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

Sunjae Lee1\*, Theo Portlock2\*, Emmanuelle Le Chatelier3\*, Jose Garcia2\*, Nicolas Pons4, Florian Plaza Onate4, Neelu Begum1, Ceri Proffitt1, Dorines Rosario1, Stefania Vaga1, Junseok Park5, Kalle von Feilitzen2, Fredric Johansson2, Azadeh Harzandi1, Cheng Zhang, Lindsey A. Edwards7, Vincent Lombard8,9, Franck Gauthier4, Claire J. Steves10, David Gomez-Cabrero1,11, Bernard Henrissat8,9,12, Doheon Lee5, Debbie L. Shawcross7, Gordon Proctor1, Jens Nielsen14,15,16, David Moyes, Adil Mardinoglu , Stanislav Dusko Ehrlich4, Mathias Uhlen2,16, Saeed Shoaie

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 School of Life Sciences, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea, 61005

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

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

6 Université Clermont Auvergne, INRAE, UMR 454 MEDIS, 28 place Henri Dunant, F-63000, Clermont-Ferrand, France

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

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

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

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

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

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

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

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

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

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

These authors contributed equally.

Email: [saeed.shoaie@kcl.ac.uk](mailto:saeed.shoaie@kcl.ac.uk), [dusko.ehrlich@inrae.fr](mailto:dusko.ehrlich@inrae.fr), [mathias.uhlen@scilifelab.se](mailto:mathias.uhlen@scilifelab.se)

# Abstract

The role of gut microbiota in humans is of great interest, and metagenomics provides the possibilities for extensively analysing bacterial diversity in health and disease. Here we explored the human gut microbiome samples across 19 countries, performing compositional, functional, and integrative analysis. We present the Human Gut Microbiome Atlas (HGMA), a tool available for public access. Through using the HGMA, the combination of species interrogation across regions, diseases, and functional clusters allows for a more complete understanding of the importance a species has in human health and disease, and in aiding in the generation of informed hypothesis.

# Introduction

Metagenomic studies of the human microbiome enable the characterization of the microbial and functional diversity in health and disease[1]. Advances in metagenome assembly and various clustering methods enabled the generation of metagenome species[2]–[6]. Specifically, the negative selective microbe-microbe and host-microbe interactions, in the context of metabolism or antimicrobial machinery, could be the main mechanism underlying microbial dysbiosis[10]. 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[11].

Herein, we integrated publicly available data from many studies across different countries from healthy and diseased individuals. The analysis is presented in an open-access Human Gut Microbiome Atlas (www.microbiomeatlas.org), allowing researchers to explore for the first time an integrative analysis on composition, functional, richness, diseases, and region signatures for the gut microbiota across 19 geographical regions and 23 diseases.

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

We performed a large-scale integrative analysis of 6,014 publicly available shotgun metagenomics stool samples with addition to one Swedish cohort of 344 samples (under submission). There were at least 10 million high-quality sequencing reads from healthy and diseased cohorts from 19 different countries across five continents (Fig. 1a-b and Supplementary Table S1). We rarefied all metagenomic samples into 10 million reads per sample, which enables comparative analysis across different cohorts. We created the Human Gut Microbiome Atlas (HGMA) using quantitative analysis of shotgun metagenomics based on microbial genomes assembled using Metagenomic Species Pan-genomes (MSPs) (Fig. 1c). The MSP number was increased from 1,661 (previous release[5]) to 1,989 (average number of genes 1,894 ± 1,616) (Methods), and their taxonomy was updated. We generated gene counts and MSP abundances for all the samples using the 10.4 million gene catalogue[12]. We also characterized the functions and phenotype of the MSPs in 7 different categories (KO, PFAM, CAZyme, Mustard, JGI-GOLD phenotype, PATRIC virulence factor, and antiSMASH biosynthetic gene clusters) and identified co-conserved functional clusters across species (7,763 clusters) (Methods).All the data are freely available in the HGMA, without restrictions, in the public open access database (www.microbiomeatlas.org) that is part of the Human Protein Atlas program (www.proteinatlas.org). All MSPs and functions are highlighted together with the 6,014 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. Enrichment of *Clostridium* and *Bacteriodes* were found to have higher mean relative abundance withing western countries whereas *Prevotella* were had higher mean relative abundance withing non-western countries (Fig. 1d). We applied the unsupervised clustering method, *monocle*, to MSP abundance profiles of all samples (Methods)[13], [14]. We 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 also identified 624 MSPs specifically enriched in certain countries (See Methods, and Supplementary Table S2). The following analysis was based on the western/non-western groups. Functional annotation-based analysis across geographical clusters revealed enrichment of CAZymes for degrading host mucins and storage carbohydrates in westernized populations, where antimicrobial resistance (AMR) and virulence factors were also more prevalent (Fig. 1f). Comparison of functions of region enriched MSPs in westernised countries revealed that genes encoding for vancomycin resistance, LPS biogenesis, and mucin degradation are overrepresented. An overrepresentation of genes encoding for complex polysaccharide binding proteins mostly belonging to *Prevotella* genus was found in the non-westernized cohorts.

# Pan-metagenomics association study across 23 diseases

To distinguish diseased and healthy microbiomes from multiple cohorts, we performed a Pan-metagenomics association (Pan-MGAS) of multiple disease cohorts (23 diseases across 43 cohorts from 14 countries). We determined the enriched and depleted species within 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 such as cohorts Non-Small Cell Lung Cancer (NSCLC, from France), renal cell carcinoma (RCC frorm France), adenoma frorm Italy (Fig. 2a). Conversely in some diseases had several enriched species, as we can see 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* areamong the most depleted species withall these examples found depleted in at least six different cohorts (Fig. 2b, Supplementary Fig. S1)*.* The firsttwo species have been described as butyrate producers and dominant species isolated from the healthy human colon[15]–[17], and the third has been associated with gut microbiota recovery after cholera infection and with normal maturation of the infant gut microbiota[18].

Between the species found enriched and 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*[19], *P. stomatis*[20]*, P. micra*[21]) and some of them have been identified in infections including bacteremia (*C. bolteae*[22], *C. clostridioforme*[23], *P. micra*[24]). Besides *F. nucleatum* and *C. symbiosum*, which were enriched in western countries and are linked with CRC[25]–[27], 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

We identified co-conserved functional clusters of the microbiome by applying an unsupervised clustering approach on MSPs (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). From the community detection algorithm, we identified 7,763 functional clusters, 6,297 singletons, and 591 representative clusters (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*” herewith) was over-represented among many commensal species, while the other (CL-10, named “*patho-cluster*”) was enriched in a few pathobionts, such as *Klebsiella* spp., *Enterobacter* spp., and *E. coli*. Interestingly, the *comm-cluster* was enriched with genes involved in the biosynthesis of amino acids indicative of functions. In contrast, the *patho-cluster* was enriched in functions associated with the uptake of several substrates. These included siderophore, ion, amino acid, and vitamin transport, thus competing with host and commensal bacteria. We also found other enriched-functional clusters, such as butyrate metabolism, propionate metabolism, vitamin B12, coenzyme metabolism, chemotaxis, ATPase, and mobile genetic elements (i.e., integrase and transposase) and the CRISPR-cas system (Fig. 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. Between them we could mention CL-1006, related to antibiotic resistance; CL-1032, a competence-related DNA transformation transport, which could provide an advantage to integrate new functions into the genome; or clusters related to metabolic pathways that could contribute indirectly to pathogenicity like the Pentose Phosphate Pathway [28] or Ethanolamine utilization [29]. Among the most frequent functional clusters that accompany the depleted species in disease we find the CL-12 comm-cluster, and some other clusters with functions related with pectate degradation and biofilm formation (Fig. 3d), all of them related with the normal function of the healthy microbiota. Global view of the function, region, and composition of disease enriched/depleted gut microbiome species

To capture a holistic view of the gut Pan-MGAS across diseases and geographical regions together with changes of the functional cluster, phylogenetic tree of the MSPs was calculated and generated with taxonomic resolution of disease and region enriched species (Fig. 4, Methods). Employing this tree as an interactive tool, it was possible to group species also by shared functional clusters that are enriched in disease in their phylogenetic context (Fig. 4). An example of how this tool can be used to analyse a metagenomic species is demonstrated with *Streptococcus* (Fig. 4).

Three different species belonging to the *Streptococcus* genus (*S. anginosus,* *S. parasanguinis* and *S. vestibularis*) and three more species from the *Veillonella* genus (*V. atypica*, *V. dispar*, *V. parvula*) were enriched in two different liver disease cohorts, all of which are common inhabitants of the oral cavity (Supplementary Table S4). Here, it’s enrichment in liver disease (ES=0.43) can be compared with its depletion in cancer (melanoma) (ES=0.30) across multiple species and cohorts. Another example can be found when looking at the species specificity of enrichment of *Fusobacterium nucleatum* in cancers. The species was not shown to be regionally enriched but shared functional clusters with this species also contribute to multiple cancers and in inflammatory diseases such as liver disease (Fig. 2c).

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

We implemented feature selection based random forest classifier prediction models trained using the MSPs 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 IGT, VKH, NSCLC, melanoma, advanced adenoma, and CD (AUCROC = 1.0, 1.0, 0.96, 0.95 0.92, 0.91 respectively).

The generalization of these models was assessed with an interstudy cross validation which demonstrated that a model trained on the CRC training cohort PRJEB10878 [30] was able to maintain high predictive precision of disease classification of the CRC test cohort PRJEB6070 [31] (Fig . 5A) (AUCROC = 0.68). Additional validation of the importance of the randomly selected healthy samples was done by selecting 30 random groups of healthy samples and repeating the cross validation. It was shown that the AUCROC of predicting the test cohort was 0.69 ± 0.04 showing low variability in predictive capabilities.

The interpretable machine learning framework, SHAP, was used to identify disease specific gut microbiome features*.* The disease classification models were able to reproduce importance of metagenomic species as biomarkers for several diseases in the HGMA. As with the effect size method for biomarker identification, the number of shared species deemed important for disease prediction between diseases was low. Of note, an increase in abundance of the commensal oral bacteria *Haemophilus parainfluenzae,* *Veillonnella dispar*, *Veillonella atypica*, and *Veillonella parvula* were shown to have high importance in predicting liver cirrhosis but not NAFLD, as found previously [32], and were found to be enriched in multiple cohorts regardless of region (Fig. 2). In the NAFLD model, an increase in abundance of *Streptococcus parasanguinis* was the most important factor for the prediction of 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 [33]. In the prediction of VKH, enrichment *Eubacteria eligens* and *Paraprevotella clara* also demonstrated high feature importance for the prediction of the disease; enrichment of which in disease was also found previously [34]. Interestingly, in the highest Z-score adjusted SHAP scoring species for the CRC predictive model, there were several absent biomarkers found previously to be important in CRC diagnosis being absent including *F. nucleatum*. However, presence of *Ruminococcus torques* was the greatest contributing factor to the prediction of CRC; an association that has been shown previously [35]. Models for the prediction of CRC and adenoma had similar profiles of species important for their predictions. A discrimination of this difference has been investigated in more detail previously [36]. Additionally, there were some shared species, such as *Acidaminococcus intestini* and *Faecalibacterium prausnitzii* with their presence and absence characterising a general dysbiotic state respectively.

# Discussion

We have performed a comprehensive integrative analysis of global and temporal gut microbiomes, and we provide an open access HMGA portal (http://microbiomeatlas.org). This tool allows for the integration of several studies simultaneously that link species to disease, region, and function. It also presents a means of contextualising gene and species enrichments phylogenetically. We demonstrated that 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 and that 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 the stratification of patients or for designing intervention studies. Beyond previous observations, our function-based analysis indicates that the western-enriched bacteria might dominate the gut microbial community with the production of antimicrobial peptides and homoserine lactone, which may inhibit their competitors.

The physiological changes caused by the 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 species enrichment 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[37], [38] or the use of drugs for treating the disease [39]–[41]).

The loss of species actively contributing to maintain 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 [42]. We also observed that most important species for the prediction of disease were commensal pathogens. Conversely, some of the enriched species might worsen the health status, 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[43], [44]. A previous report found *Flavonifractor plautii*, a species we found enriched in 6 cohorts, suppresses the Th2 immune responses in mice[45] which makes us speculate it could exert a similar effect in the humans. The Pan-MGAS we present shows a clear bias toward CRC studies due to the increased availability of these studies. We expect new studies released in the future to include more countries and diseases that would help to update our analysis and balance this bias.

The projection of functions associated to enriched/depleted species in disease support the observations made with species alone. The functions found commonly enriched in disease suggest they provide their carriers better chances to thrive in altered conditions, playing indirect roles in pathogenesis, for example using additional carbon sources (CL543-pentose phosphate pathway[46], ethanolamine [47]) or giving them the ability to survive environmental stresses CL-592 osmoprotectan cluster [48]. The enrichment of this functions does not imply they are exclusive of pathogenic organisms, anaerobic sulphite reducing activity is often used as a marker for food contamination [49] the activity is also present in several non-pathogenic bacteria. Conversely, functions depleted multiple times across different diseases could be playing an active role in health maintenance, for example, recent research has revealed that pectic substances can inhibit gut inflammation and relieve inflammatory bowel disease symptoms [50].

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

# 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 functional cluster analysis can be applied on gene counts and species abundances. The other pipeline scripts for analysis are also shared publicly and can be found 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., T.P, J.G. and S.S. led the design and analysis of the data. S.L., T.P, and J.G developed the temporal pipeline, analysis, and made the figures. L.E. and M.U. provided the wellness gut metagenomics samples. M.A., F.P., E.L., and S.D.E. generated the MSPs, performed quality check and taxonomy update. N.P. annotated the updated gut gene catalog. M.A., V.M. and F.P. performed the analysis on the Italian and American cohorts for validation. N.B., C.P., S.V., D. R. and A.H. analyzed part of the data and prepared the materials for the HGMA. K.F. and F.J. developed the HGMA website. V.L. and B.H. annotated the gut catalog with new CAZymes. J.P. and D.L. annotated the secondary metabolites of the gene catalog. M.A. and G.B. contributed to testing the pipeline, statistical and functional analysis. S.S., S.L. and T.P. wrote and drafted the manuscript. L.A.E, D.L.S, A.M., G.P. J.N. provided critical feedback on the data and manuscript. All authors read, edited and reviewed the manuscript.

# Competing interests

The authors declare no competing financial interests.

# Additional information

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

# References

[1] 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] L. A. David *et al.*, “Host lifestyle affects human microbiota on daily timescales,” *Genome Biol*, vol. 15, no. 7, p. R89, 2014, doi: 10.1186/gb-2014-15-7-r89.

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

[10] L. Dethlefsen and D. A. Relman, “Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108 Suppl 1, pp. 4554–4561, 2011, doi: 10.1073/pnas.1000087107.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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

[69] 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.

[70] 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.

[71] 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

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

[73] 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.

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

[75] 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.

[76] H. T. Nguyen, T. B. Tran, Q. M. Bui, H. H. Luong, T. P. Le, and N. C. Tran, “Enhancing disease prediction on imbalanced metagenomic dataset by cost-sensitive,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, pp. 651–657, 2020, doi: 10.14569/IJACSA.2020.0110778.

# Methods

## Metagenomics species pan-genome (MSP) creation

1601 metagenomic samples used to build the Integrated Gene Catalog of the human gut microbiome (IGC2) were downloaded from the European Nucleotide Archive [51]. Using the Meteor software suite[52], reads from each sample were mapped against the IG1C2 catalog and a raw gene abundance table was generated. This table was submitted to MSPminer[5] 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[53] 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)[54]. 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[55]. MSPs with less than 5 markers were discarded. Then, the markers were separately aligned with MUSCLE[56]. The 40 alignments were merged and trimmed with trimAl[57]. Finally, the phylogenetic tree was computed with FastTreeMP[58] and visualized with iTOL[59]. 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/)[60]. Protein sequences were aligned against 9,462 ARD sequences using *blastp* 2.7.1+ (option -evalue = 10-5). Best-hit alignments were filtered for identity ≥ 95% and bidirectional alignment coverage ≥ 90% (at query and subject level), giving a list of ARD candidates belonging to 30 families. Annotation of the carbohydrate-active enzymes (CAZymes) of the IGC2 catalog was performed by comparing the predicted protein sequences to those in the CAZy database and to Hidden Markov Models (HMMs) built from each CAZy family[61], following a procedure previously described for other metagenomics analysis[62]. Proteins of IGC2 catalog were also annotated to KEGG orthologous using Diamond (version 0.9.22.123)[63] 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[64], [65] were matched against IGC2[12] by BLASTP (best identity > 50%, e-value < 10-10). Phenotype of MSP were manually checked and annotated based on JGI-GOLD phenotype (organism metadata)[66]. We identified biosynthetic genes of MSP with the use of standalone anti-SMASH program with minimal run option, focused on core detection modules (version 5)[67]. Loading antiSMASH into Amazon cloud computing (AWS) as docker image, we executed its mining process per MSP in a massive parallel setting.

## Quality control/normalization of gene counts and species abundance profiling

We filtered out human reads and then mapped metagenomic data (Supplementary Table S1) on IGC2 catalogue of human gut metagenome by METEOR[68] 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[69]. MSP abundances were estimated by the median abundance of the 25 marker genes representing the robust centroid of gene clusters of MSP. Sample metadata of all metagenomics data such as sequencing platform, geography, age, body-mass index, gender and the data source were provided under HGMA (http://microbiomeatlas.org).

## Tracing the diversifications of healthy metagenomic samples of different geography

After the quantification and per-million scaling of MSP abundance profiles, we employed trajectory analysis in R *monocle* ver.2 package to identify how samples were clustered[14]. 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 CAZyme, 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[70] 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[71].

## 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[72]. Functional clusters within the network were identified by unsupervised community detection, short random work algorithm (*cluster\_walktrap* function)[73], [74] 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).

## Random forest classification model to predict disease phenotype

We trained a random forest classifier 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 [75]. Training and testing was performed on randomly selected samples split 70% and 30% of the full data respectively with a fixed random seed to ensure reproducibility of the model. Model performances were measured using AUCROC scoring with confusion matrices generated by applying the model to the 30% test set. Cross validation

# very similar ML approach <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7120658/>

# cross study validation for CRC https://www.nature.com/articles/s41598-021-04182-y

# will use brier score to adjust for imbalanced classifier[76]

# 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, 6,014 shotgun metagenome samples, including 344 Swedish longitudinal samples, were aligned against the gene catalogue of the human gut microbiome and quantified at the level of MSP. **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). 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 Becet’s; 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. Performance of ML models trained on multiple cohorts to discriminate between types of disease and healthy controls. A)** MeanAUCROC scores for random forest classification models for the prediction of each disease **B)** Left: Heatmap of the most important features (rows contain at least one species with > 2 Z-score adjusted SHAP values in any of the diseases in the columns) for the prediction of 17 diseases by random forest classification. Right: Species ES calculations (depleted species transformed as negative ES) filtered by Z-score adjusted SHAP values > 2 for classification of disease. **C)** ROC curves showing the validation of the PRJEB10878 training cohort against the PRJEB6070 testing cohort.

# 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 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 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 orthology 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 given cluster.