Global compositional and functional state of the human gut microbiome in health and disease

Sunjae Lee1,16\*, Theo Portlock2\*, Emmanuelle Le Chatelier3\*, Fernando Garcia-Guevara2\*, Florian Plaza Onate3, Nicolas Pons3, Azadeh Harzandi1, Neelu Begum1, Ceri Proffitt1, Dorines Rosario1, Stefania Vaga1, Junseok Park4, Kalle von Feilitzen2, Fredric Johansson2, Cheng Zhang2, Lindsey A. Edwards5, Vincent Lombard6,7, Franck Gauthier3, Claire J. Steves8, David Gomez-Cabrero1,9, Bernard Henrissat6,7,10, Doheon Lee4, Lars Engstrand11, Debbie L. Shawcross5, Gordon Proctor1, Mathieu Almeida3, Jens Nielsen12,13,14, David L. Moyes1, Adil Mardinoglu1,2 , Stanislav Dusko Ehrlich3, 15, Mathias Uhlen2,#, Saeed Shoaie1,2,#

1 Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King’s College London, SE1 9RT, UK

2 Science for Life Laboratory, KTH – Royal Institute of Technology, Stockholm, SE-171 21, Sweden

3 University Paris-Saclay, INRAE, MetaGenoPolis, 78350 Jouy-en-Josas, France

4 Department of Bio and Brain Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea

5 Institute of Liver Studies, Department of Inflammation Biology, School of Immunology and Microbial Sciences, King’s College London, London, UK

6 INRAE, USC1408 Architecture et Fonction des Macromolécules Biologiques (AFMB), Marseille, 13288, France

7 Architecture et Fonction des Macromolécules Biologiques (AFMB), CNRS, Aix-Marseille University, Marseille, 13288, France

8 Department of Twin Research & Genetic Epidemiology, King’s College London, London, UK

9 Translational Bioinformatics Unit, NavarraBiomed, Departamento de Salud-Universidad Pública de Navarra, Pamplona, 31008, Navarra, Spain

10 Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

11 Centre for Translational Microbiome Research (CTMR), Micobiology, Tumor and Cell Biology (MTC), Karolinska Institutet, 417164, Solna, Sweden

12 Department of Biology and Biological Engineering, Kemivägen 10, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden

13 BioInnovation Institute, Ole Måløes Vej 3, DK-2200 Copenhagen N, Denmark

14 Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

15 Department of Clinical and Movement Neurosciences, University College London, London, NW3 2PF, UK

16 School of Life Sciences, Gwangju Institute of Science and Technology (GIST), 61005, Gwangju, Republic of Korea

#corresponding authors: [mathias.uhlen@scilifelab.se](mailto:mathias.uhlen@scilifelab.se); [saeed.shoaie@kcl.ac.uk](mailto:saeed.shoaie@kcl.ac.uk)

\* These authors contributed equally.

# Abstract

The role of gut microbiota in humans is of great interest, and metagenomics provides key opportunities for extensively analysing bacterial diversity in health and disease. Despite increasing efforts to expand microbial gene catalogues and an increasing number of metagenomes assembled genomes, there has been few investigations in pan-metagenomics and in-depth functional analysis across different geographies and diseases. Here, we explored 5,708 human gut metagenome samples across 19 countries and 23 diseases, performing compositional, functional cluster, and integrative analysis. We identified *Fusobacterium nucleatum* and *Anaerostipes hadrus* with the highest frequencies, enriched and depleted respectively, across different disease cohorts. Distinct functional distributions were observed in the gut microbiome of westernized and non-westernized populations. Additionally, a random forest machine learning model identified key species as disease biomarkers. This compositional and functional analysis are presented in an open-access Human Gut Microbiome Atlas (www.microbiomeatlas.org), allowing for exploration of the richness, diseases, and regional signatures of the gut microbiota across different cohorts.

# Introduction

Metagenomic studies of the human microbiome enable the characterization of microbial and functional diversity in health and disease[1]. A deeper understanding of the functional potential and taxonomic composition of the microbiome will have major implications in identifying health and disease signatures across different body sites and geographic regions [2]–[4]. 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 and even new microbiome-based treatments and therapies [2], [3]. Recently, several studies have focused on the discovery of new uncultured microbes through the generation of metagenome species[6]–[10], while others have concentrated on the investigation of alterations in the microbiome compositions due to disease, geographical location, and interventions in the gut microbiome [2], [11], [12].

Key to advancing our understanding of the critical role played by the microbiome in health and disease is access to data from a wide range of studies and cohorts. Public resources collecting and processing microbiome data are essential, contributing to the laborious and necessary task of standardizing and making accessible this accumulated information. Some of them have particularly focused on the human gut microbiome (gutMDisorder [13], GIMICA [14], DISBIOME [15] and GMRepo [16]). However, there is a lack of integrative functional and compositional analysis across the cohorts and regions to provide a mechanistic understanding of the microbiome and biomarker identification. Here, we integrated publicly available data from a wide range of studies across different countries from both healthy and diseased individuals. We calculated the enrichment of microbial species in both disease and geographical regions, and performed SHAP (SHapley Additive exPlanations) interpretations on Random Forest classification models to identify biomarkers of disease associated with metagenomic species. Additionally, we present an open-access Human Gut Microbiome Atlas (www.microbiomeatlas.org), that allows researchers to explore an integrative analysis of compositional, functional, richness, disease, and regional signatures for the gut microbiota across 19 geographical regions and 23 diseases.

**RESULTS**

# The Human Gut Microbiome Atlas; A pan-metagenomics study of compositional and functional changes of the human gut microbiome

To provide a central public resource for exploring the microbiome in different settings, we performed a large-scale integrative analysis of 6014 publicly available shotgun metagenomics stool samples. The 5,708 samples were selected with at least 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 normalized all metagenomic sample abundances to enable comparative analysis across the different cohorts (Methods). Using these samples, we then 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). We generated gene counts and MSP abundances for all the samples using the IGC2 10.4 million gene catalogue[17]. We also characterized the functions and phenotype of the identified MSPs in 7 different categories: KEGG orthologs (KOs) [18], protein families (PFAM) [19] carbohydrate-active enzyme (CAZymes) [20], antimicrobial resistance (AMR) [21], microbial phenotype [22], virulence factor [23], and biosynthetic gene clusters (BGCs) [24]. We finally identified 7,763 co-conserved functional clusters across species (Methods).All these data are freely available in the HGMA without restrictions in the public open access database ([www.microbiomeatlas.org](http://www.microbiomeatlas.org)), as part of the Human Protein Atlas program (www.proteinatlas.org). All MSPs and functions are highlighted together with the 5,708 samples across 19 countries with disease and healthy cohorts.

Using all samples obtained from individuals across 19 countries, we uncovered the geographical distribution of the gut microbiome. Both *Clostridium* and *Bacteriodes* were found to have higher mean relative abundance within western countries, whereas *Prevotella* specieshad a higher mean relative abundance within non-western countries (Fig. 1d). We applied the unsupervised clustering method, *monocle*, to MSP abundance profiles of all samples (Methods)[25], [26] and observed that there were two distinct ordinations of non-westernized and European samples of subjects connected by a mixture of western/non-western samples belonging to China/Japan/US (Fig. 1e). Based on comparative analysis across different regions, we identified 624 MSPs specifically enriched in certain countries (See Methods and Supplementary Table S2). Functional annotation-based analysis across geographical clusters revealed enrichment of CAZymes for degrading host mucins, food carbohydrates from animal origins, and storage carbohydrates in westernized populations, where antimicrobial resistance (AMR) and virulence factors were also more prevalent (Fig. 1f). Comparing the functions of region enriched MSPs in westernized countries revealed that genes encoding for vancomycin resistance and lipopolysaccharide (LPS) biogenesis are overrepresented. An overrepresentation of genes encoding for complex polysaccharide binding proteins mostly belonging to the *Prevotella* genus was found in the non-westernized cohorts. Moreover, we identified that the cluster for vancomycin resistance is enriched in the westernized population while the tetracycline resistance cluster is enriched in the non-westernized population.

# Pan-metagenomics association study across 23 diseases

We performed a Pan-metagenomics association (Pan-MGAS) of multiple disease cohorts (23 diseases across 43 cohorts from 14 countries) to distinguish diseased *versus* healthy microbiomes within the multiple cohorts. We determined the enriched and depleted species within the different disease cohorts compared to healthy samples from the same country, showing an effect size greater than 0.3 (Fig. 2a, Supplementary Table S3 and Method). Some cohorts showed a depletion of multiple species, notably in cancer [Non-Small Cell Lung Cancer (NSCLC, from France), renal cell carcinoma (RCC from France), adenoma (from Italy)] (Fig. 2a). Conversely, some diseases had several enriched species, as seen for most Colorectal Cancer (CRC) cohorts.

Some species were either enriched or depleted across multiple cohorts, regardless of geographical differences. For example, *Anaerostipes hadrus,* *Coprococcus comes,* and *Blautia obeum,* which have been associated with healthy individuals, areamong the most frequently depleted species, found in at least six different disease cohorts (Fig. 2b, Supplementary Fig. S1)*.* The firsttwo species have been described as butyrate producers and dominant species isolated from the healthy human colon[27]–[29], and the third has been associated with gut microbiota recovery after cholera infection and with normal maturation of the infant gut microbiota[30].

Between the species found enriched in at least 6 different cohorts we find *Fusobacterium nucleatum, Clostridium bolteae, Clostridium clostridioforme, Clostriduium symbiosum, Peptostreptococcus stomatis, Flavonifractor plautii, Parvimonas micra,* among others (Fig. 2C, Supplementary Fig. S1). Several of them also have been isolated from oral samples (*F. nucleatum*[31], *P. stomatis*[32]*, P. micra*[33]) and some of them have been identified in infections, including bacteremia (*C. bolteae*[34], *C. clostridioforme*[35], *P. micra*[36]). Along with *F. nucleatum* and *C. symbiosum*, which were enriched in western countries and are associated with CRC[37]–[39], we also identify *P. micra* to be enriched in multiple cohorts of CRC and *P. stomatis* enriched several times only in solid tumour cohorts (Supplementary Table S4, Supplementary Fig. S1).

# Disease enriched-functional clusters showed distinct links to gut microbiome dysbiosis

To analyze the functional content in the MSP from the human microbiome, we applied an unsupervised clustering approach to the MSPs’ annotated gene content (Fig. 3a-c, Supplementary Fig. 2 and Methods). This analysis provided a better representation of microbial functions than single annotations or known pathway definitions (*e.g*., KEGG) (Fig. 3b). We identified 7,763 functional clusters and 6,297 singletons based on the community detection algorithm (Supplementary Table 5). 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 5, named “*comm-cluster*” hereafter) 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. In contrast, the *patho-cluster* was enriched in functions associated with the uptake of several substrates. These included siderophores, amino acid, and vitamin transport, thus improving competitive fitness against the commensal bacteria. We also found other functionally enriched clusters such as butyrate metabolism cluster, propionate metabolism cluster, and the CRISPR-cas system cluster (Fig. 3c); a number of these were correlated with phylum-level taxonomy (Supplementary Fig. 2c).

We next projected the functional clusters on enriched/depleted MSPs in HGMA disease cohorts (Fig. 3d: hypergeometric tests, p-value < 10-4). We found several functional clusters commonly associated with the enriched species in disease. Among these, we found CL-1006, related to antibiotic resistance; CL-1032, a competence-related DNA transformation transport, which could provide an advantage by improving the integration of new functions into the genome; or clusters related to metabolic pathways that could contribute indirectly to pathogenicities like the Pentose Phosphate Pathway [40] or Ethanolamine utilization [41]. Among the most frequent functional clusters that accompany the depleted species in disease, we found the CL-12 comm-cluster, and other clusters with functions related to pectate degradation and biofilm formation (Fig. 3d), all of them related to the normal function of the healthy microbiota.

# Global view of gut microbiome species pangenomes

To get a holistic view of the human gut MSPs we generated a phylogenetic tree displaying the taxonomic resolution of disease and region enriched species (Fig. 4, Methods) and estimated proportionality (see methods) between MSPs pairs. Most MSPs are present in both western and non-western regions. While some of them are enriched in one of the two regions, we could not identify any apparent phylogenetic pattern. When looking at the enrichment/depletion across the different cohorts, the *Streptococcus* genus shows particularly interesting features: members within this genus were found enriched in some cohorts while being depleted in others. For example, three different species within the genus (*S. anginosus,* *S. parasanguinis* and *S. vestibularis*) were enriched in two distinct liver disease cohorts, while in contrast, species including *S. salivarius* and *S. sanguinis* were depleted in cancer cohorts (Figure 2C, Supplementary Figure S3, Supplementary Table S4).

In addition, we observed proportionality between MSPs (Supplementary FigureS3). A high proportionality value between a pair of MSPs suggests they tend to increase or decrease together. Most MSPs showing the highest proportionality values belong to the same genus. Only a small subset of MSPs was found with proportionality values above the selected threshold. Many of the MSP pairs we found are inhabitants of the oral cavity, and the *Streptococcus* genus stands out again.

Bacterial infections of the *Streptococcus* genus play a central role from a clinical point of view. However, the complete role of this genus is still not fully understood and should be the subject of further research [42], [43].

# A Random Forest classification model can identify biomarkers of disease from metagenomic species

To identify disease biomarkers, we implemented feature selection-based random forest classifier models trained using the MSPs constructed from each cohort on the HGMA grouped by disease (Fig. 5). These models were able to classify between a pool of randomly selected healthy samples and disease groups with variable discriminatory performances. Prediction performance was evaluated by the Area Under the Curve (AUC) metric. The models with the highest predictive capabilities were for the prediction of Impaired Glucose Tolerance (IGT), Vogt-Koyanagi-Harada (VKH), Non-small-cell lung carcinoma (NSCLC), melanoma, advanced adenoma, and Crohn’s Disease (CD).

The generalization of these models was assessed with an interstudy cross-validation which demonstrated that a model trained on the CRC training cohort PRJEB10878 [44] was able to maintain predictive power of disease classification when applied to the CRC test cohort PRJEB6070 [45] (Figure 5A) (AUCROC = 0.74). Additional validation of the importance of randomly selected healthy samples was carried out by combining 30 random groups of 40 healthy samples with 40 random CRC samples and repeating the cross-validation. It was shown that the AUCROC of predicting the test cohort was 0.75 ± 0.04 showing conservation of predictive capabilities.

The interpretable machine learning framework, SHAP (SHapley Additive exPlanations), was used to identify disease-specific gut microbiome features [47]*.* SHAP is a state-of-the-art framework for understanding ‘black-box’ classifiers that has been recently used to explain gut microbiome classification models [48], [49]. By interpreting the disease classification models using directional mean absolute SHAP values, the importance of metagenomic species as biomarkers for 16 diseases in the HGMA was measured. As with the effect size calculation for biomarker identification, the number of shared species deemed necessary for disease prediction between diseases was low.

When comparing SHAP score calculated biomarkers to effect size biomarkers for all disease, a number of the key species were shared (Fig. 5D). The highest directional mean SHAP scoring species for the CRC predictive model were *F. nucleatum*., *Parvimonas micra, Solobacterium moorei, Streptococcus parasanguinis*; all of which have been demonstrated to be biomarkers previously [50]–[53] .

A cluster formed of *Clostridium asparagiforme, Clostridium sp KLE 1755, Eisenbergiella tayi, Clostridium phoceensis, Biolophila wadsworthia, Barnesiella intestinihominis, and Odoribacter splanchnicus* had high directional mean absolute SHAP values in models that predict multiple cancers (NSCLC, melanoma, RCC, and partly CRC). *C. asparagiforme* was also shown in be previously enriched in breast cancer [54]. This pan-cancer signal was also predictive of ME/CFS – a condition that has been linked immunologically with cancer [55].

Of note, an increase in abundance of the commensal oral bacteria *Haemophilus parainfluenzae,* *V. dispar*, *V. atypica*, and *V. parvula* were shown to have high importance in predicting liver cirrhosis but not NAFLD, as found previously [56], and were found to be enriched in multiple cohorts regardless of region (Fig. 5). In the NAFLD model, an increase in the abundance of *S. parasanguinis* was the most important factor for predicting the disease. This species was found enriched across multiple cohorts of the HGMA and is part of a cluster of oral commensal species shown previously to be biomarkers of the disease [57]. NAFLD also shared biomarkers with T1D including *Anaerostipes hadrus*, and *Eubacterium hallii.* The two diseases have been shown previously to be metabolically associated [58].

There were some diseases where the majority of high importance species for disease prediction were depleted species Chron’s Disease. Additionally, there were several shared disease predictive species, such as *Streptococcus parasanguinis* and *Doria longecatena 2,* with their presence and absence characterizing a general dysbiotic state, respectively.

# Discussion

One of the urgent requirements in applying our developing knowledge of the microbiome to predicting and treating disease are tools for assessing and analysing data from a wide range of different studies. We performed a comprehensive integrative analysis of global and temporal gut microbiomes and provided an open access HMGA portal (http://microbiomeatlas.org). This tool allows for integrating several studies simultaneously that link species to disease, region, and function. It also presents a means of contextualizing gene and species enrichments phylogenetically. We demonstrated that the difference in origin (western/non-western) is reflected by the gut microbial composition with species/genes being over/under-represented in each origin. Finally, we also found that some species and functions are enriched or depleted across multiple diseases and studies. A number of those species were important in predicting those diseases using a random forest classification model.

Confirming previous observations [6], we have described the gut microbiome regional specificity, which needs to be considered before using the gut microbiome for patient stratification or designing intervention studies. Beyond previous observations, we find there are distinctions in functions enriched in westernized and non-westernized countries including resistance to vancomycin and tetracycline respectively.

The physiological changes caused by disease might partly explain why some diseases have a pronounced compositional imbalance while others do not. Diseases affecting the bowel and CRC show a high effect size of many species, while some diseases affecting other body parts tend to produce smaller imbalances. Some other factors might also be involved in the magnitude of the imbalance, for example, the changes in diet[62], [63] or the use of drugs for treating the disease [64]–[66]).

The loss of species actively contributing to maintaining a healthy gut environment could increase the host’s vulnerability to further health complications. For example, we observed some of the more frequently depleted species have been described as butyrate producers. Butyrate has been associated with beneficial effects in the colon, such as inhibition of inflammation, reinforcing the epithelial barrier and decreasing oxidative stress [67]. Depletion of butyrate producing species were observed to be the strongest features in the random forest model that predicts Crohn’s disease. Conversely, some enriched species might induce disease pathology by leading to new infections, potentiating the disease symptoms and even weakening the immune response. Some reports suggest *F. nucleatum* promotes CRC development and metastasis[68], [69]. We have found a similar link between this species and CRC using the SHAP interpreted random forest predictive model. A previous report found *Flavonifractor plautii*, a species we found enriched in 6 cohorts, suppresses Th2 immune responses in mice[70] which leads us to speculate that this species could exert a similar effect in humans. The Pan-MGAS we present shows a clear bias toward CRC studies due to the greater number of these studies. We expect new studies released in the future to include more countries and diseases that will help to update our analyses and balance out this bias. It is interesting to note that many of the species identified in our analyses as either enriched or depleted in disease states are more commonly associated with the oral cavity. This is particularly true of the representatives of the *Streptococcus* genus. Many of the streptococcal species identified here are members of the viridans group *streptococci* – a diverse group that has members associated with disease and polymicrobial infection (*e.g*. *S. anginosus*, associated with liver and soft tissue abscesses [71], as well as members that have been proposed for use as probiotics (*e.g*. *S. salivarus* ([62], [63];)

The projection of functions associated with enriched/depleted species in disease supports the observations made with species alone. The functions found commonly enriched in disease suggest they provide their carriers increased fitness, meaning that they have a better chance of thriving in altered conditions, playing indirect roles in disease pathology for example, by utilizing additional carbon sources (CL543-pentose phosphate pathway[73], ethanolamine [74]) or increasing their ability to survive environmental stresses (CL-592 osmoprotectant cluster [62]). However, the enrichment of these functions does not imply they are exclusive to pathogenic organisms. For example, although anaerobic sulphite reducing activity is often used as a marker for food contamination [76] the activity is also present in several non-pathogenic bacteria. Conversely, functions depleted across different diseases could be playing an active role in health maintenance. This can be seen in recent research that has revealed that pectic substances can inhibit gut inflammation and relieve inflammatory bowel disease symptoms [77].

Finally, the integration of metagenomics data from many studies spanning five continents provides a valuable knowledge resource for researchers investigating the impact of the microbiome on individual health parameters. This open-access atlas will be updated routinely with the new publicly available gut metagenomics data, including the recently announced One Million Microbiome Project to provide 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 critical role of the gut microbiome in human maintaining health.

# Data availability

The datasets used in this study are available in http://www.microbiomeatlas.org, with relevant project accession codes of raw data provided in Supplementary Table S1.

# Code availability

The scripts for the functional clusters, SHAP calculation, plotting, enrichment in disease/region are publicly available at https://github.com/sysbiomelab/ATLAS.

# Acknowledgements

This study was supported by Science for Life Laboratory (SciLifeLab), Engineering and Physical Sciences Research Council (EPSRC), EP/S001301/1, Biotechnology Biological Sciences Research Council (BBSRC) BB/S016899/1, 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 assisting 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., E.L., T.P, F.G.G. and S.S. led the design and analysis of the data. S.L., E.L., T.P, and F.G.G performed the analysis, and made the figures. M.A., F.P., E.L., and S.D.E. generated the MSPs, performed quality checks and taxonomy updates. N.P. annotated the updated gut gene catalogue. N.B., C.P., S.V., D. R. and A.H. analysed part of the data and prepared the materials for the HGMA. M.U., K.F. and F.J. developed the HGMA website. V.L. and B.H. annotated the gut catalogue with new CAZymes. J.P. and D.L. annotated the secondary metabolites of the gene catalogue. S.S., S.L. and T.P. wrote and drafted the manuscript. L.A.E, D.L.S, A.M., G.P. J.N. D.L.M. 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] J. Lloyd-Price *et al.*, “Strains, functions and dynamics in the expanded Human Microbiome Project,” *Nature*, vol. 550, no. 7674, pp. 61–66, 2017, doi: 10.1038/nature23889.

[2] L. A. David *et al.*, “Host lifestyle affects human microbiota on daily timescales,” *Genome Biology*, vol. 15, no. 7, pp. 1–15, Jul. 2014, doi: 10.1186/GB-2014-15-7-R89/TABLES/1.

[3] C. A. Lozupone, J. I. Stombaugh, J. I. Gordon, J. K. Jansson, and R. Knight, “Diversity, stability and resilience of the human gut microbiota,” *Nature 2012 489:7415*, vol. 489, no. 7415, pp. 220–230, Sep. 2012, doi: 10.1038/nature11550.

[4] F. Sommer, J. M. Anderson, R. Bharti, J. Raes, and P. Rosenstiel, “The resilience of the intestinal microbiota influences health and disease,” *Nature Reviews Microbiology 2017 15:10*, vol. 15, no. 10, pp. 630–638, Jun. 2017, doi: 10.1038/nrmicro.2017.58.

[5] A. Heintz-Buschart *et al.*, “Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes,” *Nat. Microbiol.*, vol. 2, p. 16180, 2016, doi: 10.1038/nmicrobiol.2016.180.

[6] H. B. Nielsen *et al.*, “Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes,” *Nat. Biotechnol.*, vol. 32, no. 8, pp. 822–828, 2014, doi: 10.1038/nbt.2939.

[7] S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, and N. C. Kyrpides, “New insights from uncultivated genomes of the global human gut microbiome,” *Nature*, vol. 568, no. 7753, pp. 505–510, 2019, doi: 10.1038/s41586-019-1058-x.

[8] A. Almeida *et al.*, “A new genomic blueprint of the human gut microbiota,” *Nature*, vol. 568, no. 7753, pp. 499–504, 2019, doi: 10.1038/s41586-019-0965-1.

[9] F. Plaza Onate *et al.*, “MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data,” *Bioinformatics*, vol. 35, no. 9, pp. 1544–1552, 2019, doi: 10.1093/bioinformatics/bty830.

[10] E. Pasolli *et al.*, “Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle,” *Cell*, vol. 176, no. 3. 2019. doi: 10.1016/j.cell.2019.01.001.

[11] J. Jalanka-Tuovinen *et al.*, “Intestinal Microbiota in Healthy Adults: Temporal Analysis Reveals Individual and Common Core and Relation to Intestinal Symptoms,” *PLOS ONE*, vol. 6, no. 7, p. e23035, 2011, doi: 10.1371/JOURNAL.PONE.0023035.

[12] R. S. Mehta *et al.*, “Stability of the human faecal microbiome in a cohort of adult men,” *Nature Microbiology 2018 3:3*, vol. 3, no. 3, pp. 347–355, Jan. 2018, doi: 10.1038/s41564-017-0096-0.

[13] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, “gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, Jan. 2020, doi: 10.1093/NAR/GKZ843.

[14] J. Tang *et al.*, “GIMICA: host genetic and immune factors shaping human microbiota,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D715–D722, Jan. 2021, doi: 10.1093/NAR/GKAA851.

[15] Y. Janssens *et al.*, “Disbiome database: Linking the microbiome to disease,” *BMC Microbiology*, vol. 18, no. 1, pp. 1–6, Jun. 2018, doi: 10.1186/S12866-018-1197-5/TABLES/1.

[16] S. Wu *et al.*, “GMrepo: a database of curated and consistently annotated human gut metagenomes,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D545–D553, Jan. 2020, doi: 10.1093/NAR/GKZ764.

[17] C. Wen *et al.*, “Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis,” *Genome Biol.*, vol. 18, no. 1, p. 142, 2017, doi: 10.1186/s13059-017-1271-6.

[18] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, “The KEGG resource for deciphering the genome,” *Nucleic Acids Research*, vol. 32, no. suppl\_1, pp. D277–D280, Jan. 2004, doi: 10.1093/NAR/GKH063.

[19] S. El-Gebali *et al.*, “The Pfam protein families database in 2019,” *Nucleic acids research*, vol. 47, no. D1, pp. D427–D432, 2019.

[20] N. Terrapon, V. Lombard, E. Drula, P. M. Coutinho, and B. Henrissat, “The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines,” *A Practical Guide to Using Glycomics Databases*, pp. 117–131, 2017, doi: 10.1007/978-4-431-56454-6\_6.

[21] E. Ruppé *et al.*, “Prediction of the intestinal resistome by a three-dimensional structure-based method,” *Nature Microbiology 2018 4:1*, vol. 4, no. 1, pp. 112–123, Nov. 2018, doi: 10.1038/s41564-018-0292-6.

[22] S. Mukherjee *et al.*, “Genomes OnLine Database (GOLD) v.8: overview and updates,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D723–D733, Jan. 2021, doi: 10.1093/NAR/GKAA983.

[23] C. Mao, D. Abraham, A. Wattam, … M. W.-, and undefined 2015, “Curation, integration and visualization of bacterial virulence factors in PATRIC,” *academic.oup.com*, Accessed: Feb. 03, 2022. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/31/2/252/2366053

[24] K. Blin *et al.*, “antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification,” *Nucleic Acids Res*, vol. 45, no. W1, pp. W36–W41, 2017, doi: 10.1093/nar/gkx319.

[25] C. Trapnell *et al.*, “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells,” *Nat. Biotechnol.*, vol. 32, no. 4, pp. 381–386, 2014, doi: 10.1038/nbt.2859.

[26] X. Qiu *et al.*, “Reversed graph embedding resolves complex single-cell trajectories,” *Nat. Methods*, vol. 14, no. 10, pp. 979–982, 2017, doi: 10.1038/nmeth.4402.

[27] L. v. Holdeman and W. E. C. Moore, “New genus, Coprococcus, twelve new species, and emended descriptions of four previously described species of bacteria from human feces,” *International Journal of Systematic Bacteriology*, vol. 24, no. 2, pp. 260–277, Apr. 1974, doi: 10.1099/00207713-24-2-260/CITE/REFWORKS.

[28] P. Louis and H. J. Flint, “Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine,” *FEMS Microbiology Letters*, vol. 294, no. 1, pp. 1–8, May 2009, doi: 10.1111/J.1574-6968.2009.01514.X.

[29] E. Allen-Vercoe *et al.*, “Anaerostipes hadrus comb. nov., a dominant species within the human colonic microbiota; reclassification of Eubacterium hadrum Moore et al. 1976,” *Anaerobe*, vol. 18, no. 5, pp. 523–529, Oct. 2012, doi: 10.1016/J.ANAEROBE.2012.09.002.

[30] P. A. Lawson and S. M. Finegold, “Reclassification of Ruminococcus obeum as Blautia obeum comb. nov,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 65, no. 3, pp. 789–793, Mar. 2015, doi: 10.1099/IJS.0.000015/CITE/REFWORKS.

[31] S. S. Socransky, A. D. Haffajee, M. A. Cugini, C. Smith, and R. L. Kent, “Microbial complexes in subgingival plaque,” *Journal of clinical periodontology*, vol. 25, no. 2, pp. 134–144, 1998, doi: 10.1111/J.1600-051X.1998.TB02419.X.

[32] J. Downes and W. G. Wade, “Peptostreptococcus stomatis sp. nov., isolated from the human oral cavity,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, no. 4, pp. 751–754, Apr. 2006, doi: 10.1099/IJS.0.64041-0/CITE/REFWORKS.

[33] I. N. Rôças and J. F. Siqueira, “Root canal microbiota of teeth with chronic apical periodontitis,” *Journal of Clinical Microbiology*, vol. 46, no. 11, pp. 3599–3606, Nov. 2008, doi: 10.1128/JCM.00431-08/SUPPL\_FILE/RC\_CHECKERBOARD\_CHRONIC\_APICAL\_PERIODONTITIS\_TABLE\_APPENDIX\_ONLINE.PDF.

[34] S. M. Finegold *et al.*, “Clostridium clostridioforme: A mixture of three clinically important species,” *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 24, no. 5, pp. 319–324, May 2005, doi: 10.1007/S10096-005-1334-6.

[35] S. M. Finegold *et al.*, “Clostridium clostridioforme: A mixture of three clinically important species,” *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 24, no. 5, pp. 319–324, May 2005, doi: 10.1007/S10096-005-1334-6.

[36] R. García Carretero, E. Luna-Heredia, M. Olid-Velilla, and O. Vazquez-Gomez, “Bacteraemia due to Parvimonas micra, a commensal pathogen, in a patient with an oesophageal tumour,” *Case Reports*, vol. 2016, p. bcr2016217740, Nov. 2016, doi: 10.1136/BCR-2016-217740.

[37] M. Castellarin *et al.*, “Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma,” *Genome Research*, vol. 22, no. 2, pp. 299–306, Feb. 2012, doi: 10.1101/GR.126516.111.

[38] A. D. Kostic *et al.*, “Genomic analysis identifies association of Fusobacterium with colorectal carcinoma,” *Genome research*, vol. 22, no. 2, pp. 292–298, Feb. 2012, doi: 10.1101/GR.126573.111.

[39] S. Elsayed and K. Zhang, “Bacteremia caused by Clostridium symbiosum,” *Journal of Clinical Microbiology*, vol. 42, no. 9, pp. 4390–4392, Sep. 2004, doi: 10.1128/JCM.42.9.4390-4392.2004/ASSET/B89B4319-EA1D-43AE-8C46-006139A78CE7/ASSETS/GRAPHIC/ZJM0090445920001.JPEG.

[40] H. Rytter *et al.*, “The pentose phosphate pathway constitutes a major metabolic hub in pathogenic Francisella,” *PLOS Pathogens*, vol. 17, no. 8, p. e1009326, Aug. 2021, doi: 10.1371/JOURNAL.PPAT.1009326.

[41] D. A. Garsin, “Ethanolamine utilization in bacterial pathogens: roles and regulation,” *Nature Reviews Microbiology 2010 8:4*, vol. 8, no. 4, pp. 290–295, Apr. 2010, doi: 10.1038/nrmicro2334.

[42] W. Krzyściak, K. K. Pluskwa, A. Jurczak, and D. Kościelniak, “The pathogenicity of the Streptococcus genus,” *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 32, no. 11, p. 1361, Nov. 2013, doi: 10.1007/S10096-013-1914-9.

[43] Z. Marzhoseyni *et al.*, “Streptococcal bacterial components in cancer therapy,” *Cancer Gene Therapy 2021*, pp. 1–15, Mar. 2021, doi: 10.1038/s41417-021-00308-6.

[44] J. Yu *et al.*, “Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer,” *Gut*, vol. 66, no. 1, pp. 70–78, 2017, doi: 10.1136/GUTJNL-2015-309800.

[45] G. Zeller *et al.*, “Potential of fecal microbiota for early-stage detection of colorectal cancer,” *Molecular Systems Biology*, vol. 10, no. 11, p. 766, Nov. 2014, doi: 10.15252/MSB.20145645.

[46] “Colorectal cancer detection from fecal microbiota (ID 266076) - BioProject - NCBI.” https://www.ncbi.nlm.nih.gov/bioproject/266076 (accessed Jan. 21, 2022).

[47] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, Nov. 2014, doi: 10.1007/S10115-013-0679-X.

[48] N. Bar *et al.*, “A reference map of potential determinants for the human serum metabolome,” *Nature 2020 588:7836*, vol. 588, no. 7836, pp. 135–140, Nov. 2020, doi: 10.1038/s41586-020-2896-2.

[49] O. Manor and E. Borenstein, “Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome,” *Cell host & microbe*, vol. 21, no. 2, pp. 254–267, Feb. 2017, doi: 10.1016/J.CHOM.2016.12.014.

[50] Y. Y. Li *et al.*, “Association of Fusobacterium nucleatum infection with colorectal cancer in Chinese patients,” *World Journal of Gastroenterology*, vol. 22, no. 11, p. 3227, Mar. 2016, doi: 10.3748/WJG.V22.I11.3227.

[51] T. Löwenmark *et al.*, “Parvimonas micra as a putative non-invasive faecal biomarker for colorectal cancer,” *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–10, Sep. 2020, doi: 10.1038/s41598-020-72132-1.

[52] W. Sun, L. Wang, Q. Zhang, and Q. Dong, “Microbial Biomarkers for Colorectal Cancer Identified with Random Forest Model,” *http://www.xiahepublishing.com/*, vol. 5, no. 1, pp. 19–26, Jan. 2020, doi: 10.14218/ERHM.2019.00026.

[53] M. Rebersek, “Gut microbiome and its role in colorectal cancer,” *BMC Cancer*, vol. 21, no. 1, pp. 1–13, Dec. 2021, doi: 10.1186/S12885-021-09054-2/FIGURES/5.

[54] S. Terrisse *et al.*, “Intestinal microbiota influences clinical outcome and side effects of early breast cancer treatment,” *Cell Death & Differentiation 2021 28:9*, vol. 28, no. 9, pp. 2778–2796, May 2021, doi: 10.1038/s41418-021-00784-1.

[55] K. Fukuda, S. E. Straus, I. Hickie, M. C. Sharpe, J. G. Dobbins, and A. Komaroff, “The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study,” *Annals of Internal Medicine*, vol. 121, no. 12, pp. 953–959, Dec. 1994, doi: 10.7326/0003-4819-121-12-199412150-00009.

[56] V. C. Patel *et al.*, “Rifaximin-α reduces gut-derived inflammation and mucin degradation in cirrhosis and encephalopathy: RIFSYS randomised controlled trial,” *Journal of hepatology*, Sep. 2021, doi: 10.1016/J.JHEP.2021.09.010.

[57] J. Behary *et al.*, “Gut microbiota impact on the peripheral immune response in non-alcoholic fatty liver disease related hepatocellular carcinoma,” *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–14, Jan. 2021, doi: 10.1038/s41467-020-20422-7.

[58] K. Cusi *et al.*, “Non-alcoholic fatty liver disease (NAFLD) prevalence and its metabolic associations in patients with type 1 diabetes and type 2 diabetes,” *Diabetes, Obesity and Metabolism*, vol. 19, no. 11, pp. 1630–1634, Nov. 2017, doi: 10.1111/DOM.12973.

[59] Z. Ye *et al.*, “Altered gut microbiome composition in patients with Vogt-Koyanagi-Harada disease,” *Gut microbes*, vol. 11, no. 3, pp. 539–555, May 2020, doi: 10.1080/19490976.2019.1700754.

[60] “Fecal Microbiota and Gut Microbe-Derived Extracellular Vesicles in Colorectal Cancer.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8477046/ (accessed Jan. 21, 2022).

[61] Y. Wu *et al.*, “Identification of microbial markers across populations in early detection of colorectal cancer,” *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–13, May 2021, doi: 10.1038/s41467-021-23265-y.

[62] M. S. Riaz Rajoka *et al.*, “Interaction between diet composition and gut microbiota and its impact on gastrointestinal tract health,” *Food Science and Human Wellness*, vol. 6, no. 3, pp. 121–130, Sep. 2017, doi: 10.1016/J.FSHW.2017.07.003.

[63] W. Shen, H. R. Gaskins, and M. K. McIntosh, “Influence of dietary fat on intestinal microbes, inflammation, barrier function and metabolic outcomes,” *The Journal of Nutritional Biochemistry*, vol. 25, no. 3, pp. 270–280, Mar. 2014, doi: 10.1016/J.JNUTBIO.2013.09.009.

[64] A. Vich Vila *et al.*, “Impact of commonly used drugs on the composition and metabolic function of the gut microbiota,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–11, Jan. 2020, doi: 10.1038/s41467-019-14177-z.

[65] R. K. Weersma, A. Zhernakova, and J. Fu, “Interaction between drugs and the gut microbiome,” *Gut*, vol. 69, no. 8, pp. 1510–1519, Aug. 2020, doi: 10.1136/GUTJNL-2019-320204.

[66] Q. le Bastard *et al.*, “Systematic review: human gut dysbiosis induced by non-antibiotic prescription medications,” *Alimentary Pharmacology & Therapeutics*, vol. 47, no. 3, pp. 332–345, Feb. 2018, doi: 10.1111/APT.14451.

[67] H. M. Hamer, D. Jonkers, K. Venema, S. Vanhoutvin, F. J. Troost, and R. J. Brummer, “Review article: the role of butyrate on colonic function,” *Alimentary Pharmacology & Therapeutics*, vol. 27, no. 2, pp. 104–119, Jan. 2008, doi: 10.1111/J.1365-2036.2007.03562.X.

[68] S. Chen *et al.*, “Fusobacterium nucleatum promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7,” *Gut Microbes*, vol. 11, no. 3, pp. 511–525, May 2020, doi: 10.1080/19490976.2019.1695494/SUPPL\_FILE/KGMI\_A\_1695494\_SM0055.DOCX.

[69] M. A. Casasanta *et al.*, “Fusobacterium nucleatum host-cell binding and invasion induces IL-8 and CXCL1 secretion that drives colorectal cancer cell migration,” *Science Signaling*, vol. 13, no. 641, Jul. 2020, doi: 10.1126/SCISIGNAL.ABA9157/SUPPL\_FILE/ABA9157\_SM.PDF.

[70] T. Ogita, Y. Yamamoto, A. Mikami, S. Shigemori, T. Sato, and T. Shimosato, “Oral Administration of Flavonifractor plautii Strongly Suppresses Th2 Immune Responses in Mice,” *Frontiers in Immunology*, vol. 11, p. 379, Feb. 2020, doi: 10.3389/FIMMU.2020.00379/BIBTEX.

[71] G. A. Conte *et al.*, “Small Bowel Gastrointestinal Stromal Tumor as a Gateway for Streptococcus anginosus Causing Multiple Liver Abscesses,” *World journal of oncology*, vol. 11, no. 3, pp. 116–121, Jun. 2020, doi: 10.14740/WJON1270.

[72] J. D. F. Hale, R. Jain, P. A. Wescombe, J. P. Burton, R. R. Simon, and J. R. Tagg, “Safety assessment of Streptococcus salivarius M18 a probiotic for oral health,” *Beneficial microbes*, pp. 1–14, Jan. 2022, doi: 10.3920/BM2021.0107.

[73] H. Rytter *et al.*, “The pentose phosphate pathway constitutes a major metabolic hub in pathogenic Francisella,” *PLOS Pathogens*, vol. 17, no. 8, p. e1009326, Aug. 2021, doi: 10.1371/JOURNAL.PPAT.1009326.

[74] D. A. Garsin, “Ethanolamine utilization in bacterial pathogens: roles and regulation,” *Nature Reviews Microbiology 2010 8:4*, vol. 8, no. 4, pp. 290–295, Apr. 2010, doi: 10.1038/nrmicro2334.

[75] R. D. Sleator, J. Wouters, C. G. M. Gahan, T. Abee, and C. Hill, “Analysis of the Role of OpuC, an Osmolyte Transport System, in Salt Tolerance and Virulence Potential of Listeria monocytogenes,” *Applied and Environmental Microbiology*, vol. 67, no. 6, pp. 2692–2698, Jun. 2001, doi: 10.1128/AEM.67.6.2692-2698.2001/ASSET/49E38C45-4C3C-4900-8F08-81D24F36FA09/ASSETS/GRAPHIC/AM0611704004.JPEG.

[76] C. J. Doyle, P. W. O’Toole, and P. D. Cotter, “Genomic characterization of sulphite reducing bacteria isolated from the dairy production chain,” *Frontiers in Microbiology*, vol. 9, no. JUL, p. 1507, Jul. 2018, doi: 10.3389/FMICB.2018.01507/BIBTEX.

[77] D. Wu *et al.*, “Dietary pectic substances enhance gut health by its polycomponent: A review,” *Comprehensive Reviews in Food Science and Food Safety*, vol. 20, no. 2, pp. 2015–2039, Mar. 2021, doi: 10.1111/1541-4337.12723.

[78] J. Li *et al.*, “An integrated catalog of reference genes in the human gut microbiome,” *Nature biotechnology*, vol. 32, no. 8, pp. 834–841, Aug. 2014, doi: 10.1038/NBT.2942.

[79] N. Pons *et al.*, “METEOR -a plateform for quantitative metagenomic profiling of complex ecosystems,” Nov. 2010.

[80] S. Sunagawa *et al.*, “Metagenomic species profiling using universal phylogenetic marker genes,” *Nature Methods*, vol. 10, no. 12, pp. 1196–1199, 2013, doi: 10.1038/nmeth.2693.

[81] S. F. Altschul *et al.*, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/NAR/25.17.3389.

[82] J. R. Kultima *et al.*, “MOCAT: a metagenomics assembly and gene prediction toolkit,” *PLoS One*, vol. 7, no. 10, p. e47656, 2012, doi: 10.1371/journal.pone.0047656.

[83] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, 2004, doi: 10.1093/nar/gkh340.

[84] S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon, “trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses,” *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, 2009, doi: 10.1093/bioinformatics/btp348.

[85] M. N. Price, P. S. Dehal, and A. P. Arkin, “FastTree 2--approximately maximum-likelihood trees for large alignments,” *PLoS One*, vol. 5, no. 3, p. e9490, 2010, doi: 10.1371/journal.pone.0009490.

[86] I. Letunic and P. Bork, “Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation,” *Nucleic Acids Research*, vol. 49, no. W1, pp. W293–W296, Jul. 2021, doi: 10.1093/NAR/GKAB301.

[87] E. Ruppe *et al.*, “Prediction of the intestinal resistome by a three-dimensional structure-based method,” *Nat Microbiol*, vol. 4, no. 1, pp. 112–123, 2019, doi: 10.1038/s41564-018-0292-6.

[88] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat, “The carbohydrate-active enzymes database (CAZy) in 2013,” *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D490-5, 2014, doi: 10.1093/nar/gkt1178.

[89] O. Svartstrom *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*, vol. 11, no. 11, pp. 2538–2551, 2017, doi: 10.1038/ismej.2017.108.

[90] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using DIAMOND,” *Nat Methods*, vol. 12, no. 1, pp. 59–60, 2015, doi: 10.1038/nmeth.3176.

[91] C. Mao *et al.*, “Curation, integration and visualization of bacterial virulence factors in PATRIC,” *Bioinformatics*, vol. 31, no. 2, pp. 252–258, 2015, doi: 10.1093/bioinformatics/btu631.

[92] J. J. Gillespie *et al.*, “PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species,” *Infect Immun*, vol. 79, no. 11, pp. 4286–4298, 2011, doi: 10.1128/IAI.00207-11.

[93] S. Mukherjee *et al.*, “Genomes OnLine database (GOLD) v.7: updates and new features,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D649–D659, 2019, doi: 10.1093/nar/gky977.

[94] N. et al. Pons, “a platform for quantitative metagenomic profiling of complex ecosystems.,” *Journées Ouvertes en Biologie, Informatique et Mathématiques*, 2010. http://www.jobim2010.fr/sites/default/files/presentations/27Pons.pdf

[95] E. le Chatelier *et al.*, “Richness of human gut microbiome correlates with metabolic markers,” *Nature*, vol. 500, no. 7464, pp. 541–546, 2013, doi: 10.1038/nature12506.

[96] C. O. Fritz, P. E. Morris, and J. J. Richler, “Effect size estimates: current use, calculations, and interpretation,” *J Exp Psychol Gen*, vol. 141, no. 1, pp. 2–18, 2012, doi: 10.1037/a0024338.

[97] S. T.- Biorxiv and undefined 2014, “qqman: an R package for visualizing GWAS results using QQ and manhattan plots,” *biorxiv.org*, Accessed: Nov. 30, 2021. [Online]. Available: https://www.biorxiv.org/content/10.1101/005165v1.full-text

[98] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” *InterJournal, complex systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[99] P. Pons and M. Latapy, “Computing communities in large networks using random walks,” in *International symposium on computer and information sciences*, 2005, pp. 284–293.

[100] M. Uhlen *et al.*, “A pathology atlas of the human cancer transcriptome,” *Science*, vol. 357, no. 6352, 2017, doi: 10.1126/science.aan2507.

[101] T. P. Quinn, M. F. Richardson, D. Lovell, and T. M. Crowley, “propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis,” *Scientific Reports 2017 7:1*, vol. 7, no. 1, pp. 1–9, Nov. 2017, doi: 10.1038/s41598-017-16520-0.

[102] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, Accessed: Dec. 15, 2021. [Online]. Available: http://scikit-learn.sourceforge.net.

[103] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

# Methods

## Metagenomics species pan-genome (MSP) creation

The 1,601 metagenomic samples used to build the Integrated Gene Catalog of the human gut microbiome (IGC2) were downloaded from the European Nucleotide Archive [78]. Using the Meteor software suite[79] (available at https://forgemia.inra.fr/metagenopolis/meteor), reads from each sample were mapped against the IGC2 catalog and a raw gene abundance table was generated. This table was submitted to MSPminer[9] 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 were assessed by searching for 40 universal single copy marker genes[80] 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)[81]. 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 MOCAT[82]. MSPs with less than 5 markers were discarded. Next, the markers were separately aligned with MUSCLE[83]. The 40 alignments were merged and trimmed with trimAl[84]. Finally, the phylogenetic tree was computed with FastTreeMP[85] and visualized with iTOL[86]. Phylogenetic placement was further used to improve and correct taxonomic annotation. Pylogenetic data, species labels, and phylum colouring can be accessed from https://data.inrae.fr/dataset.xhtml?persistentId=doi:10.15454/FLANUP, with annotation for enriched species found at https://github.com/sysbiomelab/ATLAS.

## 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/)[87]. 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 CAZyme family[88], following a procedure previously described for other metagenomics analysis[89]. Proteins of IGC2 catalog were also annotated to KEGG orthologous using Diamond (version 0.9.22.123)[90] 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[91], [92] were matched against IGC2[17] by BLASTP (best identity > 50%, e-value < 10-10). Phenotypes of MSPs were manually checked and annotated based on JGI-GOLD phenotype (organism metadata)[93]. We identified biosynthetic genes of MSPs with the use of the standalone anti-SMASH program with minimal run option, focused on core detection modules (version 5)[24]. Loading antiSMASH into Amazon cloud computing (AWS) as a 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 S1) on IGC2 catalogue of human gut metagenome by METEOR[94] 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[95]. MSP abundances were estimated by the mean abundance of its 100 ‘marker’ genes (that is, the genes that correlate the most altogether). If less than 10% of ‘marker’ genes were seen in a sample, the abundance of the MSPs was set to 0. 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 [26]. 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 and genes from geographically distinct cohorts

Regional enrichment of species was calculated from calculating the Z-score for of each MSP. The Z-score is calculated as the difference between the mean relative abundance for each country against the whole population. Displayed are the top 20 greatest mean species Z-scores of western and non-western groups.

By selecting the top 100 overrepresented MSPs in western and non-western groups, two separate cumulative sums of their genes were filtered to have more than 90 genes. The genes in each of these lists were mapped against the CAZy, PATRIC, and CARD databases. 18 of the max differences between the western and non-western gene count lists were calculated and plotted.

## Pan-metagenomics association studies (Pan-MGAS)

First, we selected 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 S1). 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[96] and 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[97].

## Unsupervised clustering of co-conserved functions of gut microbiota

We calculated Jaccard index among functional annotations to calculate number of species that share a pair of functions. We selected highly shared pairs of functions (Jaccard index >= 0.75) and merged into functional co-occurrence network using R *igraph* package[98]. Functional clusters within the network were identified by unsupervised community detection, short random work algorithm (*cluster\_walktrap* function)[99], [100] and identified singleton functions within the network. 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. MSPs were associated to functional if the given MSP covered more than 75% functions of the functional cluster (Supplementary Table S5).

## Proportionality between MSPs

Proportionality was estimated using the *propr* R package [101]. We used as input the relative abundance matrix of all samples against MSP. Only MSPs with values above 0 in more than 50 samples were included. FDR cutoff values were estimated using the *propr* function *updateCutoffs.* We made a network representation of the resulting MSP pairs with proportionality values above 0.65.

## Random forest classification model to predict disease phenotype

We trained a random forest classifier with hyperparameters 'bootstrap': True, 'ccp\_alpha': 0.0, 'class\_weight': None, 'criterion': 'gini', 'max\_depth': None, 'max\_features': 'auto', 'max\_leaf\_nodes': None, 'max\_samples': None, 'min\_impurity\_decrease': 0.0, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'min\_weight\_fraction\_leaf': 0.0, 'n\_estimators': 500, 'n\_jobs': -1, 'oob\_score': False, 'random\_state': 1, 'verbose': 0, 'warm\_start': False to distinguish between diseases that contained a minimum of 40 samples (17 diseases) and a random, stratified sample from 100 healthy samples from all cohorts using the Scikit-learn python package [102]. Firstly, relative abundance data was standardised using the scikit-learn implementation of the StandardScaler function. Training and testing was performed on randomly selected samples split 70% and 30% of the full diseased dataset respectively with a fixed random seed to ensure reproducibility of the model. Model performances were measured using AUCROC scoring. The python implementation of the explainable AI algorithm, Shapley Additive exPlanations (SHAP), was used to show species contribution to disease classification [103]. The mean absolute SHAP score for each disease predictive model was given direction using the sign of the spearman rank correlation between the feature value and the SHAP score. Positive values indicate that higher relative abundance is more likely to classify the disease versus healthy samples. Negative values indicate that lower relative abundance is more likely to classify the disease versus healthy samples. Right colour bar indicates mean species bias for enrichment or depletion in all disease.

# Figure legends

**Fig. 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, 5,708 shotgun metagenome samples, were aligned against the gene catalogue of the human gut microbiome and quantified at the level of MSP. **D,** heatmap showing the top 15 overrepresented MSPs between western and non-western cohorts coloured by mean species Z-score for each country against all countries. **E,** *monocle* ordination of the gut microbiome. Individual samples from non-westernized countries, European countries, and US/China/Japan were coloured green, orange, and blue, respectively. **F,** stacked bar plots of contrasting functions among region-enriched species classified as non-westernized or westernized. Based on gene functional annotations of CAZyme, antimicrobial resistance (AMR), and virulence factors (PATRIC database), we calculated regional functional overrepresentation by cumulatively summing and filtering by top 18 maximal differences of gene count (Methods).

**Fig. 2. Pan-metagenomics association studies (Pan-MGAS) of 43 cohorts from 23 different diseases and 14 countries (n=2,185).** **A,** We identified significantly enriched/depleted species of cohorts based on effect sizes (ESs) of Wilcoxon one-sided tests (ES ≥ 0.3); the list is displayed in Supplementary Table 3. Acronyms are: ACVD, Acute Coronary Cardiovascular Disease; Ob, obesity; CRC, Colorectal Cancer; NSCLC, Non-Small Cell Lung Cancer; RCC Renal Cell Carcinoma; GDM, Gestational Diabetes Mellitus; T1D Type 1 Diabetes; T2D, Type 2 Diabetes; LC Liver Cirrhosis; NAFLD Non-Alcoholic Fatty Acid Liver; UC, Ulcerative Colitis; CD, Crohn's disease; BD Behçet's disease; RA, Rheumatoid Arthritis; SPA, Ankylosing Spondylitis; ME/CFS Myalgic Encephalomyelitis/ Chronic Fatigue Syndrome; PD, Parkinson Disease. **B,** Jitter plots of frequency of the significantly enriched/depleted cohorts of all MSPs (effect size >0.3) were calculated: total frequency of enriched/depleted cohorts (number of enriched cohorts + number of depleted cohorts Y axis) and subtracted frequency between enriched cohorts and depleted cohorts (number of enriched cohorts - number of depleted cohorts X axis). 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. **c**, Species found depleted (*Anaerostipes hadrus*) and enriched (*Fusobacterium nucleatum subspecies animalis*) in most studies.

**Fig. 3. Analysis of functional clusters.** 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 the microbiome. **A,** Identification of functional clusters based on co-conserved genes across species. Unlike the manually curated module database, we identified functional clusters based on high co-conservation across species using the unsupervised clustering method. **B,** among different sources of microbial functional annotations (e.g., KEGG module and pathway), we found that co-conservation of genes across different species was substantially low (Jaccard index < 0.5). **C**, Functional clusters identified by unsupervised community detection, the y-axis displays the number of genes within the functional cluster and the x-axis displays the number of MSPs possessing more than 70% of the clusters’ genes. **D**, Functional clusters projected on enriched/depleted MSPs across disease cohorts. The Jitter plot display the frequency functional of functional clusters significantly associated with the enriched/depleted species (hypergeometric test p < 0.0001) in disease cohorts. Y-axis shows the total frequency of cohorts where a functional cluster was found significantly associated with enriched/depleted species. X-axis shows the difference in the number of cohorts where a function was found enriched minus the frequency it was found depleted. 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.

**Fig. 4. Phylogenetic differences between species function, disease enrichment, and region enrichment.** Inner annotation of dendrogram is species phylum, second is enrichment of functional cluster, third is the total number of disease cohorts that the species is enriched/depleted in, and the outer annotation is the normalised, mean Z-score between western and non-western cohorts scaled between 0-1. Itol annotations and dendrogram are publicly available (Methods). Highlighted group are MSPs from the *Streptococcus* genus.

**Fig 5. Random forest models trained on multiple cohorts to discriminate between disease and healthy controls.** Acronyms are: CRC, Colorectal Cancer; NSCLC, Non-Small Cell Lung Cancer; RCC Renal Cell Carcinoma; T1D Type 1 Diabetes; T2D, Type 2 Diabetes; LC liver Cirrhosis; NAFLD Non-Alcoholic Fatty Acid Liver; CD, Crohn's disease; RA, Rheumatoid Arthritis; SPA, Ankylosing Spondylitis; ME\_CFS Myalgic Encephalomyelitis/ Chronic Fatigue Syndrome; IGT, Impaired Glucose Tolerance; VKH, Vogt-Koyanagi-Harada **A)** Schematic of Random forest classification method. **B)** AUCROC scores for each disease random forest classification model C) AUCROC curves of an inter (top) and intra (bottom) cohort validation for a RF model that predicts CRC. D) Boxplot of directional mean absolute Shapley Additive Explanations (SHAP) scores for all disease predictive models. Red and blue boxes represent species that were depleted/enriched using Effect Size calculation. D) ClusteredHeatmap of the most important species for prediction of 16 diseases by random forest classification as calculated by directional mean SHAP score (rows contain at least one species with directional mean SHAP score above 0.0125 in any of the diseases) (methods). Positive values indicate that higher relative abundance is more likely to classify the disease versus healthy samples. Negative values indicate that lower relative abundance is more likely to classify the disease versus healthy samples. Right colour bar indicates mean species bias for enrichment or depletion in all disease.

# Supplementary Figure Legends

**Supplementary Fig. 1.** Species found either depleted or enriched (effect size >0.3) in at least 6 different disease cohorts.

**Supplementary Fig. 2. Analysis of functional clusters. A,** Overview of the identification of functional clusters. **B,** community network detected from the co-occurrence network of functional annotations. All functional clusters were shown as nodes and significant co-occurrences of functional clusters across species were shown as edges in the network. “Comm-cluster” and “patho-cluster” were coloured blue and red, respectively. **C,** number of functional clusters mapped with microbial function/phenotype databases and their singleton cluster fractions (singleton and non-singleton coloured green and red, respectively).

**Supplementary Fig. 3 A)** effect size across all cohorts in the Streptococcus genus. The figure displays the phylogenetic tree branch containing the MSPs from the Streptococcus genus and the estimated effect size across all disease cohorts included in this study. The bar plots on the left side show the number of cohorts where an MSP‘s effect size was above 0.3. Red colour indicates depletion and blue colour, enrichment. The graph on the top shows the hierarchical clustering result using the Ward method on the complete Effect Size matrix. B) Network representation of proportionality values between MSPs. Nodes represent MSPs and edges represent proportionality values above 0.65 between a pair of MSPs. Node fill colour indicates phylum, node edges in red highlight MSP with effect size > 0.3 in more than two cohorts, edge width is proportional to the proportionality value. The yellow oval in the background highlights the Clusters containing MSPs from the Streptococcus genus

## **Supplementary Tables Legends**

**Supplementary Table S1.** Description of Human Gut Microbiome Atlas (HGMA) datasets used in this study. We showed the overall statistics of HGMA datasets, including total number of samples, number of samples without intervention and multiple visits, number of matched healthy controls, sequencing platform, geography, reference, and raw data link.

**Supplementary Table S2.** List of region enriched MSPs. Based on effect sizes (>0.3 in more than six countries of comparison), we identified total 782 regionally enriched MSPs among healthy samples of 17 countries.

**Supplementary Table S3.** Statistics of pan-metagenomics association study (pan-MGAS) results of 28 geographically distinct disease cohorts. We performed pan-MGAS analysis and identified enriched/depleted species in each cohort with effect sizes by each column.

**Supplementary Table S4.** Statistics of MSP common enriched/depleted in diseases. We examined the total and subtracted frequency of significantly enriched and depleted species in diseases.

**Supplementary Table S5.** Characteristics of 7,763 functional clusters. For a more detailed understanding, per functional cluster we presented the size of cluster, number of enriched species, enriched MSP identifiers, enriched species names, enriched KEGG modules (hypergeometric tests p-value < 0.01), subsystems of enriched KEGG modules, and KEGG orthologs terms, PFAM terms, virulence term, CAZyme terms, anti-microbial resistance (Mustard) terms, secondary metabolism (anti-SMASH) terms, phenotype (JGI-GOLD) terms, product names of virulence terms, and all functional terms of a given cluster.