Enterosignatures define common bacterial guilds in the human gut microbiome

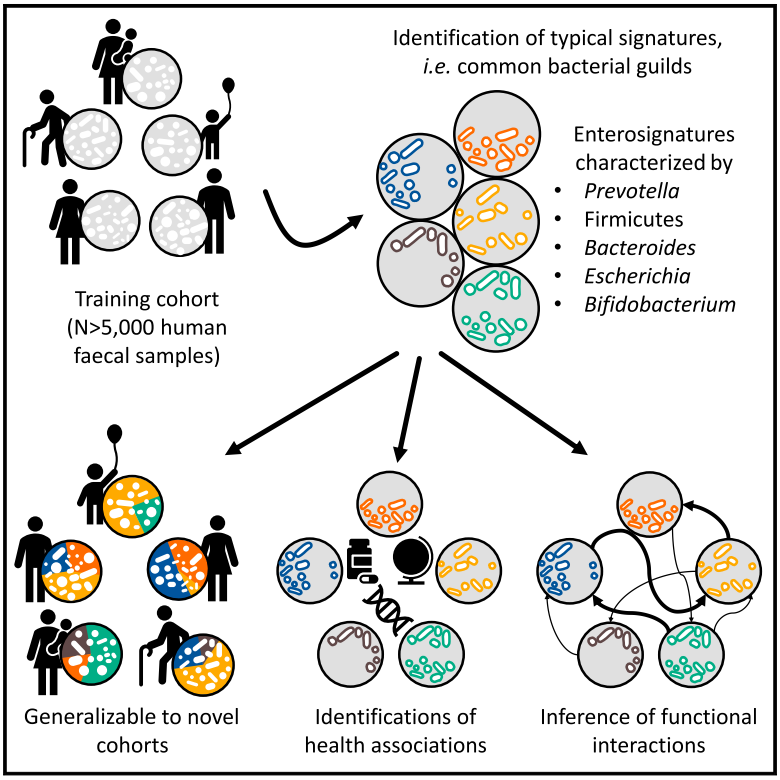
Clémence Frioux, Rebecca Ansorge, Ezgi Özkurt, Chabname Ghassemi Nedjad, Joachim Fritscher, Christopher Quince, Sebastian M Waszak, Falk Hildebrand

Article

Cell Host & Microbe

Enterosignatures define common bacterial guilds in the human gut microbiome

Graphical abstract



Highlights

- The human gut microbiome is described as combinations of five enterosignatures
- Enterosignatures represent microbial guilds, complementary in their metabolism
- Enterosignature composition is dynamic and changes with host's age
- Gut microbiomes deviating from this model could indicate dysbiotic states

Authors

Clémence Frioux, Rebecca Ansorge, Ezgi Özkurt, ..., Christopher Quince, Sebastian M. Waszak, Falk Hildebrand

Correspondence

clemence.frioux@inria.fr (C.F.), falk.hildebrand@quadram.ac.uk (F.H.)

In brief

Frioux et al. introduce enterosignatures as microbial guilds whose assemblies are accurate descriptors of the human gut microbiome composition. Enterosignature composition relates to changes and perturbations in the microbiome over a lifetime. The model generalizes to diverse human populations and can be used to detect anomalies in the gut microbiome.

Frioux et al., 2023, Cell Host & Microbe *31*, 1–15
July 12, 2023
https://doi.org/10.1016/j.chom.2023.05.024

CellPress

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

Cell Host & Microbe

CellPress
OPEN ACCESS

Article

Enterosignatures define common bacterial guilds in the human gut microbiome

Clémence Frioux,[1,2,3,*] Rebecca Ansorge,[1,2] Ezgi Özkurt,[1,2] Chabname Ghassemi Nedjad,[3] Joachim Fritscher,[1,2] Christopher Quince,[1,2] Sebastian M. Waszak,[4,5,6] and Falk Hildebrand[1,2,7,*]

[1]Food, Microbiome, and Health Institute Strategic Programme, Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ Norwich, Norfolk, UK
[2]Digital Biology, Earlham Institute NR4 7UZ Norwich, Norfolk, UK
[3]Inria, University of Bordeaux, INRAE, 33400 Talence, France
[4]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo and Oslo University Hospital, Oslo 0318, Norway
[5]Department of Neurology, University of California, San Francisco, San Francisco, CA 94148, USA
[6]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany
[7]Lead contact
*Correspondence: clemence.frioux@inria.fr (C.F.), falk.hildebrand@quadram.ac.uk (F.H.)
https://doi.org/10.1016/j.chom.2023.05.024

SUMMARY

The human gut microbiome composition is generally in a stable dynamic equilibrium, but it can deteriorate into dysbiotic states detrimental to host health. To disentangle the inherent complexity and capture the ecological spectrum of microbiome variability, we used 5,230 gut metagenomes to characterize signatures of bacteria commonly co-occurring, termed enterosignatures (ESs). We find five generalizable ESs dominated by either *Bacteroides*, Firmicutes, *Prevotella*, *Bifidobacterium*, or *Escherichia*. This model confirms key ecological characteristics known from previous enterotype concepts, while enabling the detection of gradual shifts in community structures.

Temporal analysis implies that the *Bacteroides*-associated ES is "core" in the resilience of westernized gut microbiomes, while combinations with other ESs often complement the functional spectrum. The model reliably detects atypical gut microbiomes correlated with adverse host health conditions and/or the presence of pathobionts. ESs provide an interpretable and generic model that enables an intuitive characterization of gut microbiome composition in health and disease.

INTRODUCTION

The gastrointestinal tract (GIT) microbiome typically consists of 200–300 different species, which, at the strain level, are mostly unique to their human host.[1] The taxonomic composition of this ecosystem is generally in a stable equilibrium shaped by external and internal processes, such as diet, immune system, and keystone species enabling ecosystem function.[2,3] Due to its complexity, decomposition of the gut microbiome into a few, universally applicable characteristics is a key goal for enabling accessible medical gut microbiome research. Taxonomically informed decompositions such as summarizing the microbiome at the phylum level have been used in the past to determine the ratios of dominant phyla (Bacteroidota:Firmicutes), which correlate with some physiological host characteristics[4] (Figure 1A).

However, taxonomic decompositions do not consider intrinsic ecosystem information, such as bacteria preferentially co-occurring or growing under similar conditions. Such ecologically informed decompositions of the gut microbiome can be obtained using ordinations[5] or by clustering genus-level compositions into *enterotypes* (ETs).[6] Two popular clustering algorithms used to calculate ETs are partitioning around medoids (PAMs) or Dirichlet multinomial mixture models (DMMs). PAM typically identifies three ETs enriched in key taxa *Bacteroides* (ET-B), *Prevotella* (ET-P), and *Ruminococcus* (Firmicutes-enriched, ET-F),[6] whereas the DMM typically identifies four ETs, recovering a second *Bacteroides*-dominated cluster (ET-Bact2).[7] Additionally, some studies concluded that either no ETs or only two ETs (*Bacteroides* and *Prevotella* dominated) exist in gut microbiomes.[8,9] These clustering-based approaches enforce discrete divisions,[9] implicitly assuming that each community in a sample derives from one type (Figure 1A). For example, the physiologically relevant ratio of Bacteroidota:Firmicutes[4] only reflects the presence of either a Firmicutes or *Bacteroides*/*Prevotella* ET and lacks the resolution to capture fine-grained co-existence of bacterial guilds. Instead, defining compositions of bacterial guilds encapsulates both ecologically informed bacterial assemblages and the proportional representation of these within single samples. Such assemblages can be calculated using a different form of multivariate analysis algorithm, where the latent variables are continuous and not

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

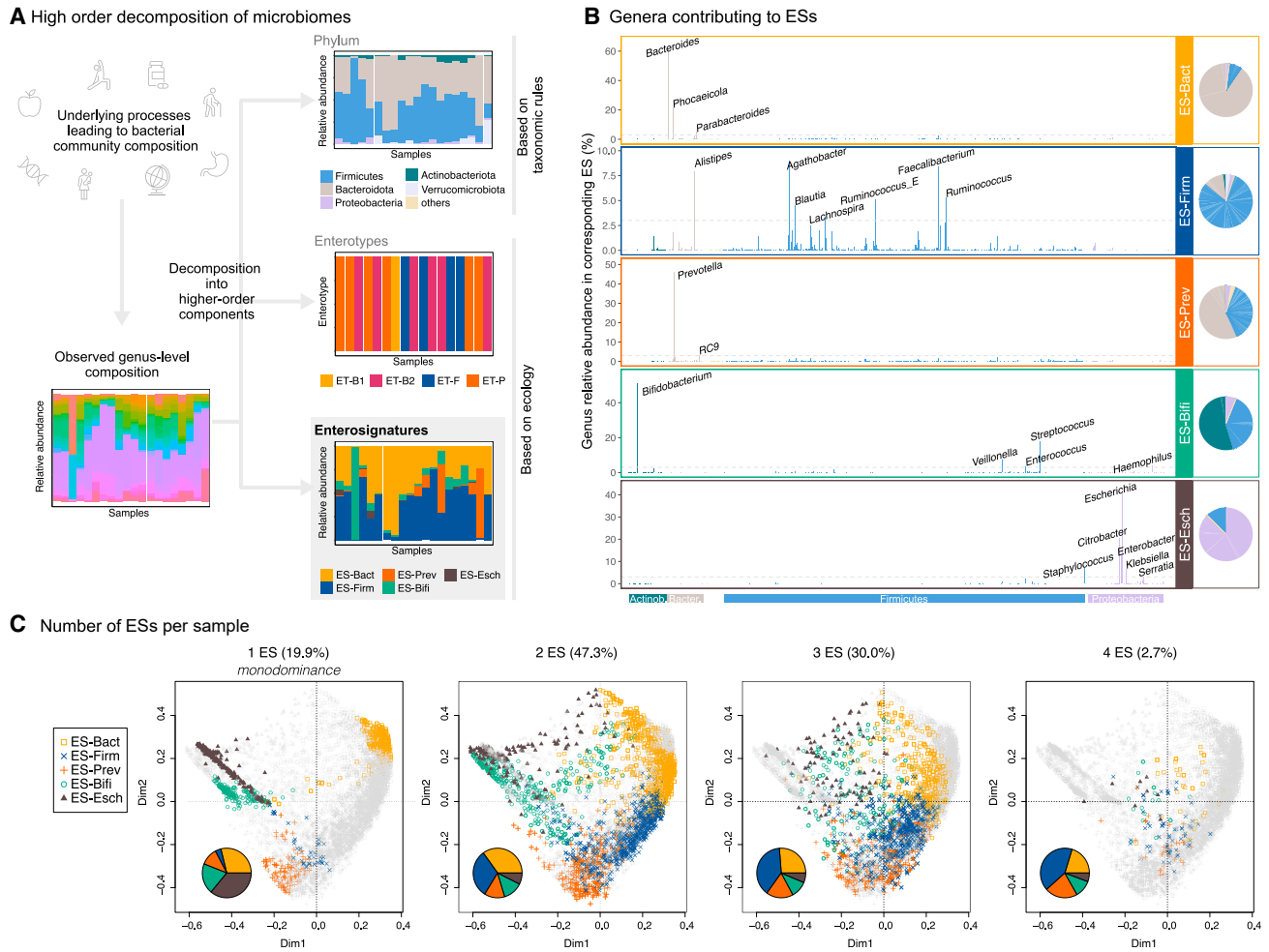


Figure 1. Composition of enterosignatures (ESs) and their diversity in samples

(A) The concept of ES: observed genus-level composition of the microbiome can be simplified using higher-order decompositions, such as taxonomically informed phylum levels, or ecology-informed enterotypes. In the here-proposed enterosignature concept, non-negative matrix factorization (NMF) is used to determine bacterial guilds driving variance in gut metagenomes.

(B) Relative genus-level composition of the five ESs identified as the optimal NMF solution.

(C) PCoA of genus Bray-Curtis's dissimilarity of 5,230 gut metagenomes from the GMR dataset, colored by their primary ES assignment. In each subplot, only samples that are explained by either 1, 2, 3, or 4 ES are colored. Pie chart shows the fraction of ES present in each subplot.

Abbreviations are as follows: ESs, enterosignatures; ETs, enterotypes; PCoA, principal coordinate analysis. See also Figure S1.

discrete, such as latent Dirichlet allocation (LDA) or non-negative matrix factorization (NMF), which have been applied to microbial ecosystems ranging from the gut to lake metagenomes.[10–14] NMF has the added advantage that pre-calculated signatures of bacterial assemblages can be reapplied to even a single metagenome, removing the need for large cohort sizes as required in ordinations (e.g., PCA) or clustering-based approaches (e.g., ET) capturing microbiome variation.

Using NMF,[15] we propose an ecologically informed decomposition of human gut microbiomes into five microbial signatures or "*enterosignatures*" (ESs). Cross-validation suggests that these five ESs were able to ubiquitously describe variation in gut metagenomes from western as well as non-western (NW) fecal metagenomes of all ages. The model is provided as a community resource at https://enterosignatures.quadram.ac.uk/. Combinations of signatures were inherent to most gut microbiomes; commonly, two or three ESs best described fecal metagenomes. We demonstrated, using metabolic modeling, the increased functional potential of ES combinations and that *Bacteroides*-ES (ES-Bact) has a central role in establishing and maintaining core gut functionality. Our ES model also enabled the detection of atypical fecal microbiomes associated with infant birth mode and antibiotic use in adults, thereby allowing the detection of potentially dysbiotic gut microbiomes.

RESULTS

The gut microbiome composition is accurately described by combinations of enterosignatures

To investigate if the human fecal microbiome could be generalized by NMF signatures, we used a diverse set of 5,230 fecal metagenomes from the study of Hildebrand et al.[1] to train our

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

data model. This dataset is referred to as gut microbiome reference (GMR) and represents human populations from 13 countries (three continents) and different age groups (2,169 samples from infants < 3 years old; 1,943 from adults ≥ 16 years old, and 1,059 > 60 years old classified as elders). To identify recurrent signatures, NMF was applied to the relative abundances of genera (Figure S1A). The algorithm identified signatures that represent guilds of co-occurring bacterial genera that we termed enterosignatures (ESs) (Figure 1A) and describing variance in the original genus-level composition space.

With increasing numbers of ESs, the NMF model reconstructed increasingly more variance of genus-level abundances (from 37.9% at k = 2 to 79.9% at k = 10). To mitigate the risk of overfitting, we used a 3×3 bi-cross validation to identify generalized ESs that represented a balance between explained microbial variance in GMR and applicability to new datasets. This approach identified five ESs (Figures S1A–S1D; STAR Methods) and explained 64% of, reaching 0.8 cosine similarity to, the original genus abundances.

Each ES is a weighted combination of several genera: some ESs are strongly dominated by one or two genera, whereas others are combinations of multiple genera with weaker associations (Figures 1B, 2B, and S1E; see Table S1 in Data S1). The most prominent ESs were found at lower signature numbers (k < 5, Figure S1G); specifically, we identified in order of emergence: (1) ES-Bact (mostly characterized by the genera *Bacteroides* and *Phocaeicola*), (2) ES-Bifi (*Bifidobacterium* and *Streptococcus*), (3) ES-Prev (*Prevotella*), (4) ES-Esch (*Escherichia* and *Citrobacter*), (5) ES-Firm (genera in the phylum Firmicutes) (Figure 1B). At higher cluster numbers, additional ESs emerged: bacterial guilds dominated by Enterobacteriaceae or *Staphylococcus* were found at k = 6 and k = 7, respectively, and a second Bacteroidota signature, mostly represented by the genus *Phocaeicola* at k = 8 (Figure S1H). Although the latter ESs represent bacterial guilds observed frequently and could therefore be relevant, our cross-validation results indicated these to be specific to our GMR dataset.

The NMF approach models coexisting ESs that represent together a particular fecal microbiome (Figure S1A; see Table S1 in Data S1), rather than assigning each sample to a single cluster type. Nonetheless, we recovered three signatures among the five ESs likely representing the community types first described as adult[6] ETs, indicating the strong biological signal exhibited by these guilds: ES-Bact, ES-Firm, and ES-Prev. The difference between assigning each fecal sample to an ET or representing a combination of ESs seems relevant, as the majority of fecal samples required two (47%) or three (30%) ESs to capture community variation (Figure 1C). Relative ES abundance varied strongly among individual samples (Figure S1F), demonstrating an additional level of ecological complexity captured in the ES model. As expected, adult ESs (ES-Bact, ES-Prev, and ES-Firm) were the dominant, or "*primary ES*" in most adult microbiomes but usually co-occurring with other ESs (85.4% of all adult samples, Figure 1C; see Table S2 in Data S1). In infant samples, this was different; ES-Bifi and ES-Esch were more often observed as the sole ES (>0.9 relative abundance in 9.4% and 17.4% of the samples, respectively). The presence of ES-Esch is of interest as it could reflect a pathological state in preterm babies where *Escherichia coli* is associated with necrotizing enterocolitis[16]; 79% of preterm samples were solely represented by ES-Esch.

Despite combinations of ESs frequently being observed in fecal samples, single genera were mostly associated with a single ES (Figure S1F; see Table S2 in Data S1). The median association strength between genera and their most strongly associated ES ranged from 90% (ES-Firm genera) to 100% (ES-Esch genera), indicating strong "loyalty" of taxa to their respective ES. This strong association is indicative of constituent bacteria in each ES forming predictable guilds. This suggests metabolic dependencies or mutual benefits among ES bacterial members, which could have co-evolved over time.

Enterosignatures are generic representations of microbial communities with strong host-phenotype associations

To compare ESs with classical ET models, we defined three ES abstraction levels: the relative abundance of each ES in a sample, the presence/absence of one or more ESs per sample, and the primary ES only that describes the strongest ES signature per sample—the latter most closely resembling classical ET concepts. Because both ET models—PAM (N = 3 clusters) and DMM (N = 4 clusters)—were only defined for adult fecal metagenomes, we restricted our comparisons to adult samples in the GMR dataset. We evaluated how well the original information in the genus matrix was represented by each approach (see Table S6 in Data S1). ET-like sample assignments via either DMM, PAM, or primary ES explained similar levels of variance in the original genus abundances ($\rho = 0.27$, $\rho = 0.41$, and $\rho = 0.41$, respectively), whereas discrete ES assignments and ES relative abundance captured up to twice the variance ($\rho = 0.61$ and $\rho = 0.73$, respectively) (Figure 2A), despite being restricted to a similar number of components.

To investigate the generalizability of our 5-ES model to a cohort from a similar background, we applied the pretrained model to the Metacardis body mass index spectrum (BMIS) cohort (n = 888 fecal metagenomes of European origin).[17] For comparison, ESs were also *de novo* calculated, as were PAM and DMM ETs. This showed a similar assignment of samples to either ET or primary ES (Figure S2A). Comparison between the two ES models showed that only a little additional information was captured in an ES model *de novo* trained on the BMIS cohort (median model fit 0.91 compared with 0.86, Figure 2B). The pre-trained 5-ES model was therefore retained in all following analyses. Following our GMR-based analysis, we calculated correlations between each type of ES and ET assignments and the original genus-level BMIS data and, as with our previous observations, found that more variance was captured by ES than ET models (Figure 2A).

The gut microbiome is correlated to different host physiological processes; thus, we hypothesized that ES and their relative abundance are correlated to different host-derived metadata variables that can explain their functional roles. In both BMIS and GMR datasets, ES explained more variation in host metadata, although most of the significantly correlated variables in these different models were similar (Figure 2C; see Table S6 in Data S1). In contrast to host physiological metadata, microbiome-intrinsic variables (gene richness and fecal microbial load) correlated more strongly with DMM (18.4%) and PAM (14.8%) ETs than ES (12%) and genus (8%) abundances

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

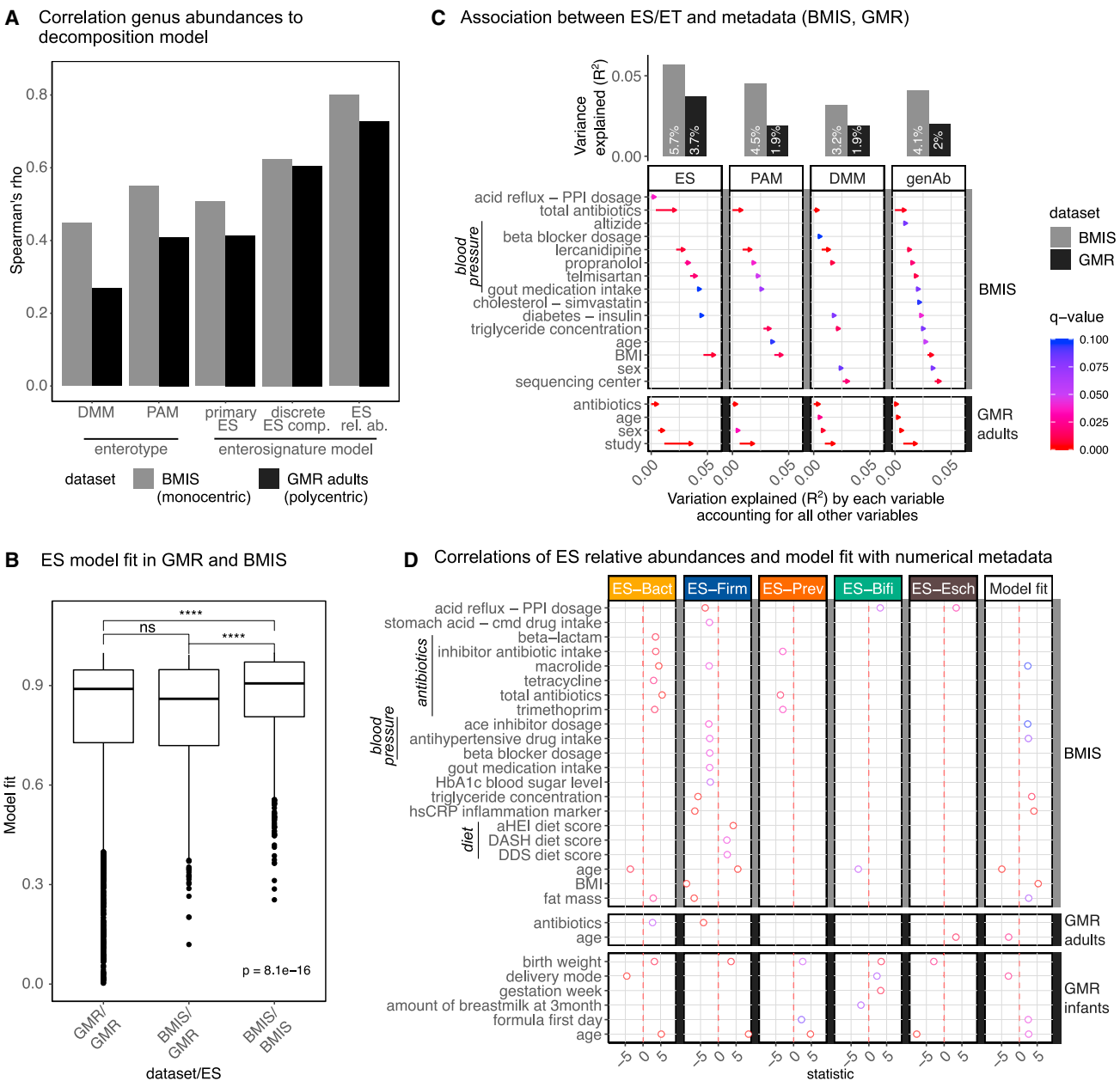


Figure 2. Comparison between enterosignatures (ESs) and enterotypes (ETs) concepts

(A) Spearman's correlation of ET (PAM, DMM) and ES concepts with the underlying genus-level abundance matrix of GMR (adult samples only) or Metacardis BMIS datasets. ES assignments were divided into primary ES assignment (similar to ET assignments), ES presence (set of ES with cumulative relative abundances > 0.9), or ES relative abundance compositions.

(B) ES can be reapplied to novel datasets with good model fit scores. This was demonstrated by measuring the model fit of ES that had been trained and applied to GMR data (GMR/GMR), ES trained on GMR and applied to BMIS (BMIS/GMR) or trained and applied to BMIS (BMIS/BMIS). P values were calculated with a Kruskal-Wallis test.

(C) Variation in ES, ET, and genera abundances were partly explained by metadata in BMIS and GMR cohorts. Only significant associations (q value < 0.1) were included.

(D) Significant Spearman correlations (q value < 0.1) between metadata from the BMIS and GMR (blocked by cohort) and the five ES. Dashed lines represent Z-statistic of zero to show negative (to the left) and positive (to the right) correlations. A positive correlation with delivery mode reflects vaginal birth, whereas a negative correlation reflects C-section birth. In (A), (C), and (D), ES applied to BMIS were those calculated using the GMR dataset.

Abbreviations are as follows: ESs, enterosignatures; ETs, enterotypes; BMIS, body mass index spectrum; GMR, gut microbiome reference; DMM, Dirichlet multinomial mixture models; PAMs, partitioning around medoids. See also Figure S2.

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

CellPress
OPEN ACCESS

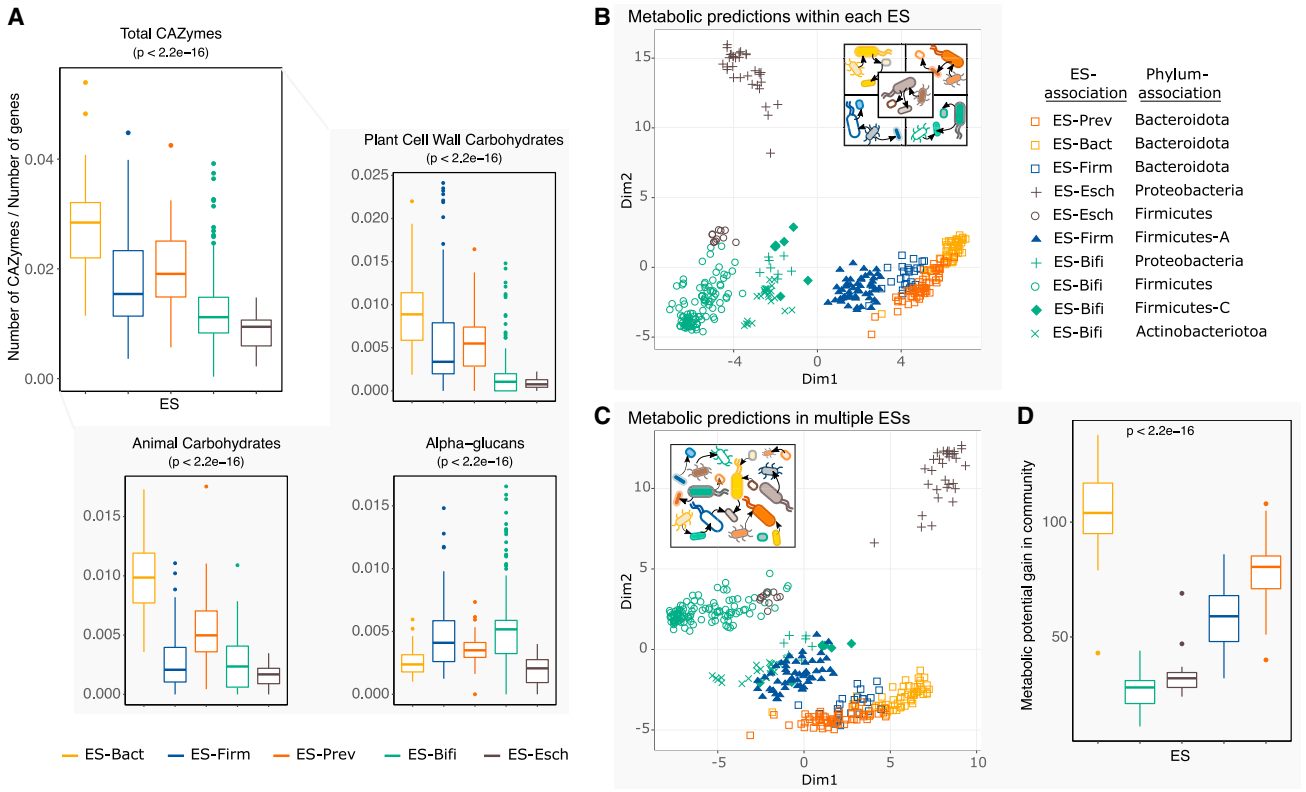


Figure 3. Metabolic potential of enterosignatures (ESs)

(A) Total number of CAZymes in ES representative genomes, normalized by the number of genes in genomes, and the number of CAZymes degrading, respectively, plant cell wall carbohydrates, animal carbohydrates, and alpha-glucans (typically in starch and glycogen) in ES representative genomes, normalized by the number of genes in genomes. Colors represent the 5 ESs.

(B and C) PCoA (Euclidean distance) of the predicted metabolic potential of each MGS computed in interaction with members of its own ES (B) or with the full community of ES. (C) Shape of points in (B) and (C) denote the phylum of each MGS, point color represents the ES each MGS is a member of.

(D) Metabolic gain per MGS when simulating cellular metabolism either within its native ES or within a community of all five ES. Metabolic gain is the difference in producible metabolites between these two conditions. ES-Bact MGS do significantly increase their metabolic spectrum, when in diverse microbial communities. Colors represent the 5 ESs. P values were calculated with a Kruskal-Wallis test.

Abbreviations are as follows: CAZyme, carbohydrate active enzyme; ES, enterosignature; MGS, metagenomic species; PCoA, principal coordinate analysis. See also Figure S3.

(Figure 2B). It is worth noting that both ES and PAM can explain more metadata variance than the original genus abundance data (Figure 2C). Similar effects have been observed before, where data reduction models can smooth observed predictions and thus lead to better predictions.[18] The increased variance in both community composition and physiological metadata that is explained by ESs demonstrates their usefulness for interpreting gut microbial datasets; the level of abstraction simplifies complex microbiomes to a few factors, thus enabling impactful health data interpretation.

For example, ES-Bact and ES-Prev were, respectively, positively and negatively correlated to antibiotic intake, in both adult and infant metagenomes (Figure 2D). It is noteworthy that genus *Prevotella* or *Prevotella* ET prevalence typically increases in rural communities and/or NW countries, linked to a diet rich in agrarian products[1,9]; our data suggest that this could alternatively be associated with reduced antibiotic intake in these communities. ES-Firm appeared to associate well with markers of host health, based on the anti-correlation with BMI, intake of various medications, fat mass, inflammation markers, blood sugar levels, and positive correlation with healthy eating index and host age. Furthermore, ES-Firm was the only ES positively correlated with fecal microbial cell counts, indicative of a thriving ecosystem and congruent with previous Firmicutes ET observations[19] (Figure S2C). In the GMR meta-cohort, all correlations were corrected for the study effect, but the metadata for adults was sparse; the few significant associations found largely agreed with the BMIS observations. However, in infants, some notable associations with ESs were detected: ES-Bact was correlated with vaginal birth, in line with previous findings[20]; there were negative correlations between ES-Esch and age and birth weight, in line with our other observation that this ES is potentially pathogenic.

Enterosignatures represent diverse and complementary metabolic landscapes

The functional potential present in the bacterial guilds represented in each ES may be vastly different, offering specific services to their host and the gut ecosystems. To investigate this, we reconstructed the bacterial genomes of key taxa in the

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

much larger GMR dataset[1,21] (see STAR Methods; Figure S3A). We first analyzed metabolism-related annotations of the genomes (Figures 3A–3D and S3; see Table S3 in Data S1) and found that dominant metabolic processes varied among the five ESs. For example, ES-Bifi genomes were enriched in amino acid, carbohydrate, lipid, and secondary metabolic functions, whereas ES-Esch was enriched in energy-related functions (Figures S3F–S3J). To understand possible nutrient fluxes, we concentrated our analysis on annotations of carbohydrate-active enzymes (CAZymes) (Figures S3A–S3D and S3K) and found that overall ES-Bact genomes were enriched in CAZymes, particularly those associated with degradation of plant- and animal-based carbohydrates. It could suggest a key role of ES-Bact in the primary degradation of host-ingested nutrients. ES-Bifi genomes were enriched in CAZymes that targeted alpha-glucans, which was consistent with previous analyses on the CAZyme content of *Bifidobacterium*.[22] By contrast, ES-Esch had a larger number of CAZymes that targeted the degradation of bacterial cell wall carbohydrates (Figure S3K), indicating a possible role in predation on other microbes.

Next, we investigated if ESs are redundant or rather complementary in their functions, hypothesizing that the first could lead to competitive exclusion and the latter to higher coincidences of the respective ES. For this, gut communities were *in silico* simulated using genome-scale metabolic networks (GSMNs) of our metagenomic reconstructed genomes (MGS).[1,21] On average, per-species metabolic profiles were similar to the metabolic profile of their associated ES (permutational multivariate analysis of variance, perMANOVA $F = 5.81$, $p < 0.001$, Figure 3B), indicating that the metabolic roles exhibited by each ES might be conserved. Although this could be due to the taxonomic relatedness of bacteria within each ES, it seems to be partly decoupled from taxonomy. For example, the predicted metabolic profiles of Proteobacteria associated with ES-Bifi are more similar to other ES-Bifi bacteria (perMANOVA F values ranging from 11.34 to 16.51), than to Proteobacteria associated with ES-Esch (F value 23.62); this was also observed, to a lesser extent, for Bacteroidota (see Table S3 in Data S1). We interpret this as each ES having a specific metabolic niche that is collectively enabled by its members.

To understand whether these metabolic niches overlapped or complemented each other, we again simulated *in silico* gut communities but with all five ESs present. Metabolic potential is still clustered by ES association to its constituent species, but the metabolic potential became more similar between ES (Figure 3C; see Table S3 in Data S1). The complementarity between ESs was generally the highest for ES-Bact and, when in combination with other ES, enabled the greatest increase in metabolic diversity (Figure 3D); this may explain why ES-Bact was frequently observed in combination with other ESs (Figures 1 and S4A–S4D; see Table S2 in Data S1). By contrast, ES-Esch, which was most commonly found as a single-dominant ES (Figure 1C), benefitted the least from the presence of other ESs in our simulations (Figure 3D). As all five ESs were rarely observed together in a single sample (Figure 1C), we also simulated the metabolic gain with stepwise increases in the number of ESs (from 1 to 5) per microbiome. Metabolic potential increased with increasing numbers of ESs but became saturated at three ESs with little further gain thereafter (Figure S3C). This observation, in conjunction with increased competition for metabolic niches with more ES, could explain why combinations of more than three ESs were rarely observed (Figure 1C).

Overall, although species within an ES exhibit a similar metabolism, the five ESs have diverse and often complementary metabolic potentials. Metabolic interactions between ESs strengthen the core metabolic functions shared by all ESs, permitting functional redundancy in metagenomes comprised of multiple ESs, likely diversifying potential responses of an ecosystem and therefore increasing resilience.

Enterosignatures capture intra-individual short- and long-term changes in the fecal microbiome

To better understand the establishment, ecological roles, and interactions between ESs in the human host, we investigated temporal ES dynamics in the GMR cohort: (1) cross-sectionally from infants to elders and (2) longitudinally using time series of n = 1,239 individuals. Consistent with our understanding of the developing human microbiome,[23] ES-Esch dominated preterm and early infant samples ($p < 2.2e{-}16$, Figures 4A and S4L) and was typically superseded by ES-Bifi in infants $\leq$ 1 year of age. From 1 year and older, and likely coinciding with weaning, adult ESs (ES-Firm, ES-Bact, and ES-Prev) were observed more frequently. It is noteworthy that ES-Bact often not only occurred in samples from infants younger than 6 months old but also decreased in prevalence in elder hosts.

The number of ESs present in metagenomes correlated strongly with genus- and species-level diversities (Spearman $\rho = 0.556$ and $\rho = 0.535$, respectively, $p < 2.2e{-}16$). As with bacterial diversity,[24] ES diversity also increased with age ($\rho = 0.374$, $p < 2.2e{-}16$): the frequency of samples colonized by a single ES decreased steadily from infants to adults to elders (26%, 12%, and 4% of the samples, respectively) and the community became more complex; two ESs was the most common state across ages (48%, 45%, and 61%, respectively, Figure 4A). In adults, particularly, we commonly observed three ESs (39%). Samples from adults were usually dominated by ES-Firm, ES-Bact, and ES-Prev (termed adult ES), although often in combination with the remaining two ESs (44% of cases). The most interconnected ES was ES-Firm, rarely occurring alone in a sample but most frequently being the dominant ES when in combination with other ESs (1.2% and 45% of adult samples, respectively). Faeces dominated by ES-Firm alone are likely unusual gut ecosystems, as 5 of the 24 ES-Firm monodominated adult samples were collected following a strong antibiotic intervention (see following section). This suggests that ES-Firm has a strong ecological and/or functional complementarity with other ESs, particularly ES-Bact (present in 80% of all ES-Firm combinations), and is likely limited in forming a stable community when occurring on its own.

To further elucidate interactions among ESs, we explored short-term intra-individual ES transitions using longitudinal time series samples from 1,239 individuals, usually covering several weeks or months (see Table S4 in Data S1). Tracking changes in primary ESs can inform on temporal attractors, i.e., ES states that the system tends to coalesce into with time. ES-Bact was an attractor in samples from all ages (Figure 4C) but especially in those from infants. In later life, a steady state between ES-Bact and ES-Firm dominance seemed to establish. This pattern might be explained by the persistence of ESs in longitudinal samples: although ES-Bact and ES-Prev as primary

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

CellPress
OPEN ACCESS

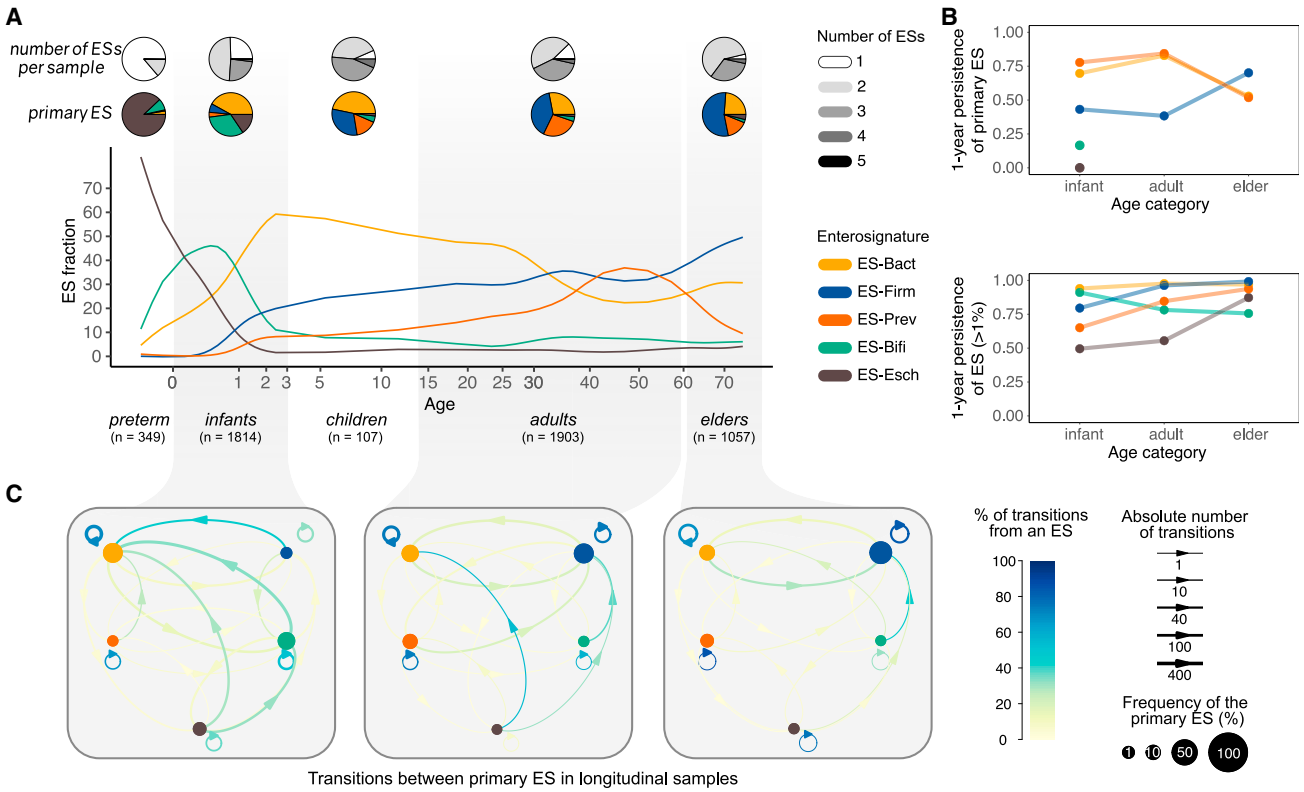


Figure 4. Long-term and short-term dynamics of ES in gut metagenomes

(A) Long-term evolution of ES relative abundances. Gray pie plots depict the frequency of the number of ESs in infant metagenomes (n = 1,811), children's metagenomes (n = 107), adult metagenomes (n = 1,904), and elder metagenomes (n = 1,057). Colored pie plots represent the frequency of each ES as primary ES in metagenomes by age groups. Samples from preterm individuals are depicted with negative values.

(B) Annual persistence of primary ES as well as ES presence (relative abundance > 1%) in faecal samples separated by host age category.

(C) Transitions between primary ES in longitudinal samples of infants (n = 1,733 samples belonging to 427 individuals), adults (n = 1,265 samples, 355 individuals), and elders (n = 964 samples, 320 individuals). Size of nodes depicts the frequency of the ES, edge width depicts the absolute number of transitions observed in the dataset, and edge color illustrates the percentage of transitions from the origin ES.

Abbreviations are as follows: ESs, enterosignatures. See also Figure S4.

ESs have high persistence in infants and adults, ES-Firm is only persistent as the most dominant ES in elders (estimates based on survival statistics, Figure 4B; see Table S4 in Data S1). Apart from ES-Bifi, ES persistence increased with age; particularly persistence of ES-Esch increased drastically in samples from elders. One hypothesis for this observation is that onset of immunosenescence[25] could lead to a reduced gut ecosystem control by the host.

Antibiotic treatments can have devastating effects on the gut microbiome[26]; we hypothesized that during recovery, typical ecosystem succession strategies can be discovered. Our GMR data suggest that ES-Bact is important in the recovery of disturbed gut ecosystems because its temporal attractor status was even stronger in antibiotic-treated individuals (Figures S4A–S4H). Following antibiotic treatments, ES-Bact relative abundance increased by 15% and 17% in samples from infants (excluding preterms) and adults (n = 1,811 and n = 1,904, respectively, $p < 1e{-}15$, Figure S4I), also observed in the BMIS cohort (Figure 2D). This is at the expense of ES-Esch and ES-Bifi in samples from infants, which decreased by 9% and 7%, respectively (Figures S4K and S4L), and ES-Firm and ES-Prev in samples from adults, which decreased by 5% and 9%, respectively, after antibiotic treatments (Figures S4J and S4M).

Collectively, these results demonstrate that both short- and long-term changes in the gut microbiome can be effectively captured using ESs at an easily interpretable level.

Atypical enterosignature composition as a marker of microbial perturbations

Not all gut microbiomes are equal, but we hypothesized that some compositions could deviate more strongly from the expected compositions the model was trained on, such as dysbiotic gut microbiomes. Mathematically, the 5-ES model was able to accurately describe the vast majority of GMR fecal microbiomes, measured by the cosine similarity between the 5-ES model and genus abundances (mean 0.80 ± 0.22, median 0.89, Figure 2B; see Table S1 in Data S1) that we refer to as ES model fit. However, among the 5,230 samples evaluated, there were 394 with an ES model fit < 0.40 (see STAR Methods), which we termed "*ES-atypical*" samples. Indeed, these samples had a lower evenness (Kruskal-Wallis: $chi^2 = 32.593$, $p = 1.14e{-}08$) and an over-representation of potential pathobionts,

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

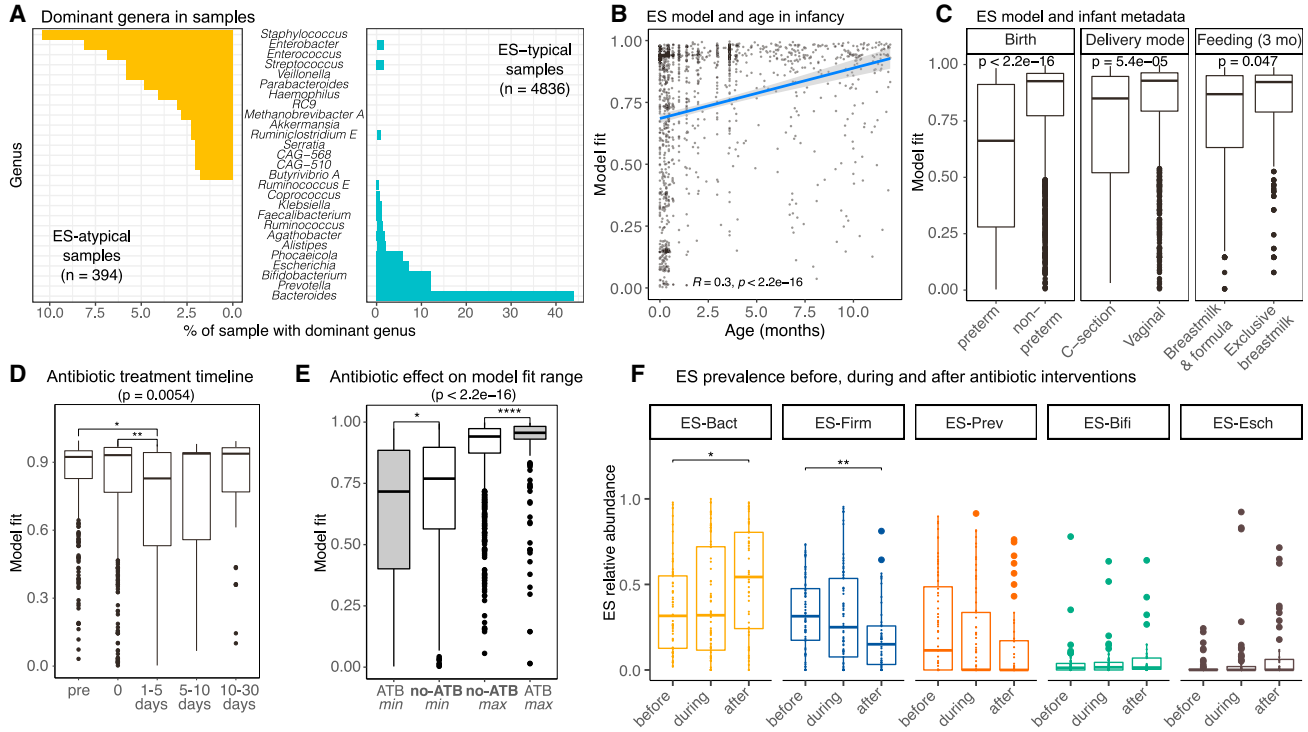


Figure 5. Enterosignatures model fit defines atypical metagenomes

(A) Dominant genera and their frequency in ES-atypical samples differ from those in ES-typical samples.

(B and C) (B) Lower model fit scores were observed in younger infants, (C) especially in preterm infant samples, C-section infant samples (excluding preterm), and infant samples (excluding preterm and C-section born infants) fed with *formula* and breastmilk.

(D–F) Impact of antibiotic treatments on the adequation of the ES model to samples. (D) Changes in ES model fit scores in antibiotic-treated individuals depended on the chronology of sampling with respect to the treatment i.e., "pre": samples preceding antibiotic treatment or obtained more than 30 days after. (E) During time series of antibiotic interventions, ES model fitting scores are either lowest or highest in antibiotic-treated individuals compared with non-treated individuals. (F) Changes in ES relative abundances in antibiotic-treated longitudinally sampled individuals, before, during, and after antibiotic treatment. (D–F) considered samples from individuals over 1 year old. P values were calculated with a Kruskal-Wallis test.

Abbreviations are as follows: ESs, enterosignatures; ATBs, antibiotics. See also Figure S5.

such as *Staphylococcus epidermidis*,[27] *Haemophilus parainfluenzae*,[28] or *Enterococcus spp*[29] (Figure 5A; supplemental information; see Table S5 in Data S1). Low evenness is indicative of monodominance (a single taxon at >60% relative abundance[26]), as would be expected in microbiomes overgrown by an invading pathogen, such as observed in *Salmonella*, *C. difficile*, or *E. coli* intestinal infections.[30] In 21.8% of ES-atypical samples, a species was monodominant, although this was only observed in 3.3% of "*ES-typical*" samples with good ES model fit.

ES-atypical samples include both low- and high-diversity metagenomes indicative that a second-order correlation might better describe the relation of ES model fit to alpha diversity (Figures S2D and S5A); thus, ES model fit appears to be complementary rather than redundant to alpha diversity measures. We found several potentially detrimental host factors that could be associated with lowered ES model fit scores in their fecal microbiomes. This was observed in metagenomes from infants born preterm (Kruskal-Wallis: $chi^2 = 205.53$, $p < 2.2e{-}16$), infants born via C-section (excluding preterm infants, $chi^2 = 16.29$, $p = 5.4e{-}5$), and infants with a formula-supplemented diet (excluding preterm and C-section born, $chi^2 = 3.95$, $p = 0.047$) (Figures 5B and 5C), in line with previous reports of perturbed gut microbiomes in such cases.[23] Maturation of the fecal microbiome translates into a gut microbial composition that fits better to known compositions, as infant age correlated positively to model fit during the first year of life ($\rho = 0.3$, $p < 2.2e{-}16$, Figure 5B).

In antibiotic-treated individuals, we would expect to observe partially dysbiotic gut microbiomes; indeed, lower ES model fits were observed 1–5 days after treatment (Figure 5D). In the time series data for these individuals, we found that the range of ES model fits was greater in individuals treated with antibiotics, being on average both higher and lower during the time series, than in reference samples (Figure 5E). We interpret this as the microbiome first becoming ES-atypical due to antibiotic exposure and then ES-typical in the recovery phase due to reduced species diversity and only the gut core composition being present. We further investigated antibiotic treatments in two sub-cohorts, one specifically treated with antibiotics to provoke dysbiosis[31] and a case study that followed one individual for 3 years with two separate antibiotic treatments.[26] In both datasets, microbiomes were strongly atypical after antibiotic treatment (mean ES model fit values after the first treatment 0.21, 0.03, respectively, Figures S5B and S5F). ES composition remained skewed even after recovery, with ES-Bact usually dominating microbiomes for >180 days (Figure S5C). This was a general

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

CellPress
OPEN ACCESS

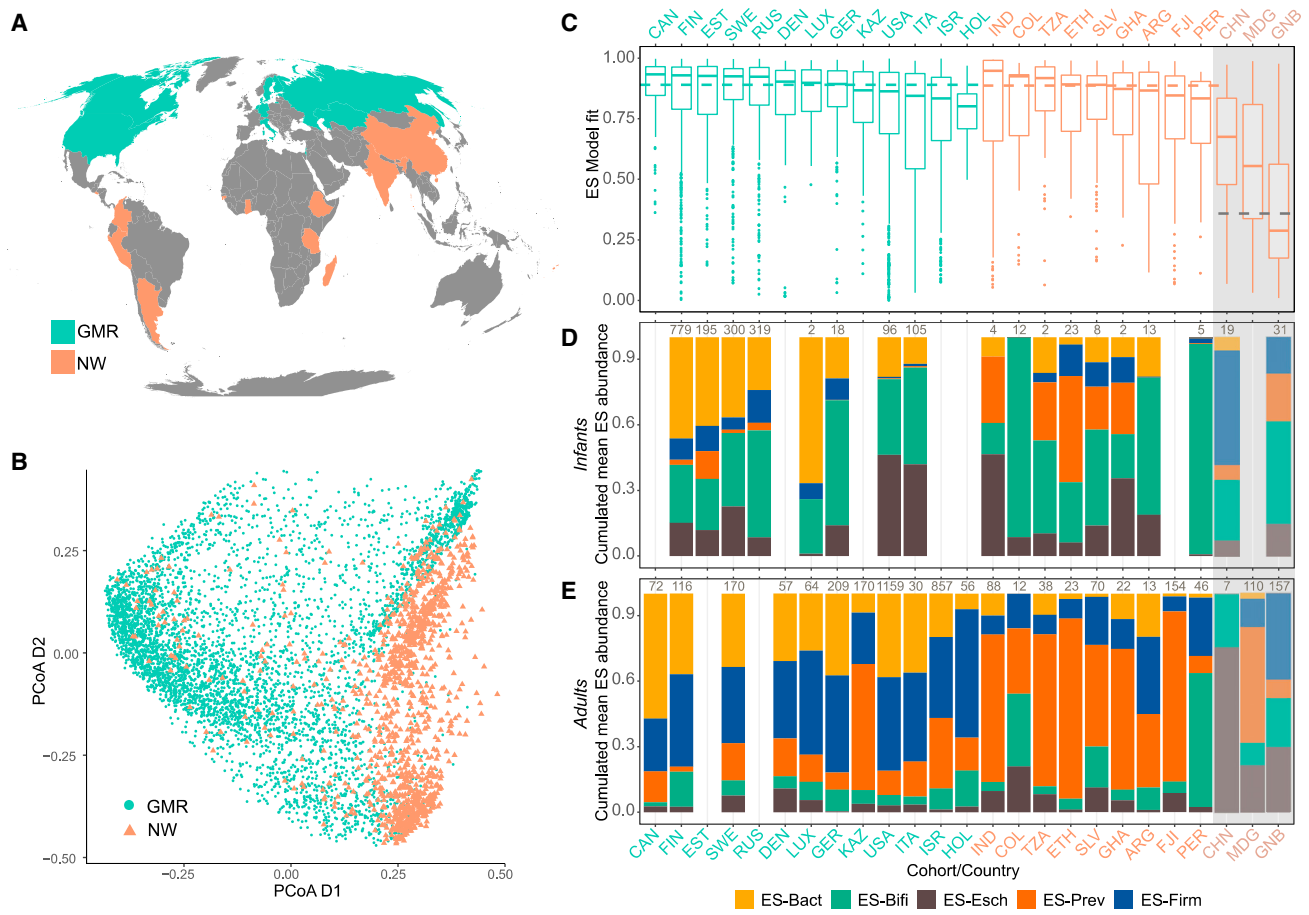


Figure 6. Applying the GMR ES to metagenomic samples associated with a non-western lifestyle validates the generalizability of the model

(A) Country of origin of the 5,230 GMR and 1,152 non-western (NW) samples.

(B) PCoA (Bray-Curtis's dissimilarity distance) of the genus-level abundance matrices of GMR and NW samples.

(C) GMR's 5-ES model fit to samples of GMR and NW cohorts, by country. Fecal samples in the CHN, MDG, and GNB cohorts were conserved in ethanol, which likely constitutes a technical bias and was detected as atypical microbiome compositions. Green, black, and orange dotted horizontal lines represent the median of model fitting score of the GMR-trained 5-ES in GMR, NW excluding CHN, MDG, GNB, and CHN, MDG, GNB cohorts, respectively.

(D and E) Average relative abundance of the five ES in infant (D) and adult (E) samples of each cohort of GMR and NW metagenomes. Numbers above the stacked bars depict the number of faecal metagenomes in each group.

Abbreviations are as follows: GMR, gut microbiome reference; NW, non-western; CAN, Canada; FIN, Finland; EST, Estonia; SWE, Sweden; RUS, Russia; DEN,= Denmark; LUX, Luxembourg; GER, Germany; KAZ, Kazakhstan; ITA, Italia; ISR, Israel; HOL, Holland; IND, India; COL, Colombia; TZA, Tanzania; ETH, Ethiopia; SLV, El Salvador; GHA, Ghana; ARG, Argentina; FJI, Fiji; PER, Peru; CHN, China; MDG, Madagascar; GNB, Guinea-Bissau. See also Figure S6.

trend for all adult individuals treated with antibiotics in our dataset; ES-Bact increased in dominance after antibiotic treatment, at the expense of ES-Firm and ES-Prev (Figure 5F). These observations are consistent with correlations observed on the BMIS dataset (Figure 2D), with ES-Bact becoming a strong temporal attractor in antibiotic-treated individuals (Figures S4E and S4F). Consistent with our previous findings, ES-Bact seems to have a core role in the western gut microbiome—recurrent and stable after perturbation—whereas ES-Firm and ES-Prev act as accessory ESs, and their relative abundances are likely to decrease in disrupted ecosystems.

Enterosignatures are generalizable to non-western cohorts

To further test the generalizability of our 5-ES model to samples diverse in age and geography, we evaluated 1,152 fecal metagenomes from 12 non-western (NW) countries that were not represented in the original GMR cohort, representing infants, children, and adults (n = 119, 278, and 740, respectively)[32,33] (Figure 6A). The NW samples were distinct in their genus-level composition from GMR samples, as expected[33] (Figure 6B). Applying the GMR-trained 5-ES model showed upon initial inspection an average ES model fit significantly lower in NW (0.64) compared with GMR (0.8) metagenomes (Kruskal-Wallis $p < 1e{-}15$). Investigating ES model fit within countries represented in the GMR and NW cohorts, we discovered that samples from China,[32] Madagascar,[34] and Guinea-Bissau[32,35] had substantially reduced values. In only these cohorts, faeces were collected in ethanol with sometimes incomplete cold-chains, a collection method known to bias community compositions.[36] Furthermore, 47.3% of the individuals from the Guinea-Bissau cohort were diagnosed with helminth infections,[35] leading to microbiomes that could be dysbiotic

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

and expected to have lowered ES model fits. When excluding these samples from further analysis, the average ES model fit for the remaining NW metagenomes—not collected in ethanol—was similar to GMR ES model fit (mean 0.78, 0.8, respectively, $p = 0.073$) (Figure 6C), demonstrating the generalizability of the 5-ES model to human samples independent of age or country.

ES-Prev was the most dominant ES in the reduced NW dataset (present in 81.3% of samples, excluding China, Madagascar, and Guinea-Bissau samples), being primary ES in 66.4% of the NW samples, having a prominent role reminiscent of ES-Bact's prominence in the GMR cohort (Figure S6A). ES-Prev established in NW infants much earlier than in GMR infants (Figures 4A, 6D, and S6B), appearing dominant as early as from 6 months of age in NW samples. In line with this, ES-Prev dominated samples often harbor ES-Firm as the second most abundant ES (49.5%) but rarely ES-Bact (6.1%). Of note is that in both GMR, BMIS, and NW cohorts, ES-Bact and ES-Prev are strongly anticorrelated ($\rho = -0.63, -0.77$, $p < 2.2e{-}16$ and $\rho = -0.11$, $p = 0.01$, respectively, Table S6 in Data S1), indicating competitive exclusion between these two bacterial guilds, plausible due to their strongly redundant functional profiles that could indicate similar ecological roles (Figure 4). ES-Prev might colonize a similar ecological niche to ES-Bact but reflect a NW lifestyle: for example, the Kazakhstan samples in the GMR cohort, arguably NW, were dominated by ES-Prev (Figures 6D and 6E).

DISCUSSION

Limitations of the study

We specifically chose an algorithm capable of decomposing complex microbial communities into assemblies of underlying microbial guilds that explained most of the gut microbial variance. Two classes of algorithms for this were previously evaluated on microbiome data: the linear algebra approach of NMF and probabilistic approaches such as generalization of DMM into LDA models. These works challenged existing methodologies[37] and demonstrated that assemblages of microbes could be captured from metagenomes,[10,13] even recovering bacterial assemblages reminiscent of ETs using LDA.[12] Here, we used NMF because the correct scaling of observed proportions to read counts in the multinomial noise term is not obvious for metagenome data, and the simpler model parametrization of NMF potentially reduces overfitting. Based on our 5-ES model, we propose a unified description of the human gut microbiome that has stronger associations with metadata and explains more variance of the original composition, than ET models. ESs rely on genus-level compositions, a taxonomic level chosen as a trade-off between the generalizability of the conserved patterns, the relevance of the taxa for functional inference, and the interpretability and number of recovered signatures.

The 5-ES model was trained on the GMR dataset consisting mostly of western individuals (74.2% western countries, 25.8% Israel, Russia, and Kazakhstan), potentially limiting generalizability of five ESs. We undertook two validation cohort studies relying on the BMIS (European samples) and NW (no-western) cohorts to evaluate generalizability: on both datasets, the GMR-trained 5-ES model generalized well. However, it is conceivable that we are missing ESs associated with specific cohorts or conditions and therefore not fully representing the spectrum of human gut ESs. Therefore, we provide the mathematical tools to identify gut metagenomes not fitting to the 5-ES model, using ES model fit scores. The pre-trained model and scripts for either applying or *de novo* calculating ES are available at https://enterosignatures.quadram.ac.uk. We explicitly want to acknowledge that additional bacterial guilds might be an important part of the healthy or diseased gut ecosystem, which can be cataloged and curated on our provided website as a community effort.

An accessible perspective on the gut microbial ecosystem

The proposed ESs augment our understanding of the gut microbial ecology, as ES profiles provide a granular classification of globally observed gut bacterial guilds. In fecal samples from most adults and many infants, multiple ESs were coexisting; therefore, using a single community type to describe a GIT sample is probably incompletely capturing the underlying gut ecosystem. Despite ESs being highly persistent in individuals, the relative ES composition rapidly changed following perturbations, indicating the model being sensitive to external and internal forces. However, the often-observed switches in dominant ES are not necessarily indicative of ecological niche changes when the overall mixture of ESs remains stable. Rather, we propose to monitor the dynamic interplay between ES combinations in future applications. For example, ES-Prev and ES-Bact are strongly co-excluding signatures, despite—or because of—highly similar ecological niches: they share metabolic functions, occur at a similar frequency in fecal samples of healthy adults, either ES often co-occurs with ES-Firm, but can also dominate a microbiome solely.

Of note was that these two co-excluding signatures dominate either NW (ES-Prev) or western (ES-Bact) gut microbiomes. We do not know why this is exactly, but features associated with a western lifestyle, such as ultra-processed foods, hygiene, and physical activity levels, are likely contributing to this.[9] However, we note that after antibiotic interventions, ES-Bact was often increasing in prevalence, at the expense of both ES-Firm and ES-Prev. Therefore, antibiotic usage, more widespread in western societies,[38] could be a contributing factor to the puzzling ES-Bact dominance in western samples.

Both ES-Bact and ES-Prev appear to provide a core functionality to the healthy gut microbiome. Although the healthy microbiome is, in both western and NW datasets, likely established by ES-Bifi, our results highlight that the latter is being replaced early in life, by either ES-Bact (western) or ES-Prev (NW). The establishment of these adult microbiomes in infants is likely supported by vaginal birth as we found a correlation between ES-Bact and vaginal birth mode (reported also in Song et al.[20]) for western samples. ES-Firm is most likely successional,[26] but in contrast to ES-Prev or ES-Bact, almost never occurring on its own. Therefore, we hypothesize that ES-Firm provides complementary functionality and relies on biotic interactions provided by either ES-Bact or ES-Prev. Although ES-Bifi, ES-Bact, ES-Prev, and ES-Firm are part of a healthy human gut microbiome, ES-Esch is often dominant in samples from preterm infants and adults undergoing drastic interventions, frequently being a single-dominant ES. It seems plausible to classify ES-Esch as an unhealthy, potentially pathogenic ES.

Based on these temporal observations, we propose the following successional model of the human gut microbiome: the

ES-Esch bacterial guild is a temporary colonizer in mostly dysbiotic infant microbiomes, succeeded by ES-Bifi and then by ES-Bact, or ES-Prev, at weaning. ES-Bact/ES-Prev forms the stable core of the healthy human microbiome, and with time, ES-Firm can complement ecosystem functionality. In the long term, ES-Prev replaces ES-Bact even in western samples, by superseding its ecological niche. The relative abundances of both ES-Firm and ES-Prev are more likely to decrease in disrupted ecosystems and the recurrence and stability of ES-Bact after perturbation is noteworthy; *Bacteroides* has already been identified as a dominant profile associated with disease,[39] *Bacteroides*/*Prevotella* were described as tenacious (highly persistent) taxa,[1] and the low diversity, low cell count associated Bact2 ET was prevalent in hosts with diseases[17,40]; ES-Bact could be the remaining core of a disrupted ecosystem in the western gut microbiome.

A unified approach to defining atypical fecal microbiomes requiring further investigation

Defining typical compositions of the healthy human gut microbiome can help to classify and quantify healthy states; conversely, it can also be used to identify microbiomes deviating from typical compositions observed in healthy hosts, thereby defining microbial dysbiosis. Detecting dysbiosis usually requires a suitably large reference dataset to allow for outlier detection in ordinations[41] or using abundances of preselected bacterial species[42] to train machine learning models.[43] Here, we demonstrated how the 5-ES model can be used to detect "atypical" compositions in fecal samples, outside of the expected ES signature spectrum. This classification of potentially dysbiotic samples relies on the "absence of evidence for normal states" and is thus fundamentally different from models that rely on the "presence of dysbiotic evidence." Because the 5-ES model generalizes well to diverse gut metagenomes and is applicable to single samples, the detection of atypical samples allows for further investigation of a wide spectrum of potentially detrimental host states.

Fecal metagenomes deviating from the 5-ES model are not necessarily dysbiotic, but rather atypical with respect to the training dataset. However, ES atypical metagenomes were often associated with detrimental conditions in the host: ES model fit decreased after antibiotic interventions, classical examples of dysbiotic microbiomes,[31] and ES atypical microbiomes were often dominated by potential pathobiont species. In addition to ES composition correlating with health-associated metadata, we demonstrated how ESs can identify detrimental changes in the gut microbiome composition and provide hypotheses about their significance. In gut microbiomes from the general population, ESs could therefore be a valuable tool to detect and assess pathological conditions, discern the effect of treatments, and monitor disease progression with greater sensitivity and interpretability by researchers and medical professionals alike.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2023.05.024.

ACKNOWLEDGMENTS

R.A., E.O., and F.H. were supported by European Research Council H2020 StG (erc-stg-948219, EPYC). C.F., R.A., E.O., and F.H. were supported by the Biotechnology and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Food Microbiome and Health BB/X011054/1 and its constituent project BBS/E/F/000PR13631, Gut Microbes and Health BB/r012490/1 and its constituent project BBS/e/F/000Pr10355, Core Capability Grant BB/CCG1720/1 and the work delivered via the Scientific Computing group, as well as support for the physical HPC infrastructure and data centre delivered via the NBI Computing infrastructure for Science (CiS) group. J.F. was supported by the UKRI Biotechnology and Biological Sciences Research Council Norwich Research Park Biosciences Doctoral Training Partnership BB/T008717/1. Some experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP, and Conseil Régional d'Aquitaine (see https://www.plafrim.fr). C.Q. was funded through the MRC Methodology grant "Strain resolved metagenomics for medical microbiology" MR/S037195/1. S.M.W. was supported by an SNSF Early Postdoc.Mobility Fellowship (P2ELP3_155365), an EMBO Long-Term Fellowship (ALTF 755-2014), and grants from the Research Council of Norway (187615), University of Oslo, and South-Eastern Norway Regional Health Authority. F.H. would like to thank Peer Bork and Jeroen Raes for insightful discussions on enterotypes. The authors thank Judith Pell for proofreading the article and all members of the Hildebrand group for valuable feedback, discussion, and support. We further want to thank the three anonymous reviewers for helping us in making the analysis more robust and impactful.

AUTHOR CONTRIBUTIONS

Conceptualization: F.H., S.M.W., and C.F.; methodology: C.F., S.M.W., and C.Q.; software: C.F., F.H., J.F., R.A., and S.M.W.; validation: C.F.; formal analysis: C.F., R.A., and F.H.; investigation: C.F., R.A., F.H., and S.M.W.; data curation: C.F., F.H., E.Ö., R.A., and C.G.N.; writing – original draft: C.F. and F.H.; writing – review & editing: F.H., C.F., R.A., E.Ö., S.M.W., and C.Q. with contributions from all authors; visualization: C.F., F.H., S.M.W., and R.A.; supervision, project administration, and funding acquisition: F.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: January 26, 2023
Revised: April 3, 2023
Accepted: May 23, 2023
Published: June 19, 2023

REFERENCES

1. Hildebrand, F., Gossmann, T.I., Frioux, C., Özkurt, E., Myers, P.N., Ferretti, P., Kuhn, M., Bahram, M., Nielsen, H.B., and Bork, P. (2021). Dispersal strategies shape persistence and evolution of human gut bacteria. Cell Host Microbe *29*. 1167.e9–1176.e9. https://doi.org/10.1016/j.chom.2021.05.008.
2. Clemente, J.C., Ursell, L.K., Parfrey, L.W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. Cell *148*, 1258–1270. https://doi.org/10.1016/j.cell.2012.01.035.
3. Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., Le Roy, C.I., Raygoza Garay, J.A., Finnicum, C.T., Liu, X., et al. (2021). Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat. Genet. *53*, 156–165. https://doi.org/10.1038/s41588-020-00763-1.
4. Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: Human gut microbes associated with obesity. Nature *444*, 1022–1023. https://doi.org/10.1038/4441022a.
5. Raman, A.S., Gehrig, J.L., Venkatesh, S., Chang, H.W., Hibberd, M.C., Subramanian, S., Kang, G., Bessong, P.O., Lima, A.A.M., Kosek, M.N., et al. (2019). A sparse covarying unit that describes healthy and impaired human gut microbiota development. Science *365*. eaau4735. https://doi.org/10.1126/science.aau4735.
6. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D.L., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., et al. (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180. https://doi.org/10.1038/nature09944.
7. Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One *7*. e30126. https://doi.org/10.1371/journal.pone.0030126.
8. Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science *334*, 105–108. https://doi.org/10.1126/science.1208344.
9. Costea, P.I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M.J., Bushman, F.D., Vos, W.M. de, Ehrlich, S.D., Fraser, C.M., Hattori, M., et al. (2018). Enterotypes in the landscape of gut microbial community composition. Nat. Microbiol. *3*, 8–16. https://doi.org/10.1038/s41564-017-0072-8.
10. Breuninger, T.A., Wawro, N., Breuninger, J., Reitmeier, S., Clavel, T., Six-Merker, J., Pestoni, G., Rohrmann, S., Rathmann, W., Peters, A., et al. (2021). Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. Microbiome *9*, 61. https://doi.org/10.1186/s40168-020-00969-9.
11. Cai, Y., Gu, H., and Kenney, T. (2017). Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. Microbiome *5*, 110. https://doi.org/10.1186/s40168-017-0323-1.
12. Hosoda, S., Nishijima, S., Fukunaga, T., Hattori, M., and Hamada, M. (2020). Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. Microbiome *8*, 95. https://doi.org/10.1186/s40168-020-00864-3.
13. Yan, J., Chuai, G., Qi, T., Shao, F., Zhou, C., Zhu, C., Yang, J., Yu, Y., Shi, C., Kang, N., et al. (2017). MetaTopics: an integration tool to analyze microbial community profile by topic model. BMC Genomics *18* (*Suppl 1*), 962. https://doi.org/10.1186/s12864-016-3257-2.
14. Raguideau, S., Plancade, S., Pons, N., Leclerc, M., and Laroche, B. (2016). Inferring aggregated functional traits from metagenomic data using constrained non-negative matrix factorization: application to fiber degradation in the human gut microbiota. PLoS Comput. Biol. *12*. e1005252. https://doi.org/10.1371/journal.pcbi.1005252.
15. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791. https://doi.org/10.1038/44565.
16. Guner, Y.S., Malhotra, A., Ford, H.R., Stein, J.E., and Kelly, L.K. (2009). Association of Escherichia coli O157:H7 with necrotizing enterocolitis in a full-term infant. Pediatr. Surg. Int. *25*, 459–463. https://doi.org/10.1007/s00383-009-2365-3.
17. Vieira-Silva, S., Falony, G., Belda, E., Nielsen, T., Aron-Wisnewsky, J., Chakaroun, R., Forslund, S.K., Assmann, K., Valles-Colomer, M., Nguyen, T.T.D., et al. (2020). Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. Nature *581*, 310–315. https://doi.org/10.1038/s41586-020-2269-x.
18. Zhou, J., Wong, M.S., Chen, W.C., Krainer, A.R., Kinney, J.B., and McCandlish, D.M. (2022). Higher-order epistasis and phenotypic prediction. Proc. Natl. Acad. Sci. USA *119*. e2204233119. https://doi.org/10.1073/pnas.2204233119.
19. Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L.D., Darzi, Y., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. Nature *551*, 507–511. https://doi.org/10.1038/nature24460.
20. Song, S.J., Wang, J., Martino, C., Jiang, L., Thompson, W.K., Shenhav, L., McDonald, D., Marotz, C., Harris, P.R., Hernandez, C.D., et al. (2021). Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. Med. *2*. 951.e5–964.e5. https://doi.org/10.1016/j.medj.2021.05.003.
21. Frioux, C., Singh, D., Korcsmaros, T., and Hildebrand, F. (2020). From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Comput. Struct. Biotechnol. J. *18*, 1722–1734. https://doi.org/10.1016/j.csbj.2020.06.028.
22. Møller, M.S., Goh, Y.J., Viborg, A.H., Andersen, J.M., Klaenhammer, T.R., Svensson, B., and Abou Hachem, M.A. (2014). Recent insight in α-glucan metabolism in probiotic bacteria. Biologia *69*, 713–721. https://doi.org/10.2478/s11756-014-0367-7.
23. Shao, Y., Forster, S.C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M.D., Rodger, A., Brocklehurst, P., et al. (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature *574*, 117–121. https://doi.org/10.1038/s41586-019-1560-1.
24. Cuesta-Zuluaga, J. de la, Kelley, S.T., Chen, Y., Escobar, J.S., Mueller, N.T., Ley, R.E., McDonald, D., Huang, S., Swafford, A.D., Knight, R., et al. (2019). Age- and sex-dependent patterns of gut microbial diversity in human adults. mSystems *4*. e00261-19. https://doi.org/10.1128/mSystems.00261-19.
25. Fu, Y.R., Yi, Z.J., Pei, J.L., and Guan, S. (2010). Effects of Bifidobacterium bifidum on adaptive immune senescence in aging mice. Microbiol. Immunol. *54*, 578–583. https://doi.org/10.1111/j.1348-0421.2010.00255.x.
26. Hildebrand, F., Moitinho-Silva, L., Blasche, S., Jahn, M.T., Gossmann, T.I., Huerta-Cepas, J., Hercog, R., Luetge, M., Bahram, M., Pryszlak, A., et al. (2019). Antibiotics-induced monodominance of a novel gut bacterial order. Gut *68*, 1781–1790. https://doi.org/10.1136/gutjnl-2018-317715.
27. Hindieh, P., Yaghi, J., Khoury, A.E., Chokr, A., Atoui, A., Louka, N., and Assaf, J.C. (2022). Lactobacillus rhamnosus and Staphylococcus epidermidis in gut microbiota: in vitro antimicrobial resistance. AMB Express *12*, 128. https://doi.org/10.1186/s13568-022-01468-w.
28. Liu, Q., Jiang, Z., Teng, Q., and Jiang, M. (2022). Pseudomembranous colitis caused by Haemophilus parainfluenzae. Inflamm. Bowel Dis. *28*. e55–e56. https://doi.org/10.1093/ibd/izab293.
29. Dubin, K., and Pamer, E.G. (2014). Enterococci and their interactions with the intestinal microbiome. Microbiol. Spectr. *5*. https://doi.org/10.1128/microbiolspec.BAD-0014-2016.
30. Yurist-Doutsch, S., Arrieta, M.C., Vogt, S.L., and Finlay, B.B. (2014). Gastrointestinal microbiota–mediated control of enteric pathogens. Annu. Rev. Genet. *48*, 361–382. https://doi.org/10.1146/annurev-genet-120213-092421.
31. Palleja, A., Mikkelsen, K.H., Forslund, S.K., Kashani, A., Allin, K.H., Nielsen, T., Hansen, T.H., Liang, S., Feng, Q., Zhang, C., et al. (2018).

Recovery of gut microbiota of healthy adults following antibiotic exposure. Nat. Microbiol. *3*, 1255–1265. https://doi.org/10.1038/s41564-018-0257-9.

32. Valles-Colomer, M., Blanco-Míguez, A., Manghi, P., Asnicar, F., Dubois, L., Golzato, D., Armanini, F., Cumbo, F., Huang, K.D., Manara, S., et al. (2023). The person-to-person transmission landscape of the gut and oral microbiomes. Nature *614*, 125–135. https://doi.org/10.1038/s41586-022-05620-1.

33. Vishnu Prasoodanan, P.K., Sharma, A.K., Mahajan, S., Dhakan, D.B., Maji, A., Scaria, J., and Sharma, V.K. (2021). Western and non-western gut microbiomes reveal new roles of Prevotella in carbohydrate metabolism and mouth–gut axis. npj Biofilms Microbiomes *V*, 7. 77. https://doi.org/10.1038/s41522-021-00248-x.

34. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*. 649.e20–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

35. Farrant, O., Marlais, T., Houghton, J., Goncalves, A., Teixeira da Silva Cassama, E.T.da S., Cabral, M.G., Nakutum, J., Manjuba, C., Rodrigues, A., Mabey, D., et al. (2020). Prevalence, risk factors and health consequences of soil-transmitted helminth infection on the Bijagos Islands, Guinea Bissau: A community-wide cross-sectional study. PLoS Negl. Trop. Dis. *14*. e0008938. https://doi.org/10.1371/journal.pntd.0008938.

36. Carruthers, L.V., Moses, A., Adriko, M., Faust, C.L., Tukahebwa, E.M., Hall, L.J., Ranford-Cartwright, L.C., and Lamberton, P.H.L. (2019). The impact of storage conditions on human stool 16S rRNA microbiome composition and diversity. PeerJ *7*. e8133. https://doi.org/10.7717/peerj.8133.

37. Sankaran, K., and Holmes, S.P. (2019). Latent variable modeling for the microbiome. Biostatistics *20*, 599–614. https://doi.org/10.1093/biostatistics/kxy018.

38. Lee, K., Raguideau, S., Sirén, K., Asnicar, F., Cumbo, F., Hildebrand, F., Segata, N., Cha, C.J., and Quince, C. (2023). Population-level impacts of antibiotic usage on the human gut microbiome. Nat. Commun. *14*, 1191. https://doi.org/10.1038/s41467-023-36633-7.

39. Hasegawa, K., Stewart, C.J., Mansbach, J.M., Linnemann, R.W., Ajami, N.J., Petrosino, J.F., and Camargo, C.A. (2017). Sphingolipid metabolism potential in fecal microbiome and bronchiolitis in infants: a case–control study. BMC Res. Notes *10*, 325. https://doi.org/10.1186/s13104-017-2659-9.

40. Fromentin, S., Forslund, S.K., Chechi, K., Aron-Wisnewsky, J., Chakaroun, R., Nielsen, T., Tremaroli, V., Ji, B., Prifti, E., Myridakis, A., et al. (2022). Microbiome and metabolome features of the cardiometabolic disease spectrum. Nat. Med. *28*, 303–314. https://doi.org/10.1038/s41591-022-01688-4.

41. Halfvarson, J., Brislawn, C.J., Lamendella, R., Vázquez-Baeza, Y., Walters, W.A., Bramer, L.M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel Disease. Nat. Microbiol. *2*, 17004. https://doi.org/10.1038/nmicrobiol.2017.4.

42. Huh, J.W., and Roh, T.Y. (2020). Opportunistic detection of Fusobacterium nucleatum as a marker for the early gut microbial dysbiosis. BMC Microbiol. *20*, 208. https://doi.org/10.1186/s12866-020-01887-4.

43. Casén, C., Vebø, H.C., Sekelja, M., Hegge, F.T., Karlsson, M.K., Ciemniejewska, E., Dzankovic, S., Frøyland, C., Nestestog, R., Engstrand, L., et al. (2015). Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with IBS or IBD. Aliment. Pharmacol. Ther. *42*, 71–83. https://doi.org/10.1111/apt.13236.

44. Kushugulova, A., Forslund, S.K., Costea, P.I., Kozhakhmetov, S., Khassenbekova, Z., Urazova, M., Nurgozhin, T., Zhumadilov, Z., Benberin, V., Driessen, M., et al. (2018). Metagenomic analysis of gut microbial communities from a Central Asian population. BMJ Open *8*. e021682. https://doi.org/10.1136/bmjopen-2018-021682.

45. Willmann, M., El-Hadidi, M., Huson, D.H., Schütz, M., Weidenmaier, C., Autenrieth, I.B., and Peter, S. (2015). Antibiotic selection pressure determination through sequence-based metagenomics. Antimicrob. Agents Chemother. *59*, 7335–7345. https://doi.org/10.1128/AAC.01504-15.

46. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., 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.

47. Chu, D.M., Ma, J., Prince, A.L., Antony, K.M., Seferovic, M.D., and Aagaard, K.M. (2017). Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. Nat. Med. *23*, 314–326. https://doi.org/10.1038/nm.4272.

48. Asnicar, F., Manara, S., Zolfo, M., Truong, D.T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A., et al. (2017). Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. mSystems *2*. e00164-16. https://doi.org/10.1128/mSystems.00164-16.

49. Lee, S.T.M., Kahn, S.A., Delmont, T.O., Shaiber, A., Esen, Ö.C., Hubert, N.A., Morrison, H.G., Antonopoulos, D.A., Rubin, D.T., and Eren, A.M. (2017). Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. Microbiome *5*, 50. https://doi.org/10.1186/s40168-017-0270-x.

50. Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., Hogan, A., Beaufort, C. de, et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat. Microbiol. *2*, 16180. https://doi.org/10.1038/nmicrobiol.2016.180.

51. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., and Fulton, R.S. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.

52. Kostic, A.D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward Type 1 diabetes. Cell Host Microbe *17*, 260–273. https://doi.org/10.1016/j.chom.2015.01.001.

53. Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.M., Härkönen, T., Ryhänen, S.J., Franzosa, E.A., Vlamakis, H., Huttenhower, C., Gevers, D., et al. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci. Transl. Med. *8*. 343ra81. https://doi.org/10.1126/scitranslmed.aad0917.

54. Vatanen, T., Kostic, A.D., d'Hennezel, E., Siljander, H., Franzosa, E.A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T.D., Hämäläinen, A.-M., et al. (2016). Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. Cell *165*, 842–853. https://doi.org/10.1016/j.cell.2016.04.007.

55. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. Cell Host Microbe *24*. 133.e5–145.e5. https://doi.org/10.1016/j.chom.2018.06.005.

56. Mehta, R.S., Abu-Ali, G.S., Drew, D.A., Lloyd-Price, J., Subramanian, A., Lochhead, P., Joshi, A.D., Ivey, K.L., Khalili, H., Brown, G.T., 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.

57. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized nutrition by prediction of glycemic responses. Cell *163*, 1079–1094. https://doi.org/10.1016/j.cell.2015.11.001.

58. Bengtsson-Palme, J., Angelin, M., Huss, M., Kjellqvist, S., Kristiansson, E., Palmgren, H., Larsson, D.G.J., and Johansson, A. (2015). The human

gut microbiome as a transporter of antibiotic resistance genes between continents. Antimicrob. Agents Chemother. *59*, 6551–6560. https://doi.org/10.1128/AAC.00933-15.

59. Raymond, F., Ouameur, A.A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.L., Giroux, R., Bérubé, È., et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. *10*, 707–720. https://doi.org/10.1038/ismej.2015.148.
60. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., et al. (2018). Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. Cell Host Microbe *24*. 146.e4–154.e4. https://doi.org/10.1016/j.chom.2018.06.007.
61. Ward, D.V., Scholz, M., Zolfo, M., Taft, D.H., Schibler, K.R., Tett, A., Segata, N., and Morrow, A.L. (2016). Metagenomic sequencing with strain-level resolution implicates uropathogenic E. coli in necrotizing enterocolitis and mortality in preterm infants. Cell Rep. *14*, 2912–2924. https://doi.org/10.1016/j.celrep.2016.03.015.
62. Li, S.S., Zhu, A., Benes, V., Costea, P.I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., Voigt, A.Y., et al. (2016). Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science *352*, 586–589. https://doi.org/10.1126/science.aad8852.
63. Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The Prevotella copri complex comprises four distinct clades underrepresented in westernized populations. Cell Host Microbe *26*. 666.e7–679.e7. https://doi.org/10.1016/j.chom.2019.08.018.
64. Pehrsson, E.C., Tsukayama, P., Patel, S., Mejía-Bautista, M., Sosa-Soto, G., Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W., Bertoli, M.T., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. Nature *533*, 212–216. https://doi.org/10.1038/nature17672.
65. Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature *535*, 435–439. https://doi.org/10.1038/nature18927.
66. R Core Team (2019). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
67. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., et al. (2011). Scikit-learn: machine learning in Python. JMLR *12*, 2825–2830. https://doi.org/10.5555/1953048.2078195.
68. Karp, P.D., Midford, P.E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W.K., Subhraveti, P., Caspi, R., Fulcher, C., et al. (2021). Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief. Bioinform. *22*, 109–126. https://doi.org/10.1093/bib/bbz104.
69. Belcour, A., Frioux, C., Aite, M., Bretaudeau, A., Hildebrand, F., and Siegel, A. (2020). Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species. eLife *9*. e61968. https://doi.org/10.7554/eLife.61968.
70. Frioux, C., Fremy, E., Trottier, C., and Siegel, A. (2018). Scalable and exhaustive screening of metabolic functions carried out by microbial consortia. Bioinformatics *34*. i934–i943. https://doi.org/10.1093/bioinformatics/bty588.
71. Saary, P., Forslund, K., Bork, P., and Hildebrand, F. (2017). RTK: efficient rarefaction analysis of large datasets. Bioinform. Oxf. Engl. *33*, 2594–2595. https://doi.org/10.1093/bioinformatics/btx206.
72. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*. e7359. https://doi.org/10.7717/peerj.7359.
73. Hildebrand, F., Tadeo, R., Voigt, A.Y., Bork, P., and Raes, J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome *2*, 30. https://doi.org/10.1186/2049-2618-2-30.
74. Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
75. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.
76. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
77. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
78. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
79. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
80. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.
81. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.
82. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment, software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.
83. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
84. Nguyen, L.T., Schmidt, H.A., Haeseler, A. von, and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.
85. Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. *44*. W242–W245. https://doi.org/10.1093/nar/gkw290.
86. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. (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.
87. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.
88. Hauswedell, H., Singer, J., and Reinert, K. (2014). Lambda: the local aligner for massive biological data. Bioinformatics *30*. i349–i355. https://doi.org/10.1093/bioinformatics/btu439.
89. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.
90. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.
91. Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., et al. (2020). Vegan: community ecology package. http://CRAN.Rproject.org/package=vegan.

92. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

93. Bahram, M., Netherway, T., Frioux, C., Ferretti, P., Coelho, L.P., Geisen, S., Bork, P., and Hildebrand, F. (2020). Metagenomic assessment of the global distribution of bacteria and fungi. Environ. Microbiol. https://doi.org/10.1111/1462-2920.15314.

94. Forslund, S.K., Chakaroun, R., Zimmermann-Kogadeeva, M., Markó, L., Aron-Wisnewsky, J., N., and Birkner, T. (2021). Data analysis pipeline for investigating drug-host-microbiome relationships in cardiometabolic disease (MetaCardis Cohort) (Zenodo). https://doi.org/10.5281/zenodo.5463864.

95. Özkurt, E., Fritscher, J., Soranzo, N., Ng, D.Y.K., Davey, R.P., Bahram, M., and Hildebrand, F. (2022). LotuS2: an ultrafast and highly accurate tool for amplicon sequencing analysis. Microbiome *10*, 176. https://doi.org/10.1186/s40168-022-01365-1.

96. Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E.L., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. *32*, 822–828. https://doi.org/10.1038/nbt.2939.

97. Mende, D.R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. Nat. Methods *10*, 881–884. https://doi.org/10.1038/nmeth.2575.

98. Owen, A.B., and Perry, P.O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. Ann. Appl. Stat. *3*, 564–594. https://doi.org/10.1214/08-AOAS227.

99. Kanagal, B., and Sindhwani, V. (2010). Rank selection in low-rank matrix approximations: a study of cross-validation for NMFs. Proceedings of the Conference on Advanced Neural Information Process 1.

100. Eng, S.W.M., Aeschlimann, F.A., Veenendaal, M. van, Berard, R.A., Rosenberg, A.M., Morris, Q., and Yeung, R.S.M.; ReACCh-Out Research Consortium (2019). Patterns of joint involvement in juvenile idiopathic arthritis and prediction of disease course: a prospective study with multilayer non-negative matrix factorization. PLoS Med. *16*. e1002750. https://doi.org/10.1371/journal.pmed.1002750.

101. Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics *11*, 367. https://doi.org/10.1186/1471-2105-11-367.

102. Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods – a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164–1167. https://doi.org/10.1093/bioinformatics/btm069.

103. Harris, C.R., Millman, K.J., Walt, S.J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

104. Therneau, T.M., and Grambsch, P.M. (2000). Modeling Survival Data: Extending the Cox Model (Springer).

105. Zabor, E. (2016). Ezfun: Emily C. Zabor's functions. https://www.emilyzabor.com/ezfun/.

106. Mende, D.R., Letunic, I., Huerta-Cepas, J., Li, S.S., Forslund, K., Sunagawa, S., and Bork, P. (2017). proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. Nucleic Acids Res. *45*. D529–D534. https://doi.org/10.1093/nar/gkw989.

107. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2018). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*. D309–D314. https://doi.org/10.1093/nar/gky1085.

108. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

109. Drula, E., Garron, M.L., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res. *50*. D571–D577. https://doi.org/10.1093/nar/gkab1045.

110. Cantarel, B.L., Lombard, V., and Henrissat, B. (2012). Complex carbohydrate utilization by the healthy human microbiome. PLoS One *7*. e28742. https://doi.org/10.1371/journal.pone.0028742.

111. The CAZypedia Consortium (2017). Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. Glycobiology *28*, 3–8. https://doi.org/10.1093/glycob/cwx089.

112. Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., and Karp, P.D. (2020). The MetaCyc database of metabolic pathways and enzymes—a 2019 update. Nucleic Acids Res. *48*. D445–D453. https://doi.org/10.1093/nar/gkz862.

113. Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., Daníelsdóttir, A.D., Krecke, M., Merten, D., Haraldsdóttir, H.S., et al. (2019). The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Res. *47*. D614–D624. https://doi.org/10.1093/nar/gky992.

114. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). Cluster: cluster analysis basics and extensions. https://CRAN.R-project.org/package=cluster.

115. Vieira-Silva, S., Sabino, J., Valles-Colomer, M., Falony, G., Kathagen, G., Caenepeel, C., Cleynen, I., Merwe, S. van der, Vermeire, S., and Raes, J. (2019). Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. Nat. Microbiol. *4*, 1826–1831. https://doi.org/10.1038/s41564-019-0483-9.

116. Martinez-Arbizu, P. (2020). pairwiseAdonis: pairwise multilevel comparison using adonis. R package version 0.4. https://github.com/pmartinezarbizu/pairwiseAdonis.

117. Kassambara, A. (2020). ggpubr: "ggplot2" based publication ready plots. GitHub. https://github.com/kassambara/ggpubr.

118. Soetaert, K. (2020). diagram: functions for visualising simple graphs (networks), plotting flow diagrams. https://rdrr.io/rforge/diagram/.

119. Sievert, C. (2020). Interactive web-based data visualization with R, plotly, and shiny. https://plotly-r.com/.

120. Brunson, J.C. (2020). ggalluvial: layered grammar for alluvial plots. J. Open Source Softw. *5*, 2017. https://doi.org/10.21105/joss.02017.

121. Larsson, J. (2020). eulerr: area-proportional Euler and Venn diagrams with ellipses. https://cran.r-project.org/web/packages/eulerr/eulerr.pdf.

122. Forslund, S.K., Chakaroun, R., Zimmermann-Kogadeeva, M., Markó, L., Aron-Wisnewsky, J., Nielsen, T., Moitinho-Silva, L., Schmidt, T.S.B., Falony, G., Vieira-Silva, S., et al. (2021). Combinatorial, additive and dose-dependent drug–microbiome associations. Nature *600*, 500–505. https://doi.org/10.1038/s41586-021-04177-9.

123. Moitinho-Silva, L., Forslund, S.K., Chakaroun, R., Zimmermann-Kogadeeva, M., Markó, L., Aron-Wisnewsky, J., Nielsen, T., and Birkner, T. (2021). grp-bork/vpthemall: release of vpthemall – assistant functions for variation partition with dbRDA method. Zenodo. https://zenodo.org/record/4719527.

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Deposited data	This paper	DOI: 10.5281/zenodo.7923559
Previously sequenced cohort	Hildebrand et al.[1]	SRA: PRJEB41102
Previously sequenced cohort (Kazakhstan)	Kushugulova et al.[44]	SRA: PRJEB17632
Previously sequenced cohort (Germany)	Willmann et al.[45]	SRA: PRJEB10391
Previously sequenced cohort (Sweden, infants)	Bäckhed et al.[46]	SRA: PRJEB6456
Previously sequenced cohort (USA, mothers and infants)	Chu et al.[47]	SRA: PRJNA322188
Previously sequenced cohort (mothers and infants)	Asnicar et al.[48]	SRA: PRJNA339914
Previously sequenced cohort	Lee et al.[49]	SRA: PRJNA353655
Previously sequenced cohort (Luxembourg, families)	Heintz-Buschart et al.[50]	SRA: PRJNA289586
Previously sequenced cohort (HMP)	Human Microbiome Project Consortium, 2012[51]	N/A
Previously sequenced cohort	Kostic et al.[52]	N/A
Previously sequenced cohort (Finland, infants)	Yassour et al.[53]	SRA: PRJNA290381
Previously sequenced cohort (Northern Europe)	Vatanen et al.[54]	SRA: PRJNA290380
Previously sequenced cohort (Italy, mothers and infants)	Ferretti et al.[55]	SRA: PRJNA352475
Previously sequenced cohort (USA)	Mehta et al.[56]	SRA: PRJNA354235
Previously sequenced cohort (Israel)	Zeevi et al.[57]	SRA: PRJEB11532
Previously sequenced cohort (Sweden)	Bengtsson-Palme et al.[58]	SRA: PRJEB7369
Previously sequenced cohort (Canada)	Raymond et al.[59]	SRA: PRJEB8094
Previously sequenced cohort (Denmark)	Palleja et al.[31]	SRA: ERP022986
Previously sequenced cohort (Finland, mothers and infants)	Yassour et al.[60]	SRA: PRJNA475246
Previously sequenced cohort (USA, preterm infants)	Ward et al.[61]	SRA: PRJNA63661
Previously sequenced cohort	Li et al.[62]	SRA: PRJEB12357
Previously sequenced cohort (Metacardis BMIS)	Vieira-Silva et al.[17]	SRA: PRJEB37249
Previously sequenced cohort (India)	Prasoodanan et al.[33]	SRA: PRJNA397112
Previously sequenced cohort (Argentina, China, Colombia, Guinea-Bissau)	Valles-Colomer et al.[32]	SRA: PRJEB45799
Previously sequenced cohort (Ethiopia)	Tett et al.[63]	SRA: PRJNA504891
Previously sequenced cohort (Peru, El Salvador)	Pehrsson et al.[64]	SRA: PRJNA300541
Previously sequenced cohort (Tanzania)	Tett et al.[63]	SRA: PRJNA529400
Previously sequenced cohort (Ghana)	Tett et al.[63]	SRA: PRJNA529124
Previously sequenced cohort (Fiji)	Brito et al.[65]	SRA: PRJNA217052
Previously sequenced cohort (Madagascar)	Pasolli et al.[34]	SRA: PRJNA485056
Software and algorithms		
R version 3.6.2	R Core Team[66]	https://www.r-project.org/
Enterosignature scripts version 0.1.0	This paper	https://gitlab.inria.fr/cfrioux/enterosignature-paper/ DOI: 10.5281/zenodo.7918185
Scikit-learn version 0.24.1	Pedregosa et al.[67]	https://github.com/scikit-learn/scikit-learn
Pathway Tools version 25.5	Karp et al.[68]	http://bioinformatics.ai.sri.com/ptools/
Mpwt version 0.7.0	Belcour et al.[69]	https://github.com/AuReMe/mpwt
Metage2Metabo version 1.5.0	Belcour et al.[69]	https://github.com/AuReMe/metage2metabo

(*Continued on next page*)

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MiSCoTo version 3.1.1	Frioux et al.[70]	https://github.com/cfrioux/miscoto
Rarefaction scripts	Saary et al.[71]	https://github.com/hildebra/Rarefaction
Shotgun metagenomic data processing pipeline	Hildebrand et al.[26]	https://github.com/hildebra/MATAFILER
Read depth windows calculation	Hildebrand et al.[26]	https://github.com/hildebra/rdCover
Frameshifts fixing in MSAs program	Hildebrand et al.[1]	https://github.com/hildebra/MSAfix
MetaBAT2 version 2.15	Kang et al.[72]	https://bitbucket.org/berkeleylab/metabat/src/master/
sdm version 1.47	Hildebrand et al.[73]	https://github.com/hildebra/sdm
MEGAHIT version 1.2.9	Li et al.[74]	https://github.com/voutcn/megahit
Prodigal version 2.6.1	Hyatt et al.[75]	https://github.com/hyattpd/Prodigal
Bowtie2 version 2.3.4.1	Langmead and Salzberg[76]	https://github.com/BenLangmead/bowtie2
Samtools version 1.3.1	Li et al.[77]	http://www.htslib.org/
BedTools version 2.21.0	Quinlan and Hall[78]	https://github.com/arq5x/bedtools2
CD-HIT version 4.6.1	Fu et al.[79]	http://weizhongli-lab.org/cd-hit/
MMseqs2 version f5a1cdb44c996d6be229226b09ecc687646c0c12	Steinegger and Söding[80]	https://github.com/soedinglab/MMseqs2
CheckM version 1.0.11	Parks et al.[81]	https://ecogenomics.github.io/CheckM/
MAFFT version 7.464	Katoh and Standley[82]	https://mafft.cbrc.jp/alignment/software/
TrimAl version 1.4.rev22	Capella-Gutiérrez et al.[83]	http://trimal.cgenomics.org/
IQ-TREE version 1.6.3a	Nguyen et al.[84]	https://github.com/iqtree/iqtree2
Itol	Letunic and Bork[85]	https://itol.embl.de/
GTDB-TK version 1.3.0	Parks et al.[86]	https://github.com/Ecogenomics/GTDBTk
Kraken2 version 2.0.9-beta	Wood et al.[87]	http://ccb.jhu.edu/software/kraken2/
Lambda version 1.9.3	Hauswedell et al.[88]	https://seqan.github.io/lambda/
EggNOG-mapper version 2.1.6	Cantalapiedra et al.[89]	https://github.com/eggnogdb/eggnog-mapper
Diamond version 2.0.14	Buchfink et al.[90]	https://github.com/bbuchfink/diamond
Vegan R package version 2.5-7	Oksanen et al.[91]	https://cran.r-project.org/web/packages/vegan/index.html
Dirichlet Multinomial version 1.28.0	Holmes et al.[7]	https://microbiome.github.io/
ggplot2 version 3.3.5	Wickham[92]	https://ggplot2.tidyverse.org/
Enterotyping tutorial	Arumugam et al.[6]	https://enterotype.embl.de/enterotypes.html
RTK version 0.2.6.1	Saary et al.[71]	https://cran.r-project.org/web/packages/rtk/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Falk Hildebrand (Falk.Hildebrand@quadram.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyses existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code and additional datasets and results have been deposited at Zenodo and are publicly available as of the date of publication. The DOIs are listed in the key resources table and the content is further available in the following repository: https://gitlab.inria.fr/cfrioux/enterosignature-paper
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

METHOD DETAILS

Metagenomic datasets

This study uses the GMR metagenomic dataset[1] (SRA: PRJEB41102) that includes 5,230 samples of individuals of all ages. This dataset is partly longitudinal with 1,239 out of 2,081 individuals sampled more than once. Overall, the dataset contains 2163 infant samples, 1903 adult samples, 1057 elder samples. Individuals originate from the 13 following countries: Canada (n = 72), Finland (n = 925), Sweden (n = 470), Estonia (n = 206), Russia (n = 321), Denmark (n = 57), Luxembourg (n = 93), Germany (n = 264), Kazakhstan (n = 170), USA (n = 1604), Israel (n = 857), Italia (n = 135), Holland (n =56). These cohorts were selected because they gather mostly healthy individuals while accounting for some physio-pathological conditions that are known to be associated to microbiome perturbations such as preterm birth and antibiotic treatment. Sequences were processed with the MATAFILER pipeline as described in Hildebrand et al.[1] and Bahram et al.[93]

The Metacardis Body Mass Index Spectrum (BMIS) raw metagenomes were retrieved from Vieira-Silva et al.[17] (ENA accession number PRJEB37249) and processed with MATAFILER using the same settings. It is a cross-sectional cohort consisting of 888 samples. Metadata was retrieved from Vieira-Silva et al.[17] and Forslund et al.[94]

We additionally gathered cohorts of individuals with a non-western (NW) lifestyle. The Indian cohort was obtained from Vishnu Prasoodanan et al.[33] (PRJNA397112). All other cohorts were retrieved from Valles-Colomer et al.[32] (Argentina, China, Colombia and Guinea-Bissau: PRJEB45799; Ethiopia: PRJNA504891[63]; Peru and El Salvador: PRJNA300541[64]; Tanzania: PRJNA529400[63]; Ghana: PRJNA529124[63]; Fiji: PRJNA217052[65]; Madagascar: PRJNA485056[34]).

Briefly, raw shotgun metagenomes were quality filtered using sdm v1.63 with default parameters,[95] assembled using MEGAHIT v 1.2.9 with parameters "–k-list 25,43,67,87,101,127 "[74] and reads mapped onto assemblies using Bowtie2 v2.3.4.1 with parameters "–end-to-end "[76], genes predicted with Prodigal v2.6.1 with parameters "-p meta"[75] and a gene catalogue clustered at 95% nt identity using MMseqs2.[80] MAGs (metagenomic assembled genomes) were binned using Metabat2 v2.15[72] and combined in MATAFILER to MGS (metagenomic species), relying on canopy clustering.[96] Matrix operations were carried out using rtk.[71] The full MATAFILER pipeline is available at https://github.com/hildebra/MATAF3 .

MGS phylogenies were calculated *de novo* based on the amino acid (AA) sequences of at most 40 marker genes[97] present in each MGS and subsequently aligned using MAFFT v v7.464[82] with default options; the multiple sequence alignment (MSA) was then trimmed and translated from nucleotide to aminoacid sequences using TrimAl[83] (options "-keepheader -ignorestopcodon -gt 0.1 -cons 60"), from these IQ-TREE v1.6.3.a[84] with parameters "-m GTR+F+I+G4 -B 1000" was used to reconstruct a phylogeny, that was visualized with iTOL.[85]

Non-negative matrix factorisation

Genus-level relative abundance matrices were normalized by the sum of abundances per sample after removal of unassigned organisms at domain level. NMF was performed with Scikit-Learn v0.24.1[67] using the multiplicative update solver, Kullback-Leibler divergence as a beta-loss function, random initialisation, and a maximal number of iterations of 2000. Nine-fold bicross validation was performed as described in Owen and Perry,[98] Kanagal and Sindhwani,[99] and Eng et al.[100] in order to determine the number of signatures and the regularization ratio (Figure S1B). Briefly, for each fold, 1/9 of the observations was withheld for validation. The basis matrix W and mixture coefficient matrix H of the validation set were calculated using the matrices obtained for the training set. 100 repetitions of this process were performed in total, each time after shuffling the matrix. All runs were performed for each number of clusters k, ranging from 2 to 10. The quality of decomposition for the validation was calculated as previously described[101,102] with the explained variance:

$$EV = 1 - \frac{\sum (x_{ij} - \hat{x}_{ij})^2}{\sum x_{ij}^2}$$

The higher the EV value, the more accurately the decomposition represents the input compositional matrix X.

The quality of the decomposition was further estimated using a cosine similarity between the reconstructed matrix and the original abundance matrix, or a sample microbial composition profile and its prediction, referred to as model fit score in the manuscript. For a contiguous flattened (ravel function from the Python numpy package[103]) matrix (whole dataset) or a vector (sample) x and its estimation $\hat{x}$:

$$CS = \frac{\sum x_i \hat{x}_i}{\sqrt{\sum x_i^2} \sqrt{\sum \hat{x}_i^2}}$$

The cosine similarity ranges from 0 to 1, higher values indicating greater similarity between x and $\hat{x}$.

The optimal number of signatures was chosen by assessing the gain of cosine similarity (median) when increasing the number of signatures: when the gain no longer increases, the optimal number of signatures is reached.

L1 regularisation parameter α applied to both W and H matrices was estimated by running the same procedure of bicross-validation with the α coefficients ranging from 0 to 100. The regularisation parameter was selected based on the method from Eng et al.[100] Briefly, we defined a threshold equal to the mean explained variance (MEV) value at $\alpha = 0$ minus the standard deviation of the mean.

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

We selected the α parameter value leading to the model whose associated MEV was greater than the threshold. We selected for ES computation an alpha value 1.0, and regularization was performed on both H and W matrices.

In order to reapply the GMR ESs to the Metacardis BMIS and non-western datasets, we reused the W matrix obtained from GMR together with the BMIS normalized relative abundance matrix as inputs to the Scikit-Learn 'non_negative_factorization' function, with the 'update_H' parameter set to 'False' and 'init' to 'custom'. Running the computation of H only consists in solving a non-negative least square problem. Prior to computation, we homogenized the taxa contents of the BMIS relative abundance matrix and W matrix by adding zero-filled rows for the genera that were absent in W or in the abundance matrix respectively.

Enterosignature analyses

The enterosignature results were calculated using the whole dataset and the parameters (number of ESs, regularization parameters) obtained with bicross-validation. The matrices W and H resulting from the NMF algorithm represent the weight of genera in enterosignatures and the presence of enterosignatures in samples, respectively (Figure 1A). Normalizing W by its columns or its rows informs on the general composition of ESs in genera, and on the association strength of genera to each ES, respectively.[101] By normalizing the H matrix column-wise, we obtained the relative abundance of ES in each sample. We define as *primary ES* the ES of a sample with the highest relative abundance.

We used the normalized H matrix to determine the number of ESs needed to best describe a sample. To that purpose, we defined an arbitrary cut-off of 90% accumulated relative abundance to describe all major ES components of a sample. Thus, our operational definition posed that the set of ESs describing a sample are ESs ordered by decreasing relative abundance, whose cumulated relative abundance is greater or equal to 0.9. This is as well the approach we used to discretely assign the composition of ESs in samples. Additionally, we describe as a "*monodominant ES*" an ES whose relative abundance is greater than 0.9, i.e., samples that are best described by a single ES according to the above definition.

The diversity of ESs in samples is also estimated, using the Shannon diversity index. The Enterosignature Shannon Diversity (ESSD) is calculated using the diversity function (Shannon index) of the vegan R package v2.5-7[91] on the normalized H matrix.

We calculated the persistence of ESs in individuals sampled longitudinally using survival analysis. We distinguished the survival of the primary ES over time from the survival of the presence of an ES. For the former case, survival analysis per signature was performed after assigning to patients in longitudinal studies the most representative signature for each of their samples. For the latter case, we considered as present all ESs with a relative abundance greater than 1% and analyzed their time of disappearance in longitudinal samples. Persistence of ES and primary ES were calculated using the Survival R package v3.2-7.[104] Statistical tests for the comparison of survival predictions between groups are log rank Chi^2 tests from the Survival and Ezfun (v0.1.3) R packages.[105]

We determined outlier samples regarding their cosine similarity to the original genus-level abundances. Such atypical samples exhibited a cosine model fit score lower than 0.4; the threshold was selected by looking at the distribution of cosine values and the interquartile range (IQR): 25th percentile – (1.5 * IQR).

Functional analysis and metabolic network reconstruction

Consensus genomes of metagenomic species (MGS) and references genomes[106] corresponding to the GMR dataset were retrieved according to the data in Hildebrand et al.[1] Proteomes were functionally-annotated using EggNOG-mapper v2.1.6[89] based on eggNOG orthology data.[107] Sequence searches were performed using Diamond v2.0.14.[90]

CAZymes and Kegg Orthology (KO) annotations were retrieved from EggNOG-mapper annotation files. Higher-level categories associated to KO were downloaded from the KEGG database.[108] For each KO found in a genome, we incremented the number of the corresponding higher level metabolic categories and we ultimately compared the metabolic functions of genomes grouped by their ES assignments. CAZymes[109] were associated to their respective substrates according to Cantarel et al.[110] and The CAZypedia Consortium[111] and genomes were compared after grouping by ES. In both CAZyme and KEGG annotations, values by genome were normalized by the number of genes in the genome.

Genome-scale metabolic networks (GSMNs) were reconstructed based on the annotated proteomes using mpwt v0.7.0[69] with Pathway Tools v25.5.[68] Analyses of metabolic pathways in the networks relied on Metacyc v25.1 database.[112] We retrieved the genera accounting for more than 3% of an enterosignature's activity and considered them as representatives of their respective enterosignature. This threshold was chosen because it allowed us to uniquely assign genera to an ES. Then, by grouping GSMNs assigned to genus-level representatives, a group of representative GSMNs was associated to each enterosignature. Of 1,737 total species, 55, 81, 60, 149 and 43 MGS were stably associated to ES-Bact, ES-Firm, ES-Prev, ES-Bifi and ES-Esch, respectively.

The metabolic potential and metabolic complementarity between enterosignatures were determined with Metage2Metabo v1.5.0.[69] The software simulates the qualitative individual and collective metabolic potentials of GSMNs provided a set of nutrients. A set of nutrients consisting of molecules described in a western diet[113] was selected. Three experiments were tested. For the first one, metabolic potential of each ES in isolation was calculated, running Metage2Metabo on GSMNs associated to each ES. For the second experiment, we computed the metabolic potential of pairwise combinations of all ES as well as all observed discrete combinations of ES in the samples. Finally, for the third experiment, all representative GSMNs for all ES were combined in the Metage2Metabo run. For each run of Metage2Metabo, the individual metabolic potential (molecules predicted to be producible by metabolic networks considered individually), the collective metabolic potential (molecules predicted to be producible by the

community while enabling mutualistic interactions) and the mutualistic potential (molecules predicted to be producible only through mutualistic interactions) were computed. In addition, MiSCoTo v3.1.1[70] was applied to the above-described communities to determine the metabolic potential of each GSMN in communities.

Enterotype calculation

Enterotypes were calculated on both the GMR (adult samples only) and the Metacardis BMIS datasets. Their computation was performed after the removal of unknown taxa on the normalized (by sample) genus-level abundance matrices multiplied by 1000 and rounded to the closest integer. The package DirichletMultinomial v1.28.0[7] was used for DMM enterotyping (https://microbiome.github.io/tutorials/DMM.html). PAM clustering was calculated on normalized (by sample) matrices after removal of unknown taxa using the Jensen-Shannon divergence and the package Cluster v2.1.1.[114] The best number of enterotypes was determined using the Calinski-Harabasz index for PAM clustering and the model fit score for DMM. While the optimal number of cluster differed (2 for PAM, 16 for DMM in GMR), we tested specifically the known 3 PAM / 4 DMM cluster models. We assigned to the clusters the known enterotypes identified by both methods, that are *Bacteroides* 1, *Bacteroides* 2 (low richness Bacteroides), *Prevotella* and Firmicutes/*Ruminococcus* for DMM, and *Bacteroides*, *Prevotella* and *Ruminococcus*/Firmicutes for PAM. The taxonomic composition of enterotypes was determined by calculating, for each ET, the average relative abundance of taxa of interest in samples assigned to this ET, as described in Vieira-Silva et al.[115]

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless stated otherwise, statistical and data analysis were performed in R v3.6.2.[66] Unless stated otherwise, the Wilcoxon rank sum test was used to compare means. When appropriate, multiple testing corrections were performed using the Benjamini-Hochberg method. In all boxplots, the box represents the interquartile (IQR) range and displays the median as a horizontal line, while the lower (respectively upper) whisker extends to the maximum (resp. minimum) between 1.5*IQR and the smallest (resp. largest) value of the data.

Correlation between the Bray-Curtis dissimilarity matrix calculated on the relative abundance matrix and the Euclidean (resp. Bray-Curtis) dissimilarity distance matrix calculated on ET or ES assignment (resp. ES relative abundance) matrices was calculated using a Mantel test from the vegan (v2.5-7) R package[91] using Spearman correlations.

Pairwise adonis was performed with the pairwiseAdonis package v0.4[116] and Benjamini Hochberg correction. Graphics were generated with ggplot2 v3.3.5,[92] ggpubr v0.4.0,[117] diagram v1.6.5[118] and plotly v4.10.0.[119] Alluvial plots were generated using ggalluvial v0.12.3[120] and Venn diagrams with eulerr v6.1.0.[121]

The multivariate effect of BMIS metadata variables including healthy factors and medications were tested according to Forslund et al.[122] using R v4.1.0. To remove highly redundant metadata variables, those with a Kendall's tau correlation of more than 0.8 were excluded. Furthermore, metadata with missing values for more than 50% of the samples were removed, as well as the variables FOSFOMYCIN, C07AB02.x, and C09CA08.x as the model failed with these. Subsequently, any samples with missing data in any of the remaining metadata variables were removed. The remaining dataset included 699/888 samples and 143/151 metadata variables. Using the vegdist package from the library vegan (v2.5-7) the Bray-Curtis dissimilarities between samples were calculated for ES and genus abundances, and DMM and PAM enterotype assignments, respectively. Variables were selected for each of the four matrices by using the function select.minimal.model (seed set to 222) from the VpThemAll package. [123] This package automatically computes a stepwise selection of metadata variables in both directions using the ordistep function from the library vegan (v2.5-5). Subsequently, all variables selected for the four matrices were combined into a single model, including also the variable "CENTER" as a potential confounding factor. Using the test.varpart.wrap function from the VpThemAll package which iteratively performs the same set of dbRDA analyses from the library vegan (v2.5-5) for each of the distances matrices (seed set to 222). The function determines the unique effect of each of the variables by conditioning for all other variables in the model. Here we separated the metadata into intrinsic (gene richness, faecal microbial load) and extrinsic (141 clinical and demographic metadata). The significance was determined using 999 permutations and the p-value was corrected for multiple testing using Benjamini-Hochberg adjustments. All models with an adjusted p-value of < 0.1 were visualized using ggplot2 v3.3.5 including the sum of total variance explained. GMR metadata were subsampled to include a single sample per individual, removing all other timepoints. In case antibiotics was taken, the first time-point between 2 and 14 days after the intake was considered, otherwise the first time-point of the individual was considered. Metadata were further filtered to remove categories with more than 50% missing values, and subsequent removal of all samples with missing data in any of the remaining variables. The remaining dataset included 1180/2088 individuals and 5/44 metadata variables. Due to the small number of variables all of these were included in the final model and the unique effects determined with the test.varpart.wrap function to the GMR dataset as described above.

Furthermore, we performed a Spearman correlation using the spearman_test function from the coin package v1.4-2 between the numeric metadata and ES abundances and cosine fit, respectively. For the BMIS dataset all numeric metadata variables filtered as above were used, resulting in 2 intrinsic (richness, faecal microbial load) and 42 extrinsic metadata variables remaining. To overcome limitation of missing metadata, for the GMR dataset all (unfiltered) numerical metadata variables were selected and binary variables (sex, antibiotics, delivery mode) were transformed to numeric resulting in 9 variables for adults. Additionally, for the GMR dataset Spearman correlations were calculated also for infant data (age below 3 years, including preterm) with 11 numeric variables. For

Please cite this article in press as: Frioux et al., Enterosignatures define common bacterial guilds in the human gut microbiome, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.05.024

GMR data the effect introduced by the cohort was accounted for by including 'Study' as a blocking factor. The p-values were adjusted for multiple testing according to the Benjamini-Hochberg method. All data with an adjusted p-value < 0.1 have been visualized using the ggplot2 package.

ADDITIONAL RESOURCES

Summary website of the project: https://enterosignatures.quadram.ac.uk/

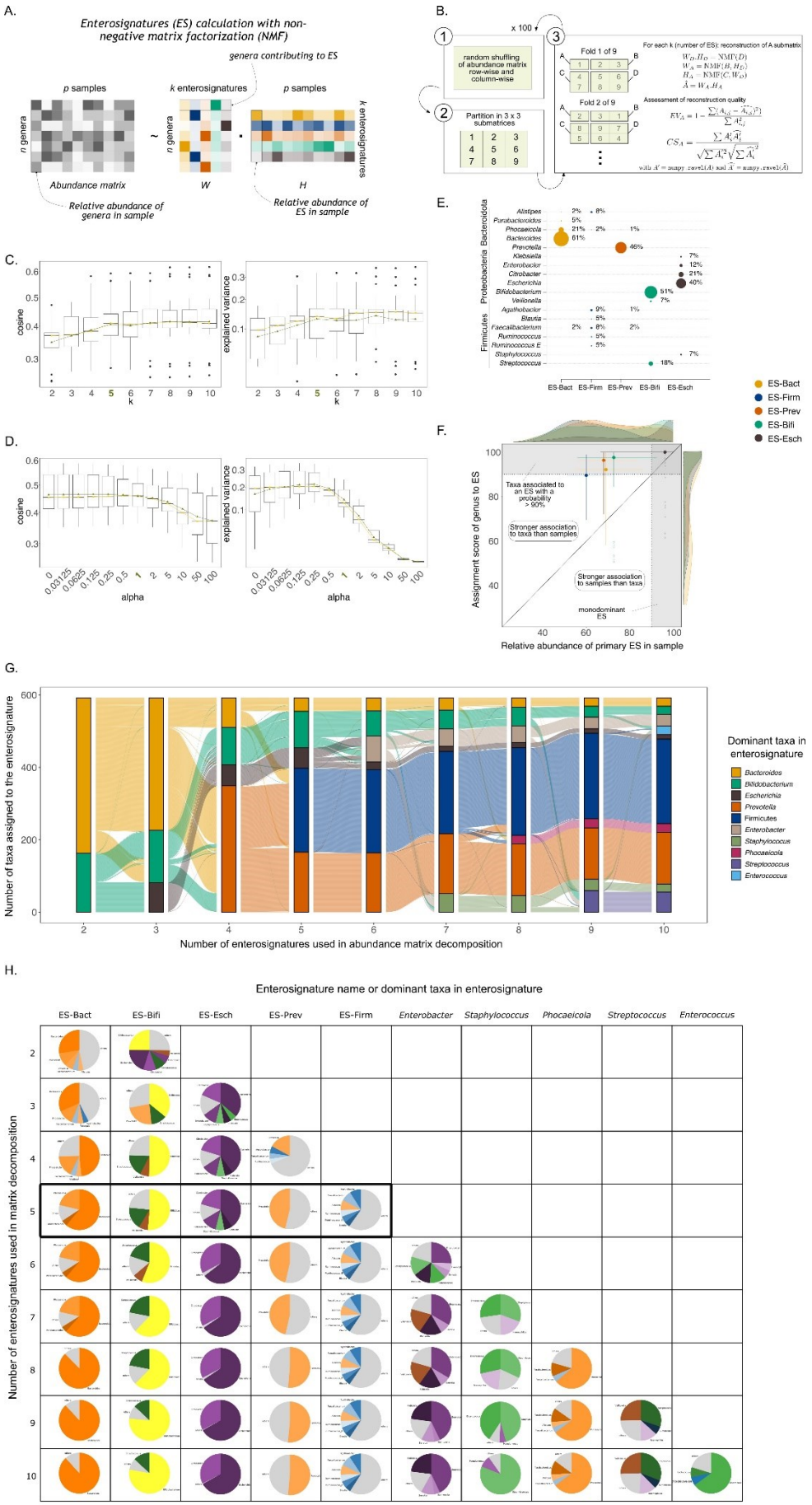
Cell Host & Microbe, Volume *31*

Supplemental information

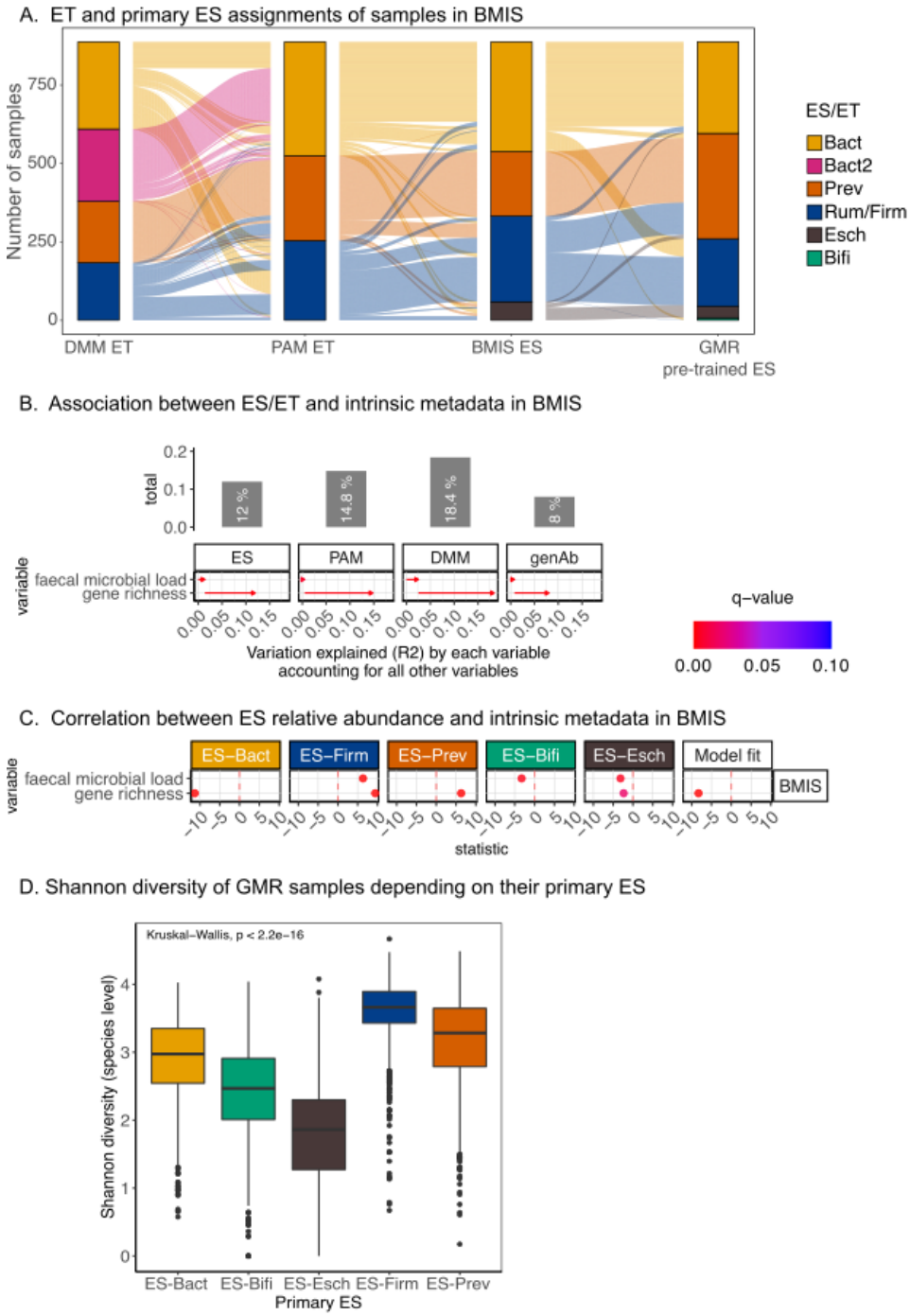
Enterosignatures define common bacterial guilds in the human gut microbiome

Clémence Frioux, Rebecca Ansorge, Ezgi Özkurt, Chabname Ghassemi Nedjad, Joachim Fritscher, Christopher Quince, Sebastian M. Waszak, and Falk Hildebrand

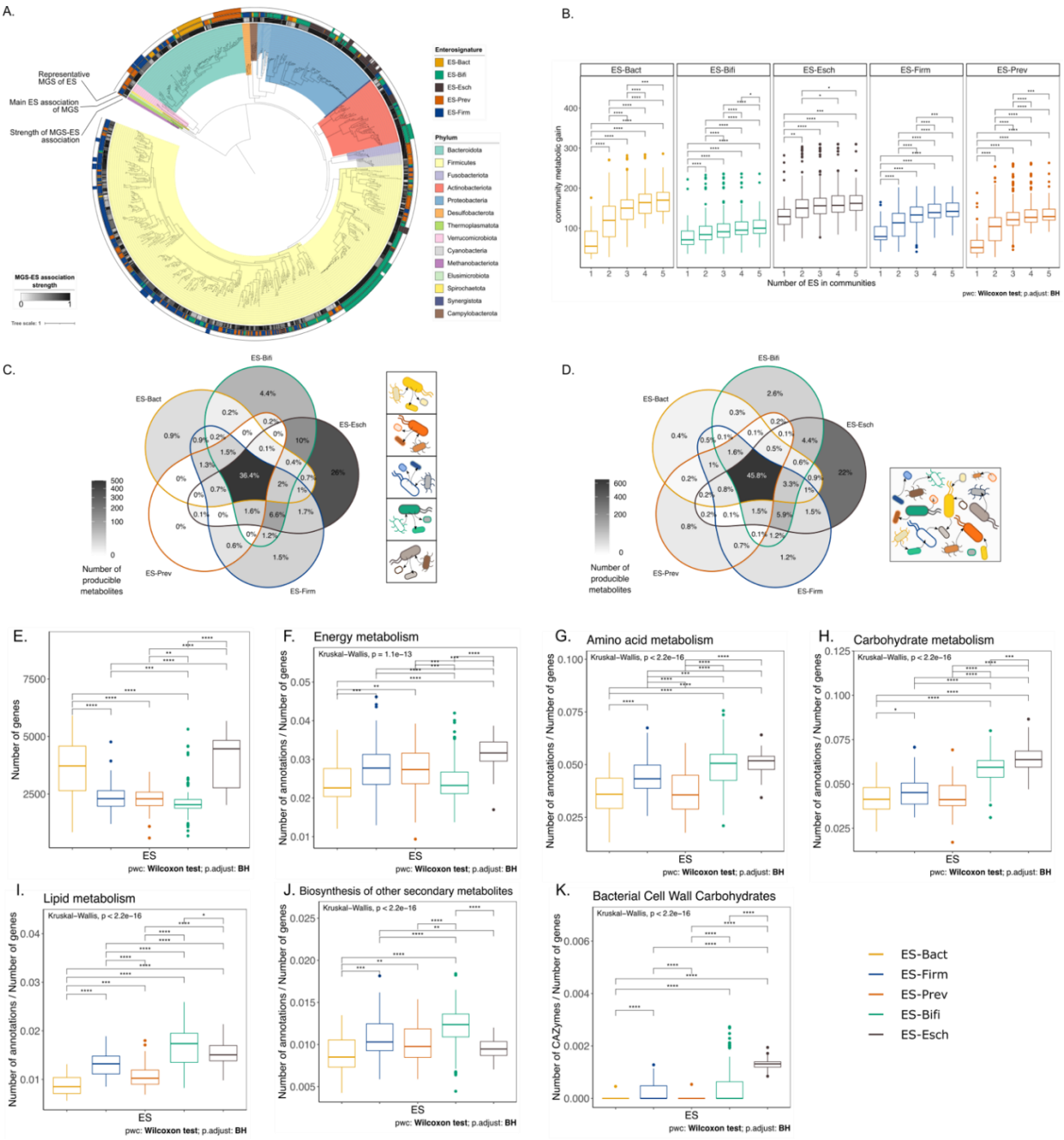
Supplementary Figures


A.
Enterosignatures (ES) calculation with non-negative matrix factorization (NMF)
genera contributing to ES
p samples
k enterosignatures
p samples
n genera
n genera
k enterosignatures
Abundance matrix
W
H
Relative abundance of genera in sample
Relative abundance of ES in sample
B.
x 100
random shuffling of abundance matrix row-wise and column-wise
Partition in 3 x 3 submatrices
Fold 1 of 9
Fold 2 of 9
For each k (number of ES): reconstruction of A submatrix
Assessment of reconstruction quality
C.
cosine
explained variance
k
D.
cosine
explained variance
alpha
alpha
E.
Bacteroidota
Proteobacteria
Firmicutes
ES-Bact
ES-Firm
ES-Prev
ES-Bifi
ES-Esch
F.
Assignment score of genus to ES
Taxa associated to an ES with a probability > 90%
Stronger association to taxa than samples
Stronger association to samples than taxa
monodominant ES
Relative abundance of primary ES in sample
G.
Number of taxa assigned to the enterosignature
Number of enterosignatures used in abundance matrix decomposition
Dominant taxa in enterosignature
Bacteroides
Bifidobacterium
Escherichia
Prevotella
Firmicutes
Enterobacter
Staphylococcus
Phocaeicola
Streptococcus
Enterococcus
H.
Enterosignature name or dominant taxa in enterosignature
ES-Bact
ES-Bifi
ES-Esch
ES-Prev
ES-Firm
Enterobacter
Staphylococcus
Phocaeicola
Streptococcus
Enterococcus
Number of enterosignatures used in matrix decomposition

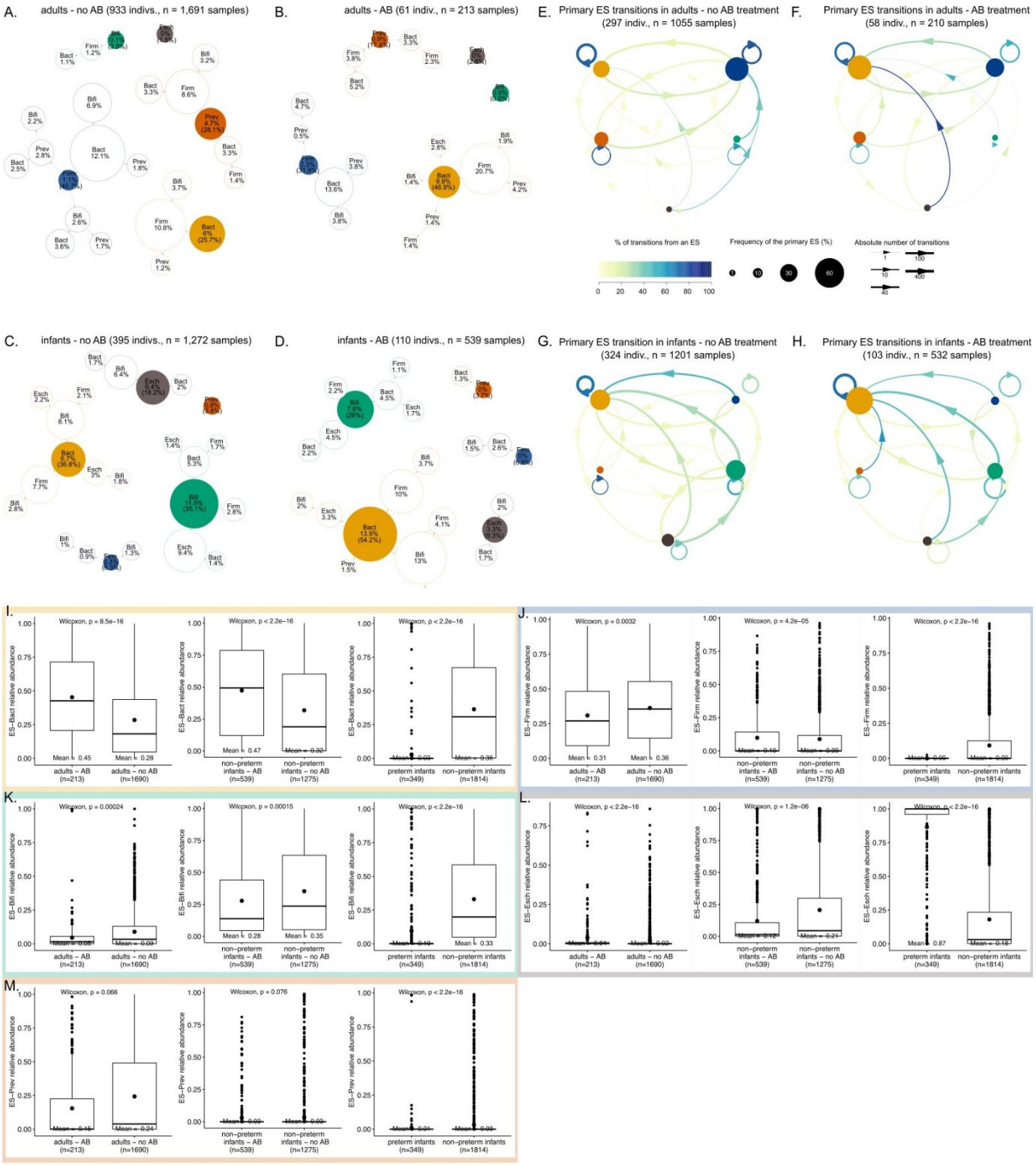
Supplementary Figure S1 related to STAR Methods and Figure 1 – Decomposition of microbiome diversity using the NMF algorithm. (A.) Concept of enterosignatures, non-negative matrix factorization (NMF) is used to determine bacterial guilds driving variance in gut metagenomes. (B.) Protocol followed for 3X3 bi-cross validation and metrics used for quality assessment. (C.) Results of 3x3 bi-cross validation for rank (number of ES) choice. (D.) Results of 3x3 bi-cross validation for regularisation ratio choice. (E.) Relative genus-level composition of the five enterosignatures identified as optimal NMF solution. Only genera accounting for >4% of any ES are shown. (F.) Assignment scores of genera to ES and relative abundance of primary ES in samples. Dots represent the median value, error bars are the whiskers of the distribution boxplots, corresponding density plots to this data are on the top and right side. (G.) Evolution of taxa assignments to ES depending on the rank of the decomposition. (H.) Genus-level composition of ES from decomposition ranks ranging from k = 2 to k = 10.
Abbreviations: EV, explained variance; CS, cosine similarity; NMF, non-negative matrix factorization.



Supplementary Figure S2 related to Figure 2 - Application of GRM ES to the BMIS dataset. (A.) Alluvial plot depicting the relation of re-applied
GMR ES, de novo trained BMIS ES or enterotypes (PAM, DMM) assignments of BMIS samples. Number of ES relates to the GMR ES reapplied to BMIS. (B.) Variation of ES, ET and genera abundances is partly explained by metadata in BMIS and GMR cohorts. Only significant associations (q-value < 0.1) are included. (C.) Significant spearman correlations (q-value < 0.1) between intrinsic metadata of the BMIS dataset and the 5 ES relative abundances. Dashed line represents Z-statistic of zero to show negative (to the left) and positive (to the right) correlations (B.) and (C.) ES are those calculated on the GMR dataset. (D.) Shannon diversity of GMR metagenomes according to their respective primary ES.

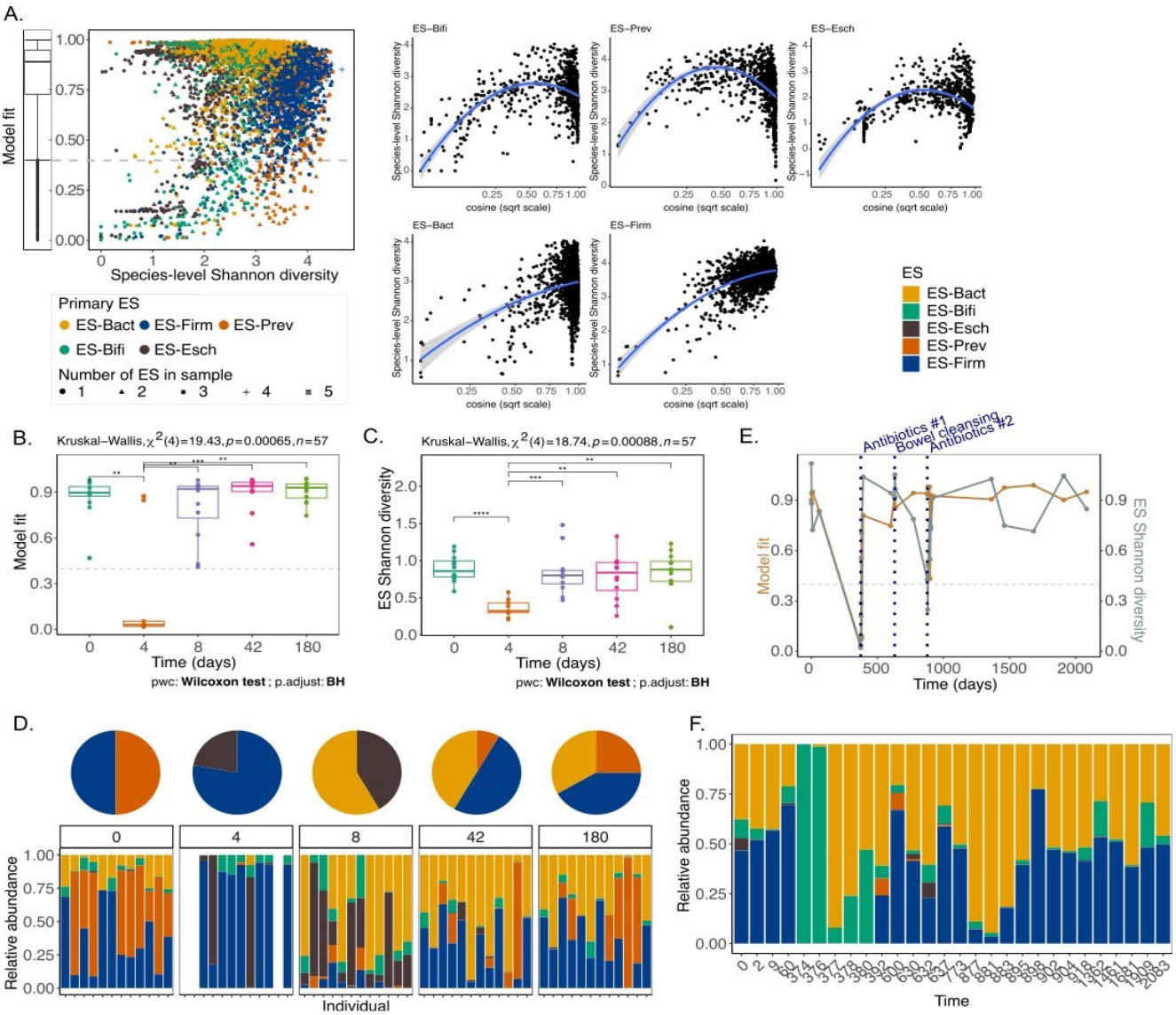


Supplementary Figure S3 related to Figure 3 – Metabolic potentials of enterosignatures and between species phylogeny. (A.) Association between ES and metagenomic species (MGS) retrieved from the GMR dataset. A phylogenetic tree was calculated for all 1,737 MGS and reference genomes. Based on genus level ES assignments, the probability that an MGS belongs to an ES was calculated (strength of association, inner ring). ES primarily associated to MGS is illustrated in the middle ring. To derive representative groups of MGS typically representing bacterial guilds for each ES, only MGS associated to genus accounting for more than 3% of the content of an enterosignature were kept (shown in outer ring). (B.) Metabolic gain of MGS of enterosignatures in communities of up to 5 ES. The metabolic potential of all observed ES combinations (1 to 5 ES) was simulated by combining ES-associated MGS into increasingly larger communities. The boxplots display the gain in metabolic potential of each MGS when simulated in the observed communities with respect to its metabolic potential computed alone (x = 1 ES in community). The 31 observed ES combinations (5 of size 1, 10 of size 2, 10 of size 3, 5 of size 4 and 1 of size 5) together with their assignments to samples are available in Supp. Table 5. (C.), (D.) Venn diagrams illustrating the redundancy in the sets of metabolites predicted to be producible by MGS of each ES in interactions with MGS of the same ES (C.) or with MGS of all ES (D.). (E.) – K.) Metabolic pathways are differentially prevalent in different ES. (E.) Genome size (gene number) in genomes associated to the 5 ES. (F.)-(J.). Number of KEGG Orthology (KO) annotations grouped by metabolic categories, normalised by genome size. (K.) Number of CAZymes degraded bacterial cell wall carbohydrates normalised by genome size. Abbreviations: PWC, pairwise comparison; BH, Benjamini-Hochberg; MGS, Metagenomic Species.

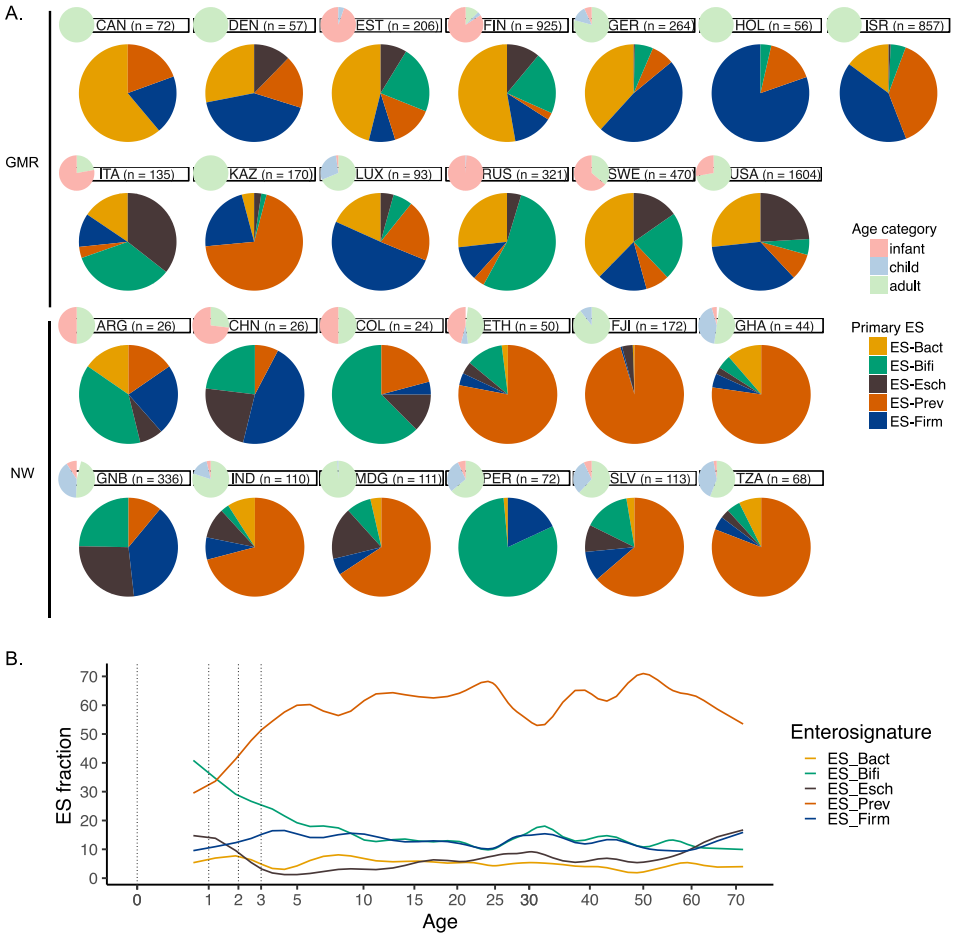


Supplementary Figure S4 related to Figure 4 - Association of ES relative abundance and composition to metadata and ES transitions with or without antibiotic treatment. (A.)-(D.) Associations of ES in samples. Coloured nodes are primary ES. The percentage in a coloured node is the frequency of the monodominant ES, the percentage between brackets indicates the proportion of samples exhibiting the corresponding primary ES. Diameter of nodes are proportional to the frequency of the corresponding ES combination. Combinations of ES with frequencies lower than 1% are not represented. Nodes linked to primary ES denote the observed secondary ES with their relative proportion out of the total number of samples. Likewise,

blank nodes connected to secondary ES are the tertiary ES. (A.) Adult metagenomes, no antibiotic (AB) treatment. (B.) Adult metagenomes with antibiotic treatment. (C.) Infant (preterm excluded) metagenomes, no antibiotic treatment. (D.) Infant (preterm excluded) metagenomes with antibiotic treatment. Example of interpretation in (A.): 40.7% of metagenomes exhibit ES-Firm as primary ES. 12.1% of metagenomes are composed of the following two ES: ES-Firm as primary, ES-Bact as secondary. 1.8% of metagenomes consist of the following 3 ES in decreasing order of abundance: ES-Firm, ES-Bact, ES-Prev. (E.)-(H.) Transitions between primary ES in longitudinal samples of adults (E.), (F.) and infants (G.)-(H.) with (F.), (H.) or without antibiotic treatment (E.), (G.). Abbreviation: AB = antibiotics. (I.)-(M.) ES relative abundance changes between antibiotic (AB) and non-treated (no AB) gut microbiomes, as well as preterm vs normal birth infants. Abbreviation: AB, antibiotic treatment.



Supplementary Figure S5 related to Figure 5 – Atypical ES composition on sub-datasets. (A.) Species-level Shannon diversity and ES model fitting scores of the GRM samples. Shape and colour of the points denote the number of ES in samples and the primary ES respectively. Subplots depict the association between both metrics by ES: only samples dominated by a specific ES are considered for computation. The blue curve on each subplot is a second-order fit with its confidence interval. The non-linear relationship between Shannon diversity and model fit suggests that both metrics convey different information. (B.), (C.), (D.): analysis of time series data from individuals with strong antibiotic interventions (Palleja et al 2018 cohort), (B.) Development of ES Shannon diversity over time, (C.) ES Shannon diversity over time, (D.) Relative abundance of ES over time in individuals (barplot) and distribution of primary ES over time (pie charts). (E.) and (F.) analysis of time series data of a single individual treated on two separate occasions with antibiotics (Hildebrand et al 2019 cohort) (E.) Development of ES model fitting score and ES Shannon diversity over time. Vertical dotted lines represent (chronologically): first antibiotic treatment, bowel cleansing, second antibiotic treatment. Horizontal line describes the threshold used to define ES-atypical samples based on ES model fitting scores of the GMR dataset. (F.) ES relative abundances in individual over timepoints in (E.).



Supplementary Figure S6 related to Figure 6 – ES composition of NW samples. (A.) Primary ES prevalence by country in the GMR and NW datasets. Smaller pie plots depict the distribution of age categories within countries. (B.) ES relative abundance in NW samples according to the age of the donors. CHN, MDG and GNB datasets are excluded; only the samples for which the age of the donor is available are considered (n = 627). Mean relative abundance of each ES for groups of samples of sliding age windows are plotted. Abbreviations. GMR = gut microbiome reference, NW = non-western. CAN = Canada; DEN = Denmark; EST = Estonia; FIN = Finland; GER = Germany; HOL = Holland; ISR = Israel; ITA = Italia; KAZ = Kazakhstan; LUX = Luxembourg; RUS = Russia; SWE = Sweden; ARG = Argentina; CHN = China; COL = Colombia; ETH = Ethiopia; FJI = Fiji; GHA = Ghana; GNB = Guinea-Bissau; IND = India; MDG = Madagascar; PER = Peru; SLV = El Salvador; TZA = Tanzania.