

# Supplemental Material

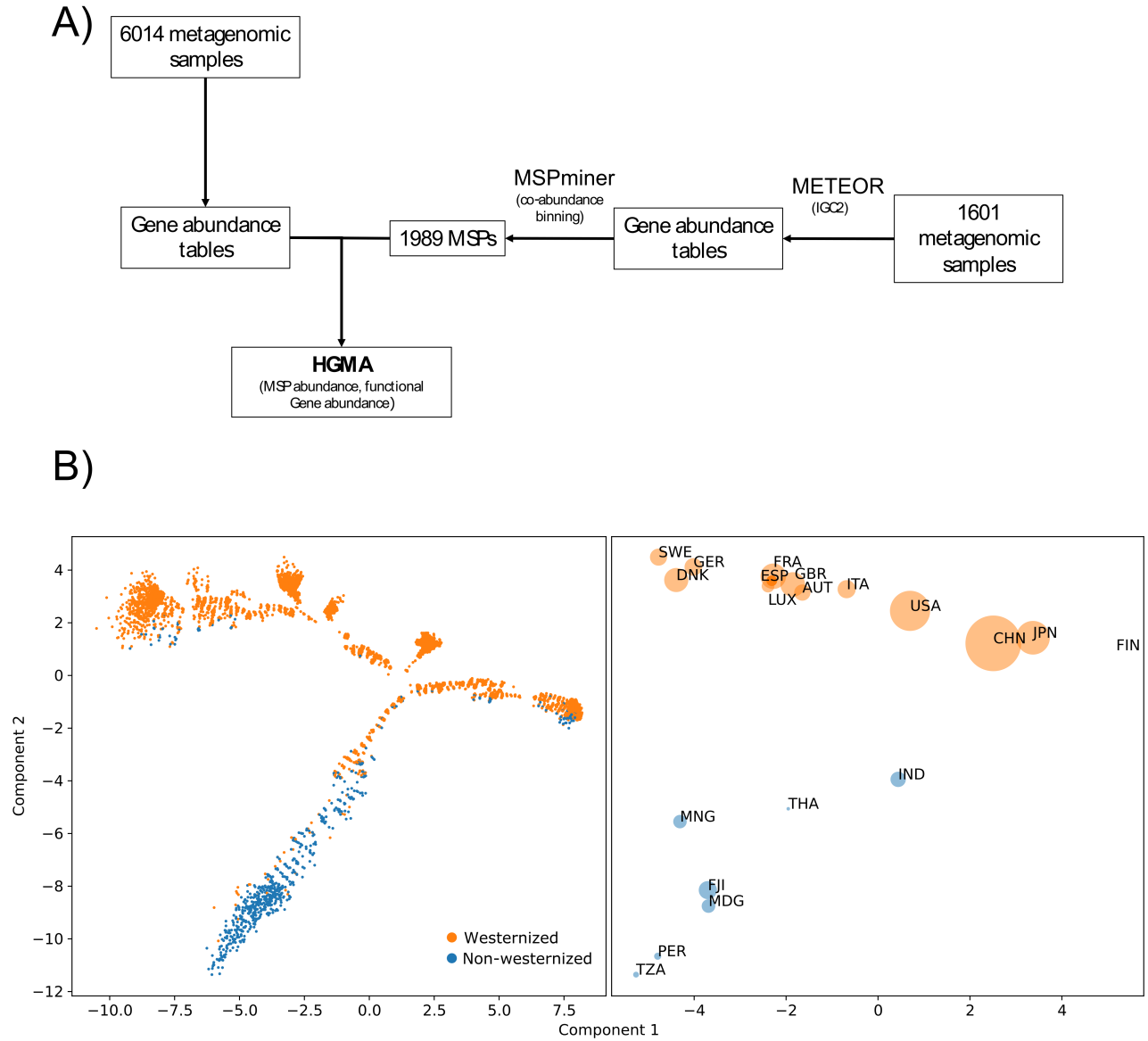
## Global compositional and functional state of the human gut microbiome in health and disease

Sunjae Lee, Theo Portlock, Emmanuelle Le Chatelier,  
Fernando Garcia-Guevara, Frederick Clasen,  
Florian Plaza Onate, Nicolas Pons, Neelu Begum,  
Azadeh Harzandi, Ceri Proffitt, Dorines Rosario,  
Stefania Vaga, Junseok Park, Kalle von Feilitzen,  
Fredric Johansson, Cheng Zhang, Lindsey A. Edwards,  
Vincent Lombard, Franck Gauthier, Claire J. Steves,  
David Gomez-Cabrero, Bernard Henrissat,  
Doheon Lee, Lars Engstrand, Debbie L. Shawcross,  
Gordon Proctor, Mathieu Almeida, Jens Nielsen  
Adil Mardinoglu, David L. Moyes, Stanislav Dusko Ehrlich,  
Mathias Uhlen, Saeed Shoaie

### Table of contents

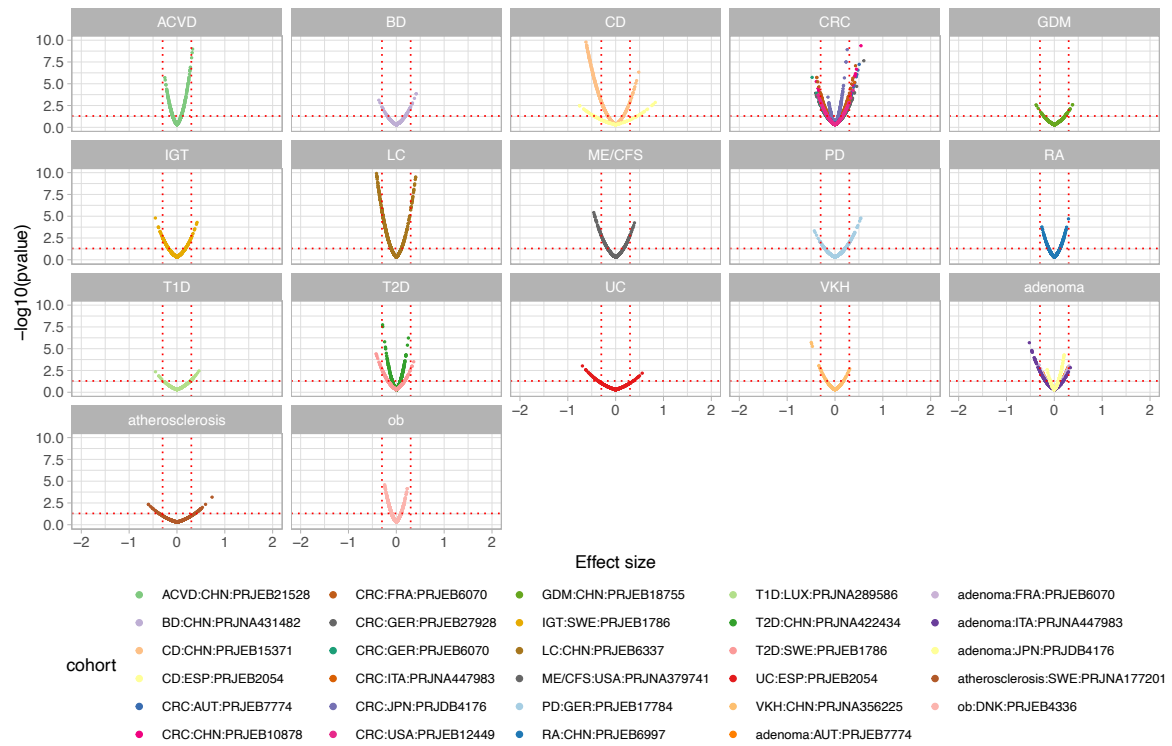
|          |                                                          |           |
|----------|----------------------------------------------------------|-----------|
| <b>1</b> | <b>Supplemental Figures</b>                              | <b>2</b>  |
| <b>2</b> | <b>Supplemental Table Legends</b>                        | <b>10</b> |
| <b>3</b> | <b>Supplemental Methods</b>                              | <b>11</b> |
| 3.1      | Metagenomics species pan-genome (MSP) creation . . . . . | 11        |

# 1 Supplemental Figures



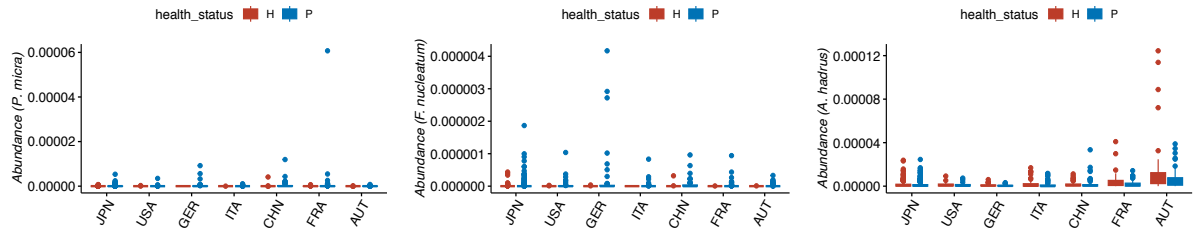
**Supplemental Figure S1.** **A**, Flow diagram depicting the workflow used to reconstruct MSPs and the Human Gut Microbiome Atlas (HGMA). **B**, Monocle ordination of the gut microbiome related to Figure 1E where nodes show country of origin.

A

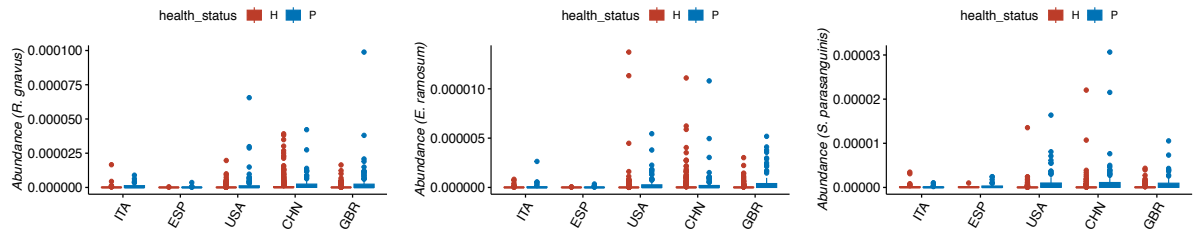


B

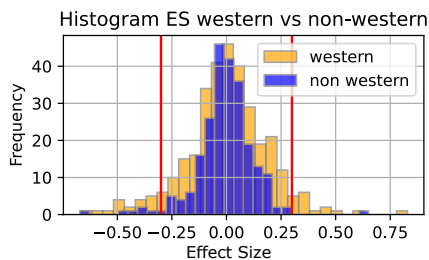
### Colorectal cancer (CRC) cohorts



### Liver disease cohorts



C

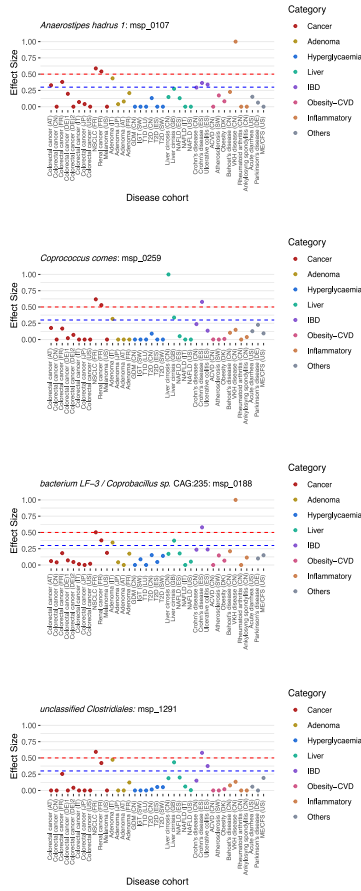


**Supplemental Figure S2.** **A**, Volcano plots display Effect Size vs  $P$ -values across disease cohorts included in this study, each dot in the plot represents an MSP within a cohort. Red horizontal dotted line indicates  $P$ -value equal to 0.05, red vertical lines indicate Effect Size value equal to 0.3 **B**, Boxplots of selected species in colorectal cancer and liver disease cohorts to highlight the underlying data used for differential abundance analysis. **C**, Histogram with the Effect Size values across disease cohorts of the species enriched Western and Non-Western healthy groups. Red vertical lines indicate Effect Size absolute values of 0.3.



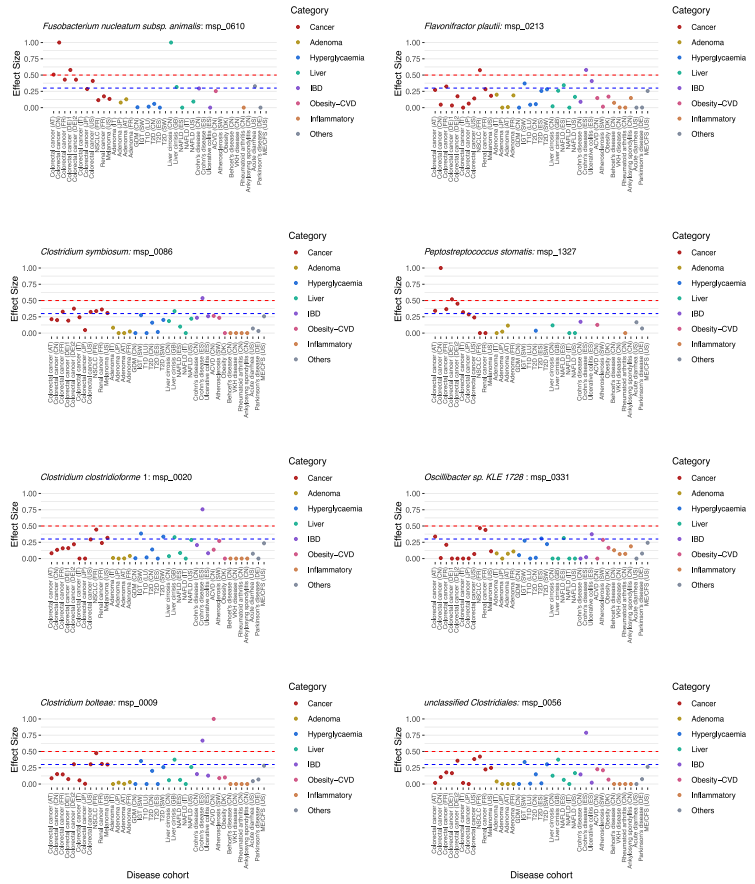
A

## Depleted in Disease

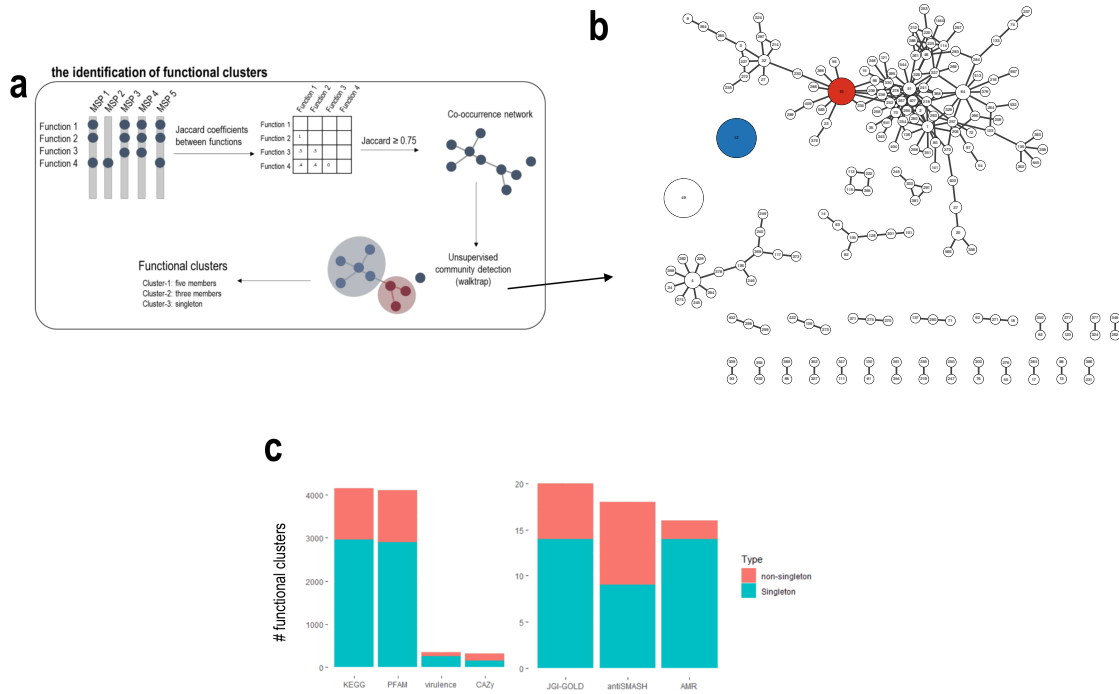


B

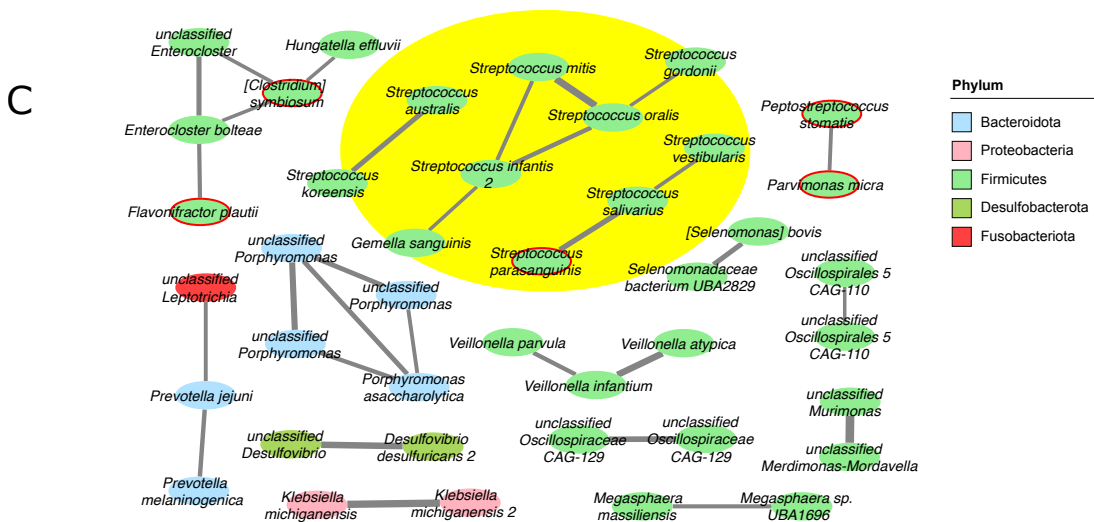
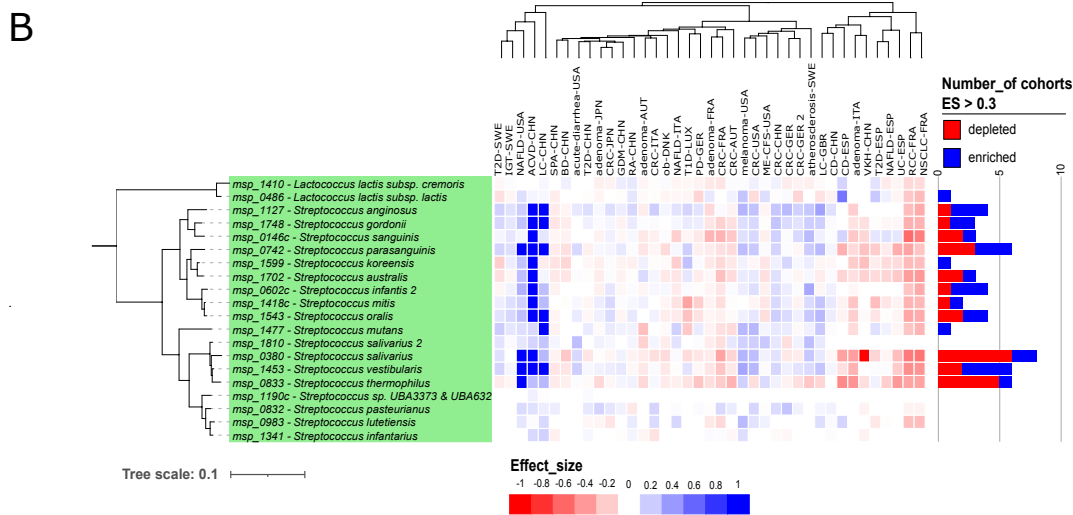
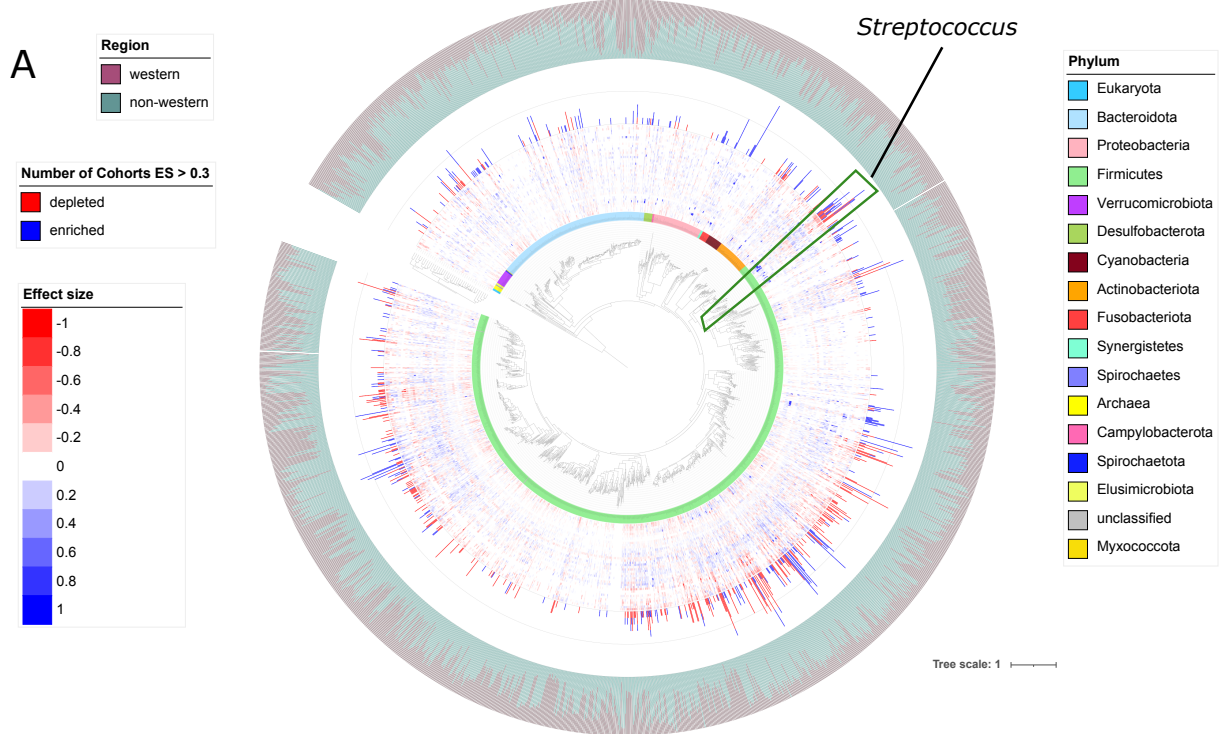
## Enriched in Disease



**Supplemental Figure S3.** Species found either **A** depleted or **B** enriched (Effect Size >0.3) in at least 6 different disease cohorts. The blue dotted line indicates Effect Size equals 0.3, the red dotted line indicates Effect Size equals 0.5, and each dot in the plot represents the Effect Size of the species within a disease cohort. Stars indicate cohorts where the Effect Size was above 0.3.



**Supplemental Figure S4.** 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 cooccurrences 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).



**Supplemental Figure S5.** *Streptococcus* genus in the MSP phylogenetic tree. **A**, The Inner annotation indicates the species' phylum, and the heatmap displays the Effect Size for each MSP 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. Highlighted in green are the MSPs from the *Streptococcus* genus. The external barplot rings show the proportion of MSP abundance between western and non-western samples. The entire phylogenetic tree can be viewed interactively <https://itol.embl.de/tree/130237251127435861638193829>. **B**, Clustermap of 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 dendrogram on the top shows the hierarchical clustering result using the Ward method on the complete Effect Size matrix. **C**, 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, and edge width is proportional to the proportionality value. The yellow oval in the background highlights the Clusters containing MSPs from the *Streptococcus* genus.



## 2 Supplemental Table Legends

**Supplemental Table S1.** Description of the Human Gut Microbiome Atlas (HGMA) datasets used in this study. We show the overall statistics of HGMA datasets, including the total number of samples, BioProject IDs, sequencing platform, geography, and reference.

**Supplemental Table S2.** List of region-enriched MSPs given as  $Z$ -score  $>2$  and P-value  $<0.05$ , we identified 743 regionally enriched MSPs among healthy samples from 17 countries.

**Supplemental Table S3.** Comparison of species relative abundance between western and non-western groups.

**Supplemental Table S4.** Substrate information of CAZyme families commonly existed in human gut microbiome (Figure 1F).

**Supplemental Table S5.** Statistics of pan-metagenomics association study (pan-MGAS) results for 23 diseases. We performed pan-MGAS analysis and identified the enriched or depleted species in each cohort with effect sizes and P-values for each column. The negative sign in the effect size values indicated that the species was depleted in disease, positive effect size values, and species richness.

**Supplemental Table S6.** Statistics of common MSP enriched/depleted diseases. The table displays the MSPs with an effect size above 0.3 in the same direction in at least three different cohorts, along with the corresponding country, disease, and disease category.

**Supplemental Table S7.** Characteristics of 7,763 functional clusters. For a more detailed understanding, per functional cluster, we presented the size of the cluster, number of enriched species, enriched MSP identifiers, enriched species names, enriched KEGG modules (hypergeometric tests  $p < 0.01$ ), subsystems of enriched KEGG modules, KEGG orthologs terms, PFAM terms, virulence term, CAZyme terms, antimicrobial resistance (Mustard) terms, secondary metabolism (antiSMASH) terms, phenotype (JGI-GOLD) terms, product names of virulence terms, and all functional terms of a given cluster. **Supplemental Table S8.** Mean directional SHAP scores for each cohort with matched controls.

**Supplemental Table S9.** Statistical results of PERMANOVA of different batches of healthy samples of same country origins

**Supplemental Table S10.** Statistical results of PERMANOVA of confounder effects on healthy samples in different geographical datasets

**Supplemental Table S11.** Meta-analysis differential abundance testing between Healthy and Disease states controlling for age and gender using MMUPHin R package.

**Supplemental Table S12.** Summary of DNA extraction protocols of cohorts used in this study.

## 3 Supplemental Methods

### 3.1 Metagenomics species pan-genome (MSP) creation

To build MSP gene clusters, MSPminer relies on the hypothesis initially introduced by Nielsen et al. that genes belonging to the same species should be co-abundant across multiple metagenomic samples. Although this assumption is mostly true for core genes, the co-abundance hypothesis has not been verified for genes shared by multiple species (horizontal transfers) or for genes subject to copy number variation. MSPminer introduces a new methodology that considers this limitation, capturing and classifying genes as part of the core genes (common to all strains of a given species) and accessory genes (specific to some strains only), with a higher clustering confidence for genes that are detected in multiple samples (*i.e.*, more than three samples).