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

Sunjae Lee[1*], Theo Portlock[2*], Emmanuelle Le Chatelier[3*], Jose Garcia[2*], Nicolas Pons[4], Florian Plaza Onate[4], Neelu Begum[1], Ceri Proffitt[1], Dorines Rosario[1], Stefania Vaga[1], Junseok Park[5], Kalle von Feilitzen[2], Fredric Johansson[2], Azadeh Harzandi[1], Cheng Zhang, Lindsey A. Edwards[7], Vincent Lombard[8,9], Franck Gauthier[4], Claire J. Steves[10], David Gomez-Cabrero[1,11], Bernard Henrissat[8,9,12], Doheon Lee[5], Debbie L. Shawcross[7], Gordon Proctor[1], Jens Nielsen[14,15,16], David Moyes, Adil Mardinoglu , Stanislav Dusko Ehrlich[4], Mathias Uhlen[2,16], Saeed Shoaie

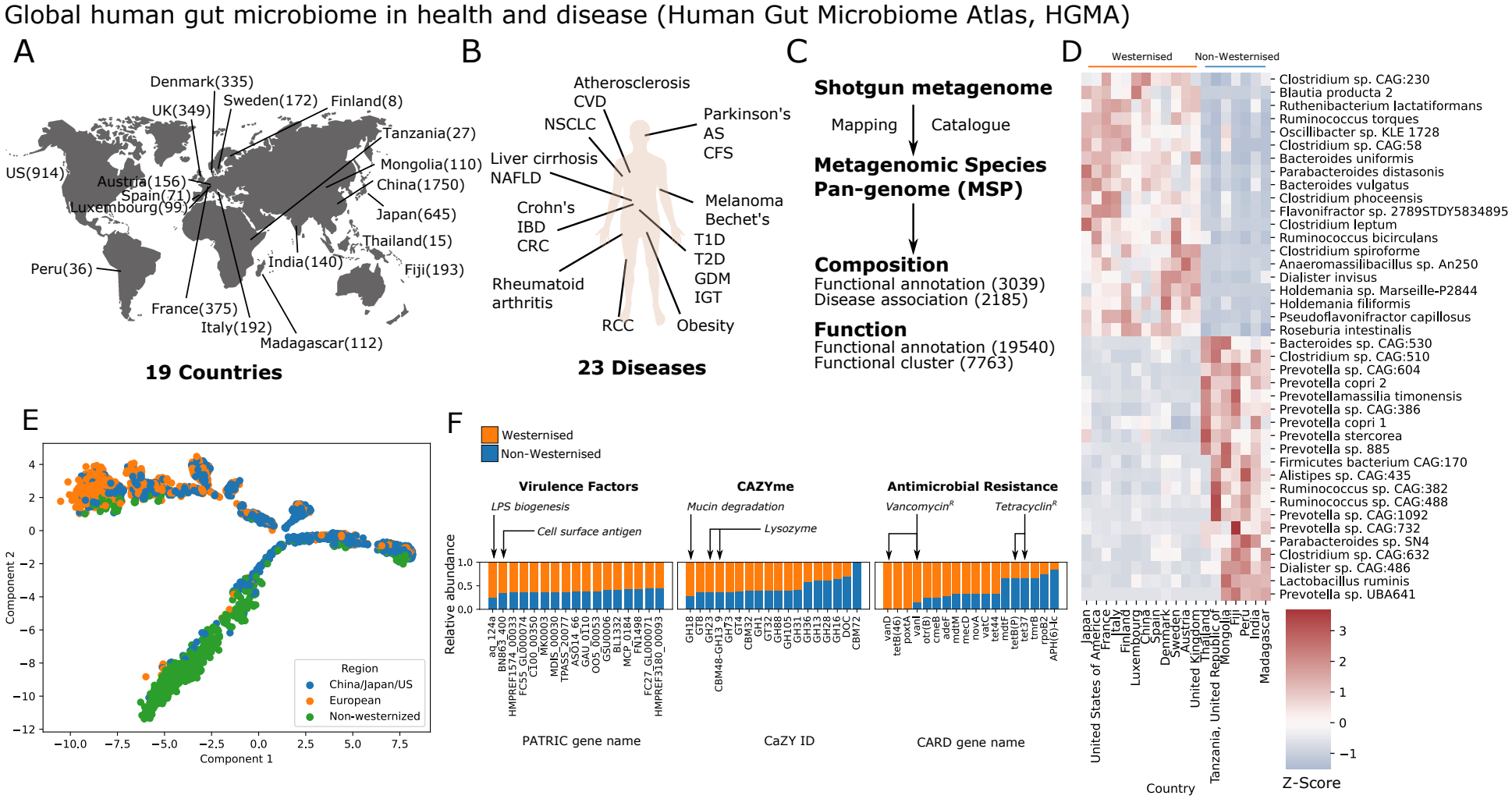


Fig. 1. Characterization of the global gut microbiome in health and disease. Pan-metagenomics studies of health and disease. Corresponding datasets were publicly shared as a resource: human gut microbiome atlas (HGMA). **A,** geographical distribution of the datasets used in this study (the number of 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).

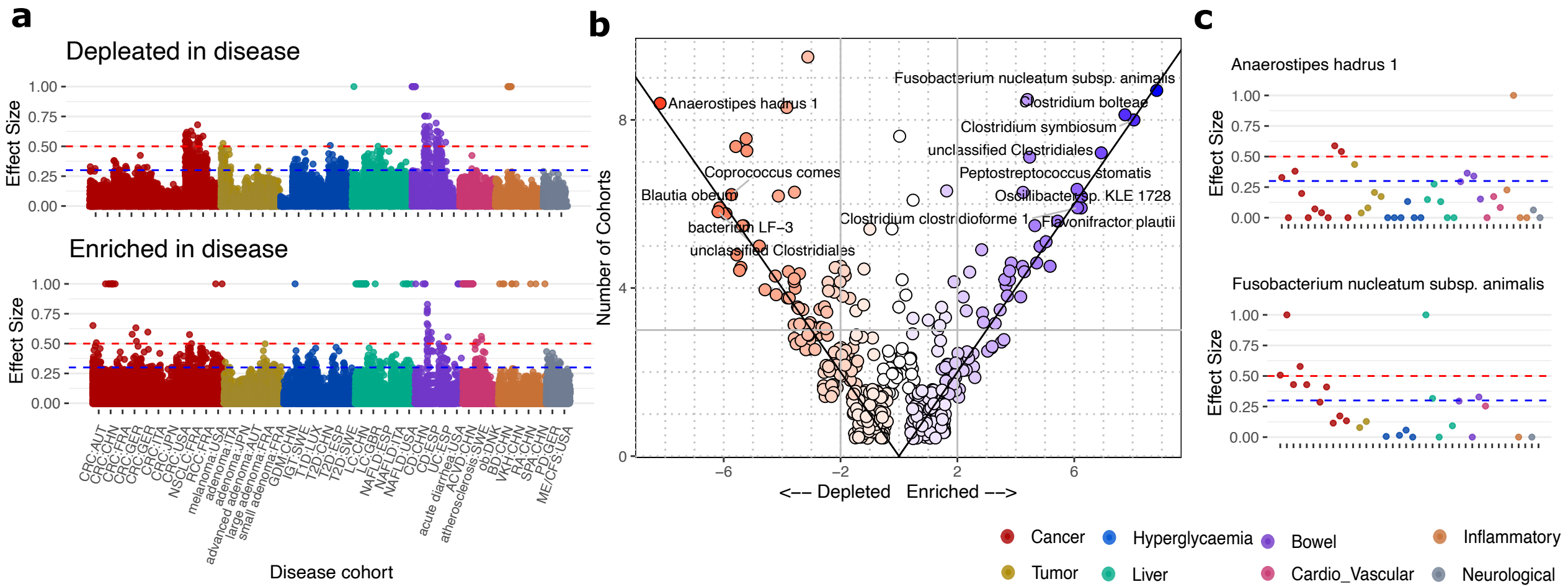


Fig. 2. Pan-metagenomics association studies (Pan-MGAS) of 43 cohorts from 23 different diseases and 14 countries (n=2,185). A, We identified significantly enriched/depleted species of cohorts based on effect sizes (ESs) of Wilcoxon one-sided tests (ES ≥ 0.3). Acronyms are: ACVD, Acute Coronary Cardiovascular 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.

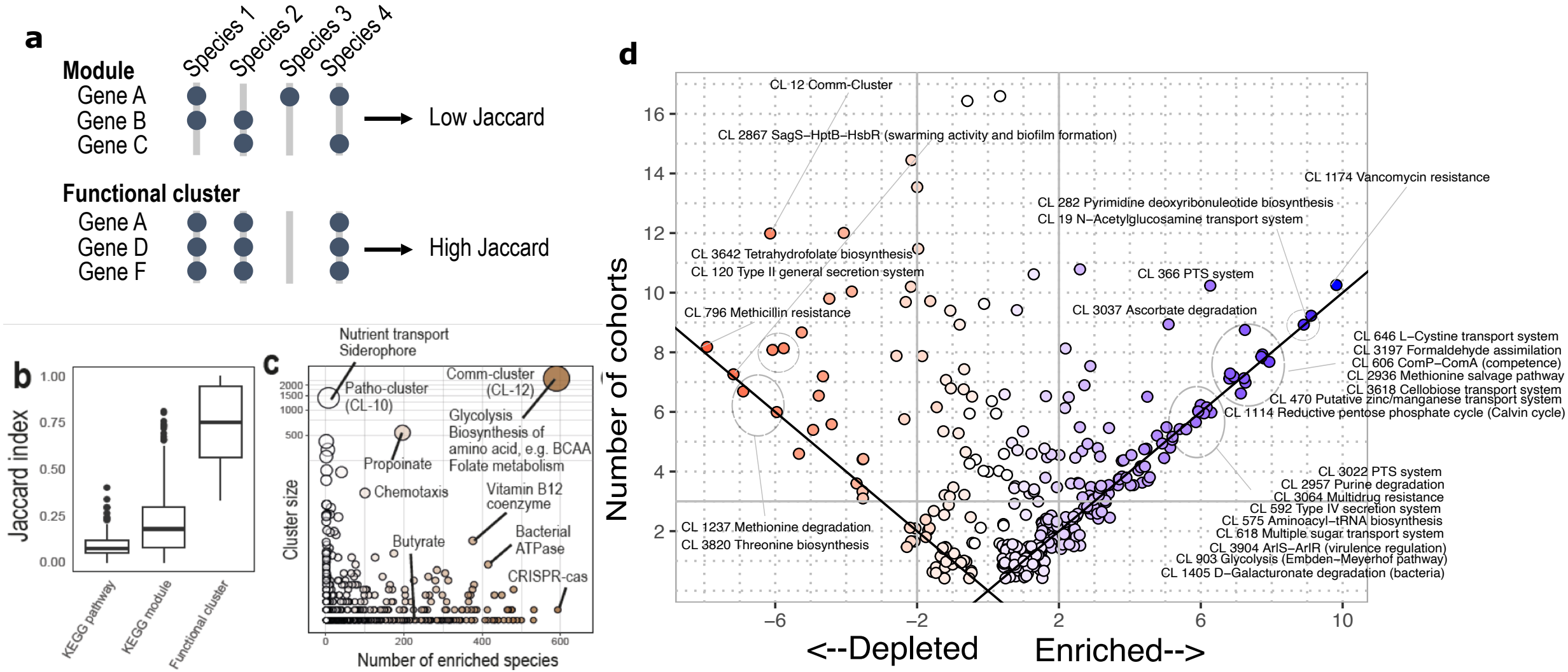


Fig. 3. Analysis of functional clusters. For the functional characterization of human gut MSPs, we annotated respective genes with 19,540 features of microbial function/phenotype databases and identified 7,763 functional clusters better representing the microbiome. **A,** Identification of functional 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 $p < 0.001$) 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.

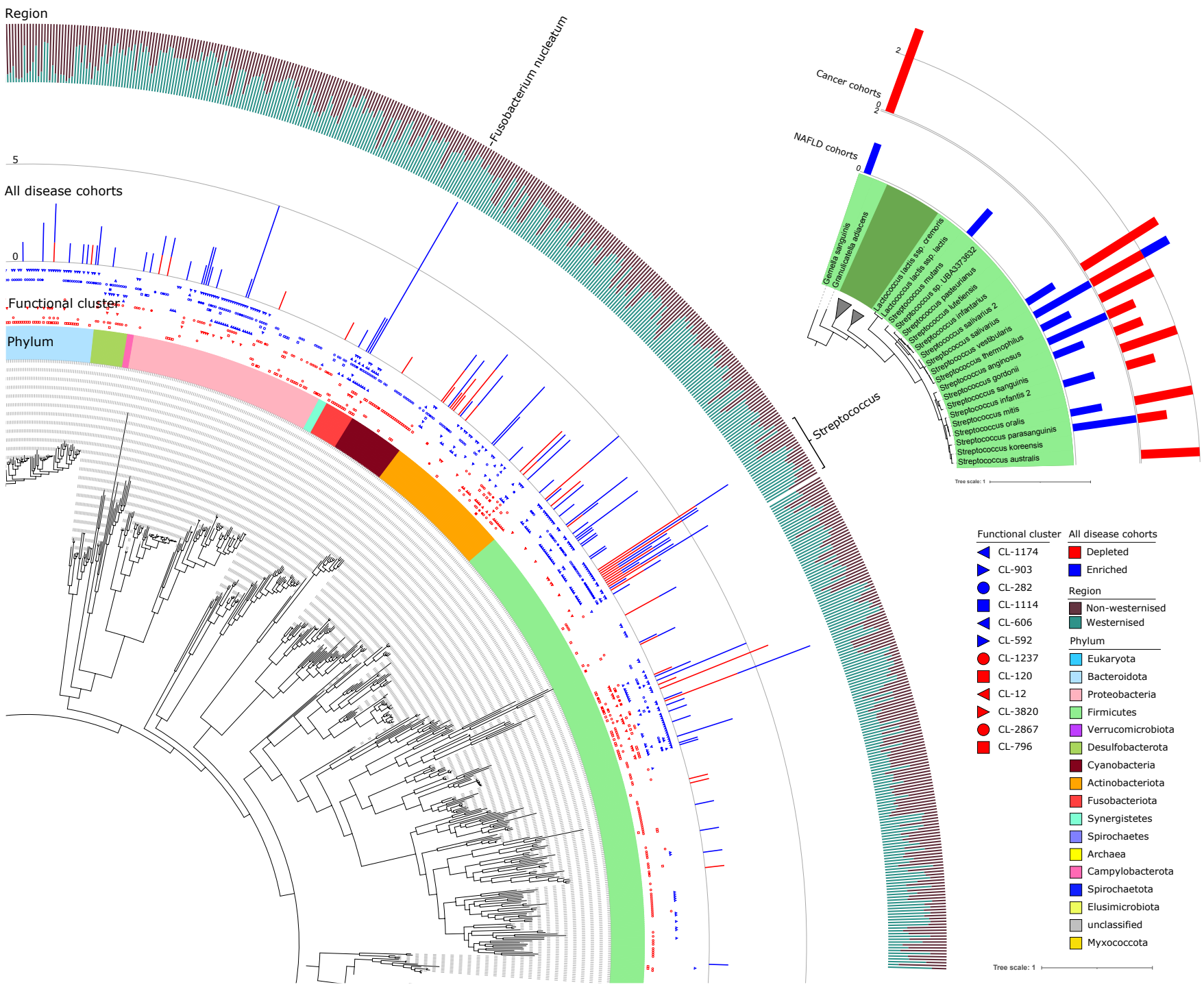
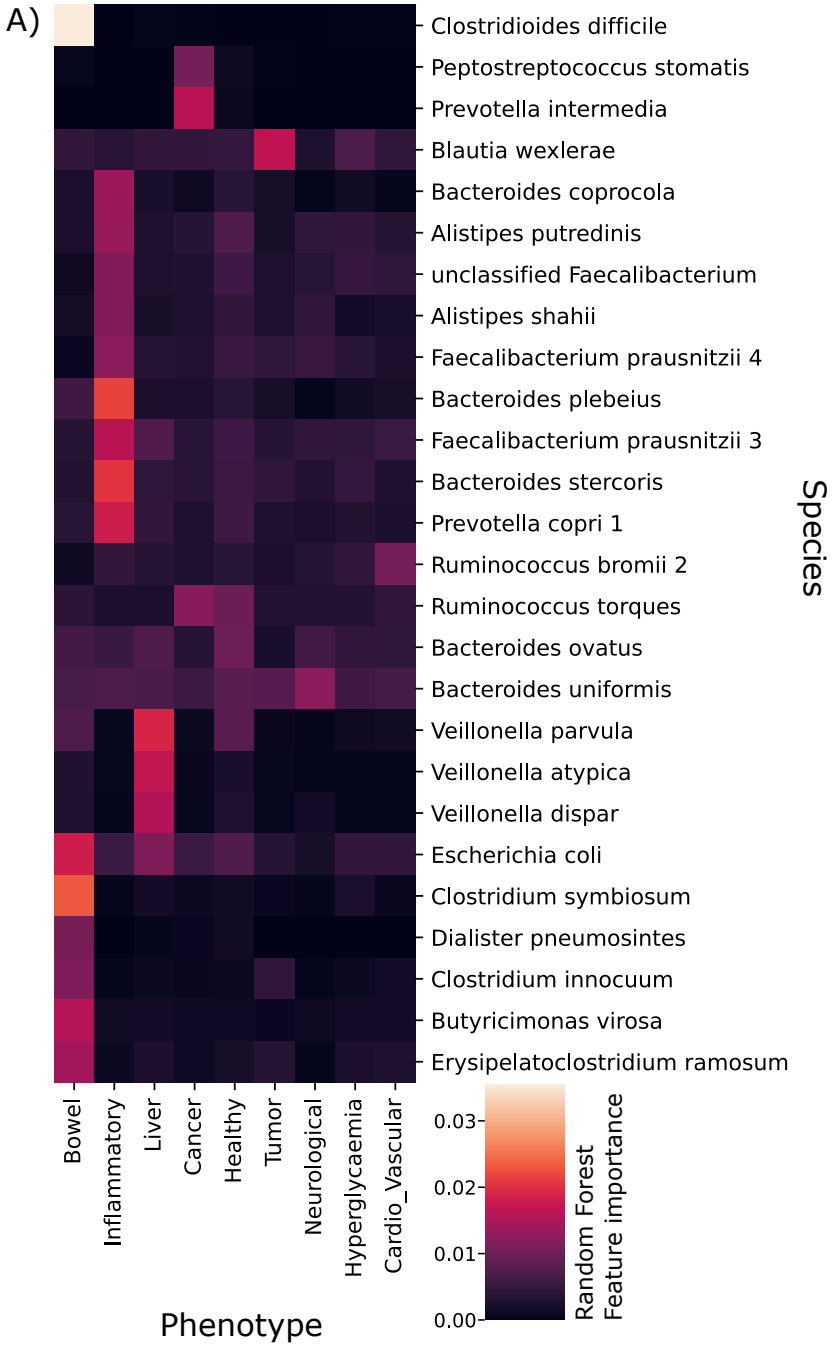


Fig. 4. Phylogenetic differences between species function, disease enrichment, and region enrichment. Inner annotation of dendrogram is species phylum, second is enrichment of functional cluster, third is the total number of disease cohorts that the species is enriched/depleted in, and the outer annotation 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.



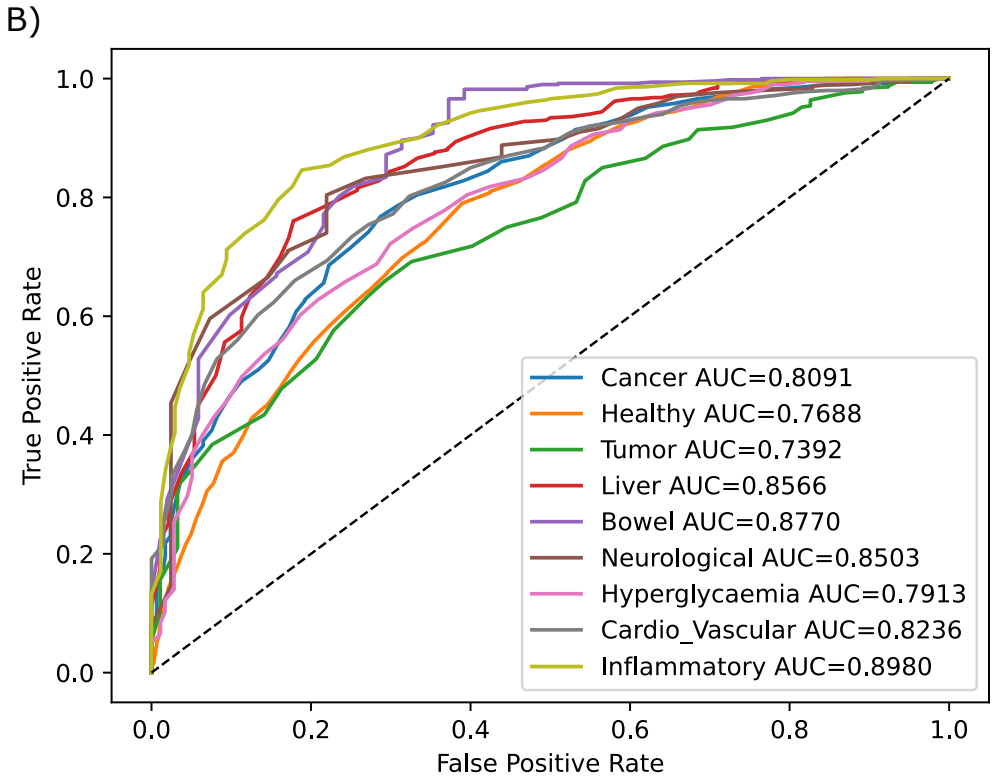
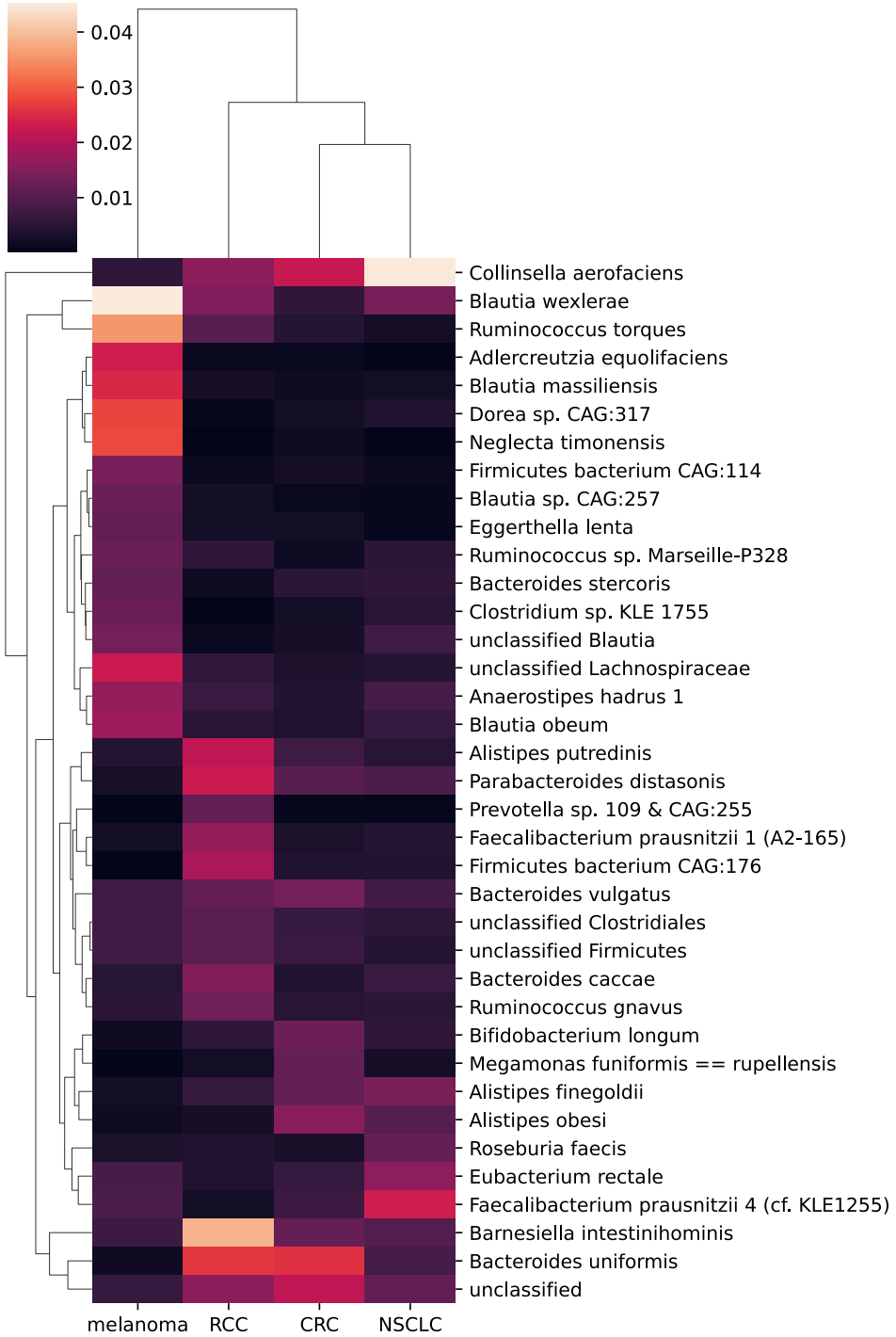
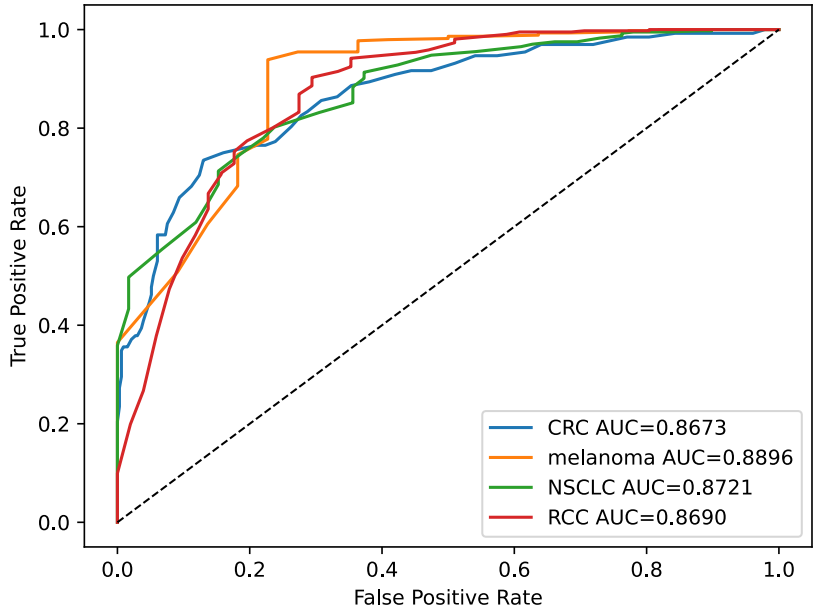


Figure 5: Random forest prediction of disease from metagenomic species. A) Heatmap of Random forest feature importance for the prediction of each disease. B) ROC curves for each prediction


0.04
0.03
0.02
0.01
Collinsella aerofaciens
Blautia wexlerae
Ruminococcus torques
Adlercreutzia equolifaciens
Blautia massiliensis
Dorea sp. CAG:317
Neglecta timonensis
Firmicutes bacterium CAG:114
Blautia sp. CAG:257
Eggerthella lenta
Ruminococcus sp. Marseille-P328
Bacteroides stercoris
Clostridium sp. KLE 1755
unclassified Blautia
unclassified Lachnospiraceae
Anaerostipes hadrus 1
Blautia obeum
Alistipes putredinis
Parabacteroides distasonis
Prevotella sp. 109 & CAG:255
Faecalibacterium prausnitzii 1 (A2-165)
Firmicutes bacterium CAG:176
Bacteroides vulgatus
unclassified Clostridiales
unclassified Firmicutes
Bacteroides caccae
Ruminococcus gnavus
Bifidobacterium longum
Megamonas funiformis == rupellensis
Alistipes finegoldii
Alistipes obesi
Roseburia faecis
Eubacterium rectale
Faecalibacterium prausnitzii 4 (cf. KLE1255)
Barnesiella intestinihominis
Bacteroides uniformis
unclassified
melanoma
RCC
CRC
NSCLC


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CRC AUC=0.8673
melanoma AUC=0.8896
NSCLC AUC=0.8721
RCC AUC=0.8690