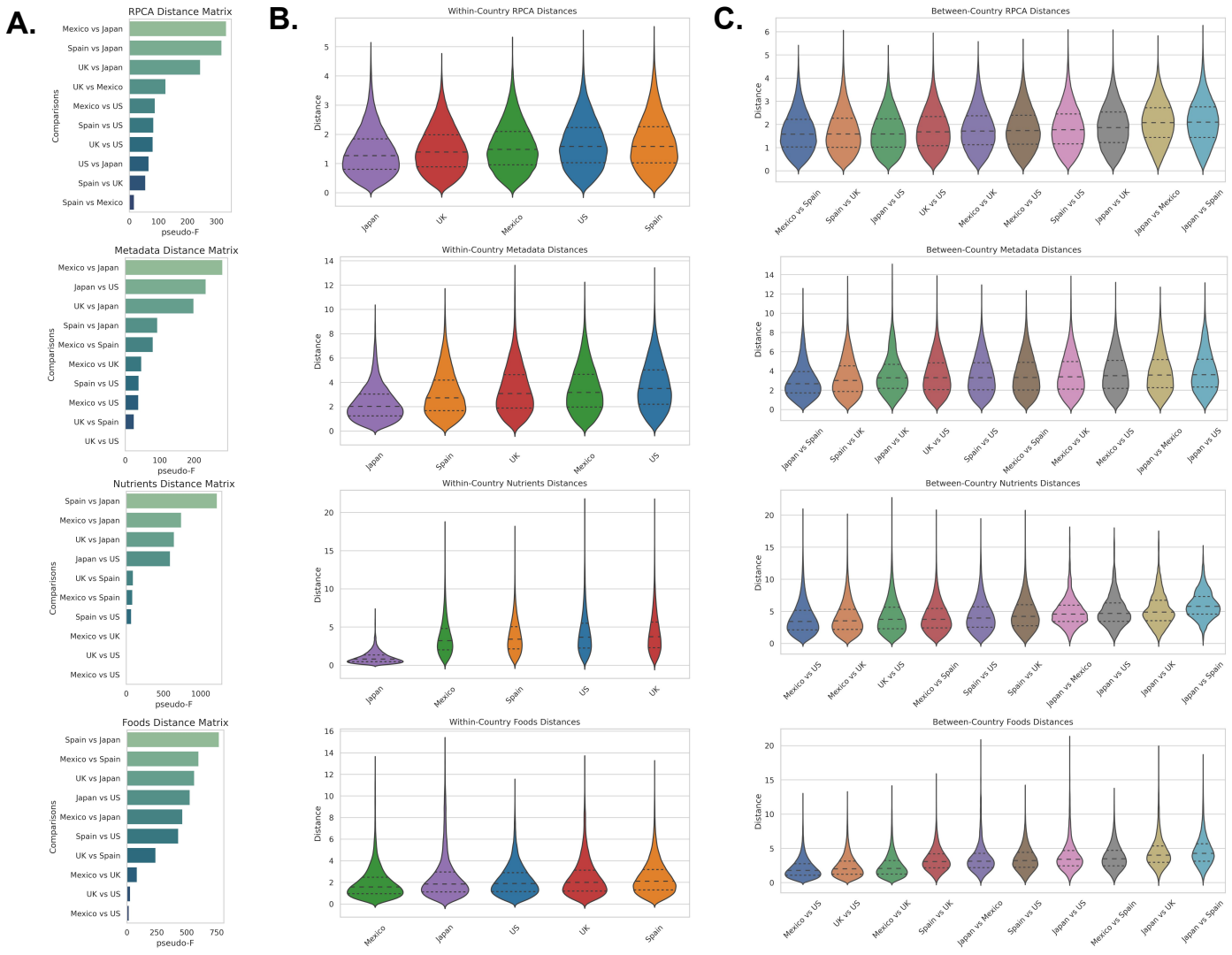

Fig S1. Cross-country differences in microbiome composition, metadata, nutrients, and foods.
(A) Pairwise PERMANOVA pseudo-F statistics comparing country-level dissimilarities across four data types: microbiome, metadata, nutrients, and foods. Higher pseudo-F values reflect greater between-country separation. (B) Within-country distributions of pairwise distances for microbiome (top), metadata (middle), nutrients (third), and foods (bottom). Violin plots illustrate the distribution density of within-country variation, with dashed lines indicating quartiles. (C) Between-country pairwise distances for microbiome (top), metadata (middle), nutrients (third), and foods (bottom). Each violin represents distances between two countries (e.g., Mexico–Spain, Spain–Japan).

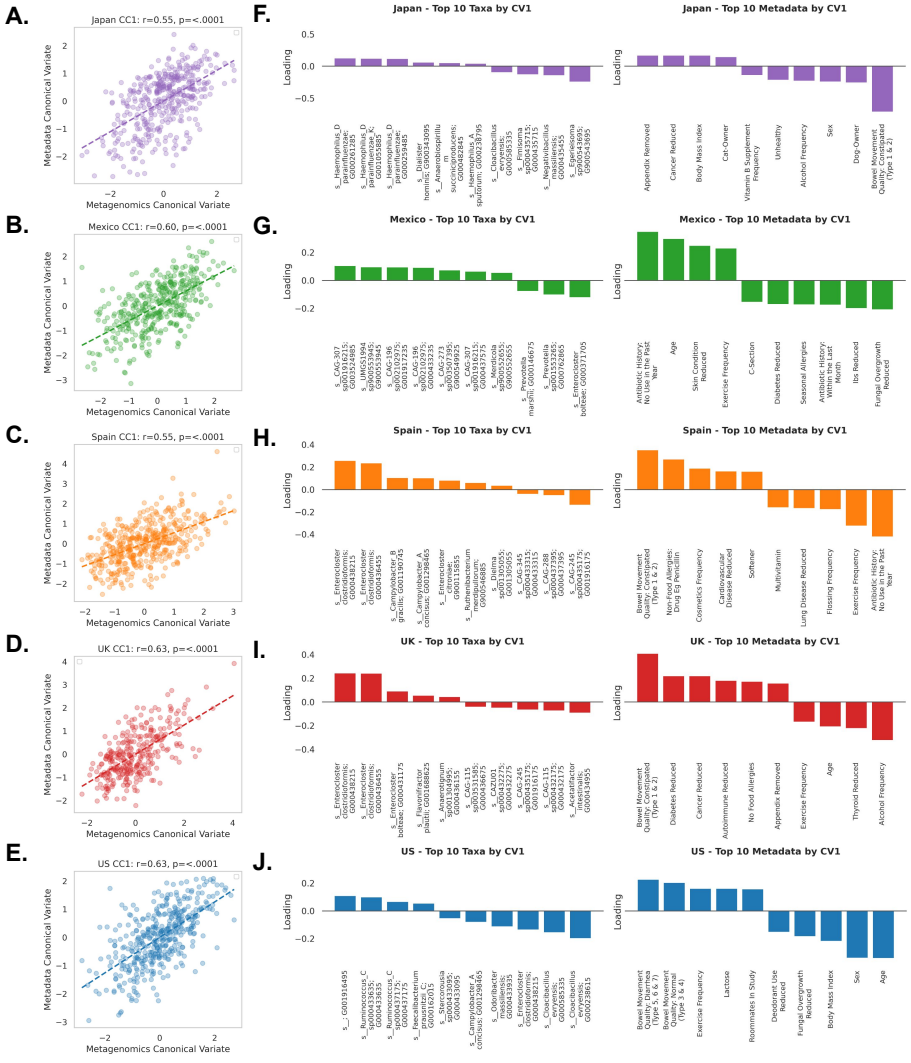

Figure S2. Sparse canonical correlation analysis (sCCA) of metagenomics with metadata within countries. (A–E) Scatterplots of canonical variates (CVs) from sCCA showing correlations between metagenomics and metadata within (A) Japan, (B) Mexico, (C) Spain, (D) UK, and (E) US. Each point represents an individual with best-fit regression lines for each cohort. Pearson correlation coefficients (r) and p-values are shown for each country. (F-J) Barplots of the top 10 features with the largest absolute loadings on CV1 for microbial taxa (left) and metadata (right) within (F) Japan, (G) Mexico, (H) Spain, (I) UK, (J) US. Positive and negative bars indicate direction of association.

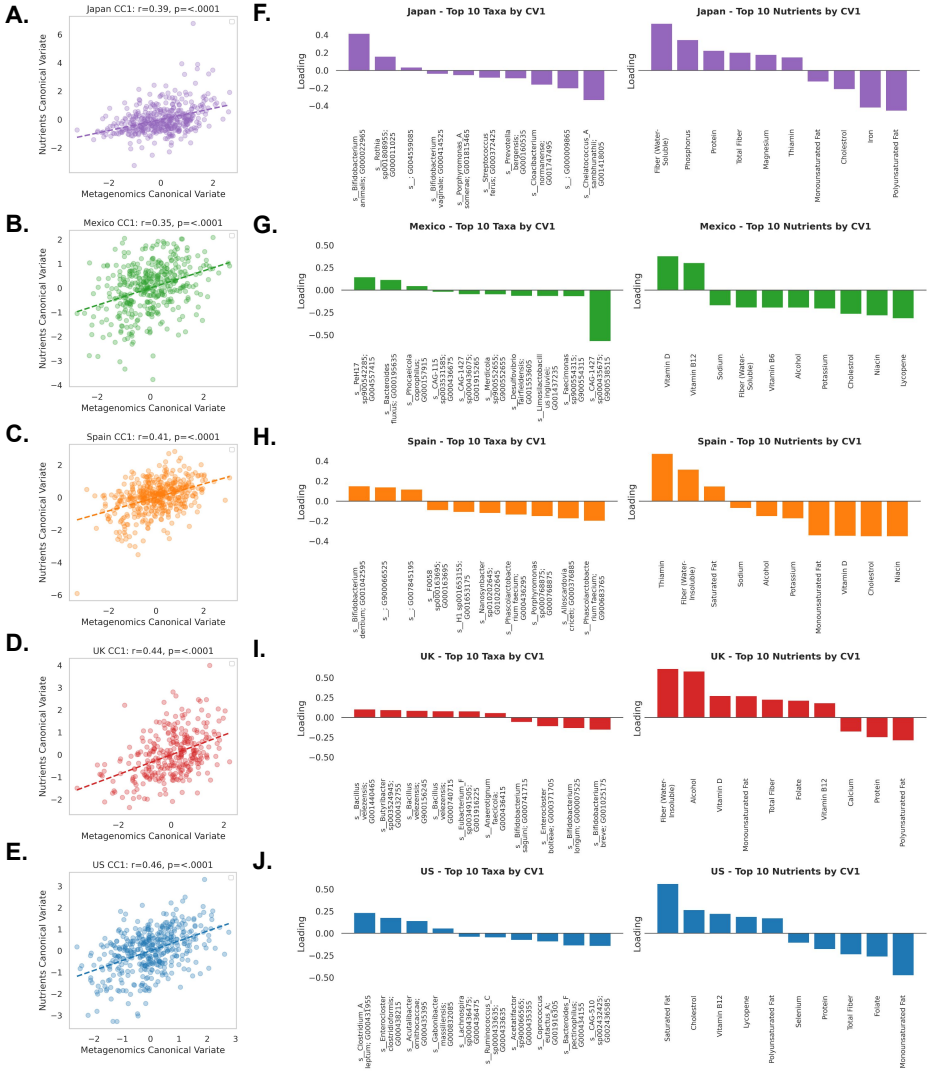

Figure S3. Sparse canonical correlation analysis (sCCA) of metagenomics with nutrients within countries. (A–E) Scatterplots of canonical variates (CVs) from sCCA showing correlations between metagenomics and nutrients within (A) Japan, (B) Mexico, (C) Spain, (D) UK, and (E) US. Each point represents an individual with best-fit regression lines for each cohort. Pearson correlation coefficients (r) and p-values are shown for each country. (F-J) Barplots of the top 10 features with the largest absolute loadings on CV1 for microbial taxa (left) and nutrients (right) within (F) Japan, (G) Mexico, (H) Spain, (I) UK, (J) US. Positive and negative bars indicate direction of association.

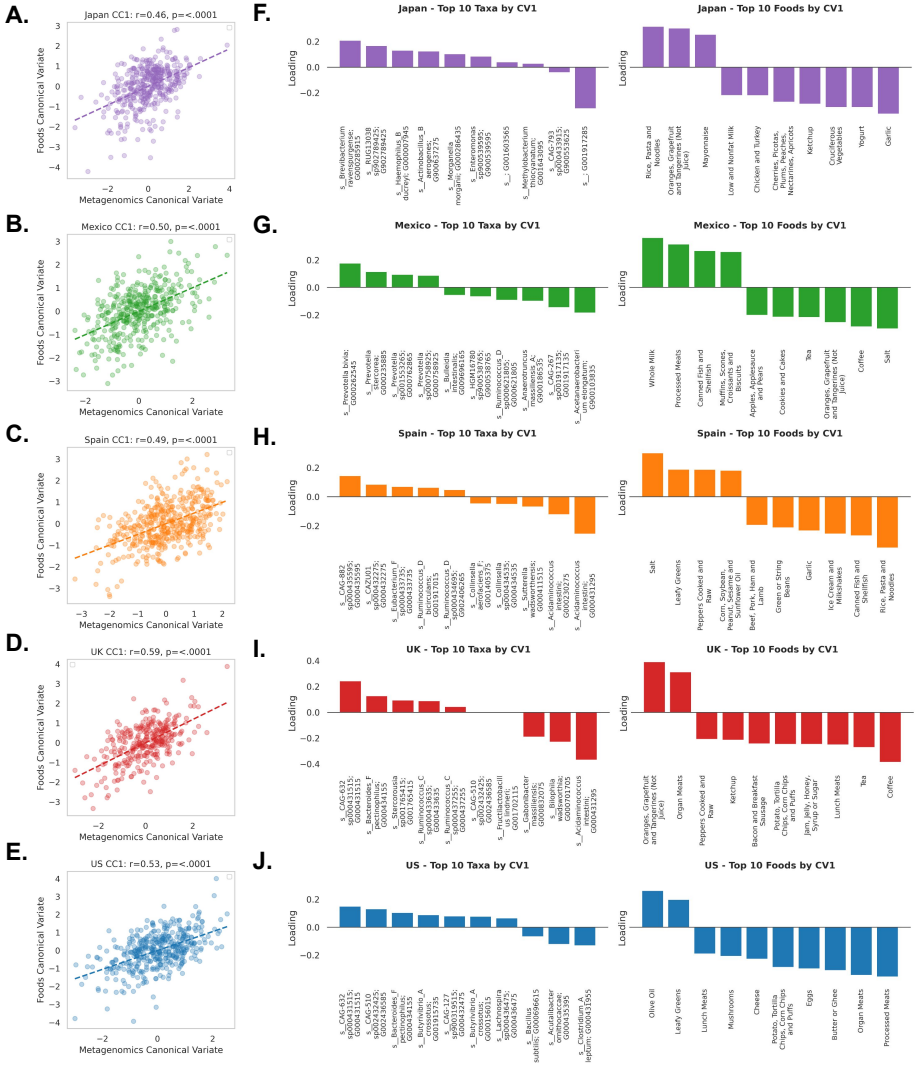

Figure S4. Sparse canonical correlation analysis (sCCA) of metagenomics with foods within countries. (A–E) Scatterplots of canonical variates (CVs) from sCCA showing correlations between metagenomics and foods within (A) Japan, (B) Mexico, (C) Spain, (D) UK, and (E) US. Each point represents an individual with best-fit regression lines for each cohort. Pearson correlation coefficients (r) and p-values are shown for each country. (F-J) Barplots of the top 10 features with the largest absolute loadings on CV1 for microbial taxa (left) and foods (right) within (F) Japan, (G) Mexico, (H) Spain, (I) UK, (J) US. Positive and negative bars indicate direction of association. (K) Scatterplot showing the relationship between the log-transformed abundance ratio of *Bacteroides finegoldii* (G000434155) to high prevalence taxa *Intestinibaculum porci* (G003925875) and pectin intake (g per kcal) in the US (blue), UK (red), and Mexico (green). Points represent individual participants, with country-specific best-fit regression lines and 95% confidence intervals overlaid.

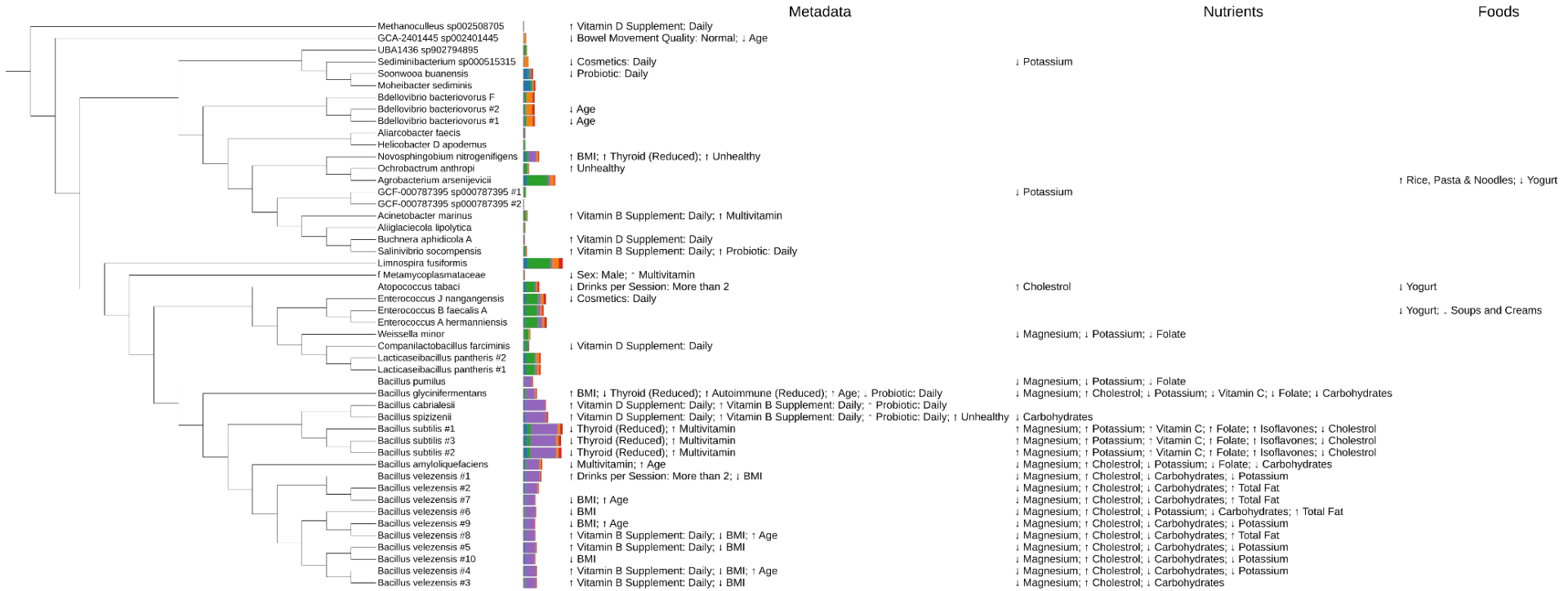

Figure S5. Phylogenetic tree of enriched microbial taxa. The phylogenetic tree depicts country-enriched microbial taxa presenting significant associations with metadata, nutrients or foods, identified using BIRDMAn. Bar plots at the tree tips show the prevalence of each taxon by country (purple = Japan, green = Mexico, orange = Spain, red = UK, blue = US). Text labels indicate associations with metadata (left), nutrients (middle), and foods (right). Upward arrows denote positive associations, while downward arrows denote negative associations. All associations were adjusted for thdmi_cohort. Branch lengths were disabled during visualization to emphasize taxon relationships rather than evolutionary distances.

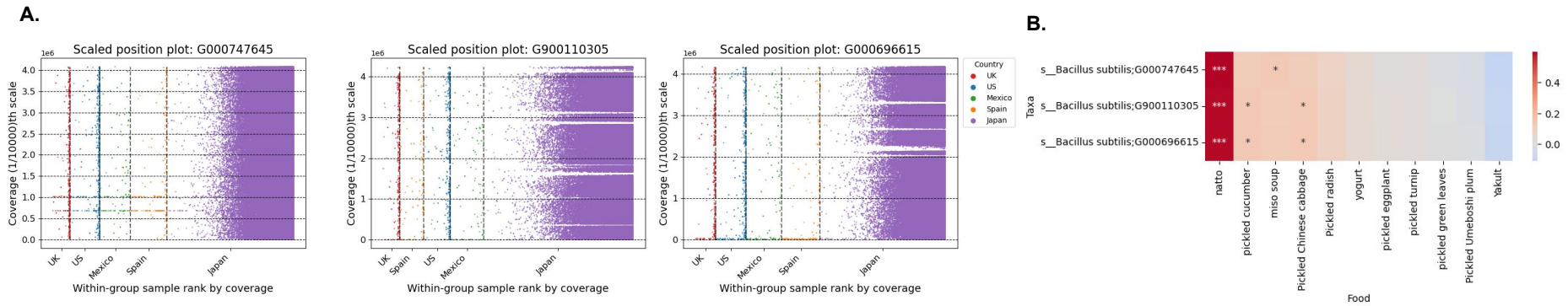

Figure S6. Coverage of *B. subtilis* OGUs and associations with fermented foods. (A) Scaled position plots (1/10,000th genome scale) showing per-sample genomic coverage for the OGUs G000747645, G900110305, and G000696615 stratified by country. Each dot represents a covered genomic bin for a sample, plotted along the x-axis by within-group sample rank (low to high coverage). Vertical dashed lines demarcate cohort boundaries. The y-axis represents genomic position bins with detected coverage. (B) Heatmap showing Spearman correlations between the CLR-abundance of three *B. subtilis* OGUs (rows) and consumption of fermented and pickled foods (columns). Warmer colors indicate positive correlations, and cooler colors indicate negative correlations. Statistical significance is denoted by asterisks based on FDR-adjusted p-values (*p < 0.05, **p < 0.01, ***p < 0.001).

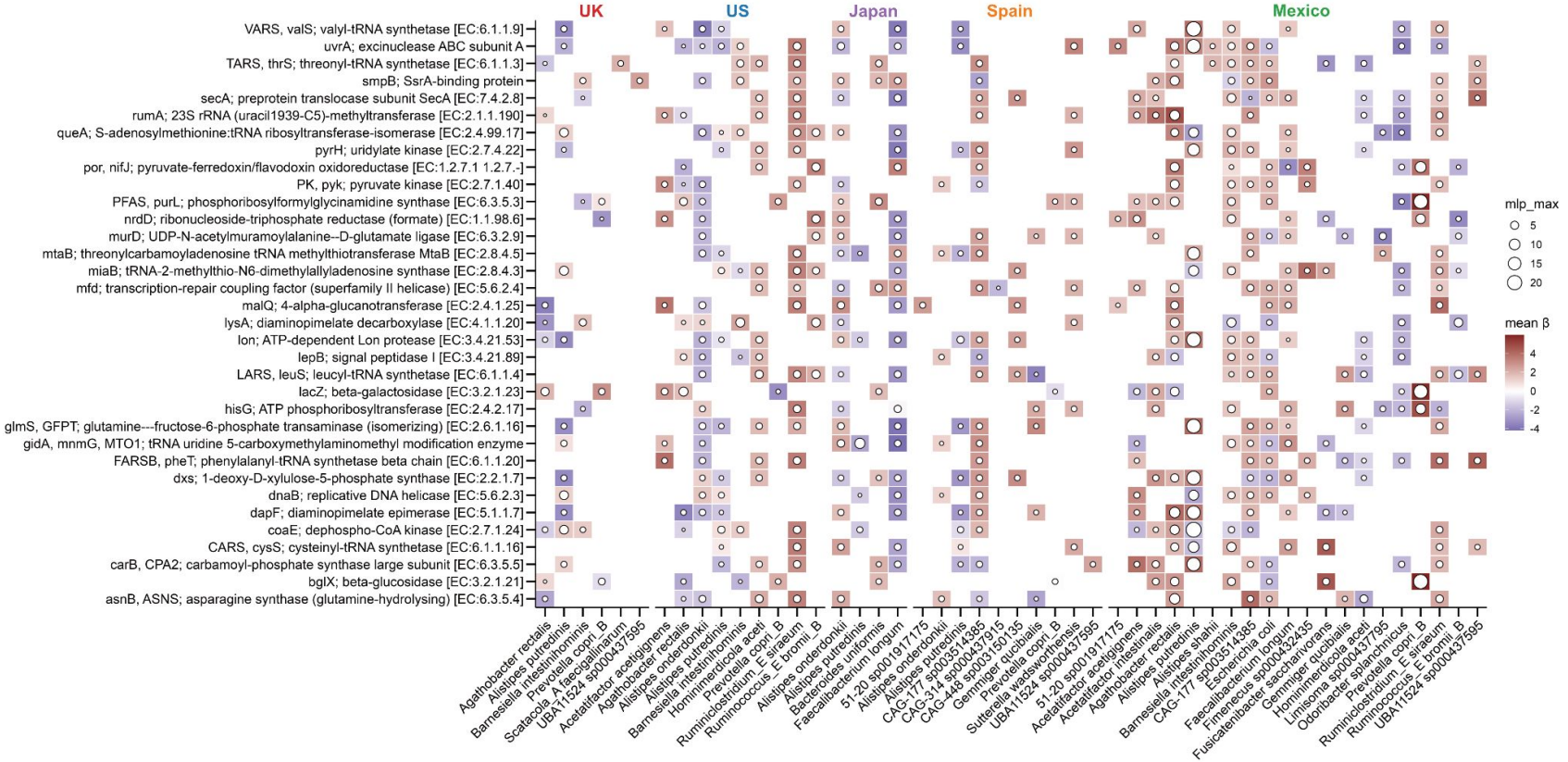

Figure S7. Cross-population enrichment of functional genes across bacterial species.
Heatmap of species by country comparison showing the mean effect size and magnitude of the association between a given bacterial KEGG group and origin country. Tile color encodes the average effect size (mean β) from a genome-wide association study, where positive values (red) indicate enrichment in that country and negative values (blue) indicate depletion relative to the remaining cohorts. Circle overlays mark the statistical signal strength: circle size corresponds to the maximum $-\log_{10}$ p value (mlp_max) observed for that KEGG group in the respective species. Only gene–species pairs with ≥ 1 significant variant (FDR < 0.05) in at least one population are shown.