Global and temporal state of the human gut microbiome in health and disease

Saeed Shoaie (saeed.shoaie@kcl.ac.uk)
Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London https://orcid.org/0000-0001-5834-4533

Sunjae Lee
King's College London

Mathieu Almeida
INRAE

Gholamreza Bidkhori
King's College London

Nicolas Pons
INRAE

Florian Onate
INRAE

Emmanuelle Chatelier
INRAE

Neelu Begum
King's College London

Ceri Proffitt
King's College London

Dorinês Rosário
King's College London

Stefania Vaga
King's College London

Junseok Park
KAIST

Kalle von Feilitzen
Science for Life Laboratory, KTH

Fredric Johansson
Science for Life Laboratory, KTH – Royal Institute of Technology

Victoria Meslier
INRAE

Azadeh Harzandi
King's College London

Lucie Etienne-Mesmin
INRAE
Lindsey Edwards
King's College London https://orcid.org/0000-0002-8222-4555
Vincent Lombard
INRAE
Franck Gauthier
INRAE
Claire Steves
King's College London
David Gomez-Cabrero
King's College London
Bernard Henrissat
Aix-Marseille University https://orcid.org/0000-0002-3434-8588
Doheon Lee
Korea Advanced Institute of Science and Technology https://orcid.org/0000-0001-9070-4316
Debbie Shawcross
King's College London
Stéphanie Blanquet-Diot
INRAE
Gordon Proctor
King's College London
Lars Engstrand
Karolinska Institutet
Adil Mardinoglu
Science for Life Laboratory, KTH – Royal Institute of Technology
Jens Nielsen
Chalmers University of Technology https://orcid.org/0000-0002-9955-6003
Stanislav Ehrlich
INRAE
Mathias Uhlen
Science for Life Laboratory https://orcid.org/0000-0002-4858-8056

Article

Keywords: human gut microbiota, microbiome, bacterial diversity, health, disease

Posted Date: April 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-339282/v1

Abstract

The role of gut microbiota in humans is of great interest, and metagenomics provided the possibilities for extensively analysing bacterial diversity in health and disease. Here we explored the human gut microbiome samples across 19 countries, performing compositional, functional and integrative analysis. To complement these data and analyse the stability of the microbiome, we followed 86 healthy Swedish individuals over one year, with four sampling times and extensive clinical phenotyping. The integrative analysis of temporal microbiome changes shows the existence of two types of species with a tendency to vary in abundance with time, here called outflow and inflow species. Importantly, the former tends to be enriched in disease, while the latter is enriched in health. We suggest that the decrease of disease-associated outflow and the increase of health-associated inflow species with time may be a fundamental albeit previously unrecognized aspect of the homeostasis maintenance in a healthy microbiome.

Introduction

Metagenomic studies of the human microbiome enable the characterization of the microbial and functional diversity in health and disease[1]. Advances in metagenome assembly and various clustering methods enabled the generation of metagenome species[2-6]. Most of these studies focused on unveiling new uncultured genomes, while only few focused on investigating the functional potentials and dynamic changes of the gut microbiome[7-9]. Understanding the functional and temporal behaviour of the microbiome may have great implications for the identification of its global signature in health and disease[10-12]. Additionally, short-term perturbations may trigger gut microbiota dysbiosis and changes at compositional and functional levels. Specifically, the negative selective microbe-microbe and host-microbe interactions, in the context of metabolism or antimicrobial machinery, could be the main mechanism underlying microbial dysbiosis[13]. Large-scale integration of microbiome functional changes and their associations with clinical data may provide novel information on temporal changes in the microbiome and host physiology[14].

Herein, we integrated publicly available data from a large number of studies across different countries from healthy and diseased individuals. The analysis is presented in an open-access Human Gut Microbiome Atlas (www.microbiomeatlas.org), allowing researchers to explore for the first time an integrative analysis on composition, functional, richness, diseases and region signatures for the gut microbiota across 19 geographical regions and 20 diseases. This analysis was followed by investigating the gut microbiome of healthy Swedish individuals with four times sampling across one year to study the longitudinal variability of the microbiota. This revealed the tendency of disease-associated species to decrease in abundance. In contrast, the health-associated species tended to increase in abundance. We suggest that this dynamic contributes to the maintenance of gut microbial homeostasis in healthy individuals. These findings were further validated by follow up sampling from same cohort with additional two time points across 6 months.

Human Gut Microbiome Atlas; Pan-metagenomics study on compositional and functional changes of human gut microbiome

We performed large-scale integrative analysis of 4,880 publicly available shotgun metagenomics stool samples, with at least 10 million high-quality sequencing reads from healthy and diseased cohorts from 19 different countries across five continents (Fig. 1a-b and Supplementary Table S1). We rarefied all metagenomic samples into 10 million reads per sample, which enable comparative analysis across different cohorts. We created the Human Gut Microbiome Atlas (HGMA) using quantitative analysis of shotgun metagenomics based on microbial genomes assembled using Metagenomic Species Pan-genomes (MSPs) (Fig. 1c). The MSP number was increased from 1,661 (previous release[5]) to 1,989 (average number of genes 1,894 ± 1,616) (Methods), and their taxonomy was updated. We generated gene counts and MSP abundances for all the samples using the 10.4 million gene catalogue[15]. We also characterized the functions and phenotype of the MSPs in 7 different categories (KO, PFAM, CAZyme, Mustard, JGI-GOLD phenotype, PATRIC virulence factor, and antiSMASH biosynthetic gene clusters) and identified co-conserved functional clusters across species (7,763 clusters) (Methods). This information was completed with 344 newly sequenced longitudinal samples from 86 Swedish individuals, described in detail in a subsequent section (Fig. 1d). All the data are freely available in the HGMA, without restrictions, in the public open access database (www.microbiomeatlas.org) that is part of the Human Protein Atlas program (https://www.proteinatlas.org). All MSPs and functions are highlighted together with the 5,224 samples across 19 countries with disease and healthy cohorts.

Using the 3,039 samples obtained from healthy individuals across 18 countries, including westernized and non-westernized regions, we uncovered the geographical distribution of the healthy gut microbiome. To this end, we applied the unsupervised clustering method, *monocle*, to MSP abundance profiles of the 3,039 samples (Methods)[16,17]. We observed that there were two distinct ordinations of non-westernized and European samples of healthy subjects (Fig. 1e and Supplementary Fig. 1). Based on comparative analysis across different regions, we also identified 783 MSPs specifically enriched in certain countries (See Methods, Extended Fig 1a-d, Supplementary Fig. 2 and Supplementary Table S2). Functional annotation-based analysis across geographical clusters, revealed an enrichment of CAZymes for degrading host mucins and storage carbohydrates in westernized population, where antimicrobial resistance (AMR) and virulence factors were also more prevalent (Fig. 1f). Comparison of functions of region-enriched MSPs in European countries and in US/China/Japan revealed that fosfomycin and aminoglycoside resistance were the significant AMR, as deduced from the explained variances of ANOVA (Methods). Among the biosynthetic genes encoding the production of secondary metabolites, resorcinol, lanthipeptide, bacteriocin and homoserine lactone were enriched in the European and US/China/Japan populations. These secondary metabolites, together with AMR, appeared to be the key feature in the westernized countries.

To distinguish diseased and healthy microbiome from multiple cohorts, we performed a pan-metagenomics association study (Pan-MGAS) of multiple disease cohorts (18 diseases across 28 cohorts from 11 westernized countries). We determined the enriched and depleted species in diseased

compared to healthy control samples, with an effect size > 0.3 (Fig. 1g and Supplementary Table S3). Specific species were either enriched or depleted in a certain disease(s), regardless of geographical differences. For instance, in individuals with fatty liver disease, *Gordonibacter urolithinfaciens* and *Allisonella histaminifomans* were depleted and enriched, respectively. We also found that species associated with low gene richness[18], such as *Clostridium bolteae,* were enriched in several diseases. Similarly, *Parvimonas micra* was enriched in colorectal cancers, regardless of geographical differences. Furthermore, *Akkermansia muciniphila* was often depleted in several diseases. *A. muciniphila* is associated with improved intestinal barrier function and its depletion suggests intestinal barrier disruption in these different diseases[19]. The analysis of the frequency of the enriched/depleted MSPs among diseased cohorts showed that there were common species that could initially disrupt the microbiome balance and cause gut dysbiosis or foster microbial symbiosis that promotes health (top-left and top-right boxes in Fig. 1h, Supplementary Table S4).

Interestingly, we observed that MSPs commonly depleted in diseases were highly country-specific, while commonly enriched MSPs were usually shared by several diseases and were less related to geographical origins (Extended Fig. 1e-i and Supplementary Table S4). Moreover, we observed that MSPs commonly enriched or depleted in diseases were associated with different temporal behaviours among healthy individuals, as detailed in the following sections.

Dynamic changes of gut microbiome composition; inflow and outflow species

We investigated the temporal changes in microbiome composition at the individual level by applying Markov chain models (MCMs) to the MSPs identified in the longitudinal cohort of 86 healthy Swedes (Methods). This analysis enabled us to estimate the transition probability of individual MSPs from presence to absence (outflow probability) and vice versa (inflow probability) across different sampling times. We identified two groups of MSPs preferably transiting from presence to absence or from absence to presence; for brevity, we term them “outflow species (OFS)” and “inflow species (IFS)” respectively (Fig. 2a, Extended Fig. 2a-f, Supplementary Table S5). Clearly, declaring a species absent or present depends on the detection threshold, which is in turn determined by sequencing depth. We performed the analysis at three depth levels for 15, 10, and 5 million reads, and observed largely concordant results (Extended Fig. 3). For instance, 35 IFS (90%) were detected at both 10 and 15 million reads levels, while 4 and 6 species were detected only at former and latter, respectively. Similar results were observed for OFS: 447 (88%) were detected at both levels, while 62 and 27 species were detected at 10 and 15 million reads only. Overall, inflow and outflow probabilities were highly correlated at three different depth levels, with a slight reduction for outflows at 5 million reads (Supplementary Table S5).

To determine the robustness of these findings, we compared species-retaining probability (Kaplan-Meier estimates) with outflow probability, expected to be inversely proportional (Extended Fig. 4a and Methods). We observed that the species that had lower outflow probability, such as *Bacteroides vulgatus* and *Prevotella copri,* indeed had the highest retaining probability, whereas those of higher outflow probability (e.g. *Veillonella infantium and Streptococcus parasanguinis)* had reduced retaining probability. The

species retaining probabilities were correlated with their mean abundance (Extended Fig. 4b), even if the associations did not appear significant for any individual species based on Cox regression (p-values > 0.1, Extended Fig. 4c).

We observed that the changes of OFS abundances among 86 individuals (Δ_{OFS}) were inversely correlated with those of IFS (Δ_{IFS}), i.e. suggesting competition between IFS and OFS (Spearman's correlation = -0.334, p-value = 4.6×10^{-8}; Fig. 2b-c, Extended Fig. 2g). A higher abundance of OFS increased the gut microbiome imbalance between different visits, as the similarity of the gut microbiome compositions between the visits was decreased in 31 OFS-enriched individuals compared to 37 OFS-depleted individuals (Wilcoxon one-sided test p-value < 0.01; Fig. 2d). IFS-enriched individuals maintained their gut microbial composition such that it was similar between different time points (Extended Fig. 2h). Interestingly, increasing abundance of IFS was associated with increasing gene richness, known to be related to better health[18], suggesting that IFS may be beneficial (Spearman's correlation = 0.206, p-value = 9.0×10^{-4}; Fig. 2e and Extended Fig. 2i).

We hypothesised that IFS and OFS may differ in their growth rates, the former outgrowing the latter, and tested this hypothesis in three ways. First, we estimated species growth rates from metagenomic samples by Growth Rate InDex (GRiD) analysis [20] (Methods). For that, we stratified individuals into groups, enriched in IFS or OFS and found that in both groups GRiD scores of IFS were higher than OFS; they were the highest among IFS-enriched group (Fig. 2f-g). Second, we assessed species growth rates in bioreactors inoculated with healthy human stool samples, via GRiD analysis (Fig. 2h-i, Methods). We observed that the growth of IFS increased significantly over 24 hours, whereas that of OFS did not change, demonstrating that IFS could outgrow the OFS. Third, we used genome scale modelling to simulate species growth rates[21-24] (Methods) on four different putative diets (high fibre and high protein for plant- and animal-based diets) for highly prevalent 34 OFS and 30 IFS. The predicted growth rates of the IFS were significantly higher than of OFS (Fig. 2j-k, Supplementary Table S6). Furthermore, the reaction essentiality analysis indicated that the outflow GEMs were significantly dependent on the substrate, often displaying amino acid auxotrophy (Supplementary Figure 3 and 4). We suggest that the differences in growth rates and substrate dependence between IFS and OFS could underlie the directionality of the gut microbiome dynamics we report.

To test whether the inflow and outflow assignments of species deduced from the analysis of the four time points in our cohort persist over time, we collected and analysed two additional time points with 6 months intervals from the same individuals (Fig. 2l-m). Furthermore, to examine whether the assignments defined from a Swedish study are also found in other, geographically different regions, we analysed two publicly available cohorts, from Italy and US[9,25](Fig. 2n-q). In all cases, for both IFS and OFS, the transition probabilities were well correlated with those found for the first time points of our longitudinal cohort (Spearman's correlation coefficients > 0.56) for all comparisons). We conclude that IFS and OFS are largely conserved and are thus a global feature of the human gut microbiome.

Enrichment of IFS and OFS species determines richness, dysbiosis, and host physiology

To further explore links between gene richness and IFS and OFS, we determined the gene richness distribution of HGMA samples and grouped healthy samples as either high or low gene richness (HGR and LGR) based on the top and bottom 25% gene counts of HGMA samples, respectively (Extended Fig. 5a-c). We then identified species enriched in HGR and LGR by comparing MSP abundances: total 759 MSPs were enriched in HGR and 95 MSPs were enriched in LGR ($|\text{Log}_2$ fold change$| > 2$, Wilcoxon rank two-sided test adjusted p-value $< 10^{-3}$) (Supplementary Table S7). Interestingly, LGR-enriched MSPs were significantly enriched in OFS (i.e. higher outflow probability), while no enrichment was observed for HGR-enriched MSPs (Fig. 3a-b). Low gene richness was previously associated with a risk of developing chronic diseases related to the metabolic syndrome; the enrichment of disease associated OFS in LGR individuals was thus coherent with that observation[18]. We observed similar species enriched with low gene richness and associations with metabolic phenotypes (Extended Fig. 5d).

To investigate the impacts of IFS and OFS species on host physiology, we next examined the association of IFS and OFS with clinical parameters among healthy individuals (Supplementary Table S8). We traced the changes in 40 clinical parameters of 86 Swedish individuals (Supplementary Table S9) and linked them with the abundances of IFS and OFS species using linear mixed effect models (Methods). We found that IFS abundances were associated with muscle composition (p-value = 0.018, Fig. 3c), thus showing associations with microbial core metabolism such as amino acid metabolism, whereas OFS abundances were associated with a more diverse spectrum of clinical parameters, such as blood glucose, urate, B-type natriuretic peptide (BNP), ApoA1, and erythrocyte particles (p-value < 0.05). Interestingly, the OFS abundances were positively associated with BNP as a key heart failure marker (Extended Fig. 6c-g**,** Supplementary Table 10).

Next, we associated the common depleted and enriched species in diseases **(**Fig. 1h**)**, from Pan-MGAS analysis of 18 diseases**,** to the IFS and OFS. The commonly depleted MSPs had a greater tendency to be IFS (Fig. 4a-b). On the contrary, the commonly enriched species (top-right, Fig. 1h) were likely to be OFS, compared to the depleted species across all diseases. Out of the 23 commonly enriched species in diseases (enriched in at least 3 different disease cohorts), 14 (61%) were OFS species, significantly more than the OFSs species within all species detected in the cohorts (36%) (*Chi-square test, p-value =* 0.015, Supplementary Table S4). In addition, among the 14 OFS species, 5 (36%) were opportunistic pathogens reported previously [26] (*Chi-square test, p-value =* 10^{-9}). These observations support the view of microbiome dynamics lowering the abundance of potentially harmful species in healthy adults.

Functional understanding of region-enriched species, IFS and OFS

Our functional analysis indicated that the IFS species were enriched in core metabolism, essential for energy homeostasis and for biosynthesis of macromolecules (i.e., amino acids, carbohydrates and fatty acids; Fig. 4c and Supplementary Table S11 and S12, Methods). They were also enriched in: (i) processes associated with increased survival, such as sporulation, cobalamin biosynthesis (CobS), and sirohydrochlorin cobaltochelatase (CbiK); (ii) secondary metabolites (bacteriocin); (iii) proteins related to starch and plant-based fibre use (CAZymes GT5, GH13, GH51); (iv) anaerobic phenotypes

(Supplementary Table S11). By contrast, the OFS species were enriched in accessory metabolism, such as biodegradation of xenobiotics (benzene, toluene, ethylbenzene, and xylenes - BTEX), paralleled by that of ABC transporters, possibly involved in the import of xenobiotics, suggesting that exposure to pollutants may promote their appearance (Fig. 4d, Extended Fig. 6a-b, and Supplementary Table S11). They were also enriched in (i) active sugar transport (i.e., phosphotransferase system (PTS); (ii) virulence factors (VFs) and trigger factors; (iii) putative competence protein ComGF and type IV secretion systems, the latter two being important mechanisms for horizontal gene transfer[27]. Finally, those microbes tended to be facultative anaerobes and microaerophiles.

Outflow enriched-functional clusters showed distinct links to gut microbiome dysbiosis

To better understand potential impact of IFS and OFS at the functional module level, we identified co-conserved functional clusters of microbiome by applying an unsupervised clustering approach on MSPs (Fig. 4e, Extended Fig. 7 and Methods). This analysis provided a better representation of microbial functions than single annotations or known pathway definitions (e.g. KEGG) (Extended Fig. 8). From the community detection algorithm, we identified 7,763 functional clusters, 6,297 singletons, and 591 representative clusters (Methods, Supplementary Table 13). For example, antimicrobial resistance and secondary biosynthetic genes were found to be singletons and not co-conserved with other functional genes. After excluding singletons and unreliable functional clusters detected in less than three species, we retained 591 representative clusters of microbial functions. One of the two largest clusters (CL-12 in Supplementary Table 13, named "*comm-cluster*" herewith) was over-represented among many commensal species, while the other (CL-10, named "*patho-cluster*") was enriched in a few pathobionts, such as *Klebsiella* spp., *Enterobacter* spp., and *E. coli*. Interestingly, the *comm-cluster* was enriched with genes involved in the biosynthesis of amino acids indicative of functions enriched in IFS species. In contrast, the *patho-cluster* was enriched in functions associated with the uptake of several substrates, again indicative of transporters enriched in OFS species. These included siderophore, ion, amino acid, and vitamin transport, thus competing with host and commensal bacteria. We also found other enriched-functional clusters, such as butyrate metabolism, propionate metabolism, vitamin B12, coenzyme metabolism, chemotaxis, ATPase, and mobile genetic elements (i.e., integrase and transposase) and the CRISPR-cas system (Fig. 4e); a number of these were correlated with phylum-level taxonomy (Extended Fig. 7c).

We next projected the functional clusters on enriched/depleted MSPs in HGMA disease cohorts (Fig. 4f and Supplementary Fig. 5: hypergeometric tests, p-value < 10^{-3}). We found that many of disease-enriched functional clusters were enriched in the OFS species, for example, isoprenoid biosynthesis, competence proteins for DNA transformation, key signatures of OFS species, virulence factors, and nutrient uptake (e.g. ascorbate and mannose). It has been previously shown that isoprenoid biosynthesis initiates the majority of secondary metabolism[28]. However, we found a few functional clusters associated with species depleted in diseases, such as the CRISPR-cas system (i.e., the bacterial immune system) and teichoic acid transport.

Discussion

We have performed a comprehensive integrative analysis of global and temporal gut microbiomes, and we provide an open access HMGA portal (http://microbiomeatlas.org). Confirming previous observations[29], we have described the gut microbiome regional specificity, which needs to be taken into account before using the gut microbiome for the stratification of patients or for designing intervention studies. Beyond previous observations, our function-based analysis indicates that the western-enriched bacteria might dominate the gut microbial community with the production of antimicrobial peptides and homoserine lactone, which may inhibit their competitors.

Previous studies reported the temporal stability of the gut microbiome composition in an individual[7–9], implying oscillations around an average value. Our integrative analysis of temporal microbiome changes in a longitudinal study of healthy individuals has shown the existence of directionality of compositional variations: there are two types of species with a tendency to either increase or decrease in abundance with time, termed inflow or outflow species, respectively. Importantly, outflow species include most of the known opportunistic pathogens, while inflow species are essentially devoid of them. Remarkably, our function-based analysis indicates that outflow species might have a negative impact on host physiology, as they have enriched accessory metabolism and secretion of virulence factors. Most interestingly, outflow species tend to be enriched in different diseases while, in contrast, the inflow ones tend to be enriched in healthy individuals. We suggest that the tendency for the former to decrease and the latter to increase in healthy individuals is a previously unrecognized facet of the gut microbiome homeostasis.

The outflow species tend to be facultative anaerobes and to have an oral origin (e.g., *Streptococcus* spp.). This observation suggests an increase in oral microbial transmission to the gut, possibly due to a decrease in gut microbiome resilience. Enrichment of oral species in the gut has been observed in several diseases[30,31], and we suggest that increased mouth to gut microbial flow could be one of the global features of dysbiosis.

We have described the temporal dynamics of the gut microbiome through the discovery of outflow and inflow species. The enrichment of inflow species in healthy populations could possibly be due to their involvement in the storage carbohydrate degradation, such as starch and fibre, accounting for their higher persistence. We consider two mechanisms of outflow species enrichment in disease. *First*, the outflow species were enriched in competence mechanisms, facilitating import of genetic elements such as AMR, possibly conferring selective advantage in the gut. Enrichment of drug efflux mechanisms in outflow species might also confer resistance to antibiotics and other medications used in disease treatments. *Second*, the outflow species may plunder nutrient uptake from the host and utilize simple monosaccharides to increase their abundance. Metabolic modelling indicates that IFS are favoured by common diets. It is known that short-chain fatty acids (SCFAs) have an important impact on microbe-microbe interactions and host physiology. However, we also observed that the *comm-cluster* that comprises the biosynthesis of amino acids and folate metabolism may have a larger impact on the resilience of the gut microbiome and host physiology. We observed that the enrichment of the *patho-*

cluster could gradually or instantly perturb the gut ecosystem and shift it to either short-term gut imbalance or a persistent dysbiotic state.

Finally, the integration of metagenomics data from a large number of studies spanning five continents provides valuable knowledge for researchers interested in the impact of the microbiome on individual health parameters. The open-access atlas will be updated routinely with the new publicly available gut metagenomics data, including the recently announced one million microbiome project aimed at providing comprehensive open-access metagenomics data from multiple research centres. In this manner, in-depth analysis of the impact of the gut microbiome on health and disease will be used to facilitate future studies to reveal the key role of the gut microbiome in human maintaining health.

Methods

Metagenomics species pan-genome (MSP) creation

1601 metagenomic samples used to build the Integrated Gene Catalog of the human gut microbiome (IGC2) were downloaded from the European Nucleotide Archive. Using the Meteor software suite[1], reads from each sample were mapped against the IGC2 catalog and a raw gene abundance table was generated. This table was submitted to MSPminer[2] that reconstituted 1,989 clusters of co-abundant genes named Metagenomic-Species Pangenomes (MSPs). Quality control of each MSP was manually performed by visualizing heatmaps representative of the normalized gene abundance profiles. In addition, MSPs completeness and contamination was assessed by searching for 40 universal single copy marker genes[3] and by checking taxonomic homogeneity.

MSP taxonomic annotation with phylogenetic tree

MSPs taxonomic annotation was performed by aligning all core and accessory genes against *nt* and NCBI WGS (version of September 2018 restricted to the taxa Bacteria, Archaea, Fungi, Viruses and Blastocystis) using *blastn* (version 2.7.1, task = megablast, word_size = 16)[4]. The 20 best hits for each gene were kept. A species-level assignment was given if more than 50% of the genes matched the RefSeq reference genome of a given species, with a mean identity ≥ 95% and mean gene length coverage ≥ 90%. The remaining MSPs were assigned to a higher taxonomic level (genus to superkingdom), if more than 50% of their genes had the same annotation.

40 universal phylogenetic markers genes were extracted from the MSPs with[5]. MSPs with less than 5 markers were discarded. Then, the markers were separately aligned with MUSCLE[6]. The 40 alignments were merged and trimmed with trimAl[7]. Finally, the phylogenetic tree was computed with FastTreeMP[8] and visualized with iTOL[9].

Phylogenetic placement was further used to improve and correct taxonomic annotation.

Wellness study population, sample collection, extraction, library prep and sequencing

The wellness study is an ongoing prospective cohort study based on the Swedish CArdioPulmonary bioImage Study (SCAPIS) with 30,154 individuals enrolled at ages between 50 and 64 years recruited from random sampling of the general Swedish population. A total of 101 healthy individuals were recruited in the study and followed longitudinally for two years. Examinations in SCAPIS include imaging to assess coronary and carotid atherosclerosis, clinical chemistry, anthropometry, and extensive questionnaires, as previously described[10]. All participants provided written informed consent. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

Total genomic DNA was isolated from 100-120 mg of faces using a repeated bead beating method. Briefly, faces samples were placed in Lysing Matrix E tubes (MP Biomedicals) and extracted twice in lysis buffer (4% w/v SDS; 500 mmol/L NaCl; 50 mmol/L EDTA; 50 mmol/L Tris·HCl; pH 8) with bead beating at 5.0 m/s for 60 s in a FastPrep®-24 Instrument (MP Biomedicals). After each bead-beating cycle, samples were heated at 95ºC for 5 min and then centrifuged at full speed for 5 min at 4ºC. Supernatants from the two extractions were pooled and a 600 µL aliquot from each sample was purified using the QIAamp DNA Mini kit (QIAGEN) in the QIAcube (QIAGEN) instrument using the procedure for human DNA analysis. Samples were eluted in 200 µL of AE buffer (10 mmol/L Tris·Cl; 0.5 mmol/L EDTA; pH 9.0).. 1 ug of extracted DNA from each faeces sample were prepared for sequencing using Illumina TruSeq DNA PCR-Free sample prep kit and sequenced paired-end, 125bp on an Illumina HiSeq 2500 sequencer.

Functional annotation of the gut gene catalog and MSP

IGC2 catalog was annotated for the Antibiotic Resistant Determinants (ARD) described in Mustard database (v1.0) (http://www.mgps.eu/Mustard/)[11]. Protein sequences were aligned against 9,462 ARD sequences using *blastp* 2.7.1+ (option -evalue = 10^{-5}). Best-hit alignments were filtered for identity ≥ 95% and bidirectional alignment coverage ≥ 90% (at query and subject level), giving a list of ARD candidates belonging to 30 families. Annotation of the carbohydrate-active enzymes (CAZymes) of the IGC2 catalog was performed by comparing the predicted protein sequences to those in the CAZy database and to Hidden Markov Models (HMMs) built from each CAZy family[12], following a procedure previously described for other metagenomics analysis[13]. Proteins of IGC2 catalog were also annotated to KEGG orthologous using Diamond (version 0.9.22.123)[14] against KEGG database (version 82). Best-hit alignments with e-value ≤ 10^{-5} and bit score ≥ 60 were considered. Proteins involved in virulence factors of PATRIC[15,16] were matched against IGC2[17] by BLASTP (best identity > 50%, e-value < 10^{-10}). Phenotype of MSP were manually checked and annotated based on JGI-GOLD phenotype (organism metadata)[18]. We identified biosynthetic genes of MSP with the use of standalone anti-SMASH program with minimal run option, focused on core detection modules (version 5)[19]. Loading antiSMASH into Amazon cloud computing (AWS) as docker image, we executed its mining process per MSP in a massive parallel setting.

Quality control/normalization of gene counts and species abundance profiling

We filtered out human reads and then mapped metagenomic data (Supplementary Table 1) on IGC2 catalogue of human gut metagenome by METEOR[1] and based on the aligned reads, we estimated the

abundance of each reference gene of the catalogue, normalizing multiple mapped reads by their numbers and summing up normalized counts for a given gene. Reducing the variability by sequencing depths, gene count values were downsized into 10 million reads per sample; and any samples less than 10 million mapped reads were excluded from our dataset. Normalized gene counts were used for the quantification of MSP abundance by R *momr* (*MetaOMineR*) package[20,21]. MSP abundances were estimated by the median abundance of the 25 marker genes representing the robust centroid of gene clusters of MSP. Sample metadata of all metagenomics data such as sequencing platform, geography, age, body-mass index, gender and the data source were provided under HGMA (http://microbiomeatlas.org).

Tracing the diversifications of healthy metagenomic samples of different geography

After the quantification and per-million scaling of MSP abundance profiles, we employed trajectory analysis in R *monocle* ver.2 package to identify how samples were clustered[22]. In short, we selected the species profiles of all normal samples from different geographical origins and reduced the sample profiles into two dimensions by advanced nonlinear reconstruction algorithm, *DDRTree*. Based on the reduced two-dimensional components, we presented how samples were closely clustered as branches in scatter plots. Based on reduced profiles, we also calculated centroids and standard deviations of samples of given countries, except Finland population in toddlers (2 years).

Identification of region-enriched species from geographically distinct cohorts

We selected healthy samples of 17 countries after excluding matched controls of two-year old subjects of Finland T1D cohort and redundant samples of subjects with multiple measurements (i.e. multiple visits). Among 17 countries, we estimated effect sizes for Wilcoxon signed rank (one-sided) tests[23] of different MSP abundances of two different countries. As one-sided tests were used, we set the lower bound of effect sizes as zero and the upper bound of effect sizes as one, avoiding negative and infinite values. Based on estimated effect sizes, we identified significantly enriched species having medium effect sizes of specific country (effect size ≥ 0.3), compared to six or more countries, and defined those species as “region-enriched” species.

Next we categories species if enriched in 1) European countries, 2) non-westernized countries, and 3) China/Japan/US and identified contrasted functions among those three clusters of countries by multivariate regressions as follows:

$$Y_i = E_i \beta_{Ei} + N_i \beta_{Ni} + C_i \beta_{Ci} + \epsilon$$

where Y_i indicates a function regard to species i like CAZyme, antibiotics resistance, anti-SMASH, and virulence factor (if a given function exists in species i, Y_i =1, otherwise Y_i = 0), ϵ indicates an intercept, β_{Ei}^2, β_{Ni}^2, *and* β_{Ci}^2 are regression coefficients for E_i, N_i, and C_i, respectively and E_i, N_i, and C_i are categorial variables that indicate the region-enrichment of species i:

E_i = 1 if $i \in$ species enriched in any of European countries, otherwise E_i = 0

N_i = 1 if $i \in$ species enriched in any of non-westernized countries, otherwise N_i = 0

C_i = 1 if $i \in$ species enriched in China, Japan, or US, otherwise, C_i = 0)

Functions significantly associated with enrichment of any of three geographical clusters were identified based on *F*-tests of regressions (p-value < 0.01; β_{Ei}, β_{Ni}, $\beta_{Ci} > 0$) and quaternary plots were shown based on squared regression coefficients (β_{Ei}^2, β_{Ni}^2, β_{Ci}^2) normalized by their total sum.

Modelling temporal changes of normal gut microbiota during a year

First, we chose samples with sequential visits of given subjects and counted presence/absence of all MSPs detected in samples. To decide detection limit here, we fitted all non-zero abundance of MSPs into gamma distribution after per-million scaling and log2-transformation using R *fitdistrplus* package. Based on estimated shape and rate parameters from fitted gamma distribution, we counted species presence only when its abundance exceeded a percentile (>1%) based on the gamma distribution. Presence/absence profiles were fitted into two-state Markov chain model (i.e. states of presence and absence) to estimate state transition probabilities between presence and absence (R *markovchain* package). We did not include species of 100% prevalence, i.e., *Blautia wexlerae* (msp_0076) to Markov chain model. Here we estimated inflow probability of state transition from absence to presence, and outflow probability of state transition from presence to absence. For the estimation of species-retaining probabilities, we modeled presence/absence profiles as “events” and estimated the retaining probability from the survival rates of Kaplan Meier estimates using R *survival* and *survminer* packages.

For the validation of inflow and outflow from same Swedish wellness cohort, we additionally followed the two more visits (by every three months) and processed metagenomics data of 67 subjects (134 samples) after excluding subjects of missing visits and low sequencing depth less than 10 million reads. For the validation of inflow and outflow from independent cohorts, we processed metagenomics data from Italy (DINAMIC cohort) and US (HPFS cohort) after excluding subjects of missing visits and low sequencing depth less than 10 million reads. In HPFS cohort, we only took six-months interval samples of individuals, excluding one-day interval samples. We counted presence/absence of MSPs from the abundance profiles in a similar way of calculation in Swedish wellness cohort, and calculated state transition probabilities between presence and absence (i.e. inflow and outflow) after fitting presence/absence profiles into two-state Markov chain model.

Based on estimated inflow and outflow probabilities, we identified IFS ($P_{inflow} > 0.3$, and $P_{outflow} < 0.3$) and OFS ($P_{outflow} > 0.3$ and $P_{inflow} < 0.3$) and calculated scaled abundance of IFS (Z_{IFS}) and OFS populations (Z_{OFS}) like below.

$$z_{ij} = \frac{A_{ij} - \mu_i}{\sigma_i}$$

$$Z_{IFS} \text{ or } Z_{OFS}(\mathrm{j}) = \frac{1}{\sqrt{n}} \sum_i z_{ij}$$

where i is a given MSP belonging to IFS or OFS, A_i is the abundance of species i, u_i is mean abundance of species i over all wellness cohort samples (344 samples), σ_i is the standard deviation of species i over all wellness cohort samples, j is a given sample of wellness cohort, and n is the total number of IFS or OFS. Based on scaled abundance of single MSP (z_{ij}), we calculated the aggregated z-score of all IFS species and OFS species (Z_{IFS} and Z_{OFS}, respectively) by summing scaled MSP abundance for n species, where Z_{IFS} and Z_{OFS} follows standard normal distribution, independent of n value[24].

Microbial functions associated with IFS and OFS

Inflow/outflow scores of MSPs were tested their associations with function/phenotype annotations of given MSPs (i.e. presence/absence of functions) using univariate linear regressions. We selected significant associations of microbial functions to inflow/outflow scores when adjusted p-values of predictor variables (i.e. microbial functions) < 10^{-3} and regression coefficients > 0.

Associations between MSP abundance profiles and clinical metadata

Scaled abundance of IFS and OFS species populations together (Z_{IFS} and Z_{OFS}, respectively) were tested their associations with clinical parameters with considering random effects of individuals by linear mixed-effect models using R *lme4* packages (p-values < 0.05) like below:

$Y_i = Z_{IFS}\,\beta_{IFS} + Z_{OFS}\,\beta_{OFS} + u_i + \epsilon$

where Y is clinical parameter, β_{IFS} and β_{OFS} are coefficients of fixed effect variables, Z_{IFS} and Z_{OFS}, respectively, u_i is a random intercept for subject i, and ϵ is residual.

In addition, we tested associations of single MSP with clinical parameters of given samples of wellness cohorts by linear mixed effect models like below:

$Y_{ij} = A_i\,\beta_i + u_j + \epsilon$, $i \in$ IFS or OFS, A_i = species abundance

where Y is clinical parameter, β_i is coefficient of fixed effect variable, A_i, u_j is a random intercept for subject j, and ϵ is residual. We identified significant associations between MSP abundance and clinical

parameters based on explained variance of fixed effect calculated using R *MuMIn* package (explained variance > 10%).

Fecal fermentation in ARCOL bioreactor

M-ARCOL is a one-stage fermentation system run under semi-continuous conditions that simulates the main physicochemical and microbial conditions encountered in the human colonic ecosystem[25]. It consists of pH and temperature controlled, stirred (400 rpm), airtight glass vessels inoculated with fecal samples from human volunteers and maintained under anaerobic conditions by the sole activity of resident microbiota. The set-up in this study consisted in a main bioreactor containing the luminal-associated microbiota and a connected glass compartment with mucin beads to simulate the mucus-associated microbiota. The system was operated to simulate the colonic conditions of healthy human adults as described earlier (temperature 37°C, pH 6.3, retention time 24 h)[25,26]**. T**he experiments were conducted in duplicate with fecal samples from two donors (one male and one female, ranging in age from 24 to 50 years, with no history of antibiotic or probiotic treatment 3 months prior the beginning of the study)[25]. Following fecal inoculation of the bioreactor, fermentations were conducted for a total duration of 9 days, including 1 day under fed batch and the following 8 days under semi-continuous mode. Samples were collected daily in the bioreactor.

In situ metagenomic measurement of growth rate by Growth Rate Index (GRiD) scores

The GRiD software (v1.3)[27] was used to calculate the growth rate index from the metagenomic samples from Swedish wellness cohorts and fecal samples inoculated into bioreactor and fermented after 24 hours. Briefly, this software calculates the growth rate by mapping the metagenomics reads to microbial genomes and calculates the coverage ratio between the origin and terminus of replication as a proxy for the growth rate. Since GRiD is sensitive to the representativeness and quality of the genome used, we created a GRiD custom database representative to the gut microbiota, consisting only of high-quality draft genomes from the MGnify database[28]. First, we matched the msp gene clusters to the MGnify genomes using a BLASTN procedure, with a 95% identity threshold. Then we kept only the MGnify genomes passing these criteria: (1) >=95% gene completion, <=5% contamination, (3) <=100 contigs. This resulted in a GRiD database of 36 inflow genomes (92% of all IFS) and 194 outflow genomes (38% of all outflow species). Finally, the GRiD growth rate values were considered only when: (1) the genome displayed at least 1X coverage in the metagenome (using the –c 1.0 parameter), (2) the genome displayed a species heterogeneity of <=0.3 (as recommended by the authors), in order to remove spurious growth rate index.

Reconstruction of Genome Scale metabolic Model (GEM) and constrained based modelling for inflow and outflow MSPs

We used the GEMs of 30 IFS species and 34 OFS species with high prevalence (≥ 10%) and taxonomy annotated at species-level (i.e. excluding unclassified MSPs) using our recently reconstructed GEMs[29]

and KEGG orthology (KO) annotation of the gut catalogue. The KO profile of each MSP were mapped into KBase metabolic model [30] as reference model to provide reaction profiles. Regarding the reaction profiles the context specific GEMs were reconstructed and the functionality of the models was checked based provided biomass objective function and the gap filling was done using the COBRA toolbox and the reference model. To investigate the response of the IFS and OFS MSPs to environmental changes and calculate the perturbations, we used four different diets i.e. high protein- and fibre- plant based diets and high-protein and fibre omnivorous diets. The composition of the diet was converted to mmol/gDW*hour for the simulation in anaerobic situation and the growth rate for each model were predicted for each diet using constraint-based modelling. To check the dependence of the IFS and OFS species to the compounds as input or medium and autotrophy, we performed an essentiality analysis in which the inability of each MSP to synthesize the metabolites was simulated by closing the corresponding exchange reactions; decreased growth rate shows the dependence of the MSP to the metabolites for growth.

Gene richness and species associated with high and low gene richness

Gene counts (i.e. mapped read counts) of all dataset samples were downsized into 10 million reads by R *momr* (*MetaOMineR*) package. Based on detected genes from downsized gene count profiles, we identified gene richness of given samples. We first examined top-25% and bottom-25% gene richness of all HGMA samples and by Wilcoxon two-sided tests we compared species abundance between two groups of healthy samples: high gene richness (HGR) group for samples < first quartile of richness (Q_1); and low gene richness group (LGR) for samples > third quartile of richness (Q3).

Pan-metagenomics association studies (Pan-MGAS)

First, we selected of healthy and disease samples without interventions and redundant measurement (i.e. multiple visits) and performed comparative analyses of chosen samples (number of selected samples were shown in Supplementary Table 1). We estimated the effect sizes of Wilcoxon signed rank (one-sided) tests for MSP enrichment and depletion in diseases, compared to healthy controls of given country[23] and identified we identified significantly enriched or depleted species having medium effect sizes (effect size ≥ 0.3). Manhattan plots of pan-MGAS based on effect sizes were plotted with R *qqman* package.

Unsupervised clustering of co-conserved functions of gut microbiota

We calculated Jaccard index among functional annotations to check how many species were sharing given a pair of functions together. We selected highly shared pairs of functions (Jaccard index >= 0.75) and merged into functional co-occurrence network using R *igraph* package[31]. Functional clusters within the network were identified by unsupervised community detection, short random work algorithm (*cluster_walktrap* function)[32,33] and identified singleton functions within the network as well. Among non-singleton functional clusters, we selected representative functional clusters if functions of given

functional clusters were found more than three species, thereby excluding functional clusters sparsely annotated over MSPs. Associated MSPs to functional clusters were chosen if given MSP covered more than 75% functions of given functional cluster.

Declarations

Data availability

All primary HGMA data together with the newly sequenced data are freely available without restrictions in the public open access database (https://www.microbiomeatlas.org) that is part of the Human Protein Atlas program (https://www.proteinatlas.org).

Code availability

The R package used to perform modelling temporal changes of microbiome for inflow and outflow analysis together with functional clusters including unsupervised clustering of co-conserved functions of gut microbiota can be found at our GitHub repository link: https://github.com/sysbiomelab/mPackage. The modeling of temporal changes can be applied

directly to any sets of longitudinal microbiome data. The functional cluster analysis can be applied on gene counts and species abundances. The other pipeline scripts for analysis are also publicly shareable and available upon reasonable request from the corresponding author.

Acknowledgements

This study primarily was supported by Engineering and Physical Sciences Research Council (EPSRC), EP/S001301/1, Biotechnology Biological Sciences Research Council (BBSRC) BB/S016899/1, Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Erling Persson Foundation. Additional funding was from the Metagenopolis grant ANR-11-DPBS-0001. DL and JP were supported by the Bio-Synergy Research Project (2012M3A9C4048758) of the Ministry of Science and ICT through the National Research Foundation. SL was supported by Global University Project, “GIST Research Institute (GRI) IIBR” grants funded by the GIST in 2021, and the Bio-Synergy Research Project (2021M3A9C4000991) and the National Research Foundation of Korea (NRF) grant (NRF-2021R1C1C1006336) of the Ministry of Science, ICT through the National Research Foundation. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (CDRF), Zoe Global Ltd and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust in partnership with King’s College London. We thank the entire staff of the MetaGenoPolis at INRAE, Human Protein Atlas program (HPA), Centre for Host-Microbiome Interactions, the Science for Life Laboratory, the National Genomics Infrastructure for providing assistance in massive parallel sequencing, and Swedish National Infrastructure for Computing at SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2020-5-222,

SNIC 2019/3-226, SNIC 2020/6-153 and King's College London computational infrastructure facility, Rosalind (https://rosalind.kcl.ac.uk) for high performance computing.

Author contributions

S.S., S.D.E. and M.U. conceived the project. S.L. and S.S. led the design and analysis of the data. S.L. developed the temporal pipeline, analysis and made the figures. L.E. and M.U. provided the wellness gut metagenomics samples. M.A., F.P., E.L. and S.D.E. generated the MSPs, performed quality check and taxonomy update. N.P. annotated the updated gut gene catalog. L.E.M and S.B.D performed the bioreactor fermentation experiment on healthy human stool samples. M.A. performed the GRiD analysis on bioreactor. G.B. applied metabolic models and performed simulations. M.A., V.M. and F.P. performed the analysis on the Italian and American cohorts for validation. N.B., C.P., S.V., D. R. and A.H. analyzed part of the data and prepared the materials for the HGMA. K.F. and F.J. developed the HGMA website. V.L. and B.H. annotated the gut catalog with new CAZymes. J.P. and D.L. annotated the secondary metabolites of the gene catalog. M.A. and G.B. contributed to testing the pipeline, statistical and functional analysis. S.S., S.L., M.U. and S.D.E wrote and drafted the manuscript. L.A.E, D.L.S, A.M., G.P. J.N. provided critical feedback on the data and manuscript. All authors read, edited and reviewed the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Correspondence and requests for materials should be addressed to S.S. or D.E. or M.U.

References

1 Pons, N. e. a. *a platform for quantitative metagenomic profiling of complex ecosystems.*, <http://www.jobim2010.fr/sites/default/files/presentations/27Pons.pdf> (2010).

2 Plaza Onate, F. *et al.* MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* **35**, 1544-1552, doi:10.1093/bioinformatics/bty830 (2019).

3 Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* **10**, 1196-1199, doi:10.1038/nmeth.2693 (2013).

4 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

5 Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7**, e47656, doi:10.1371/journal.pone.0047656 (2012).

6 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

7 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

8 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

9 Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242-245, doi:10.1093/nar/gkw290 (2016).

10 Bergstrom, G. *et al.* The Swedish CArdioPulmonary BioImage Study: objectives and design. *J Intern Med* **278**, 645-659, doi:10.1111/joim.12384 (2015).

11 Ruppe, E. *et al.* Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat Microbiol* **4**, 112-123, doi:10.1038/s41564-018-0292-6 (2019).

12 Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-495, doi:10.1093/nar/gkt1178 (2014).

13 Svartstrom, O. *et al.* Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation. *ISME J* **11**, 2538-2551, doi:10.1038/ismej.2017.108 (2017).

14 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

15 Mao, C. *et al.* Curation, integration and visualization of bacterial virulence factors in PATRIC. *Bioinformatics* **31**, 252-258, doi:10.1093/bioinformatics/btu631 (2015).

16 Gillespie, J. J. *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* **79**, 4286-4298, doi:10.1128/IAI.00207-11 (2011).

17 Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol* **18**, 142, doi:10.1186/s13059-017-1271-6 (2017).

18 Mukherjee, S. *et al.* Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Res* **47**, D649-D659, doi:10.1093/nar/gky977 (2019).

19 Blin, K. *et al.* antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36-W41, doi:10.1093/nar/gkx319 (2017).

20 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).

21 Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59-64, doi:10.1038/nature13568 (2014).

22 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982, doi:10.1038/nmeth.4402 (2017).

23 Fritz, C. O., Morris, P. E. & Richler, J. J. Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen* **141**, 2-18, doi:10.1037/a0024338 (2012).

24 Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1**, S233-240, doi:10.1093/bioinformatics/18.suppl_1.s233 (2002).

25 Deschamps, C. *et al.* Comparative methods for fecal sample storage to preserve gut microbial structure and function in an in vitro model of the human colon. *Appl Microbiol Biotechnol* **104**, 10233-10247, doi:10.1007/s00253-020-10959-4 (2020).

26 Thevenot, J. *et al.* Enterohemorrhagic Escherichia coli O157:H7 survival in an in vitro model of the human large intestine and interactions with probiotic yeasts and resident microbiota. *Appl Environ Microbiol* **79**, 1058-1064, doi:10.1128/AEM.03303-12 (2013).

27 Emiola, A. & Oh, J. High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nat Commun* **9**, 4956, doi:10.1038/s41467-018-07240-8 (2018).

28 Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **48**, D570-D578, doi:10.1093/nar/gkz1035 (2020).

29 Bidkhori, G. *et al.* The Reactobiome Unravels a New Paradigm in Human Gut Microbiome Metabolism. *bioRxiv*, 2021.2002.2001.428114, doi:10.1101/2021.02.01.428114 (2021).

30 Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* **36**, 566-569, doi:10.1038/nbt.4163 (2018).

31 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1-9 (2006).

32 Pons, P. & Latapy, M. in *International symposium on computer and information sciences.* 284-293 (Springer).

33 Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, doi:10.1126/science.aan2507 (2017).

Figures

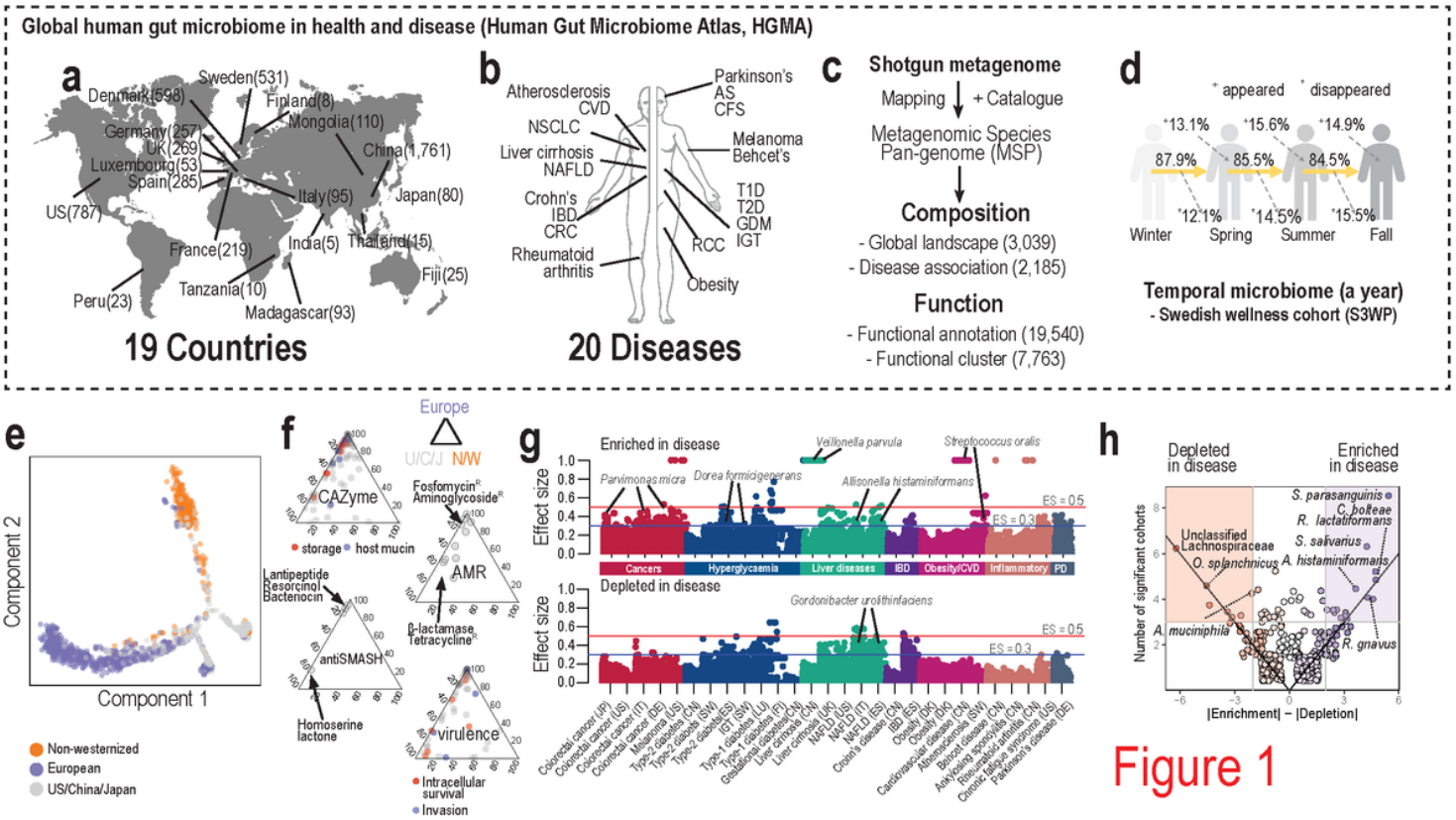

Figure 1

Characterization of the global gut microbiome in health and disease. Pan-metagenomics studies of health and disease. Corresponding datasets were publicly shared as a resource: human gut microbiome atlas (HGMA). a, geographical distribution of the datasets used in this study (the number of the samples is shown in parentheses). b, types of disease datasets of shotgun metagenomics used in this study. c, the workflow of the metagenomic species pan-genome (MSP) quantification together with functional characterization. In total, 5,224 shotgun metagenome samples, including 344 Swedish longitudinal samples, were aligned against the gene catalogue of the human gut microbiome and quantified at the level of MSP. All healthy samples (3,039) were used for the analysis of the global gut microbiome of healthy individuals, and all disease samples (2,185) were used for pan-disease analysis. For the functional characterization of human gut MSPs, we annotated respective genes with 19,540 features of microbial function/phenotype databases and identified 7,763 functional clusters better representing microbial functions. d, characterization of temporal changes of 86 Swedish healthy individuals (Swedish wellness cohort, S3WP) during a year (total 344 samples). e, monocle ordination of the gut microbiome in healthy samples. Individual samples from non-westernized countries, European countries, and US/China/Japan were coloured yellow, purple, and grey, respectively. f, contrasted functions among region-enriched species originated from three different geographical clusters, that is, non-westernized countries, European countries, and US/China/Japan. Based on functional annotations of CAZyme,

antimicrobial resistance (AMR), secondary metabolism (antiSMASH), and virulence factors (PATRIC database), we checked the enrichment of functions in a given geographical cluster. Enrichment of species with respective functions was checked by ANOVA of multiple regressions, considering association of geographical clusters as dependent variables of regressions. g, pan-metagenomics association studies (Pan-MGAS) of 28 cohorts from 18 different diseases and 11 countries (n=2,185). We identified significantly enriched/depleted species of cohorts based on effect sizes (ESs) of Wilcoxon one-sided tests (ES ≥ 0.3). We found species enriched with diseases in different countries such as A. histaminiformans (NAFLD), and species depleted, such as G. urolithinfaciens (NAFLD). h, Jitter plots of frequency of the significantly enriched/depleted cohorts of all MSPs were calculated: total frequency of enriched/depleted cohorts (|number of enriched cohorts| + |number of depleted cohorts|) and subtracted frequency between enriched cohorts and depleted cohorts (|number of enriched cohorts| - |number of depleted cohorts|). Point colours changed from red (left) to blue (right) according to x-axis values. Common enriched/depleted species among cohorts were identified when total frequency ≥ 3 and absolute subtracted frequency ≥ 2. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

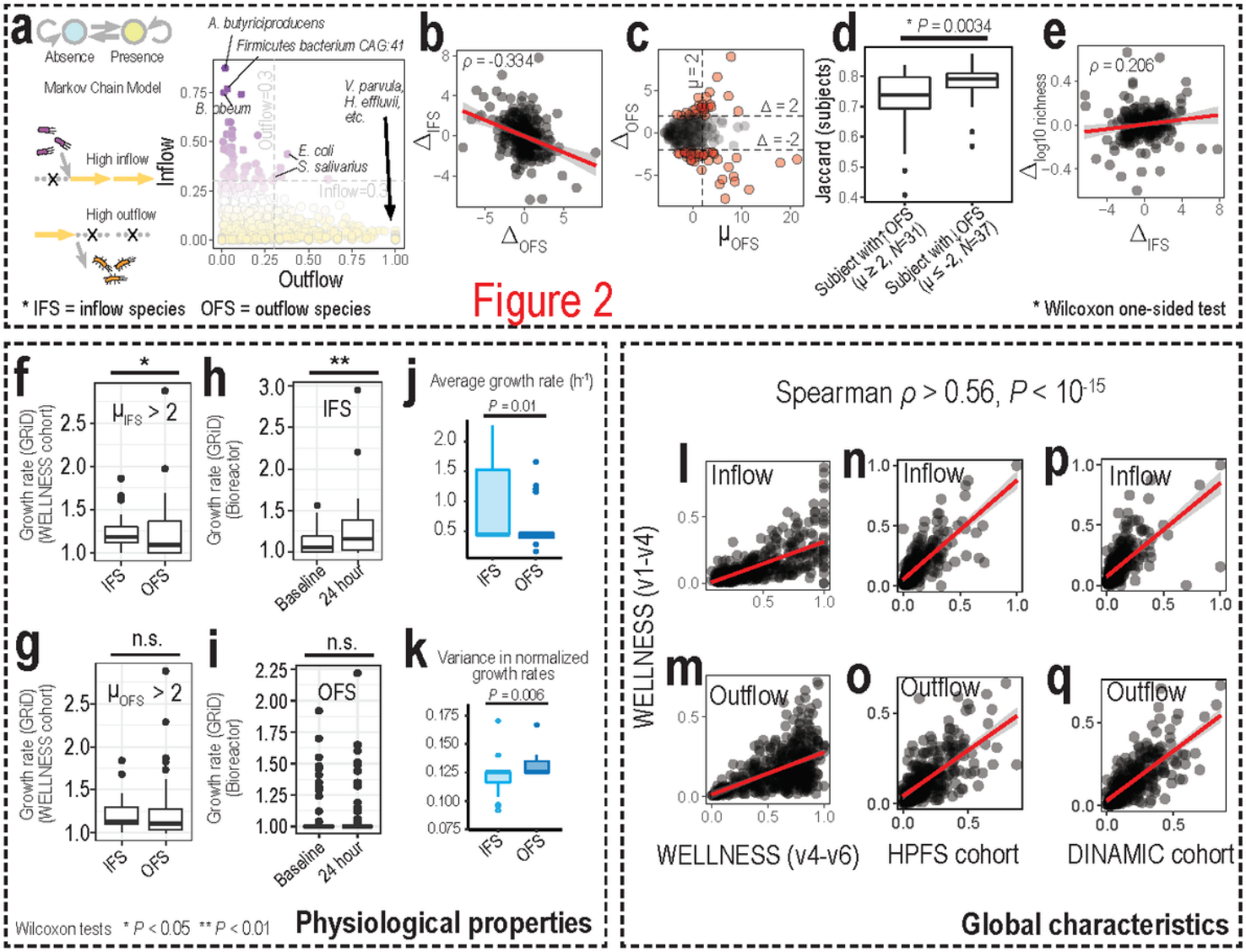

Figure 2

Changes in inflow and outflow species populations linked to gene richness and host physiology. a, modelling the temporal changes of species by Markov chain models (MCMs) estimating the state transition probabilities between absence and presence under detection limit (left top), inflow and outflow; dotted lines with a cross and gold arrows represent a failure and success to detect a species, respectively (left bottom). Inflow vs outflow score plot (right panel) identifies inflow species (IFS) (e.g. A. butyriciproducens and B. obeum), outflow species (OFS) (e.g. V. parvula and H. effluvia) or species that often change state between visits (E. coli, S. salivarius). b, we determined abundance changes between visits (t and t+1) and compared the total changes of IFS and OFS (ΔIFS and ΔOFS, respectively). In short, ΔIFS and ΔOFS were determined by differences of scaled total abundance between visits (i.e. Δ = Zt+1 - Zt) (See Methods). We observed negative correlation of abundance changes by visit between IFS species (ΔIFS) and OFS species populations (ΔOFS) (Spearman's ρ = -0.334, p-value = 4.6×10-8). c, we compared the mean abundance of OFS species in visits (μOFS) and abundance changes between visits (ΔOFS). The mean abundance of OFS species in visits was determined by the mean of scaled total abundance of

OFS species in two sequential visits (i.e. μOFS = ½ × (Zt + Zt+1)). The higher the mean OFS species abundance (μOFS ≥ 2), the more its abundance changes between visits either increasing (ΔOFS > 2) or decreasing (ΔOFS < -2). d, Individuals enriched with OFS species at certain visits (μOFS ≥ 2) had significantly lower intra-individual similarity (Wilcoxon one-sided test P = 0.0034). e, we observed that individuals with an increase in IFS species abundance were significantly positively correlated with an increase in richness (Spearman's ρ = 0.206, p-value = 9.0×10-4). We estimated physiological properties of IFS and OFS species by growth rate estimations by GRiD scores (f-i) and genome-scale metabolic modelling (j-k). We estimated GRiD scores of IFS and OFS species from (f) individuals highly enriched with IFS species and (g) individuals highly enriched with OFS species and observed higher GRiD scores of IFS species. In additional experiments of bioreactor fermentation of human faeces, we observed higher GRiD scores of IFS species after 24 hours, compared to original feces samples, whereas OFS species remained to be lower in GRiD scores at 24 hours. We evaluated our findings inflow/outflow scores estimated from Swedish cohort of four visits with those from (l-m) additional two more visits, (n-o) American HPFS cohort and (p-q) Italian DINAMIC cohort. We compared inflow scores (l, n, p) and outflow scores (m, o, q) between different datasets.

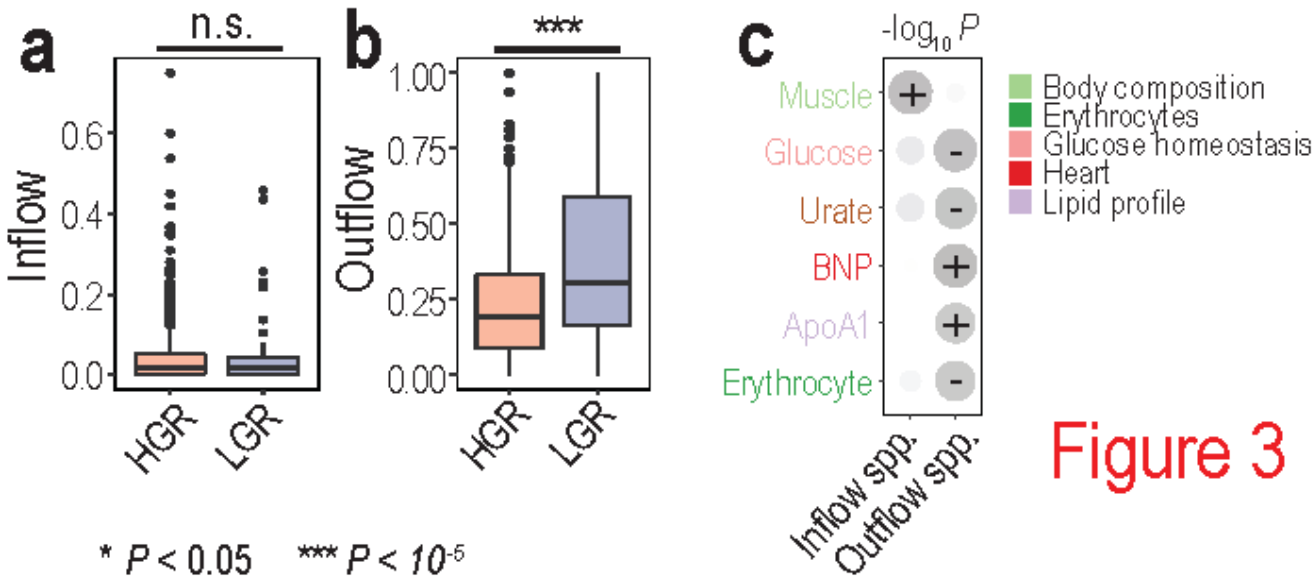

Figure 3

Inflow and outflow were associated with signatures of gene richness, and clinical parameters among healthy individuals. a-b, comparison of inflow (a) and outflow (b) probability between high gene richness (HGR)- and low gene richness (LGR)-enriched MSPs; LGR-enriched species have a higher probability of outflow (Wilcoxon one-sided test P = 1.1×10-6). c, IFS/OFS species abundances were significantly associated with clinical parameters by linear mixed effect models (p-value < 0.05). Significant positive (+) and negative associations (-) are marked on a heatmap (size proportional to significance). Increase of IFS

species abundance was associated with increase of muscle mass, whereas increase of OFS species was associated with increase of BNP, a heart failure marker.

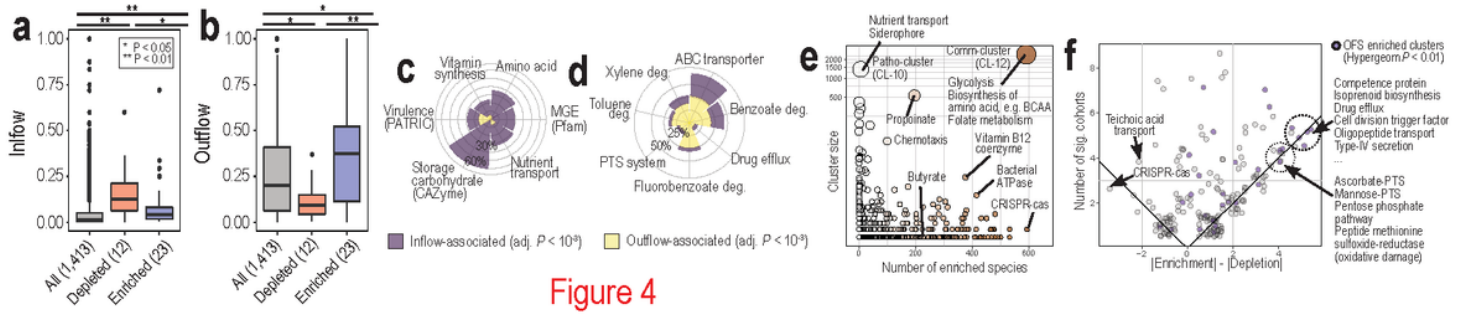

Figure 4

Figure 4

Inflow and outflow were associated with disease signatures from pan-disease analysis, together with biodegradation of xenobiotics and host nutrient hijacking. a-b, we observed significant differences in inflow (a) and outflow (b) probability between common enriched and depleted MSPs in diseases (from Fig. 1h) (Wilcoxon one-sided tests p-values < 0.05). we observed that common depleted species were higher in inflow, whereas common enriched species were higher in outflow. c-d, radar plots showing the fraction of functional classes enriched in either (c) inflow or (d) outflow, tested by linear mixed effect models (adjusted p-value < 10-3). The inflows were enriched in core metabolism and outflow in accessory metabolism (e.g., BTEX contaminants). e, we identified functional clusters co-conserved across different species and distinguished functional clusters found in many species or few species (x-axis). Y-axis indicates the size of functional clusters. Here we identified two largest clusters enriched in pathobionts (CL-10, patho-cluster) and commensal bacteria (CL-12, comm-cluster), respectively. f, frequency of functional clusters associated with enriched/depleted MSPs in diseases: total frequency of enriched cohorts and depleted cohorts per functional cluster (|number of enriched cohorts| + |number of depleted cohorts|, y-axis) and subtracted frequency between enriched cohorts and depleted cohorts per functional cluster (|number of enriched cohorts| - |number of depleted cohorts|, x-axis). Functional clusters significantly associated with outflow species (hypergeometric tests P < 0.01) were mostly associated with depleted MSPs in diseases (points coloured purple). Points were plotted in a jittered setting to avoid overlaid points.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesatlasversion121.xlsx
- SupplementaryFigure1atlasversion129.pdf
- SupplementaryFigure2atlasversion129.pdf
- SupplementaryFigure3atlasversion1291.pdf
- SupplementaryFigure4atlasversion129.pdf

- SupplementaryFigure5atlasversion1291.pdf
- ExtendedFigure1atlasversion129.pdf
- ExtendedFigure2atlasversion129.pdf
- ExtendedFigure3atlasversion129.pdf
- ExtendedFigure4atlasversion129.pdf
- ExtendedFigure5atlasversion129.pdf
- ExtendedFigure6atlasversion129.new.pdf
- ExtendedFigure7atlasversion129.pdf
- ExtendedFigure8atlasversion129.pdf