



OPEN ACCESS

EDITED BY
Pradip Pradip Behare,
National Dairy Research Institute (ICAR), India

REVIEWED BY
Swarna Kanchan,
Marshall University, United States
Yihua Wang,
Shandong University, China

*CORRESPONDENCE
Huizhen Xiong
✉ xionghzh3@mail.sysu.edu.cn
Guangyi Li
✉ guangyili@hnu.edu.cn
Honglei Chen
✉ chl_830830@163.com

[†]These authors have contributed equally to this work and share first authorship

[†]These authors have contributed equally to this work and share second authorship

RECEIVED 29 June 2025

ACCEPTED 01 September 2025

PUBLISHED 24 September 2025

CITATION

Li Q, Liu F, Zhong J, Fang X, Zhang X, Xiong H, Li G and Chen H (2025) Multi-cohort metagenomics reveals strain functional heterogeneity and demonstrates fecal microbial load correction improves colorectal cancer diagnostic models.
Front. Microbiol. 16:1656016.
doi: 10.3389/fmicb.2025.1656016

COPYRIGHT

© 2025 Li, Liu, Zhong, Fang, Zhang, Xiong, Li and Chen. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Multi-cohort metagenomics reveals strain functional heterogeneity and demonstrates fecal microbial load correction improves colorectal cancer diagnostic models

Qiucheng Li^{1†}, Fang Liu^{1†}, Jianfeng Zhong^{1†}, Xiaoling Fang^{1†}, Xinyi Zhang^{1†}, Huizhen Xiong^{1*}, Guangyi Li^{2*} and Honglei Chen^{1*}

¹Digestive Endoscopy Center, The Eighth Affiliated Hospital, Sun Yat-sen University, Shenzhen, China,

²College of Biology, Hunan University, Changsha, China

Introduction: Colorectal cancer (CRC) is strongly associated with alterations in the gut microbiome. While numerous studies have examined this association, most focus on genus- or species-level taxonomic classifications, overlooking functional heterogeneity at the strain level.

Methods: We integrated 1,123 metagenomic samples from seven global CRC cohorts to conduct multi-level metagenome-wide association studies (MWAS). Fecal microbial load (FML) correction was applied to mitigate technical confounding. We evaluated the performance of taxonomic models at various resolutions strain, species, and genus levels in classifying CRC status both within and across cohorts.

Results: Strain-level analysis revealed conspecific strains with divergent associations to CRC. For instance, distinct strains of *Bacteroides thetaiotaomicron* exhibited both protective and risk-increasing effects across different cohorts. Genomic functional annotation suggested potential mechanistic bases for these opposing roles. Correction for FML reduced confounding and significantly improved the performance of within-cohort and cross-cohort CRC classification models. Interestingly, genus- and species-level models demonstrated superior predictive robustness compared to strain-level models, likely due to higher microbial abundance and greater cross-population conservation at these taxonomic ranks.

Conclusion: Our study underscores the biological relevance of strain level analysis in elucidating functional diversity within the microbiome. However, higher taxonomic levels provide more robust and clinically translatable diagnostic markers for CRC. Integrating FML correction with multi-level taxonomic profiling enhances both mechanistic insight into microbiome CRC interactions and the generalizability of diagnostic models across diverse populations.

KEYWORDS

colorectal cancer, gut microbiome, strain, metagenomics, fecal microbial load, classification models

Introduction

Colorectal cancer (CRC), comprising approximately 10% of all cancer cases worldwide, stands as the second leading cause of cancer-related mortality (Sung et al., 2021). Due to its typically asymptomatic early stages, CRC is often diagnosed at advanced phases when therapeutic options are limited. Accumulating evidence over recent years has firmly established a critical association between CRC development and the human gut microbiome, with interactions primarily mediated through mechanisms such as microbial metabolism, inflammatory regulation, immune dysbalance, and intestinal barrier dysfunction (White and Sears, 2024; Wong and Yu, 2023). Studies have further highlighted the potential of gut microbes as diagnostic biomarkers, demonstrating high accuracy in predicting gastrointestinal diseases including CRC and inflammatory bowel disease (IBD) (Wong et al., 2017; Wang and Jia, 2016).

Notably, most existing research has focused on taxonomic analysis at the genus or species level, overlooking the functional heterogeneity that may exist among different strains within the same genus or species. Strains from the same microbial species can exhibit divergent phenotypes or even opposing biological functions in host environments. For example, *Escherichia coli* (*E. coli*), a common commensal in mammalian intestines, includes the probiotic strain Nissle 1917, which synthesizes essential vitamins, alongside highly pathogenic variants like *E. coli* STEC O26:H11 and EHEC O104:H4, associated with hemolytic uremic syndrome and fatal diarrhea (Marx, 2016; Bonanno et al., 2015). Similarly, distinct strains of *Staphylococcus aureus* and *Streptococcus pyogenes* elicit markedly different immune responses in humans (Sela et al., 2018; Van Rossum et al., 2020).

Advances in metagenomic sequencing depth and high-resolution taxonomic profiling tools have enabled strain-level metagenome-wide association studies (MWAS), providing new avenues to dissect microbiome functional characteristics (Shi et al., 2022; Olm et al., 2021). However, a gap remains in strain-resolved MWAS of CRC, particularly regarding systematic cross-cohort comparisons across diverse geographical populations.

Additionally, fecal microbial load (FML), an important factor influencing microbial composition analysis, has gained increasing attention. A recent study demonstrated that neglecting FML correction can lead to spurious associations between microbial taxa and diseases, with effect sizes and significance metrics changing substantially after load adjustment (Nishijima et al., 2025). This suggests FML represents a potential confounder in MWAS, yet its impact on disease classification model performance—especially across taxonomic levels (strain, species, and genus) and geographical cohorts—remains uncharacterized.

To address these knowledge gaps, our study integrated 1,123 samples from seven independent CRC cohorts across seven countries, employing a standardized analytical pipeline to conduct systematic MWAS at strain, species, and genus levels. We evaluated the effects of FML correction on the identification of disease-associated microbial features and the performance of classification models for CRC. Through multi-cohort, multi-level comparisons, we aimed to: (1) demonstrate the unique value of strain-level analysis in resolving biological heterogeneity, (2) assess the robustness of genus/species-level features in diagnostic models

for clinical utility, and (3) determine whether FML correction enhances disease prediction performance. Our findings provide critical insights for mechanistic studies of the gut microbiome in CRC and its application in early clinical diagnosis.

Materials and methods

Cohort selection

We collected published fecal whole metagenome sequencing (WMS) data from seven cohorts consisting of CRC patients and healthy controls, spanning seven countries. Raw sequencing data for these samples were downloaded from the Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) using the following accession IDs: ERP008729 (AUT cohort) from Feng et al. (2015), PRJEB10878. (CHI cohort) from Yu et al. (2017), PRJNA531273 and PRJNA397112 (IND cohort) from Gupta et al. (2019), ERP005534 (FRA cohort) from Zeller et al. (2014), SRP136711 (ITA cohort) from Thomas et al. (2019), PRJEB12449 (USA cohort) from Vogtmann et al. (2016), and DRA006684/DRA008156 (JPN cohort) from Yachida et al. (2019). Metadata were manually curated from original studies, excluding samples with missing Age, BMI, or Gender information, resulting in 1,123 samples. Only colorectal cancer and healthy control samples were included in downstream analyses, excluding adenoma cases.

Sample preprocessing and metagenomic profiling

For raw sequencing data preprocessing, KneadData (<https://github.com/biobakery/kneaddata>, V0.12.0) was used for quality control and host contamination removal. Trimmomatic (V0.39, integrated in KneadData) performed sequence quality filtering and adapter trimming with parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50. Host-derived reads were removed by aligning to the human reference genome (GRCh38_p14) using Bowtie2 (V2.4.1) with parameters: `-very-sensitive -dovetail -reorder`.

Strain-level abundance analysis of preprocessed sequences was conducted using Sylph (V0.6.1) (Shaw and Yu, 2024) against a custom non-redundant strain database (c200_gtdb_strain.sylldb, compression parameter $c = 200$). Genomes were downloaded from the Genome Taxonomy Database (GTDB) using genome_updater (https://github.com/pirovc/genome_updater), with a limit of 100 genomes per species to constrain computational costs, yielding 343,362 strains. For each species, pairwise average nucleotide identity (ANI) matrices were calculated by FastANI (v1.33) (Jain et al., 2018), followed by custom graph-based clustering at ANI thresholds of 95%–99.9%. Here, the 95% ANI threshold is widely used for microbial species delineation to differentiate interspecies boundaries (Jain et al., 2018; Konstantinidis and Tiedje, 2005; Goris et al., 2007), while the 99.9% upper bound is designed to capture intraspecies strain-level genetic variations, preventing the loss of biologically meaningful diversity due to overly stringent

thresholds. The refined database contained 206,273 strains (GTDB:206273).

Species-level analysis utilized MetaPhlAn4 (V4.1.1) (Blanco-Míguez et al., 2023) with the mpa_vJan21_CHOCOPhAnSGB_202103 reference database. Taxonomic results from both tools were merged at genus, species, and strain levels using MetaPhlAn's merge_metaphlan_tables.py script for downstream differential analysis.

Fecal microbial load prediction

Fecal microbial load (total microbial cells per gram or cell density) was predicted using the Microbial Load Predictor (MLP, <https://microbiome-tools.embl.de/mlp/>) (Nishijima et al., 2025), an R-based computational tool designed to estimate FML from species-level taxonomic profiles of the human gut microbiome. Input files consisted of species-level taxonomic feature tables generated by classification tools including mOTUs v2.5, mOTUs v3.0, MetaPhlAn3, MetaPhlAn4, or RDP-based 16S rRNA annotations. Given the demonstrated congruence between species-level classifications from MetaPhlAn4 and Sylph, we utilized the metagenomic profiling outputs of MetaPhlAn4 to predict the fecal microbial load.

Microbiome diversity and community structure analysis

Alpha-diversity metrics (Shannon index, Richness) were calculated using the vegan package (V2.6-8) (Oksanen, 2022). Multivariate linear regression models [lm() function] analyzed covariate effects on diversity, reporting coefficients, standard errors, and p-values with residual diagnostics. Group differences were tested via Wilcoxon rank-sum tests. Beta-diversity was assessed using Bray-Curtis distance-based permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2014) via adonis2 in vegan, evaluating independent contributions of covariates (Disease, Age, Gender, BMI, and FML) with 999 permutations. Pairwise comparisons of significant variables (e.g., Disease) used pairwise.adonis() from the pairwiseAdonis package, with p-values corrected for false discovery rate (FDR) via Benjamini-Hochberg. These analyses were conducted using R scripts.

Training-test set partitioning

Samples from each country were partitioned into training and test sets at an 8:2 ratio, with the random grouping process repeated 100 times to construct diverse datasets. Stratified sampling was employed in each partition to ensure balanced class representation in the test set, maintaining proportional distribution of colorectal cancer (CRC) and non-CRC cases. To preserve consistency between the feature matrix (microbial abundance data) and metadata, microbial abundance matrices were extracted according to the partition results. All partitioned datasets (including metadata

and microbial profiles) were stored in a predefined directory structure for subsequent model training and validation.

Differential abundance analysis

Multivariate Association with Linear Models 2 (MaAsLin2, V1.20.0) (Mallick et al., 2020) was used to identify microbial features associated with CRC status. We utilized OTU abundance tables generated by Sylph and MetaPhlAn4, combined with sample metadata including disease status, age, gender, BMI, and total fecal microbial load (FML). Two model types were constructed: one excluding FML as a covariate and another including it to assess its regulatory effect. Linear regression models were applied with log-transformed feature data, using raw relative abundances (where the sum of relative abundances for each taxonomic level OTU in a single sample equals 1) without normalization. Features were filtered to require a minimum occurrence frequency of 10%, and multiple hypothesis testing was corrected using the Benjamini-Hochberg method with a significance threshold FDR < 0.25.

Functional annotation of strain genomes

Functional annotation of bacterial genomes was performed using three databases: VFDB (Zhou et al., 2025), CARD (Alcock et al., 2023), and KEGG (Kanehisa and Goto, 2000). For VFDB, Abricate (V1.0.1) (Seemann, 2017) was applied to genome FASTA files to identify virulence factors, with a minimum sequence identity and coverage set to 50%. For CARD, Resistance Gene Identifier (RGI, V6.0.4) (Alcock et al., 2023) was used with the CARD database to predict antibiotic resistance genes, using contig input and DIAMOND for sequence alignment. For KEGG, gene prediction and general functional annotation were performed with Prokka (V1.14.6) (Seemann, 2014), followed by functional annotation of predicted protein sequences and KEGG pathway assignment using EGGNOG-mapper (V2.0.1) (Cantalapiedra et al., 2021). Differential pathway enrichment was evaluated by Fisher's exact test with FDR-adjusted p-values.

Batch effect correction

Microbial abundance data were first filtered to remove low-abundance OTUs (retaining only OTUs present in at least 10% of samples at the genus and species levels; for strains, due to low abundance and high specificity, OTUs present in at least 1% of samples were retained) and then matched with sample metadata. To correct for cohort-associated batch effects while preserving disease-related signals, compositional data were subjected to centered log-ratio (CLR) transformation, which is suitable for handling the compositionality, sparsity, and skewness of microbiome data. Batch effects were subsequently adjusted using the ComBat (implemented in the sva R package, V3.54.0) (Johnson et al., 2007) method, with cohort as the batch variable and disease status as a covariate. The effectiveness of batch correction was assessed using PERMANOVA, and PCA was applied to visualize

the data before and after correction. The batch-corrected relative abundance data were then used for downstream analyses.

Disease classifier construction and validation

Random Forest (RF) (Breiman, 2001) models were employed in conjunction with Recursive Feature Elimination with Cross-Validation (RFECV) and hyperparameter optimization to evaluate the classification performance of microbiome features in predicting colorectal cancer (CRC). Two independent modeling approaches were implemented based on feature subsets derived from differential abundance analysis (with and without fecal microbial load correction). For each modeling approach, RFECV was first applied to the training set to identify stable discriminatory features. Hyperparameters of the RF models were then optimized via randomized search (RandomizedSearchCV() function in Python package of sklearn) over a predefined parameter grid. Model performance was evaluated on the test set, and confidence intervals for the Area Under the Receiver Operating Characteristic Curve (AUC) were estimated using bootstrap resampling. Pairwise comparisons of AUC distributions were conducted using the non-parametric Mann-Whitney U test: (i) between models incorporating vs. excluding total microbial load adjustment, and (ii) across taxonomic levels (genus, species, and strain). Statistical significance cutoff was set as $P \leq 0.05$. Cross-cohort external validation employed a leave-one-country-out strategy, where models were trained on data from one country and independently validated on each of the remaining six countries. This approach ensured that each national cohort served sequentially as the training set, with the other six cohorts functioning as distinct validation sets to assess model generalizability across geographical populations. Only features retained during training were used in validation; missing features were imputed with zeros to maintain dimensional consistency. If all required features were absent in a validation set, the corresponding model-validation combination was excluded.

Results

Cohort characteristics and multilevel microbiota diversity analysis

Raw metagenomic sequencing data from all 1,123 samples across seven independent cohorts were first processed using a uniform standardized pipeline to ensure comparability, involving quality control, adapter trimming, and metagenomic profiling (detailed in Methods, see Figure 1 and Supplementary Table 1). Strain-level analysis leveraged the metagenomic classifier Sylph to construct a non-redundant reference genome database from the Genome Taxonomy Database (GTDB), while genus- and species-level taxonomic annotations were performed using MetaPhlAn4 and Sylph, respectively. Fecal microbial load (FML) was estimated via the Microbial Load Predictor pipeline (MLP, <https://microbiome-tools.embl.de/mlp/>) (Nishijima et al., 2025), which requires input in mOTU or MetaPhlAn format; given the

demonstrated consistency between Sylph and MetaPhlAn4 at the species annotation level, we utilized MetaPhlAn4 outputs for FML estimation and included load as a covariate in subsequent statistical models.

In the subsequent experiments, to mitigate biases from random dataset partitioning, each cohort was stratified and randomly divided into training (80%) and test (20%) sets using an 8:2 ratio, with this process repeated 100 times. Training data were used for identifying differential microbial features via MaAsLin2 and constructing random forest classifiers, while test sets enabled within-cohort validation of model performance.

Analysis of within-sample diversity (α -diversity) revealed heterogeneous patterns across geographical cohorts (Figure 2A). In Indian and Austrian samples, CRC cases exhibited significantly higher Shannon diversity and richness indices than controls at the genus, species, and strain levels ($P < 0.05$ for all comparisons), whereas most other cohorts showed minimal or non-significant differences in microbial diversity between groups. These findings highlight the geographical dependency and complexity of CRC-associated gut microbiota alterations. Linear regression models further showed that FML exerted a significant effect on both Shannon and richness metrics across all taxonomic levels in all cohorts except India ($P < 0.05$, Supplementary Tables 2, 3), underscoring the need to account for microbial load when interpreting diversity indices.

As for β -diversity analysis, principal coordinate analysis (PCoA) based on Bray-Curtis dissimilarity mirrored α -diversity trends, with notable distinctions in microbial community structure between CRC cases and controls (Figure 2B). Significant β -diversity differences (permutation test, $P < 0.05$) were observed at all taxonomic levels in most cohorts, except for Italian and US samples, indicating region- or population-specific shifts in gut microbiota associated with CRC. When FML was incorporated as a covariate in permutational multivariate analysis of variance (PERMANOVA), it emerged as a significant factor influencing community structure across all taxonomic levels and cohorts ($P < 0.05$, Supplementary Tables 4, 5), confirming its role as a critical confounder in microbiome compositional analyses.

Contrasting effects of conspecific strains in colorectal cancer

Previous studies have referred to this metric as strain richness (SR), defined as the number of strains of a given microbial species j present in the gut of an individual i , denoted as SR_{ij} . Typically, an individual harbors no more than two strains per species. To distinguish this from the “richness” measure used in Result 1, we hereafter denote this concept as strain number (SN) (Chen-Liaw et al., 2025). In our analysis, the proportion of samples with $SN \geq 2$ across all cohorts was consistently below 15%, and 91% of 1,123 total samples exhibited fewer than 20% species with $SN \geq 2$. The majority of $SN \geq 2$ cases (overwhelmingly $SN = 2$) aligned with historical observations (Figure 3A, Supplementary Table 6). Using MaAsLin2, we identified strain-level associations with CRC across cohorts and observed striking functional dichotomy within species. In all cohorts except the USA, two or more strains belonging

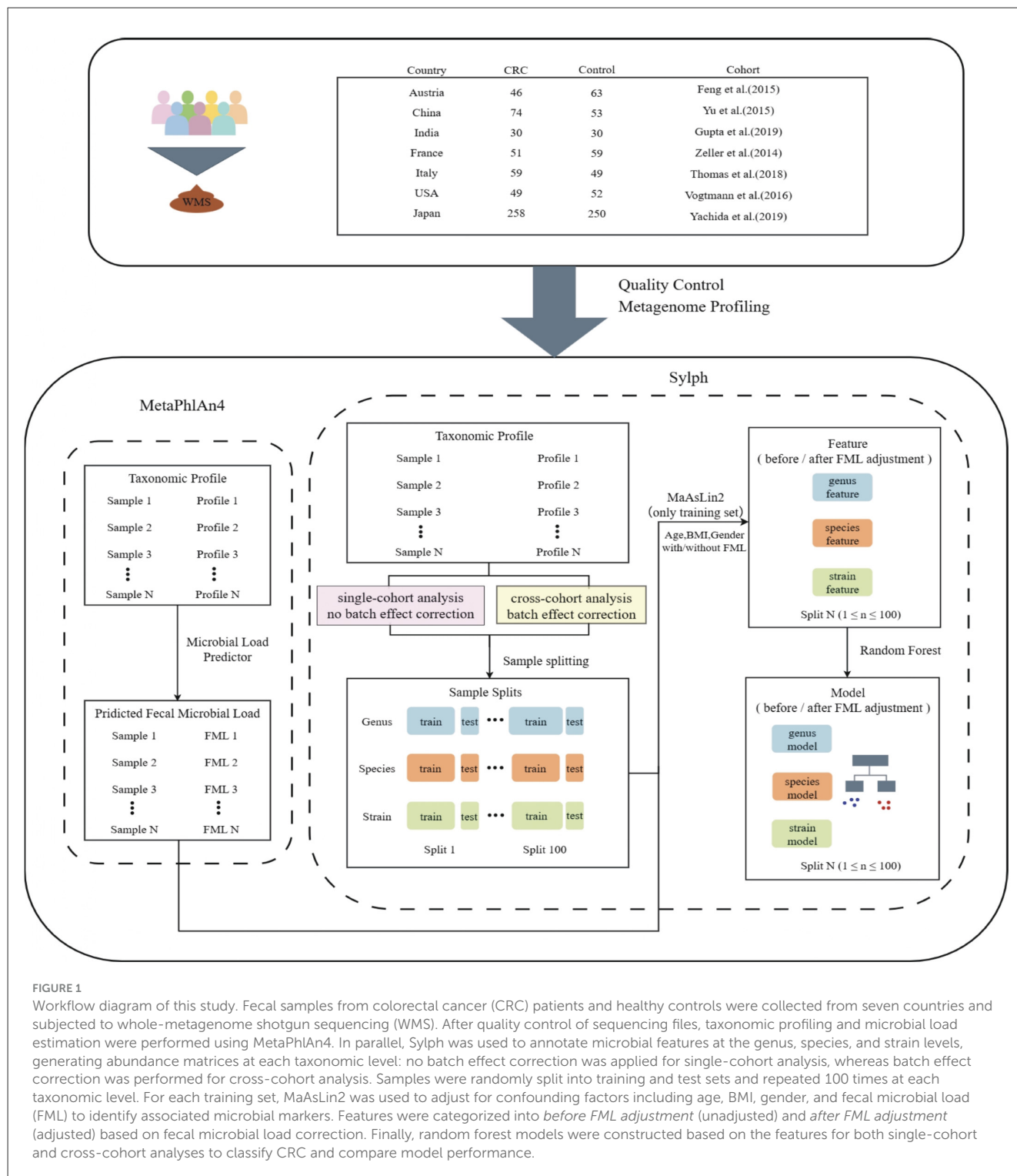


FIGURE 1

Workflow diagram of this study. Fecal samples from colorectal cancer (CRC) patients and healthy controls were collected from seven countries and subjected to whole-metagenome shotgun sequencing (WMS). After quality control of sequencing files, taxonomic profiling and microbial load estimation were performed using MetaPhlAn4. In parallel, Sylph was used to annotate microbial features at the genus, species, and strain levels, generating abundance matrices at each taxonomic level: no batch effect correction was applied for single-cohort analysis, whereas batch effect correction was performed for cross-cohort analysis. Samples were randomly split into training and test sets and repeated 100 times at each taxonomic level. For each training set, MaAsLin2 was used to adjust for confounding factors including age, BMI, gender, and fecal microbial load (FML) to identify associated microbial markers. Features were categorized into *before FML adjustment* (unadjusted) and *after FML adjustment* (adjusted) based on fecal microbial load correction. Finally, random forest models were constructed based on the features for both single-cohort and cross-cohort analyses to classify CRC and compare model performance.

to the same species were detected. Notably, in the Indian and Japanese cohorts, certain species such as *Vescimonas* sp900555735, *Avimicrobium caecorum*, *Bacteroides thetaiotaomicron*, and *Dorea formicigenerans* exhibited opposing strain-level associations with CRC risk, as indicated by their regression coefficients (Figure 3B, Supplementary Table 7).

To further explore the biological mechanisms underlying opposing effects on CRC among strains from the same species,

we performed functional annotation of their genomes using VFDB (Virulence Factor Database) (Zhou et al., 2025), CARD (Comprehensive Antibiotic Resistance Database) (Alcock et al., 2023), and KEGG (metabolic pathway database) (Kanehisa and Goto, 2000). We compared “risk strains” and “protective strains” in terms of metabolic pathways and virulence gene content.

Based on presence/absence data of genes as annotated in the VFDB and CARD (Supplementary Tables 8, 9), the results show

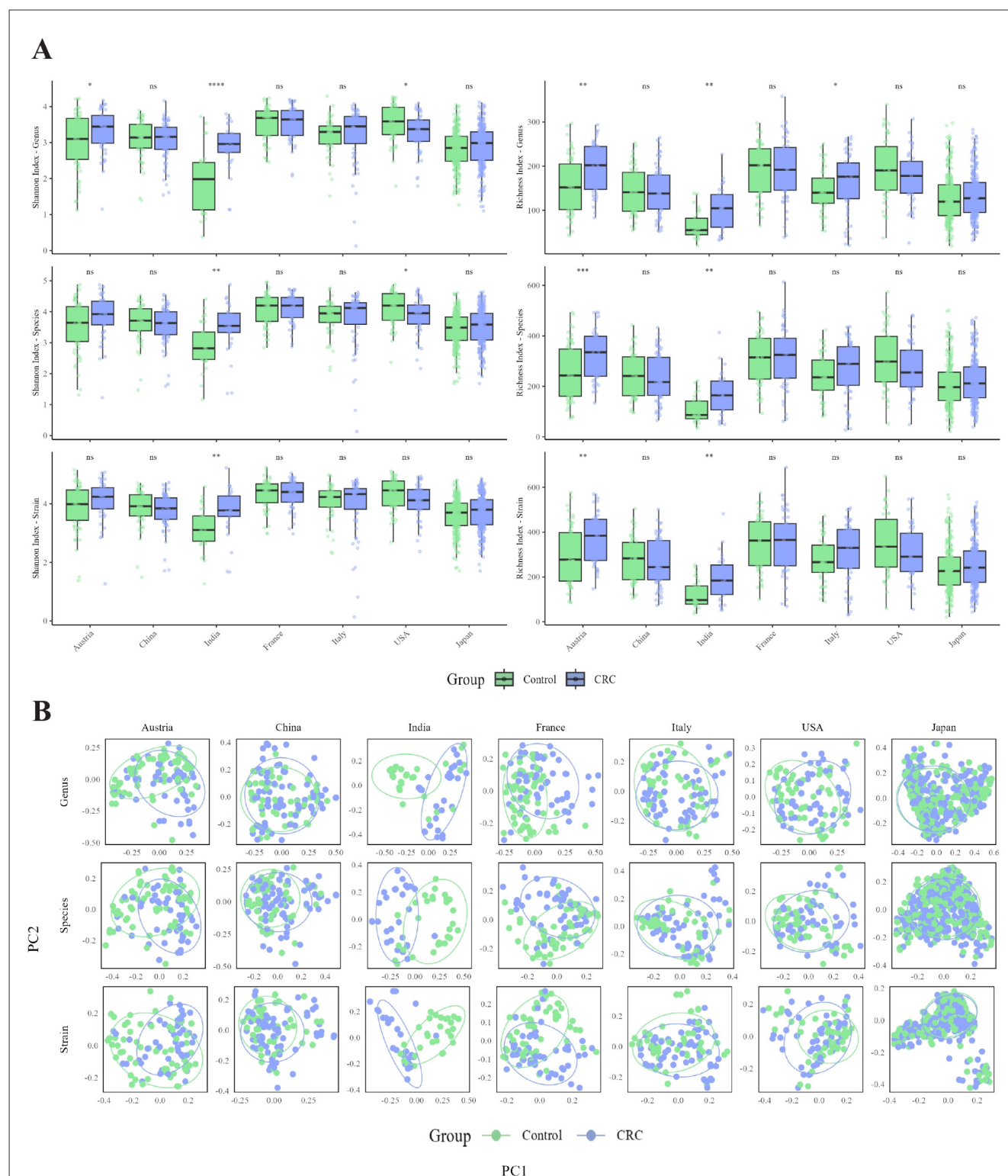
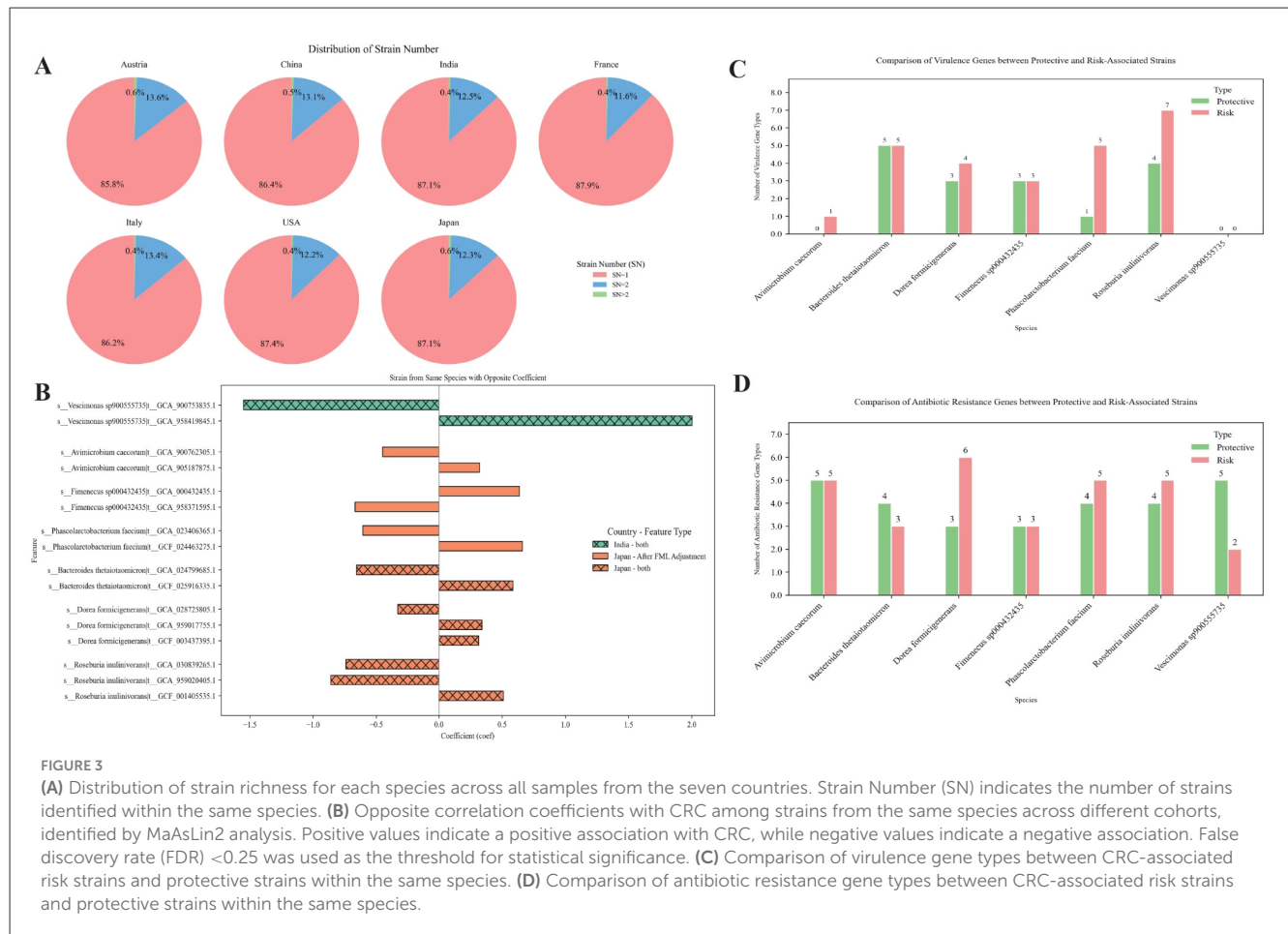


FIGURE 2

(A) Comparison of alpha diversity indices (Shannon and Richness) between CRC and control groups across different taxonomic levels (genus, species, strain) in each country. The Shannon index reflects both the richness and evenness of taxa, while the Richness index represents the number of unique taxa. Statistical significance was assessed using the non-parametric Wilcoxon rank-sum test. Significance levels: **** $p < 0.0001$; *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns, not significant. (B) Principal component analysis (PCA) of genus-, species-, and strain-level microbial profiles in CRC and control samples from different countries. Each subplot shows the distribution of samples along the first two principal components (PC1 and PC2) with 95% confidence ellipses for each group.



that risk strains generally carry a more diverse set of virulence factor genes (VFGs) and antibiotic resistance genes (ARGs) (Figures 3C, D). Specifically, VFDB data showed that risk strains carry VFGs including *flpP* and *rfaD*. *flpP* is involved in flagellar protein transport and assembly and may induce chronic inflammation via the TLR5/NF- κ B pathway (Song et al., 2017). *rfaD* is involved in the biosynthesis of bacterial lipopolysaccharide (LPS). LPS can activate the host immune system and trigger inflammatory responses; under chronic inflammatory conditions, LPS may continuously stimulate the intestinal immune system, thereby leading to persistent inflammation (Zhang et al., 2024). Other VFGs, such as *cps4I* and *pseB*, may promote bacterial colonization, immune evasion, and host inflammatory responses. Although these genes are not classical oncogenes, their presence in the gut microbiota or specific risk strains may indicate a microenvironment conducive to low-grade chronic inflammation, which is a known risk factor for CRC (Shimomura et al., 2023). CARD data showed that ARGs such as *ACI-1*, *dfrF*, and *tet(O/W/40)* may reflect community structures associated with dysbiosis and increased microbial activity.

In the KEGG pathway analysis, we compared the enrichment of pathways between risk strains and protective strains (Supplementary Table 10). Considering the high ANI similarity among strains of the same species, the FDR threshold was relaxed to 0.25 to avoid missing potentially relevant differences. The results showed that *map00540*, which is related to lipopolysaccharide biosynthesis, may induce chronic inflammation via the

TLR4/NF- κ B signaling pathway (Hu et al., 2021; Luo and Zhang, 2017), thereby increasing CRC risk, and *map05111*, which is related to *Vibrio cholerae* infection, may contribute to carcinogenesis by sustaining chronic infection and local inflammation (Ou et al., 2009). By contrast, *map00511*, which is related to other glycan degradation, may help maintain normal glycosylation levels and reduce abnormal glycosylation-associated changes in cell adhesion and metastasis (Bangarh et al., 2023), whereas *map00600*, which is related to sphingolipid metabolism, may modulate the ceramide-S1P balance (Karmelić et al., 2024), promoting apoptosis and exerting anti-inflammatory effects.

These findings highlight the utility of strain-level analysis in resolving microbiome functional heterogeneity, which is obscured at the species level. The identification of conspecific strains with diametrically opposed effects on CRC provides a rational basis for developing strain-targeted therapeutic interventions. Although we proposed hypotheses for some key pathways and genes, other pathways and genes not detailed here may also be involved in microbiota-mediated inflammation and carcinogenesis, and future studies are needed to clarify their specific roles and mechanisms.

FML correction improves performance of disease classifiers

Significant features selected by MaAsLin2 at the strain, species, and genus levels from the training set were input into classification

models, and the test sets were used for within-cohort validation of model performance. Across cohorts, fecal microbial load (FML) correction increased the number of detected differential features in all groups except Italy, with genus- and species-level features consistently outnumbering strain-level equivalents (Figure 4A, Supplementary Table 11). To address feature count variability, recursive feature elimination with cross-validation (RFECV) was applied to select subsets optimizing area under the receiver operating characteristic curve (AUC).

We systematically assessed model performance using multiple metrics, including AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. At the strain level, models constructed from FML-corrected features consistently outperformed their uncorrected counterparts across all cohorts, although the degree of statistical significance varied (Figure 4B, Supplementary Table 12). Similar trends were observed at the genus and species levels, where FML normalization consistently enhanced discriminative accuracy.

To further validate the reliability of conclusions derived from single-cohort models, we additionally performed cross-cohort validation. Batch effects were first corrected, and their influence on disease status at the genus, species, and strain levels was reduced from 0.123, 0.073, 0.058 to 0.016, 0.017, 0.010, respectively (Supplementary Table 13). We then applied principal component analysis (PCA) to assess clustering patterns before and after correction, and observed a markedly increased overlap across cohorts after correction, indicating that batch effects had been effectively controlled (Figure 4C). Based on the corrected data, we re-performed differential feature analysis and classifier construction, with cross-cohort results further corroborating the robustness of our initial conclusions (Figure 4D, Supplementary Table 14).

Higher taxonomic levels outperform strain-level disease classifiers

We evaluated the impact of taxonomic resolution on classifier performance by comparing models built from genus-, species-, and strain-level features across multiple metrics. In all cohorts except the FML-corrected Indian dataset, genus- and species-level models showed significantly better performance in AUC compared to strain-level models, regardless of microbial load adjustment (Figure 5). This trend was also observed in cross-cohort validation, where genus- and species-level models demonstrated greater generalizability across different geographic populations. All complete evaluation metrics, including accuracy, sensitivity, specificity, PPV, NPV, and F1 score, are provided in the shared Supplementary Table 12.

Two putative mechanisms may explain this observation: (1) Strain-level features typically exhibit lower relative abundances than higher taxonomic levels, leading to increased technical noise and reduced signal-to-noise ratios during metagenomic profiling; and (2) Strain-specific markers are highly influenced by individual host backgrounds (e.g., genetics, lifestyle) and geographical factors, limiting their transferability across diverse cohorts (Andreu-Sánchez et al., 2025). These findings highlight that while strain-level

analysis uncovers biological heterogeneity, genus-, and species-level features offer more robust and reproducible signals for clinical diagnostic applications, balancing mechanistic insight with practical utility in multi-cohort settings.

Discussion

This study leveraged 1,123 metagenomic samples from seven independent cohorts to systematically evaluate gut microbiome associations with colorectal cancer (CRC) across taxonomic levels (genus, species, and strain). We also investigated the impact of fecal microbial load (FML) correction on disease classification model performance and compared predictive capabilities across taxonomic resolutions.

Most cohorts showed higher Shannon diversity and richness indices in CRC patients compared to healthy controls, aligning with prior studies suggesting increased microbiota diversity in CRC (Figure 2A). However, this trend was not statistically significant in all cohorts, likely reflecting complex influences of population background, geographical environment, and study design on microbial community structure. Beta-diversity analysis revealed limited significant compositional differences between CRC cases and controls, implying that pronounced microbiome structural alterations may primarily occur in advanced disease stages (Figure 2B).

We observed that the majority of samples harbored no more than two strains per species (Figure 3A), and across multiple cohorts, distinct strains within the same species exhibited opposing associations with CRC risk (Figure 3B). This finding underscores strain-level functional heterogeneity, where conspecific strains can influence host health through divergent mechanisms. Strain-level analysis provides finer biological resolution than higher taxonomic levels, enabling identification of potentially pathogenic or protective strains. Furthermore, by functionally annotating the genomes of risk and protective strains within the same species, we proposed hypotheses regarding their potential biological mechanisms.

Fecal microbial load (FML) correction significantly improved both within-cohort and cross-cohort predictive performance of CRC classifiers (Figures 4B, D). Uncorrected models may be confounded by total microbial biomass, which distorts relative abundance measurements and masks true biological signals. Load adjustment mitigates this confound, allowing models to more accurately identify CRC-associated microbial features. These results advocate for routine inclusion of FML as a covariate in metagenomic analyses.

While strain-level analysis offers high biological resolution, genus- and species-level models outperformed strain-resolved counterparts in predictive accuracy (Figure 5). This discrepancy arises from: (1) lower strain-level abundances and associated technical noise, reducing signal-to-noise ratios; and (2) high inter-individual and geographical specificity of strains, limiting cross-cohort reproducibility. In contrast, genus/species-level features exhibit greater conservation across populations, making them more robust for clinical prediction. These findings highlight the need to balance research objectives when selecting taxonomic resolution:

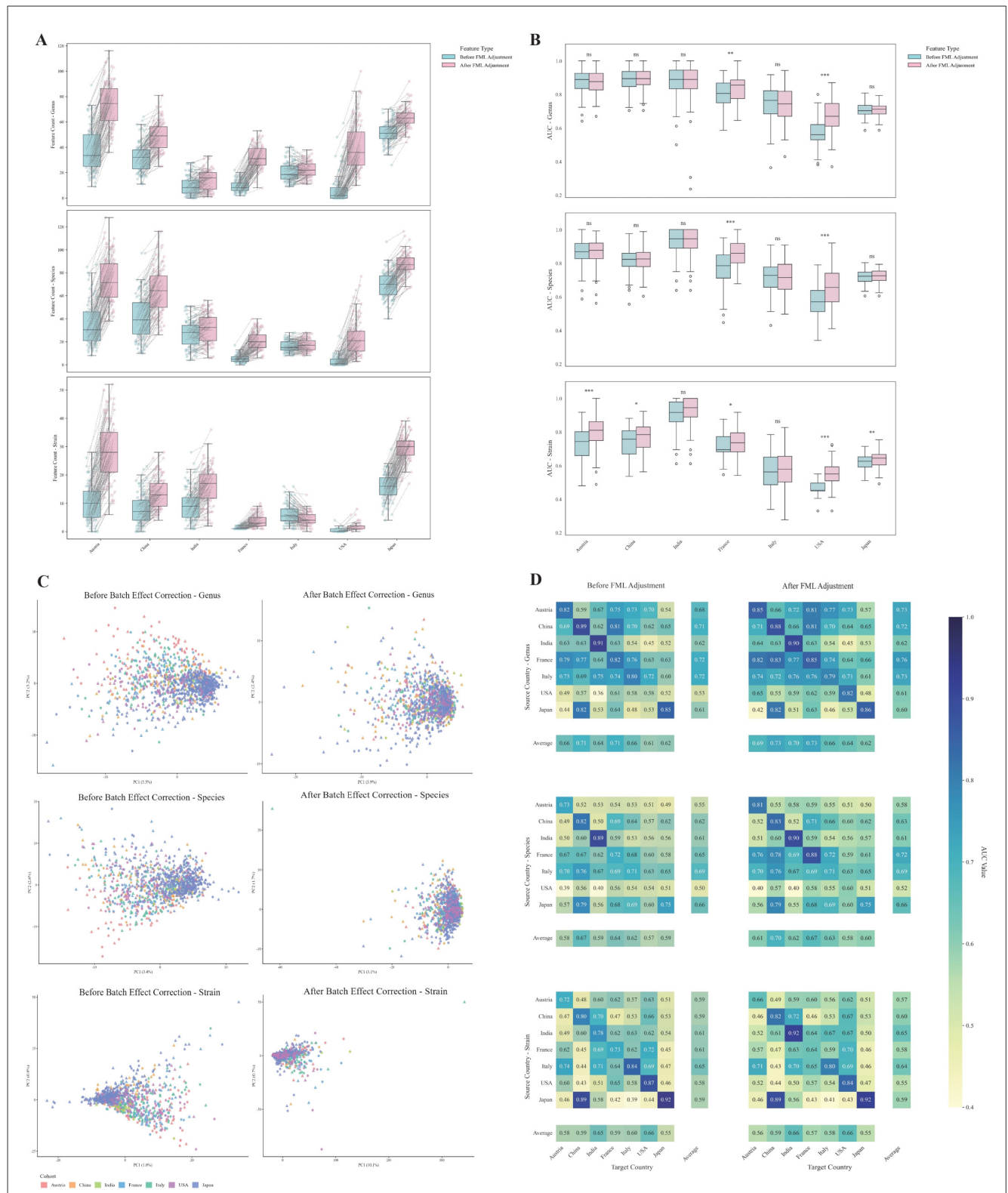
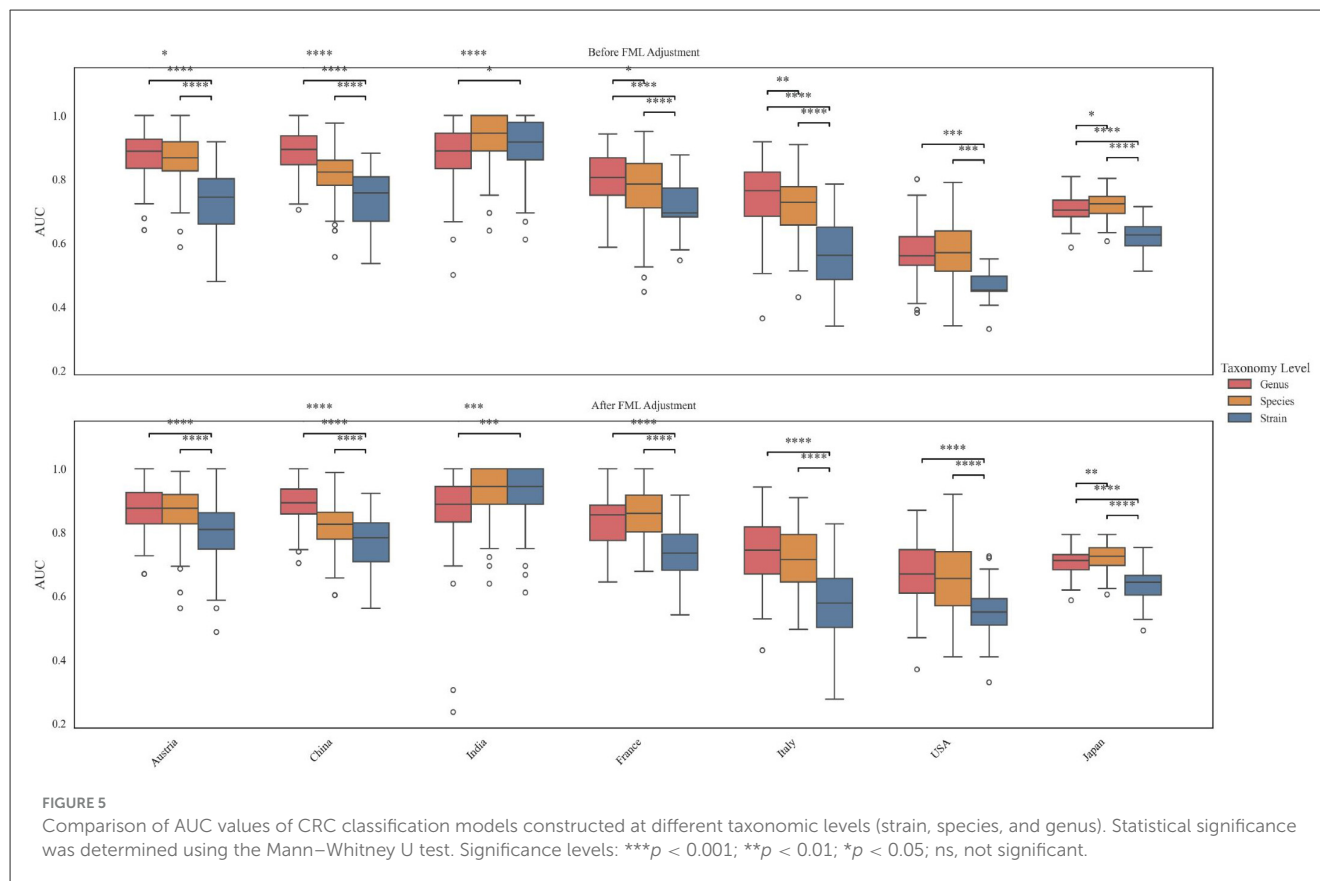


FIGURE 4

(A) Number of differential features between CRC and control groups identified by MaAsLin2 at the genus, species, and strain levels before and after microbial load adjustment across countries. (B) Comparison of classification performance (area under the ROC curve, AUC) of CRC prediction models before and after microbial load adjustment across countries. Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns, not significant. Statistical significance was assessed using the non-parametric Mann-Whitney U test. (C) PCA of the samples after batch effect correction. Each point represents a sample, and colors correspond to different study cohorts. (D) Cross-cohort validation performance of CRC classification models before and after microbial load adjustment.



strain-level analysis for mechanistic insights, versus higher levels for stable diagnostic markers.

This work has several limitations: although we proposed hypotheses for some key pathways and genes, their specific roles and mechanisms still need to be further clarified; host genetic or clinical variables were not integrated into the models; and the findings are associative rather than causal. Future studies could leverage metagenome-assembled genomes (MAGs) and culturomics to characterize strain functions, and integrate multi-omics data to develop more comprehensive predictive models.

Conclusion

Through multi-cohort integrative analysis, this study reveals taxon-level specificities in gut microbiome–CRC associations. Strain-level analysis uncovers functional heterogeneity invisible at higher taxonomic scales, but genus/species-level features currently offer greater stability for clinical translation. To further advance the field, future studies should leverage metagenome-assembled genomes (MAGs) and culturomics to better characterize strain functions, and integrate multi-omics data to develop more comprehensive and robust predictive models. Ultimately, improving high-sensitivity strain detection and functional validation methods will be critical for translating strain-resolved microbiome insights into precision medicine.

Data availability statement

Cohort metadata, analysis results, and code are available at Zenodo (<https://doi.org/10.5281/zenodo.16899759>) and GitHub (https://github.com/gygyli/CRC_analysis).

Ethics statement

The studies involving humans were approved by this study did not involve new data collection from human participants. All data analyzed were obtained from previously published studies and publicly available databases (e.g., SRA, ENA), where ethical approval had been granted by the respective original authors' institutional review boards. Therefore, additional ethical approval was not required by our institution. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

QL: Writing – original draft, Writing – review & editing, Formal analysis. FL: Writing – original draft, Formal analysis. JZ: Writing – original draft, Formal analysis. XF: Writing – original draft, Formal analysis. XZ: Writing – original draft, Formal

analysis. HX: Writing – review & editing, Writing – original draft, Conceptualization. GL: Writing – original draft, Writing – review & editing, Conceptualization, Formal analysis. HC: Writing – review & editing, Writing – original draft, Conceptualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the 2024 Futian Healthcare Research Project (Grant Nos.: FTWS053 and FTWS094).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

References

- Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., et al. (2023). Card 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 51, D690–D699. doi: 10.1093/nar/gkac920
- Anderson, M. J. (2014). “Permutational multivariate analysis of variance (permanova),” in *Wiley Statsref: Statistics Reference Online*, 1–15.
- Andreu-Sánchez, S., Blanco-Míguez, A., Wang, D., Golzato, D., Manghi, P., Heidrich, V., et al. (2025). Global genetic diversity of human gut microbiome species is related to geographic location and host health. *Cell* 188, 3942–3959. doi: 10.1016/j.cell.2025.04.014
- Bangarh, R., Khatana, C., Kaur, S., Sharma, A., Kaushal, A., Siwal, S. S., et al. (2023). Aberrant protein glycosylation: Implications on diagnosis and immunotherapy. *Biotechnol. Adv.* 66, 108149. doi: 10.1016/j.biotechadv.2023.108149
- Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4. *Nat. Biotechnol.* 41, 1633–1644. doi: 10.1038/s41587-023-01688-w
- Bonanno, L., Loukiadis, E., Mariani-Kurkdjian, P., Oswald, E., Garnier, L., Michel, V., et al. (2015). Diversity of Shiga toxin-producing *Escherichia coli* (stec) o26: H11 strains examined via stx subtypes and insertion sites of STX and ESPK bacteriophages. *Appl. Environ. Microbiol.* 81, 3712–3721. doi: 10.1128/AEM.00077-15
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). EGGNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293
- Chen-Liaw, A., Aggarwala, V., Mogno, I., Haifer, C., Li, Z., Eggers, J., et al. (2025). Gut microbiota strain richness is species specific and affects engraftment. *Nature* 637, 422–429. doi: 10.1038/s41586-024-08242-x
- Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6:6528. doi: 10.1038/ncomms7528
- Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0
- Gupta, A., Dhakan, D. B., Maji, A., Saxena, R., PK, V. P., Mahajan, S., et al. (2019). Association of flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *MSystems* 4, 10–1128. doi: 10.1128/msystems.00438-19
- Hu, L., Liu, Y., Kong, X., Wu, R., Peng, Q., Zhang, Y., et al. (2021). Fusobacterium nucleatum facilitates m2 macrophage polarization and colorectal carcinoma progression by activating tlr4/nf- κ b/s100a9 cascade. *Front. Immunol.* 12:658681. doi: 10.3389/fimmu.2021.658681
- Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9
- Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kjx037
- Kanehisa, M., and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27
- Karmelić, I., Jurilj Sajko, M., Sajko, T., Rotim, K., and Fabris, D. (2024). The role of sphingolipid rheostat in the adult-type diffuse glioma pathogenesis. *Front. Cell Dev. Biol.* 12:1466141. doi: 10.3389/fcell.2024.1466141
- Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Nat. Acad. Sci.* 102, 2567–2572. doi: 10.1073/pnas.0409727102
- Luo, C., and Zhang, H. (2017). The role of proinflammatory pathways in the pathogenesis of colitis-associated colorectal cancer. *Mediat. Inflamm.* 2017:5126048. doi: 10.1155/2017/5126048
- Mallick, H., Rahnavard, A., and McIver, L. (2020). *Maaslin2: Maaslin2*. R package version 1(0).
- Marx, V. (2016). Microbiology: the road to strain-level identification. *Nat. Methods* 13, 401–404. doi: 10.1038/nmeth.3837
- Nishijima, S., Stankevici, E., Aasmets, O., Schmidt, T. S., Nagata, N., Keller, M. I., et al. (2025). Fecal microbial load is a major determinant of gut microbiome variation and a confounder for disease associations. *Cell* 188, 222–236. doi: 10.1016/j.cell.2024.10.022
- Oksanen, J. (2022). *_vegan: Community ecology package_*. r package version 2.6–4. (No Title).
- Olm, M. R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B. A., Morowitz, M. J., and Banfield, J. F. (2021). Instrain profiles population microdiversity from

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2025.1656016/full#supplementary-material>

- metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* 39, 727–736. doi: 10.1038/s41587-020-00797-0
- Ou, G., Rompikuntal, P. K., Bitar, A., Lindmark, B., Vaitkevicius, K., Wai, S. N., et al. (2009). *Vibrio cholerae* cytotoxin causes an inflammatory response in human intestinal epithelial cells that is modulated by the PRTV protease. *PLoS ONE* 4:e7806. doi: 10.1371/journal.pone.0007806
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Seemann, T. (2017). *Abricate: Mass screening of contigs for antimicrobial and virulence genes*. Available online at: <https://github.com/tseemann/abricate> (Accessed August 18, 2025).
- Sela, U., Euler, C. W., Correa da Rosa, J., and Fischetti, V. A. (2018). Strains of bacterial species induce a greatly varied acute adaptive immune response: the contribution of the accessory genome. *PLoS Pathog.* 14:e1006726. doi: 10.1371/journal.ppat.1006726
- Shaw, J., and Yu, Y. W. (2024). Rapid species-level metagenome profiling and containment estimation with sylph. *Nat. Biotechnol.* 43, 1348–1359. doi: 10.52843/cassyni.gq2yrj
- Shi, Z. J., Dimitrov, B., Zhao, C., Nayfach, S., and Pollard, K. S. (2022). Fast and accurate metagenotyping of the human gut microbiome with gt-pro. *Nat. Biotechnol.* 40, 507–516. doi: 10.1038/s41587-021-01102-3
- Shimomura, K., Hattori, N., Iida, N., Muranaka, Y., Sato, K., Shiraiishi, Y., et al. (2023). Sleeping beauty transposon mutagenesis identified genes and pathways involved in inflammation-associated colon tumor development. *Nat. Commun.* 14:6514. doi: 10.1038/s41467-023-42228-z
- Song, W., Jeon, Y., Namgung, B., Hong, M., and Yoon, S. (2017). A conserved tlr5 binding and activation hot spot on flagellin. *Sci. Rep.* 7:40878. doi: 10.1038/srep40878
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660
- Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7
- Van Rossum, T., Ferretti, P., Maistrenko, O. M., and Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* 18, 491–506. doi: 10.1038/s41579-020-0368-1
- Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., et al. (2016). Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS ONE* 11:e0155362. doi: 10.1371/journal.pone.0155362
- Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-tuning the microbiome. *Nat. Rev. Microbiol.* 14, 508–522. doi: 10.1038/nrmicro.2016.83
- White, M. T., and Sears, C. L. (2024). The microbial landscape of colorectal cancer. *Nat. Rev. Microbiol.* 22, 240–254. doi: 10.1038/s41579-023-00973-4
- Wong, C. C., and Yu, J. (2023). Gut microbiota in colorectal cancer development and therapy. *Nat. Rev. Clin. Oncol.* 20, 429–452. doi: 10.1038/s41571-023-00766-x
- Wong, S. H., Kwong, T. N., Chow, T.-C., Luk, A. K., Dai, R. Z., Nakatsu, G., et al. (2017). Quantitation of faecal fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. *Gut* 66, 1441–1448. doi: 10.1136/gutjnl-2016-312766
- Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7
- Yu, J., Feng, Q., Wong, S. H., Zhang, D., Yi Liang, Q., Qin, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/gutjnl-2015-309800
- Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645
- Zhang, X., Tian, X., Wang, Y., Yan, Y., Wang, Y., Su, M., et al. (2024). Application of lipopolysaccharide in establishing inflammatory models. *Int. J. Biol. Macromol.* 279:135371. doi: 10.1016/j.ijbiomac.2024.135371
- Zhou, S., Liu, B., Zheng, D., Chen, L., and Yang, J. (2025). Vifdb 2025: an integrated resource for exploring anti-virulence compounds. *Nucleic Acids Res.* 53, D871–D877. doi: 10.1093/nar/gkae968