation/synthesis pathways associated with age. Curves show 95% confidence intervals and were modeled with LOESS regression. **(G)** For the pathways correlated with age in E and F, in the scatter plot, dots represent the species that were significantly correlated with that pathway. Of those, species that were also significantly correlated with age are colored by their phyla label. Species not associated with age are colored gray. Size of the point corresponds to the $-\log_{10}$(FDR) of the species by age correlation. See also Data S1, tables 12–14.



Variable	Estimate	Std. Error	p-value	FDR
Baseline Shannon	0.2800	0.0710	8.8e-05	0.014
Raw fruit/mo	0.0033	0.0012	5.7e-03	0.340
Triglycerides (mmol/L)	-0.1200	0.0440	6.6e-03	0.340
Salts food	0.1700	0.0660	9.3e-03	0.360
Wears corrective lenses	0.1200	0.0520	2.1e-02	0.600
Sweet items/mo	0.0034	0.0015	2.6e-02	0.600
Sodium (mmol/L)	-0.0370	0.0170	2.7e-02	0.600
BMI (kg/m^2) V1	-0.0160	0.0078	3.6e-02	0.700
Total bilirubin	0.0092	0.0045	4.0e-02	0.700
Whooping cough vaccine	0.1200	0.0590	4.8e-02	0.700

Figure 6. **Bacterial profiles are stable over the short term; however, the degree of stability is variable across donors. (A)** Boxplots of species BC distances (left), pathway BC distances (middle), and pathway binary Jaccard distances (right) between donors (n = 946, for each individual, the average distance from all other individuals at V1 was used) and within a donor over time (n = 413, time between samples = 17 ± 3.3 d); 1 = samples are completely different, 0 = samples are identical. ****, $P < 2.2 \times 10^{-16}$ by Wilcoxon rank sum. **(B)** Histogram of 413 donors' species stability (1 – BC). **(C)** MDS plot of BC distance of bacterial species composition. Each dot corresponds to one donor who had a second sample. Color indicates longitudinal species stability of that donor (1 – within-sample BC). **(D)** Scatter plots show the top phyla associated with species stability (FDR < 0.05). **(E)** Scatter plots show the top phyla associated with pathway stability (FDR < 0.05). **(D and E)** Each point corresponds to one donor with a longitudinal sample (n = 413). Trend lines show 95% confidence intervals and were modeled with lm. Statistics based on Spearman correlation. **(F)** Results of the series of GLM fits aimed at identifying factors associated with intraindividual species stability. Only factors with P value < 0.05 are shown. Std., standard. **(G)** Scatter plots show the stability of individual species (1 – BC) by their mean baseline relative abundance and prevalence. Each point corresponds to a bacterial species and is colored by the species Phyla. 533 species with mean prevalence >5% are shown. **(H)** Scatter plots show the stability of individual pathways (1 – BC) by their mean baseline relative abundance and prevalence. Each point corresponds to a pathway. 364 pathways with mean prevalence >5% are shown. **(I)** Boxplots compare stability of individual species and pathways (1 – BC). Points are colored by the mean prevalence of the species or pathway in both samples. ****, $P < 2.2 \times 10^{-16}$ by Wilcoxon rank sum. See also Data S1, tables 15–18.

we found that stability was strongly associated with mean abundance and prevalence across donors (Fig. 6, G and H). For example, the Firmicutes *Enterococcus_B faecium_B* had a mean abundance of 0.15%, prevalence 7.5%, and low stability (0.025). Additionally, many species known to be present in yogurt and probiotics were also highly unstable, e.g., *Lactobacillus_D sakei*, *Lactococcus lactis*, and *B. animalis* (Fig. 6 G; Fijan, 2014), in agreement with previous observations that probiotics often face colonization resistance (Zmora et al., 2018). Comparison of overall stability of pathways and species revealed that individual pathways were on average more stable than individual species (Fig. 6 I), consistent with pathways being more conserved across individuals (Fig. 6 A). Moving forward, these data can be leveraged to prioritize microbial pathways/species that will make reliable biomarkers as well as persistent colonizers if incorporated into a microbial therapeutic.

Across cohorts, patients with non-GI cancers have altered gut microbial communities

Given our success integrating results of the MI cohort with those from Zeevi et al. (2015), we sought to determine whether similar congruence was observed in the gut microbiome of cancer patients. We focused on those with non-GI tumors, for whom recent publications have demonstrated associations between microbiome composition and positive responses to checkpoint inhibitors (Frankel et al., 2017; Gopalakrishnan et al., 2018; Matson et al., 2018; Peters et al., 2019; Routy et al., 2018). In contrast to colorectal cancer (Thomas et al., 2019), there remains a large gap in our knowledge detailing how the microbiome composition of cancer patients with non-GI indications compares to that of healthy donors. To investigate this, we applied our Kraken2/GTDB pipeline to an additional 375 samples from 283 cancer patients across five published cohorts (Frankel et al., 2017; Gopalakrishnan et al., 2018; Matson et al., 2018; Peters et al., 2019; Routy et al., 2018; Fig. S4, A–E). Despite technical and geographic differences, we identified that cancer patients have significantly altered gut bacterial communities compared with their healthy counterparts, as quantified by differences in principal coordinates (PCs) 1 and 2 (Fig. 7, A and B). When comparing datasets across cohorts, even those processed with identical analytical methods, there is always a risk of differences being driven by technical artifacts, for example collection or sequencing method. However, in this case, the differences between healthy donors and those with cancer were consistent across age groups (Fig. 7 C) and diverse cohorts (Fig. S4, F and G), supporting our conclusion.

Specifically, when compared with healthy age-matched controls, cancer patients had increased Bacteroidota/Firmicutes_A ratios (driver of PC2; Fig. 7, D–F) and decreased Shannon diversity (driver of PC1; Fig. 7, G and H). Across measures (Fig. S4, F–I), patients in the Frankel et al. (2017) cohort were closer to healthy donors than patients in the other cancer cohorts. Given that the indication, geographic location, and age distribution were similar to Gopalakrishnan et al. (2018), additional information is needed to understand why this may be the case. To determine if these microbial shifts were consistent in individuals with a chronic disease other than cancer, we analyzed an additional 520 samples from 257 patients with IBD, either ulcerative colitis or Crohn's disease, and 80 controls from two published studies (Fig. S4 J; Franzosa et al., 2019; Schirmer et al., 2018). While PC1 values were increased and Shannon values were consistently reduced in the donors with IBD, PC2 varied more by study than health status, with the integrative Human Microbiome Project (iHMP) samples consistently having higher PC2 and Bacteroidota/Firmicutes_A ratios than donors from the Lifelines and Prism cohorts (Fig. S4, K and L). This indicates that while high Bacteroidota/Firmicutes_A may be a conserved feature of the gut microbiome in patients with cancer, this trend is not found in other chronic diseases, as suggested by the results from two IBD cohorts.

To determine if receiving checkpoint inhibitors dramatically alters the gut microbiome, we compared longitudinal stability of the 60 patients with an on-treatment sample to that of the MI donors (Fig. S4 A). Counter to their low Shannon diversity values, we found that cancer patients on checkpoint inhibitor treatment were on average significantly more stable than the healthy MI donors (Fig. 7 I), consistent with our observation that Bacteroidota levels are associated with greater community stability (Fig. 7 J). Although these patients have relatively stable microbial communities on treatment, checkpoint inhibitors are rapidly being tested in combination with other agents (Tang et al., 2018), including chemotherapeutics, which themselves have been shown to alter microbial communities (Montassier et al., 2015). Therefore, additional studies are needed to understand if and how these emerging therapeutic combinations alter the gut microbiome.

In the absence of extensive metadata for the patients with cancer, we leveraged the detailed characterization of the MI cohort to understand within healthy donors which factors contribute to a more "cancer-like" microbiome, as characterized by greater PC1 and PC2 values. Unlike PC1, primarily driven by Shannon diversity (Fig. 7 G), PC2 was associated with multiple factors (Data S1, table 19). Notably, PC2 was consistently elevated in females across the two healthy cohorts (Fig. S5 A). By contrast, these differences were diminished in the cancer cohorts (Fig. 8, A and B), where the average PC2 value was higher despite an enrichment of male patients (Fig. S4, B and G). From this, we can hypothesize that factors driving gut microbiome differences between males and females may be diminished in cancer patients.

After controlling for sex and age, we found that factors generally associated with good health, such as fruit consumption and mean corpuscular volume, were elevated with PC2, while factors associated with poor health, including BMI and circulating levels of the liver proteins alanine aminotransferase, and C-reactive protein, were lower in donors with a more cancer-like, PC2-high microbiome (Fig. 8, C and D; and Data S1, table 19). Many of these correlations were also true when looking at the Bacteroidota/Firmicutes_A ratio (Fig. S5 B and Data S1, table 19), which has been previously linked to lower BMI (Castaner et al., 2018).

Given the inherent batch effects between cohorts, when exploring differentially present species we prioritized those that were consistently elevated in cancer patients versus healthy donors across cohorts (Fig. 8 E and Data S1, tables 20 and 21).

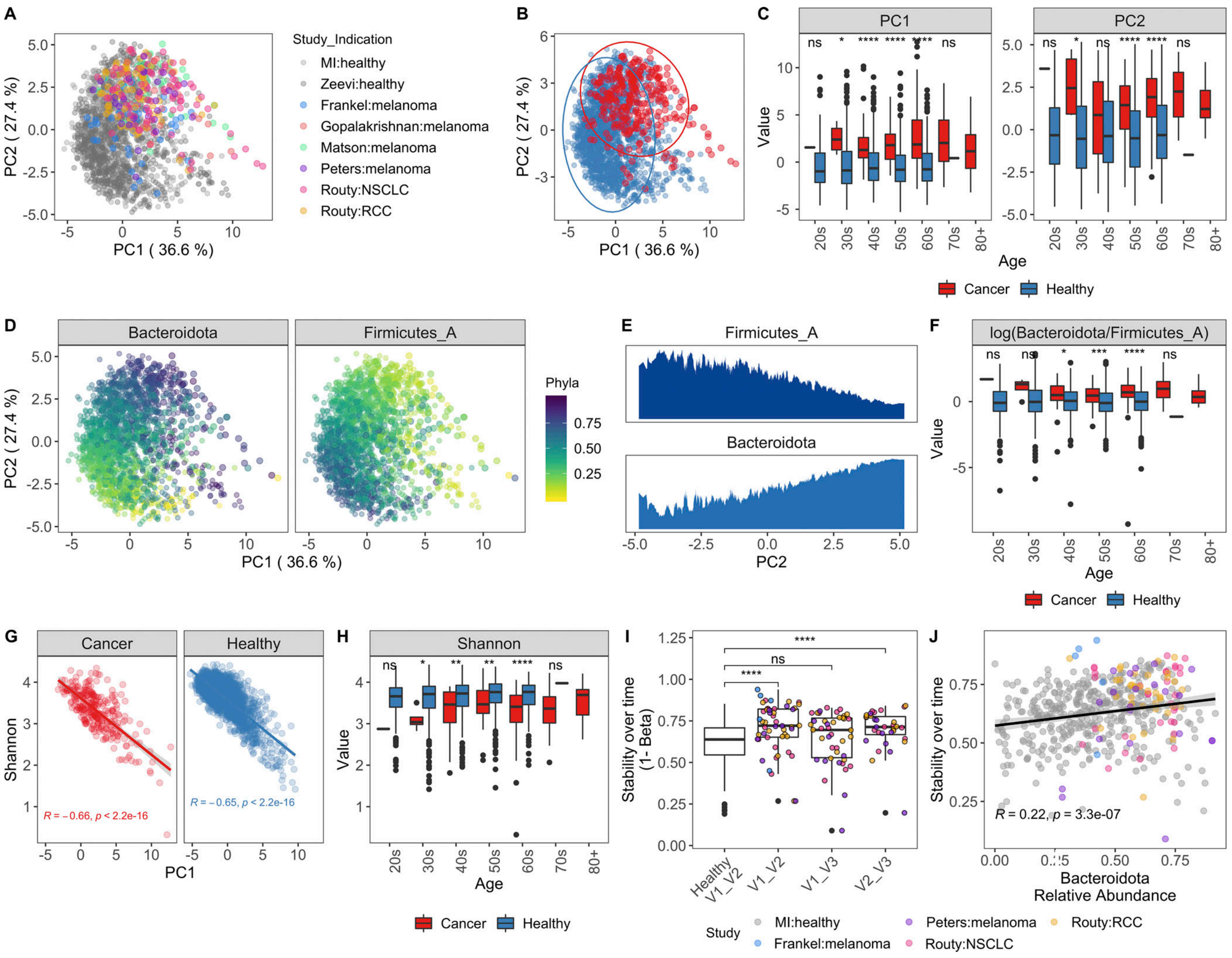


Figure 7. **Cancer patients have altered gut bacterial profiles associated with higher Bacteroidota/Firmicutes_A ratios and lower Shannon diversity. (A)** PC plot of BC distance of bacterial species composition of one sample per donor in the table in Fig. S4 A. When available, the baseline sample was prioritized. Each dot corresponds to one donor, while color indicates the study and indication. **(B)** Same PC analysis plot as A, with dots colored by health status of the donor. Red, cancer; blue, healthy. **(C)** Boxplots compare PC1 and PC2 values of samples from cancer patients and healthy donors stratified by age. **(D)** Same PC analysis plot as A with dots colored by relative phyla abundance. **(E)** Density plots of Bacteroidota and Firmicutes_A were generated using a moving average of the abundance of the phyla within the communities along PC2, with a scale from 0 to the maximum moving average. **(F)** Boxplots compare the log(Bacteroidota/Firmicutes_A) ratio of samples from cancer patients and healthy donors stratified by age. **(G)** Scatter plots show PC1 versus Shannon for cancer patients and healthy donors. Trend lines show 95% confidence intervals and were modeled with lm. Statistics based on Spearman correlation. **(H)** Boxplots compare bacterial Shannon diversity of samples from cancer patients and healthy donors stratified by age. **(I)** Boxplots of species stability (1 – BC distance) within a donor over time points indicated in the x axis. For cancer patients, V1 = baseline, pretreatment; V2-Frankel = within 1 mo of starting ICT; V2-Peters = week 6; V2-Routy = after second injection, ∼1 mo; V2-Peters = week 12; V3-Routy = after fourth injection, ∼2 mo. **(J)** Scatter plots shows Bacteroidota abundance versus stability (1 – BC distance) for cancer and healthy patients. Trend lines show 95% confidence intervals and were modeled with lm. Statistics based on Spearman correlation. For C, F, H, and I, ns, not significant (P > 0.05); *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001 by Wilcoxon rank sum with FDR correction. See also Fig. S4. ICT, immune checkpoint inhibitor; NSCLC, non–small cell lung cancer; RCC, renal cell carcinoma.

Consistent with the greater overall abundance of Bacteroidota, many species including the well-studied *Bacteroides fragilis* and *Bacteroides thetaiotaomicron* were more abundant in cancer patients (Fig. 8 F). In addition, several Firmicutes_A species, including many previously identified as potent regulatory T cell inducers (Atarashi et al., 2013; Narushima et al., 2014), particularly Clostridium_M bolteae, were also elevated in cancer patients versus healthy donors (Fig. S5, C and D). Finally, several species in the *Enterobacteriaceae* family, including the pathobionts *Escherichia coli_D*, *Escherichia dysenteriae*, and *Klebsiella pneumoniae*, were more prevalent in cancer patients across cohorts (Fig. 8 G and Data S1, tables 20 and 21). Previously observed as enriched in patients with IBD and colorectal cancer (Duvallet et al., 2017), *Enterobacteriaceae* are typically dominant in the upper GI tract and may become enriched with the faster stool transit time that occurs during diarrhea (Donaldson et al., 2016); without prior treatment and stool consistency information, we cannot tease apart if that is the cause for these patients.

In total, this analysis revealed that compared with their healthy counterparts, cancer patients' gut microbiomes are less

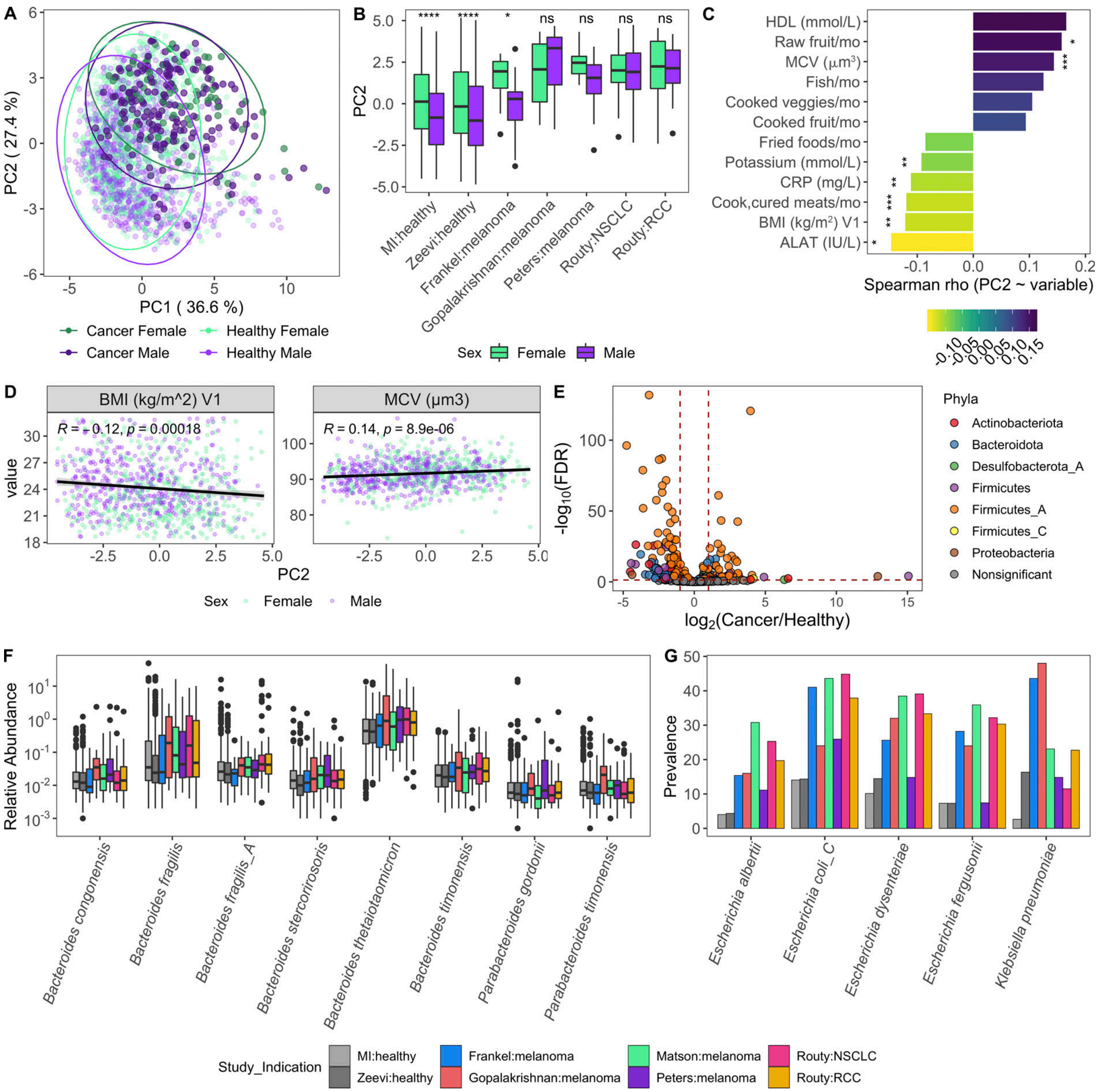


Figure 8. **Variables associated with a more cancer-like microbiome. (A)** Same PC analysis plot as Fig. 7 A. Each dot corresponds to one donor, while color indicates health status of the donor and their sex. **(B)** Boxplots compare PC2 values of samples from males and females stratified by study; ns, not significant ($P > 0.05$); *, $P < 0.05$; ****, $P < 0.0001$ by Wilcoxon rank sum with FDR correction. **(C)** Length and color of the bars indicates the Spearman ρ value of each of the variables by PC2. Only those variables with an FDR < 0.05 are shown. Stars indicate the P value after correcting for sex and age in a linear model. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. **(D)** Scatter plots show PC2 value versus variables from C. Trend lines show 95% confidence intervals and were modeled with lm. Statistics based on Spearman correlation. **(E)** Volcano plot of species differentially abundant between cancer patients and healthy donors based on DESeq2. To ensure that results were not driven by rare species, particularly in the large healthy cohorts, we first removed any species not present in at least three of the cancer cohorts at >10% prevalence, and then kept only the 448 species with a mean relative abundance >0.01%. Each species is colored by its taxonomic phyla. **(F)** Boxplot of Bacteroidota species more abundant in cancer patients than healthy donors. Y axis is on log10 scale. **(G)** Prevalence of selected Enterobacteriaceae species across healthy and cancer cohorts. Color corresponds to study. See also Fig. S5 and Data S1, tables 19–21. NSCLC, non–small cell lung cancer; RCC, renal cell carcinoma.

diverse and populated by more *Enterobacteriaceae*. However, counterintuitively, cancer patients also had higher Bacteroidota/Firmicutes_A ratios, which are associated with features of good health such as lower BMI, alanine aminotransferase, and C-reactive protein, as well as greater stability. While these observations are important for contextualizing findings of the recent cancer immunology/microbiome literature, future sampling efforts are needed to deconvolute which of these changes are driven by the cancer itself versus lifestyle changes or therapeutic agents that are taken after a cancer diagnosis and can themselves account for shifts in the gut microbiome. In addition, further experiments are needed to understand how these shifts may influence a patient's underlying cancer immune set point and subsequent response to therapy (Chen and Mellman, 2017).

Discussion

As the number of microbial intervention trials and biomarker studies continues to grow, it is increasingly important to develop a robust understanding of the gut microbiome across individuals in the steady state. In this study, we use the statistical power of a large cohort, the resolution afforded by deep shotgun sequencing, and an updated microbial database to expand our understanding of the gut microbiome in health and disease. More specifically, we identified sex as the strongest driver of community composition, with many *Prevotella* species enriched in men compared with women (Fig. 3 B); many of which were absent in previous databases (Truong et al., 2015) and thus not detectable in prior analyses. Given the recent literature on the strain-level variability within *Prevotella* species (De Filippis et al., 2019; Fehlner-Peach et al., 2019), particularly *Prevotella copri*, follow-up analyses should compare if there are also strain-level differences between the sexes.

Additionally, we identified 192 species associated with age (Fig. 4 B), greatly expanding what was known about the effects of aging on the gut microbiome (An et al., 2018). The changes seen here are particularly striking because they occur in the absence of underlying diseases or medication usage. Given the cross-sectional nature of this cohort, it is difficult to tease apart which of these associations is mediated by the variables correlated with age (Fig. S3 E), for example increased raw fruit, reduced fast food consumption, and increased BMI, in contrast to physiological changes associated with aging, such as thinning of the mucosal layer or altered pH levels. To definitively understand how the gut microbiome matures with aging will require longitudinal sampling of donors throughout their lifetimes, since in the short term, fluctuations are negligible in most donors (Fig. 6 A). Despite these caveats, the knowledge that bacterial communities are strongly shaped by age and sex encourages additional analysis into whether matching the age and sex of FMT donor and recipient could promote durable engraftment of bacteria.

These sex and age results could also have multiple implications for the interpretation of microbial biomarker studies. For example, several bacterial biomarkers have been reported for response to checkpoint inhibitors in non-GI cancer indications (Frankel et al., 2017; Gopalakrishnan et al., 2018; Matson et al., 2018; Peters et al., 2019; Routy et al., 2018). Notably, however, age—a prognostic biomarker of response in some indications and strong correlate of microbial composition—remains unaccounted for in some of those analyses. While our own attempts to associate PC1 and PC2 with response to therapy were unsuccessful (Fig. S5 E), these results should be interpreted with caution given the different definitions of response across cancer indications and interventional studies. Moving forward, statistically robust signatures of response will require larger cohort sizes and will benefit from analysis of continuous variables such as progression-free survival versus binary groupings (Peters et al., 2019). Additionally, while our observations concerning the differences between cancer patients and healthy donors should be carefully considered when designing FMT trials for this population, further experimental evidence is needed to determine whether the optimal donor is a patient who previously responded to therapy or a healthy donor with a responder-like signature.

In addition to cancer, these data will be valuable for designing microbial therapeutics for individuals of all ages. For instance, interventions containing *Bifidobacterium* species may need to be dosed more frequently in individuals older than 50 yr, in whom *Bifidobacterium* appears to colonize less effectively (Fig. 5, A, B, and D). Similarly, consortia with *Prevotella* species may work less effectively in females, and species which demonstrated low stability in the short term may also require additional dosing. Finally, beyond the findings in this paper, the rich metadata and 1,000-plus deep shotgun metagenomic samples provided here will be a valuable resource on which future microbiome studies can build new computational tools as well as generate and test new hypotheses.

Materials and methods

Experimental model and subject details

The MI cohort

The 1,000 healthy donors of the MI cohort were recruited by BioTrial in the suburban Rennes area (Ille-et-Vilaine, Bretagne, France). The cohort included 500 men and 500 women; 200 individuals were from each decade of life, between 20 and 69 yr of age. Participants were selected based on stringent inclusion and exclusion criteria, detailed elsewhere (Thomas et al., 2015). Donor BMI was restricted to ≥18.5 and ≤32 kg/m^2. Briefly, the donors had no evidence of any severe/chronic/recurrent pathological conditions. Primary exclusion criteria were seropositivity for HIV or hepatitis C virus, travel to tropical or subtropical countries within the previous 6 mo, recent vaccine administration, and alcohol abuse. Subjects were also excluded if they took nasal, intestinal, or respiratory antibiotics or antiseptics any time in the 3 mo preceding enrollment. Additionally, anyone following a doctor- or dietician-prescribed diet for medical reasons (e.g., calorie-controlled diet in overweight patients) and volunteers with food intolerance or allergy were excluded. To avoid the influence of hormonal fluctuations in women during the perimenopausal phase, only pre- or postmenopausal women were included. To minimize the influence of population substructure, the study was restricted to individuals of self-reported metropolitan French origin for three generations (i.e., with parents and grandparents born in continental France).

Demographic, environmental, dietary, and clinical variables

Multiple demographic, environmental, and clinical variables were collected for each of the donors in an electronic case report form (Thomas et al., 2015). For example, donors were asked about their family medical history, smoking habits, sleeping habits, and infection and vaccination history. Additionally, donors completed a food-frequency questionnaire (FFQ) administered by trained investigators and comprising 19 food groups (Data S1, table 5). Participants estimated their "usual consumption" selecting from six intake frequencies ranging from "twice per day or more" to "never" (except for alcohol, which offered five intake frequencies ranging from "every day" to "never").

Investigators administering the FFQ invited participants to declare their "usual" diet, rather than focusing on their latest dietary consumption. The detailed FFQ is available in Partula et al. (2019). For clinical chemistry, hematologic, and serologic assessments, 20 ml of blood was collected from each donor and analyzed at the certified Laboratoire de Biologie Médicale, Centre Eugene Marquis (Rennes, France). For microbiome profiles, stool samples were produced by the participant at home within 24 h before the scheduled visits (visit 1 [V1] and V2). For individuals who provided two stool samples, V1 and V2 were on average 17 ± 3.3 d apart, minimum 8 d and maximum 45 d.

After manual curation and removal of variables that were (a) variable in <5% of participants, (b) missing in >25% of donors, or (c) correlated with another variable (Spearman ρ greater than −0.6 or < 0.6), 154 metadata variables were considered for future correlations. In the case of correlated variables (Data S1, tables 5 and 6), the variable with fewer missing values was prioritized and kept, while the other variable was removed. When the pair had equivalent numbers of missing values, one from the pair was randomly selected. Notably, circulating levels of creatinine were so strongly correlated with sex (Spearman $\rho = 0.72$, $P = 3.5 \times 10^{-115}$), this variable was excluded from the 154.

Ethics statement

The clinical study was approved by the Comité de Protection des Personnes–Ouest 6 on June 13, 2012, and by the French Agence Nationale de Sécurité du Médicament on June 22, 2012, and was performed in accordance with the Declaration of Helsinki. The study was sponsored by the Institut Pasteur (Pasteur ID-RCB no. 2012-A00238-35) and conducted as a single-center study without any investigational product. The original protocol is registered under ClinicalTrials.gov (study number NCT01699893). Informed consent was obtained from the participants after the nature and possible consequences of the studies were explained. The samples and data used in this study were formally established as the Milieu Interieur biocollection (NCT03905993), with approvals by the Comité de Protection des Personnes–Sud Méditerranée and the Commission nationale de l'informatique et des libertés on April 11, 2018.

Details

Fecal DNA extraction and shotgun metagenomic sequencing

Stool specimens were collected in a double-lined sealable bag containing a GENbag Anaer atmosphere generator (Aerocult; Biomerieux) to maintain anaerobic conditions. Upon reception at the clinical site, fresh samples were aliquoted into cryotubes and stored at −80°C.

Stool aliquots were shipped to the CRO Diversigen for DNA extraction and shotgun metagenomic sequencing. At Diversigen, genomic DNA was extracted using PowerMag Soil DNA Isolation Kit (27100; Qiagen, MO BIO Laboratories). Libraries were prepared using Beckman robotic workstations (Biomek FX and FXp models) in batches of 96 samples. DNA (10–500 ng) was sheared into fragments of ∼300–400 bp in a Covaris E210 system (96-well format; Covaris) followed by purification of the fragmented DNA using AMPure XP beads. DNA end repair, 3′-adenylation, ligation to Illumina multiplexing PE adaptors, and ligation-mediated PCR were all completed using automated processes. To amplify high GC-rich and low AT-rich regions at greater efficiency, KAPA HiFi polymerase (KAPA Biosystems) was used for PCR amplification (6–10 cycles). Fragment Analyzer (Advanced Analytical Technologies) electrophoresis system was used for library quantification and size estimation. Prepared libraries were then pooled and sequenced on an Illumina HiSeq 2500.

In the end, we obtained 21 trillion raw paired-end reads from 1,359 samples from 946 of the donors. On average per sample, there were 2.4 Gbp, 15.5 million reads, with 358-bp insert size. To process the reads, Illumina TruSeq adapters were trimmed with Trimmomatic v0.36 (Bolger et al., 2014); low-quality and low-complexity reads were removed with prinseq-lite 0.20.4 (Schmieder and Edwards, 2011); and Bowtie2 v2.1.0 (Langmead and Salzberg, 2012) was used to remove reads mapping to PhiX or the PacBio human genome (parameters specified in Fig. S1 C). After processing, there were on average 13.9 ± 2.9 million reads per sample (Fig. S1, D and E). Of an initial 1,000 recruited donors, 44 were excluded from this analysis because of lack of consent for sharing their data outside of the MI consortium. An additional 10 donors were excluded because of technical issues in the extraction and sequencing steps (e.g., low DNA extraction yield), resulting in a sample size of 946 donors.

Quantification and statistical analysis

Building the Kraken-GTDB database

To build the Kraken-GTDB database, first the following files were downloaded ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt and https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/bac120_taxonomy_r89.tsv (Parks et al., 2018) on June 25, 2019. These files were merged based on accession number, and only those genomes present in both databases were considered, i.e., RefSeq genomes with a GTDB taxonomy. To avoid biasing the database toward those species with large numbers of genomes (Fig. S1 A), while balancing the added information provided by additional isolates per species, we selected up to five genomes per GTDB species to include in our database. Genomes were first ordered by their assembly quality, i.e., reference genome, representative genome, complete genome, chromosome, contig, and scaffold, and then randomly selected. Based on these criteria, 23,505 genomes representing 13,446 unique bacterial species were downloaded and formatted into a Kraken2 database (Wood and Salzberg, 2014). To incorporate the GTDB taxonomy into the Kraken2 database, files mimicking the NCBI-like taxonomy files from ftp://ftp.ncbi.nih.gov/pub/taxonomy/new_taxdump/new_taxdump.zip were created for names.dmp, complete_names.dmp, nodes.dmp, and accession2taxid. A matching Bracken database was then generated with bracken-build -k 35 -l 126 (Lu et al., 2017).

Metagenomic data analysis

First, putative reagent contaminants identified by species co-correlation analysis were filtered using Kraken2's *unclassified-out* option and a custom database of contaminant genomes (Data S1, table 22). Using our custom GTDB-Kraken database, Kraken2 v2.0.8 (Wood and Salzberg, 2014) and Bracken v2.5 (Lu et al.,

2017) were run on the 1,359 quality-controlled samples (parameters specified in Fig. S1 F) to generate bacterial profiles. With the exception of longitudinal results in Fig. 6, all analyses were based on bacterial profiles from the V1 samples. In the case of eight donors from whom no V1 sample was available, the sample from V2 was used.

To complement results from the MI donors, this pipeline was run on an additional 1,159 shotgun metagenomics samples from 851 donors downloaded from ENA: PRJEB11532 (Zeevi et al., 2015), as well as 44 samples from 39 donors from SRP115355 (Frankel et al., 2017), 39 samples from 39 donors from SRP116709 (Matson et al., 2018), 25 samples from 25 donors from PRJEB22893 (Gopalakrishnan et al., 2018), 219 samples from 153 donors from PRJEB22863 (Routy et al., 2018), and 48 samples from 27 donors from PRJNA541981 (Fig. S4 A; Peters et al., 2019). Samples from PRJNA541981 had a large variability in sampling depths, 29 ± 31 million reads; therefore, 11 samples with >40 million reads were subsampled to 40 million reads. For the IBD comparison, an additional 220 samples from 220 donors were downloaded from the Lifelines and PRISM cohorts, PRJNA400072 (Franzosa et al., 2019), as well as 300 samples from 117 donors from iHMP, PRJNA389280 (Fig. S4 J; Schirmer et al., 2018). Donors in the iHMP <20 yr old were grouped in the pediatric cohort. For the Zeevi samples, metadata (sex, age, and BMI) was obtained by emailing the authors; no time point was provided, so an average of the microbial profiles across samples of the same donor were used in further analyses. For the oncology cohorts, one sample per donor was used for all but the longitudinal analysis. When available, the baseline sample was prioritized; otherwise an on-treatment sample was used. For the iHMP samples, an average of the microbial profiles across samples of the same donor were used.

To go beyond taxa-based calls, HUMANn2 v0.11.2 (Franzosa et al., 2018) with default parameters including Uniref90 was run on all MI samples where the forward and reverse reads were concatenated into a single file. Raw output values were converted from *rpks* to *cpms* with *humann2_renorm_table*. Outputs of all samples were joined into a single merged table with the *humann2_join_tables* function. Using *humann2_regroup_table*, individual gene families were regrouped with multiple different databases including COG, GO, KEGG, and MetaCyc. Ultimately, MetaCyc pathways (Caspi et al., 2018) were selected for the correlations because of the additional steps implemented in HUMANn2 to check for completeness of the pathways.

Statistical analysis

All correlations and statistical tests were performed in R v3.6.1 (R Core Team, 2019), documented via rmarkdown documents (Allaire et al., 2019), and compiled with knitr (Xie, 2019). Within R, tables were manipulated with functions of the dplyr package (Wickham et al., 2019). The majority of figures were rendered with ggplot2 (Wickham, 2016), adjusted with geasy (Carroll et al., 2020), and arranged with cowplot (Wilke, 2019). Colors were selected with the help of RColorBrewer (Neuwirth, 2014) and viridis (Garnier, 2018). The cladogram in Fig. 4 B was generated with GraPhlAn (Asnicar et al., 2015). Correlation plots in Fig. S3, D and E and Fig. 5 C were generated with ggcorrplot (Kassambara, 2019). Supplemental tables were generated with Openxlsx (Walker, 2019). When comparing values between two or more groups, Wilcoxon rank sum tests were used.

Bacterial α- and β-diversity measures including Shannon and BC were calculated using the R package vegan (Oksanen et al., 2019). To identify which of the 154 metadata variables were significantly associated with BC β-diversity, we used the adonis function in vegan to run PERMANOVA tests with 999 permutations (Oksanen et al., 2019). When repeated with 1,500, 2,000, 5,000, and 10,000 permutations, the same 52 variables were identified as significant (FDR ≤ 0.06). For all species-based analyses, we tested only those with prevalence >5% and a mean relative abundance >0.01% in the respective cohort. Similarly, for pathways, we prioritized those with prevalence >5%. To identify bacterial species differentially abundant between males/females and cancer patients/healthy donors, we input raw counts into DESeq2 (Love et al., 2014). Differences between sexes were adjusted for age and BMI, while differences between healthy and cancer were adjusted for age and sex. Because the latter led to the exclusion of samples from Matson et al. (2018), for which age and sex were not available, the comparison was also run without covariates (Data S1, table 21). Default values for the results function of DESeq were applied, including using Benjamini–Hochberg FDR for adjusting P values. To account for any lack of independence in the hypotheses being tested, multiple hypothesis correction using the Benjamini–Yekutieli adjustment was also performed and has been included in Data S1, tables 9, 10, and 21. Pathways differentially abundant between males and females were identified with Wilcoxon tests.

To identify bacterial species and pathways differentially prevalent between males and females, we used prop.test in R. Bacterial taxa and pathways associated with age were identified using Spearman correlations. To complement these analyses, GLM was also used. After performing arcsine square-root transformation of the relative abundances, models were fitted using the glm function in R, with sex and BMI included as covariates (model: species ~ age + sex + BMI; Data S1, table 12). For all correlations, species abundances were normalized with total sum scaling, and P values were adjusted with Benjamini–Hochberg. For conclusions based on species richness (e.g., Fig. 3 C and Fig. 5 B), we used linear models with sequencing depth as a covariate to validate that results were not an artifact of unequal library sizes.

To determine the stability of a donor's bacterial species and pathways between V1 and V2, the BC distance was calculated and subtracted from 1. Individuals with a stability of 1 had samples that were identical across time points, while a stability of 0 meant the samples were nothing alike. Similarly, 1 – the Jaccard index was used to determine pathway stability based on presence/absence. To calculate between donor dissimilarities as shown in Fig. 6 A, for each individual we averaged the distance, as measured by BC, between that individual's sample and all other individuals' samples at V1. To identify the phyla associated with stability, Spearman correlation coefficients were calculated between the relative abundance of the phyla at baseline and stability as measured by 1 – BC. We then used GLM fit with the R package betareg v3.1-3 (Cribari-Neto and Zeileis, 2012) with a β response to identify which metadata factors were associated with intraindividual stability. To calculate stability of individual features (species and pathways), we again applied the BC metric, but this time compared the relative abundance of a single feature across all donors at V1 and V2.

To determine which continuous variables were associated with PC1, PC2, and Bacteroidota/Firmicutes_A ratios within the MI donors, we used Spearman correlations. When then applied linear models with age and sex as covariates to determine which of the associations was still significant after correcting for age and sex.

For all statistical tests, P values were corrected with the R function *p.adjust* using the Benjamini–Hochberg (FDR) method (Benjamini and Hochberg, 1995). To account for possible dependence between the hypotheses being tested across all species, multiple hypothesis correction using the Benjamini–Yekutieli adjustment was also performed, and those values have been included in Data S1, tables 9, 10, 12, 13, and 21 (Benjamini and Yekutieli, 2001).

Data and software availability

Data availability

Sequence data have been deposited in the European Genome-Phenome Archive under accession code EGAS00001004437. Donor metadata and code used in this paper will also be available.

Online supplemental material

Fig. S1 shows the quality control and Kraken analysis pipelines. Fig. S2 shows results from all the different program–database combinations tested. Fig. S3 shows distributions of the 154 metadata variables across the MI donors. Fig. S4 shows how cancer patients have altered gut bacterial profiles consistently across studies and indications. Fig. S5 shows additional analysis of which variables were associated with a cancer-like microbiome. 22 tables are provided in Data S1 that present summary statistics and other information.

Acknowledgments

We thank all the donors for their participation in the study. We thank J. Segre and F. Tamburini for critical reading of the manuscript and helpful comments. We also thank J. Paulson and O. Mayba for expert statistical support.

This work benefited from support of the French government's Invest in the Future program. This program is managed by the Agence Nationale de la Recherche, reference ANR-10-LABX-69-01.

Author contributions: Conceptualization, A.L. Byrd, O.J. Harrison, D. Duffy, M.L. Albert, L. Quintana-Murci; Methodology, A.L. Byrd, M. Liu, J. Bergstedt; Software, A.L. Byrd, M. Liu; Formal analysis, A.L. Byrd, S. Lyalina, M. Liu, K.E. Fujimura; Investigation, B. Charbit; Resources, D. Duffy, M.L. Albert, L. Quintana-Murci; Data curation, A.L. Byrd, J. Bergstedt; Writing (original draft), A.L. Byrd, O.J. Harrison; Writing (review and editing), D. Duffy, S. Lyalina, E. Patin, D.R. Nagarkar, M. Liu, K.E. Fujimura, J. Bergstedt, I. Mellman, M.L. Albert; Visualizations, A.L. Byrd, M. Liu; Supervision, A.L. Byrd, D.R. Nagarkar, D. Duffy, I. Mellman, M.L. Albert; Project administration, D.R. Nagarkar, D. Duffy; Funding acquisition, D. Duffy, M.L. Albert, L. Quintana-Murci.

Disclosures: A.L. Byrd, K.E. Fujimura, S. Lyalina, D.R. Nagarkar, and I. Mellman reported being employees of Genentech. M. Liu reported being a former employee of Genentech. The current work does not pertain to any marketed Genentech product. D. Duffy reported grants from Genentech during the conduct of the study. M.L. Albert reported being an employee of Insitro. No other disclosures were reported.

Submitted: 1 April 2020
Revised: 20 July 2020
Accepted: 9 September 2020

References

Allaire, J., Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, W. Chang, and R. Iannone. 2019. rmarkdown: Dynamic Documents for R. https://rmarkdown.rstudio.com/docs/index.html (accessed August 18, 2020).

Allegretti, J.R., B.H. Mullish, C. Kelly, and M. Fischer. 2019. The evolution of the use of faecal microbiota transplantation and emerging therapeutic indications. *Lancet.* 394:420–431. https://doi.org/10.1016/S0140-6736(19)31266-8

An, R., E. Wilms, A.A.M. Masclee, H. Smidt, E.G. Zoetendal, and D. Jonkers. 2018. Age-dependent changes in GI physiology and microbiota: time to reconsider? *Gut.* 67:2213–2222. https://doi.org/10.1136/gutjnl-2017-315542

Ananthakrishnan, A.N., C. Luo, V. Yajnik, H. Khalili, J.J. Garber, B.W. Stevens, T. Cleland, and R.J. Xavier. 2017. Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases. *Cell Host Microbe.* 21:603–610.e3. https://doi.org/10.1016/j.chom.2017.04.010

Asnicar, F., G. Weingart, T.L. Tickle, C. Huttenhower, and N. Segata. 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ.* 3:e1029. https://doi.org/10.7717/peerj.1029

Atarashi, K., T. Tanoue, K. Oshima, W. Suda, Y. Nagano, H. Nishikawa, S. Fukuda, T. Saito, S. Narushima, K. Hase, et al. 2013. Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature.* 500:232–236. https://doi.org/10.1038/nature12331

Bäckhed, F., J. Roswall, Y. Peng, Q. Feng, H. Jia, P. Kovatcheva-Datchary, Y. Li, Y. Xia, H. Xie, H. Zhong, et al. 2015. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe.* 17:690–703. https://doi.org/10.1016/j.chom.2015.04.004

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.* 57: 289–300.

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29:1165–1188.

Biagi, E., C. Franceschi, S. Rampelli, M. Severgnini, R. Ostan, S. Turroni, C. Consolandi, S. Quercia, M. Scurti, D. Monti, et al. 2016. Gut Microbiota and Extreme Longevity. *Curr. Biol.* 26:1480–1485. https://doi.org/10.1016/j.cub.2016.04.016

Biagi, E., L. Nylund, M. Candela, R. Ostan, L. Bucci, E. Pini, J. Nikkïla, D. Monti, R. Satokari, C. Franceschi, et al. 2010. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One.* 5:e10667. https://doi.org/10.1371/journal.pone.0010667

Bolger, A.M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Browne, H.P., S.C. Forster, B.O. Anonye, N. Kumar, B.A. Neville, M.D. Stares, D. Goulding, and T.D. Lawley. 2016. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature.* 533: 543–546. https://doi.org/10.1038/nature17645

Carroll, J., A. Schep, and J. Sidi. 2020. geasy: Easy Access to 'ggplot2' Commands. https://rdrr.io/cran/ggeasy/ (accessed August 18, 2020).

Caspi, R., R. Billington, C.A. Fulcher, I.M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P.E. Midford, Q. Ong, W.K. Ong, et al. 2018. The

MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 46(D1):D633–D639. https://doi.org/10.1093/nar/gkx935

Castaner, O., A. Goday, Y.M. Park, S.H. Lee, F. Magkos, S.T.E. Shiow, and H. Schröder. 2018. The Gut Microbiome Profile in Obesity: A Systematic Review. *Int. J. Endocrinol.* 2018:4095789.

Chen, D.S., and I. Mellman. 2017. Elements of cancer immunity and the cancer-immune set point. *Nature.* 541:321–330. https://doi.org/10.1038/nature21349

Costello, E.K., C.L. Lauber, M. Hamady, N. Fierer, J.I. Gordon, and R. Knight. 2009. Bacterial community variation in human body habitats across space and time. *Science.* 326:1694–1697. https://doi.org/10.1126/science.1177486

Cribari-Neto, F., and A. Zeileis. 2012. Beta Regression in R. *J. Stat. Softw.* 34: 1–24.

De Filippis, F., E. Pasolli, A. Tett, S. Tarallo, A. Naccarati, M. De Angelis, E. Neviani, L. Cocolin, M. Gobbetti, N. Segata, and D. Ercolini. 2019. Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host Microbe.* 25:444–453.e3. https://doi.org/10.1016/j.chom.2019.01.004

Ding, T., and P.D. Schloss. 2014. Dynamics and associations of microbial community types across the human body. *Nature.* 509:357–360. https://doi.org/10.1038/nature13178

Donaldson, G.P., S.M. Lee, and S.K. Mazmanian. 2016. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* 14:20–32. https://doi.org/10.1038/nrmicro3552

Dubin, K., M.K. Callahan, B. Ren, R. Khanin, A. Viale, L. Ling, D. No, A. Gobourne, E. Littmann, C. Huttenhower, et al. 2016. Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. *Nat. Commun.* 7:10391. https://doi.org/10.1038/ncomms10391

Duvallet, C., S.M. Gibbons, T. Gurry, R.A. Irizarry, and E.J. Alm. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8:1784. https://doi.org/10.1038/s41467-017-01973-8

Faith, J.J., J.L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A.L. Goodman, J.C. Clemente, R. Knight, A.C. Heath, R.L. Leibel, et al. 2013. The long-term stability of the human gut microbiota. *Science.* 341: 1237439. https://doi.org/10.1126/science.1237439

Falony, G., M. Joossens, S. Vieira-Silva, J. Wang, Y. Darzi, K. Faust, A. Kurilshikov, M.J. Bonder, M. Valles-Colomer, D. Vandeputte, et al. 2016. Population-level analysis of gut microbiome variation. *Science.* 352: 560–564. https://doi.org/10.1126/science.aad3503

Fehlner-Peach, H., C. Magnabosco, V. Raghavan, J.U. Scher, A. Tett, L.M. Cox, C. Gottsegen, A. Watters, J.D. Wiltshire-Gordon, N. Segata, et al. 2019. Distinct Polysaccharide Utilization Profiles of Human Intestinal Prevotella copri Isolates. *Cell Host Microbe.* 26:680–690.e5. https://doi.org/10.1016/j.chom.2019.10.013

Fijan, S. 2014. Microorganisms with claimed probiotic properties: an overview of recent literature. *Int. J. Environ. Res. Public Health.* 11:4745–4767. https://doi.org/10.3390/ijerph110504745

Flores, G.E., J.G. Caporaso, J.B. Henley, J.R. Rideout, D. Domogala, J. Chase, J.W. Leff, Y. Vázquez-Baeza, A. Gonzalez, R. Knight, et al. 2014. Temporal variability is a personalized feature of the human microbiome. *Genome Biol.* 15:531. https://doi.org/10.1186/s13059-014-0531-y

Forster, S.C., N. Kumar, B.O. Anonye, A. Almeida, E. Viciani, M.D. Stares, M. Dunn, T.T. Mkandawire, A. Zhu, Y. Shao, et al. 2019. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* 37:186–192. https://doi.org/10.1038/s41587-018-0009-7

Frankel, A.E., L.A. Coughlin, J. Kim, T.W. Froehlich, Y. Xie, E.P. Frenkel, and A.Y. Koh. 2017. Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients. *Neoplasia.* 19:848–855. https://doi.org/10.1016/j.neo.2017.08.004

Franzosa, E.A., K. Huang, J.F. Meadow, D. Gevers, K.P. Lemon, B.J. Bohannan, and C. Huttenhower. 2015. Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci. USA.* 112:E2930–E2938. https://doi.org/10.1073/pnas.1423854112

Franzosa, E.A., L.J. McIver, G. Rahnavard, L.R. Thompson, M. Schirmer, G. Weingart, K.S. Lipson, R. Knight, J.G. Caporaso, N. Segata, and C. Huttenhower. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods.* 15:962–968. https://doi.org/10.1038/s41592-018-0176-y

Franzosa, E.A., A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H.J. Haiser, S. Reinker, T. Vatanen, A.B. Hall, H. Mallick, L.J. McIver, et al. 2019. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* 4:293–305. https://doi.org/10.1038/s41564-018-0306-4

Gao, C., B.P. Ganesh, Z. Shi, R.R. Shah, R. Fultz, A. Major, S. Venable, M. Lugo, K. Hoch, X. Chen, et al. 2017. Gut Microbe-Mediated Suppression of Inflammation-Associated Colon Carcinogenesis by Luminal Histamine Production. *Am. J. Pathol.* 187:2323–2336. https://doi.org/10.1016/j.ajpath.2017.06.011

Garnier, S. 2018. viridis: Default Color Maps from ‘matplotlib.’ https://rdrr.io/cran/viridis/ (accessed August 18, 2020).

Gingold-Belfer, R., S. Levy, O. Layfer, L. Pakanaev, Y. Niv, R. Dickman, and T.T. Perets. 2020. Use of a Novel Probiotic Formulation to Alleviate Lactose Intolerance Symptoms-a Pilot Study. *Probiotics Antimicrob. Proteins.* 12:112–118. https://doi.org/10.1007/s12602-018-9507-7

Goodrich, J.K., E.R. Davenport, A.G. Clark, and R.E. Ley. 2017. The Relationship Between the Human Genome and Microbiome Comes into View. *Annu. Rev. Genet.* 51:413–433. https://doi.org/10.1146/annurev-genet-110711-155532

Gopalakrishnan, V., C.N. Spencer, L. Nezi, A. Reuben, M.C. Andrews, T.V. Karpinets, P.A. Prieto, D. Vicente, K. Hoffman, S.C. Wei, et al. 2018. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science.* 359:97–103. https://doi.org/10.1126/science.aan4236

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature.* 486:207–214. https://doi.org/10.1038/nature11234

Jackson, M.A., S. Verdi, M.E. Maxan, C.M. Shin, J. Zierer, R.C.E. Bowyer, T. Martin, F.M.K. Williams, C. Menni, J.T. Bell, et al. 2018. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat. Commun.* 9:2655. https://doi.org/10.1038/s41467-018-05184-7

Johnson, A.J., P. Vangay, G.A. Al-Ghalith, B.M. Hillmann, T.L. Ward, R.R. Shields-Cutler, A.D. Kim, A.K. Shmagel, A.N. Syed, J. Walter, et al. Personalized Microbiome Class Students. 2019. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe.* 25:789–802.e5. https://doi.org/10.1016/j.chom.2019.05.005

Kang, D.W., J.B. Adams, D.M. Coleman, E.L. Pollard, J. Maldonado, S. McDonough-Means, J.G. Caporaso, and R. Krajmalnik-Brown. 2019. Long-term benefit of Microbiota Transfer Therapy on autism symptoms and gut microbiota. *Sci. Rep.* 9:5821. https://doi.org/10.1038/s41598-019-42183-0

Kassambara, A. 2019. ggcorrplot: Visualization of a Correlation Matrix using ‘ggplot2’. http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2 (accessed August 18, 2020).

Kato, K., T. Odamaki, E. Mitsuyama, H. Sugahara, J.Z. Xiao, and R. Osawa. 2017. Age-Related Changes in the Composition of Gut Bifidobacterium Species. *Curr. Microbiol.* 74:987–995. https://doi.org/10.1007/s00284-017-1272-4

Kearney, S.M., S.M. Gibbons, M. Poyet, T. Gurry, K. Bullock, J.R. Allegretti, C.B. Clish, and E.J. Alm. 2018. Endospores and other lysis-resistant bacteria comprise a widely shared core community within the human microbiota. *ISME J.* 12:2403–2416. https://doi.org/10.1038/s41396-018-0192-z

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 9:357–359. https://doi.org/10.1038/nmeth.1923

Love, M.I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15: 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, J., F.P. Breitwieser, P. Thielen, and S.L. Salzberg. 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3: e104. https://doi.org/10.7717/peerj-cs.104

Maier, L., M. Pruteanu, M. Kuhn, G. Zeller, A. Telzerow, E.E. Anderson, A.R. Brochado, K.C. Fernandez, H. Dose, H. Mori, et al. 2018. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature.* 555: 623–628. https://doi.org/10.1038/nature25979

Matson, V., J. Fessler, R. Bao, T. Chongsuwat, Y. Zha, M.L. Alegre, J.J. Luke, and T.F. Gajewski. 2018. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science.* 359:104–108. https://doi.org/10.1126/science.aao3290

McCann, K.S. 2000. The diversity-stability debate. *Nature.* 405:228–233. https://doi.org/10.1038/35012234

McIntyre, A.B.R., R. Ounit, E. Afshinnekoo, R.J. Prill, E. Hénaff, N. Alexander, S.S. Minot, D. Danko, J. Foox, S. Ahsanuddin, et al. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* 18:182. https://doi.org/10.1186/s13059-017-1299-7

Mehta, R.S., G.S. Abu-Ali, D.A. Drew, J. Lloyd-Price, A. Subramanian, P. Lochhead, A.D. Joshi, K.L. Ivey, H. Khalili, G.T. Brown, et al. 2018. Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3:347–355. https://doi.org/10.1038/s41564-017-0096-0

Montassier, E., T. Gastinne, P. Vangay, G.A. Al-Ghalith, S. Bruley des Varannes, S. Massart, P. Moreau, G. Potel, M.F. de La Cochetière, E. Batard, and D. Knights. 2015. Chemotherapy-driven dysbiosis in the intestinal microbiome. *Aliment. Pharmacol. Ther.* 42:515–528. https://doi.org/10.1111/apt.13302

Mueller, S., K. Saunier, C. Hanisch, E. Norin, L. Alm, T. Midtvedt, A. Cresci, S. Silvi, C. Orpianesi, M.C. Verdenelli, et al. 2006. Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl. Environ. Microbiol.* 72: 1027–1033. https://doi.org/10.1128/AEM.72.2.1027-1033.2006

Mullard, A. 2018. Oncologists tap the microbiome in bid to improve immunotherapy outcomes. *Nat. Rev. Drug Discov.* 17:153–155. https://doi.org/10.1038/nrd.2018.19

Nakagawa, S., H. Fujii, Y. Machida, and K. Okuda. 1994. A longitudinal study from prepuberty to puberty of gingivitis. Correlation between the occurrence of Prevotella intermedia and sex hormones. *J. Clin. Periodontol.* 21:658–665. https://doi.org/10.1111/j.1600-051X.1994.tb00783.x

Narushima, S., Y. Sugiura, K. Oshima, K. Atarashi, M. Hattori, M. Suematsu, and K. Honda. 2014. Characterization of the 17 strains of regulatory T cell-inducing human-derived Clostridia. *Gut Microbes.* 5:333–339. https://doi.org/10.4161/gmic.28572

Nasko, D.J., S. Koren, A.M. Phillippy, and T.J. Treangen. 2018. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* 19:165. https://doi.org/10.1186/s13059-018-1554-6

Neuwirth, E. 2014. RColorBrewer: ColorBrewer Palettes. http://applied-r.com/rcolorbrewer-palettes/ (accessed August 18, 2020).

Oksanen, J., F.G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, et al 2019. vegan: Community Ecology Package. https://rdrr.io/rforge/vegan/ (accessed August 18, 2020).

Parks, D.H., M. Chuvochina, D.W. Waite, C. Rinke, A. Skarshewski, P.A. Chaumeil, and P. Hugenholtz. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36:996–1004. https://doi.org/10.1038/nbt.4229

Partula, V., S. Mondot, M.J. Torres, E. Kesse-Guyot, M. Deschasaux, K. Assmann, P. Latino-Martel, C. Buscail, C. Julia, P. Galan, et al. Milieu Intérieur Consortium. 2019. Associations between usual diet and gut microbiota composition: results from the Milieu Intérieur cross-sectional study. *Am. J. Clin. Nutr.* 109:1472–1483. https://doi.org/10.1093/ajcn/nqz029

Patin, E., M. Hasan, J. Bergstedt, V. Rouilly, V. Libri, A. Urrutia, C. Alanio, P. Scepanovic, C. Hammer, F. Jönsson, et al. Milieu Intérieur Consortium. 2018. Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat. Immunol.* 19:302–314. https://doi.org/10.1038/s41590-018-0049-7

Peters, B.A., M. Wilson, U. Moran, A. Pavlick, A. Izsak, T. Wechter, J.S. Weber, I. Osman, and J. Ahn. 2019. Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome Med.* 11:61. https://doi.org/10.1186/s13073-019-0672-4

Poyet, M., M. Groussin, S.M. Gibbons, J. Avila-Pacheco, X. Jiang, S.M. Kearney, A.R. Perrotta, B. Berdy, S. Zhao, T.D. Lieberman, et al. 2019. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25:1442–1452. https://doi.org/10.1038/s41591-019-0559-3

R Core Team. 2019. R: A Language and Environment for Statistical Computing. In R Foundation for Statistical Computing, Vienna, Austria.

Rothschild, D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, D. Zeevi, P.I. Costea, A. Godneva, I.N. Kalka, N. Bar, et al. 2018. Environment dominates over host genetics in shaping human gut microbiota. *Nature.* 555:210–215. https://doi.org/10.1038/nature25973

Routy, B., E. Le Chatelier, L. Derosa, C.P.M. Duong, M.T. Alou, R. Daillère, A. Fluckiger, M. Messaoudene, C. Rauber, M.P. Roberti, et al. 2018. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science.* 359:91–97. https://doi.org/10.1126/science.aan3706

Savaiano, D.A., A.J. Ritter, T.R. Klaenhammer, G.M. James, A.T. Longcore, J.R. Chandler, W.A. Walker, and H.L. Foyt. 2013. Improving lactose digestion and symptoms of lactose intolerance with a novel galacto-oligosaccharide (RP-G28): a randomized, double-blind clinical trial. *Nutr. J.* 12:160. https://doi.org/10.1186/1475-2891-12-160

Scepanovic, P., F. Hodel, S. Mondot, V. Partula, A. Byrd, C. Hammer, C. Alanio, J. Bergstedt, E. Patin, M. Touvier, et al. Milieu Intérieur Consortium. 2019. A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. *Microbiome.* 7:130. https://doi.org/10.1186/s40168-019-0747-x

Schindler, D.E., R. Hilborn, B. Chasco, C.P. Boatright, T.P. Quinn, L.A. Rogers, and M.S. Webster. 2010. Population diversity and the portfolio effect in an exploited species. *Nature.* 465:609–612. https://doi.org/10.1038/nature09060

Schirmer, M., E.A. Franzosa, J. Lloyd-Price, L.J. McIver, R. Schwager, T.W. Poon, A.N. Ananthakrishnan, E. Andrews, G. Barron, K. Lake, et al. 2018. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat. Microbiol.* 3:337–346. https://doi.org/10.1038/s41564-017-0089-z

Schmieder, R., and R. Edwards. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics.* 27:863–864. https://doi.org/10.1093/bioinformatics/btr026

Stewart, C.J., N.J. Ajami, J.L. O'Brien, D.S. Hutchinson, D.P. Smith, M.C. Wong, M.C. Ross, R.E. Lloyd, H. Doddapaneni, G.A. Metcalf, et al. 2018. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature.* 562:583–588. https://doi.org/10.1038/s41586-018-0617-x

Tang, J., A. Shalabi, and V.M. Hubbard-Lucey. 2018. Comprehensive analysis of the clinical immuno-oncology landscape. *Ann. Oncol.* 29:84–91. https://doi.org/10.1093/annonc/mdx755

Thomas, A.M., P. Manghi, F. Asnicar, E. Pasolli, F. Armanini, M. Zolfo, F. Beghini, S. Manara, N. Karcher, C. Pozzi, et al. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25: 667–678. https://doi.org/10.1038/s41591-019-0405-7

Thomas, S., V. Rouilly, E. Patin, C. Alanio, A. Dubois, C. Delval, L.G. Marquier, N. Fauchoux, S. Sayegrih, M. Vray, et al. Milieu Intérieur Consortium. 2015. The Milieu Intérieur study - an integrative approach for study of human immunological variance. *Clin. Immunol.* 157:277–293. https://doi.org/10.1016/j.clim.2014.12.004

Tilman, D. 1999. The ecological consequences of changes in biodiversity: A search for general principles. *Ecology.* 80:1455–1474.

Truong, D.T., E.A. Franzosa, T.L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods.* 12:902–903. https://doi.org/10.1038/nmeth.3589

Turroni, F., E. Foroni, P. Pizzetti, V. Giubellini, A. Ribbera, P. Merusi, P. Cagnasso, B. Bizzarri, G.L. de'Angelis, F. Shanahan, et al. 2009. Exploring the diversity of the bifidobacterial population in the human intestinal tract. *Appl. Environ. Microbiol.* 75:1534–1545. https://doi.org/10.1128/AEM.02216-08

Walker, A. 2019. openxlsx: Read, Write and Edit XLSX Files. https://rdrr.io/cran/openxlsx/ (accessed August 18, 2020).

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis, 2nd Edition. Springer International, Basel, Switzerland. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., R. François, L. Henry, and K. Müller. 2019. dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org/reference/dplyr-package.html (accessed August 18, 2020).

Wilke, C.O. 2019. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2.' https://wilkelab.org/cowplot/ (accessed August 18, 2020).

Wood, D.E., and S.L. Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xie, Y. 2019. knitr: A General-Purpose Package for Dynamic Report Generation in R. https://rdrr.io/cran/knitr/ (accessed August 18, 2020).

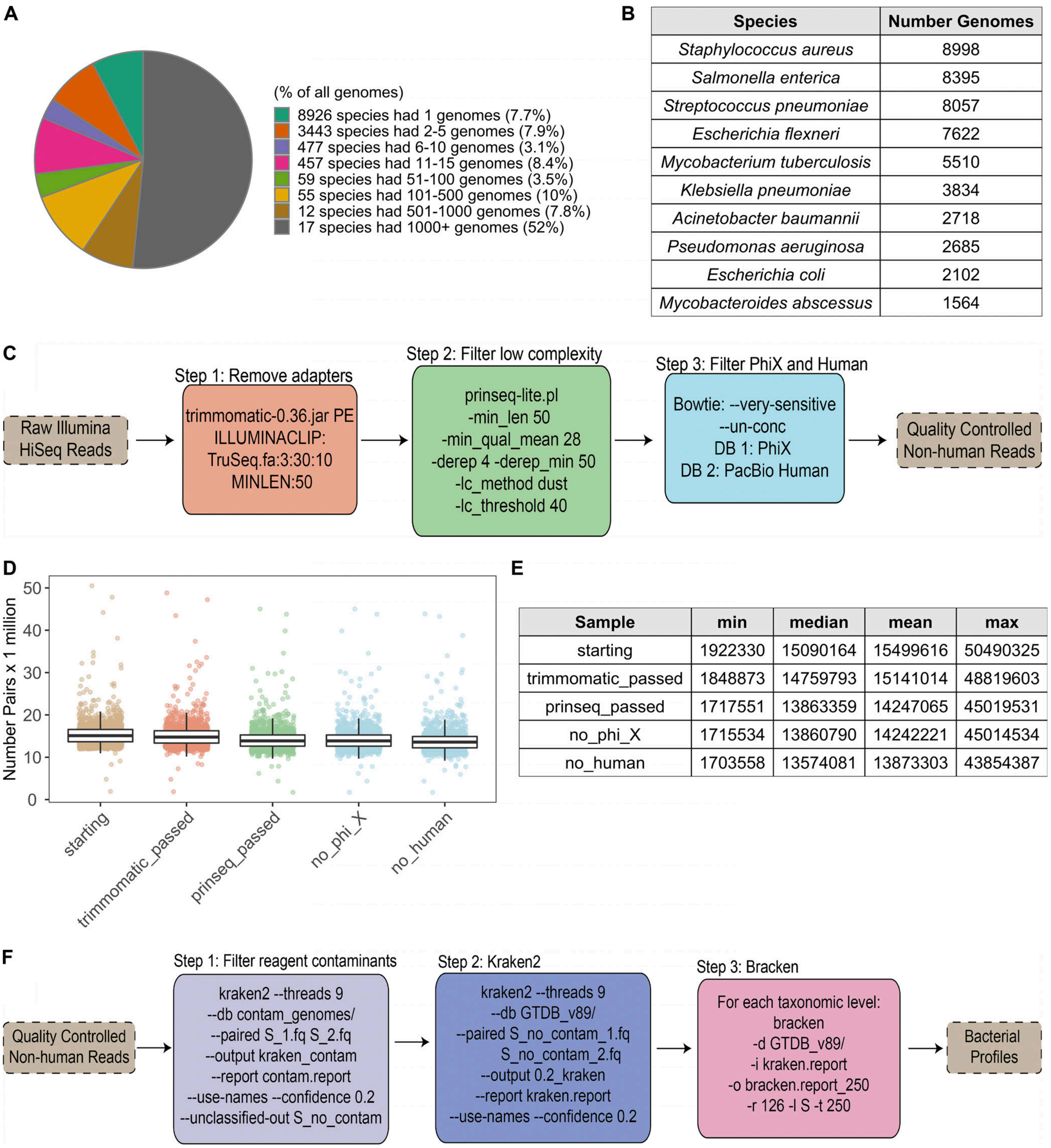
Zeevi, D., T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, et al. 2015. Personalized Nutrition by Prediction of Glycemic Responses. *Cell.* 163: 1079–1094. https://doi.org/10.1016/j.cell.2015.11.001

Zhernakova, A., A. Kurilshikov, M.J. Bonder, E.F. Tigchelaar, M. Schirmer, T. Vatanen, Z. Mujagic, A.V. Vila, G. Falony, S. Vieira-Silva, et al. LifeLines cohort study. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science.* 352: 565–569. https://doi.org/10.1126/science.aad3369

Zmora, N., G. Zilberman-Schapira, J. Suez, U. Mor, M. Dori-Bachash, S. Bashiardes, E. Kotler, M. Zur, D. Regev-Lehavi, R.B. Brik, et al. 2018. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell.* 174: 1388–1405.e21. https://doi.org/10.1016/j.cell.2018.08.041

Zou, Y., W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai, et al. 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* 37: 179–185. https://doi.org/10.1038/s41587-018-0008-8

Supplemental material



Species	Number Genomes
Staphylococcus aureus	8998
Salmonella enterica	8395
Streptococcus pneumoniae	8057
Escherichia flexneri	7622
Mycobacterium tuberculosis	5510
Klebsiella pneumoniae	3834
Acinetobacter baumannii	2718
Pseudomonas aeruginosa	2685
Escherichia coli	2102
Mycobacteroides abscessus	1564

Step 1: Remove adapters

```
trimmomatic-0.36.jar PE
ILLUMINACLIP:
TruSeq.fa:3:30:10
MINLEN:50
```

Step 2: Filter low complexity

```
prinseq-lite.pl
-min_len 50
-min_qual_mean 28
-derep 4 -derep_min 50
-lc_method dust
-lc_threshold 40
```

Step 3: Filter PhiX and Human

```
Bowtie: --very-sensitive
--un-conc
DB 1: PhiX
DB 2: PacBio Human
```

Sample	min	median	mean	max
starting	1922330	15090164	15499616	50490325
trimmomatic_passed	1848873	14759793	15141014	48819603
prinseq_passed	1717551	13863359	14247065	45019531
no_phi_X	1715534	13860790	14242221	45014534
no_human	1703558	13574081	13873303	43854387

Step 1: Filter reagent contaminants

```
kraken2 --threads 9
--db contam_genomes/
--paired S_1.fq S_2.fq
--output kraken_contam
--report contam.report
--use-names --confidence 0.2
--unclassified-out S_no_contam
```

Step 2: Kraken2

```
kraken2 --threads 9
--db GTDB_v89/
--paired S_no_contam_1.fq
S_no_contam_2.fq
--output 0.2_kraken
--report kraken.report
--use-names --confidence 0.2
```

Step 3: Bracken

```
For each taxonomic level:
bracken
-d GTDB_v89/
-i kraken.report
-o bracken.report_250
-r 126 -l S -t 250
```

Figure S1. **Quality control and Kraken analysis pipeline.** Related to Fig. 1. **(A)** Pie chart indicating the distribution of genomes in RefSeq (as of June 25, 2019) to different species, based on GTDB taxonomy. 29 species represent >60% of the reference genomes. **(B)** Table of the 10 species with the greatest number of genomes in RefSeq. **(C)** Steps taken to quality control shotgun metagenomic reads. Additional details can be found in Materials and methods. **(D)** Bar plots show the number of paired-end reads remaining after each filtration step in C. Each point corresponds to one sample (*n* = 1,359). Colors correspond to steps in C. **(E)** Summary table for the number of paired-end reads in D. max, maximum; min, minimum. **(F)** Steps taken to run Kraken2 and Bracken2 on the quality-controlled reads, including removal of putative reagent contaminants. Additional details can be found in Materials and methods. See also Data S1, table 2.

A

Level	Metaphlan2-NCBI	Kraken2-NCBI	Kraken2-GTDB
Phylum	15	43	52
Class	26	95	112
Order	46	211	287
Family	164	493	670
Genus	1098	2397	3205
Species	6278	18795	13446

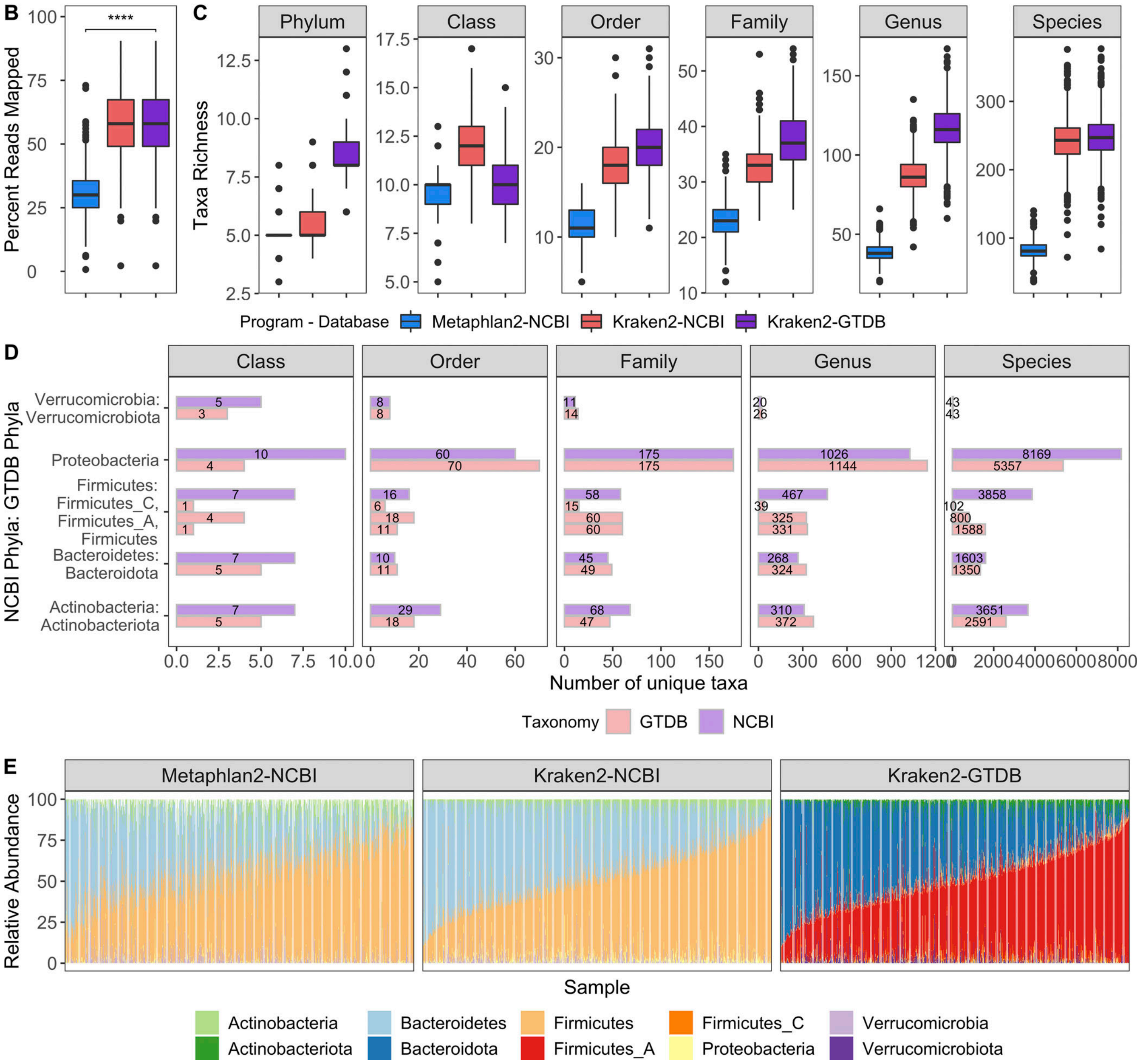


Figure S2. **Comparison of all program–database combinations tested.** Related to Fig. 1. **(A)** Table indicates the richness across taxonomic levels of the different databases tested. The NCBI and GTDB databases used by Kraken2 include the same 23,505 RefSeq genomes with taxonomies assigned either by NCBI or the GTDB, respectively. GTDB taxonomic assignments often lead to collapsing or splitting of those taxa originally in NCBI; e.g., Firmicutes was split into Firmicutes, Firmicutes_A, Firmicutes_B, Firmicutes_C, etc. **(B)** Boxplots compare the percentage of shotgun metagenomic reads mapped by the different program–database combinations. Percentage of mapped reads for Metaphlan2 was based on its estimated number of reads from the clade; ****, $P < 2.22 \times 10^{-16}$ by Wilcoxon rank sum. **(C)** Boxplots compare the richness of taxa identified per sample with the different program–database combinations tested. All pairwise comparisons had $P < 0.0001$ by Wilcoxon rank sum. **(D)** Bars indicate the number of unique taxa present in the NCBI (purple) and GTDB (red) databases for each of the top phyla across taxonomic levels. **(E)** Relative abundance of the top phyla across different program–database combinations. Each bar corresponds to one sample. Samples are ordered based on the relative abundance of Firmicutes_A found by Kraken2-GTDB.

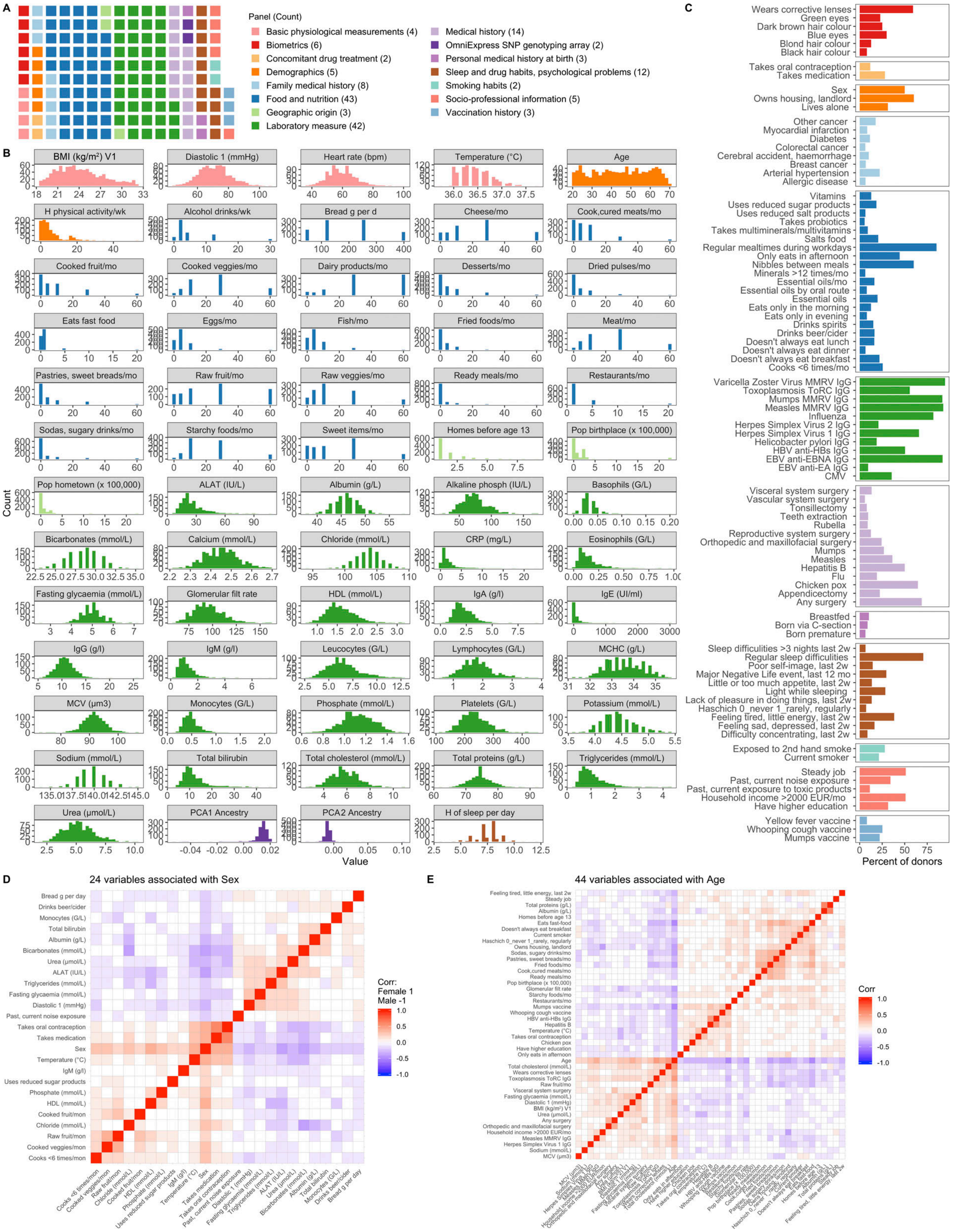
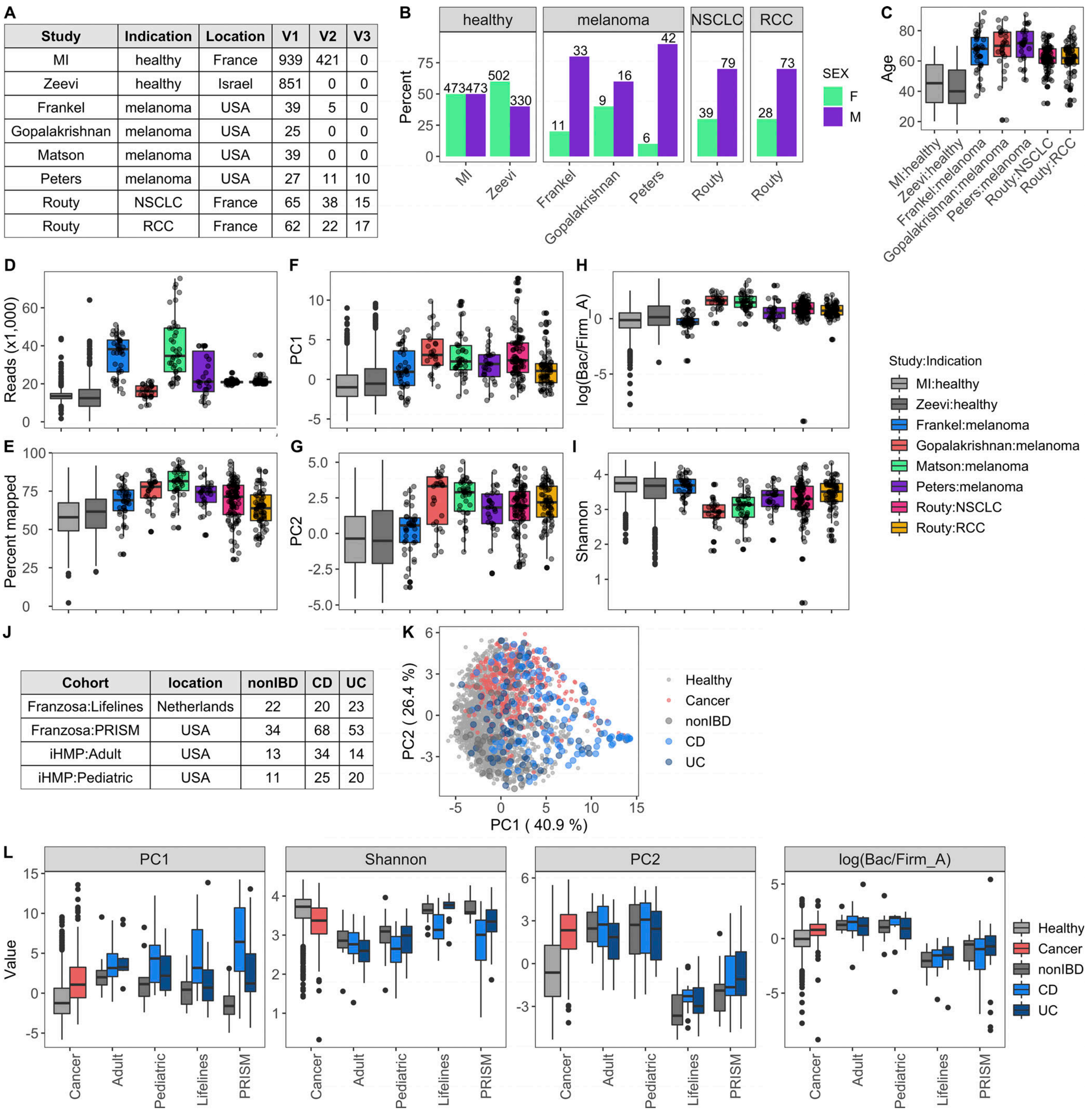


Figure S3. **154 variables were associated with bacterial profiles.** Related to Fig. 1. **(A)** Distribution of variables across broad categories. **(B)** Distribution of 64 continuous variables across donors. Colors correspond to those in A. **(C)** Percentage of donors for whom each of the 90 binary variables was true. Colors correspond to those in A. 2w, 2 wk; G/L, billion cells per liter; HBV, hepatitis B virus; MCHC, mean corpuscular hemoglobin concentration; mon, month; PC, principal component of genetics SNP array. **(D)** 24 variables were associated with sex with Spearman ρ > 0.2 or less than −0.2. Corr, correlation. **(E)** 44 variables were associated with age with Spearman ρ > 0.2 or less than −0.2. See also Data S1, tables 5 and 6. CMV, cytomegalovirus; CRP, C-reactive protein; EUR, Euro; HDL, high density lipoprotein; MCV, mean corpuscular volume; MMRV, measles, mumps, rubella, and varicella vaccine; PCA, principal component ancestry.



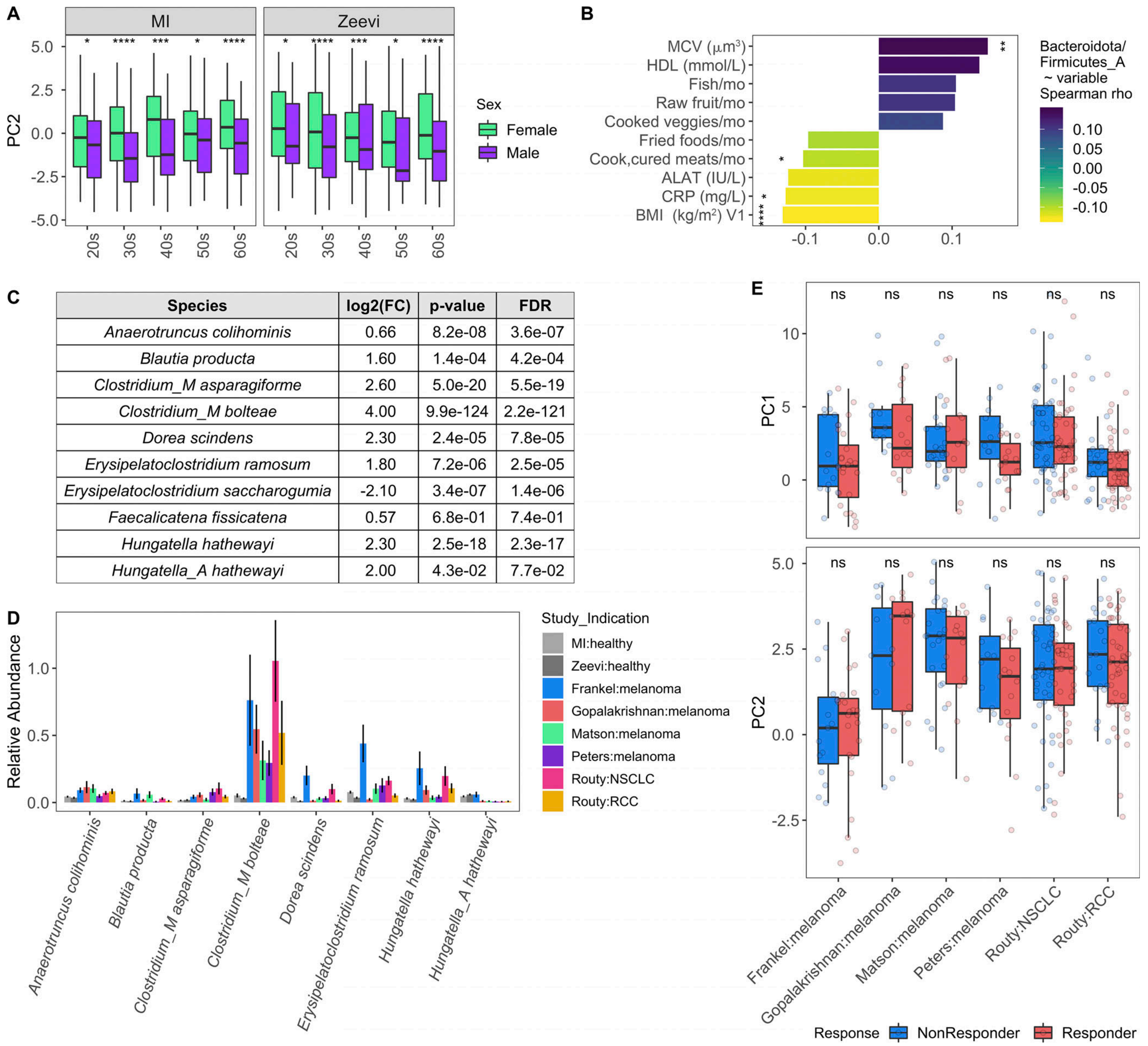
A

Study	Indication	Location	V1	V2	V3
MI	healthy	France	939	421	0
Zeevi	healthy	Israel	851	0	0
Frankel	melanoma	USA	39	5	0
Gopalakrishnan	melanoma	USA	25	0	0
Matson	melanoma	USA	39	0	0
Peters	melanoma	USA	27	11	10
Routy	NSCLC	France	65	38	15
Routy	RCC	France	62	22	17

J

Cohort	location	nonIBD	CD	UC
Franzosa:Lifelines	Netherlands	22	20	23
Franzosa:PRISM	USA	34	68	53
iHMP:Adult	USA	13	34	14
iHMP:Pediatric	USA	11	25	20

Figure S4. **Across studies and indications, cancer patients have altered gut bacterial profiles.** Related to Fig. 7. **(A)** Table indicating the studies included in all cancer versus healthy analyses. For all but the longitudinal analysis, one sample per patient was included. When available, the baseline sample was prioritized. For cancer patients, V1 = baseline, pretreatment; V2-Frankel = within 1 mo of starting ICT; V2-Peters = week 6; V2-Routy = after second injection, ∼1 mo; V2-Peters = week 12; V3-Routy = after fourth injection, ∼2 mo. **(B)** Bars indicate percentage of males and females in each of the studies. Number indicates the actual number of males and females. Sex data were not provided by Matson et al. (2018). **(C)** Boxplots compare age of donors across the studies. Age data were not provided by Matson et al. (2018). **(D–I)** Boxplots compare values indicated on the y axis across the healthy and cancer cohorts. Color corresponds to the study. **(J)** Table indicating the number of donors in each of the IBD cohorts from the two studies. Patients in iHMP <20 yr old were grouped in the pediatric cohort. **(K)** PC plot of BC distance of bacterial species composition of donors in Fig. 7 A as well as donors in the IBD cohorts. For the HMP samples, multiple samples from the same donor were aggregated together, so there is only one dot per donor. Color corresponds to the health status. **(L)** Boxplots compare PC1, PC2, Shannon, and Bacteroidota (Bac)/Firmicutes_A (Firm) ratios of samples in K. On the x axis, "cancer" is all the samples in Fig. 7; the remaining are the IBD cohorts in J. CD, Crohn's disease; ICT, immune checkpoint inhibitor; NSCLC, non–small cell lung cancer; RCC, renal cell carcinoma; UC, ulcerative colitis.



Species	log2(FC)	p-value	FDR
Anaerotruncus colihominis	0.66	8.2e-08	3.6e-07
Blautia producta	1.60	1.4e-04	4.2e-04
Clostridium_M asparagiforme	2.60	5.0e-20	5.5e-19
Clostridium_M bolteae	4.00	9.9e-124	2.2e-121
Dorea scindens	2.30	2.4e-05	7.8e-05
Erysipelatoclostridium ramosum	1.80	7.2e-06	2.5e-05
Erysipelatoclostridium saccharogumia	-2.10	3.4e-07	1.4e-06
Faecalicatena fissicatena	0.57	6.8e-01	7.4e-01
Hungatella hathewayi	2.30	2.5e-18	2.3e-17
Hungatella_A hathewayi	2.00	4.3e-02	7.7e-02

Figure S5. **Variables associated with cancer-like microbiome.** Related to Fig. 8. **(A)** Boxplots compare PC2 values of samples from males and females stratified by study and age. *, P < 0.05; ***, P < 0.001; ****, P < 0.0001 by Wilcoxon rank sum with FDR correction. **(B)** Length and color of the bars indicates the Spearman ρ value of each of the variables by ratio of Bacteroidota/Firmicutes_A. Only those variables with an FDR < 0.05 are shown. Stars indicate the P value after correcting for sex and age in a linear model. *, P < 0.05; **, P < 0.01; ****, P < 0.0001. **(C)** Table shows DESeq2 results for species identified by Atarashi et al. (2013) and Narushima et al. (2014) as potent inducers of regulatory T cells. Only species assigned a name in Narushima et al. (2014) that could then be linked to a representative in our database were included. **(D)** For the most abundant species in C, bars indicate the mean relative abundance of the species in each study. Error bars indicate mean ± SE. **(E)** Boxplots compare PC1 and PC2 values of samples from responders and nonresponders across studies. P values by Wilcoxon rank sum with FDR correction; ns, not significant (P > 0.05). Definitions of responders and nonresponders by study: Frankel R, response, stable disease; Frankel NR, progressed; Matson R, complete response, partial response; Matson NR, stable disease, progressed; Routy R, complete response, partial response, or stable disease; Routy NR, progressed or died; Gopalakrishnan R, complete response, partial response, or stable disease for ≥6 mo; Gopalakrishnan NR, progressed or stable disease <6 mo; Peters R, did not progress; Peters NR, progressed. See also Data S1, tables 19–21. NSCLC, non–small cell lung cancer; RCC, renal cell carcinoma.

Provided online is a data file which contains 22 tables. Table 1 lists the species and genomes included in the GTDB database. Table 2 gives basic metadata for each sample, stats on the number of reads, and species and pathway diversity. Table 3 provides summary information for each of the taxa identified in the MI cohort. Table 4 lists summary information for each metabolic pathway identified in the MI cohort. Table 5 displays summary stats for each of the metadata variables. Table 6 shows information on how

the different metadata variables are correlated with one another: Spearman rho values, P values, and FDR correct P values. Table 7 provides values from the PERMANOVA analysis. Table 8 gives summary information for each of the taxa identified in the Zeevi cohort. Table 9 lists summary stats as well as DESeq2 and prop.test results for taxa associated with sex in the MI cohort. Table 10 shows summary stats as well as DESeq2 and prop.test results for taxa associated with sex in the Zeevi cohort. Table 11 provides summary stats as well as Wilcox and prop.test results for pathways associated with sex in the MI cohort and gives correlations of pathways associated with sex × species. Table 12 displays Spearman results for all taxa × age correlations in the MI and GLM results for all taxa × age correlations in the MI cohort. Table 13 lists Spearman and GLM results for all taxa × age correlations in the Zeevi cohort. Table 14 gives Spearman results for all microbial pathways × age correlations in the MI cohort and Spearman correlations of pathways associated with age × species. Table 15 shows Spearman results for species and pathway stability × baseline phyla relative abundances. Table 16 provides GLM results for metadata variables × species stability. Table 17 lists stability values for each species. Table 18 gives stability values for each pathway. Table 19 displays correlations between metadata variables × PC1, PC2, and Bacteroidota/Firmicutes_A ratios. Table 20 shows summary information for species across the different healthy and cancer cohorts. Table 21 provides DESeq2 and prop.test results for taxa associated with health status. Table 22 lists species identified as putative reagent contaminants in the MI samples.