AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Meta-Analysis of Altered Gut Microbiota Reveals Microbial and Metabolic Biomarkers for Colorectal Cancer

Nagavardhini Avuthu,[a] **Chittibabu Guda**[a,b]

[a]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, Nebraska, USA
[b]Center for BioInformatics Research and Innovation (CBIRI), University of Nebraska Medical Center, Omaha, Nebraska, USA

ABSTRACT Colorectal cancer (CRC) is the second leading cause of cancer mortality worldwide. The dysbiotic gut microbiota and its metabolite secretions play a significant role in CRC development and progression. In this study, we identified microbial and metabolic biomarkers applicable to CRC using a meta-analysis of metagenomic datasets from diverse geographical regions. We used LEfSe, random forest (RF), and co-occurrence network methods to identify microbial biomarkers. Geographic dataset-specific markers were identified and evaluated using area under the ROC curve (AUC) scores and random effect size. Co-occurrence networks analysis showed a reduction in the overall microbial associations and the presence of oral pathogenic microbial clusters in CRC networks. Analysis of predicted metabolites from CRC datasets showed the enrichment of amino acids, cadaverine, and creatine in CRC, which were positively correlated with CRC-associated microbes (*Peptostreptococcus stomatis*, *Gemella morbillorum*, *Bacteroides fragilis*, *Parvimonas* spp., *Fusobacterium nucleatum*, *Solobacterium moorei*, and *Clostridium symbiosum*), and negatively correlated with control-associated microbes. Conversely, butyrate, nicotinamide, choline, tryptophan, and 2-hydroxybutanoic acid showed positive correlations with control-associated microbes ($P < 0.05$). Overall, our study identified a set of global CRC biomarkers that are reproducible across geographic regions. We also reported significant differential metabolites and microbe-metabolite interactions associated with CRC. This study provided significant insights for further investigations leading to the development of noninvasive CRC diagnostic tools and therapeutic interventions.

IMPORTANCE Several studies showed associations between gut dysbiosis and CRC. Yet, the results are not conclusive due to cohort-specific associations that are influenced by genomic, dietary, and environmental stimuli and associated reproducibility issues with various analysis approaches. Emerging evidence suggests the role of microbial metabolites in modulating host inflammation and DNA damage in CRC. However, the experimental validations have been hindered by cost, resources, and cumbersome technical expertise required for metabolomic investigations. In this study, we performed a meta-analysis of CRC microbiota data from diverse geographical regions using multiple methods to achieve reproducible results. We used a computational approach to predict the metabolomic profiles using existing CRC metagenomic datasets. We identified a reliable set of CRC-specific biomarkers from this analysis, including microbial and metabolite markers. In addition, we revealed significant microbe-metabolite associations through correlation analysis and microbial gene families associated with dysregulated metabolic pathways in CRC, which are essential in understanding the vastly sporadic nature of CRC development and progression.

KEYWORDS biomarkers, colorectal cancer, gut dysbiosis, meta-analysis, microbial metabolites, microbiome

CRC is the third most diagnosed and the second leading cause of cancer-related deaths for men and women combined, globally (1). Genetic and environmental factors influence CRC incidences. Most CRCs are sporadic (70%), while about 25% are

Editor Jan Claesen, Lerner Research Institute

Address correspondence to Chittibabu Guda, babu.guda@unmc.edu.

The authors declare no conflict of interest.

Received 10 January 2022
Accepted 6 June 2022

familial, and about 5% are hereditary (2). Environmental factors consistently associated with sporadic CRC incidences include a low-fiber diet, tobacco, alcohol, lack of physical activity, obesity, age, and diabetes mellitus (3–8) which could modify gut microbial composition and function (9–11). Gut microbiota is a complex and dynamic microbial community with diverse functions that significantly contribute to human health and metabolic and immune functions. Dysbiosis of gut microbiota is associated with numerous gastrointestinal and extraintestinal disorders, including cancers (12–14). It could contribute to the etiology of CRC by altering the inflammatory, genomic, and metabolic processes in the host. To date, only a few pathogenic species such as *F. nucleatum*, *B. fragilis*, and *Escherichia coli* were experimentally shown to be involved in CRC carcinogenesis through inflammatory and genotoxic activities (15–17). The dysbiotic gut microbiota also promotes inflammation and alters the host metabolism through its metabolite secretions (18). For instance, the red meat diet enriches the sulfate-reducing gut bacteria that are involved in the production of hydrogen sulfide, a genotoxic agent (19, 20). Long-term dependence on sulfur microbial diet is associated with increased CRC risk (21).

Several studies have explored the gut microbial composition to identify CRC biomarkers and linked pathogenic bacteria such as *B. fragilis*, *F. nucleatum*, *Streptococcus bovis*, *E. coli*, *Enterococcus faecalis*, and *Porphyromonas* spp. to CRC (7, 22–24). However, consistent CRC biomarkers are lacking as most of the studies are associated with specific cohorts that are highly influenced by dietary factors. For example, increased CRC risk in African Americans was shown to be associated with secondary bile acids production by enriched *Bacteroides* under a high-fat and low-fiber diet (25). Conversely, decreased CRC risk in native Africans was associated with increased short-chain fatty acid (SCFAs) production by enriched *Prevotella* members under high-fiber diet conditions (25). The altered gut microbial metabolite profiles, such as a decrease in SCFAs (acetate, propionate, and butyrate) and an increase in secondary bile acids are shown to promote carcinogenesis through a proinflammatory mechanism (26), and diet-derived microbial metabolites, such as N-nitroso compounds, azo compounds, and nitrates lead to genotoxic and carcinogenic effects in the host (27). Despite the vital role played by the microbe-derived metabolites in CRC, global biomarkers are not available due to cumbersome experimental limitations that involve testing in animal models.

We identified a set of global biomarkers for CRC in this study through a comprehensive meta-analysis of existing CRC datasets from diverse geographical regions and identified a novel CRC biomarker, *A. onderdonkii*. Using a computational approach, we predicted the metabolomic profiles of CRC datasets and identified metabolite biomarkers, which are consistent with previous metabolomic studies, such as the enrichment of butanoic acid in controls but different amino acids in the CRC cohort. And in this study, we also showed the significant microbial and metabolite correlations in CRC pathogenesis. Next, we identified potential functional capabilities in CRC pathogens and their differences across the geographical regions based on gene family analysis. This study enhances our understanding of the role of CRC-associated microbes and their metabolites in CRC development and progression.

RESULTS

Composition and diversity of gut microbial communities associated with CRC. The taxonomic composition of CRC gut communities was analyzed using publicly available shotgun metagenomic sequencing data from three different geographical regions, the USA, China, and France (Table S1). High-quality sequencing reads were selected after preprocessing and quality control and then quantified for taxonomic composition using *MetaPhlAn* software (Data Set S1). We identified archaea, bacteria, eukaryotes, and viruses in most of the samples. However, bacterial taxa dominated with more than 97% of all species identified in each sample. We selected a total of 418, 410, and 469 microbial species with average relative abundance $>$0.1% from the samples of the USA, China, and France

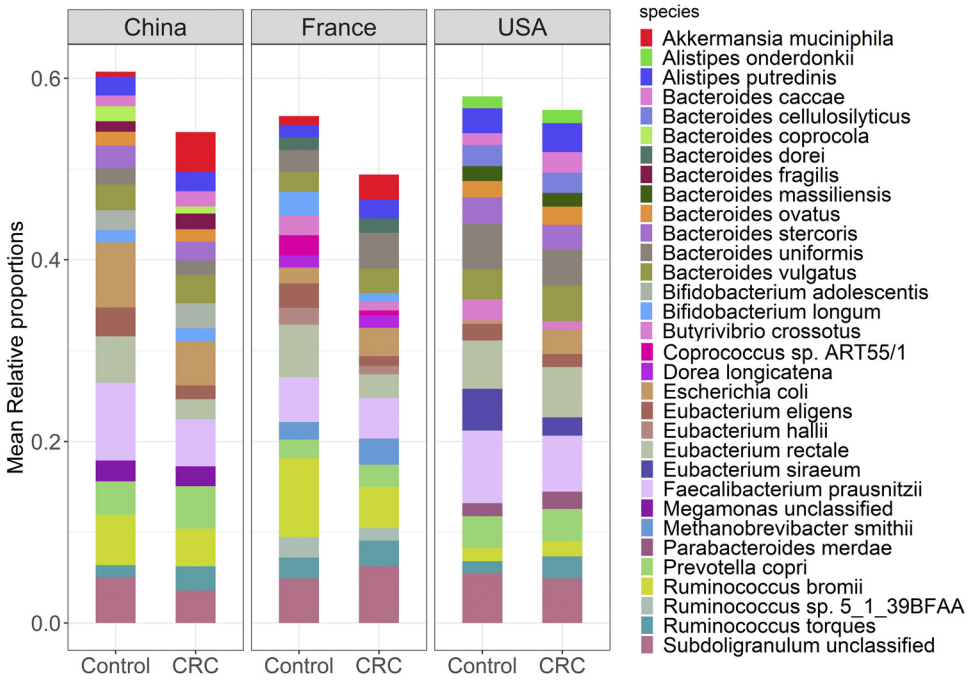

FIG 1 Stacked bar plot shows the mean relative proportions of the top 20 species-level taxa of gut microbial communities in CRC and healthy controls.

datasets, respectively. Mean relative abundances of the top 20 species covered >50% of total microbial abundance (Fig. 1, Data Set S2). And most of these species, including *Bacteroides* spp., and *Eubacterium* spp., belong to phyla Bacteroidetes and Firmicutes, respectively. High proportions of *Faecalibacterium prausnitzii*, *Eubacterium rectale*, and *Eubacterium eligens* were observed in the control group compared to the CRC group, whereas high proportions of *Prevotella copri*, *Ruminococcus torques*, and *Bacteroides vulgatus* were observed in CRC group compared to control group across the three geographic regions. At the species level, the alpha diversity of metagenomes was not significantly different between CRC and control groups across geographic regions (Fig. S1A). The species diversity in CRC and control groups was independent of potential confounding factors like age, BMI, and sex, as shown in Fig. S1B. The non-metric multidimensional scaling (NMDS) along with PERMANOVA evaluation based on Bray-Curtis dissimilarity measures showed significant differences between CRC and control metagenomes in France and China datasets and no differences between CRC and control metagenomes in the USA dataset ($P < 0.05$) (Fig. S2A to D). Similar comparisons based on the Bray-Curtis dissimilarity measures showed significant differences in CRC and control metagenomes across the geographical regions ($P < 0.05$) (Fig. S3A and B).

Altered microbial associations in CRC across geographic regions. Microbial associations in CRC were analyzed at the species level using co-occurrence networks, *LEfSe* algorithm, and RF methods. Differential microbial species were selected based on the geographic region using each method and evaluated using RF models.

In the co-occurrence network analysis, we calculated the correlation coefficients between microbes in CRC cases and controls separately and constructed networks by selecting the significant positive correlations ($r > 0.4$, $q < 0.05$) for each geographic region. A cluster analysis of networks showed various microbial interactions in CRC and control networks. We identified a cluster formed from oral microbes such as *G. morbillorum*, *Porphyromonas asaccharolytica*, *Parivimonas* spp., *P. stomatis*, *Prevotella intermedia*, *Parvimonas micra*, and *F. nucleatum* in all CRC networks (Fig. S4). The matched cluster was observed in the Chinese control network, however, it showed a high abundance of these microbes in CRC cases compared to healthy controls (Fig. S5), and the results were consistent with previous studies (28, 29).

Further, we synchronized the CRC and control networks of each geographic region using *DyNet* software and identified 156, 65, and 81 rewired nodes between CRC and control networks of USA, French, and Chinese datasets, respectively. The rewired node score (D_n-score) between the two networks indicates the changed interactions among the microbes (e.g., DyNet visualization of synchronized networks for China dataset shown in Fig. 2A, and corresponding unsynchronized networks shown in Fig. S6A and B). Similar networks for the USA and France datasets were shown in Fig. S7 and S8. The rewired nodes were identified based on the D_n-score (Data Set S3) and compared among the datasets. A set of 40 microbes with changed node perturbations were found in all datasets, and most of these include commensals such as *F. pratsnitzii*, *Roseburia hominis*, *Eubacteriun halli*, and *Blautia producta* and pathogens, such as *S. moorei*, *Ruminococcus gnavus*, and *Clostridium sp*. A high number of unique rewired nodes (87 differential species) were identified from the synchronized network of the USA dataset (Fig. 2B). We selected different sets of rewired nodes for assessing the performance of DyNet markers, and those include differential taxa identified in each geographic region by DyNet (DyNet dataset-specific markers) and a set of 78 rewired nodes that were common in 3 or 2 geographic regions (common DyNet-specific markers).

LEfSe analysis revealed the significant differences in microbial species between CRC and control groups (LDA score $>$ 2.0, $P < 0.05$) (Fig. 3A and Fig. S9A and B). A total of 110 unique microbial species differed between CRC and control groups among geographic regions. Among those, 8 pathogenic microbes, *B. fragilis*, *C. symbiosum*, *F. nucleatum*, *G. morbillorum*, *Parvimonas* spp., *P. stomatis*, *P. asaccharolytica*, and *Prevotella intermedia*, were identified as CRC enriched microbes in all three geographic regions. Other species were identified as unique to a geographic region or as common in at least two geographic regions (Fig. 3B, Data Set S3). For performance evaluation, we selected different sets of markers, those included LEfSe identified markers for each geographic region separately (LEfSe dataset-specific markers), and a set of 24 common LEfSe dataset-specific markers present in 3 or 2 geographic regions (common LEfSe-specific markers).

We built RF models by training on the individual dataset from each geographic region with 10-fold cross-validations and identified 40 top-ranked species from each geographic region (more than one species could get the same rank) as a differential set of markers. Together, these markers included 119 unique species, and among those, 15 were found in all three geographic regions. Others were randomly distributed among the geographic regions (Fig. 4B, Data Set S3). Some of the RF-identified microbial markers with ranks below 20 in at least one geographic region are shown in Fig. 4A. For performance evaluation, we selected different sets of RF-identified differential taxa for each geographic region (RF dataset-specific markers) and a set of 58 RF dataset-specific differential markers common in 3 or 2 geographic regions (common RF-specific markers).

Overall, differential analysis using these three methods showed that both dysbiotic and healthy gut microbiota differ based on geographic location, and the three methods used in this analysis confirmed that the results are also sensitive to analysis methods.

Identification of global biomarkers for CRC. RF classifiers were built based on the dataset-specific (markers identified from each dataset using three different methods described above) and common method-specific markers (combined dataset markers commonly found in at least two out of three datasets identified by the same method). To test the hypothesis that the set of CRC-associated taxa commonly present across different geographical regions could improve the prediction performance, we compared the performance score (AUC values) from the above classifiers, i.e., dataset-specific (Fig. 5A) and common method-specific (Fig. 5B). Results showed common DyNet-specific and common RF-specific markers performed better than the respective dataset-specific markers. The classifiers based on DyNet-identified markers in individual datasets (USA-, France- and China-specific markers) showed average performance scores ranging from 0.57 to 0.60 whereas classifiers based on common DyNet-specific markers' average performance scores ranged

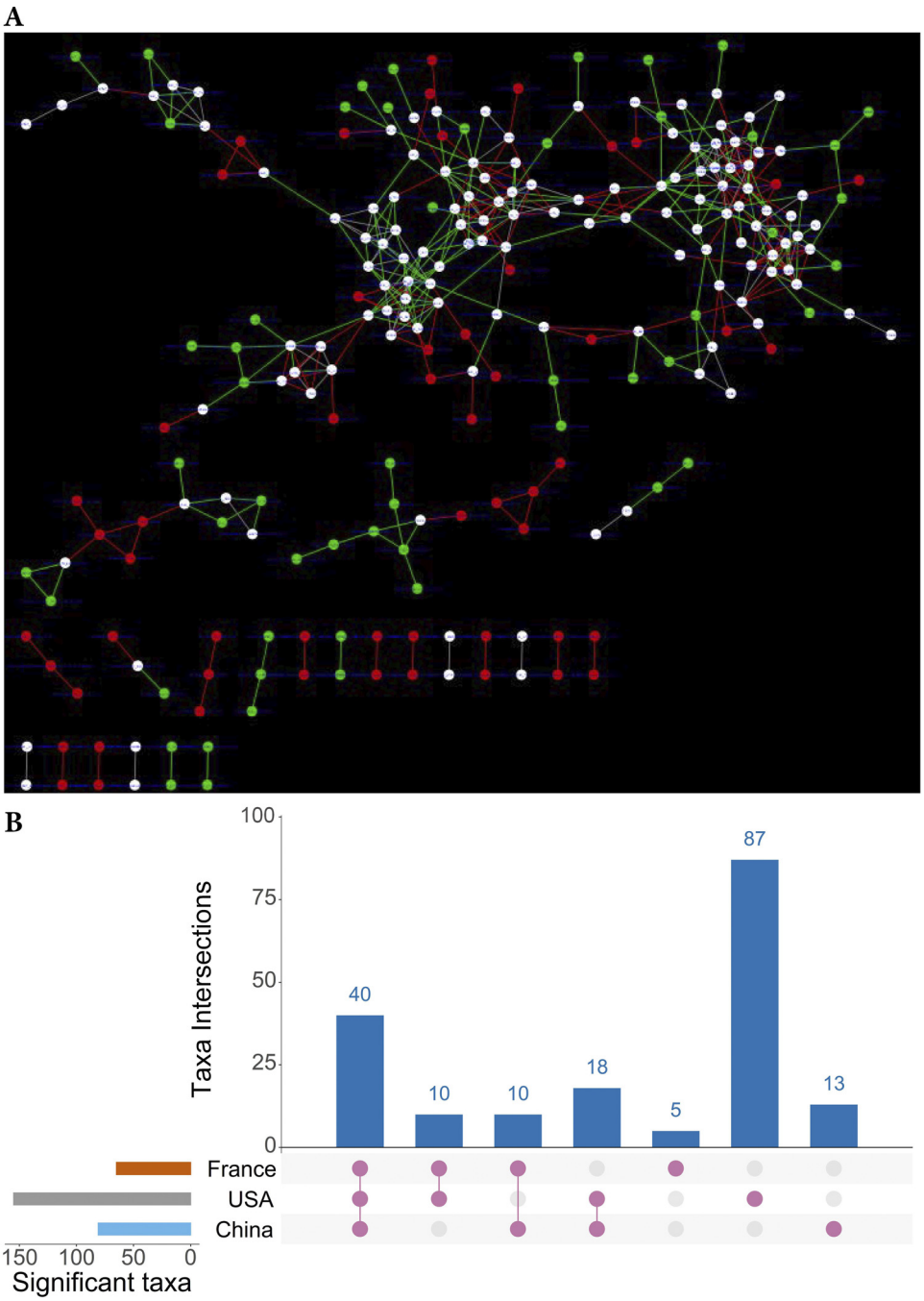

FIG 2 Co-occurrence networks and DyNet dataset-specific markers. (A) DyNet visualization of synchronized CRC and healthy control co-occurrence networks of microbial species from the China dataset. Red nodes and red edges are present only in the CRC network, green nodes and green edges are present only in the control network, and white nodes are present in both. (B) An Upset plot visualization of DyNet dataset-specific markers intersections across France, USA, and China datasets. Each bar represents the number of dataset markers in that category and orange dots below the bar indicates their conservation across the datasets. For instance, the 1st bar shows 40 Dynet dataset-specific markers that are common in all three datasets.

from 0.54 to 0.70 across the datasets. The dataset-specific markers identified by RF showed average performance ranging from 0.58 to 0.66, whereas common RF-specific markers average performance ranged from 0.59 to 0.72 across the datasets. Among the dataset-specific markers, LEfSe dataset-specific markers have average performance scores (0.58 to 069) across the geographic regions with a maximum value for the French population (0.82) and the average performance scores were

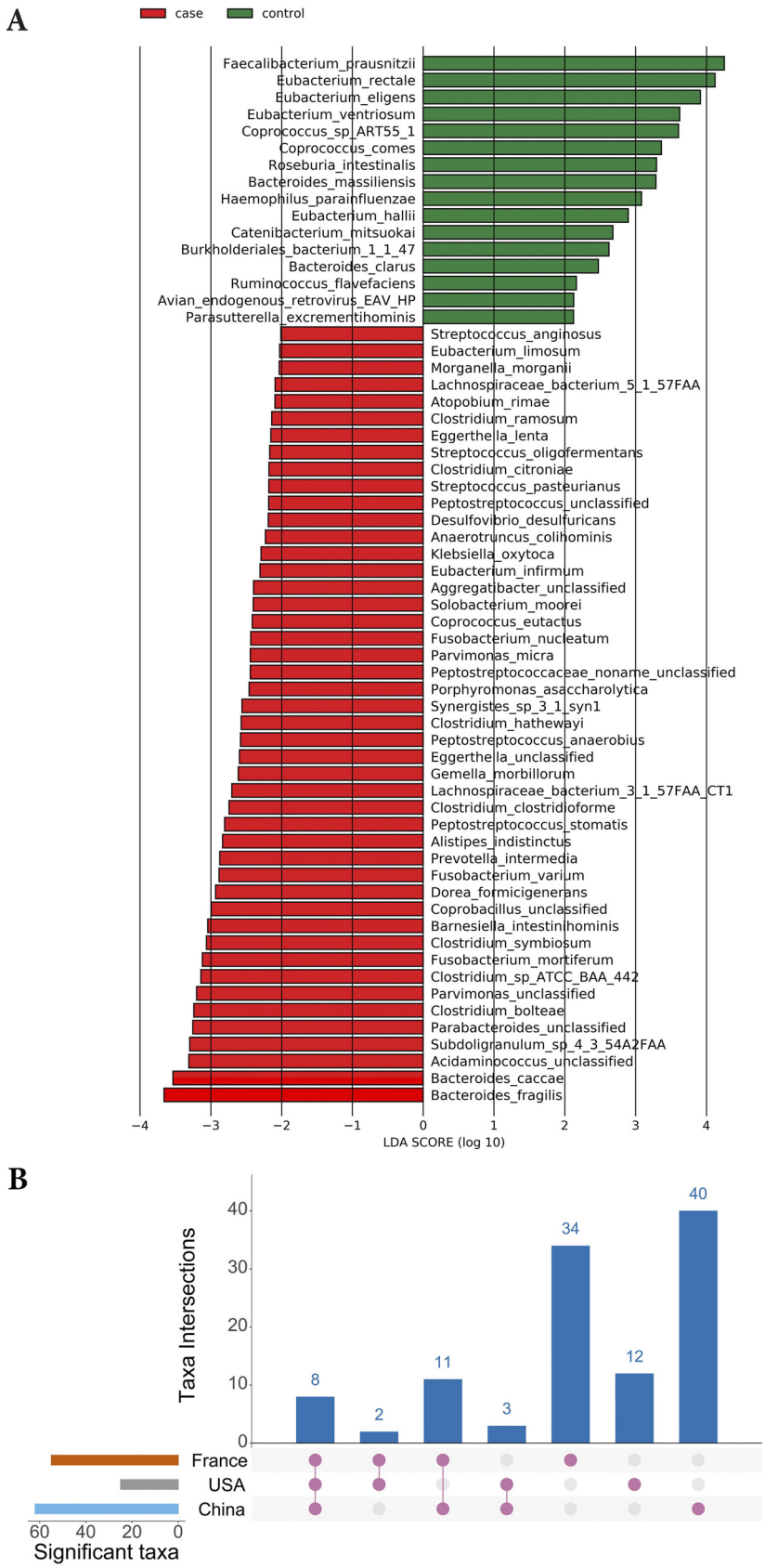

FIG 3 LEfSe analysis. (A) Histograms of differential species in Chinese dataset. CRC enriched species are indicated with a negative LDA score (red), and species enriched in healthy controls are indicated

(Continued on next page)

nearly equal to average performance scores of common LEfSe-specific markers (0.57 to 0.70). Overall, the common method-specific markers performed better than dataset-specific markers. Among the common method-specific biomarkers classifiers, RF-specific classifiers showed the highest AUC scores in cross-validation and cross-cohort validation (range of AUC score 0.56 to 0.78), whereas LEfSe-specific markers showed AUC scores ranging from 0.53 and 0.74 and DyNet-specific markers showed performance values ranging from 0.50 to 0.73 on cross-validation of datasets. Hence, we considered common RF-specific markers for further validation on Austrian and Japanese datasets, which were not part of the training datasets.

The routine clinical procedures require a minimum set of diagnostic markers that are cost-effective and advantageous. For this purpose, we took 58 RF-specific markers and another 20 species having RF ranking below 30 in at least one dataset. Then calculated the random effect size of each marker based on standardized mean differences and selected the 21 biomarkers with the largest effect size, and are associated with CRC in all three datasets or associated with controls in all three datasets (Fig. 6A) and validated on the Japanese dataset (258 CRC cases and 246 healthy controls) and the Austrian dataset (46 CRC cases and 57 healthy controls). The small set of RF-specific markers (21 species) has similar performance compared with the larger set of RF-specific markers (58 species) on the USA, Chinese, and French datasets (cross-validation AUC score range from 0.62 to 0.78 AUC) (Fig. 6B) and cross-validation AUC scores on Austrian and Japanese datasets were 0.66 and 0.61. Among the 21 microbial markers, 14 species have the largest effect size in CRC samples, whereas 7 species have the largest effect size in control samples across the USA, China, and France regions (Random-effect model fit, $P < 0.0001$). The species associated with CRC include *C. symbiosum*, *F. nucleatum*, *R. torque*, *G. morbillorum*, *S. moorei*, *P. micra*, *Clostridium citroniae*, and others. In contrast, most of the nonpathogenic microbial species associated with controls include *E. eligens*, *Eubacterium ventriosum*, *E. hallii*, *Bifidobacterium catenlatum*, and others (Fig. 6A). Most of the CRC biomarkers reported in this study are consistent with previous studies, where *F. nucleatum* reported as an oral pathogen was shown enriched in CRC patients (28, 30–32), *P. stomatis* and *S. moorei* were reported as enriched in saliva and stool of CRC patients (33), *P. micra*, an obligate anaerobe was associated with CRC and *E. ventriosum* was shown to be associated with healthy controls (7), and *B. fragilis* was identified as a CRC biomarker in the previous studies (29, 34).

Gene families of individual strains of selected microbial species. Strain-level gene families of microbial species were investigated to better understand the genomic differences of CRC pathogens across geographic regions using PanPhlAn software. For this analysis, we considered CRC-associated microbes, *F. nucleatum*, and *B. fragilis* and analyzed the strain-specific genes present in 346 (180 CRC cases and 166 controls) metagenomes from the USA, French and Chinese populations. *F. nucleatum* was found in 74 CRC and 11 control metagenomes, and its strain-level profiles were identified based on its 15,239 pangenome gene families, whereas *B. fragilis* was found in 149 CRC and 110 control metagenomes, and its strain-level gene-families were identified based on its 29, 335 pangenome gene-families. Individual strains of *F. nucleatum* were detected in 11 metagenomes (10 CRC and 1 control) (Data Set S4). Statistical analysis of gene families showed significant differences in 80 gene families across *F. nucleatum* strains of USA, Chinese, and French populations ($P < 0.05$) (Fig. 7). The *F. nucleatum* strains of the USA were separated from those of the French and Chinese populations, whereas few strains from the French population were similar to strains from the Chinese population. Further mapping with the UniProt database showed that LPS core

FIG 3 Legend (Continued)
with a positive LDA score (green). Only species with an LDA score >2 at $P < 0.05$ are shown. (B) An Upset plot visualization of LEfSe dataset-specific markers intersections across the three datasets (France, USA, and China). Each bar represents the number of differential species in that category and orange dots below the bar indicate their conservation across the datasets. For instance, the 1st bar shows eight LEfSe identified differential species that are common in all three datasets.

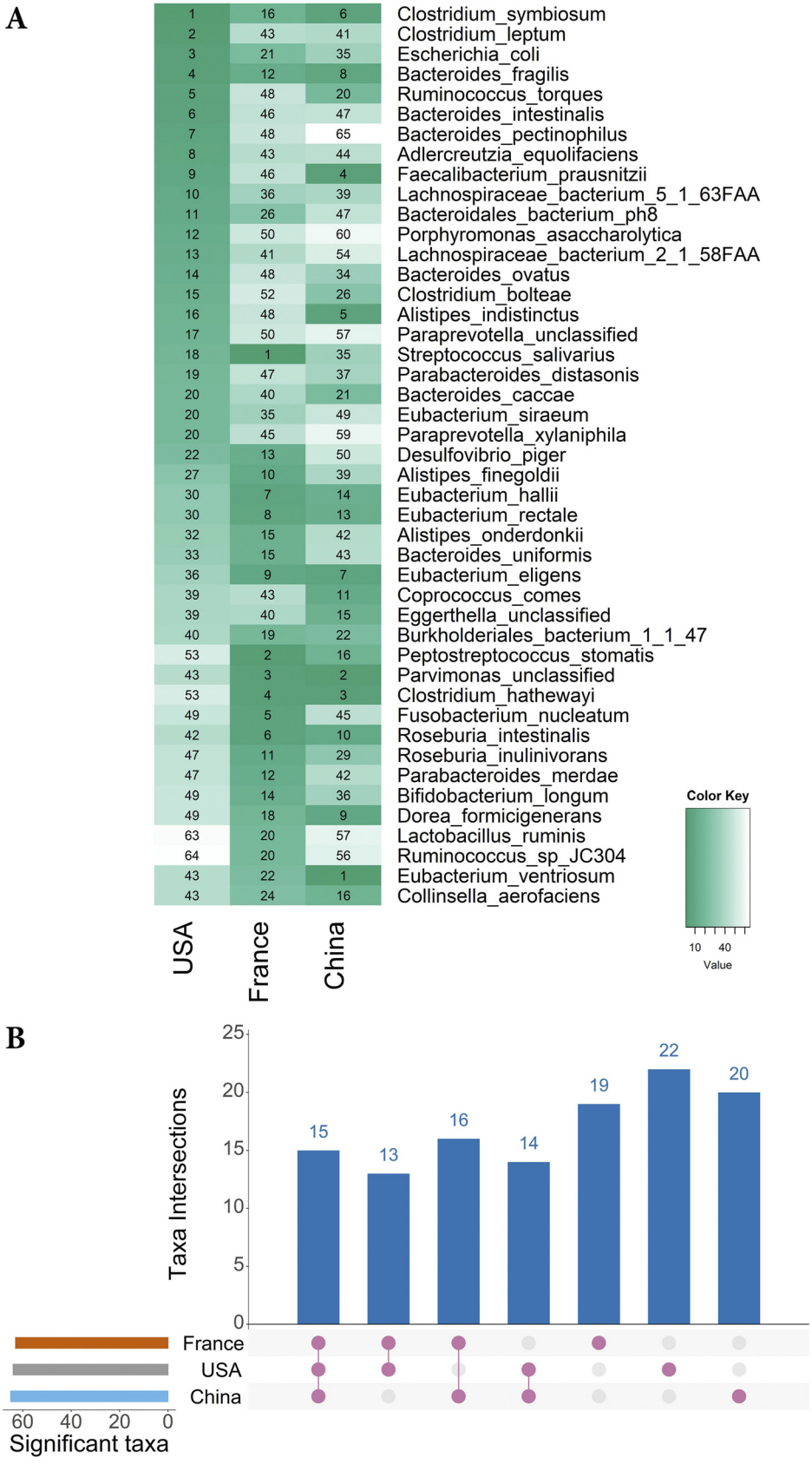

FIG 4 RF identified microbial markers of CRC in USA, French, and Chinese datasets. (A) In the RF cross-validations, the prediction performance of each species was scored based on internal RF rankings. Rankings

(Continued on next page)

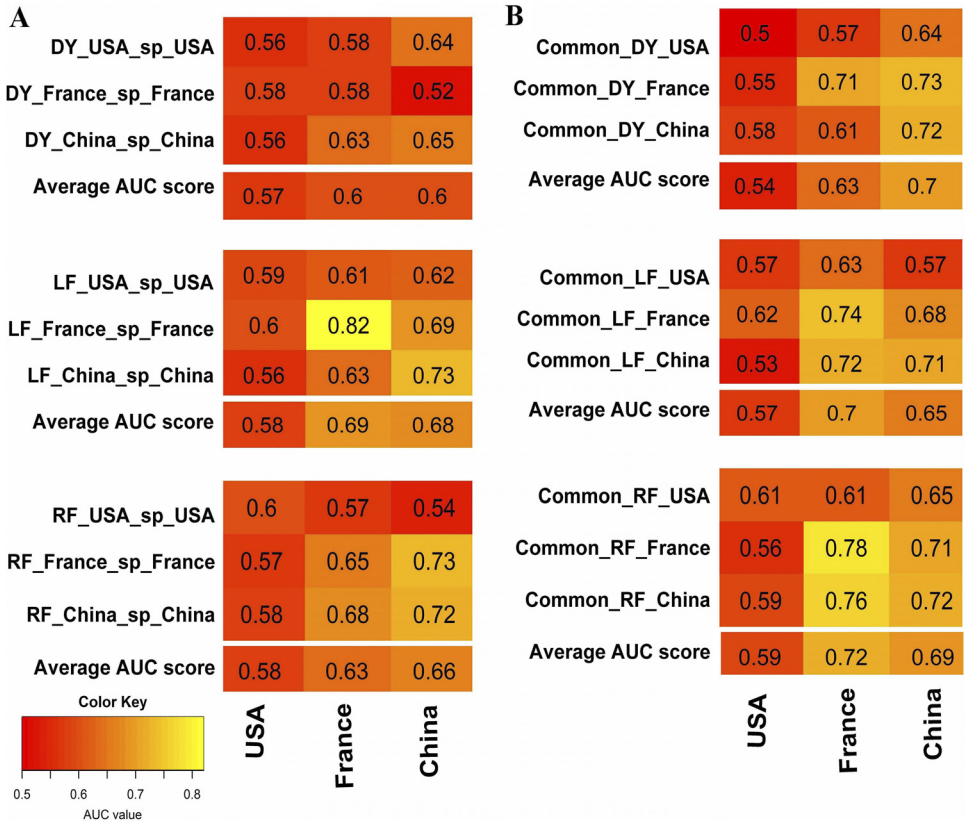

FIG 5 Prediction performance of the RF classifiers. Row indicates the RF classifier trained on the dataset-specific or common method-specific CRC markers; column indicates the classifier applying to the dataset of the corresponding column. In each three by three matrix of AUC values, diagonal values represent the AUC values of cross-validation obtained by using the trained row RF classifier on the column dataset, and off-diagonal values represent the AUC values of cross-cohort validation obtained by applying the trained row RF classifier on corresponding column dataset, (A) RF classifier was built from each dataset-specific markers (row). (B) RF classifier was built from the common markers present in at least two datasets from the USA, France, and China (common method-specific markers). 'Average AUC score' row represents the column average of the corresponding three-by-three AUC score matrix. Notation: e.g., DY_USA_sp_USA means classifier trained on the USA data based on the USA-specific markers identified by DyNet method, common_DY_USA means classifier trained on the USA data based on the common markers identified by DyNet method those are present in at least in two datasets DY, DyNet; LF, LEfSe; RF, random forest; and sp, specific.

biosynthesis, adenosylcobalamin biosynthesis, NAD (+) biosynthesis, isoprenoid biosynthesis, phospholipid metabolism, and tRNA modification pathways were significantly different across the strains of three geographical regions. PanPhlAn analysis of *B. fragilis* identified strains in 32 metagenomic samples (23 CRC and 9 controls) (Data Set S4), and 605 gene families were significantly different in the strains from the USA, French, and Chinese populations ($P < 0.05$) (Fig. S10). Mapping with the UniProt database showed the functional differences in gene families related to choloylglycine hydrolase, TonB-dependent receptors, OmpA/MotB domain proteins, histidine kinase, BfmA, and ArsR family transcriptional regulator proteins.

Predicted metabolome changes associated with CRC. The metabolite profiles of the gut microbial communities from the USA, French, and Chinese populations were predicted using *MelonnPan* software. This analysis predicted 135 metabolites from each dataset. The predicted metabolite profiles were filtered to remove <0.01% rela-

FIG 4 Legend (Continued)
of the RF-identified species markers with a rank below 20 in at least one dataset are shown in the figure. (B) An Upset plot visualization of RF dataset-specific markers intersections across the three datasets. Each bar represents the number of dataset-specific markers in that category and the orange dots below the bar indicate their conservation across the datasets. For instance, 1st bar shows the 15 RF dataset-specific markers that are common in all three datasets.

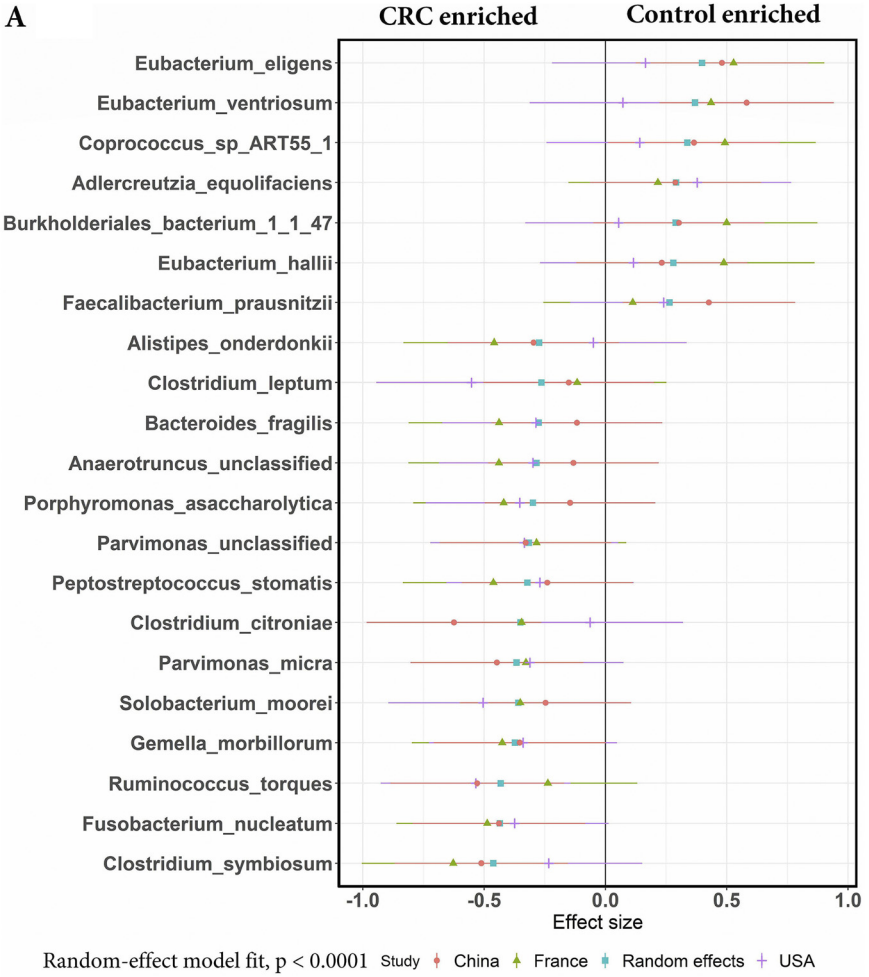

FIG 6 (A) Meta-analysis of selected RF CRC biomarkers markers using MetaPhlAn2 profiles from USA, China, and France geographic regions. The colored lines represent the 95% confidence interval for each dataset and random effect model estimate. (B) Cross-validations of a minimum set of RF CRC biomarkers on USA, France, and China datasets. The AUC values on each cell of the heatmap were obtained by the RF classifier (built from selected RF features) trained on the dataset row and applying the classifier on the dataset column.

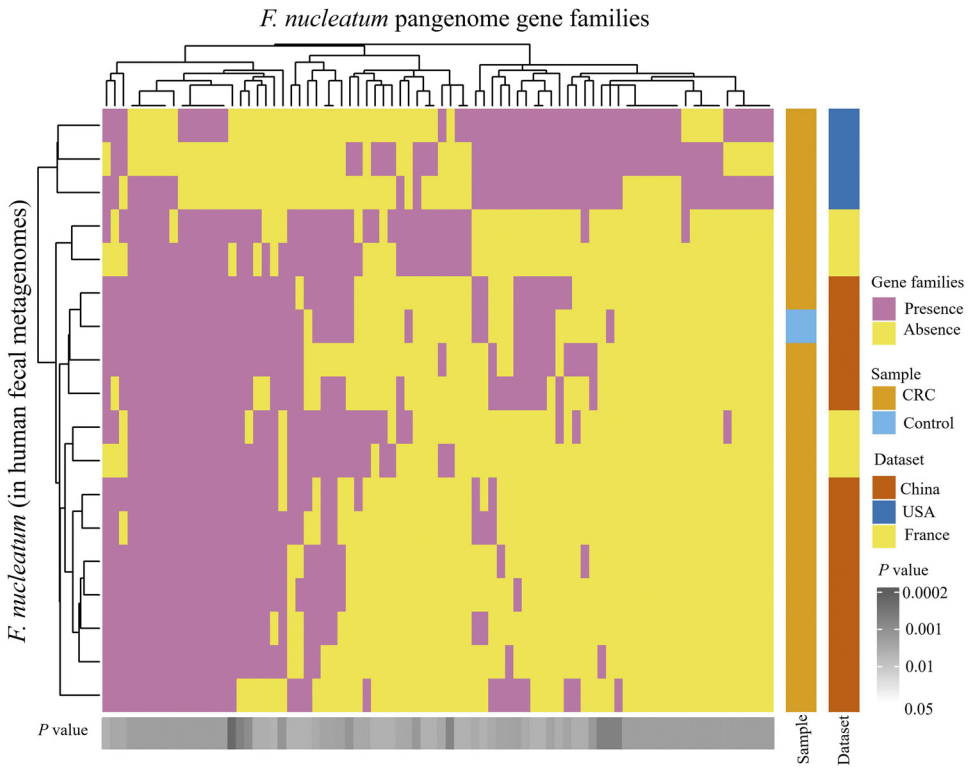

FIG 7 Heat maps of the strain-level genomic diversity of *F. nucleatum* across three geographic regions: USA, China, and France. The significant differential gene families ($P < 0.05$) were identified using Fisher's exact test on presence and absence gene family profiles.

tive abundance in ≥10% of the samples as suggested by the developers of the *MelonnPan* software (35). After filtering we were left with 70 metabolites in each dataset. The filtered metabolite profiles of all datasets were merged and normalized and analyzed using *edgeR* and *limma* software. Differential analysis showed significant changes in the metabolite profiles between CRC and control samples (*limma*, $P < 0.05$; Fig. S11 and Data Set S4). Amino acids, including phenylalanine, valine, leucine, alanine, isoleucine, and tyrosine, and other metabolites, including cadaverine, succinate, and creatine, were highly enriched in CRC, whereas butanoic acid, L-glutamate, and L-aspartate were abundant in the controls (Fig. 8A). Similarly, pathway enrichment analysis of differential metabolites showed pathways related to amino acid metabolism enriched in CRC (Fig. 8C), whereas arginine biosynthesis, nicotinate, and nicotinamide metabolism, D-glutamine, and D-glutamate metabolism, and butanoate metabolism were enriched in controls ($P < 0.05$, Fig. 8B). The metabolite profiles and pathway associations in CRC indicate altered energy sources required for the high metabolic growth rates of CRC cells. In addition to glucose, the cancer cells use amino acids like glutamine, valine, leucine, and isoleucine as alternate energy sources to meet the high energy demand for growth and as biosynthetic molecules required for tumor growth (36).

We also estimated the correlations between gut microorganisms and metabolites in CRC samples from three geographic regions USA, France, and China using Spearman correlation and examined the correlations between differential CRC microbial biomarkers and metabolites identified in our study. The heatmap (Fig. 9) showed the separation of CRC microbial markers into clusters (vertical axis). The top clusters are mostly control-associated organisms like *F. prausnitzii*, *E. eligens*, *E. ventriosum*, and *E. halli* and the bottom clusters are CRC-associated such as *F. nucleatum*, *S. moorei*, and *C. symbiosum* in one cluster and *R. torques*, *B. fragilis*, *Parvimonas*

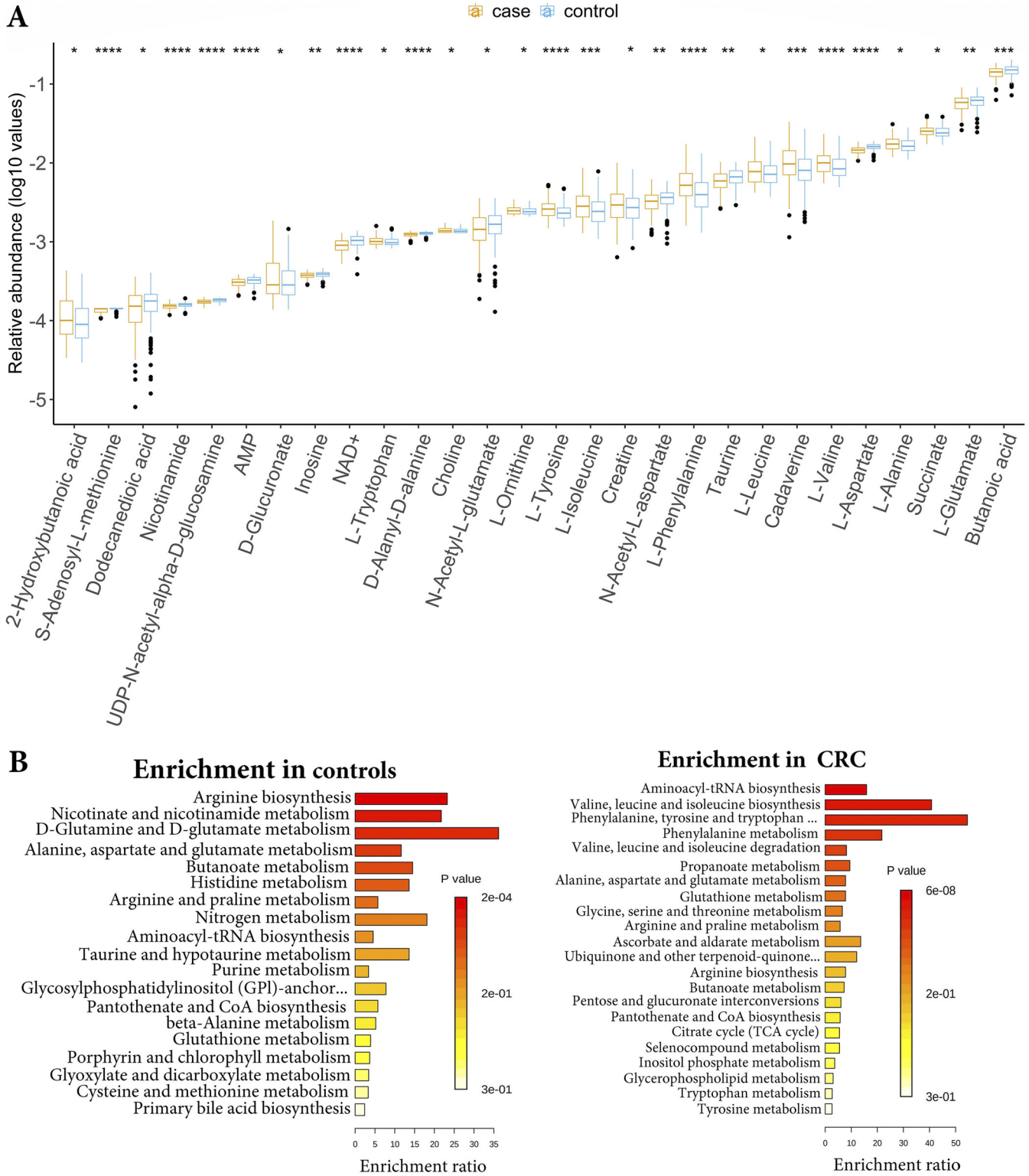

FIG 8 Analysis of MelonnPan predicted metabolites from the USA, China, and France datasets. (A) The relative abundance of significantly different ($P < 0.05$) metabolites between CRC and healthy control groups. Blue indicates the control samples and red indicates the CRC samples from all three datasets. Enrichment of pathways based on predicted metabolites in (B) healthy controls and (C) CRC samples.

spp, *G. morbillorum*, and *P. stomatis* into another cluster. Even though *Alistipes onderdonkii*, a CRC-associated microorganism, was clustered with *F. prausnitzii* and *E. eligens* (control-associated), they differed in correlations with butanoic acid, glutamate, aspartate, and tyrosine. However, the CRC-associated microorganisms, *C.*

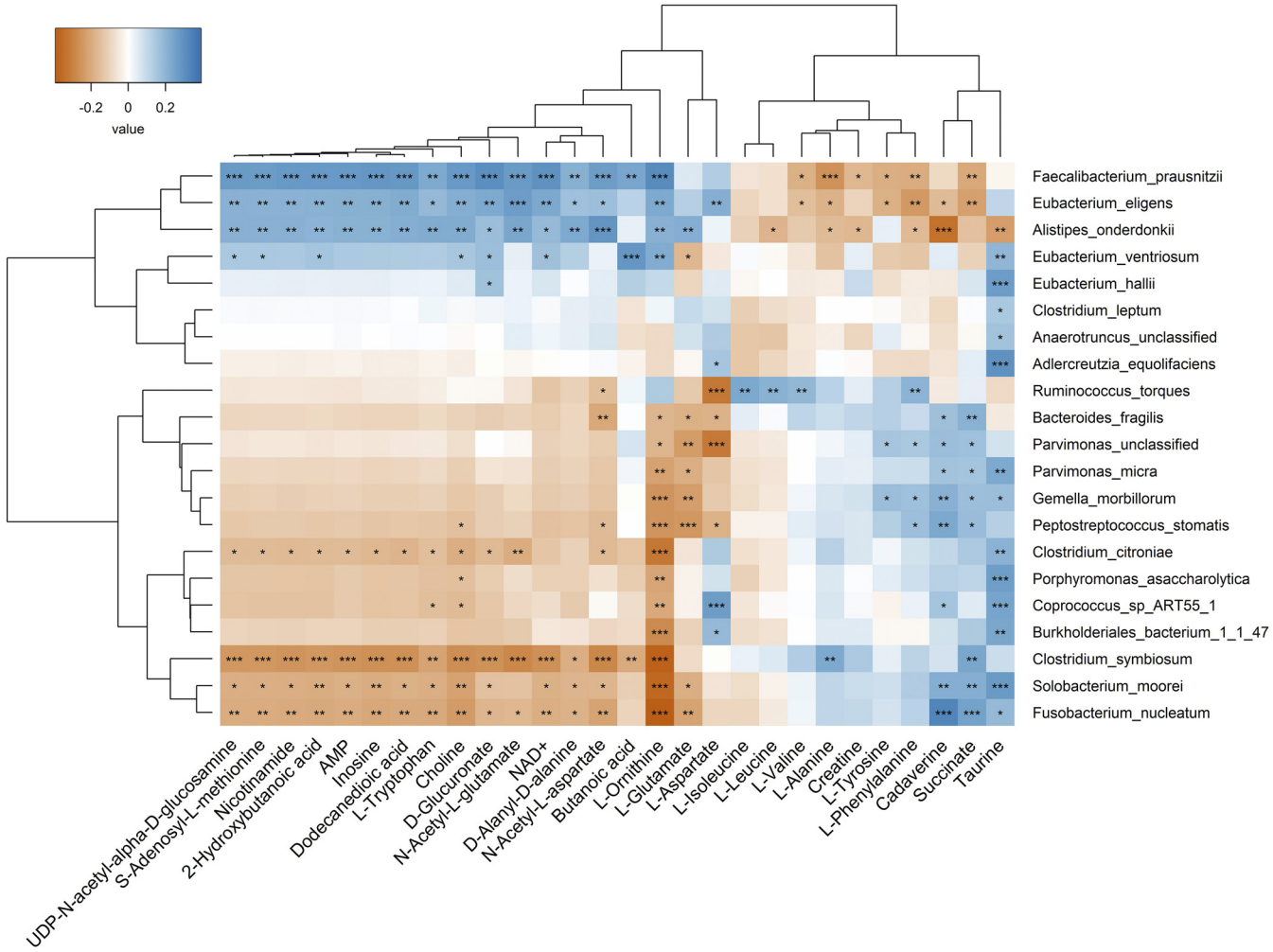

FIG 9 Correlation analysis between 21 CRC microbial markers and 28 metabolites that were significantly different between CRC and healthy gut communities, 14 were enriched in CRC cases and 14 were enriched in healthy controls. Red indicates the positive correlation and blue indicates the negative correlation. *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$.

citroniae, and *P. asaccharolytica* grouped with control-associated microorganisms like *Coprococcus* sp. *ART55_1* and *Burkholderiales bacterium 1_1_47* differed in correlations with aspartate.

Metabolites such as cadaverine, succinate, and phenylalanine were enriched in CRC samples and showed significant positive correlations with most CRC-associated microbes like *P. stomatis*, *G. morbillorum*, *B. fragilis*, *Parvimonas* spp., *F. nucleatum*, *S. moorei*, and *C. symbiosum*, while showing significant negative correlations with control associated microbes, *F. prausnitzii*, and *E. eligens* ($P < 0.05$). Branched-chain amino acids, including valine, leucine, and isoleucine, were positively correlated with the presence of *R. torques*. Butyrate, a short-chain fatty acid metabolite, was positively associated with *F. prausnitzii* and *E. ventriosum*, and negatively associated with *C. symbiosum* ($P < 0.05$). Alanine was enriched in CRC samples and showed a significant positive correlation with *C. symbiosum*, while taurine was positively correlated with both CRC- and control-associated gut microorganisms. Ornithine showed significant positive correlations with most of the control-associated microbes and negative associations with CRC pathogens ($P < 0.05$) (Fig. 9, Data Set S4). Ornithine is an important metabolite in the arginine metabolism produced by lactobacilli, which helps in gut mucosal barrier function (37). A major cluster of metabolites such as nicotinamide, choline, tryptophan, 2-hydroxybutanoic acid, and others showed positive correlations with control-associated microbes, *F. prausnitzii* and *E. eligenes* (top cluster) whereas

negatively correlated with *F. nucleatum*, *S. moorei*, and *C. symbiosum* (bottom cluster). CRC-associated bacteria, *A. onderdonkii* showed positive correlations with most of these metabolites except butanoic acid and aspartate ($P < 0.05$).

DISCUSSION

Various studies were focused on the dysbiotic gut microbiota to identify specific microbial associations as etiological agents of CRC (22, 23, 38, 39), but the results were inconsistent across the studies. Because the composition of the gut microbiome varies with different environmental stimuli, dietary habits, and genetic traits, the independent geographic region/cohort-specific dataset analysis would result in biased associations related to dominant factors that cannot be generalized. In this study, we carried out a meta-analysis of multiple datasets from diverse geographic regions and identified a set of global CRC biomarkers associated with all datasets. By selecting common CRC-microbial associations across geographic regions, we can avoid nonspecific associations introduced by the heterogeneity factors. Our results showed that the RF method performed better than the other methods, followed by LEfSe and co-occurrence network methods. RF-specific markers showed a consistent increase in performance from dataset-specific markers to common method-specific markers, whereas the LEfSe method performed better with independent dataset-specific biomarkers than using common LEfSe-specific biomarkers. Co-occurrence networks are often used to examine microbial associations among key taxa in ecological communities. A recent microbiome study used this approach to identify the key taxa associated with acute pulmonary exacerbations (40). In our study, we implemented a co-occurrence network for the first time to identify and compare CRC biomarkers from the rewired nodes between CRC and control networks. Rewired nodes across the networks indicate altered interactions among the community members. In our analysis, co-occurrence network-specific features showed a lower performance score compared to the other two methods. However, it helped to identify significant interactions between oral pathogens in CRC and a healthy gut, and these results are consistent with the previous study (28). RF was one of the most widely used methods for the identification of biomarkers in microbiome studies. In our study, RF models showed a consistent increase in performance from RF dataset-specific markers to common RF-specific markers. Due to this reason, we selected the global CRC markers identified by the RF method.

Most of the CRC biomarkers that we identified in this study are consistent with the previous investigations (39, 41). We identified a new CRC biomarker, *A. onderdonkii*. The genus *Alistipes* is generally considered a commensal organism. However, it may contribute to inflammatory effects in the host under dysbiotic conditions due to the presence of putrefaction pathways such as histidine degradation and production of tetrahydrofolate, indole, and phenol (42). Similarly, we also identified *C. citroniae* as a biomarker, it belongs to the newly defined *Lachnoclostridium* genus. It has been reported in early stage to later stage CRC samples with incremental abundances (43). On the other hand, the control-associated biomarkers are gut commensal organisms, which include *F. prausnitzii*, *E. ventriosum*, and *E. hallii*. They are known to produce SCFAs and regulate gut mucosal health (44).

PanPhlAn analysis of CRC-associated microbe, *F. nucleatum*, showed a significant difference in gene families related to isoprenoid biosynthesis and phospholipid metabolism. These pathways are involved in lipid metabolism, and dysregulation of lipid metabolism was identified as a characteristic feature and correlated with shorter survival rates in CRC (45, 46). Similarly, other studies showed increased phospholipids were observed in colonic neoplasms (47) and upregulation of the mevalonate-isoprenoid (MIB) pathway was identified in CRC stem cells (48). Enterotoxigenic *B. fragilis* has been shown to contribute to colon carcinogenesis, and its pathogenicity is mainly due to its capsule, outer membrane proteins, and a metalloprotease protein toxin (*B. fragilis* toxin) (49). In this study, we identified that the *B. fragilis* strains present in the USA, Chinese, and French populations mainly differed in the gene families related to TonB-dependent receptor, OmpA/MotB domain protein, BfmA, choloylglycine hydrolase, MobA, multidrug efflux MFS transporter, type IV

secretion system needle protein Hcp and other putative proteins. Most of these proteins play important roles in bacterial pathogenicity like toxin transport, colonization, biofilm formation, nutrient uptake, and multidrug resistance and are produced from bacterial pathogenicity islands (50). The type IV secretion system is involved in the transfer of virulence factors to the host and TonB-dependent receptors and OmpA are involved in the uptake of nutrients and other small molecules into the cell (51). Choloylglycine hydrolase is a conjugated bile salt hydrolase known to be produced by gut microbes (strains from Bacteroidetes) that has the potential to alter the host fatty acid metabolism by mediating bile salt hydrolysis (52). In strain-level metagenomic analysis, both *F. nucleatum* and *B. fragilis* strains showed differences in gene families that can affect the host fatty acid metabolism. This indicates the crosstalk between the gut microbes and host metabolism through microbial metabolites in CRC.

The metabolite analysis showed enrichment of SCFAs such as butanoic acid in control samples than CRC samples and positively correlated with control-associated microbes. It is one of the by-products of the fermentation of fiber by gut microbiota, that is linked to gut homeostasis and the prevention of tumor growth (26, 53). Succinate, one of the intermediates in the tricarboxylic acid (TCA) cycle was enriched in CRC samples and positively correlated with CRC-associated microbes. It is also known to be produced by gut microbes such as *B. fragilis*, *F. prausnitzii*, *Alistipes* spp., *Prevotellaceae*, and others. Gut microbial dysbiosis can lead to the accumulation of succinate and intestinal inflammation (54). Both CRC- and control-associated gut microorganisms showed a positive correlation with taurine. Meat-rich diets promote taurine conjugation, which leads to increased taurocholic acid formation in the small intestine. Later the deconjugation of taurocholic acid by diverse gut bacteria (in the large intestine) generates cholic acids and taurine, which in turn are converted to carcinogenic secondary bile acid, deoxycholic acid, and a cytotoxic compound, hydrogen sulfide, respectively (55).

Interestingly amino acids valine, leucine, isoleucine, and phenylalanine showed enrichment in CRC than controls, which has been previously observed in CRC (56, 57). Glutamate was enriched in control samples compared to CRC samples, which contrasts with previous studies (57) that showed various interactions with both CRC- and control-associated microbes. Polyamine-cadaverine was enriched in CRC and positively associated with CRC-associated microbes, consistent with a previous study (56). Polyamines are essential for normal cell growth, they are produced by the host, gut bacteria, and dietary origin, however, dysregulation of polyamine metabolism is linked to colon cancer (58). Choline was enriched in CRC which is consistent with a previous study by Thomas et al. (41) where they showed an abundance of choline degradation genes (*cutC* and *cutD*) in the CRC gut microbiome. However, in our study, choline is positively correlated with control-associated microbes. Overall, our study provided a reliable set of global microbial biomarkers for CRC identification that can be used across different populations. Moreover, we reported the differentiating metabolites and important gut microbe-metabolite interactions in CRC, which may have the potential to influence host metabolism. This study would pave the way for further investigations that could lead to the development of noninvasive diagnostic tools and therapeutic interventions for CRC management.

MATERIALS AND METHODS

Fecal shotgun metagenomics sequencing datasets of CRC patients and healthy controls belonging to three different geographical regions USA (24), China (7), and France (23) were downloaded from the public database, the European Nucleotide Archive (ENA). The other two fecal shotgun datasets of CRC and control samples from Austria (22) and Japan (38) were downloaded from ENA and DDBJ Sequence Read Archive (DRA), respectively, for validation purposes. Details of the samples with accession numbers of the datasets used in this study are provided in Table S1, Metadata for USA, China, France, and Austria datasets were obtained as JSON-formatted files from EBI BioSamples and parsed using Perl/Python scripts into tables with different meta fields that include sample ID, study ID, secondary ID, sequencing type, country, age, BMI, and diagnosis. For the Japanese dataset, metadata was obtained from the original publication (38). For metabolomic profiling of gut communities, the metabolomic profiles of the Japan dataset were obtained from the original work by Yachda and group (38), and the list of 250 samples of the Japan dataset used in this analysis is in Data Set S4.

Taxonomic profiling of metagenomic datasets. Metagenomic sequencing datasets obtained from all five geographical regions were quality filtered using *FastQC* software and aligned against Coliphage phi-X174 (PhiX) and GRCh38 human reference genomes to remove PhiX and human read contaminations using *BBduk* and *BBmap* tools (59). After preprocessing, samples from all datasets were quantified for taxonomic composition using *MetaPhlAn2* (60) software. *MetaPhlAn2* relies on unique clade-specific markers identified from about 17,000 reference genomes from bacterial, archaeal, viral, and eukaryotic microorganisms for microbial profiling and quantification. In this study, we used species-level taxonomic profiles in all analysis methods. The taxonomic profiles of each dataset (each dataset represents a geographic region considered in this study) were filtered to remove species present at a relative abundance value of $< 0.1\%$ across the samples to reduce potential false-positives (61). For the downstream analyses, the species abundance of each dataset was mapped with related metadata using the *Phyloseq* package in R. Alpha and beta diversity indices of the gut microbiome for each dataset were estimated based on the relative abundance profiles of species using the *Vegan* package in R and plotted using the *ggplot2* package in R.

Identification of microbes specific to CRC or healthy gut communities. Statistical, machine learning, and network-based methods were used in this study to identify CRC biomarkers. To describe briefly, linear discriminant analysis (LDA) effect size (*LEfSe*) is a statistical method to identify key taxa that are significantly different between CRC cases and controls. Species-level relative abundance data along with sample details were analyzed separately for each dataset using *LEfSe* software (62). It uses nonparametric tests to identify significant features, performs subclass comparisons, and then LDA to estimate the effect size of identified features. The differential taxa for each dataset were considered based on the Effect size LDA score > 2 and FDR-adjusted $P \leq 0.05$, (ii) RF algorithm is the most used machine learning method on microbiome data to identify differential microbial features (41). The random forest and caret packages in R were used for the RF model building. Standardized relative abundance data (Z-score) of microbial species in all samples and group information were used as input for RF analysis. RF classifiers were built by training on each dataset separately with 10-fold cross-validation. RF features with a higher mean decrease in the Gini index (top-ranked microbial species) from each dataset were considered differential markers. The performance of each classifier was evaluated using 10-fold cross-validation and AUC metrics, (iii) Co-occurrence networks for CRC cases and controls were inferred separately for each dataset based on the correlation coefficients calculated using *SparCC* v 0.1 software (63) from the species-level relative abundance data (normalized to 1 million counts) of corresponding datasets. *SparCC* estimates the Pearson correlation between species from the log-ratio transformation values while accounting for the compositionality of metagenomics data (64). The correlation coefficients were estimated by the permutation-based approach in *SparCC*, and the Pseudo *P* values were calculated for the bootstrapped correlation coefficients. Network plots were constructed considering significant co-occurrence correlations (FDR < 0.05) in *Cytoscape* 3.7.1 (65). Then, the highly connected dense regions in the network were detected using *the MCODE* plugin. The CRC and control networks of each dataset were synchronized using the *DyNet* plugin (66) and then identified the most rewired nodes across the two network states. The highly rewired nodes between CRC and control networks (rewiring metric or D_n-score ≥ 2.0 and an edge count ≥ 4) from each dataset were considered differential taxa for CRC.

Assessment of predicted biomarkers. Differential taxa from each method were tested among the datasets using RF models with 10-fold cross-validations. The performance of each classifier was measured in terms of the AUC metric and selected biomarkers with high AUC scores. The random effect size of each selected biomarker was calculated based on standardized mean differences using the *Metafor* package (67) in R. Biomarkers with the highest random effect size associated with either CRC or control groups were selected as CRC global biomarkers and further analyzed for metabolite interactions in CRC gut communities.

Strain-level metagenomic profiling of CRC pathogens. The gene composition of individual strains of CRC-associated microbe in CRC and healthy gut metagenomes from the USA, China, and France were identified using *PanPhlAn* v 3.0 software (68). *PanPhAn* identifies the gene presence and absence within different strains of species in metagenomes based on the entire set of the species' pangenomes. Differential analysis of gene families was performed across the geographical regions using Fisher's exact test (fisher.test() in R). Then significant differential gene families were mapped to UniProt Knowledgebase (69) to understand their functional roles in CRC.

Metabolomic profiling of gut communities. Metabolomic profiles of CRC and healthy gut metagenomes were predicted using *MelonnPan* v1.0.0 software (35). In brief, *MelonnPan* builds a model based on paired metabolomic and metagenomics features of a community and uses that model to predict the metabolite relative abundances for a given metagenomic community based on its set of features derived from sequencing data. Paired microbial and metabolite relative abundance data of the Japanese CRC dataset (250 samples listed in Data Set S4) were used to build a predictable model and predicted the metabolite profiles of the USA, China, and France datasets based on species relative abundances. Predicted metabolomic profiles were combined and analyzed using the *edgeR* and *limma* package in R (70, 71), and the significant differential metabolites were analyzed for pathways using *MetaboAnalyst* v5.0 (72). Correlations between differential metabolites and CRC microbial biomarkers were estimated using the Spearman correlation method and results were plotted using *ggplot2* and *gplot* packages in R.

Statistical analyses. Nonparametric Fisher's exact test was used for the differential analysis of gene families. Microbe-microbe or microbes-metabolites correlations were estimated using Spearman's correlation coefficients. For statistical significance $P < 0.05$ and Benjamini-Hochberg corrected FDR values were considered appropriate. Standardized (Z score) data was used for metabolite analyses. The covariate effect on species diversity was tested using a multivariate linear regression method. The batch effect on metabolite

profiles was tested using *edgeR* and *limma* in R. All data analyses and visualizations were conducted in R v 4.0.0. and above (73).

Data availability. The original contributions presented in the study are included in the article/ Supplemental Material. Further inquiries can be directed to the corresponding author.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLSX file, 0.03 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.8 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 4, XLSX file, 2.6 MB.

SUPPLEMENTAL FILE 5, PDF file, 4 MB.

ACKNOWLEDGMENTS

We thank the Bioinformatics and System Biology Core at the University of Nebraska Medical Center (UNMC) for providing the computational resources, the Holland Computing Center, University of Nebraska-Lincoln (UNL) for providing access to supercomputers, and Neetha Nanoth Vellichirammal for proofreading.

This study was supported by National Science Foundation, United States EPSCoR Award (grant OIA-1557417).

C.G. supervised the study, N.A. and C.G. developed the concept and methodology, N.A. generated and analyzed the data, N.A. made the figures, N.A. wrote the original manuscript, and N.A. and C.G. reviewed and edited the manuscript. All authors have read and approved the final manuscript.

We declare no conflict of interest.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin 71:209–249. https://doi.org/10.3322/caac.21660.
2. Mokarram P, Albokashy M, Zarghooni M, Moosavi MA, Sepehri Z, Chen QM, Hudecki A, Sargazi A, Alizadeh J, Moghadam AR, Hashemi M, Movassagh H, Klonisch T, Owji AA, Łos MJ, Ghavami S. 2017. New frontiers in the treatment of colorectal cancer: autophagy and the unfolded protein response as promising targets. Autophagy 13:781–819. https://doi.org/10.1080/15548627.2017.1290751.
3. Karahalios A, English DR, Simpson JA. 2015. Weight change and risk of colorectal cancer: a systematic review and meta-analysis. Am J Epidemiol 181:832–845. https://doi.org/10.1093/aje/kwu357.
4. Wu XC, Chen VW, Steele B, Ruiz B, Fulton J, Liu L, Carozza SE, Greenlee R. 2001. Subsite-specific incidence rate and stage of disease in colorectal cancer by race, gender, and age group in the United States, 1992–1997. Cancer 92:2547–2554. https://doi.org/10.1002/1097-0142(20011115)92:10%3C2547::AID-CNCR1606%3E3.0.CO;2-K.
5. Boyle T, Keegel T, Bull F, Heyworth J, Fritschi L. 2012. Physical activity and risks of proximal and distal colon cancers: a systematic review and meta-analysis. J Natl Cancer Inst 104:1548–1561. https://doi.org/10.1093/jnci/djs354.
6. Yuhara H, Steinmaus C, Cohen SE, Corley DA, Tei Y, Buffler PA. 2011. Is diabetes mellitus an independent risk factor for colon cancer and rectal cancer. Am J Gastroenterol 106:1911–1921. https://doi.org/10.1038/ajg.2011.301.
7. Yu J, Feng Q, Wong SH, Zhang D, Yi Liang Q, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, Wang X, Xu X, Chen N, Wu WKK, Al-Aama J, Nielsen HJ, Kiilerich P, Jensen BAH, Yau TO, Lan Z, Jia H, Li J, Xiao L, Lam TYT, Ng SC, Cheng ASL, Wong VWS, Chan FKL, Xu X, Yang H, Madsen L, Datz C, Tilg H, Wang J, Brünner N, Kristiansen K, Arumugam M, Sung JJY, Wang J. 2017. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 66:70–78. https://doi.org/10.1136/gutjnl-2015-309800.
8. Fedirko V, Tramacere I, Bagnardi V, Rota M, Scotti L, Islami F, Negri E, Straif K, Romieu I, La Vecchia C, Boffetta P, Jenab M. 2011. Alcohol drinking and colorectal cancer risk: an overall and dose-Response meta-analysis of published studies. Ann Oncol 22:1958–1972. https://doi.org/10.1093/annonc/mdq653.
9. Wong SH, Yu J. 2019. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. Nat Rev Gastroenterol Hepatol 16:690–704. https://doi.org/10.1038/s41575-019-0209-8.
10. Li K, Peng W, Zhou Y, Ren Y, Zhao J, Fu X, Nie Y. 2020. Host genetic and environmental factors shape the composition and function of gut microbiota in populations living at high altitude. Biomed Res Int 2020.
11. Moran-Ramos S, Lopez-Contreras BE, Villarruel-Vazquez R, Ocampo-Medina E, Macias-Kauffer L, Martinez-Medina JN, Villamil-Ramirez H, León-Mimila P, Del Rio-Navarro BE, Ibarra-Gonzalez I, Vela-Amieva M, Gomez-Perez FJ, Velazquez-Cruz R, Salmeron J, Reyes-Castillo Z, Aguilar-Salinas C, Canizales-Quinteros S. 2020. Environmental and intrinsic factors shaping gut microbiota composition and diversity and its relation to metabolic health in children and early adolescents: a population-based study. Gut Microbes 11:900–917. https://doi.org/10.1080/19490976.2020.1712985.
12. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. 2005. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102:11070–11075. https://doi.org/10.1073/pnas.0504978102.
13. Brugère JF, Borrel G, Gaci N, Tottey W, O'Toole PW, Malpuech-Brugère C. 2014. Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. Gut Microbes 1:138–147. https://doi.org/10.4161/gmic.1.3.12360.
14. Joossens M, Huys G, Cnockaert M, De Preter V, Verbeke K, Rutgeerts P, Vandamme P, Vermeire S. 2011. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut 60:631–637. https://doi.org/10.1136/gut.2010.223263.
15. Buc E, Dubois D, Sauvanet P, Raisch J, Delmas J, Darfeuille-Michaud A, Pezet D, Bonnet R. 2013. High prevalence of mucosa-associated e. coli producing cyclomodulin and genotoxin in colon cancer. PLoS One 8:e56964. https://doi.org/10.1371/journal.pone.0056964.
16. Brennan CA, Garrett WS. 2019. Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. Nat Rev Microbiol 17:156–166. https://doi.org/10.1038/s41579-018-0129-6.
17. Wu S, Lim KC, Huang J, Saidi RF, Sears CL. 1998. Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin. Proc Natl Acad Sci U S A 95:14979–14984. https://doi.org/10.1073/pnas.95.25.14979.
18. Candela M, Turroni S, Biagi E, Carbonero F, Rampelli S, Fiorentini C, Brigidi P. 2014. Inflammation and colorectal cancer, when microbiota-host

mutualism breaks. World J Gastroenterol 20:908–922. https://doi.org/10.3748/wjg.v20.i4.908.
19. Albenberg LG, Wu GD. 2014. Diet and the intestinal microbiome: associations, functions, and implications for health and disease. Gastroenterology 146:1564–1572. https://doi.org/10.1053/j.gastro.2014.01.058.
20. Han S, Gao J, Zhou Q, Liu S, Wen C, Yang X. 2018. Role of intestinal flora in colorectal cancer from the metabolite perspective: a systematic review. Cancer Manag Res 10:199–206. https://doi.org/10.2147/CMAR.S153482.
21. Nguyen LH, Ma W, Wang DD, Cao Y, Mallick H, Gerbaba TK, Lloyd-Price J, Abu-Ali G, Hall AB, Sikavi D, Drew DA, Mehta RS, Arze C, Joshi AD, Yan Y, Branck T, DuLong C, Ivey KL, Ogino S, Rimm EB, Song M, Garrett WS, Izard J, Huttenhower C, Chan AT. 2020. Association between sulfur-metabolizing bacterial communities in stool and risk of distal colorectal cancer in men. Gastroenterology 158:1313–1325. https://doi.org/10.1053/j.gastro.2019.12.029.
22. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, Su L, Li X, Li X, Li J, Xiao L, Huber-Schönauer U, Niederseer D, Xu X, Al-Aama JY, Yang H, Wang J, Kristiansen K, Arumugam M, Tilg H, Datz C, Wang J. 2015. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 6:6528. https://doi.org/10.1038/ncomms7528.
23. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10:766. https://doi.org/10.15252/msb.20145645.
24. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, Goedert JJ, Shi J, Bork P, Sinha R. 2016. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. PLoS One 11:e0155362. https://doi.org/10.1371/journal.pone.0155362.
25. Ou J, Carbonero F, Zoetendal EG, DeLany JP, Wang M, Newton K, Gaskins HR, O'Keefe SJD. 2013. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. Am J Clin Nutr 98:111–120. https://doi.org/10.3945/ajcn.112.056689.
26. Louis P, Hold GL, Flint HJ. 2014. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol 12:661–672. https://doi.org/10.1038/nrmicro3344.
27. Gill CIR, Rowland IR. 2002. Diet and cancer: assessing the risk. Br J Nutr 88:s73–s87. https://doi.org/10.1079/BJN2002632.
28. Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F, O'Toole PW. 2018. The oral microbiota in colorectal cancer is distinctive and predictive. Gut 67:1454–1463. https://doi.org/10.1136/gutjnl-2017-314814.
29. Nakatsu G, Li X, Zhou H, Sheng J, Wong SH, Wu WKK, Ng SC, Tsoi H, Dong Y, Zhang N, He Y, Kang Q, Cao L, Wang K, Zhang J, Liang Q, Yu J, Sung JJY. 2015. Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat Commun 6:30–32. https://doi.org/10.1038/ncomms9727.
30. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, Holt RA. 2012. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res 22:299–306. https://doi.org/10.1101/gr.126516.111.
31. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M. 2012. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res 22:292–298. https://doi.org/10.1101/gr.126573.111.
32. Komiya Y, Shimomura Y, Higurashi T, Sugi Y, Arimoto J, Umezawa S, Uchiyama S, Matsumoto M, Nakajima A. 2019. Patients with colorectal cancer have identical strains of Fusobacterium nucleatum in their colorectal cancer and oral cavity. Gut 68:1335–1337. https://doi.org/10.1136/gutjnl-2018-316661.
33. Uchino Y, Goto Y, Konishi Y, Tanabe K, Toda H, Wada M, Kita Y, Beppu M, Mori S, Hijioka H, Otsuka T, Natsugoe S, Hara E, Sugiura T. 2021. Colorectal cancer patients have four specific bacterial species in oral and gut microbiota in common–a metagenomic comparison with healthy subjects. Cancers (Basel) 13:3332. https://doi.org/10.3390/cancers13133332.
34. Dai Z, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, Chan FKL, Kristiansen K, Sung JJY, Wong SH, Yu J. 2018. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome 6:70. https://doi.org/10.1186/s40168-018-0451-2.
35. Mallick H, Franzosa EA, McIver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier RJ, Huttenhower C. 2019. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun 10:1–11. https://doi.org/10.1038/s41467-019-10927-1.
36. Lieu EL, Nguyen T, Rhyne S, Kim J. 2020. Amino acids in cancer. Exp Mol Med 52:15–30. https://doi.org/10.1038/s12276-020-0375-3.
37. Qi H, Li Y, Yun H, Zhang T, Huang Y, Zhou J, Yan H, Wei J, Liu Y, Zhang Z, Gao Y, Che Y, Su X, Zhu D, Zhang Y, Zhong J, Yang R. 2019. Lactobacillus maintains healthy gut mucosa by producing L-Ornithine. Commun Biol 2:171. https://doi.org/10.1038/s42003-019-0424-4.
38. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, Hosoda F, Rokutan H, Matsumoto M, Takamaru H, Yamada M, Matsuda T, Iwasaki M, Yamaji T, Yachida T, Soga T, Kurokawa K, Toyoda A, Ogura Y, Hayashi T, Hatakeyama M, Nakagama H, Saito Y, Fukuda S, Shibata T, Yamada T. 2019. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 25:968–976. https://doi.org/10.1038/s41591-019-0458-7.
39. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 25:679–689. https://doi.org/10.1038/s41591-019-0406-6.
40. Layeghifard M, Li H, Wang PW, Donaldson SL, Coburn B, Clark ST, Caballero JD, Zhang Y, Tullis DE, Yau YCW, Waters V, Hwang DM, Guttman DS. 2019. Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. NPJ Biofilms Microbiomes 5:4. https://doi.org/10.1038/s41522-018-0077-y.
41. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, Segata N. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 25:667–678. https://doi.org/10.1038/s41591-019-0405-7.
42. Kaur H, Das C, Mande SS. 2017. In silico analysis of putrefaction pathways in bacteria and its implication in colorectal cancer. Front Microbiol 8:2166. https://doi.org/10.3389/fmicb.2017.02166.
43. Liang JQ, Li T, Nakatsu G, Chen Y-X, Yau TO, Chu E, Wong S, Szeto CH, Ng SC, Chan FKL, Fang J-Y, Sung JJY, Yu J. 2020. A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. Gut 69:1248–1257. https://doi.org/10.1136/gutjnl-2019-318532.
44. Canani RB, Di Costanzo M, Leone L, Pedata M, Meli R, Calignano A. 2011. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol 17:1519–1528. https://doi.org/10.3748/wjg.v17.i12.1519.
45. Zaytseva YY, Rychahou PG, Gulhati P, Elliott VA, Mustain WC, O'Connor K, Morris AJ, Sunkara M, Weiss HL, Lee EY, Evers BM. 2012. Inhibition of fatty acid synthase attenuates CD44-associated signaling and reduces metastasis in colorectal cancer. Cancer Res 72:1504–1517. https://doi.org/10.1158/0008-5472.CAN-11-4057.
46. Menendez JA, Lupu R. 2007. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. Nat Rev Cancer 7:763–777. https://doi.org/10.1038/nrc2222.
47. Dueck DA, Chan M, Tran K, Wong JT, Jay FT, Littman C, Stimpson R, Choy PC. 1996. The modulation of choline phosphoglyceride metabolism in human colon cancer. Mol Cell Biochem 162:97–103. https://doi.org/10.1007/BF00227535.
48. Sharon C, Baranwal S, Patel NJ, Rodriguez-Agudo D, Pandak WM, Majumdar APN, Krystal G, Patel BB. 2015. Inhibition of insulin-like growth factor receptor/AKT/mammalian target of rapamycin axis targets colorectal cancer stem cells by attenuating mevalonate-isoprenoid pathway in vitro and in vivo. Oncotarget 6:15332–15347. https://doi.org/10.18632/oncotarget.3684.
49. Sears CL, Geis AL, Housseau F. 2014. Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis. J Clin Invest 124:4166–4172. https://doi.org/10.1172/JCI72334.
50. Zakharzhevskaya NB, Vanyushkina AA, Altukhov IA, Shavarda AL, Butenko IO, Rakitina DV, Nikitina AS, Manolov AI, Egorova AN, Kulikov EE, Vishnyakov IE, Fisunov GY, Govorun VM. 2017. Outer membrane vesicles

secreted by pathogenic and nonpathogenic Bacteroides fragilis represent different metabolic activities. Sci Rep 7:5008. https://doi.org/10.1038/s41598-017-05264-6.
51. Wilson MM, Anderson DE, Bernstein HD. 2015. Analysis of the outer membrane proteome and secretome of Bacteroides fragilis reveals a multiplicity of secretion mechanisms. PLoS One 10:e0117732. https://doi.org/10.1371/journal.pone.0117732.
52. Yao L, Seaton SC, Ndousse-Fetter S, Adhikari AA, Dibenedetto N, Mina AI, Banks AS, Bry L, Devlin AS. 2018. A selective gut bacterial bile salt hydrolase alters host metabolism. Elife 7:e37182. https://doi.org/10.7554/eLife.37182.
53. den Besten G, Lange K, Havinga R, van Dijk TH, Gerding A, van Eunen K, Müller M, Groen AK, Hooiveld GJ, Bakker BM, Reijngoud DJ. 2013. Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. Am J Physiol - Gastrointest Liver Physiol 305:900–910. https://doi.org/10.1152/ajpgi.00265.2013.
54. Fernández-Veledo S, Vendrell J. 2019. Gut microbiota-derived succinate: friend or foe in human metabolic diseases? Rev Endocr Metab Disord 20:439–447. https://doi.org/10.1007/s11154-019-09513-z.
55. Ridlon JM, Kang DJ, Hylemon PB. 2006. Bile salt biotransformations by human intestinal bacteria. J Lipid Res 47:241–259. https://doi.org/10.1194/jlr.R500013-JLR200.
56. Yang Y, Misra BB, Liang L, Bi D, Weng W, Wu W, Cai S, Qin H, Goel A, Li X, Ma Y. 2019. Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. Theranostics 9:4101–4114. https://doi.org/10.7150/thno.35186.
57. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. 2013. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. PLoS One 8:e70803. https://doi.org/10.1371/journal.pone.0070803.
58. Casero RA, Murray Stewart T, Pegg AE. 2018. Polyamine metabolism and cancer: treatments, challenges and opportunities. Nat Rev Cancer 18:681–695. https://doi.org/10.1038/s41568-018-0050-3.
59. Brushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner (No. LBNL-7065E). Lawrence Berkeley National Lab (LBNL), Berkeley, CA (United States).
60. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12:902–903. https://doi.org/10.1038/nmeth.3589.
61. Peabody MA, Van Rossum T, Lo R, Brinkman FSL. 2015. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics 16:363. https://doi.org/10.1186/s12859-015-0788-5.
62. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60–18. https://doi.org/10.1186/gb-2011-12-s1-p47.
63. Friedman J, Alm EJ. 2012. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol 8:e1002687. https://doi.org/10.1371/journal.pcbi.1002687.
64. Jackson MA, Bonder MJ, Kuncheva Z, Zierer J, Fu J, Kurilshikov A, Wijmenga C, Zhernakova A, Bell JT, Spector TD, Steves CJ. 2018. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. PeerJ 6:e4303. https://doi.org/10.7717/peerj.4303.
65. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 1971. Cytoscape: a software environment for integrated models. Genome Res 13:426.
66. Goenawan IH, Bryan K, Lynn DJ. 2016. DyNet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics 32:2713–2715. https://doi.org/10.1093/bioinformatics/btw187.
67. Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor. J Stat Softw 36:1–48.
68. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 13:435–438. https://doi.org/10.1038/nmeth.3802.
69. Bateman A. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515.
70. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616.
71. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47. https://doi.org/10.1093/nar/gkv007.
72. Pang Z, Chong J, Zhou G, De Lima Morais DA, Chang L, Barrette M, Gauthier C, Jacques PÉ, Li S, Xia J. 2021. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Res 49:W388–W396. https://doi.org/10.1093/nar/gkab382.
73. R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 10:11–18.