

Lineage-specific microbial protein prediction enables large-scale exploration of protein ecology within the human gut

Matthias Schmitz¹, Nicholas J. Dimonaco^{2,3,4,5}, Thomas Clavel¹, Thomas C.A. Hitch^{1,#}

¹Functional Microbiome Research Group, RWTH University Hospital, Aachen, Germany

²Department of Medicine, McMaster University, Hamilton, ON, Canada

³Farncombe Family Digestive Health Research Institute, McMaster University, Hamilton, ON, Canada

⁴School of Biological Sciences, Queen's University Belfast, Belfast, BT7 1NN, Northern Ireland, UK

⁵Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB, Wales, UK

Corresponding author: thitch@ukaachen.de

Abstract

Microbes use a range of genetic codes and gene structures, yet these are ignored during metagenomic analysis. This causes spurious protein predictions, preventing functional assignment which limits our understanding of ecosystems. To resolve this, we developed a lineage-specific gene prediction approach that uses the correct genetic code based on the taxonomic assignment of genetic fragments, removes partial predictions, and optimises prediction of small proteins. Applied to 9,634 metagenomes and 3,594 genomes from the human gut, this approach increased the landscape of captured expressed microbial proteins by 78.9%, including previously hidden functional groups. Optimised small protein prediction captured 3,772,658 small protein clusters, many with antimicrobial activity. Integration of the protein sequences and sample metadata into a tool, InvestiGUT, enables association of protein prevalence with host parameters. Accurate prediction of proteins is critical for understanding the functionality of microbiomes, hence this work will enhance understanding mechanistic interactions between microbes and hosts.

Introduction

Microbiome analysis has focused on the study of taxonomic groups, although it is the functionality of these taxa that is of interest^{1,2}. Microbes that perform the same function, or work together to form a functional group, are referred to as members of the same 'guild'³. Guild-based analysis of metagenomic datasets provides some functional insight, but represents an indirect method of studying functionality, as the unit of study remains the taxa, rather than directly studying the function of interest. Such inference of functionality from taxonomy is reductionist given the variability between strains of the same species, thereby expanding the functional capacity of microbial communities⁴.

Methods derived from functional ecology⁴⁻⁹ can be applied to the study of proteins and their functions. We term this 'protein ecology', which aims to study the ecological distribution of proteins or functions as the unit of study^{6,10}. The direct study of proteins/functions takes into account horizontal gene transfer, which can lead to the sharing of the protein of interest outside the taxa being studied, although transcriptional activity and genomic context can modify the functionality of a protein¹¹⁻

¹³. It has been shown that horizontal gene transfer occurs at high rates in the human gut microbiome, particularly in industrialised populations¹⁴, facilitating the spread of antibiotic resistance genes, virulence factors and other traits that affect the progression of human disease¹⁵.

Computational methods such as AnnoTree have been developed to study the distribution of functions across the microbial tree of life, highlighting functional variation between taxa¹⁶. Studying the ecology of proteins, rather than species, can provide insights into the ecological importance of protein-bearing species and the ecological pressures that drive protein evolution¹⁷. In the context of the human gut microbiome, species-independent functional studies have provided insights into the association of bile acid and sulphur metabolic genes with colorectal cancer¹⁸, and type 3 secretion system effectors with human health conditions^{19,20}.

Gene catalogues also place the proteins/genes as the unit of study by providing a reference source of non-redundant genes for either descriptive studies of a microbiome²¹, or targeted analysis of microbial functionality^{22,23}. So far, each gene catalogue has ignored the diversity of genetic codes used by bacteria^{24–26} and the existence of eukaryotic multiple exon genes²⁷. This is indicative of the trend in human gut microbiota research to focus on prokaryotes, neglecting other taxa²⁸. This is detrimental as eukaryotes and viruses directly influence human health, including via immune modulation²⁹.

We present a workflow using the taxonomy of metagenomic contigs to inform protein sequence prediction. Application of this approach to the human gut uncovered a multitude of previously missed proteins and provided an improved gene catalogue. To facilitate protein ecology studies of the human gut, we developed InvestiGUT, a tool that identifies associations between protein sequence prevalence and host parameters.

Methods

Selection of optimal gene prediction tools

To test the performance of gene prediction tools for large-scale metagenomic predictions, 13 different tools (**Supplementary Table 1**) were evaluated individually as well as in combinations of two or three. Among these, three tools were designed for eukaryotic sequences^{30–32}, nine focused on prokaryotes^{24–27,33–39}, and one for viruses⁴⁰. Both Prodigal, and Pyrodigal³⁹ (v2.1.0), an actively maintained successor to Prodigal²⁷ (v2.6.3), were included in the comparisons. However, as Pyrodigal was determined to give identical results to Prodigal, Prodigal was removed from the comparisons. Three of the eukaryotic tools (AUGUSTUS³⁰ v3.3, GlimmerHMM³¹ v3.0.4, and SNAP³² v2006-07-28) rely on the selection of a model before genes can be predicted. AUGUSTUS was run with the built-in *saccharomyces_cerevisiae*_S288C model on all eukaryotic genomes. Both other tools had no default fungal model, so a model was built based on *Saccharomyces cerevisiae* S288C (GCF_000146045.2).

The genomes and corresponding gene predictions included in the tool comparison are detailed in **Supplementary Table 2**. The bacteria studied included

the six Bacteria from the original ORForise comparison, supplemented with bacterial species deemed pathogenic^{41–45} or health-associated^{46–51}. In addition to the bacterial (n = 17) genomes, archaeal (n = 3), fungal (n = 3), and viral (n = 3) species (total species = 26) were included. Genomes were selected for strains which originate from the mammalian gut. Only the nuclear genomes of the fungi were included.

The under- and overprediction of the tools was quantified using ORForise⁵², which provides metrics on the performance of gene prediction software, by comparing predicted results to reference annotations (**Supplementary Table 3**).

Datasets

A total of 9,677 metagenomes from 43 studies were analysed (**Supplementary Table 4**). Among those, 7,735 metagenomic assemblies from 41 studies were available from Pasolli *et al.* (2019)⁸ and 1,326 assemblies from the Integrative Human Microbiome Project⁹. FASTQ files for 616 metagenomes from a Japanese cohort⁵³ were also downloaded and processed. Human sequences were removed from the latter metagenomic reads by running BMap⁵⁴ (v38.18) with Genome Reference Consortium Human Build 38 (GRCh38)⁵⁵ as a reference. Reads were assembled using MEGAHIT⁵⁶ (v1.2.9) with a range of k-mer sizes from 21 to 99 and a minimum count of 5 to filter out low-quality reads. Metagenomes with less than 1,000 proteins predicted were removed (n = 43), meaning 9,634 metagenomic samples were retained. Metadata for metagenomic samples was either obtained directly, or the curatedMetagenomicData R package⁵⁷. Additionally, a collection of 3,594 non-redundant high-quality genomes from Leviatan *et al.* (2022) were studied, this included both metagenomic assembled genomes (MAGs) and isolate genomes. These non-redundant genomes had a reported median completeness of 95% and median contamination of 0.67%⁵⁸.

Taxonomic assignment

For the taxonomic classification of all contigs, Kraken 2 (v2.1.2) was applied with a confidence threshold of 0.15⁵⁹. To enhance the taxonomy-assigned fraction of each metagenome, a custom Kraken 2 database was formed using all complete genomes for Archaea, Bacteria, Eukaryota, and Viruses from the NCBI RefSeq database (April 2023)⁶⁰. The inclusion of the human genome (Genome Reference Consortium Human Build 38 (GRCh38)) within the Kraken 2 database facilitated the removal of host contamination from downstream gene prediction. Based on the taxonomic ID assigned to each contig, the correct genetic code was assigned according to the NCBI Taxonomy Database (August 2023)⁶¹. For representative genomes, the taxonomic classifications of Leviatan *et al.* were computed with GTDB-Tk (v2.3.2; r214)^{58,62}. Contigs from each metagenome/genome were subdivided into files based on their inferred taxonomy and genetic code with the selected combination of three gene prediction tools being run on the corresponding subset files. Contigs without an assigned taxonomy were annotated with the bacteria-optimised selection of tools and genetic code 11.

Gene prediction

Variability in tool outputs was accounted for with a script to parse the start, stop, and strand information of prediction outputs. When multiple tools predicted genes with the same stop position, but different start positions, the longest sequence was selected. In cases of differing stop codons, all sequences were retained. Secondly, all partial genes were removed from consideration, including those with a start codon in the first or last three bases of a contig, where assembly errors are common⁶³, to prevent the inclusion of truncated genes. When possible, the minimum prediction length setting was set to 21 bases (six amino acids plus stop codon).

Based on this information, protein sequences were generated using the genetic code corresponding to the Kraken 2 taxonomy ID predictions. No protein count threshold was applied for the number of predicted genes in the studied representative genomes.

Protein ecology across human gut samples (InvestiGUT)

InvestiGUT (v0.1) accepts either a single protein sequence or a set of sequences, as defined by the options ‘-s’ or ‘-m’ respectively. User-provided protein sequences are matched to the collection of human gut microbiome proteins via DIAMOND (v2.1.8.162)⁶⁴ alignment with a default minimum query and subject coverage of 90% and identity of 90%, which can be defined by the user. User proteins can be analysed individually (-s), or as a group (-m) where all proteins must be present within a sample to be considered. The group analysis allows users to determine how frequently all enzymes in a certain pathway or all subunits in a protein complex are present. Output is divided into metagenomic and species-based analysis results, with both being available as raw data in the form of TSV files and as automatically generated vector figures.

Metadata for each metagenomic sample included age, gender, BMI, smoker-status, westernised-diet status, use of antibiotics, and geographical location. Additionally, disease prevalence includes output for Crohn’s disease, ulcerative colitis, colorectal cancer, *Clostridioides difficile* infection, type-2 diabetes, rheumatoid arthritis, and fatty liver disease. Metadata vocabulary is consistent with that used in the curatedMetagenomeData repository. Prevalence of the queried sequence in each of these groups is quantified, and compared either between categories, or healthy controls, using Fisher’s exact test from the scipy.stats⁶⁵ module in Python.

Querying the user input protein against the 3,594 nonredundant genomes provides a list of species which contain the protein, termed “function-positive species”. The taxonomic range of the protein is determined by the prevalence of the protein with each genus, family, and phylum of positive species. The function-positive fraction is obtained as the cumulative relative abundance of function-positive genomes across 4,624 individuals for which pre-computed relative abundance values are available. InvestiGUT is available at: <https://github.com/Matt-Schmitz/InvestiGut>

Gene catalogue creation

The protein sequences from the 9,634 filtered human gut metagenomes and 3,594 genomes were clustered using MMseqs2 (v14-7e284) linclust using options ‘--cov-mode 1 -c 0.8’ (minimum coverage threshold of 80% the length of the shortest

sequence), '--kmer-per-seq 80' (number of k-mers selected per sequence), and '--min-seq-id 0.9' to cluster at 90% protein identity. Thresholds were selected as previously used to generate the UHGP-90⁶⁶. Overlap between the MiProGut-90 and UHGP-90 was identified by combining the proteins of both catalogues and reclustering with the same options.

Functional annotation

The MiProGut gene-catalogue was annotated using eggNOG-mapper (v2.1.12) with the inbuilt DIAMOND blastp search option and default parameters^{67,68}. Further annotation was conducted with MANTIS using the integrated annotation file⁶⁹. When proteins had more than one functional category or general category letter association, the corresponding functional descriptions of each category were evenly weighted. Antimicrobial activity was predicted using the DBAASP toolkit⁷⁰. Positive activity against *Escherichia coli* ATCC 25922, *Pseudomonas aeruginosa* ATCC 27853, *Klebsiella pneumoniae*, *Staphylococcus aureus* ATCC 25923, *Candida albicans*, and *Saccharomyces cerevisiae* was determined as a predicted MIC <25 µg/ml. Annotation against the NCBI nr database was conducted using BLASTP with default options⁷¹.

Metatranscriptome expression analysis

DIAMOND (v2.1.8.162)⁶⁴ mapped metatranscriptomic reads^{7,72,73} against the MiProGut-90 and UHGP-90 protein catalogues using the 'blastx' command with options '--id 90 --evaluate 1e-6 -k 1 --max-hsps 1'. The aligned fraction was calculated per-sample as: (aligned reads / total reads) × 100. Expression of specific proteins was quantified as reads per kilobase per million mapped reads (RPKM) and calculated using: (aligned reads / ((gene length / 1,000)) × (total reads / 1,000,000).

Results

Lineage-specific gene prediction expands the human gut protein repertoire

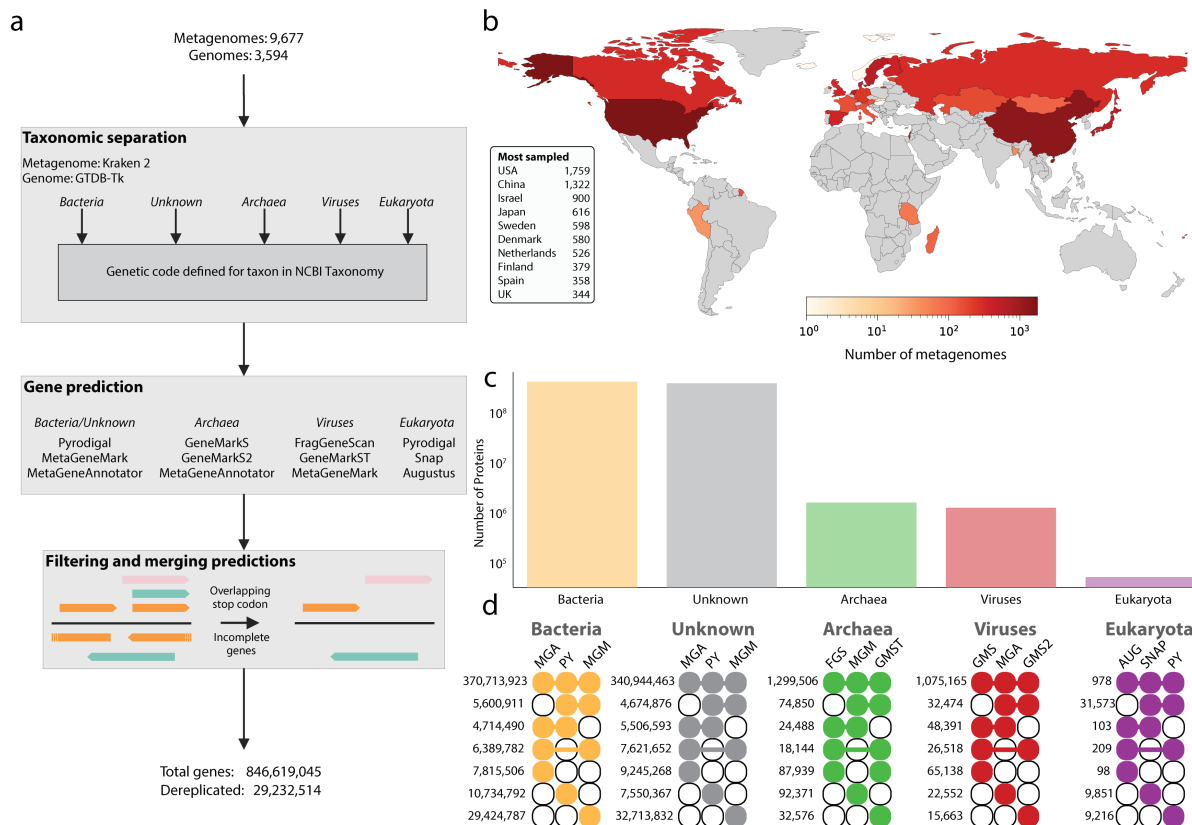


Figure 1: Taxonomy-informed gene prediction workflow and its application to the human gut. **a:** The workflow consists of three major steps; Taxonomic separation of either metagenomic contigs, or input genomes, into their respective domains. The genetic code utilised by each lineage is then identified at the family level, to account for variation within domains of life. For each domain-genetic code grouping, the respective gene prediction tools are run, and then merged, removing redundant predictions and incomplete gene predictions occurring at the edges of contigs. **b:** Geographical distribution of the human gut metagenomes used, with darker colours indicating a greater number of samples originate from the country. **c:** The number of proteins predicted for each taxonomic domain **d:** The overlap between gene prediction tools was identified for each domain to quantify each tool's variation from consensus predictions. Visualisation is based on a variation of an upset plot, where the number on the right is the number of overlapping protein predictions from the tools indicated by the connected coloured circles. Abbreviations for tools are: MGA = MetaGeneAnnotator; MGM = MetaGeneMark; FGS = FragGeneScan; GMS = GeneMarkS; GMS2 = GeneMarkS2; GMST = GeneMarkST; PY = Pyrodigal; SNAP = SNAP, AUG = AUGUSTUS.

A workflow for lineage-optimised gene prediction was developed using targeted selection of gene prediction tools based on the taxonomic assignment of each metagenomic sequence and customisation of each tool's options (genetic code, gene size) (**Figure 1a**). The selection of tools for gene prediction was informed by initial testing of 13 tools on archaeal ($n = 3$), bacterial ($n = 17$), fungal ($n = 3$), and viral ($n = 3$) species (total species = 26). The quality of each tool's annotations was quantified and compared using ORForise⁵². Given the variability of individual tools to predict all genes, we investigated the potential synergy of combining two or three tools in tandem to predict genes (**Supplementary Table 3**). Given the small, but consistent increase in both full and partial genes predicted by the combination of three tools, and the low cost in terms of spurious genes compared to two tools, we chose to use the combination of three tools that performed best for each taxonomic group.

This workflow was applied to 9,677 metagenomes from 28 countries (**Figure 1b**). In addition to the metagenomes, a non-redundant collection of genomes representing the prokaryotic diversity within the human gut was included for downstream analysis of the taxonomic occurrence of proteins⁵⁸. Taxonomic profiling of the metagenomic samples was consistent with previous observations that identified Bacteria as the dominant microbial domain in the human gut (**Supplementary Table**

5). The predicted proteins were dominated by those originating from bacterial contigs ($58.4 \pm 18.9\%$), followed by proteins on contigs that could not be assigned to a specific domain by Kraken 2, termed unknown ($41.2 \pm 18.8\%$), then viruses ($0.19 \pm 0.409\%$), archaea ($0.154 \pm 0.645\%$), and eukaryotes ($0.0320 \pm 1.31\%$) (**Supplementary Table 6**). The high percentage of taxonomically unassigned proteins, the unknown group, is in line with current estimates that $>50\%$ of the gut microbiota has yet to be cultured, hence complete genomes, which were used for taxonomic assignment, do not exist for many gut microbes^{74,75}.

In total, 846,619,045 genes (metagenomes: 838,528,977, genomes: 8,090,068) were predicted, with the majority originating from contigs of bacteria or unknown assignment (**Figure 1c**). In comparison, the exclusive use of Pyrodigal across all metagenomes identified 737,874,876 genes (metagenomes: 730,237,038, genomes: 7,637,838), meaning that the lineage-specific workflow identified an additional 108,744,169 genes (14.7%). While the use of multiple gene prediction tools might be expected to inflate the number of predicted genes, the majority of predictions were consistent across tools, except for eukaryotes. For eukaryotic contigs, AUGUSTUS³⁰ resulted in substantially lower gene prediction numbers, reducing the harmonious prediction of eukaryotic genes across all three tools (**Figure 1d**). In addition to the lower number of genes predicted by AUGUSTUS, the overlap between SNAP and Pyrodigal also highlights inconsistency. This is likely due to the inability of Pyrodigal to predict multi-exon genes, whereas both SNAP and AUGUSTUS predict exons and introns, which is crucial for accurate gene prediction in eukaryotes⁷⁶.

Enhanced coverage of the protein landscape from the human gut

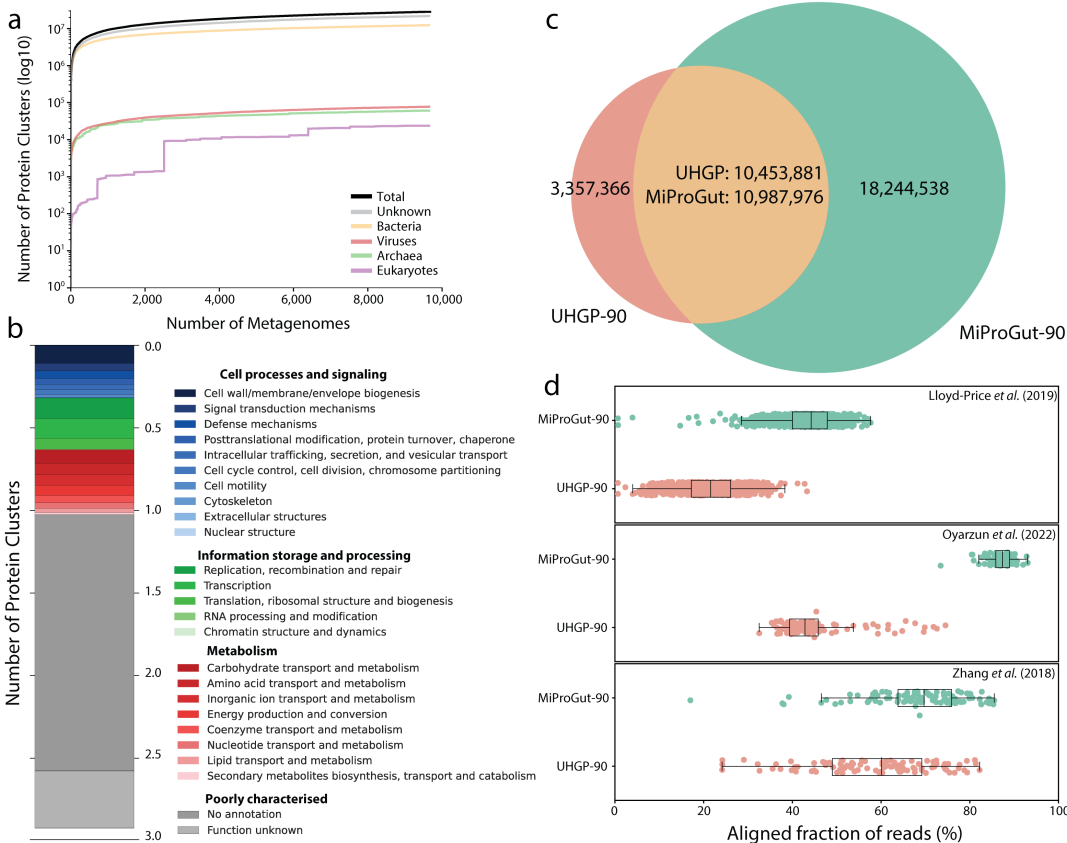


Figure 2: Taxonomy-informed gene prediction expands the protein landscape of the human gut microbiome. a: For each taxonomic group, as well as the cumulative protein repertoire of the of proteins within the MiProGut (log10 scale), refraction

analysis was conducted across all studied metagenomes. **b**: Functional annotation of the MiProGut based on COG functional groupings. **c**: Comparison of MiProGut (green) to UHGP (orange), both clustered at 90% amino acid identity. **d**: The aligned fraction of metatranscriptomic reads from multiple human gut studies (Lloyd-Price *et al.* 2019: n = 682, Oyarzun *et al.* 2022: n = 100, Zhang *et al.* 2018: n = 100) by both MiProGut and UHGP.

Lineage-specific gene prediction yielded a larger number of proteins compared to any single approach. Therefore, we sought to confirm that these were real proteins and not spurious predictions. To facilitate comparison with a previously established protein catalogue, the Unified Human Gastrointestinal Protein (UHGP) catalogue⁶⁶, we dereplicated our >800 million proteins at 90% similarity to 29,232,514 protein clusters, increasing the human gut protein landscape by 210.2% compared to UHGP. We termed this protein catalogue the Microbial Protein Catalogue of the Human Gut (MiProGut).

Nearly 10,000 samples were used to generate MiProGut, but the rarefaction of MiProGut suggests that greater diversity has yet to be captured for all taxonomic groups (**Figure 2a**). This was supported by analysis of 100 permutations where 2,397 samples were needed to cover 50% of proteins in MiProGut, and for 90% coverage an average of 7824 samples were needed. Eukaryotic proteins were observed least frequently, with the majority of these proteins being identified within a few samples that were dominated by Eukaryotic contigs (**Supplementary Table 5**). The Western bias of included samples, as well as low number of samples from developing countries may explain the rarefaction observation of further diversity. The inclusion of additional samples may also reduce the number of protein clusters consisting of a single protein sequence, known as singletons (n = 14,043,436). This frequency of singleton clusters is consistent with previous observations that most protein clusters are rare⁷⁷, leading to the formation of singletons. While singletons were rarely captured by metagenomic sequencing, expression in human gut samples was observed for 39.1% of singletons, confirming they are not spurious and are functionally relevant to the microbiota.

MiProGut provides an improved resource for protein sequence identification, but functional annotation is required to understand the role of these proteins, both in relation to the microbiome, and their impact on the host^{78,79}. The majority (64.4%) of MiProGut proteins lacked informative functional annotation when annotated with eggNOG-mapper (**Figure 2b**). Further annotation with MANTIS, which integrates multiple functional databases, increased the annotated fraction by 4.9% compared to all eggNOG assigned COG categories. However, this included assignments to general functions. These values re-highlight the high number of microbial proteins in the human gut that lack meaningful functional annotation⁷⁴.

Significant overlap was observed between MiProGut and UHGP, with 75.7% of UHGP proteins being covered by MiProGut, whereas only 37.6% of MiProGut proteins were covered by UHGP (**Figure 2c**). Compared to UHGP, MiProGut was enriched for eukaryotic functions, especially nuclear structure proteins (3.5 Vs 227.3 protein clusters respectively).

Identification of overlapping proteins between the two catalogues provides support that these proteins are not spurious. Expression was quantified in 862 metatranscriptomic samples from three studies (**Figure 2c**)^{7,72,73}. We identified transcriptional evidence for 13,756,670 MiProGut proteins (47.1%), compared to 7,689,106 UHGP proteins. Given the increased number of expressed proteins, we

calculated the aligned fraction of reads for each of the 862 metatranscriptomes. Across all samples, MiProGut covered an additional $22.9 \pm 9.9\%$ reads compared to UHGP. The largest increase occurred in a Spanish cohort where MiProGut increased read coverage by an additional $41.6 \pm 8.6\%$ ⁷².

Identification of commonly expressed small protein clusters within the human gut

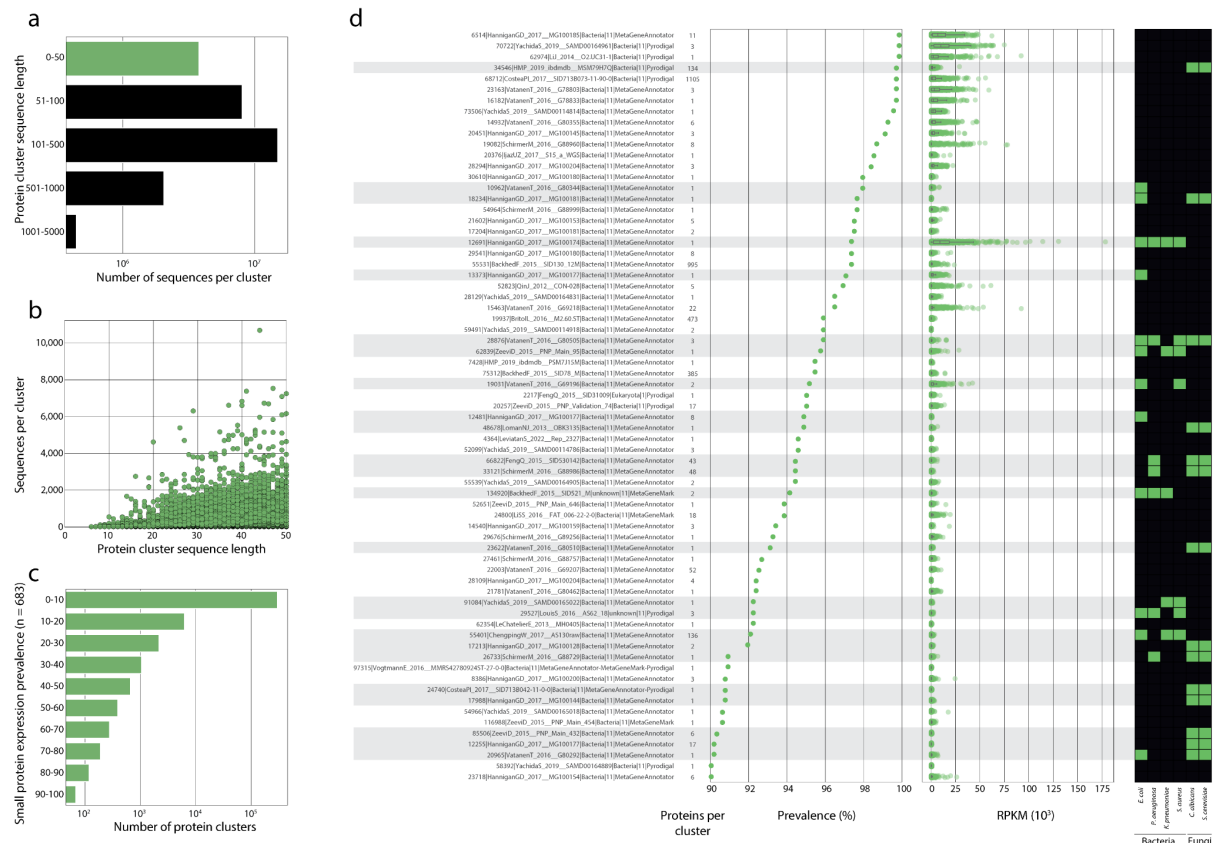


Figure 3: Identification of highly prevalent, and expressed small proteins within the gut and their antimicrobial potential. **a:** The size of protein cluster representative sequences, with small proteins (≤ 50 amino acids) in green. **b:** The number of members within each protein cluster, plotted against the size of each protein clusters representative sequence. **c:** The prevalence of each protein clusters expression across the 687 metatranscriptomic samples from Lloyd-Price *et al.* (2019)⁷. **d:** The name, number of member sequences, expression prevalence, RPKM and predicted antimicrobial activity for proteins expressed within $\geq 90\%$ of metatranscriptomic samples (n = 69).

Small proteins have been understudied in the human gut as many gene prediction tools exclude them by default⁸⁰. To improve the prediction of these proteins, we modified the parameters to predict proteins of >5 amino acids. This resulted in the prediction of 44,164,853 small proteins (5.2% of total proteins), represented by 3,571,095 protein clusters within MiProGut, called small protein clusters (SPCs) (Figure 3a). Of these, 1,104,393 (30.9%) clusters were singletons (Figure 3b), while the largest SPC contained 10,700 proteins and was identified as a member of the proposed 'family 350024' of crosstalk proteins⁸⁰. Studying the expression of the SPCs within 687 metatranscriptomic samples revealed that, while many were rarely expressed ($<10\%$ of samples), 69 SPCs were highly prevalent, being expressed in $>90\%$ of samples (Figure 3c).

The gastrointestinal tract has been identified as a rich source of small proteins with antimicrobial activity^{81,82}. To determine whether this included the most commonly expressed SPCs, we predicted each SPCs antimicrobial activities⁸³ (Figure 3d). By

predicting the antimicrobial activity against four ESKAPE pathogens and two fungi of relevance to human health, we found that 35% (24/69) of these SPCs could be antimicrobial. The most highly expressed antimicrobial SPC ($13,707 \pm 17,886$ RPKM) was predicted to be active against all four ESKAPE pathogens (*Escherichia coli* ATCC 25922, *Pseudomonas aeruginosa* ATCC 27853, *Klebsiella pneumoniae*, *Staphylococcus aureus* ATCC 25923). Although expressed in 97.1% of the samples, this cluster contained only a single sequence with similarity to an unknown protein from *Bifidobacterium* spp. The coding of this protein by a dominant genus of human gut commensals, in addition to its high expression and wide range of predicted activity, suggests this protein may be of importance to the gut ecosystem. Of the 24 antimicrobial SPCs, 14 were predicted to target both fungal species, *Candida albicans* and *Saccharomyces cerevisiae*.

Ecological distribution of proteins across individuals uncovers associations with host health status

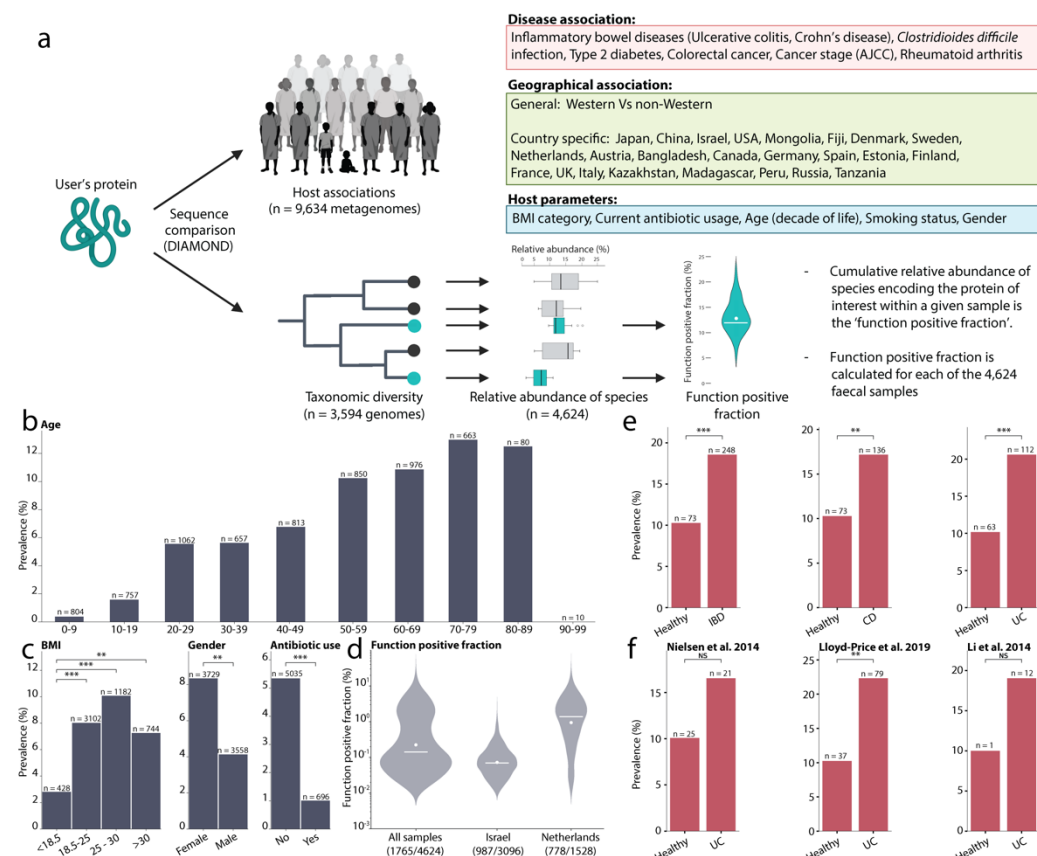


Figure 4: Application of protein ecology to the human gut identifies associations between microbial functions and host parameters. **a:** Workflow of InvestiGUT, detailing the host associations that are studied, as well as detailing the process for calculation of the function positive fraction. Functional positive species and their relative abundances are coloured turquoise, along with the final functional positive fraction. Methyl-co-reductase was used as a use-case of the multiple-protein option within InvestiGUT, identifying the complexes prevalence across host age (**b**), BMI, gender, and antibiotic use (**c**), and determine the MCR function positive fraction of the microbiome from 4,624 individuals (**d**). As a second use-case, the small antimicrobial peptide '55401|ChengpingW_2017__AS130raw|Bacteria|11|MetaGeneAnnotator' was studied. Significant associations with IBD, CD, and UC were observed across studies (**e**). To explore the association with UC further, the three individual studies into UC were examined, each showing increased prevalence in UC patients compared to healthy controls (**f**). Significance from Fisher's exact test, after Benjamini-Hochberg correction, are shown as: NS = not significant, * < 0.05, ** < 0.01, *** < 0.001.

The application of lineage-specific gene prediction enhances the ability to study the ecological distribution of proteins, referred to in this manuscript as protein ecology. Using the metadata associated with both the metagenomes studied and the genomes

analysed, we created a tool, InvestiGUT, to facilitate protein ecology studies of the human gut (**Figure 4a**).

InvestiGUT was developed to accept multiple sequences at once and examines them either individually or as a collection, where only samples that are positive for all queried sequences are reported. The input sequences are compared against all 846,619,045 predicted protein sequences, representing those from both the metagenomes and genomes. The sequence similarity for annotation is by default 90%, but can be defined by the user. Integration of metadata for the metagenomes is used to quantify the prevalence of the queried proteins with disease and geographic locations, as well as host parameters. The inclusion of genomes from human gut commensals allows the taxonomic range of the protein to be determined. Additionally, we can determine the 'function positive fraction', which represents the proportion of an individual's microbiota that is positive for the function/protein sequence of interest. This is calculated by determining which genomes encode the protein sequence of interest and then determining the cumulative relative abundance of these genomes across 4,464 individuals⁵⁸.

Methanogenesis in the human gut is restricted to archaea and has been linked to several human diseases⁸⁴. Methyl-coenzyme M reductase (MCR) catalyses the formation of methane in methanogens and consists of an alpha, beta, and gamma chain. Due to MCRs multiple chains, the multiple sequence approach was used in InvestiGUT, ensuring only those samples in which all three chains were detected to be studied. The prevalence of methanogenic functionality increased with host age (**Figure 4b**), confirming previous reports of this association⁸⁵. Increased methane production has been reported in individuals with higher BMI, supporting our observation that MCR was more prevalent in individuals with higher BMI⁸⁶ (**Figure 4c**). Variation with gender has also been previously observed, although age is a confounding variable in this association⁸⁵. The decreased prevalence of MCR in patients noted as having used antibiotics provides the first metagenomic insight into the observation that antibiotic therapy can eliminate methane in the breath of individuals⁸⁷. This may be due to indirect interactions, such as the dependence of methanogenic archaea on bacteria that are themselves susceptible to the antibiotics. The MCR positive fraction revealed a country-specific association, with individuals from the Netherlands having a higher relative abundance of species positive for MCR than individuals from Israel (**Figure 4d**).

Country-specific associations can be explained by variation in diet. For example, the enrichment prevalence of seaweed-degrading porphyranases (*Bp1689*), and agarases (*Bp1670*) in the gut of Japanese individuals, compared to North Americans². The original study was limited to 31 individuals (13 Japanese vs 18 North Americans), hence our analysis expands the study of these proteins across 28 countries. We confirmed that both proteins are most prevalent in Japanese samples (56.8% and 79.4% respectively), followed by China (18.5% and 27.4% respectively) (**Supplementary Figure 1**). Interestingly, 41 USA samples were positive for *Bp1670* and 39 for *Bp1689*, perhaps reflecting changes in diet, travel, or immigration over the last decade.

Given that small proteins are highly expressed in the human gut, we investigated the ecological distribution of the 69 most highly expressed SPCs. Of

these, we observed that 7 were enriched in inflammatory bowel disease (IBD), 9 in ulcerative colitis (UC), and 8 in Crohn's disease (CD). In particular, '55401|ChengpingW_2017__AS130raw|Bacteria|11|MetaGeneAnnotator' showed a significant change in prevalence with all three conditions across the studies (**Figure 4e**). When we explored this association further, we found that the association with UC was only significant in one of the three studies, although the increased prevalence in UC patients was observed in all three studies (**Figure 4f**). Analysis of the 69 SPCs expression in patients with UC and Crohn's disease CD patients compared to control patients identified that 31 and 34 proteins respectively were enriched in IBD sub-types, including '55401|ChengpingW_2017__AS130raw|Bacteria|11|MetaGeneAnnotator' (**Supplementary Table 7, 8**). This association may be explained by the predicted antimicrobial activity of the protein, which targets *E. coli*⁸⁸, *K. pneumoniae*⁸⁹, and *S. aureus*⁹⁰, all of which have been studied as causative agents of colitis.

The largest SPC (n = 10,700) was also of interest due to its prevalence. It contained a highly conserved core sequence across all sequences, forming a continuous helix (**Supplementary Figure 2a**). The taxonomic range of this small protein was restricted to the *Bacteroidaceae*, in which it is prolific, with 76.3% (45/59) of *Bacteroides* spp. containing a matching protein sequence (**Supplementary Figure 2b**). Functional annotation of the protein suggested that it has antimicrobial activity against *S. aureus* (**Supplementary Figure 2c**). The prevalence of this protein was associated with Westernised individuals and decreased with antibiotic use (**Supplementary Figure 2d**). These use cases highlight InvestiGUT's ability to facilitate protein ecology studies, confirm associations previously identified in smaller cohorts, and identify associations uncovered by integrating multiple data sources.

Discussion

Integrating the taxonomic assignment of a sequence into the choice of gene prediction tools, and optimising the parameters used, revealed millions of protein sequences previously missing from human gut gene catalogues. These included proteins specific to groups that have previously been overlooked, such as the eukaryotes, archaea and viruses²⁸. The lack of functional annotation by both MANTIS and eggno-mapper for 47.9% of the proteins is consistent with previous observations that nearly half of the identified proteins in the human gut have yet to be characterised⁷⁴. Integration of metatranscriptomic datasets facilitated the identification of highly expressed small proteins with potential antimicrobial activity against key human pathogens. Further analysis of these proteins revealed that the majority (49/69) were differentially expressed between IBD subtypes and non-IBD patients, including those with predicted antimicrobial activity. The association with human health conditions, as well as their antimicrobial activity, warrants further investigation of these proteins to characterise their impact on the microbiota and potential therapeutic application.

The increased coverage of the human gut protein landscape by MiProGut, compared to other gene catalogues, facilitates greater analysis of omics datasets, doubling the aligned fraction in a metatranscriptomic study. As gene catalogues are often used as reference databases for the metagenomic study of an environment, the discovery of these missing proteins will facilitate the identification of disease-specific

biomarkers that were previously overlooked. However, the bias of metagenomic samples included towards Westernised countries (USA, China) suggests that the human gut landscape could be further enhanced with the inclusion of additional samples from underrepresented countries and age groups (infants, elderly)²¹.

We present the concept of protein ecology, which focuses on the study of proteins rather than taxonomic species and should be applied more broadly to the study of the human gut. InvestiGUT facilitates protein ecology studies by providing statistical analysis of the prevalence of a protein sequence/sequences of interest across the samples examined in this work. The application of this approach identified both previously observed associations, but across a larger number of samples, providing independent validation of these associations. In addition to protein prevalence, the relative abundance of microbial species containing the queried protein, referred to as the 'function positive fraction', is of interest for protein ecology, as it allows the estimation of protein distribution within the ecosystem. The current genome collection and approach does not account for strain-level diversity, instead assuming that the presence within the reference genome is representative. Currently, this limits our ability to truly capture the gut microbiotas diversity. Future integration of a larger collection of genomes that captures the strain-level diversity of each species would facilitate the calculation of a modified functional positive fraction.

Data availability

Gene and protein predictions, along with the nonredundant MiProGut catalogue, are available at: <https://zenodo.org/doi/10.5281/zenodo.10988030>. InvestiGUT is available at: <https://github.com/Matt-Schmitz/InvestiGut>

Funding

TCAH received funding from the RWTH Start-up grant, titled ProtoBIOME and University Hospital of RWTH Aachen START grant, titled LeakyGut (021/23). TC received funding from the German Research Foundation (DFG): project no. 403224013 – SFB1382 Gut-liver axis, subproject Q02; project no. 460129525 – NFDI4Microbiota.

Acknowledgements

The authors would like to thank Susan Jennings and Charlie Pauvert for reviewing the paper prior to submission, and Amy Coates for designing the logo.

References

- 523 1. Haiser, H. J. *et al.* Predicting and Manipulating Cardiac Drug Inactivation by the Human
524 Gut Bacterium *Eggerthella lenta*. *Science* **341**, 295–298 (2013).
- 525 2. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to
526 Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
- 527 3. Wu, G., Zhao, N., Zhang, C., Lam, Y. Y. & Zhao, L. Guild-based analysis for understanding
528 gut microbiome in human health and diseases. *Genome Med.* **13**, 22 (2021).
- 529 4. Afrizal, A. *et al.* Enhanced cultured diversity of the mouse gut microbiota enables custom-
530 made synthetic communities. *Cell Host Microbe* **30**, 1630-1645.e25 (2022).
- 531 5. Mazmanian, K., Sargsyan, K. & Lim, C. How the Local Environment of Functional Sites
532 Regulates Protein Function. *J. Am. Chem. Soc.* **142**, 9861–9871 (2020).
- 533 6. Konopka, A. What is microbial community ecology? *ISME J.* **3**, 1223–1230 (2009).
- 534 7. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel
535 diseases. *Nature* **569**, 655–662 (2019).
- 536 8. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over
537 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*
538 **176**, 649-662.e20 (2019).
- 539 9. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human
540 Microbiome Project. *Nature* **569**, 641–648 (2019).
- 541 10. Hitch, T. C. A. *et al.* Microbiome-based interventions to modulate gut ecology and the
542 immune system. *Mucosal Immunol.* **15**, 1095–1113 (2022).
- 543 11. Nagy-Staron, A. *et al.* Local genetic context shapes the function of a gene regulatory
544 network. *eLife* **10**, e65993.
- 545 12. Eguchi, Y., Bilollikar, G. & Geiler-Samerotte, K. Why and how to study genetic changes
546 with context-dependent effects. *Curr. Opin. Genet. Dev.* **58–59**, 95–102 (2019).

13. Bedhomme, S. *et al.* Evolutionary Changes after Translational Challenges Imposed by Horizontal Gene Transfer. *Genome Biol. Evol.* **11**, 814–831 (2019).
14. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053–2067.e18 (2021).
15. Zhou, H., Beltrán, J. F. & Brito, I. L. Functions predict horizontal gene transfer and the emergence of antibiotic resistance. *Sci. Adv.* **7**, eabj5056 (2021).
16. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).
17. Pold, G. *et al.* Phylogenetics and environmental distribution of nitric oxide-forming nitrite reductases reveal their distinct functional and ecological roles. *ISME Commun.* **4**, ycae020 (2024).
18. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
19. Wolf, P. G. *et al.* Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer. *Microbiome* **10**, 64 (2022).
20. Young, V. *et al.* A gut meta-interactome map reveals modulation of human immunity by microbiome effectors. 2023.09.25.559292 Preprint at <https://doi.org/10.1101/2023.09.25.559292> (2023).
21. Zeng, S. *et al.* A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat. Commun.* **13**, 5139 (2022).
22. Pellock, S. J. *et al.* Discovery and Characterization of FMN-Binding β -Glucuronidases in the Human Gut Microbiome. *J. Mol. Biol.* **431**, 970–980 (2019).

23. Pollet, R. M. *et al.* An Atlas of β -Glucuronidases in the Human Intestinal Microbiome. *Struct. Lond. Engl.* 1993 **25**, 967-977.e5 (2017).
24. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
25. Besemer, J. & Borodovsky, M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27**, 3911–3920 (1999).
26. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
27. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
28. Guzzo, G. L., Andrews, J. M. & Weyrich, L. S. The Neglected Gut Microbiome: Fungi, Protozoa, and Bacteriophages in Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **28**, 1112–1122 (2022).
29. Li, X. V. *et al.* Immune regulation by fungal strain diversity in inflammatory bowel disease. *Nature* **603**, 672–678 (2022).
30. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
31. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
32. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
33. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191 (2010).

34. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).
35. Lomsadze, A., Gemayel, K., Tang, S. & Borodovsky, M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.* **28**, 1079–1089 (2018).
36. Gemayel, K., Lomsadze, A. & Borodovsky, M. MetaGeneMark-2: Improved Gene Prediction in Metagenomes. 2022.07.25.500264 Preprint at <https://doi.org/10.1101/2022.07.25.500264> (2022).
37. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34**, 5623–5630 (2006).
38. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res.* **15**, 387–396 (2008).
39. Larralde, M. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *J. Open Source Softw.* **7**, 4296 (2022).
40. McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* **35**, 4537–4542 (2019).
41. Young, K. T., Davis, L. M. & DiRita, V. J. *Campylobacter jejuni*: molecular biology and pathogenesis. *Nat. Rev. Microbiol.* **5**, 665–679 (2007).
42. Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H. & Kuijper, E. J. *Clostridium difficile* infection. *Nat. Rev. Dis. Primer* **2**, 1–20 (2016).

43. Galán, J. E. Salmonella Typhimurium and inflammation: a pathogen-centric affair. *Nat. Rev. Microbiol.* **19**, 716–725 (2021).
44. Cho, J. Y., Liu, R., Macbeth, J. C. & Hsiao, A. The Interface of *Vibrio cholerae* and the Gut Microbiome. *Gut Microbes* **13**, 1937015 (2021).
45. Swaminathan, B., Harmon, M. C. & Mehlman, I. J. A Review: *Yersinia enterocolitica**. *J. Appl. Bacteriol.* **52**, 151–183 (1982).
46. Zhang, T., Li, Q., Cheng, L., Buch, H. & Zhang, F. *Akkermansia muciniphila* is a promising probiotic. *Microb. Biotechnol.* **12**, 1109–1125 (2019).
47. Merenstein, D. *et al.* *Bifidobacterium animalis* subsp. *lactis* BB-12 Protects against Antibiotic-Induced Functional and Compositional Changes in Human Fecal Microbiome. *Nutrients* **13**, 2814 (2021).
48. Sokol, H. *et al.* *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci.* **105**, 16731–16736 (2008).
49. Beghetti, I. *et al.* Probiotics for Preventing Necrotizing Enterocolitis in Preterm Infants: A Network Meta-Analysis. *Nutrients* **13**, 192 (2021).
50. Berni Canani, R. *et al.* *Lactobacillus rhamnosus* GG-supplemented formula expands butyrate-producing bacterial strains in food allergic infants. *ISME J.* **10**, 742–750 (2016).
51. Vitetta, L., Llewellyn, H. & Oldfield, D. Gut Dysbiosis and the Intestinal Microbiome: *Streptococcus thermophilus* a Key Probiotic for Reducing Uremia. *Microorganisms* **7**, 228 (2019).

- 637 52. Dimonaco, N. J., Aubrey, W., Kenobi, K., Clare, A. & Creevey, C. J. No one tool to rule
638 them all: prokaryotic gene prediction tool annotations are highly dependent on the
639 organism of study. *Bioinformatics* **38**, 1198–1207 (2022).
- 640 53. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific
641 phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
- 642 54. BBMap. *SourceForge* <https://sourceforge.net/projects/bbmap/> (2023).
- 643 55. Homo sapiens genome assembly GRCh38. *NCBI* [https://www.ncbi.nlm.nih.gov/data-](https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000001405.26/)
644 [hub/assembly/GCF_000001405.26/](https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000001405.26/).
- 645 56. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node
646 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
647 *Bioinforma. Oxf. Engl.* **31**, 1674–1676 (2015).
- 648 57. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat.*
649 *Methods* **14**, 1023–1024 (2017).
- 650 58. Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M. & Segal, E. An expanded reference
651 map of the human gut microbiome reveals hundreds of previously unknown species.
652 *Nat. Commun.* **13**, 3863 (2022).
- 653 59. Wright, R. J., Comeau, A. M. & Langille, M. G. I. From defaults to databases: parameter
654 and database choice dramatically impact the performance of metagenomic taxonomic
655 classification tools. *Microb. Genomics* **9**, mgen000949 (2023).
- 656 60. Smith, R. H., Glendinning, L., Walker, A. W. & Watson, M. Investigating the impact of
657 database choice on the accuracy of metagenomic read classification for the rumen
658 microbiome. *Anim. Microbiome* **4**, 57 (2022).
- 659 61. Index of /pub/taxonomy. <https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/>.

62. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to
classify genomes with the Genome Taxonomy Database. *Bioinforma. Oxf. Engl.* **36**,
1925–1927 (2019).
63. Ruiz, J. L. *et al.* From contigs towards chromosomes: automatic improvement of long
read assemblies (ILRA). *Brief. Bioinform.* **24**, bbad248 (2023).
64. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
65. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python.
Nat. Methods **17**, 261–272 (2020).
66. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
67. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J.
eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain
Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
68. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic
Acids Res.* **47**, D309–D314 (2019).
69. Queirós, P., Delogu, F., Hickl, O., May, P. & Wilmes, P. Mantis: flexible and consensus-
driven genome annotation. *GigaScience* **10**, giab042 (2021).
70. Vishnepolsky, B. *et al.* Comparative analysis of machine learning algorithms on the
microbial strain-specific AMP prediction. *Brief. Bioinform.* **23**, bbac233 (2022).
71. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

72. Oyarzun, I. *et al.* Human gut metatranscriptome changes induced by a fermented milk product are associated with improved tolerance to a flatulogenic diet. *Comput. Struct. Biotechnol. J.* **20**, 1632–1641 (2022).
73. Zhang, Y. *et al.* Compositional and Functional Differences in the Human Gut Microbiome Correlate with Clinical Outcome following Infection with Wild-Type *Salmonella enterica* Serovar Typhi. *mBio* **9**, 10.1128/mbio.00686-18 (2018).
74. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC Biol.* **17**, 48 (2019).
75. Hitch, T. C. A. *et al.* Recent advances in culture-based gut microbiome research. *Int. J. Med. Microbiol.* **311**, (2021).
76. Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O. & Thompson, J. D. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics* **21**, 293 (2020).
77. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
78. Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937 (2022).
79. Van den Broeck, L. *et al.* Functional annotation of proteins for signaling network inference in non-model species. *Nat. Commun.* **14**, 4654 (2023).
80. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* **178**, 1245-1259.e14 (2019).
81. Petruschke, H. *et al.* Discovery of novel community-relevant small proteins in a simplified human intestinal microbiome. *Microbiome* **9**, 55 (2021).

82. Ma, Y. *et al.* Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **40**, 921–931 (2022).
83. Pirtskhalava, M. *et al.* DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).
84. Hoegenauer, C., Hammer, H. F., Mahnert, A. & Moissl-Eichinger, C. Methanogenic archaea in the human gastrointestinal tract. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 805–813 (2022).
85. Polag, D., Leiß, O. & Keppler, F. Age dependent breath methane in the German population. *Sci. Total Environ.* **481**, 582–587 (2014).
86. Basseri, R. J. *et al.* Intestinal Methane Production in Obese Individuals Is Associated with a Higher Body Mass Index. *Gastroenterol. Hepatol.* **8**, 22–28 (2012).
87. Mathur, R. *et al.* Metabolic Effects of Eradicating Breath Methane using Antibiotics in Prediabetic Subjects with Obesity. *Obes. Silver Spring Md* **24**, 576–582 (2016).
88. Yang, H. *et al.* Ulcerative Colitis-associated E. coli pathobionts potentiate colitis in susceptible hosts. *Gut Microbes* **12**, 1847976 (2020).
89. Lee, I.-A. & Kim, D.-H. Klebsiella pneumoniae increases the risk of inflammation and colitis in a murine model of intestinal bowel disease. *Scand. J. Gastroenterol.* **46**, 684–693 (2011).
90. Pecha, R. E., Prindiville, T., Drell, E. R. & Redor, M. W. Severe Colitis Caused by Staphylococcus aureus Associated with Antibiotic Use: 2 Cases: 436. *Off. J. Am. Coll. Gastroenterol. ACG* **100**, S167 (2005).

Supplementary Tables

Supplementary Table 1: Gene Prediction Software. For each gene prediction tool studied, the version, target taxa, and reference are provided.

Supplementary Table 2: Species studied for optimisation of gene prediction. For each species studied, the genome name, domain, and database are defined. Those species placed in the 'probiotic' and 'pathogen' groups are also defined.

Supplementary Table 3: Comparison of gene prediction tools. For each genome assigned to the studied domain (Archaeal, Bacterial, Eukaryotic, Viral) tools were tested individually, in pairs, and triplets. The perfect, partial, and spurious gene predictions as determined by ORForise are reported.

Supplementary Table 4: Human gut metagenomic studies analysed. For metagenomic study used to create MiProGut and the InvestiGUT database, the 'curated metagenomic data' identifier, number of included metagenomes (post filtering), and the studies PubMed identifier are provided.

Supplementary Table 5: Taxonomic assignment of metagenomic contigs. The number of contigs assigned to Archaea, Bacteria, Eukaryota, Viruses, or defined as unknown are provided, both for each individual sample, and grouped by study.

Supplementary Table 6: Taxonomic assignment of predicted proteins. The number of proteins predicted from contigs assigned to each taxonomic group is stated for each metagenomic study.

Supplementary Table 7: Differential expression analysis of frequently expressed small proteins between ulcerative colitis and healthy samples. For each of the 69 small protein clusters expressed within at least 90% of studied metatranscriptomic samples, the RPKM within ulcerative colitis and healthy samples was calculated and compared. Significantly differentially expressed (adj. P-value < 0.05; Benjamini-Hochberg correction) protein clusters are reported along with their mean RPKM within each group and log2 fold change between groups.

Supplementary Table 8: Differential expression analysis of frequently expressed small proteins between Crohn's disease and healthy samples. For each of the 69 small protein clusters expressed within at least 90% of studied metatranscriptomic samples, the RPKM within Crohn's disease and healthy samples was calculated and compared. Significantly differentially expressed (adj. P-value < 0.05; Benjamini-Hochberg correction) protein clusters are reported along with their mean RPKM within each group and log2 fold change between groups.

Supplementary Figures

Supplementary Figure 1: Geographical prevalence of seaweed degrading porphyranases (*Bp1689*), and agarases (*Bp1670*). Each protein was studied using InvestiGUT at the default 90% identity threshold.

Supplementary Figure 2: Analysis of the small protein cluster '33977|YachidaS_2019__SAM00114748|Bacteria|11|MetaGeneAnnotator-

MetaGeneMark-Pyrodigal'. The conservation of sequence between members of the cluster, along with an Alphafold2 structural model were generated (**a**). The frequency of the cluster across *Bacteroides* spp. (n = 59) (**b**), and predicted antimicrobial activity against key pathogens and fungi (**c**). The prevalence of the cluster with host parameters, including westernised status, and antibiotic use, was studied (**d**), along with the prevalence in IBD and Crohn's disease (CD) patients (**e**), and *Clostridioides difficile* infection (CDI) patients compared to healthy controls (**f**). Significance from Fisher's exact test, after Benjamini-Hochberg correction, are shown as: NS = not significant, * < 0.05, ** < 0.01, *** < 0.001.

a

Metagenomes: 9,677
Genomes: 3,594

Taxonomic separation

Metagenome: Kraken 2
Genome: GTDB-Tk

Bacteria Unknown Archaea Viruses Eukaryota

Genetic code defined for taxon in NCBI Taxonomy

Gene prediction

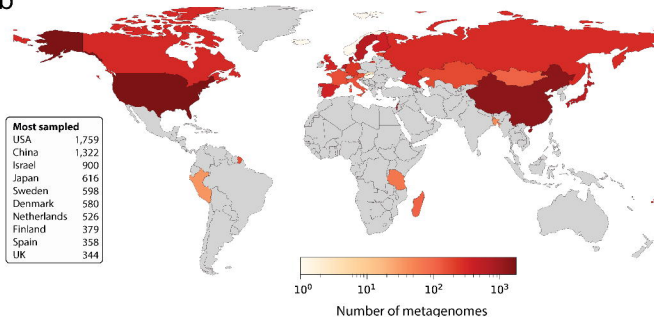
<i>Bacteria/Unknown</i>	<i>Archaea</i>	<i>Viruses</i>	<i>Eukaryota</i>
Pyrodigal	GeneMarkS	FragGeneScan	Pyrodigal
MetaGeneMark	GeneMarkS2	GeneMarkST	Snap
MetaGeneAnnotator	MetaGeneAnnotator	MetaGeneMark	Augustus

Filtering and merging predictions

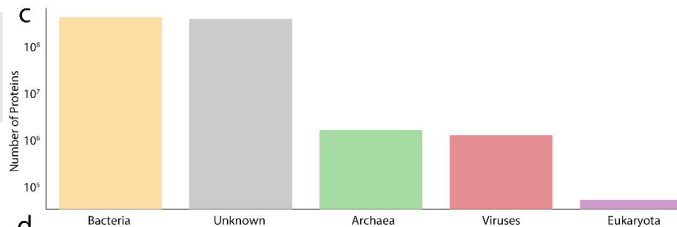


Total genes: 846,619,045
Dereplicated: 29,232,514

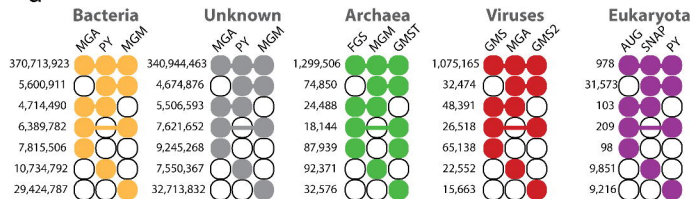
b

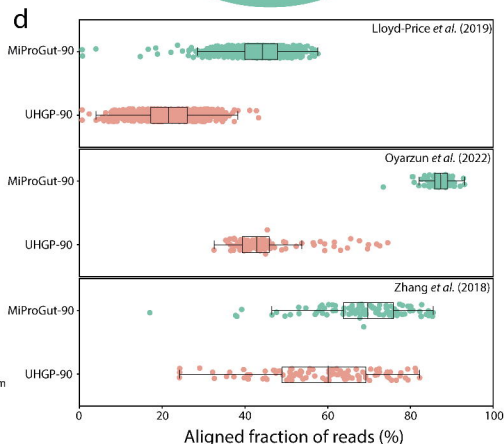
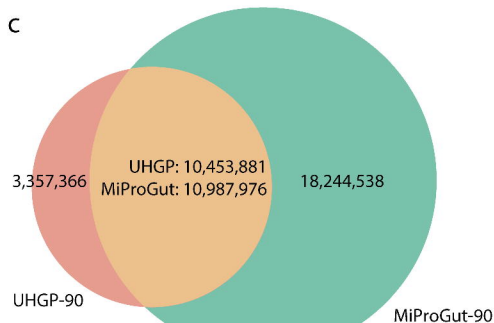
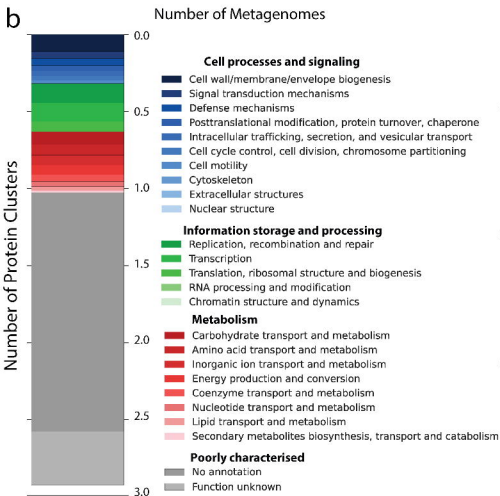
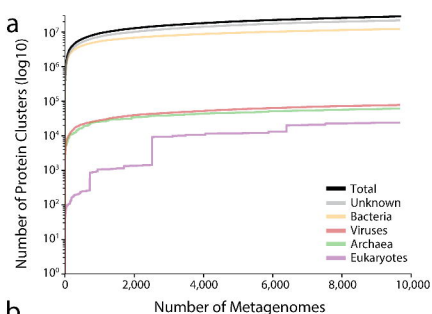


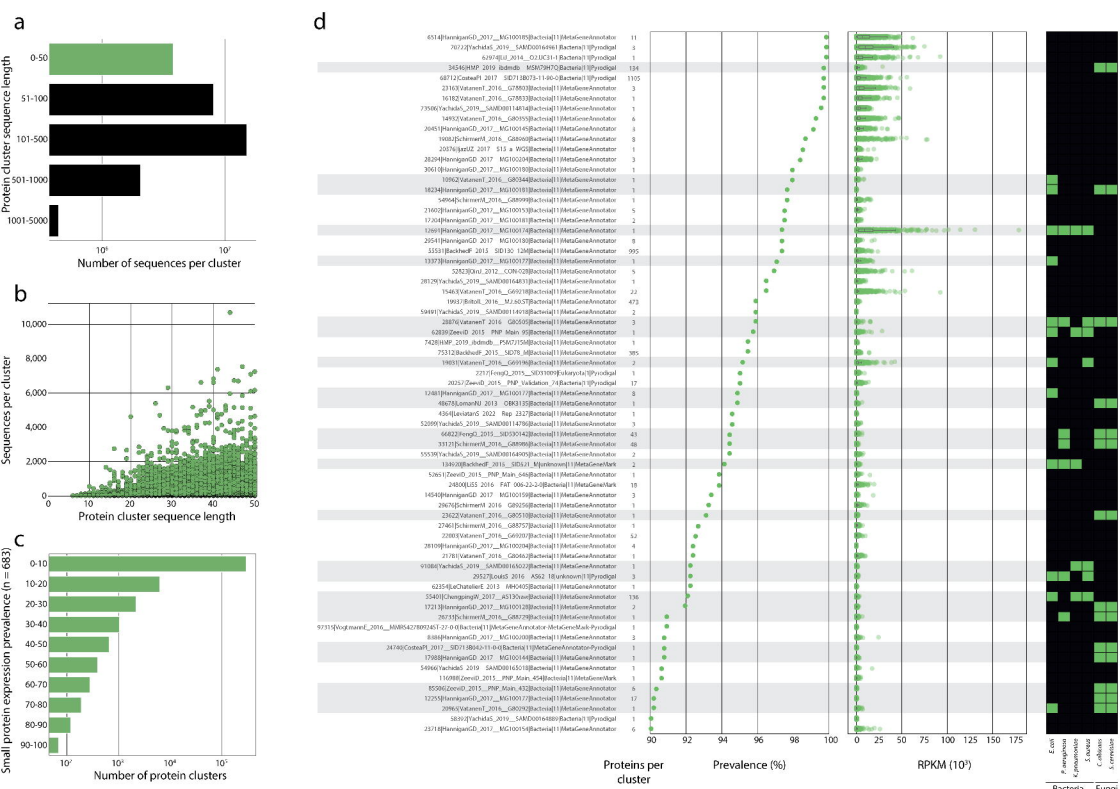
c



d





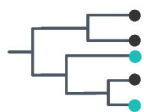


a

User's protein



Sequence comparison (DIAMOND)

Host associations
(n = 9,634 metagenomes)Taxonomic diversity
(n = 3,594 genomes)**Disease association:**Inflammatory bowel diseases (Ulcerative colitis, Crohn's disease), *Clostridioides difficile* infection, Type 2 diabetes, Colorectal cancer, Cancer stage (AJCC), Rheumatoid arthritis**Geographical association:**

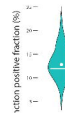
General: Western Vs non-Western

Country specific: Japan, China, Israel, USA, Mongolia, Fiji, Denmark, Sweden, Netherlands, Austria, Bangladesh, Canada, Germany, Spain, Estonia, Finland, France, UK, Italy, Kazakhstan, Madagascar, Peru, Russia, Tanzania

Host parameters:

BMI category, Current antibiotic usage, Age (decade of life), Smoking status, Gender

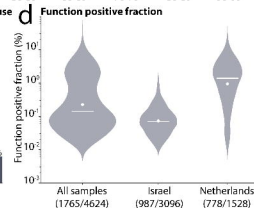
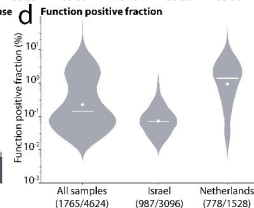
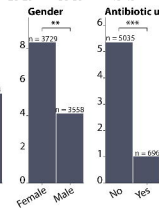
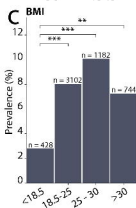
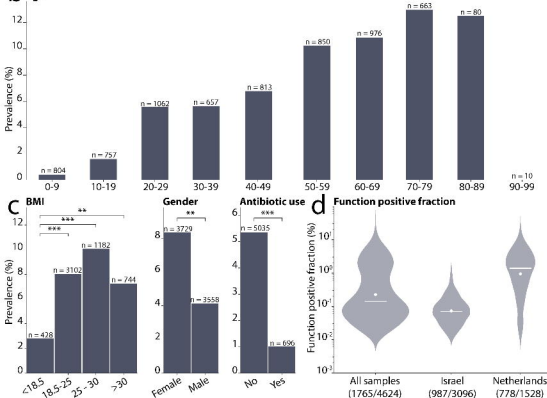
Relative abundance (%)

Relative abundance of species
(n = 4,624)

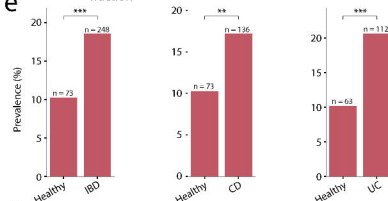
Function positive fraction

- Cumulative relative abundance of species encoding the protein of interest within a given sample is the 'function positive fraction'.
- Function positive fraction is calculated for each of the 4,624 faecal samples

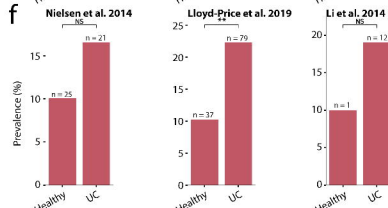
b Age



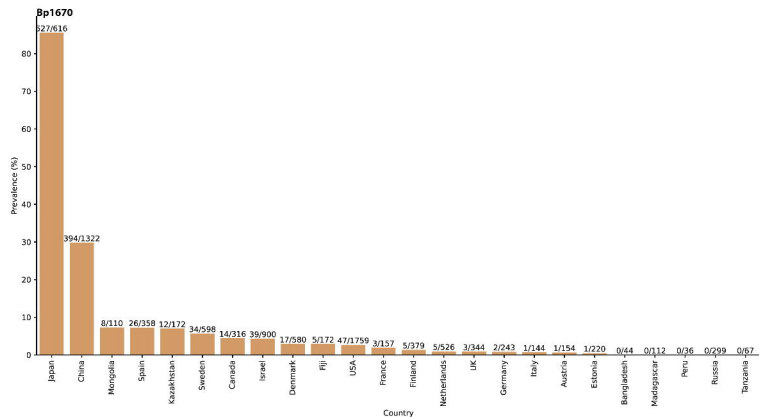
e



f



a



b

